

Computational Biology: Genomes, Networks, Evolution  
MIT course 6.047/6.878

Taught by Prof. Manolis Kellis

September 6, 2018



## TODO LIST

This should be in its own chapter . . . . .	9
EXPAND . . . . .	60
EXPAND . . . . .	60
EXPAND . . . . .	60
missing . . . . .	185
Add reference to this figure. . . . .	187
add picture of ribosome density vs. protein abundance . . . . .	189
add reference to Figures 5-8 . . . . .	190
expand . . . . .	251
MAKE A NEW INTRO: This intro is used to introduce networks and the concept of networks. Please refer to chapter 12B: Networks . . . . .	291
This section is in chapter 12B now . . . . .	300
Scribe: Expand - maybe talk about folding and free energy? . . . . .	309
Should this chapter be included? . . . . .	325
missing . . . . .	341
Reference . . . . .	341
figure . . . . .	344
figure . . . . .	345
missing . . . . .	347
SCRIBE: Elaborate upon intricacies of the model itself: <a href="http://www.ncbi.nlm.nih.gov/pubmed/21270390">http://www.ncbi.nlm.nih.gov/pubmed/21270390</a> . . . . .	401
Find a figure to replace missing figure . . . . .	462
Find a figure to replace missing figure . . . . .	462
Find a figure to replace missing figure . . . . .	463
Find a figure to replace missing figure . . . . .	463
Find a figure to replace missing figure . . . . .	464
Find a figure to replace missing figure . . . . .	464
Find a figure to replace missing figure . . . . .	465



## CONTENTS

<b>1</b>	<b>Introduction to the Course</b>	<b>5</b>
1.1	Introduction and Goals . . . . .	5
1.1.1	A course on computational biology . . . . .	5
1.1.2	Duality of Goals: Foundations and Frontiers . . . . .	5
1.1.3	Duality of disciplines: Computation and Biology . . . . .	6
1.1.4	Why Computational Biology? . . . . .	6
1.1.5	Finding Functional Elements: A Computational Biology Question . . . . .	8
1.2	Final Project - Introduction to Research In Computational Biology . . . . .	9
1.2.1	Final project goals . . . . .	9
1.2.2	Final project milestones . . . . .	9
1.2.3	Project deliverables . . . . .	10
1.2.4	Project grading . . . . .	10
1.3	Additional materials . . . . .	11
1.3.1	Online Materials for Fall 2016 . . . . .	11
1.3.2	Textbooks . . . . .	11
1.4	Crash Course in Molecular Biology . . . . .	11
1.4.1	The Central Dogma of Molecular Biology . . . . .	11
1.4.2	DNA . . . . .	12
1.4.3	Transcription . . . . .	13
1.4.4	RNA . . . . .	14
1.4.5	Translation . . . . .	15
1.4.6	Protein . . . . .	16
1.4.7	Regulation: from Molecules to Life . . . . .	16
1.4.8	Metabolism . . . . .	17
1.4.9	Systems Biology . . . . .	18
1.4.10	Synthetic Biology . . . . .	18
1.4.11	Model organisms and human biology . . . . .	19
1.5	Introduction to algorithms and probabilistic inference . . . . .	20
1.5.1	Probability distributions . . . . .	21
1.5.2	Graphical probabilistic models . . . . .	21
1.5.3	Bayes rules: priors, likelihood, posterior . . . . .	21
1.5.4	Markov Chains and Sequential Models . . . . .	21
1.5.5	Probabilistic inference and learning . . . . .	21
1.5.6	Max Likelihood and Max A Posteriori Estimates . . . . .	21
<b>I</b>	<b>Comparing Genomes</b>	<b>23</b>
<b>2</b>	<b>Sequence Alignment and Dynamic Programming</b>	<b>25</b>
2.1	Introduction . . . . .	25

2.2	Aligning Sequences . . . . .	26
2.2.1	Example Alignment . . . . .	26
2.2.2	Solving Sequence Alignment . . . . .	26
2.3	Dynamic Programming . . . . .	28
2.3.1	Theory of Dynamic Programming . . . . .	28
2.3.2	Fibonacci Numbers . . . . .	29
2.3.3	Sequence Alignment using Dynamic Programming . . . . .	31
2.4	Problem Formulations . . . . .	31
2.4.1	Formulation 1: Longest Common Substring . . . . .	32
2.4.2	Formulation 2: Longest Common Subsequence (LCS) . . . . .	32
2.4.3	Formulation 3: Sequence Alignment as Edit Distance . . . . .	33
2.4.4	Formulation 4: Varying Gap Cost Models . . . . .	34
2.4.5	Enumeration . . . . .	34
2.5	The Needleman-Wunsch Algorithm . . . . .	34
2.5.1	Memoization vs tabulation . . . . .	35
2.5.2	Problem Statement . . . . .	35
2.5.3	Index space of subproblems . . . . .	35
2.5.4	Local optimality . . . . .	35
2.5.5	Optimal Solution . . . . .	36
2.5.6	Solution Analysis . . . . .	36
2.5.7	Needleman-Wunsch in practice . . . . .	36
2.5.8	Optimizations . . . . .	38
2.6	Multiple alignment . . . . .	40
2.6.1	Aligning three sequences . . . . .	40
2.6.2	Heuristic multiple alignment . . . . .	40
2.7	Current Research Directions . . . . .	41
2.8	Further Reading . . . . .	41
2.9	Tools and Techniques . . . . .	41
2.10	What Have We Learned? . . . . .	41
2.11	Appendix . . . . .	41
2.11.1	Homology . . . . .	41
2.11.2	Natural Selection . . . . .	42
2.11.3	Dynamic Programming v. Greedy Algorithms . . . . .	42
2.11.4	Pseudocode for the Needleman-Wunsch Algorithm . . . . .	43
<b>3</b>	<b>Rapid Sequence Alignment and Database Search</b> . . . . .	<b>45</b>
3.1	Introduction . . . . .	46
3.2	Global alignment vs. Local alignment vs. Semi-global alignment . . . . .	47
3.2.1	Using Dynamic Programming for local alignments . . . . .	48
3.2.2	Algorithmic Variations . . . . .	49
3.2.3	Generalized gap penalties . . . . .	51
3.3	Linear-time exact string matching . . . . .	51
3.3.1	Karp-Rabin Algorithm . . . . .	52
3.4	The BLAST algorithm (Basic Local Alignment Search Tool) . . . . .	54
3.4.1	The BLAST algorithm . . . . .	55
3.4.2	Extensions to BLAST . . . . .	56
3.5	Probabilistic Foundations of Sequence Alignment . . . . .	56
3.6	Deterministic linear-time exact string matching . . . . .	58
3.6.1	Pre-processing . . . . .	58
3.6.2	String matching algorithms . . . . .	60
3.7	Current Research Directions . . . . .	62
3.8	Further Reading . . . . .	62
3.9	Tools and Techniques . . . . .	63
3.10	What Have We Learned? . . . . .	63

<b>4 Comparative Genomics I: Genome Annotation</b>	<b>65</b>
4.1 Introduction . . . . .	66
4.1.1 Motivation and Challenge . . . . .	66
4.1.2 Importance of many closely-related genomes . . . . .	67
4.1.3 Comparative genomics and evolutionary signatures . . . . .	68
4.2 Conservation of genomic sequences . . . . .	69
4.2.1 Functional elements in <i>Drosophila</i> . . . . .	69
4.2.2 Rates and patterns of selection . . . . .	69
4.3 Excess Constraint . . . . .	70
4.3.1 Causes of Excess Constraint . . . . .	71
4.3.2 Modeling Excess Constraint . . . . .	72
4.3.3 Excess Constraint in the Human Genome . . . . .	73
4.3.4 Examples of Excess Constraint . . . . .	74
4.3.5 Measuring constraint at individual nucleotides . . . . .	74
4.4 Diversity of Evolutionary Signatures: An Overview of Selection Patterns . . . . .	75
4.4.1 Selective Pressures On Different Functional Elements . . . . .	75
4.5 Protein-Coding Signatures . . . . .	77
4.5.1 Reading-Frame Conservation (RFC) . . . . .	78
4.5.2 Codon-Substitution Frequencies (CSFs) . . . . .	79
4.5.3 Classification of <i>Drosophila</i> Genome Sequences . . . . .	82
4.5.4 Leaky Stop Codons . . . . .	82
4.5.5 Overlapping Selection . . . . .	83
4.6 microRNA (miRNA) Gene Signatures . . . . .	83
4.6.1 Computational Challenge . . . . .	84
4.6.2 Unusual miRNA Genes . . . . .	84
4.6.3 Example: Re-examining 'dubious' protein-coding genes . . . . .	85
4.7 Regulatory Motifs . . . . .	85
4.7.1 Computationally Detecting Regulatory Motifs . . . . .	85
4.7.2 Individual Instances of Regulatory Motifs . . . . .	86
4.8 Further Reading . . . . .	86
4.9 Tools and Techniques . . . . .	86
4.10 Bibliography . . . . .	86
<b>5 Genome Assembly and Alignment</b>	<b>99</b>
5.1 Introduction . . . . .	101
5.2 Genome Assembly I: Overlap-Layout-Consensus Approach . . . . .	101
5.2.1 Setting up the experiment . . . . .	101
5.2.2 Finding overlapping reads . . . . .	104
5.2.3 Merging reads into contigs . . . . .	104
5.2.4 Laying out contig graph into scaffolds . . . . .	106
5.2.5 Deriving consensus sequence . . . . .	107
5.3 Genome Assembly II: String graph methods . . . . .	108
5.3.1 String graph definition and construction . . . . .	108
5.3.2 Flows and graph consistency . . . . .	110
5.3.3 Feasible flow . . . . .	110
5.3.4 Dealing with sequencing errors . . . . .	111
5.3.5 Resources . . . . .	111
5.4 Whole-Genome Alignment . . . . .	111
5.4.1 Global, local, and 'glocal' alignment . . . . .	111
5.4.2 Lagan: Chaining local alignments . . . . .	113
5.5 Gene-based region alignment . . . . .	114
5.6 Mechanisms of Genome Evolution . . . . .	116
5.6.1 Regions of Rapid Change . . . . .	117
5.6.2 Chromosomal Rearrangements . . . . .	118

5.6.3	Rapid Protein Change . . . . .	118
5.7	Whole Genome Duplication . . . . .	119
5.8	Additional figures . . . . .	121
<b>6</b>	<b>Bacterial Genomics– Molecular Evolution at the Level of Ecosystems</b>	<b>125</b>
6.1	Introduction . . . . .	125
6.1.1	Evolution of microbiome research . . . . .	126
6.1.2	Data generation for microbiome research . . . . .	126
6.2	Study 1: Evolution of life on earth . . . . .	126
6.3	Study 2: Pediatric IBD study with Athos Boudvaros . . . . .	127
6.4	Study 3: Human Gut Ecology (HuGE) project . . . . .	128
6.5	Study 4: Microbiome as the connection between diet and phenotype . . . . .	132
6.6	Study 5: Horizontal Gene Transfer (HGT) between bacterial groups and its effect on antibiotic resistance . . . . .	133
6.7	Study 6: Identifying virulence factors in Meningitis . . . . .	133
6.8	Q/A . . . . .	135
6.9	Current research directions . . . . .	136
6.10	Further Reading . . . . .	136
6.11	Tools and techniques . . . . .	136
6.12	What have we learned? . . . . .	136
<b>II</b>	<b>Coding and Non-Coding Genes</b>	<b>139</b>
<b>7</b>	<b>Hidden Markov Models I</b>	<b>141</b>
7.1	Introduction . . . . .	141
7.2	Motivation: . . . . .	142
7.2.1	We have a new sequence of DNA, now what? . . . . .	142
7.2.2	Why probabilistic sequence modeling? . . . . .	143
7.3	Markov Chains and HMMs . . . . .	143
7.3.1	Motivating Example: Weather Prediction . . . . .	143
7.3.2	Formalizing the Markov Chain and HMM . . . . .	143
7.4	Applications of HMMs: From the Casino to Biology . . . . .	145
7.4.1	The Dishonest Casino . . . . .	145
7.4.2	Back to Biology . . . . .	148
7.5	Algorithmic Settings for HMMs . . . . .	150
7.5.1	Scoring . . . . .	151
7.5.2	Decoding . . . . .	152
7.5.3	Evaluation . . . . .	154
7.6	An Interesting Question: Can We Incorporate Memory in Our Model? . . . . .	155
7.7	Further Reading . . . . .	156
7.7.1	Length Distributions of States and Generalized Hidden Markov Models . . . . .	156
7.7.2	Conditional random fields . . . . .	157
7.8	Current Research Directions . . . . .	157
7.9	Tools for implementing HMMs . . . . .	157
7.10	What Have We Learned? . . . . .	157
<b>8</b>	<b>Hidden Markov Models II - Posterior Decoding and Learning</b>	<b>159</b>
8.1	Review of previous lecture . . . . .	159
8.1.1	Introduction to Hidden Markov Models . . . . .	159
8.1.2	Genomic Applications of HMMs . . . . .	160
8.1.3	Viterbi decoding . . . . .	161
8.1.4	Forward Algorithm . . . . .	162
8.1.5	This lecture . . . . .	163

8.2	Posterior Decoding . . . . .	164
8.2.1	Motivation . . . . .	164
8.2.2	Backward Algorithm . . . . .	164
8.2.3	The Big Picture . . . . .	166
8.3	Encoding Memory in a HMM: Detection of CpG islands . . . . .	167
8.4	Learning . . . . .	169
8.4.1	Supervised Learning . . . . .	170
8.4.2	Unsupervised Learning . . . . .	170
8.5	Using HMMs to align sequences with affine gap penalties . . . . .	173
8.6	Current Research Directions . . . . .	174
8.7	Further Reading . . . . .	176
8.8	Tools and Techniques . . . . .	176
8.9	What Have We Learned? . . . . .	176
<b>9</b>	<b>Gene Identification: Gene Structure, Semi-Markov, CRFs</b>	<b>177</b>
9.1	Introduction . . . . .	177
9.2	Overview of Chapter Contents . . . . .	178
9.3	Eukaryotic Genes: An Introduction . . . . .	178
9.4	Assumptions for Computational Gene Identification . . . . .	178
9.5	Hidden Markov Models . . . . .	179
9.6	Conditional Random Fields . . . . .	180
9.7	Other Methods . . . . .	181
9.8	Conclusion . . . . .	182
9.8.1	HMM . . . . .	182
9.8.2	CRF . . . . .	182
9.9	Current Research Directions . . . . .	183
9.10	Further Reading . . . . .	183
9.11	Tools and Techniques . . . . .	183
9.12	What Have We Learned? . . . . .	183
<b>10</b>	<b>RNA Modifications</b>	<b>185</b>
10.1	Introduction . . . . .	185
10.2	Post-Transcriptional Regulation . . . . .	187
10.2.1	Basics of Protein Translation . . . . .	187
10.2.2	Measuring Translation . . . . .	188
10.2.3	Codon Evolution . . . . .	190
10.2.4	Translational Regulation . . . . .	191
10.3	Current Research Directions . . . . .	192
10.4	Further Reading . . . . .	192
10.5	Tools and Techniques . . . . .	192
10.6	What Have We Learned? . . . . .	192
<b>11</b>	<b>Large Intergenic non-Coding RNAs</b>	<b>193</b>
11.1	Introduction . . . . .	193
11.2	Noncoding RNAs from Plants to Mammals . . . . .	194
11.2.1	Long non-coding RNAs . . . . .	194
11.3	Practical topic: RNAseq . . . . .	195
11.3.1	How it works . . . . .	195
11.3.2	Aligning RNA-Seq reads to genomes and transcriptomes . . . . .	195
11.3.3	Calculating expression of genes and transcripts . . . . .	197
11.3.4	Differential analysis with RNA-Seq . . . . .	198
11.4	Long non-coding RNAs in Epigenetic Regulation . . . . .	199
11.5	Intergenic Non-coding RNAs: missing lincs in Stem/Cancer cells? . . . . .	201
11.5.1	An example: XIST . . . . .	201

11.6 Technologies: in the wet lab, how can we find these? . . . . .	201
11.6.1 Example: p53 . . . . .	201
11.7 Current Research Directions . . . . .	202
11.8 Further Reading . . . . .	202
11.9 Tools and Techniques . . . . .	202
11.10 What Have We Learned? . . . . .	202
<b>III Gene and Genome Regulation</b>	<b>205</b>
<b>12 mRNA sequencing for Expression Analysis and Transcript discovery</b>	<b>207</b>
12.1 Introduction . . . . .	207
12.2 Functional Diversity of RNA . . . . .	208
12.3 Microarrays for Gene Expression . . . . .	208
12.4 mRNA Sequencing . . . . .	208
12.4.1 RNA-seq . . . . .	208
12.4.2 Read Mapping . . . . .	208
12.4.3 RNA Transcript Reconstruction - Map to Genome Approach . . . . .	211
12.4.4 Quantification . . . . .	213
12.4.5 The Align-de-novo Approach . . . . .	214
12.5 Differential Expression Analysis . . . . .	216
<b>13 Gene Regulation 1 –Gene Expression Clustering</b>	<b>219</b>
13.1 Introduction . . . . .	220
13.1.1 Clustering vs Classification . . . . .	220
13.1.2 Applications . . . . .	220
13.2 Methods for Measuring Gene Expression . . . . .	221
13.2.1 Microarrays . . . . .	221
13.2.2 RNA-seq . . . . .	222
13.2.3 Gene Expression Matrices . . . . .	222
13.3 Clustering Algorithms . . . . .	224
13.3.1 <i>K</i> -Means Clustering (Clustering by Partitioning) . . . . .	224
13.3.2 Fuzzy <i>K</i> -Means Clustering . . . . .	225
13.3.3 <i>K</i> -Means as a Generative Model . . . . .	226
13.3.4 Expectation Maximization . . . . .	227
13.3.5 The limitations of the <i>K</i> -Means algorithm . . . . .	227
13.3.6 Hierarchical Clustering (Clustering by Agglomeration) . . . . .	228
13.3.7 Evaluating Cluster Performance . . . . .	229
13.4 Current Research Directions . . . . .	229
13.5 Further Reading . . . . .	230
13.6 Resources . . . . .	230
13.7 What Have We Learned? . . . . .	230
<b>14 Gene Regulation 2 –Classification</b>	<b>233</b>
14.1 Introduction . . . . .	233
14.2 Classification - Bayesian Techniques . . . . .	233
14.2.1 Single Feature and Bayes? Rule . . . . .	234
14.2.2 Multiple features and Naïve Bayes . . . . .	235
14.2.3 Collecting Data . . . . .	236
14.2.4 Estimating Priors . . . . .	236
14.2.5 Testing a classifier . . . . .	237
14.2.6 MAESTRO ? Mitochondrial Protein Classification . . . . .	237
14.3 Classification - Support Vector Machines . . . . .	238
14.3.1 Kernels . . . . .	238

14.3.2 Open Problems . . . . .	239
14.4 Tumor Classification with SVMs . . . . .	240
14.5 Semi-Supervised Learning . . . . .	240
14.6 Current Research Directions . . . . .	241
14.7 Further Reading . . . . .	241
14.8 Resources . . . . .	241
<b>15 Regulatory Motifs, Gibbs Sampling, and EM</b> . . . . .	<b>243</b>
15.1 Introduction to regulatory motifs and gene regulation . . . . .	243
15.1.1 The regulatory code: Transcription Factors and Motifs . . . . .	244
15.1.2 Challenges of motif discovery . . . . .	244
15.1.3 Motifs summarize TF sequence specificity . . . . .	245
15.2 Expectation maximization . . . . .	246
15.2.1 The key idea behind EM . . . . .	246
15.2.2 The E step: Estimating $Z_{ij}$ from the PWM . . . . .	247
15.2.3 M step: Finding the maximum likelihood motif from starting positions $Z_{ij}$ . . . . .	248
15.3 Gibbs Sampling: Sample from joint (M,Z <sub>ij</sub> ) distribution . . . . .	249
15.3.1 Sampling motif positions based on the Z vector . . . . .	249
15.3.2 More likely to find global maximum, easy to implement . . . . .	250
15.4 De novo motif discovery . . . . .	250
15.4.1 Motif discovery using genome-wide conservation . . . . .	251
15.4.2 Validation of discovered motifs with functional datasets . . . . .	251
15.5 Evolutionary signatures for regulatory motifs . . . . .	251
15.6 Phylogenies, Branch length score as Confidence score . . . . .	251
15.6.1 Foreground vs. background. Real vs. control motifs. . . . .	251
15.7 Possibly deprecated stuff below: . . . . .	251
15.7.1 Greedy . . . . .	251
15.8 Comparing different Methods . . . . .	252
15.9 OOPS,ZOOPS,TCM . . . . .	252
15.10 Extension of the EM Approach . . . . .	253
15.10.1 ZOOPS Model . . . . .	253
15.10.2 Finding Multiple Motifs . . . . .	254
15.11 Motif Representation and Information Content . . . . .	254
<b>16 Regulatory Genomics</b> . . . . .	<b>257</b>
16.1 Introduction . . . . .	257
16.1.1 History of the Field . . . . .	257
16.1.2 Open Problems . . . . .	257
16.2 <i>De Novo</i> Motif Discovery . . . . .	258
16.2.1 TF Motif Discovery . . . . .	258
16.2.2 Validating Discovered Motifs . . . . .	259
16.2.3 Summary . . . . .	259
16.3 Predicting Regular Targets . . . . .	259
16.3.1 Motif Instance Identification . . . . .	259
16.3.2 Validating Targets . . . . .	260
16.4 MicroRNA Genes and Targets . . . . .	260
16.4.1 MiRNA Gene Discovery . . . . .	260
16.4.2 Validating Discovered MiRNAs . . . . .	260
16.4.3 MiRNA's 5' End Identification . . . . .	261
16.4.4 Functional Motifs in Coding Regions . . . . .	261
16.5 Current Research Directions . . . . .	261
16.6 Further Reading . . . . .	261
16.7 Tools and Techniques . . . . .	261
16.8 What Have We Learned? . . . . .	261

<b>17 Epigenomics/Chromatin States</b>	<b>263</b>
17.1 Introduction . . . . .	264
17.2 Epigenetic Information in Nucleosomes . . . . .	265
17.2.1 Epigenetic Inheritance . . . . .	266
17.3 Epigenomic Assays . . . . .	267
17.3.1 ChIP: a method for determining where proteins bind to DNA or where histones are modified . . . . .	267
17.3.2 Bisulfite Sequencing: a method for determining where DNA is methylated . . . . .	268
17.4 Primary data processing of ChIP data . . . . .	269
17.4.1 Read mapping . . . . .	269
17.4.2 Quality control metrics . . . . .	270
17.4.3 Peak Calling and Selection . . . . .	273
17.5 Annotating the Genome Using Chromatin Signatures . . . . .	276
17.5.1 Data Binarization . . . . .	277
17.5.2 HMMs for Chromatin State Annotation . . . . .	278
17.5.3 Model Complexity: Choosing the Number of States to Model . . . . .	278
17.5.4 Results . . . . .	281
17.5.5 Learning Chromatin States Jointly Across Multiple Cell Types . . . . .	283
17.5.6 Epigenomic Imputation . . . . .	284
17.6 Current Research Directions . . . . .	285
17.7 Further Reading . . . . .	285
17.8 Tools and Techniques . . . . .	288
17.9 What Have We Learned? . . . . .	288
<b>18 Regulatory Networks: Inference, Analysis, Application</b>	<b>291</b>
18.1 Introduction . . . . .	291
18.1.1 Introducing Biological Networks . . . . .	292
18.1.2 Interactions Between Biological Networks . . . . .	292
18.1.3 Studying Regulatory Networks . . . . .	293
18.2 Structure Inference . . . . .	293
18.2.1 Key Questions in Structure Inference . . . . .	293
18.2.2 Abstract Mathematical Representations for Networks . . . . .	294
18.3 Overview of the PGM Learning Task . . . . .	295
18.3.1 Parameter Learning for Bayesian Networks . . . . .	295
18.3.2 Learning Regulatory Programs for Modules . . . . .	296
18.3.3 Conclusions in Network Inference . . . . .	296
18.4 Applications of Networks . . . . .	296
18.4.1 Overview of Functional Models . . . . .	296
18.4.2 Functional Prediction for Unannotated Nodes . . . . .	297
18.5 Structural Properties of Networks . . . . .	298
18.5.1 Degree distribution . . . . .	298
18.5.2 Network motifs . . . . .	299
18.6 Network clustering . . . . .	299
18.6.1 An algebraic view to networks . . . . .	300
18.6.2 The spectral clustering algorithm . . . . .	303
<b>19 Chromatin Interactions</b>	<b>307</b>
19.1 Introduction . . . . .	307
19.1.1 What's already known . . . . .	308
19.1.2 What we don't know . . . . .	308
19.1.3 Why do we study it? . . . . .	308
19.2 Relevant terminology . . . . .	309
19.2.1 Nuclear lamina . . . . .	309
19.2.2 Lamina Associated Domains(LADs) . . . . .	309

19.2.3 Histones . . . . .	309
19.2.4 Chromatin . . . . .	309
19.2.5 Chromosome territories (CT) . . . . .	309
19.2.6 Gross folding principles . . . . .	309
19.3 Molecular Methods for Studying Nuclear Genome Organization . . . . .	309
19.3.1 Methods for measuring DNA-Nuclear Lamina interactions . . . . .	310
19.3.2 Measuring DNA-DNA contacts . . . . .	312
19.4 Mapping Genome-Nuclear Lamina Interactions (LADs) . . . . .	314
19.4.1 Interpreting DamID Data . . . . .	314
19.4.2 Interpreting Hi-C Data . . . . .	315
19.5 Computational Methods for Studying Nuclear Genome Organization . . . . .	316
19.5.1 Sources of Bias . . . . .	316
19.5.2 Bias Correction . . . . .	316
19.5.3 3D-modeling of 3C-based data . . . . .	317
19.6 Architecture of Genome Organization . . . . .	318
19.6.1 Multiple cell types influence on determining architecture . . . . .	318
19.6.2 Inter-species comparison of lamina associations . . . . .	318
19.6.3 A-T Content Rule . . . . .	318
19.7 Mechanistic Understanding of Genome Architecture . . . . .	318
19.7.1 Understanding Mitosis and LADs . . . . .	318
19.7.2 Modeling . . . . .	319
19.7.3 Mechanisms of Loop Extrusion . . . . .	320
19.8 Current Research Directions: . . . . .	321
19.8.1 LADs . . . . .	321
19.8.2 TADs and Other Compartments: . . . . .	321
19.8.3 Other/Miscellaneous: . . . . .	322
19.9 Further Reading . . . . .	323
19.10 Available Tools and Techniques . . . . .	323
19.11 What Have We Learned? . . . . .	323
<b>20 Introduction to Steady State Metabolic Modeling</b> . . . . .	<b>325</b>
20.1 Introduction . . . . .	325
20.1.1 What is Metabolism? . . . . .	325
20.1.2 Why Model Metabolism? . . . . .	326
20.2 Model Building . . . . .	326
20.2.1 Chemical Reactions . . . . .	326
20.2.2 Steady-State Assumption . . . . .	327
20.2.3 Reconstructing Metabolic Pathways . . . . .	327
20.3 Metabolic Flux Analysis . . . . .	328
20.3.1 Mathematical Representation . . . . .	328
20.3.2 Null Space of S . . . . .	328
20.3.3 Constraining the Flux Space . . . . .	329
20.3.4 Linear Programming . . . . .	329
20.4 Applications . . . . .	331
20.4.1 <i>In Silico</i> Detection Analysis . . . . .	331
20.4.2 Quantitative Flux <i>In Silico</i> Model Predictions . . . . .	332
20.4.3 Quasi Steady State Modeling (QSSM) . . . . .	332
20.4.4 Regulation via Boolean Logic . . . . .	333
20.4.5 Coupling Gene Expression with Metabolism . . . . .	335
20.4.6 Predicting Nutrient Source . . . . .	336
20.5 Current Research Directions . . . . .	339
20.6 Further Reading . . . . .	339
20.7 Tools and Techniques . . . . .	339
20.8 What Have We Learned? . . . . .	339

<b>21 The ENCODE project: Systematic experimentation and integrative genomics</b>	<b>341</b>
21.1 Introduction . . . . .	341
21.2 Experimental Techniques . . . . .	341
21.3 Computational Techniques . . . . .	343
21.4 Current Research Directions . . . . .	345
21.5 Further Reading . . . . .	345
21.6 Tools and Techniques . . . . .	345
21.7 What Have We Learned? . . . . .	345
<b>22 Synthetic Biology</b>	<b>347</b>
22.1 Introduction . . . . .	347
22.2 Current Research Directions . . . . .	348
22.3 Further Reading . . . . .	349
22.4 Tools and Techniques . . . . .	349
22.5 What Have We Learned? . . . . .	351
<b>IV Phylogenomics and Population Genomics</b>	<b>353</b>
<b>23 Molecular Evolution and Phylogenetics</b>	<b>355</b>
23.1 Introduction . . . . .	357
23.2 Basics of Phylogeny . . . . .	357
23.2.1 Traits . . . . .	357
23.2.2 Trees . . . . .	358
23.2.3 Methods for Tree Reconstruction . . . . .	359
23.3 Distance Based Methods . . . . .	361
23.3.1 From alignment to distances . . . . .	361
23.3.2 Distances to Trees . . . . .	366
23.3.3 Summary of Distance Methods Pros and Cons . . . . .	371
23.4 Molecular Clocks and themakers . . . . .	371
23.5 Character-Based Methods . . . . .	373
23.5.1 Scoring . . . . .	374
23.5.2 Search . . . . .	377
23.6 Possible Theoretical and Practical Issues with Discussed Approach . . . . .	379
23.6.1 Major Issues . . . . .	379
23.6.2 Forest of Life . . . . .	379
23.7 Towards final project . . . . .	380
23.7.1 Project Ideas . . . . .	380
23.7.2 Project Datasets . . . . .	381
23.8 What Have We Learned? . . . . .	381
<b>24 Phylogenomics II</b>	<b>383</b>
24.1 Introduction . . . . .	384
24.2 Inferring Orthologs/Paralogs, Gene Duplication and Loss . . . . .	384
24.2.1 Species Tree . . . . .	384
24.2.2 Gene Tree . . . . .	384
24.2.3 Gene Family Evolution . . . . .	385
24.2.4 Reconciliation . . . . .	385
24.2.5 Interpreting Reconciliation Examples . . . . .	389
24.3 Reconstruction . . . . .	389
24.3.1 Species Tree Reconstruction . . . . .	390
24.3.2 Improving Gene Tree Reconstruction and Learning Across Gene Trees . . . . .	390
24.4 Modeling Population and Allele Frequencies . . . . .	392
24.4.1 The Wright-Fisher Model . . . . .	392

24.4.2 The Coalescent Model . . . . .	393
24.4.3 The Multispecies Coalescent Model . . . . .	397
24.5 SPIDIR . . . . .	397
24.5.1 Background . . . . .	397
24.5.2 Method and Model . . . . .	398
24.6 Ancestral Recombination Graphs . . . . .	399
24.6.1 The Sequentially Markov Coalescent . . . . .	400
24.7 Conclusion . . . . .	401
24.8 Current Research Directions . . . . .	402
24.9 Further Reading . . . . .	402
24.10 Tools and Techniques . . . . .	402
24.11 What Have We Learned? . . . . .	402
<b>25 Population History</b> . . . . .	<b>403</b>
25.1 Introduction . . . . .	403
25.2 Quick Survey of Human Genetic Variation . . . . .	404
25.3 African and European Gene Flow . . . . .	405
25.4 Gene Flow on the Indian Subcontinent . . . . .	405
25.4.1 Almost All Mainland Indian Groups are Mixed . . . . .	406
25.4.2 Population structure in India is different from Europe . . . . .	406
25.4.3 Discussion . . . . .	408
25.5 Gene Flow Between Archaic Human Populations . . . . .	408
25.5.1 Background . . . . .	408
25.5.2 Evidence of Gene Flow between Humans and Neanderthals . . . . .	408
25.5.3 Gene Flow between Humans and Denisovans . . . . .	409
25.5.4 Analysis of High Coverage Archaic Genomes . . . . .	409
25.5.5 Discussion . . . . .	410
25.6 European Ancestry and Migrations . . . . .	410
25.6.1 Tracing the Origins of European Genetics . . . . .	410
25.6.2 Migration from the Steppe . . . . .	411
25.6.3 Screening for Natural Selection . . . . .	412
25.7 Tools and Techniques . . . . .	413
25.7.1 Techniques for Studying Population Relationships . . . . .	413
25.7.2 Extracting DNA from Neanderthal Bones . . . . .	413
25.7.3 Reassembling Ancient DNA . . . . .	414
25.8 Research Directions . . . . .	415
25.9 Further Reading . . . . .	415
<b>26 Population Genetic Variation</b> . . . . .	<b>417</b>
26.1 Introduction . . . . .	417
26.2 Population Selection Basics . . . . .	418
26.2.1 Polymorphisms . . . . .	418
26.2.2 Allele and Genotype Frequencies . . . . .	419
26.2.3 Ancestral State of Polymorphisms . . . . .	422
26.2.4 Measuring Derived Allele Frequencies . . . . .	423
26.3 Genetic Linkage . . . . .	423
26.3.1 Correlation Coefficient $r^2$ . . . . .	424
26.4 Haplotype Phasing . . . . .	425
26.4.1 Phasing of related individuals . . . . .	425
26.4.2 Phasing of unrelated individuals . . . . .	425
26.4.3 Fine-mapping Disease Associations . . . . .	426
26.5 Natural Selection . . . . .	426
26.5.1 Genomics Signals of Natural Selection . . . . .	426
26.6 Human Evolution . . . . .	429

26.6.1 A History of the Study of Population Dynamics . . . . .	429
26.6.2 Understanding Disease . . . . .	432
26.6.3 Understanding Recent Population Admixture . . . . .	433
26.7 Current Research . . . . .	434
26.7.1 HapMap project . . . . .	434
26.7.2 1000 genomes project . . . . .	434
26.8 Further Reading . . . . .	434
<b>v Medical Genomics</b>	<b>435</b>
<b>27 Variation 2: : Quantitative trait mapping, eQTLs, molecular trait variation</b>	<b>437</b>
27.1 Introduction . . . . .	437
27.2 Motivations for studying eQTLs . . . . .	437
27.2.1 Effects of noncoding regions on phenotype . . . . .	437
27.2.2 Understanding and treating diseases at multiple phenotypic layers . . . . .	438
27.3 eQTL Basics . . . . .	438
27.3.1 Cis-eQTLs . . . . .	439
27.3.2 Trans-eQTLs . . . . .	440
27.4 Structure of an eQTL Study . . . . .	441
27.4.1 Considerations for Expression Data . . . . .	441
27.4.2 Considerations for Genomic Data . . . . .	442
27.4.3 Covariate Adjustment . . . . .	442
27.4.4 Insights on eQTL's and biology . . . . .	444
27.5 Current Research Directions . . . . .	446
27.5.1 Quantifying Trait Variation . . . . .	446
27.5.2 New Applications . . . . .	446
27.6 What Have We Learned? . . . . .	447
27.7 Further Reading . . . . .	447
27.8 Tools and Resources . . . . .	448
<b>28 Personal Genomes, Synthetic Genomes, Computng in C vs. Si</b>	<b>451</b>
28.1 Introduction . . . . .	451
28.2 Reading and Writing Genomes . . . . .	451
28.3 Personal Genomes . . . . .	452
28.4 Current Research Directions . . . . .	452
28.5 Further Reading . . . . .	452
28.6 Tools and Techniques . . . . .	453
28.7 What Have We Learned? . . . . .	453
<b>29 Personal Genomics</b>	<b>455</b>
29.1 Introduction . . . . .	455
29.2 Epidemiology: An Overview . . . . .	456
29.3 Genetic Epidemiology . . . . .	457
29.4 Molecular Epidemiology . . . . .	458
29.4.1 meQTLs . . . . .	459
29.4.2 EWAS . . . . .	460
29.5 Causality Modeling and Testing . . . . .	460
29.5.1 Polygenic Risk Prediction . . . . .	461
29.6 Current Research Directions . . . . .	461
29.6.1 PheWAS motivation . . . . .	461
29.6.2 Modeling PheWAS with genetic information . . . . .	462
29.6.3 Modeling PheWAS with EHR data . . . . .	463
29.6.4 Missing Mechanism . . . . .	463

29.6.5 EHR with MAR . . . . .	464
29.6.6 EHR with NMAR . . . . .	464
29.6.7 Modeling longitudinal EHR data . . . . .	465
29.7 Further Reading . . . . .	465
29.8 Tools and Techniques . . . . .	465
29.9 What Have We Learned? . . . . .	465
<b>30 Cancer Genomics</b> . . . . .	<b>467</b>
30.1 Introduction . . . . .	467
30.2 Characterization . . . . .	468
30.3 Background Mutation Models . . . . .	469
30.4 Immunotherapy . . . . .	470
30.5 Current Research Directions . . . . .	471
30.6 Further Reading . . . . .	471
30.7 Tools and Techniques . . . . .	471
30.8 What Have We Learned? . . . . .	471
<b>31 Genome Editing</b> . . . . .	<b>473</b>
31.1 Introduction . . . . .	473
31.1.1 What is CRISPR/Cas? . . . . .	473
31.1.2 Why is CRISPR/Cas important to us? . . . . .	473
31.1.3 Cas9 . . . . .	473
31.2 Engineering Cas9 for Mammalian Genome Editing . . . . .	474
31.2.1 Cas9 Specificity . . . . .	474
31.2.2 Cas9 for transcriptional activation and repression . . . . .	474
31.2.3 Cas9 as a tool for screening . . . . .	474
31.3 Novel CRISPR enzymes . . . . .	475
31.3.1 Cpf1 . . . . .	475
31.3.2 C2c2 . . . . .	475
31.4 Further Reading . . . . .	476
31.5 Tools and Techniques . . . . .	476
31.6 What Have We Learned? . . . . .	476
31.6.1 Why is CRISPR/Cas important to us? . . . . .	476



## Preface and Acknowledgements

These notes summarize the material taught in the MIT course titled “Computational Biology: Genomes, Networks, Evolution”, also cross-listed with Harvard, HST, HSPH and BU over the years. The course was listed as MIT course 6.047/6.878 in 2007-2011 (and under the temporary numbers 6.085/6.095/6.895 in Fall 2005-2006, and 6.096 in Spring 2005). It was cross-listed with MIT/Harvard Health Sciences and Technology (HST) course HST.507 in 2007-2011, Boston University Biological Engineering course BE-562 in 2008 and 2009, and Harvard School of Public Health course IMI231 in 2009-2011.

The course was originally developed by Prof. Manolis Kellis at MIT, with advice from **Silvio Micali**. It was first taught in Spring 2005 as a half-course extension to the Introduction to Algorithms Course (6.046), and as an independent full-credit course in Fall 2005-2011. The course was co-lectured with Prof. **Piotr Indyk** in Fall 2005-2006, who contributed to the material on hashing and dimensionality reduction techniques. It was co-taught with Prof. **James Galagan** in Fall 2007-2009 who contributed to the lectures on expression clustering, supervised learning and metabolic modelling, and who continued teaching the course independently at BU.

The material in the course has benefited tremendously from courses by **Bonnie Berger** at MIT, whose course “Introduction to Computational Biology (18.417)” was co-taught by Manolis Kellis as a student in Fall 2001, and **Serafim Batzoglou** at Stanford whose course “Computational Genomics (CS262)” was an inspiration for clarity and style and a source of figures and diagrams for the early chapters on alignment and HMMs. Lastly, the material in the course also benefited from two books used extensively in the course in the last several years, titled “Biological Sequence Analysis” by **Durbin, Eddy, Krogh, and Mitchison**, and “Bioinformatics Algorithms” by **Jones and Pevzner**.

The material of several chapters was initially developed by guest lecturers who are experts in their field and contributed new material, figures, slides, organization, and thoughts in the form of one or more lectures. Without them, the corresponding chapters would not have been possible. They are: **Pardis Sabeti** (Population Genetic Variation), **Mark Daly** (Medical Genetics), **David Reich** (Population History), **Eric Alm** (Bacterial Genomics), **John Rinn** (Long Non-Coding RNAs), **James Galagan** (Steady State modeling), **Matt Rasmussen** (Phylogenomics), **Mike Lin** (Gene finding), **Stefan Washietl** (RNA folding), **Jason Ernst** (Epigenomics), **Sushmita Roy** (Regulatory Networks), **Pouya Kheradpour** (Regulatory Genomics).

The Teaching Assistants who taught recitations and helped develop the course problem sets have been **Reina Reimann** (Spring 2005), **Pouya Kheradpour** (Fall 2005), **Matt Rasmussen** and **Mike Lin** (Fall 2006), **Mike Lin** and **David Sontag** (Fall 2007), **Matt Rasmussen** and **Pouya Kheradpour** (Fall 2008), **Ed Reznik** and **Bob Altshuler** (Fall 2009), **Matt Edwards** (Fall 2010), **Melissa Gymrek** (Fall 2011), **Rachel Sealfon** (Fall 2012), **Max Wolf** (Fall 2013), **Abhishek Sarkar** (Fall 2014), **Michal Grzadkowski** (Fall 2015), **Connor Duffy** and **Tejas Sundaresan** (Fall 2016) and **Sitara Persad** (Fall 2017). The notes were originally compiled in a uniform format by **Anna Shcherbina** (Fall 2011).

The current and past members of the **MIT CompBio Lab** (<http://compbio.mit.edu/people.html>), who have taught me as they grew into experts in their own fields. They are: Matt Rasmussen, Mike Lin, Pouya Kheradpour, Alexander Stark, Xiaohui Xie, Jason Ernst, Sushmita Roy, Luke Ward, Chris Bristow, Abdoulaye Diallo, David Hendrix, Loyal Goff, Stefan Washietl, Daniel Marbach, Mukul Bansal, Matthew Eaton, Irwin Jungreis, Rachel Sealfon, Bob Altshuler, Jessica Wu, Angela Yen, Soheil Feizi, Luis Barrera, Ben Holmes, Anna Ayuso, Wouter Meuleman, Ferhat Ay, Rogerio Candeias, Patrick Meyer, Tom Morgan, Wes Brown, Will Gibson, Rushil Goel, Luisa Di Stefano, Stephan Ossowski, Aviva Presser, Erez Lieberman, Joshua Grochow, Yuliya Kodysh, Leopold Parts, Ameya Deoras, Matt Edwards, Adrian Dalca.

The students taking the class and contributing to the scribe notes are:

- Spring 2005: Dan Arlow, Arhab Battacharyya, Punyashloka Biswal, Adam Bouhenguel, Dexter Chan, Shuvo Chatterjee, Tiffany Dohzen, Lyric Doshy, Robert Figueiredo, Edena Gallagher, Josh Grochow, Aleksas Hauser, Blanca Himes, George Huo, Xiaoming Jia, Scott Johnson, Steven Kannan, Faye Kasemset, Jason Kelly, Daniel Kim, Yuliya Kodysh, Nate Kushner, Lucy Mendel, Jose Pacheco, Sejal Patel, Haiharan Rahul, Gireeja Ranade, Sophie Rapoport, Aditya Rastogi, Shubhangi Saraf, Oded Shaham, Walter Stiehl, Kevin Stolt, James Sun, Xin Sun, Kah Tai, Kah Tay, Chester Tse, Verlik Tzanov, Brian Wu

- Fall 2005: Ebad Ahmed, Christophe Falling, Michael Farry, Elaine Gee, Luke Hutchison, Michael Lin, Grigore Pintilie, Asfandyar Qureshi, Matthew Rasmussen, Alexandru Salcianu, Zeeshan Syed, Hayden Taylor, Velin Tzanov, Grant Wang
- Fall 2006: Mats Ahlgren, Zhu Ailing, Bob Altshuler, Nada Amin, Shay Artzi, Solomon Bisker, Allen Bryan, Sumeet Gupta, Adam Kiezun, Richard Koche, Mieszko Lis, Ryan Newton, Michael O'Kelly, Chris Reeder, Jonathan Rhodes, Michael Schnall-Levin, Alex Tsankov, Tarak Upadhyaya, Kush Varshney, Sam Volchenboum, Jon Wetzel, Amy Williams
- Fall 2007: Anton Aboukhalil, Matthew Belmonte, Ellenor Brown, Brad Cater, Alal Eran, Guilherme Fujiwara, Saba Gul, Kate Hoff, Shannon Iyo, Eric Jonas, Peter Kruskall, Michael Lee, Ben Levick, Fulu Li, Alvin Liang, Joshua Lim, Chit-Kwan Lin, Po-Ru Loh, Kevin Modzelewski, Georgis Papachristoudis, Michalis Potamias, Emmanuel Santos, Alex Schwendner, Maryam Shanechi, Timo Somervuo, James Sun, Xin Sun, Robert Toscano, Qingqing Wang, Ning Xie, Qu Zhang, Blaine Ziegler
- Fall 2008: Burak Alver, Tural Badirkhanli, Arnab Bhattacharyya, Can Cenik, Clara Chan, Lydia Chilton, Arkajit Dey, Ardavan Farjadpour, Jeremy Fineman, Bernhard Haeupler, Arman Hajati, Ethan Heilman, Joe Herman, Irwin Jungreis, Arjun Manrai, Nilah Monnier, Christopher Rohde, Rachel Sealoff, Daniel Southern, Paul Steiner, David Stiebel, Mengdi Wang
- Fall 2009: Layla Barkal, Michael Bennie, David Charlton, Guoliang Chew, John Dong, Matthew Edwards, Eric Eisner, Subha Gollakota, Nathan Haseley, Allen Lin, Christopher McFarland, Michael Melgar, Anrae Motes, Anand Oza, Elizabeth Perley, Brianna Petrone, Arya Tafvizi Zavareh, Yi-Chieh Wu, Angela Yen, Morteza Zadimoghaddam, Chelsea Zhang, James Zou
- Fall 2010: Minjeong Ahn, Andreea Bodnari, Wesley Brown, Jenny Cheng, Bianca Dumitrascu, Sam Esfahani, Amer Fejzic, Talitha Forcier, Maria Frendberg, Dhruv Garg, Rushil Goel, Melissa Gymrek, Benjamin Holmes, Wui Ip, Isaac Joseph, Geza Kovacs, Gleb Kuznetsov, Adam Marblestone, Alexander McCauley, Sheida Nabavi, Jacob Shapiro, Andrew Shum, Ashutosh Singhal, Mark Smith, Mashaal Sohail, Eli Stickgold, Tahin Syed, Lance Wall, Albert Wang, Fulton Wang, Jerry Wang
- Fall 2011: Asa Adadey, Leah Alpert, Ahmed Bakkar, Rebecca Bianco, Brett Boval, Kelly Brock, Peter Carr, Efrain Cermenio, Alex Chernyakhovsky, Diana Chien, Akashnil Dutta, Temuge Enkhbaatar, Maha Farhat, Alec Garza-Galindo, Fred Grober, Gabriel Ha, Marc Hafner, Neel Hajare, Timothy Helbig, Ivan Imaz, Yarden Katz, Gwang Ko, David Ku, Yu-Chi Kuo, Dan Landay, Yingqing Li, Mark Mimee, Selene Mota, Hyun Ji Noh, Chrisantha Perera, Aleksey Pesterev, Michael Quintin, Maria Rodriguez, Megan Roystman, Abhishek Sarkar, Angela Schwarz, Meriem Sefta, Anna Shcherbina, Mindy Shi, Noam Shores, Eric Soderstrom, Ying Qi Soh, Sarah Spencer, Derrick Sund, Ruqi Tang, Zenna Tavares, Arvind Thiagarajan, Paul Tillberg, Christos Tzamos, Leonardo Urbina, Manasi Vartak, Nathan Villagaray-Carski, Sajith Wickramasekara, Thomas Willems, Maxim Wolf, Lok Sang Wong, Iris Xu, Johannes Yeh, Deniz Yorukoglu, Boyang Zhao.
- Fall 2013: Maria Alexis, Polina Binder, Jake Bograd-Denton, Orhan Tunc Celiker, Hyunghoon Cho, Brian Cleary, David Danko, Vivek Dasari, Dalesh Dharamshi, Atray Dixit, Joseph Driscoll, John Froberg, Themistoklis Gouleakis, Carissa Jansen, Yuta Kato, Hanna Levitin, Brendan Liu, Quanquan Liu, Yang Li, Julianna Mello, Hayden Metsky, Peter Nguyen, Luke O'Connor, Alexander Pagan, Sebastian Palacios, Peter Palmedo, Jr., Staphany Park, Nicole Power, Emma Seropian, Meena Subramaniam, Nirvan Tyagi, Joseph Vitti, Timothy Wall, Deena Wang, James Weis, Iris Xu, Haoyang Zeng, Sidi Zhang
- Fall 2014: Abdulaziz Alghunaim, Sahar Alkhairy, Benjamin Bauchwitz, Tristan Bepler, Silvia Canas Duarte, Kevin Chen, Michael Coulombe, Lei (Jerry) Ding, Gabriel Filsinger, Matthew Fox, Kristjan Kasenitit, Joseph Kim, David Lazar, William Leiserson, Jenny Lin, Kathy Lin, Yunpeng Liu, Nicolai Ludvigsen, Eric Mazumdar, Hilary Mulholland, Pavel Muravyev, Muneeza Patel, Divya Pillai, Clément Pit-Claudel, Adam Sealoff, Ha Kyung (Kris) Shin, Aradhana Sinha, Daniel Sosa, Yi-Shiuan Tung, Margaret Walker, Sarah Walker, Yuhao Wang, Hui Ting Grace Yeo, Catherine Yun

- Fall 2015: Jonathan Li, Jesse Tordoff, Thrasyvoulos Karydis, Heather Sweeney, Eric Bartell, Anastasiya Belyaeva, Justin Gullingsrud, Cara Weisman, Robert Hunt, Alex Genshaft, Ge Liu, Richard Hsu, Karthik Murugadoss, Sagar Indurkhy, Max Shen, Kevin Tian, Alvin Shi, Connor Duffy, Narek Dshkhunyan, Joyce Hong, Gil Goldshlager, Sophia Liu, Aurora Alvarez-Buylla, Giri Anand, Tejas Sundaresan, Nolan Kamitaki, Bryce Hwang, Hunter Gatewood, Misha Jamy, Nadia Wallace, Carles Boix, Ava Soleimany, Brock Wooldridge, Sadik Yildiz, Anne Kim, Divya Shanmugam, Deniz Aksel, Molly Schmidt, Jonahtan Uesato, Joseph Cunningham, Suganya Sridharma, Oleksandr Chaykovskyy, Eunice Wu, Sam Johnson, Ye Tao
- Fall 2016: Claire Simpson, Maria Fabre, Ariya Shajii, Eadaoin Harney, Nathan Nakatsuka, Casper Enghuis, Jan-Christian Huetter Tiffany Chen, Minyi Lee, Ryan Chung, Sarah Nyquist, Sitara Persad, Joseph Lin, Omar Abudayyeh, Julia Joung, Varun Mangalick, Yilun Du Peitong Duan, Nathan Hunt, Sang-Woo Jun, Dina Levy-Lambert, Katy Johnson, Jen Hammelman, Erin Hong, Casey Hong, Jackie Vahey, Friederike Buck, Jacqui Liu, Yuge Ji, Isabel Chien, Tina Quach, Meseret Kebede, Aasavari Phanse , Ruth Park, Zi-Ning Choo Alex Smith, Max Gold, Miriam Schiffman, Tony Cao, Ang Cui, Roberto Soto, Larry Zhang, Melissa Slaughter, Sabrina Ibarra, Yi Wang, Margalit Glasgow, Maria Karelina, Alap Sahoo
- Fall 2017: Alyssa Y Chen, Andrew Robert Moorman, Anna E Sappington, Austin T Wang, Chengzhen L Dai, Fangyuan Hong, Helen Abadiotakis, Jesse D Gibson, King Him Ching, Maha Shady, Michael Alexander Murphy, Nemanja Marjanovic, Pranam Chatterjee, Riker F Bixby, Roger Jin, Rohil Verma, Shahul Alam, Skyler E Kaufman, Samuel Sungil Kim, Ellen Dee Zhong ,Kevin R Cuneo, Alexander W Derry, Ryan S Hays, Samuel Joseph Hendel



---

CHAPTER  
**ONE**

---

## INTRODUCTION TO THE COURSE

### Figures

---

1.1	In this computational biology problem, we are provided with a sequence of bases, and wish to locate genes and regulatory motifs.	8
1.2	The double-helix structure of DNA. Nucleotides are in the center, and the sugar-phosphate backbone lies on the outside.	13
1.3	DNA is packed over several layers of organization into a compact chromosome.	14
1.4	RNA is produced from a DNA template during transcription. A “bubble” is opened in the DNA, allowing the RNA polymerase to enter and place down bases complementary to the DNA.	14
1.5	This codon table shows which of the 20 amino acid each of the 3-nucleotide codons in mRNA are translated into. In red are the stop codons, which terminate translation.	16
1.6	Operon Lac illustrates a simple biological regulatory system. In the presence of glucose, genes to lactose metabolism are turn out because glucose inactivates an activator protein. In the absence of lactose, a repressor protein also turns out the operon. Lactose metabolism genes are expressed only in the presence of lactose and absence of glucose.	17
1.7	Metabolic pathways and regulation can be studied by Computational biology. Models are made from genome scale information and used to predict metabolic function and to metabolic engineering. An example of biological engineering is modifying bacteria genome to overproduce artemesinin, an antibiotic used to treat malaria.	18

---

## 1.1 Introduction and Goals

### 1.1.1 A course on computational biology

These lecture notes are aimed to be taught as a term course on computational biology, each 1.5 hour lecture covering one chapter, coupled with bi-weekly homework assignments and mentoring sessions to help students accomplish their own independent research projects. The notes grew out of MIT course 6.047/6.878, and very closely reflect the structure of the corresponding lectures.

### 1.1.2 Duality of Goals: Foundations and Frontiers

There are two goals for this course. The first goal is to introduce you to the **foundations** of the field of computational biology. Namely, introduce the fundamental biological problems of the field, and learn the algorithmic and machine learning techniques needed for tackling them. This goes beyond just learning how to use the programs and online tools that are popular any given year. Instead, the aim is for you to understand the underlying principles of the most successful techniques that are currently in use, and provide you with the capacity to design and implement the next generation of tools. That is the reason why an introductory

algorithms class is set as a pre-req; the best way to gain a deeper understanding for the algorithms presented is to implement them yourself.

The second goal of the course is to tackle the research **frontiers** of computational biology, and that's what all the advanced topics and practical assignments are really about. We'd actually like to give you a glimpse of how research works, expose you to current research directions, guide you to find the problems most interesting to you, and help you become an active practitioner in the field. This is achieved through guest lectures, problem sets, labs, and most importantly, a term-long independent research **project**, where you carry out your independent research.

The **modules** of the course follow that pattern, each consisting of lectures that cover the foundations and the frontiers of each topic. The foundation lectures introduce the classic problems in the field. These problems are very well understood and elegant solutions have already been found; some have even been taught for well over a decade. The frontiers portion of the module cover advanced topics, usually by tackling central questions that still remain open in the field. These chapters frequently include guest lectures by some of the pioneers in each area who speak both about the general state of the field as well as their own lab's research.

The **assignments** for the course follow the same foundation/frontiers pattern. Half of the assignments are going to be about working out the methods with pencil on paper, and diving deep into the algorithmic and machine learning notions of the problems. The other half are actually going to be practical questions consisting of programming assignments, where real data sets are provided. You will analyze this data using the techniques you have learned and interpret your results, giving you a real hands on experience. The assignments build up to the final project, where you will propose and carry out an original research project utilizing or expanding upon the methods taught in the course, and present your findings in conference format. Overall, the assignments are designed to give you the opportunity to apply computational biology methods to real problems in biology.

### 1.1.3 Duality of disciplines: Computation and Biology

In addition to aiming to cover both foundations and frontiers, the other important duality of this course is between computation and biology.

From the **biological** perspective of the course, we aim to teach topics that are fundamental to our understanding of biology, medicine, and human health. We therefore shy away from any computationally-interesting problems that are biologically-inspired, but not relevant to biology. We're not just going to see something in biology, get inspired, and then go off into computer science and do a lot of stuff that biology will never care about. Instead, our goal is to work on problems that can make a significant change in the field of biology. We'd like you to publish papers that actually matter to the biological community and have real biological impact. This goal has therefore guided the selection of topics for the course, and each chapter focuses on a fundamental biological problem.

From the **computational** perspective of the course, being after all a computer science class, we focus on exploring general techniques and principles that are certainly important in computational biology, but nonetheless can be applied in any other fields that require data analysis and interpretation. Hence, if what you want is to go into cosmology, meteorology, geology, or any such, this class offers computational techniques that will likely become useful when dealing with real-world data sets related to those fields.

### 1.1.4 Why Computational Biology?

#### [lecture1\\_transcript.html#Motivations](#)

There are many reasons why Computational Biology has emerged as an important discipline in recent years, and perhaps some of these have led you to pick up this book or register for this class. Although we have our own opinions on what these reasons are, we have asked the students year after year for their own views on what has enabled the field of Computational Biology to expand so rapidly in the last few years. Their responses fall into several broad themes, which we summarize here.

1. Perhaps the most fundamental reason why computational approaches are so well-suited to the study of biological data is that at their core, biological systems are **fundamentally digital in nature**. To be blunt, humans are not the first to build a digital computer – our ancestors *are* the first digital computer,

as the earliest DNA-based life forms were already storing, copying, and processing digital information encoded in the letters A,C,G, and T. The major evolutionary advantage of a digital medium for storing genetic information is that it can persist across thousands of generations, while analog signals would be diluted from generation to generation from basic chemical diffusion.

2. Besides DNA, many other aspects of biology are digital, such as **biological switches**, which ensure that only two discrete possible states are achieved by feedback loops and metastable processes, even though these are implemented by levels of molecules. Extensive feedback loops and other diverse regulatory circuits implement discrete decisions through otherwise unstable components, again with design principles similar to engineering practice, making our quest to understand biological systems from an engineering perspective more approachable.
3. Sciences that heavily benefit from data processing, such as Computational Biology, follow a virtuous cycle involving the data available for processing. The more that can be done by processing and analyzing the available data, the more funding will be directed into developing technologies to obtain, process and analyze even more data. New technologies such as sequencing, and high-throughput experimental techniques like microarray, yeast two-hybrid, single cell RNA-seq, and ChIP-chip assays are creating **enormous and increasing amounts of data** that can be analyzed and processed using computational techniques. The \$1000 and \$100 genome projects are evidence of this cycle. Over ten years ago, when these projects started, it would have been ludicrous to even imagine processing such massive amounts of data. However, as more potential advantages were devised from the processing of this data, more funding was dedicated into developing technologies that would make these projects feasible.
4. The ability to process data has greatly improved in the recent years, owing to: 1) the massive computational power available today (due to Moore's law, among other things), and 2) the advances in the algorithmic techniques at hand.
5. Optimization approaches can be used to solve, via computational techniques, that are otherwise intractable problems.
6. **Running time & memory** considerations are critical when dealing with huge datasets. An algorithm that works well on a small genome (for example, a bacteria) might be too time or space inefficient to be applied to 1000 mammalian genomes. Also, combinatorial questions dramatically increase algorithmic complexity.
7. Biological datasets can be **noisy**, and filtering signal from noise is a computational problem.
8. **Machine learning** approaches are useful to make inferences, classify biological features, & identify robust signals.
9. As our understanding of biological systems deepens, we have started to realize that such systems cannot be analyzed in isolation. These systems have proved to be intertwined in ways previously unheard of, and we have started to shift our analyses to techniques that consider them all as a whole.
10. It is possible to use computational approaches to find correlations in an unbiased way, and to come up with conclusions that transform biological knowledge and facilitate active learning. This approach is called **data-driven discovery**.
11. Computational studies can **predict** hypotheses, mechanisms, and theories to explain experimental observations. These falsifiable hypotheses can then be tested experimentally.
12. Computational approaches can be used not only to analyze existing data but also to **motivate data collection** and suggest useful experiments. Also, computational filtering can narrow the experimental search space to allow more focused and efficient experimental designs.
13. Biology has **rules**: Evolution is driven by two simple rules: 1) random mutation, and 2) brutal selection. Biological systems are constrained to these rules, and when analyzing data, we are looking to find and interpret the emerging behavior that these rules generate.

14. **Datasets can be combined** using computational approaches, so that information collected across multiple experiments and using diverse experimental approaches can be brought to bear on questions of interest.
15. Effective **visualizations** of biological data can facilitate discovery.
16. Computational approaches can be used to **simulate & model** biological data.
17. Computational approaches can be more **ethical**. For example, some biological experiments may be unethical to perform on live subjects but could be simulated by a computer.
18. Large scale, systems engineering approaches are facilitated by computational technique to obtain global views into the organism that are too complex to analyze otherwise.

### 1.1.5 Finding Functional Elements: A Computational Biology Question

[lecture1\\_transcript.html#Codons](#)

Several computational biology problems refer to finding biological signals in DNA data (e.g. coding regions, promoters, enhancers, regulators, ...).

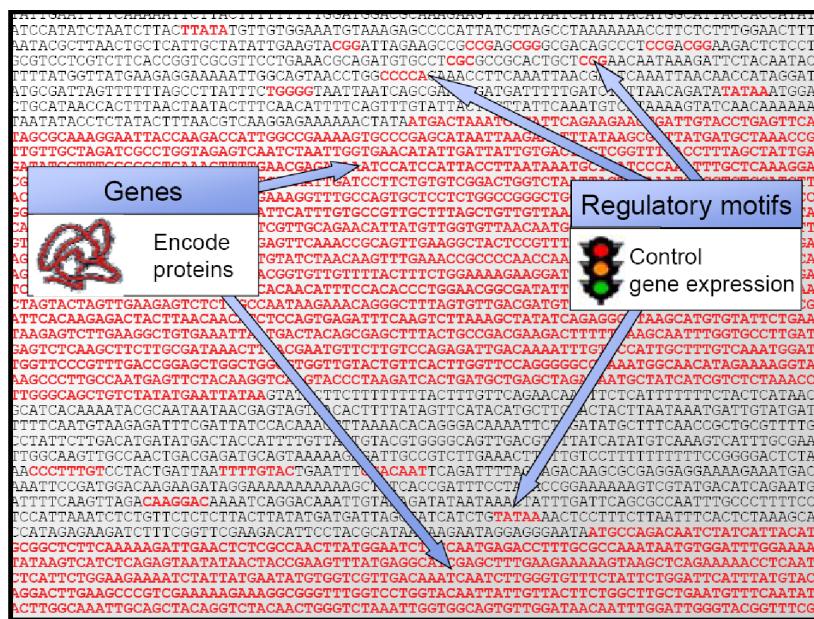


Figure 1.1: In this computational biology problem, we are provided with a sequence of bases, and wish to locate genes and regulatory motifs.

We now diss a specific question that computational biology can be used to address: how can one find functional elements in a genomic sequence? Figure 1.1 shows part of the sequence of the yeast genome. Given this sequence, we can ask:

**Q:** What are the genes that encode proteins?

**A:** During translation, the start codon marks the first amino acid in a protein, and the stop codon indicates the end of the protein. However, as indicated in the “Extracting signal from noise” slide, only a few of these ATG sequences in DNA actually mark the start of a gene which will be expressed as protein. The others are “noise”; for example, they may have been part of introns (non-coding sequences which are spliced out after transcription).

**Q:** How can we find features (genes, regulatory motifs, and other functional elements) in the genomic sequence?

**A:** These questions could be addressed either experimentally or computationally. An experimental approach to the problem would be creating a knockout, and seeing if the fitness of the organism is affected. We could also address the question computationally by seeing whether the sequence is conserved across the genomes of multiple species. If the sequence is significantly conserved across evolutionary time, it's likely to perform an important function.

There are caveats to both of these approaches. Removing the element may not reveal its function—even if there is no apparent difference from the original, this could be simply because the right conditions have not been tested. Also, simply because an element is not conserved doesn't mean it isn't functional. (Also, note that "functional element" is an ambiguous term. Certainly, there are many types of functional elements in the genome that are not protein-encoding. Intriguingly, 90-95% of the human genome is transcribed (used as a template to make RNA). It isn't known what the function of most of these transcribed regions are, or indeed if they are functional).

This should be chapter

## 1.2 Final Project - Introduction to Research In Computational Biology

[lecture1\\_transcript.html#FinalProject](#)

### 1.2.1 Final project goals

An important component of being a computational biologist is the ability to carry out independent research in the area. The skills for a successful researcher differ from one person to the next, but in the process of teaching this course, we have identified several aspects that are all needed, and laid out activities for a term-long project, that enable students to carry out their independent research.

The project mirrors real world scientific process: come up with an idea → frame it → propose it → revise it → carry it out → present your results. Students are expected to think critically about their own project, and also evaluate peer research proposals, and lastly respond to feedback from their peers.

Students are expected to use real data and present their results in conference format. The ultimate goal is publishable research. Students are encouraged to talk with the course staff while formulating a final project idea, look head through the various chapters and modules, and get an idea of what areas will interest you most.

### 1.2.2 Final project milestones

Instead of waiting until the end of the term to begin brainstorming or provide feedback, we begin project activities with the first problem set, to identify problems of interest and types of projects, find partners, speak with current students and postdocs in computational biology that can serve as mentors, and lay out a research plan in the style of an NIH proposal to identify potential pitfalls early and address them or work around them before they become a bottleneck.

By setting up several incremental progress milestones throughout the term, coupled with mentoring and feedback throughout the semester, we have achieved consistent progress in previous years, which can be useful to students taking on a new project at any stage of their career. Research projects from this course in the past have been used as the starting point for a published paper, have led to Masters and PhD theses, and earned awards both academically and in conferences.

The timeline for the final project is as follows:

1. **Set-up:** a brief overview of your experience and interest. Due 9/27
2. **Brainstorming:** a list of initial project ideas and partners. Due 10/4
3. **Proposal:** submit a project proposal in the form of an NIH proposal. Due 10/18
4. **Proposal presentation:** present slides to class and mentors on the proposal. Due 10/21

5. **Review:** review and critique 3 peer proposals. Due 10/28
6. **Midterm Progress Report:** write outline of final report. Due 11/23
7. **Final Project Report:** write report in conference paper format. Due 12/11
8. **Final Class Presentation:** 10min conference talk. Due 12/13

There will be Friday mentoring sessions before each portion of the final project is due, and you are encouraged to find a mentor at the first few sessions who is actively interested in your project and could help you more frequently. The mentoring sessions can be helpful in identifying if unexpected results are the result of a bug or are instead a discovery.

Make sure you start working on the project even while waiting for peer reviews, so that you will have 4-5 weeks to complete the research itself.

### 1.2.3 Project deliverables

The final project will include the following two deliverables:

1. A written presentation, due Sunday at 8pm, last week of classes. The written presentation can contain the following elements:
  - Who did what (to reflect trend in publications)
  - The overall project experience
  - Your discoveries
  - What you learned from the experience (introspection)
2. An oral presentation, due Tuesday after the written presentation. This allows students three days to prepare the oral presentation.

### 1.2.4 Project grading

Selecting a project that will be successful can be difficult. To help students optimize for a successful project, we let them know in advance the grading scheme, designed to maximize the project impact by being original, challenging, and relevant to the field, but of course the grade is ultimately dependent on the overall achievement and the clarity of presentation.

Briefly, the grading equation for the final project is:

$$\min(O, C, R) \times A + P$$

where

**Originality** - unoriginal computational experiments don't get published

**Challenge** - the project needs to be sufficiently difficult

**Relevance** - it needs to be from biology, can't just reuse something from another field

**Achievement** - if you don't accomplish anything you won't get a good grade

**Presentation** - even if you've achieved a good project you have to be able to present it so everyone knows that, and make it look easy. The presentation should show how the project is *O*, *C*, and *R*.

Originality, Challenge, Relevance are each out of 5 points, Achievement and Presentation are each out of 10.

## 1.3 Additional materials

### 1.3.1 Online Materials for Fall 2016

[lecture1\\_transcript.html#Handouts](#)

In addition to these *static* notes, the course has several online resources:

- The course website can be accessed at: <http://compbio.mit.edu/6.047>, where course material and information from each term can be found
- The course calendar on Google Calendar. You can add "6.047 Lectures", a public calendar.
- The NB note-taking system for annotating these notes <http://nb.mit.edu/>
- Lastly, videos and audio recordings of all lectures from Fall 2014 are freely available on Stellar.

### 1.3.2 Textbooks

[lecture1\\_transcript.html#CourseInformation](#) The following three (optional) reference textbooks are recommended for the class.

1. Richard Durbin, Sean R. Eddy, Anders Krogh and Graeme Mitchison, Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.
2. Neil Jones and Pavel Pevzner, An Introduction to Bioinformatics Algorithms.
3. Richard Duda, Peter Hart, David Stork, Pattern Classification.

Each book has a different advantage. The first book is a classic one. It is heavy in math and covers much of what is in class. The book is focused on sequence alignment. As part of sequence alignment theory, the book approaches Hidden Markov Models (HMM), pairwise and multiple alignment methods, phylogenetic trees as well as a short background in probability theory.

The second book intends to balance between mathematical rigor and biological relevance. According to the author, it is a good book for undergrad students. The book includes a table that associates algorithms to biological problems.

The third book is about machine learning. It takes more of an engineering approach. It includes machine learning theory, neural network and, as the name suggests, pattern recognition.

## 1.4 Crash Course in Molecular Biology

For the primarily computational students, we provide a brief introduction to the key notions of molecular biology that we will encounter throughout the term.

### 1.4.1 The Central Dogma of Molecular Biology

[lecture1\\_transcript.html#CentralDogma](#)  $DNA \rightarrow RNA \rightarrow Protein$

The central dogma of molecular biology describes how genetic information is stored and interpreted in the cell: The genetic code of an organism is stored in DNA, which is transcribed into RNA, which is finally translated into protein. Proteins carry out the majority of cellular functions such as motility, metabolism, DNA regulation, replication, and many more.

Though the central dogma holds true in most situations, there are a number of notable exceptions to the model. For instance, retroviruses are able to generate DNA from RNA via reverse-transcription. In addition, some viruses are so primitive that they do not even have DNA, instead only using RNA to make protein.

### Did You Know?

The central dogma is sometimes **incorrectly** interpreted too strongly as meaning that DNA only stores immutable information from one generation to the next that remains identical within a generation, that RNA is only used as a temporary information transfer medium, and that proteins are the only molecule that can carry out complex actions.

Again, there are many exceptions to this interpretation. For example:

- Somatic mutations can alter the DNA within a generation, and different cells can have different DNA content.
- Some cells undergo programmed DNA alterations during maturation, resulting in different DNA content: most famously, the B and T immunity while blood cells.
- Epigenetic modifications of the DNA can be inherited from one generation to the next.
- RNA can play many diverse roles in gene regulation, metabolic sensing, and enzymatic reactions. These functions were previously thought to be limited to proteins.
- Proteins themselves can undergo conformational changes that are epigenetically inherited; notably, prion states were famously responsible for mad cow disease.

## 1.4.2 DNA



### DNA function

The DNA molecule stores the genetic information of an organism. DNA contains regions called genes, which encode proteins to be produced. Other regions of the DNA contain regulatory elements, which partially influence the level of expression of each gene. Within the genetic code of DNA lies both the data about the proteins that need to be encoded and the control circuitry in the form of regulatory motifs.

### DNA structure

DNA is composed of four **nucleotides**: A (**adenine**), C (**cytosine**), T (**thymine**), and G (**guanine**). A and G are purines, which have two rings, while C and T are pyrimidines, with one ring. A and T are connected by two hydrogen bonds, while C and G are connected by three bonds. Therefore, the A-T pairing is weaker than the C-G pairing. (For this reason, the genetic composition of bacteria that live in hot springs is 80% G-C). [lecture1\\_transcript.html#Complementarity](#)

The two DNA strands in the double helix are **complementary**, meaning that if there is an A on one strand, it will be bonded to a T on the other, and if there is a C on one strand, it will be bonded to a G on the other. The DNA strands also have **directionality**, which refers to the positions of the pentose ring where the phosphate backbone connects. This directionality convention comes from the fact that DNA and RNA polymerase synthesize in the 5' to 3' direction. With this in mind, we can say that the DNA strands are **anti-parallel**, as the 5' end of one strand is adjacent to the 3' end of the other. As a result, DNA can be read both in the 3' to 5' direction and the 5' to 3' direction, and genes and other functional elements can be found in each. By convention, DNA is written from 5' to 3'. The 5' and 3' directions refer to the positions on the pentose ring where the phosphate backbone connects.

Base pairing between nucleotides of DNA constitutes its primary and secondary structure. In addition to DNA's secondary structure, there are several extra levels of structure that allow DNA to be tightly compacted and influence gene expression (Figure 3). The tertiary structure describes the twist in the DNA ladder that forms a helical shape. In the quaternary structure, DNA is tightly wound around small proteins called histones. These DNA-histone complexes are further wound into tighter structures seen in chromatin.

Before DNA can be replicated or transcribed into RNA, the chromatin structure must be locally "unpacked". Thus, gene expression may be regulated by modifications to the chromatin structure, which make it easier

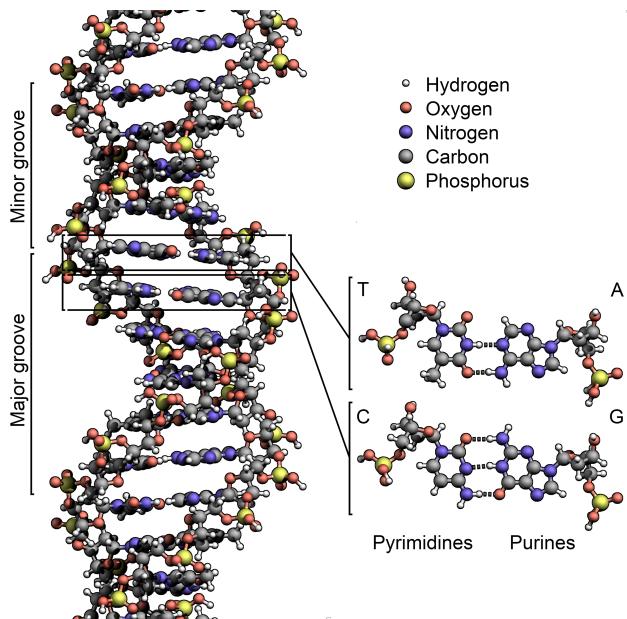


Figure 1.2: The double-helix structure of DNA. Nucleotides are in the center, and the sugar-phosphate backbone lies on the outside.

or harder for the DNA to be unpacked. This regulation of gene expression via chromatin modification is an example of epigenetics.

### DNA replication

The structure of DNA, with its weak hydrogen bonds between the bases in the center, allows the strands to easily be separated for the purpose of DNA replication (the capacity for DNA strands to be separated also allows for transcription, translation, recombination, and DNA repair, among others). This was noted by Watson and Crick as “It has not escaped our notice that the specific pairing that we have postulated immediately suggests a possible copying mechanism for the genetic material.” In the replication of DNA, the two complementary strands are separated, and each of the strands are used as templates for the construction of a new strand.

DNA polymerases attach to each of the strands at the origin of replication, reading each existing strand from the 3' to 5' direction and placing complementary bases such that the new strand grows in the 5' to 3' direction. Because the new strand must grow from 5' to 3', one strand (the leading strand) can be copied continuously, while the other (the lagging strand) grows in pieces which are later glued together by DNA ligase. The end result is 2 double-stranded pieces of DNA, where each is composed of 1 old strand, and 1 new strand; for this reason, DNA replication is semiconservative.

Many organisms have their DNA broken into several chromosomes. Each chromosome contains two strands of DNA, which are complementary to each other but are read in opposite directions. Genes can occur on either strand of DNA. The DNA before a gene (in the 5' region) is considered “upstream” whereas the DNA after a gene (in the 3' region) is considered “downstream”.

### 1.4.3 Transcription

[lecture1\\_transcript.html#Transcription](#)

DNA → RNA → Protein

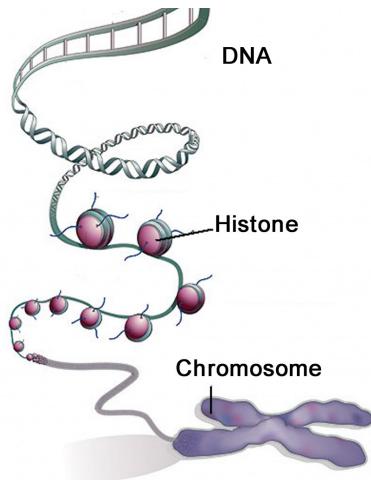


Figure 1.3: DNA is packed over several layers of organization into a compact chromosome.

Figure 1.4: RNA is produced from a DNA template during transcription. A “bubble” is opened in the DNA, allowing the RNA polymerase to enter and place down bases complementary to the DNA.

### mRNA generation

Transcription is the process by which RNA is produced using a DNA template. The DNA is partially unwound to form a “bubble”, and RNA polymerase is recruited to the transcription start site (TSS) by regulatory protein complexes. RNA polymerase reads the DNA from the 3’ to 5’ direction and placing down complementary bases to form messenger RNA (mRNA). RNA uses the same nucleotides as DNA, except Uracil is used instead of Thymine.

### Post-transcriptional modifications

mRNA in eukaryotes experience post-translational modifications, or processes that edit the mRNA strand further. Most notably, a process called splicing removes **introns**, intervening regions which don’t code for protein, so that only the coding regions, the **exons**, remain. Different regions of the primary transcript may be spliced out to lead to different protein products (alternative splicing). In this way, an enormous number of different molecules may be generated based on different splicing permutations.

In addition to splicing, both ends of the mRNA molecule are processed. The 5’ end is capped with a modified guanine nucleotide. At the 3’ end, roughly 250 adenine residues are added to form a poly(A) tail.

### 1.4.4 RNA

[lecture1\\_transcript.html#RNA](#)



RNA is produced when DNA is transcribed. It is structurally similar to DNA, with the following major differences:

1. The nucleotide uracil (U) is used instead of DNA’s thymine (T).
2. RNA contains ribose instead of deoxyribose (deoxyribose lacks the oxygen molecule on the 2’ position found in ribose).
3. RNA is single-stranded, whereas DNA is double-stranded.

RNA molecules are the intermediary step to code a protein. RNA molecules also have catalytic and regulatory functions. One example of catalytic function is in protein synthesis, where RNA is part of the ribosome.

There are many different types of RNA, including:

1. **mRNA** (messenger RNA) contains the information to make a protein and is translated into protein sequence.
2. **tRNA** (transfer RNA) specifies codon-to-amino-acid translation. It contains a 3 base pair anti-codon complementary to a codon on the mRNA, and carries the amino acid corresponding to its anticodon attached to its 3' end.
3. **rRNA** (ribosomal RNA) forms the core of the ribosome, the organelle responsible for the translation of mRNA to protein.
4. **snRNA** (small nuclear RNA) is involved in splicing (removing introns from) pre- mRNA, as well as other functions.
5. **miRNA** (micro RNA) is involved in regulation of post transcriptional gene expression.

Other functional kinds of RNA exist and are still being discovered. Though proteins are generally thought to carry out essential cellular functions, RNA molecules can have complex three-dimensional structures and perform diverse functions in the cell.

According to the “RNA world” hypothesis, early life was based entirely on RNA. RNA served as both the information repository (like DNA today) and the functional workhorse (like protein today) in early organisms. Protein is thought to have arisen afterwards via ribosomes, and DNA is thought to have arisen last, via reverse transcription.

#### 1.4.5 Translation

[lecture1\\_transcript.html#Translation](#)

DNA → RNA → Protein

##### Translation

Unlike transcription, in which the nucleotides remained the means of encoding information in both DNA and RNA, when RNA is translated into protein, the primary structure of the protein is determined by the sequence of amino acids of which it is composed. Since there are 20 amino acids and only 4 nucleotides, 3-nucleotides sequences in mRNA, known as codons, encode for each of the 20 amino acids.

Each of the 64 possible 3-sequences of nucleotides (codon) uniquely specifies either a particular amino acid, or is a stop codon that terminates protein translation (the start codon also encodes methionine). Since there are 64 possible codon sequences, the code is degenerate, and some amino acids are specified by multiple encodings. Most of the degeneracy occurs in the 3rd codon position.

##### Post-translational modifications

Like mRNA, protein also undergo further modifications that affect its structure and function. One type of post-translational modification (PTM) involves introducing new functional groups to the amino acids. Most notably, phosphorylation is the process by which a phosphate group is added onto an amino acid which can activate or deactivate the protein entirely. Many signaling pathways utilize the phosphorylation mechanism (or dephosphorylation) to turn on or off the pathway in response to stimulus. Another type of PTM is cleavage of peptide bonds. For example, the hormone insulin is cleaved twice following the formation of disulfide bonds within the original protein.

SECOND POSITION					
U	U	serine	tyrosine	cysteine	U C A G
	leucine		stop	stop	
			stop	tryptophan	
C	leucine	proline	histidine	arginine	U C A G
			glutamine		
	isoleucine		asparagine	serine	U C A G
A	* methionine	threonine	lysine	arginine	
			aspartic acid	glycine	U C A G
G	valine	alanine	glutamic acid		

\* and start

Figure 1.5: This codon table shows which of the 20 amino acid each of the 3-nucleotide codons in mRNA are translated into. In red are the stop codons, which terminate translation.

#### 1.4.6 Protein

DNA → RNA → Protein

Protein is the molecule responsible for carrying out most of the tasks of the cell, and can have many functions, such as enzymatic, contractile, transport, immune system, signal and receptor to name a few. Like RNA and DNA, proteins are polymers made from repetitive subunits. Instead of nucleotides, however, proteins are composed of amino acids.

Each amino acid has special properties of size, charge, shape, and acidity. As such, additional structure emerges beyond simply the sequence of amino acids (the primary structure), as a result of interactions between the amino acids. As such, the three-dimensional shape, and thus the function, of a protein is determined by its sequence. However, determining the shape of a protein from its sequence is an unsolved problem in computational biology.

#### 1.4.7 Regulation: from Molecules to Life

[lecture1\\_transcript.html#Regulation](#)

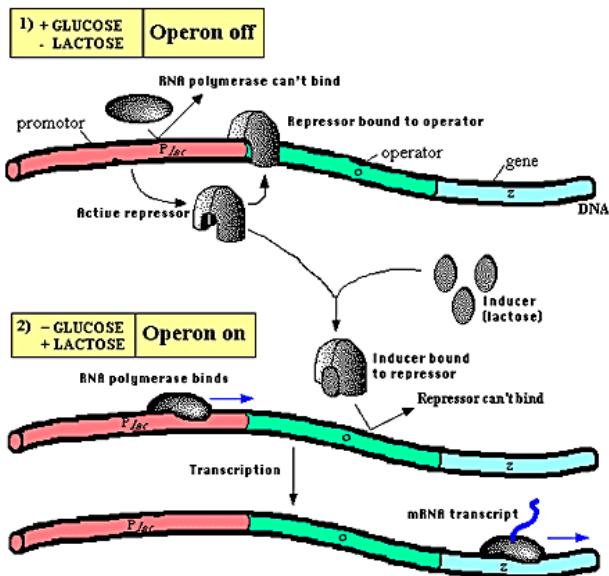
Not all genes are expressed at the same time in a cell. For example, cells would waste energy if they produced lactose transporter in the absence of lactose. It is important for a cell to know which genes it should express and when. A regulatory network is involved to control expression level of genes in a specific circumstance.

Transcription is one of the steps at which protein levels can be regulated. The promoter region, a segment of DNA found upstream (past the 5' end) of genes, functions in transcriptional regulation. The promoter region contains motifs that are recognized by proteins called transcription factors. When bound, transcription factors can recruit RNA polymerase, leading to gene transcription. However, transcription factors can also participate in complex regulatory interactions. There can be multiple binding sites in a promoter, which can act as a logic gate for gene activation. Regulation in eukaryotes can be extremely complex, with gene expression affected not only by the nearby promoter region, but also by distant enhancers and repressors.

We can use probabilistic models to identify genes that are regulated by a given transcription factor. For example, given the set of motifs known to bind a given transcription factor, we can compute the probability

that a candidate motif also binds the transcription factor (see the notes for precept #1). Comparative sequence analysis can also be used to identify regulatory motifs, since regulatory motifs show characteristic patterns of evolutionary conservation.

The lac operon in *E. coli* and other bacteria is an example of a simple regulatory circuit. In bacteria, genes with related functions are often located next to each other, controlled by the same regulatory region, and transcribed together; this group of genes is called an operon. The lac operon functions in the metabolism of the sugar lactose, which can be used as an energy source. However, bacteria prefer to use glucose as an energy source, so if there is glucose present in the environment the bacteria would not want to make the proteins that are encoded by the lac operon. Therefore, transcription of the lac operon is regulated by an elegant circuit in which transcription occurs only if there is lactose but not glucose present in the environment.



### Induction of the *lac* Operon

Figure 1.6: Operon Lac illustrates a simple biological regulatory system. In the presence of glucose, genes to lactose metabolism are turned off because glucose inactivates an activator protein. In the absence of lactose, a repressor protein also turns off the operon. Lactose metabolism genes are expressed only in the presence of lactose and absence of glucose.

#### 1.4.8 Metabolism

##### [lecture1\\_transcript.html#](#)

Live organisms are made from self-organizing building blocks. Energy source is necessary for organizing blocks. The basic mechanism involved in building blocks is degrading small molecules to get energy to build big molecules. The process of degrading molecules to release energy is called catabolism and the process of using energy to assemble more complex molecules is called anabolism. Anabolism and catabolism are both metabolic processes. Metabolism regulates the flow of mass and energy in order to keep an organism in a state of low entropy.

Enzymes are a critical component of metabolic reactions. The vast majority of (but not all!) enzymes are proteins. Many biologically critical reactions have high activation energies, so that the uncatalyzed reaction would happen extremely slowly or not at all. Enzymes speed up these reactions, so that they can happen at a rate that is sustainable for the cell. In living cells, reactions are organized into metabolic pathways. A reaction may have many steps, with the products of one step serving as the substrate for the next. Also, metabolic reactions often require an investment of energy (notably as a molecule called ATP), and energy

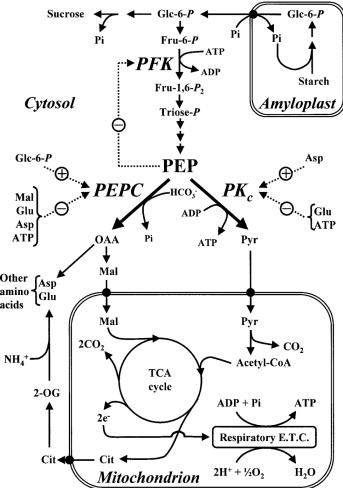


Figure 1.7: Metabolic pathways and regulation can be studied by Computational biology. Models are made from genome scale information and used to predict metabolic function and to metabolic engineering. An example of biological engineering is modifying bacteria genome to overproduce artemesinin, an antibiotic used to treat malaria.

released by one reaction may be captured by a later reaction in the pathway. Metabolic pathways are also important for the regulation of metabolic reactions—if any step is inhibited, subsequent steps may lack the substrate or the energy that they need to proceed. Often, regulatory checkpoints appear early in metabolic pathways, since if the reaction needs to be stopped, it is obviously better to stop it before much energy has been invested.

### 1.4.9 Systems Biology

#### [lecture1\\_transcript.html#SystemsBiology](#)

Systems biology strives to explore and explain the behavior that emerges from the complex interactions among the components of a biological system. One interesting recent paper in systems biology is “Metabolic gene regulation in a dynamically changing environment” (Bennett et al., 2008). This work makes the assumption that yeast is a linear, time invariant system, and runs a signal (glucose) through the system to observe the response. A periodic response to low-frequency fluctuations in glucose level is observed, but there is little response to high-frequency fluctuations in glucose level. Thus, this study finds that yeast acts as a low-pass filter for fluctuations in glucose level.

### 1.4.10 Synthetic Biology

#### [lecture1\\_transcript.html#SyntheticBiology](#)

Not only can we use computational approaches to model and analyze biological data collected from cells, but we can also design cells that implement specific logic circuits to carry out novel functions. The task of designing novel biological systems is known as synthetic biology.

A particularly notable success of synthetic biology is the improvement of artemesinin production. Artemesinin is a drug used to treat malaria. However, artemisinin was quite expensive to produce. Recently, a strain of yeast has been engineered to synthesize a precursor to artemisinic acid at half of the previous cost.

The field of synthetic biology is very promising, because it has the potential to solve many problems existing in our society today, and we will most likely need to leverage the power of computation to engineer these biological mechanisms, genes, and systems.

### 1.4.11 Model organisms and human biology

Diverse model organisms exist for all aspects of human biology. Importance of using model organisms at appropriate level of complexity.

Note: In this particular book, we'll focus on human biology, and we'll use examples from baker's yeast *Saccharomyces cerevisiae*, the fruitfly *Drosophila melanogaster*, the nematode worm *Caenorhabditis elegans*, and the house mouse *Mus musculus*. We'll deal with bacterial evolution only in the context of metagenomics of the human microbiome.

## 1.5 Introduction to algorithms and probabilistic inference

- We will quickly review some basic probability by considering an alternate way to represent motifs: a *position weight matrix* (PWM). We would like to model the fact that proteins may bind to motifs that are not fully specified. That is, some positions may require a certain nucleotide (e.g. A), while others positions are free to be a subset of the 4 nucleotides (e.g. A or C). A PWM represents the set of all DNA sequences that belong to the motif by using a matrix that stores the probability of finding each of the 4 nucleotides in each position in the motif. For example, consider the following PWM for a motif with length 4:

	1	2	3	4
A	0.6	0.25	0.10	1.0
G	0.4	0.25	0.10	0.0
T	0.0	0.25	0.40	0.0
C	0.0	0.25	0.40	0.0

We say that this motif can generate sequences of length 4. PWMs typically assume that the distribution of one position is not influenced by the base of another position. Notice that each position is associated with a probability distribution over the nucleotides (they sum to 1 and are nonnegative).

- We can also model the *background distribution* of nucleotides (the distribution found across the genome):

A	0.1
G	0.4
T	0.1
C	0.4

Notice how the probabilities for A and T are the same and the probabilities of G and C are the same. This is a consequence of the complementarity DNA which ensures that the overall composition of A and T, G and C is the same overall in the genome.

- Consider the sequence  $S = \text{GCAA}$ .

The probability of the motif generating this sequence is  $P(S|M) = 0.4 \times 0.25 \times 0.1 \times 1.0 = 0.01$ .

The probability of the background generating this sequence  $P(S|B) = 0.4 \times 0.4 \times 0.1 \times 0.1 = 0.0016$ .

- Alone, this isn't particularly interesting. However, given fraction of sequences that are generated by the motif, e.g.  $P(M) = 0.1$ , and assuming all other sequences are generated by the background ( $P(B) = 0.9$ ) we can compute the probability that the motif generated the sequence using Bayes' Rule:

$$\begin{aligned}
 P(M|S) &= \frac{P(S|M)P(M)}{P(S)} \\
 &= \frac{P(S|M)P(M)}{P(S|B)P(B) + P(S|M)P(M)} \\
 &= \frac{0.01 \times 0.1}{0.0016 \times 0.9 + 0.01 \times 0.1} = 0.40984
 \end{aligned}$$

- 1.5.1 Probability distributions
- 1.5.2 Graphical probabilistic models
- 1.5.3 Bayes rules: priors, likelihood, posterior
- 1.5.4 Markov Chains and Sequential Models
- 1.5.5 Probabilistic inference and learning
- 1.5.6 Max Likelihood and Max A Posteriori Estimates

## Bibliography

[1] lec1test. lec1test, lec1test.



# **Part I**

## **Comparing Genomes**



---

CHAPTER  
**TWO**

---

## SEQUENCE ALIGNMENT AND DYNAMIC PROGRAMMING

Guilherme Issao Fujiwara, Pete Kruskal (2007)  
Arkajit Dey, Carlos Pards (2008)  
Victor Costan, Marten van Dijk (2009)  
Andreea Bodnari, Wes Brown (2010)  
Sarah Spencer (2011)  
Nathaniel Parrish (2012)  
Clément Pit-Claudel (2014)  
Jesse Tordoff, Thrasyvoulos Karydis (2015)  
Heleen Abadiotakis, Rohil Verma (2017)

### Figures

---

2.1	Sequence alignment of Gal10-Gal1 between four yeast strains. Asterisks mark conserved nucleotides.	27
2.2	Evolutionary changes of a genetic sequence	27
2.3	Aligning human to mouse sequences is analogous to tracing	28
2.4	Examples of Fibonacci numbers in nature are ubiquitous.	29
2.5	The recursion tree for the fib procedure showing repeated subproblems. The size of the tree is $O(\phi(n))$ , where $\phi$ is the golden ratio.	30
2.6	Example of longest common substring	32
2.7	Example of longest common subsequence formulation	33
2.8	Cost matrix for matches and mismatches	33
2.9	(Example) Initial setup for Needleman-Wunsch	37
2.10	(Example) Half-way through the second step of Needleman-Wunsch	37
2.11	(Example) Tracing the optimal alignment	38
2.12	Bounded dynamic programming example	39
2.13	Recovering the sequence alignment with $O(m + n)$ space	39
2.14	Ortholog and paralog sequences	42

---

### 2.1 Introduction

Sequence alignment is a powerful tool that assesses the similarity of sequences in order to learn about their function or their evolutionary relationship. If two genetic regions are similar or identical, sequence alignment can demonstrate the conserved elements or differences between them. Evolution has preserved two broad classes of functional elements in the genome. Such preserved elements between species are often homologs<sup>1</sup> – either orthologous or paralogous sequences (refer to Appendix 2.11.1). Both classes of conserved

---

<sup>1</sup>Homologous sequences are genomic sequences descended from a common ancestor.

elements can help demonstrate the function or evolutionary history of a gene sequence. Primarily solved using computational methods (most frequently dynamic programming), sequence alignment is a fast and powerful way to find similarities among genes or genomes. These notes discuss the sequence alignment problem, the technique of dynamic programming, and a specific solution to the problem using this technique.

## 2.2 Aligning Sequences

Sequence alignment represents the method of comparing two or more genetic strands, such as DNA or RNA. These comparisons help with the discovery of genetic commonalities and with the implicit tracing of strand evolution. There are two main types of alignment:

- Global alignment: an attempt to align every element in a genetic strand, most useful when the genetic strands under consideration are of roughly equal size. Global alignment can also end in gaps.
- Local alignment: an attempt to align regions of sequences that contain similar sequence motifs within a larger context.

### 2.2.1 Example Alignment

Within orthologous gene sequences, there are islands of conservation (relatively large stretches of nucleotides that are preserved between generations). These conserved regions typically imply functional elements and vice versa. As an example, we considered the alignment of the Gal10-Gal1 intergenic region for four different yeast species, the first cross-species whole genome alignment (Figure 2.1). As we look at this alignment, we note that some areas are more similar than others, suggesting that these areas have been conserved through evolution. In particular, we note some small conserved motifs such as CGG and CGC, which in fact are functional elements in the binding of Gal4[8].<sup>2</sup> This example highlights how evolutionary data can help locate functional areas of the genome: per-nucleotide levels of conservation denote the importance of each nucleotide, and exons are among the most conserved elements in the genome.

We have to be cautious with our interpretations, however, because conservation does sometimes occur by random chance. In order to extract accurate biological information from sequence alignments we have to separate true signatures from noise. The most common approach to this problem involves modeling the evolutionary process. By using known codon substitution frequencies and RNA secondary structure constraints, for example, we can calculate the probability that evolution acted to preserve a biological function.

### 2.2.2 Solving Sequence Alignment

Genomes change over time, and the scarcity of ancient genomes makes it virtually impossible to compare the genomes of living species with those of their ancestors. Thus, we are limited to comparing just the genomes of living descendants. The goal of sequence alignment is to infer the ‘edit operations’ that change a genome by looking only at these endpoints.

We must make some assumptions when performing sequence alignment, if only because we must transform a biological problem into a computationally feasible one and we require a model with relative simplicity and tractability. In practice, sequence evolution is mostly due to nucleotide mutations, deletions, and insertions (Figure 2.2). Thus, our sequence alignment model will only consider these three operations and will ignore other realistic events that occur with lower probability (e.g. duplications).<sup>3</sup>

1. A nucleotide **mutation** occurs when some nucleotide in a sequence changes to some other nucleotide during the course of evolution.
2. A nucleotide **deletion** occurs when some nucleotide is deleted from a sequence during the course of evolution.

<sup>2</sup>Gal4 in fact displays a particular structure, comprising two arms that each bind to the same sequence, in reversed order.

<sup>3</sup>Interestingly, modeling decisions taken to improve tractability do not necessarily result in diminished relevance; for example, accounting for directionality in the study of chromosome inversions yields polynomial-time solutions to an otherwise NP problem.[6]

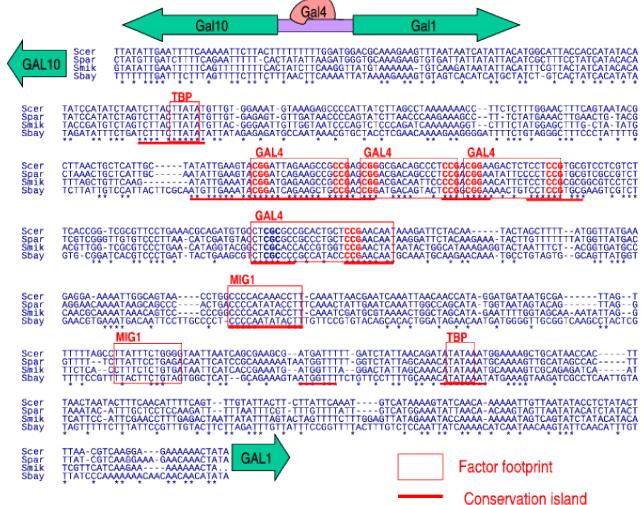


Figure 2.1: Sequence alignment of Gal10-Gal1 between four yeast strains. Asterisks mark conserved nucleotides.

3. A nucleotide **insertion** occurs when some nucleotide is added to a sequence during the course of evolution.

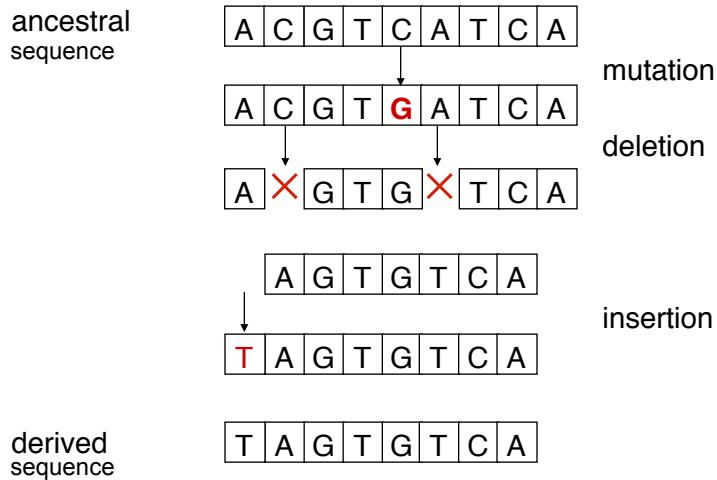


Figure 2.2: Evolutionary changes of a genetic sequence

Note that these three events are all reversible. For example, if a nucleotide N mutates into some nucleotide M, it is also possible that nucleotide M can mutate into nucleotide N. Similarly, if nucleotide N is deleted, the event may be reversed if nucleotide N is (re)inserted. Clearly, an insertion event is reversed by a corresponding deletion event.

This reversibility is part of a larger design assumption: time-reversibility. Specifically, any event in our model is reversible in time. For example, a nucleotide deletion going forward in time may be viewed as a nucleotide insertion going backward in time. This is useful because we will be aligning sequences which both exist in the present. In order to compare evolutionary relatedness, we will think of ourselves following one sequence backwards in time to a common ancestor and then continuing forward in time to the other sequence. In doing so, we can avoid the problem of not having an ancestral nucleotide sequence.

Note that time-reversibility is useful in solving some biological problems but does not actually apply to

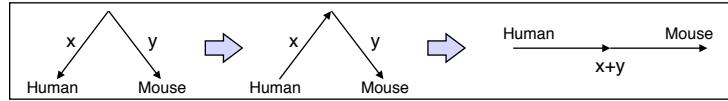


Figure 2.3: Aligning human to mouse sequences is analogous to tracing backward from the human to a common ancestor, then forward to the mouse

biological systems. For example, CpG<sup>4</sup> may incorrectly pair with a TpG or CpA during DNA replication, but the reverse operation cannot occur; hence this transformation is not time-reversible. To be very clear, time-reversibility is simply a design decision in our model; it is not inherent to the biology<sup>5</sup>.

We also need some way to evaluate our alignments. There are many possible sequences of events that could change one genome into another. Perhaps the most obvious ones minimize the number of events (i.e., mutations, insertions, and deletions) between two genomes, but sequences of events in which many insertions are followed by corresponding deletions are also possible. We wish to establish an optimality criterion that allows us to pick the ‘best’ series of events describing changes between genomes.

We choose to invoke Occam’s razor and select a maximum parsimony method as our optimality criterion. That is, in general, we wish to minimize the number of events used to explain the differences between two nucleotide sequences. In practice, we find that point mutations are more likely to occur than insertions and deletions, and certain mutations are more likely than others[11]. Our parsimony method must take these and other inequalities into account when maximizing parsimony. This leads to the idea of a substitution matrix and a gap penalty, which are developed in the following sections. Note that we did not need to choose a maximum parsimony method for our optimality criterion. We could choose a probabilistic method, for example using Hidden Markov Models (HMMs), that would assign a probability measure over the space of possible event paths and use other methods for evaluating alignments (e.g., Bayesian methods). Note the duality between these two approaches: our maximum parsimony method reflects a belief that mutation events have low probability, thus in searching for solutions that minimize the number of events we are implicitly maximizing their likelihood.

## 2.3 Dynamic Programming

Before proceeding to a solution of the sequence alignment problem, we first discuss dynamic programming, a general and powerful method for solving problems with certain types of structure.

### 2.3.1 Theory of Dynamic Programming

Dynamic programming may be used to solve problems with:

1. **Optimal Substructure:** The optimal solution to an instance of the problem contains optimal solutions to subproblems.
2. **Overlapping Subproblems:** There are a limited number of subproblems, many/most of which are repeated many times.

Dynamic programming is often used to solve optimization problems, similar to greedy algorithms. Unlike greedy algorithms, which require a greedy choice property to be valid, dynamic programming works on a range of problems in which locally optimal choices may not produce globally optimal results. Appendix 2.11.3 discusses the distinction between greedy algorithms and dynamic programming in more detail. Generally speaking, greedy algorithms solve a smaller class of problems than dynamic programming, as they usually yield the locally optimal solution.

<sup>4</sup>p denotes the phosphate backbone in a DNA strand

<sup>5</sup>This is an example where understanding the biology helps the design greatly, and illustrates the general principle that success in computational biology requires strong knowledge of the foundations of both CS and biology. Warning: computer scientists who ignore biology will work too hard.

In practice, solving a problem using dynamic programming involves two main parts: Setting up dynamic programming and then performing computation. Setting up dynamic programming usually requires the following 5 steps:

1. Find a 'matrix' parameterization of the problem. Determine the number of dimensions (variables).
2. Ensure the subproblem space is polynomial (not exponential). Note that if a small portion of subproblems are used, then memoization may be better; similarly, if subproblem reuse is not extensive, dynamic programming may not be the best solution for the problem. Memoization is the storing of results from smaller subproblems to be reused for later computations.
3. Determine an effective transversal order. Subproblems must be ready (solved) when they are needed, so computation order matters.
4. Determine a recursive formula: A larger problem is typically solved as a function of its subparts.
5. Remember choices: Typically, the recursive formula involves a minimization or maximization step. Moreover, a representation for storing transversal pointers is often needed, and the representation should be polynomial.

Once dynamic programming is setup, computation is typically straight-forward:

1. Systematically fill in the table of results (and usually traceback pointers) to find an optimal score.
2. Traceback from the optimal score through the pointers to determine an optimal solution.

### 2.3.2 Fibonacci Numbers

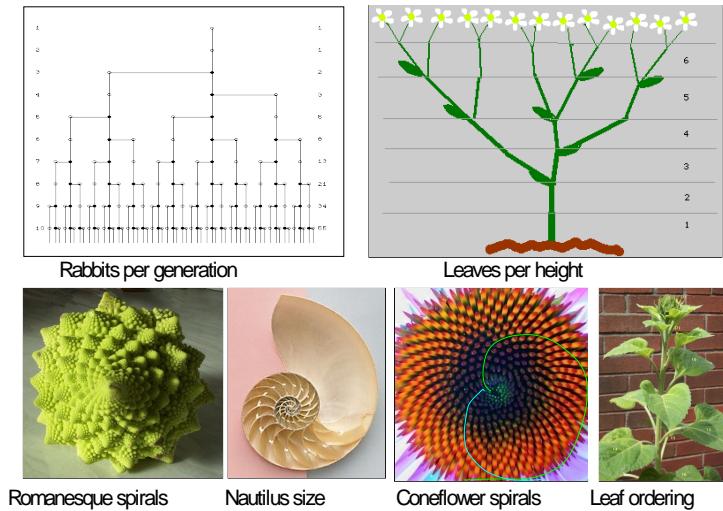


Figure 2.4: Examples of Fibonacci numbers in nature are ubiquitous.

The Fibonacci numbers provide an instructive example of the benefits of dynamic programming. The Fibonacci sequence is recursively defined as  $F_0 = F_1 = 1, F_n = F_{n-1} + F_{n-2}$  for  $n \geq 2$ . We develop an algorithm to compute the  $n^{\text{th}}$  Fibonacci number, and then refine it first using memoization and later using dynamic programming to illustrate key concepts.

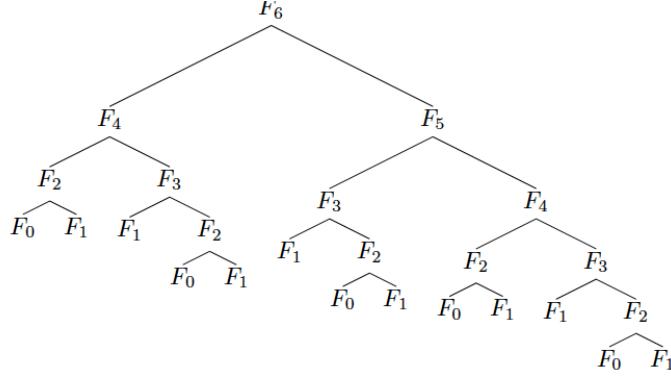


Figure 2.5: The recursion tree for the fib procedure showing repeated subproblems. The size of the tree is  $O(\phi(n))$ , where  $\phi$  is the golden ratio.

## The Naïve Solution

The simple top-down approach is to just apply the recursive definition. Listing 1 shows a simple Python implementation.

```

1 # Assume n is a non-negative integer.
2 def fib(n):
3     if n == 0 or n == 1:
4         return 1
5     else:
6         return fib(n - 1) + fib(n - 2)
  
```

Listing 2.1: Python implementation for computing Fibonacci numbers recursively.

However, this top-down algorithm runs in exponential time. That is, if  $T(n)$  is how long it takes to compute the  $n^{\text{th}}$  Fibonacci number, we have that  $T(n) = T(n - 1) + T(n - 2) + O(1)$ , so  $T(n) = O(\phi^n)$ <sup>6</sup>. The problem is that we are repeating work by solving the same subproblem many times.

## The Memoization Solution

A better solution that still utilizes the top-down approach is to memoize the answers to the subproblems. Listing 2 gives a Python implementation that uses memoization.

```

1 # Assume n is a non-negative integer.
2 fibs = {0: 1, 1: 1} # stores subproblem answers
3 def fib(n):
4     if n not in fibs:
5         x = fib(n - 2)
6         y = fib(n - 1)
7         fibs[n] = x + y
8     return fibs[n]
  
```

Listing 2.2: Python implementation for computing Fibonacci numbers using memoization.

Note that this implementation now runs in  $T(n) = O(n)$  time because each subproblem is computed at most once.

## The Dynamic Programming Solution

For calculating the  $n^{\text{th}}$  Fibonacci number, instead of beginning with  $F(n)$  and using recursion, we can start computation from the bottom (base case(s)) since we know we are going to need all of the subproblems

<sup>6</sup> $\phi$  is the **golden ratio**, i.e.  $\frac{1+\sqrt{5}}{2}$

anyway. In this way, we will omit much of the repeated work that would be done by the naïve top-down approach, and we will be able to compute the  $n^{\text{th}}$  Fibonacci number in  $O(n)$  time.

As a formal exercise, we can apply the steps outlined in section 2.3.1:

1. **Find a 'matrix' parameterization:** In this case, the matrix is one-dimensional; there is only one parameter to any subproblem  $F(x)$ .
2. **Ensure the subproblem space is polynomial:** Since there are only  $n - 1$  subproblems, the space is polynomial.
3. **Determine an effective transversal order:** As mentioned above, we will apply a bottom-up transversal order (that is, compute the subproblems in ascending order).
4. **Determine a recursive formula:** This is simply the well-known recurrence  $F(n) = F(n-1) + F(n-2)$ .
5. **Remember choices:** In this case there is nothing to remember, as no choices were made in the recursive formula.

Listing 3 shows a Python implementation of this approach.

```
1 # Assume n is a non-negative integer
2 def fib(n):
3     x = y = 1
4     for i in range(1, n):
5         x, y = y, x + y
6     return x
```

Listing 2.3: Python implementation for computing Fibonacci numbers iteratively using dynamic programming.

This method is optimized to only use constant space instead of an entire table since we only need the answer to each subproblem once. But in general dynamic programming solutions, we want to store the solutions to subproblems in a table since we may need to use them multiple times without recomputing their answers. Such solutions would look somewhat like the memoization solution in Listing 2, but they will generally be bottom-up instead of top-down. In this particular example, the distinction between the memoization solution and the dynamic programming solution is minimal as both approaches compute all subproblem solutions and use them the same number of times. In general, memoization is useful when not all subproblems will be computed, while dynamic programming saves the overhead of recursive function calls, and is thus preferable when all subproblem solutions must be calculated<sup>7</sup>. Additional dynamic programming examples may be found online [7].

### 2.3.3 Sequence Alignment using Dynamic Programming

Note that the key insight in solving the sequence alignment problem is that alignment scores are additive. This allows us to create a matrix  $M$  indexed by  $i$  and  $j$ , which are positions in two sequences  $S$  and  $T$  to be aligned. The best alignment of  $S$  and  $T$  corresponds with the best path through the matrix  $M$  after it is filled in using a recursive formula.

By using dynamic programming to solve the sequence alignment problem, we achieve a provably optimal solution, that is far more efficient than brute-force enumeration.

## 2.4 Problem Formulations

In this section, we introduce a simple problem, analyze it, and iteratively increase its complexity until it closely resembles the sequence alignment problem. This section should be viewed as a warm-up for Section 2.5 on the Needleman-Wunsch algorithm.

<sup>7</sup>In some cases dynamic programming is virtually the only acceptable solution; this is the case in particular when dependency chains between subproblems are long: in this case, the memoization-based solution recurses too deeply, and causes a stack overflow

### 2.4.1 Formulation 1: Longest Common Substring

As a first attempt, suppose we treat the nucleotide sequences as strings over the alphabet A, C, G, and T. Given two such strings, S1 and S2, we might try to align them by finding the longest common substring between them. In particular, these substrings cannot have gaps in them.

As an example, if S1 = ACGTCATCA and S2 = TAGTGTCA (refer to Figure 2.6), the longest common substring between them is GTCA. So in this formulation, we could align S1 and S2 along their longest common substring, GTCA, to get the most matches. A simple algorithm would be to try aligning S1 with different offsets of S2 and keeping track of the longest substring match found thus far. Note that this algorithm is quadratic in the length of the shortest sequence, which is slower than we would prefer for such a simple problem.

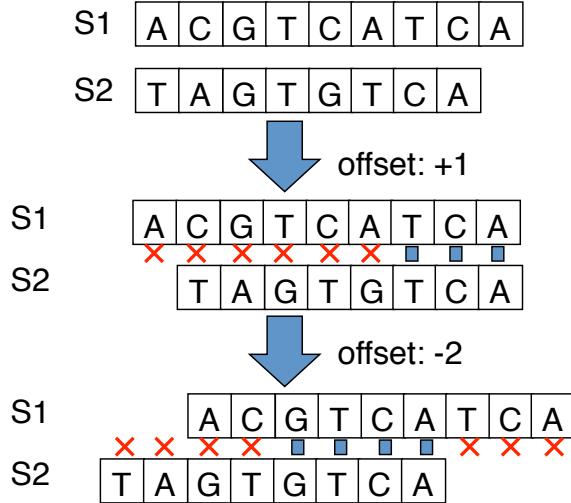


Figure 2.6: Example of longest common substring formulation

### 2.4.2 Formulation 2: Longest Common Subsequence (LCS)

Another formulation is to allow gaps in our subsequences and not just limit ourselves to substrings with no gaps. Given a sequence  $X = (x_1 \dots, x_m)$ , we formally define  $Z = (z_1, \dots, z_k)$  to be a subsequence of X if there exists a strictly increasing sequence  $i_1 < i_2 < \dots < i_k$  of indices of X such that for all  $j$ ,  $1 \leq j \leq k$ , we have  $x_{i_j} = z_j$  (CLRS 350-1).

In the longest common subsequence (LCS) problem, we're given two sequences X and Y and we want to find the maximum-length common subsequence Z. Consider the example of sequences S1 = ACGTCATCA and S2 = TAGTGTCA (refer to Figure 2.7). The longest common subsequence is AGTTCA, a longer match than just the longest common substring.

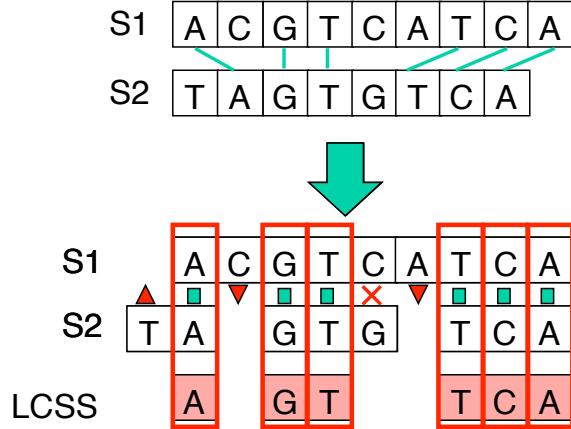


Figure 2.7: Example of longest common subsequence formulation

### 2.4.3 Formulation 3: Sequence Alignment as Edit Distance

#### Formulation

The previous LCS formulation is close to the full sequence alignment problem, but so far we have not specified any cost functions that can differentiate between the three types of edit operations (insertion, deletions, and substitutions). Implicitly, our cost function has been uniform, implying that all operations are equally likely. Since substitutions are much more likely, we want to bias our LCS solution with a cost function that prefers substitutions over insertions and deletions.

We recast sequence alignment as a special case of the classic Edit-Distance<sup>8</sup> problem in computer science (CLRS 366). We add varying penalties for different edit operations to reflect biological occurrences. One biological reasoning for this scoring decision is the probabilities of bases being transcribed incorrectly during polymerization. Of the four nucleotide bases, A and G are purines (larger, two fused rings), while C and T are pyrimidines (smaller, one ring). Thus DNA polymerase<sup>9</sup> is much more likely to confuse two purines or two pyrimidines since they are similar in structure. The scoring matrix in Figure 2.8 models the considerations above. Note that the table is symmetric - this supports our time-reversible design.

	A	G	T	C
A	+1	-½	-1	-1
G	-½	+1	-1	-1
T	-1	-1	+1	-½
C	-1	-1	-½	+1

Figure 2.8: Cost matrix for matches and mismatches

Calculating the scores implies alternating between the probabilistic interpretation of how often biological events occur and the algorithmic interpretation of assigning a score for every operation. The problem is to find the least expensive (as per the cost matrix) operation sequence which can transform the initial nucleotide sequence into the final nucleotide sequence.

<sup>8</sup>Edit-distance or Levenshtein distance is a metric for measuring the amount of difference between two sequences (e.g., the Levenshtein distance applied to two strings represents the minimum number of edits necessary for transforming one string into another).

<sup>9</sup>DNA polymerase is an enzyme that helps copy a DNA strand during replication.

### Complexity of Edit Distance

All algorithms to solve the edit distance between two strings operate in near-polynomial time. In 2015, Backurs and Indyk [?] published a proof that edit distance cannot be solved faster than  $O(n^2)$  in the general case. This result depends on the Strong Exponential Time Hypothesis (SETH), which states that NP-complete problems cannot be solved in subexponential time in the worse case.

#### 2.4.4 Formulation 4: Varying Gap Cost Models

Biologically, the cost of creating a gap is more expensive than the cost of extending an already created gap. Thus, we could create a model that accounts for this cost variation. There are many such models we could use, including the following:

- **Linear gap penalty:** Fixed cost for all gaps (same as formulation 3).
- **Affine gap penalty:** Impose a large initial cost for opening a gap, then a small incremental cost for each gap extension.
- **General gap penalty:** Allow any cost function. Note this may change the asymptotic runtime of our algorithm.
- **Frame-aware gap penalty:** Tailor the cost function to take into account disruptions to the coding frame (indels that cause frame-shifts in functional elements generally cause important phenotypic modifications).

#### 2.4.5 Enumeration

Recall that in order to solve the Longest Common Substring formulation, we could simply enumerate all possible alignments, evaluate each one, and select the best. This was because there were only  $O(n)$  alignments of the two sequences. Once we allow gaps in our alignment, however, this is no longer the case. It is a known issue that the number of all possible gapped alignments cannot be enumerated (at least when the sequences are lengthy). For example, with two sequences of length 1000, the number of possible alignments exceeds the number of atoms in the universe.

Given a metric to score a given alignment, the simple brute-force algorithm enumerates all possible alignments, computes the score of each one, and picks the alignment with the maximum score. This leads to the question, ‘How many possible alignments are there?’ If you consider only NBAs<sup>10</sup>  $n > m$ , the number of alignments is

$$\binom{n+m}{m} = \frac{(n+m)!}{n!m!} \approx \frac{(2n)!}{(n!)^2} \approx \frac{\sqrt{4\pi n} \frac{(2n)^{2n}}{e^{2n}}}{(\sqrt{2\pi n} \frac{(n)^n}{e^n})^2} = \frac{2^{2n}}{\sqrt{\pi n}} \quad (2.1)$$

This number grows extremely fast, and for values of  $n$  as small 30 is too big ( $> 10^{17}$ ) for this enumeration strategy to be feasible. Thus, using a better algorithm than brute-force is a necessity.

## 2.5 The Needleman-Wunsch Algorithm

We will now use dynamic programming to tackle the harder problem of general sequence alignment. Given two strings  $S = (S_1, \dots, S_n)$  and  $T = (T_1, \dots, T_m)$ , we want to find the longest common subsequence, which may or may not contain gaps. Rather than maximizing the length of a common subsequence we want to compute the common subsequence that optimizes the score as defined by our scoring function. Let  $d$  denote the gap penalty cost and  $s(x; y)$  the score of aligning a base  $x$  and a base  $y$ . These are inferred from insertion/deletion and substitution probabilities which can be determined experimentally or by looking at sequences that we know are closely related. The algorithm we will develop in the following sections to solve sequence alignment is known as the Needleman-Wunsch algorithm. Note that this algorithm defines a ‘global’ alignment, choosing the best possible total alignment. This is different from a ‘local’ alignment, which reaches the best possible score within regions, which we will explore later.

---

<sup>10</sup>Non-Boring Alignments, or alignments where gaps are always paired with nucleotides.

### 2.5.1 Memoization vs tabulation

Before we dive into the algorithm, a final note on memoization is in order. Much like the Fibonacci problem, the sequence alignment problem can be solved in either a top-down or bottom-up approach.

In a *top-down recursive approach* we can use memoization to create a dictionary indexed by each of the subproblems that we are solving (aligned sequences). This requires  $O(n^2m^2)$  space if we index each subproblem by the starting and end points of the subsequences for which an optimal alignment needs to be computed. The advantage is that we solve each subproblem at most once: if it is not in the dictionary, the problem gets computed and then inserted into dictionary for further reference. The disadvantage is that there will be some recursive overhead.

In a *bottom-up iterative approach* we can use tabulation. We define the order of computing sub-problems in such a way that a solution to a problem is computed once the relevant sub-problems have been solved. In particular, simpler sub-problems will come before more complex ones. This removes the need for keeping track of which sub-problems have been solved (the dictionary in memoization turns into a matrix) and ensures that there is no duplicated work (each sub-alignment is computed only once).

Note that the asymptotic time requirement in both cases is the same, but in practice with large datasets the recursive overhead may create issues. Along these same lines, the space requirement need not be the same. To build intuition for this, observe that setting an order in which problems are solved can only be done for tabulation: it cannot be easily done using memoization, since it is recursive. This gives us additional flexibility that recursive memoization lacks, as we can easily drop solutions to subproblems that we know we won't need again later. A concrete example for this is the linear space algorithm for this problem, that is difficult to implement using memoization.

### 2.5.2 Problem Statement

Suppose we have an optimal alignment for two sequences  $S_{1\dots n}$  and  $T_{1\dots m}$  in which  $S_i$  matches  $T_j$ . The key insight is that this optimal alignment is composed of an optimal alignment between  $(S_1, \dots, S_{i-1})$  and  $(T_1, \dots, T_{j-1})$  and an optimal alignment between  $(S_{i+1}, \dots, S_n)$  and  $(T_{j+1}, \dots, T_m)$ . This follows from a cut-and-paste argument: if one of these partial alignments is suboptimal, then we cut-and-paste a better alignment in place of the suboptimal one. This achieves a higher score of the overall alignment and thus contradicts the optimality of the initial global alignment. In other words, every subpath in an optimal path must also be optimal. More generally, this 'optimal substructure' is a key property of problems well suited to dynamic programming.

Notice that the scores are additive, so the score of the overall alignment equals the addition of the scores of the alignments of the subsequences. This implicitly assumes that the sub-problems of computing the optimal scoring alignments of the subsequences are independent. We need to biologically motivate that such an assumption leads to meaningful results.

### 2.5.3 Index space of subproblems

We now need to index the space of subproblems. Let  $F_{i,j}$  be the score of the optimal alignment of  $(S_1, \dots, S_i)$  and  $(T_1, \dots, T_j)$ . The space of subproblems is  $\{F_{i,j}, i \in [0, |S|], j \in [0, |T|]\}$ . This allows us to maintain an  $(m+1) \times (n+1)$  matrix  $F$  with the solutions (i.e. optimal scores) for all the subproblems.

### 2.5.4 Local optimality

We can compute the optimal solution for a subproblem by making a locally optimal choice based on the results from the smaller sub-problems. Thus, we need to establish a recursive function that shows how the solution to a given problem depends on its subproblems. And we use this recursive definition to fill up the table  $F$  in a bottom-up fashion.

We can consider the 4 possibilities (insert, delete, substitute, match) and evaluate each of them based on the results we have computed for smaller subproblems. To initialize the table, we set  $F_{0,j} = -j \cdot d$  and  $F_{i,0} = -i \cdot d$  since those are the scores of aligning  $(T_1, \dots, T_j)$  with  $j$  gaps and  $(S_1, \dots, S_i)$  with  $i$  gaps (aka zero overlap between the two sequences). Then we traverse the matrix column by column computing the optimal score for each alignment subproblem by considering the four possibilities:

- Sequence S has a gap at the current alignment position.
- Sequence T has a gap at the current alignment position.
- There is a mutation (nucleotide substitution) at the current position.
- There is a match at the current position.

We then use the possibility that produces the maximum score. We express this mathematically by the recursive formula for  $F_{i,j}$ :

$$\begin{aligned} \text{Initialization} & : \quad F(0, 0) = 0 \\ & \quad F(i, 0) = F(i - 1, 0) - d \\ & \quad F(0, j) = F(0, j - 1) - d \\ \\ \text{Iteration} & : \quad F(i, j) = \max \left\{ \begin{array}{ll} F(i - 1, j) - d & \text{insert gap in } S \\ F(i, j - 1) - d & \text{insert gap in } T \\ F(i - 1, j - 1) + s(x_i, y_j) & \text{match or mutation} \end{array} \right. \\ \\ \text{Termination} & : \quad \text{Bottom right} \end{aligned}$$

After traversing the matrix, the optimal score for the global alignment is given by  $F_{m,n}$ . The traversal order needs to be such that we have solutions to given subproblems when we need them. Namely, to compute  $F_{i,j}$ , we need to know the values to the left, up, and diagonally above  $F_{i,j}$  in the table. Thus we can traverse the table in row or column major order or even diagonally from the top left cell to the bottom right cell. Now, to obtain the actual alignment we just have to remember the choices that we made at each step. This is generally done using a separate traceback matrix. This traceback matrix stores, for each cell, the previous cell on the optimal path to that cell. We simply fill this matrix as we fill in our score matrix, so this does not take any additional asymptotic time or space.

### 2.5.5 Optimal Solution

Paths through the matrix  $F$  correspond to optimal sequence alignments. In evaluating each cell  $F_{i,j}$  we make a choice by selecting the maximum of the three possibilities. Thus the value of each (uninitialized) cell in the matrix is determined either by the cell to its left, above it, or diagonally to the left above it. A match and a substitution are both represented as traveling in the diagonal direction; however, a different cost can be applied for each, depending on whether the two base pairs we are aligning match or not. To construct the actual optimal alignment, we need to traceback through our choices in the matrix. It is helpful to maintain a pointer for each cell while filling up the table that shows which choice was made to get the score for that cell. Then we can just follow our pointers backwards to reconstruct the optimal alignment.

### 2.5.6 Solution Analysis

The runtime analysis of this algorithm is very simple. Each update takes  $O(1)$  time, and since there are  $mn$  elements in the matrix  $F$ , the total running time is  $O(mn)$ . Similarly, the total storage space is  $O(mn)$ . For the more general case where the update rule is more complicated, the running time may be more expensive. For instance, if the update rule requires testing all sizes of gaps (e.g. the cost of a gap is not linear), then the running time would be  $O(mn(m + n))$ .

### 2.5.7 Needleman-Wunsch in practice

Assume we want to align two sequences S and T, where

$$S = AGT$$

$T = \text{AAGC}$

The first step is placing the two sequences along the margins of a matrix and initializing the matrix cells. To initialize we assign a 0 to the first entry in the matrix and then fill in the first row and column based on the incremental addition of gap penalties, as in Figure 2.9 below. Although the algorithm could fill in the first row and column through iteration, it is important to clearly define and set boundaries on the problem.

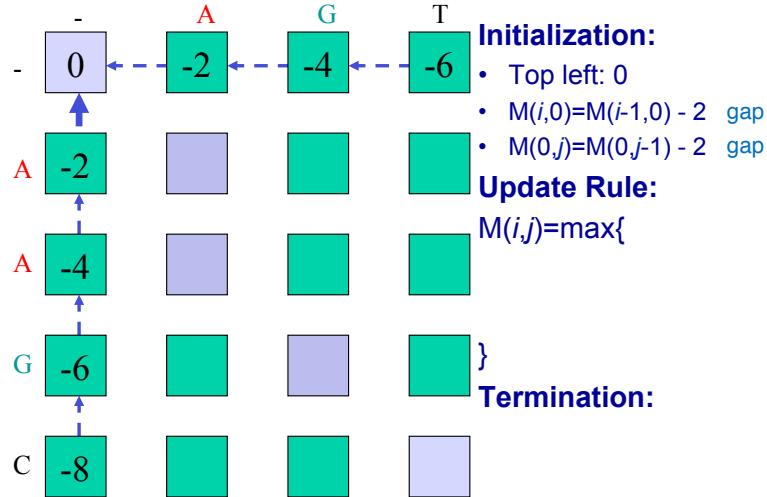


Figure 2.9: (Example) Initial setup for Needleman-Wunsch

The next step is iteration through the matrix. The algorithm proceeds either along rows or along columns, considering one cell at time. For each cell three scores are calculated, depending on the scores of three adjacent matrix cells (specifically the entry above, the one diagonally up and to the left, and the one to the left). The maximum score of these three possible tracebacks is assigned to the entry and the corresponding pointer is also stored. Termination occurs when the algorithm reaches the bottom right corner. In Figure 2.10 the alignment matrix for sequences S and T has been filled in with scores and pointers.

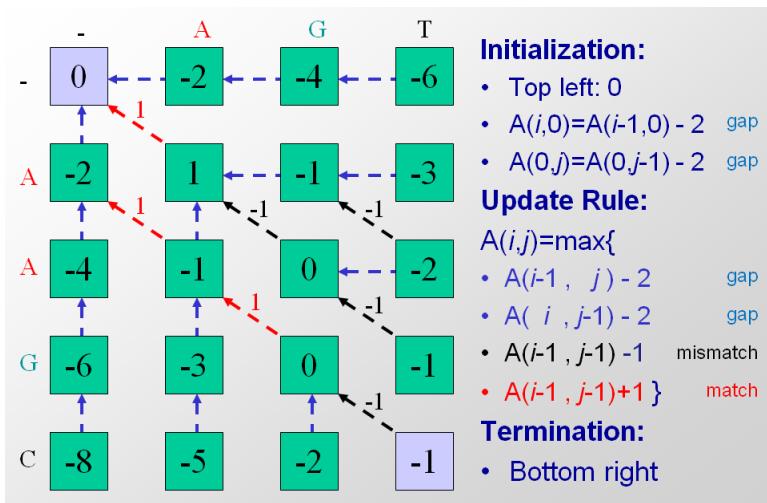


Figure 2.10: (Example) Half-way through the second step of Needleman-Wunsch

The final step of the algorithm is optimal path traceback. In our example we start at the bottom right corner and follow the available pointers to the top left corner. By recording the alignment decisions made at each cell during traceback, we can reconstruct the optimal sequence alignment from end to beginning and then invert it. Note that in this particular case, multiple optimal pathways exist (Figure 2.11). A

pseudocode implementation of the Needleman-Wunsch algorithm is included in Appendix 2.11.4

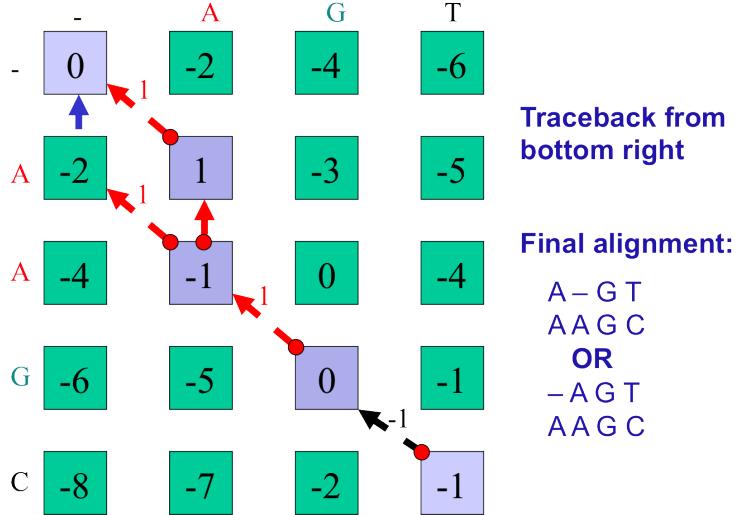


Figure 2.11: (Example) Tracing the optimal alignment

### 2.5.8 Optimizations

The dynamic algorithm we presented is much faster than the brute-force strategy of enumerating alignments and it performs well for sequences up to 10 kilo-bases long. Nevertheless, at the scale of whole genome alignments the algorithm given is not feasible. In order to align much larger sequences we can make modifications to the algorithm and further improve its performance.

#### Dynamic Programming

One possible optimization is to ignore Mildly Boring Alignments (MBAs), or alignments that have too many gaps. Explicitly, we can limit ourselves to stay within some distance  $W$  from the diagonal in the matrix  $F$  of subproblems. That is, we assume that the optimizing path in  $F$  from  $F_{0,0}$  to  $F_{m,n}$  is within distance  $W$  along the diagonal. This means that recursion (2.2) only needs to be applied to the entries in  $F$  within distance  $W$  around the diagonal, and this yields a time/space cost of  $O((m+n)W)$  (refer to Figure 2.12).

Note, however, that this strategy is heuristic and no longer guarantees an optimal alignment. Instead it attains a lower bound on the optimal score. This can be used in a subsequent step where we discard the recursions in matrix  $F$  which, given the lower bound, cannot lead to an optimal alignment.

#### Linear Space Alignment

Recursion (2.2) can be solved using only linear space: we update the columns in  $F$  from left to right during which we only keep track of the last updated column which costs  $O(m)$  space. However, besides the score  $F_{m,n}$  of the optimal alignment, we also want to compute a corresponding alignment. If we use trace back, then we need to store pointers for each of the entries in  $F$ , and this costs  $O(mn)$  space.

It is also possible to find an optimal alignment using only linear space! The goal is to use divide and conquer in order to compute the structure of the optimal alignment for one matrix entry in each step. Figure 2.13 illustrates the process. The key idea is that a dynamic programming alignment can proceed just as easily in the reverse direction, starting at the bottom right corner and terminating at the top left. So if the matrix is divided in half, then both a forward pass and a reverse pass can run at the same time and converge in the middle column. At the crossing point we can add the two alignment scores together; the cell in the middle column with the maximum score must fall in the overall optimal path.

We can describe this process more formally and quantitatively. First compute the row index  $u \in \{1, \dots, m\}$  that is on the optimal path while crossing the  $\frac{n}{2}$ th column. For  $1 \leq i \leq m$  and  $\frac{n}{2} \leq j \leq n$

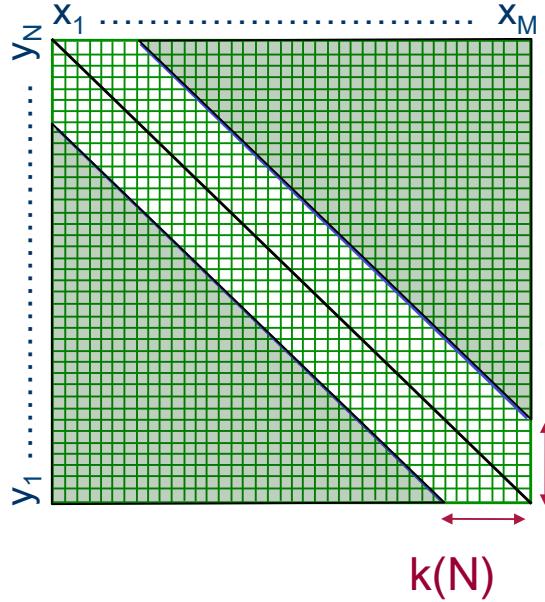
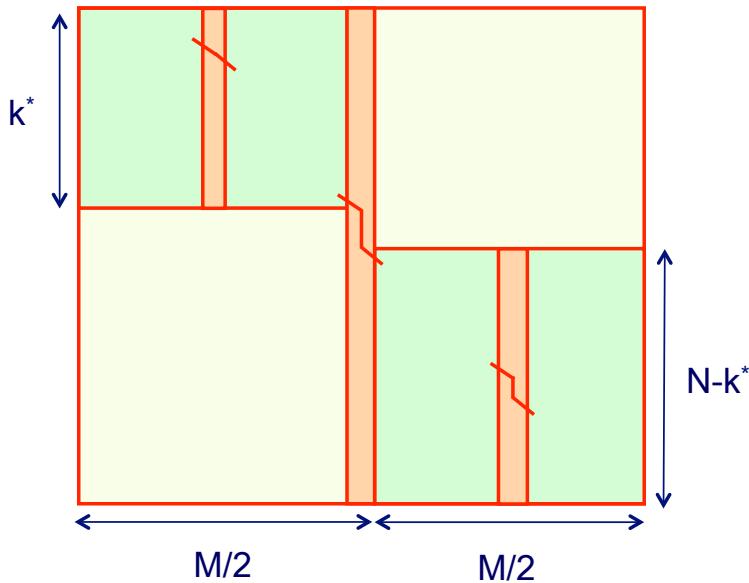


Figure 2.12: Bounded dynamic programming example


 Figure 2.13: Recovering the sequence alignment with  $O(m + n)$  space

let  $C_{i,j}$  denote the row index that is on the optimal path to  $F_{i,j}$  while crossing the  $\frac{n}{2}$ <sup>th</sup> column. Then, while we update the columns of  $F$  from left to right, we can also update the columns of  $C$  from left to right. So, in  $O(mn)$  time and  $O(m)$  space we are able to compute the score  $F_{m,n}$  and also  $C_{m,n}$ , which is equal to the row index  $u \in \{1, \dots, m\}$  that is on the optimal path while crossing the  $\frac{n}{2}$ <sup>th</sup> column.

Now the idea of divide and conquer kicks in. We repeat the above procedure for the upper left  $u \times \frac{n}{2}$  submatrix of  $F$  and also repeat the above procedure for the lower right  $(m-u) \times \frac{n}{2}$  submatrix of  $F$ . This can be done using  $O(m+n)$  allocated linear space. The running time for the upper left submatrix is  $O(\frac{un}{2})$  and the running time for the lower right submatrix is  $O(\frac{(m-u)n}{2})$ , which added together gives a running time of  $O(\frac{mn}{2}) = O(mn)$ .

We keep on repeating the above procedure for smaller and smaller submatrices of  $F$  while we gather more and more entries of an alignment with optimal score. The total running time is  $O(mn) + O(\frac{mn}{2}) + O(\frac{mn}{4}) + \dots = O(2mn) = O(mn)$ . So, without sacrificing the overall running time (up to a constant factor), divide and conquer leads to a linear space solution (see also Section ?? on Lecture 3).

## 2.6 Multiple alignment

### 2.6.1 Aligning three sequences

Now that we have seen how to align a pair of sequences, it is natural to extend this idea to *multiple* sequences. Suppose we would like to find the optimal alignment of 3 sequences. How might we proceed?

Recall that when we align two sequences  $S$  and  $T$ , we choose the maximum of three possibilities for the final position of the alignment (sequence  $T$  aligned against a gap, sequence  $S$  aligned against a gap, or sequence  $S$  aligned against sequence  $T$ ):

$$F_{i,j} = \max \begin{cases} F_{i,j-1} + d \\ F_{i-1,j} + d \\ F_{i-1,j-1} + s(S_i, T_j) \end{cases}$$

For three sequences  $S, T$ , and  $U$ , there are seven possibilities for the final position of the alignment. That is, there are three ways to have two gaps in the final position, three ways to have one gap, and one way to have all three sequences aligned ( $\binom{3}{1} + \binom{3}{2} + \binom{3}{3} = 7$ ). The update rule is now:

$$F_{i,j,k} = \max \begin{cases} F_{i-1,j,k} + s(S_i, -, -) \\ F_{i,j-1,k} + s(-, T_j, -) \\ F_{i,j,k-1} + s(-, -, U_k) \\ F_{i-1,j-1,k} + s(S_i, T_j, -) \\ F_{i-1,j,k-1} + s(S_i, -, U_k) \\ F_{i,j-1,k-1} + s(-, T_j, U_k) \\ F_{i-1,j-1,k-1} + s(S_i, T_j, U_k) \end{cases}$$

where  $s$  is the function describing gap, match, and mismatch scores.

This approach, however, is exponential in the number of sequences we are aligning. If we have  $k$  sequences of length  $n$ , computing the optimal alignment using a  $k$ -dimensional dynamic programming matrix takes  $O((2n)^k)$  time (the factor of 2 results from the fact that a  $k$ -cube has  $2^k$  vertices, so we need to take the maximum of  $2^k - 1$  neighboring cells for each entry in the score matrix). As you can imagine, this algorithm quickly becomes impractical as the number of sequences increases.

### 2.6.2 Heuristic multiple alignment

One commonly used approach for multiple sequence alignment is called *progressive multiple alignment*. Assume that we know the evolutionary tree relating each of our sequences. Then we begin by performing a pairwise alignment of the two most closely-related sequences. This initial alignment is called the *seed alignment*. We then proceed to align the next closest sequence to the seed, and this new alignment replaces the seed. This process continues until the final alignment is produced.

In practice, we generally do not know the evolutionary tree (or *guide tree*), this technique is usually paired with some sort of clustering algorithm that may use a low-resolution similarity measure to generate an estimation of the tree.

While the running time of this heuristic approach is much improved over the previous method (polynomial in the number of sequences rather than exponential), we can no longer guarantee that the final alignment is optimal.

Note that we have not yet explained how to align a sequence against an existing alignment. One possible approach would be to perform pairwise alignments of the new sequence with each sequence already in the

seed alignment (we assume that any position in the seed alignment that is already a gap will remain one). Then we can add the new sequence onto the seed alignment based on the best pairwise alignment (this approach was previously described by Feng and Doolittle[4]). Alternatively, we can devise a function for scoring the alignment of a sequence with another alignment (such scoring functions are often based on the pairwise sum of the scores at each position).

Design of better multiple sequence alignment tools is an active area of research. Section 2.9 details some of the current work in this field.

## 2.7 Current Research Directions

## 2.8 Further Reading

## 2.9 Tools and Techniques

**Lalign** finds local alignments between two sequences. **Dotlet** is a browser-based Java applet for visualizing the alignment of two sequences in a dot-matrix.

The following tools are available for multiple sequence alignment:

- **Clustal Omega** - A multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments.[10]
- **MUSCLE** - MULTiple Sequence Comparison by Log-Expectation[3]
- **T-Coffee** - Allows you to combine results obtained with several alignment methods[2]
- **MAFFT** - (Multiple Alignment using Fast Fourier Transform) is a high speed multiple sequence alignment program[5]
- **Kalign** - A fast and accurate multiple sequence alignment algorithm[9]

## 2.10 What Have We Learned?

## 2.11 Appendix

### 2.11.1 Homology

One of the key goals of sequence alignment is to identify homologous sequences (e.g., genes) in a genome. Two sequences that are homologous are evolutionarily related, specifically by descent from a common ancestor. The two primary types of homologs are orthologous and paralogous (refer to Figure 2.14<sup>11</sup>). Other forms of homology exist (e.g., xenologs), but they are outside the scope of these notes.

*Orthologs* arise from speciation events, leading to two organisms with a copy of the same gene. For example, when a single species A speciates into two species B and C, there are genes in species B and C that descend from a common gene in species A, and these genes in B and C are orthologous (the genes continue to evolve independent of each other, but still perform the same relative function).

*Paralogs* arise from duplication events within a species. For example, when a gene duplication occurs in some species A, the species has an original gene B and a gene copy B', and the genes B and B' are paralogs.

Generally, orthologous sequences between two species will be more closely related to each other than paralogous sequences. This occurs because orthologous typically (although not always) preserve function over time, whereas paralogous often change over time, for example by specializing a gene's (sub)function or by evolving a new function. As a result, determining orthologous sequences is generally more important than identifying paralogous sequences when gauging evolutionary relatedness.

---

<sup>11</sup>R.B. - BIOS 60579

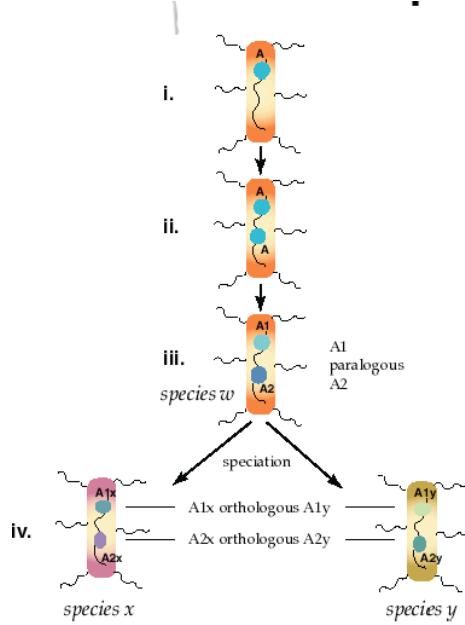


Figure 2.14: Ortholog and paralog sequences

### 2.11.2 Natural Selection

The topic of natural selection is a too large topic to summarize effectively in just a few short paragraphs; instead, this appendix introduces three broad types of natural selection: positive selection, negative selection, and neutral selection.

- *Positive selection* occurs when a trait is evolutionarily advantageous and increases an individual's fitness, so that an individual with the trait is more likely to have (robust) offspring. It is often associated with the development of new traits.
- *Negative selection* occurs when a trait is evolutionarily disadvantageous and decreases an individual's fitness. Negative selection acts to reduce the prevalence of genetic alleles that reduce a species' fitness. Negative selection is also known as purifying selection due to its tendency to 'purify' genetic alleles until only the most successful alleles exist in the population.
- *Neutral selection* describes evolution that occurs randomly, as a result of alleles not affecting an individual's fitness. In the absence of selective pressures, no positive or negative selection occurs, and the result is neutral selection.

### 2.11.3 Dynamic Programming v. Greedy Algorithms

Dynamic programming and greedy algorithms are somewhat similar, and it behooves one to know the distinctions between the two. Problems that may be solved using dynamic programming are typically optimization problems that exhibit two traits:

1. **optimal substructure** and
2. **overlapping subproblems**.

Problems solvable by greedy algorithms require both these traits as well as (3) the **greedy choice property**. When dealing with a problem “in the wild,” it is often easy to determine whether it satisfies (1) and (2) but difficult to determine whether it must have the greedy choice property. It is not always clear whether locally optimal choices will yield a globally optimal solution.

For computational biologists, there are two useful points to note concerning whether to employ dynamic programming or greedy programming. First, if a problem may be solved using a greedy algorithm, then it may be solved using dynamic programming, while the converse is not true. Second, the problem structures that allow for greedy algorithms typically do not appear in computational biology.

To elucidate this second point, it could be useful to consider the structures that allow greedy programming to work, but such a discussion would take us too far afield. The interested student (preferably one with a mathematical background) should look at matroids and greedoids, which are structures that have the greedy choice property. For our purposes, we will simply state that biological problems typically involve entities that are highly systemic and that there is little reason to suspect sufficient structure in most problems to employ greedy algorithms.

#### 2.11.4 Pseudocode for the Needleman-Wunsch Algorithm

The first problem in the first problem set asks you to finish an implementation of the Needleman-Wunsch (NW) algorithm, and working Python code for the algorithm is intentionally omitted. Instead, this appendix summarizes the general steps of the NW algorithm (Section 2.5) in a single place.

Problem: Given two sequences S and T of length m and n, a substitution matrix vU of matching scores, and a gap penalty G, determine the optimal alignment of S and T and the score of the alignment.

Algorithm:

1. Create two  $m + 1$  by  $n + 1$  matrices A and B. A will be the scoring matrix, and B will be the traceback matrix. The entry  $(i, j)$  of matrix A will hold the score of the optimal alignment of the sequences  $S[1, \dots, i]$  and  $T[1, \dots, j]$ , and the entry  $(i, j)$  of matrix B will hold a pointer to the entry from which the optimal alignment was built.
2. Initialize the first row and column of the score matrix A such that the scores account for gap penalties, and initialize the first row and column of the traceback matrix B in the obvious way.
3. Go through the entries  $(i, j)$  of matrix A in some reasonable order, determining the optimal alignment of the sequences  $S[1, \dots, i]$  and  $T[1, \dots, j]$  using the entries  $(i - 1, j - 1)$ ,  $(i - 1, j)$ , and  $(i, j - 1)$ . Set the pointer in the matrix B to the corresponding entry from which the optimal alignment at  $(i, j)$  was built.
4. Once all entries of matrices A and B are completed, the score of the optimal alignment may be found in entry  $(m, n)$  of matrix A.
5. Construct the optimal alignment by following the path of pointers starting at entry  $(m, n)$  of matrix B and ending at entry  $(0, 0)$  of matrix B.

## Bibliography

- [1] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. The MIT Press, London, third edition, 1964.
- [2] Paolo Di Tommaso, Sébastien Moretti, Ioannis Xenarios, Miquel Orobitg, Alberto Montanyola, Jia-Ming Chang, Jean-François Taly, and Cedric Notredame. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Research*, 39(Web Server issue):W13–W17, 2011.
- [3] Robert C Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–7, January 2004.
- [4] D F Feng and R F Doolittle. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution*, 25(4):351–360, 1987.
- [5] Kazutaka Katoh, George Asimenos, and Hiroyuki Toh. Multiple alignment of DNA sequences with MAFFT. *Methods In Molecular Biology Clifton Nj*, 537:39–64, 2009.

- [6] John D. Kececioglu and David Sankoff. Efficient bounds for oriented chromosome inversion distance. In *Proceedings of the 5th Annual Symposium on Combinatorial Pattern Matching*, CPM '94, pages 307–325, London, UK, UK, 1994. Springer-Verlag.
- [7] Manolis Kellis. Dynamic programming practice problems. <http://people.csail.mit.edu/bdean/6.046/dp/>, September 2010.
- [8] Manolis Kellis, Nick Patterson, Matthew Endrizzi, Bruce Birren, and Eric S Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–254, 2003.
- [9] Timo Lassmann and Erik L L Sonnhammer. Kalign—an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics*, 6(1):298, 2005.
- [10] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, Julie D Thompson, and Desmond G Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(539):539, 2011.
- [11] Zhaolei Zhang and Mark Gerstein. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Research*, 31(18):5338–5348, 2003.

---

CHAPTER  
**THREE**

---

## RAPID SEQUENCE ALIGNMENT AND DATABASE SEARCH

Isabel Chien (Sep 21, 2016)  
Heather Sweeney (Sep 20, 2015)  
Eric Bartell (Sep 20, 2015)  
Kathy Lin (Sep 11, 2014)  
Maria Rodriguez (Sep 13, 2011)  
Rushil Goel (Sep 16, 2010)  
Eric Eisner -Guilhelm Richard (Sep 17, 2009)  
Tural Badirkhnali (Sep 11, 2008)

### Figures

---

3.1	Global Alignment . . . . .	46
3.2	Global Alignment . . . . .	48
3.3	Local Alignment . . . . .	48
3.4	Local alignments to detect rearrangements . . . . .	48
3.5	Semi-global Alignment . . . . .	48
3.6	Bounded-space computation . . . . .	50
3.7	Linear-space computation for optimal alignment score . . . . .	50
3.8	Space-saving optimization for finding the optimal alignment . . . . .	50
3.9	Divide and Conquer . . . . .	51
3.10	Naive Karp-Rabin algorithm . . . . .	52
3.11	Final Karp-Rabin algorithm . . . . .	53
3.12	Pigeonhole Principle . . . . .	54
3.13	The BLAST Algorithm . . . . .	55
3.14	Nucleotide match scores . . . . .	57
3.15	BLOSUM62 matrix for amino acids . . . . .	59
3.16	Suffix tree representation of "BANANA" . . . . .	59
3.17	Sample Z vector . . . . .	61
3.18	Educated String Matching . . . . .	61
3.19	Final String Matching . . . . .	61
3.20	Knuth-Morris-Pratt Algorithm Step . . . . .	62
3.21	Boyer-Moore Algorithm Step . . . . .	62
3.22	Z Algorithm Step . . . . .	63

---

### 3.1 Introduction

In the previous chapter, we used dynamic programming to compute sequence alignments in  $O(n^2)$ . In particular, we learned the **Needleman-Wunsch** algorithm for global alignment, which matches complete sequences with one another at the nucleotide level. We usually apply this when the sequences are known to be homologous (i.e. the sequences come from organisms that share a common ancestor).

The biological significance of finding sequence alignments is to be able to infer the most likely set of evolutionary events such as point mutations/mismatches and gaps (insertions or deletions) that occurred in order to transform one sequence into the other. To do so, we first assume that the set of transformations with the lowest cost is the most likely sequence of transformations. By assigning costs to each transformation type (mismatch or gap) that reflect their respective levels of evolutionary difficulty, finding an optimal alignment reduces to finding the set of transformations that result in the lowest overall cost.

We achieve this by using a dynamic programming algorithm known as the Needleman-Wunsch algorithm. Dynamic programming uses optimal substructures to decompose a problem into similar sub-problems. The problem of finding a sequence alignment can be nicely expressed as a dynamic programming algorithm since alignment scores are additive, which means that finding the alignment of a larger sequence can be found by recursively finding the alignments of smaller subsequences. The scores are stored in a matrix, with one sequence corresponding to the columns and the other sequence corresponding to the rows. Each cell represents the transformation required between two nucleotides corresponding to the cell's row and column. An alignment is recovered by tracing back through the dynamic programming matrix (shown below). The dynamic programming approach is preferable to a greedy algorithm that simply chooses the transition with minimum cost at each step because a greedy algorithm does not guarantee that the overall result will give the optimal or lowest-cost alignment.

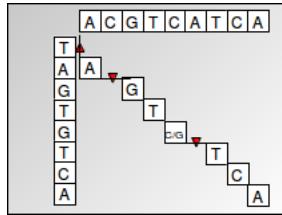


Figure 3.1: Global Alignment

To summarize the Needleman-Wunsch algorithm for global alignment:

We compute scores corresponding to each cell in the matrix and record our choice (memoization) at that step i.e. which one of the top, left or diagonal cells led to the maximum score for the current cell. We are left with a matrix full of optimal scores at each cell position, along with pointers at each cell reflecting the optimal choice that leads to that particular cell.

We can then recover the optimal alignment by tracing back from the cell in the bottom right corner (which contains the score of aligning one complete sequence with the other) by following the pointers reflecting locally optimal choices, and then constructing the alignment corresponding to an optimal path followed in the matrix.

The runtime of Needleman-Wunsch algorithm is  $O(n^2)$  since for each cell in the matrix, we do a finite amount of computation. We calculate 3 values using already computed scores and then take the maximum of those values to find the score corresponding to that cell, which is a constant time ( $O(1)$ ) operation.

To guarantee correctness, it is necessary to compute the cost for every cell of the matrix. It is possible that the optimal alignment may be made up of a bad alignment (consisting of gaps and mismatches) at the start, followed by many matches, making it the best alignment overall. These are the cases that traverse the boundary of our alignment matrix. Thus, to guarantee the optimal global alignment, we need to compute every entry of the matrix.

Global alignment is useful for comparing two sequences that are believed to be homologous. It is less useful for comparing sequences with rearrangements or inversions or aligning a newly-sequenced gene against reference genes in a known genome, known as database search. In practice, we can also often restrict the alignment space to be explored if we know that some alignments are clearly sub-optimal.

This chapter will address other forms of alignment algorithms to tackle such scenarios. It will first introduce the Smith-Waterman algorithm for local alignment for aligning subsequences as opposed to complete sequences, in contrast to the Needleman-Wunsch algorithm for global alignment. Later on, an overview will be given of hashing and semi-numerical methods like the Karp-Rabin algorithm for finding the longest (contiguous) common substring of nucleotides. These algorithms are implemented and extended for inexact matching in the BLAST program, one of the most highly cited and successful tools in computational biology. Finally, this chapter will go over BLAST for database searching as well as the probabilistic foundation of sequence alignment and how alignment scores can be interpreted as likelihood ratios.

Outline:

## 1. Introduction

- Review of global alignment (Needleman-Wunsch)

## 2. Global alignment vs. Local alignment vs. Semi-global alignment

- Initialization, termination, and update rules for Global alignment (Needleman-Wunsch) vs. Local alignment (Smith-Waterman) vs. Semi-global alignment
- Varying gap penalties, algorithmic speedups

## 3. Linear-time exact string matching

- Karp-Rabin algorithm and semi-numerical methods
- Hash functions and randomized algorithms

## 4. The BLAST algorithm and inexact matching

- Hashing with neighborhood search
- Two-hit blast and hashing with combs

## 5. Deterministic linear-time string matching

- Pre-processing
- String matching algorithms

## 6. Probabilistic foundations of sequence alignment

- Mismatch penalties, BLOSUM and PAM matrices
- Statistical significance of an alignment score

## 3.2 Global alignment vs. Local alignment vs. Semi-global alignment

A **global alignment** is defined as the *end-to-end* alignment of two strings  $s$  and  $t$  (Figure 3.2). A **local alignment** of strings  $s$  and  $t$  is an alignment of *substrings* of  $s$  with *substrings* of  $t$  (Figure 3.3). Often, we are more interested in finding local alignments because we normally do not know the boundaries of genes and because long segments often undergo rearrangements. In addition, sometimes only a small domain of the gene may be conserved. In such cases, we do not want to enforce that other (potentially non-homologous) parts of the sequence also align. Local alignment is also useful when searching for a small gene in a large chromosome (Figure 3.4). A **semi-global alignment** is essentially a global alignment in which starting and ending gaps are not penalized (Figure 3.5).

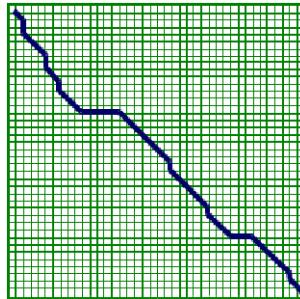


Figure 3.2: Global Alignment

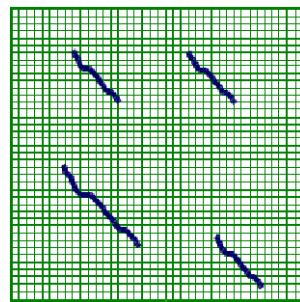


Figure 3.3: Local Alignment

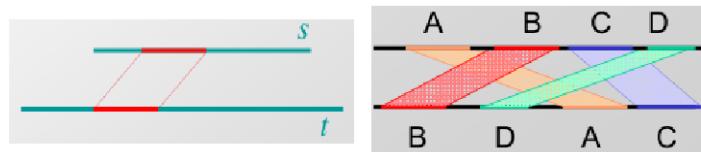


Figure 3.4: Local alignments to detect rearrangements

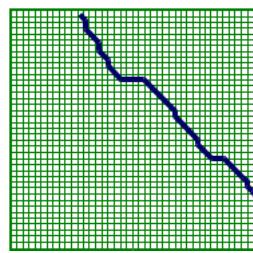


Figure 3.5: Semi-global Alignment

### 3.2.1 Using Dynamic Programming for local alignments

In this section we will see how to find local alignments by modifying the Needleman-Wunsch algorithm discussed in the previous chapter.

To find global alignments, we used the following dynamic programming algorithm (Needleman-Wunsch algorithm):

$$\begin{aligned}
 \text{Initialization} & : F(0, 0) = 0 \\
 \text{Iteration} & : F(i, j) = \max \begin{cases} F(i-1, j) - d \\ F(i, j-1) - d \\ F(i-1, j-1) + s(x_i, y_j) \end{cases} \\
 \text{Termination} & : \text{Bottom right}
 \end{aligned}$$

For finding local alignments, we only need to modify the Needleman-Wunsch algorithm slightly to start over and find a new local alignment whenever the existing alignment score goes negative. Since a local alignment can start anywhere, we initialize the first row and column in the matrix to zeros. The iteration step is modified to include a zero to include the possibility that starting a new alignment would be cheaper than having many mismatches. Furthermore, since the alignment can end anywhere, we need to traverse the entire matrix to find the optimal alignment score (not only in the bottom right corner). The rest of the algorithm, including traceback, remains unchanged, with traceback indicating an end at a zero, indicating the start of the optimal alignment.

These changes result in the following dynamic programming algorithm for local alignment, which is also known as the Smith-Waterman algorithm :

$$\begin{aligned}
 \text{Initialization} & : F(i, 0) = 0 \\
 & F(0, j) = 0 \\
 \text{Iteration} & : F(i, j) = \max \begin{cases} 0 \\ F(i-1, j) - d \\ F(i, j-1) - d \\ F(i-1, j-1) + s(x_i, y_j) \end{cases} \\
 \text{Termination} & : \text{Anywhere}
 \end{aligned}$$

For finding a semi-global alignment, the key idea is that we do not wish to penalize starting and ending gaps. Thus, we initialize the top row and leftmost column to zero to indicate an equal cost regardless of starting point, and terminate at either the bottom row or rightmost column to prevent penalizing end gaps.

The algorithm is as follows:

$$\begin{aligned}
 \text{Initialization} & : F(i, 0) = 0 \\
 & F(0, j) = 0 \\
 \text{Iteration} & : F(i, j) = \max \begin{cases} F(i-1, j) - d \\ F(i, j-1) - d \\ F(i-1, j-1) + s(x_i, y_j) \end{cases} \\
 \text{Termination} & : \text{Bottom row or Right column}
 \end{aligned}$$

### 3.2.2 Algorithmic Variations

Sometimes it can be costly in both time and space to run these alignment algorithms. Therefore, this section presents some algorithmic variations to save time and space that work well in practice.

One method to save time, is the idea of bounding the space of alignments to be explored. The idea is that good alignments generally stay close to the diagonal of the matrix. Thus we can just explore matrix cells within a radius of  $k$  from the diagonal. The problem with this modification is that this is a heuristic and can lead to a sub-optimal solution as it doesn't include the boundary cases mentioned at the beginning of the chapter. Nevertheless, this works very well in practice. In addition, depending on the properties of the scoring matrix, it may be possible to argue the correctness of the bounded-space algorithm. This algorithm requires  $O(k * m)$  space and  $O(k * m)$  time.

We saw earlier that in order to compute the optimal solution, we needed to store the alignment score in each cell as well as the pointer reflecting the optimal choice leading to each cell. However, if we are only

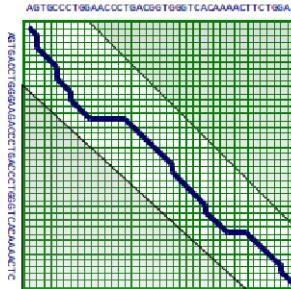


Figure 3.6: Bounded-space computation

interested in the *optimal alignment score*, and not the actual alignment itself, there is a method to compute the solution while saving space. To compute the score of any cell we only need the scores of the cell above, to the left, and to the left-diagonal of the current cell. By saving the previous and current column in which we are computing scores, the optimal solution can be computed in linear space.

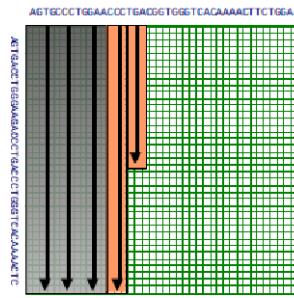


Figure 3.7: Linear-space computation for optimal alignment score

If we use the principle of divide and conquer, we can actually find the *optimal alignment* with linear space. The idea is that we compute the optimal alignments from both sides of the matrix i.e. from the left to the right, and vice versa. Let  $u = \lfloor \frac{n}{2} \rfloor$ . Say we can identify  $v$  such that cell  $(u, v)$  is on the optimal alignment path. That means  $v$  is the row where the alignment crosses column  $u$  of the matrix. We can find the optimal alignment by concatenating the optimal alignments from  $(0,0)$  to  $(u, v)$  plus that of  $(u, v)$  to  $(m, n)$ , where  $m$  and  $n$  is the bottom right cell (note: alignment scores of concatenated subalignments using our scoring scheme are additive). So we have isolated our problem to two separate problems in the top left and bottom right corners of the DP matrix. Then we can recursively keep dividing up these subproblems to smaller subproblems, until we are down to aligning 0-length sequences or our problem is small enough to apply the regular DP algorithm. To find  $v$  the row in the middle column where the optimal alignment crosses we simply add the incoming and outgoing scores for that column.

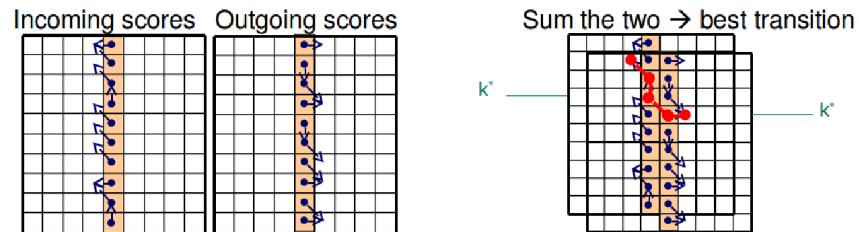


Figure 3.8: Space-saving optimization for finding the optimal alignment

One drawback of this divide-and-conquer approach is that it has a longer runtime. Nevertheless, the runtime is not dramatically increased. Since  $v$  can be found using one pass of regular DP, we can find  $v$

for each column in  $O(mn)$  time and linear space since we don't need to keep track of traceback pointers for this step. Then by applying the divide and conquer approach, the subproblems take half the time since we only need to keep track of the cells diagonally along the optimal alignment path (half of the matrix of the previous step). That gives a total run time of  $O(mn(1 + \frac{1}{2} + \frac{1}{4} + \dots)) = O(2MN) = O(mn)$  (using the sum of geometric series), to give us a quadratic run time (twice as slow as before, but still same asymptotic behavior). The total time will never exceed  $2MN$  (twice the time as the previous algorithm). Although the runtime is increased by a constant factor, one of the big advantages of the divide-and-conquer approach is that the space is dramatically reduced to  $O(N)$ .

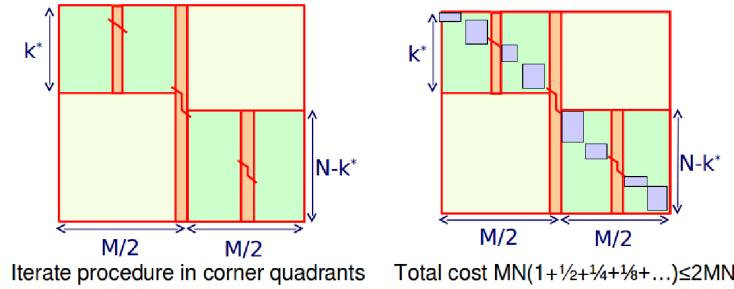


Figure 3.9: Divide and Conquer

**Q:** Why not use the bounded-space variation over the linear-space variation to get both linear time and linear space?

**A:** The bounded-space variation is a heuristic approach that can work well in practice but does not guarantee the optimal alignment.

### 3.2.3 Generalized gap penalties

Gap penalties determine the score calculated for a subsequence and thus affect which alignment is selected. There are several models that can be used, with different algorithmic tradeoffs for each.

- **Linear gap penalty:**  $w(k) = k * p$

The normal model is to use a *linear gap penalty* where each individual gap in a sequence of gaps of length  $k$  is penalized equally with value  $p$ .

- **Quadratic gap penalty:**  $w(k) = p + q * k + r * k^2$

Depending on the situation, it could be a good idea to penalize differently for, say, gaps of different lengths. In the *quadratic gap penalty*, the incremental penalty decreases quadratically as the size of the gap grows. However, the trade-off is that there is also cost associated with using more complex gap penalty function; the runtime increases substantially. This cost can be mitigated by using simpler approximations to the gap penalty functions.

- **Affine gap penalty:**  $w(k) = p + q * k$ , where  $q < p$

The *affine gap penalty* is a fine intermediate: you have a fixed penalty to start a gap and a linear cost to add to a gap.

- **Length (mod 3) gap penalty**

You can also consider more complex functions that take into consideration the properties of protein coding sequences. In the case of protein coding region alignment, a gap of length mod 3 can be less penalized because it would not result in a frame shift.

## 3.3 Linear-time exact string matching

While we have looked at various forms of alignment and algorithms used to find such alignments, these algorithms are not fast enough for some purposes. For instance, we may have a 100 nucleotide sequence

which we want to search for in the whole genome, which may be over a billion nucleotides long. In this case, we want an algorithm with a run-time that depends on the length of query sequence, possibly with some pre-processing on the database, because processing the entire genome for every query would be extremely slow. For such problems, we enter the realm of randomized algorithms where instead of worrying about the worst-case performance, we are more interested in making sure that the algorithm is linear in the expected case.

### 3.3.1 Karp-Rabin Algorithm

The problem is as follows: in text  $T$  of length  $n$  we are looking for pattern  $P$  of length  $m$ . A naive approach would be an  $O(nm)$  algorithm that compares the pattern to the text for each possible offset. The key idea of the Karp-Rabin algorithm is to interpret strings as numbers, which enables fast  $O(1)$  comparison. A naive version of the approach involves mapping the pattern  $P$  and  $m$ -length substring of  $T$  into numbers  $x$  and  $y_i$ , respectively, and comparing  $x$  and  $y_i$  for each possible subset.

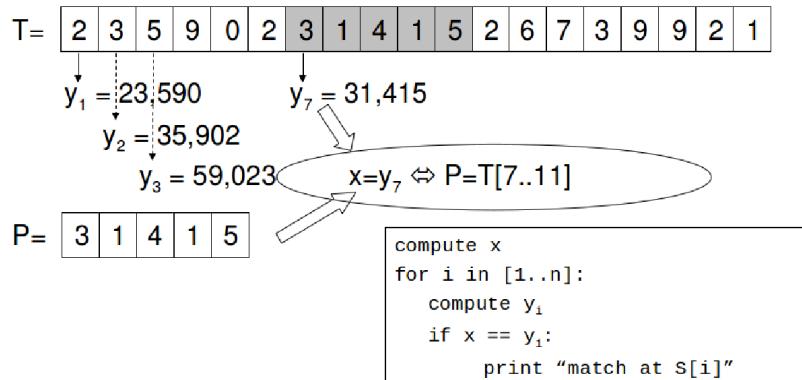


Figure 3.10: Naive Karp-Rabin algorithm

However, this approach poses a problem, because the algorithm is not actually linear for two reasons:

1. Mapping a string to a number does not take constant time, but rather linear time.
2. If the pattern to be matched is very long and almost infinite, the numbers become very long. Comparisons no longer take linear time, and space limitations become an issue.

To tackle the first problem, we first modify the procedure to calculate  $y_i$  in constant time by using the previously computed number,  $y_{i-1}$ . We can do this using some bit operations: a subtraction to remove the high-order bit, a multiplication to shift the characters left, and an addition to append the low-order digit. Since all three operations take  $O(1)$  time, this procedure takes constant time. For example, in Figure 3.10, we can compute  $y_2$  from  $y_1$  by

- removing the highest order bit:  $23590 \bmod 10000 = 3590$
- shifting left:  $3590 * 10 = 35900$
- adding the new low-order digit:  $35900 + 2 = 35902$

Our next issue arises when we have very long sequences to compare. This causes our calculations to be with very large numbers, which becomes no longer linear time. In general, in order to reduce the time for operations on arguments like numbers or strings that are really long, it is necessary to reduce the number range to something more manageable. **Hashing** is a general solution to this which involves mapping keys  $k$  from a large universe  $U$  of strings/numbers into a hash of the key  $h(k)$  which lies in a smaller range, say  $[1..m]$ . There are many hash function that can be used, all with different theoretical and practical properties. The two key properties that we need for hashing are:

1. Reproducibility – if  $x = y$ , then  $h(x) = h(y)$ . This is essential for our mapping to make sense.
2. Uniform output distribution – This implies that regardless of the input distribution, the output distribution is uniform. i.e. if  $x \neq y$ , then  $P(h(x) = h(y)) = 1/m$ , irrespective of the input distribution. This is a desirable property to reduce the chance of spurious hits.

In this algorithm, we will perform all our computations modulo  $p$ , where  $p$  reflects the word length available to us for storing numbers, but small enough such that the comparison between  $x$  and  $y_i$  is doable in constant time.

Because we are using hashing, mapping to the space of numbers modulo  $p$  can result in spurious hits due to hashing collisions, so we modify the algorithm to deal with such spurious hits by explicitly verifying reported hits of the hash values. Hence, the final version of the Karp-Rabin algorithm is:

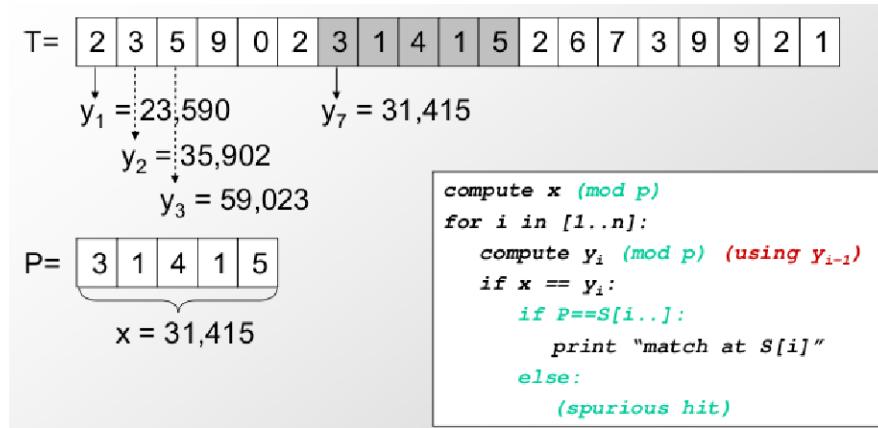


Figure 3.11: Final Karp-Rabin algorithm

Also, it is to be noted that modulo doesn't satisfy property 2 above because it is possible to have input distributions (e.g. all multiples of the number vis-à-vis which the modulo is taken) that result in a lot of collisions. Nevertheless, choosing a random number as the divisor of the modulo can avoid many collisions.

Working with hashing increases the complexity of analyzing the algorithm since now we need to compute the expected run time by including the cost of verification. To show that the expected run time is linear, we need to show that the probability of spurious hits is small.

#### Questions:

**Q:** What if there are more than 10 characters in the alphabet?

**A:** In such a case, we can just modify the above algorithm by including more digits i.e. by working in a base other than 10, e.g. say base 256. But in general, when hashing is used, strings are mapped into a space of numbers and hence the strings are interpreted numerically.

**Q:** How do we apply this to text?

**A:** A hash function is used that changes the text into numbers that are easier to compare. For example, if the whole alphabet is used, letters can be assigned a value between 0 and 25, and then be used similar to a string of numbers.

**Q:** Why does using modulus decrease the computation time?

**A:** Modulus can be applied to each individual part in the computation while preserving the answer. For instance: imagine our current text is "314152" and word length is 5. After making our first computation on "31415", we move our frame over to make our second computation, which is:

$$\begin{aligned} 14152 &= (31415 - 3 * 10000) * 10 + 2 \pmod{13} \\ &= (7 - 3 * 3) * 10 + 2 \pmod{13} \end{aligned}$$

$$= 8 \pmod{13}$$

This computation can be done now in linear time.

**Q:** Are there provisions in the algorithm for inexact matches?

**A:** The above algorithm only works when there are regions of exact similarity between the query sequence and the database. However, the BLAST algorithm, which we look at later, extends the above ideas to include the notion of searching in a biologically meaningful neighborhood of the query sequence to account for some inexact matches. This is done by searching in the database for not just the query sequence, but also some variants of the sequence up to some fixed number of changes.

### 3.4 The BLAST algorithm (Basic Local Alignment Search Tool)

The BLAST algorithm looks at the problem of sequence database search, wherein we have a query, which is a new sequence, and a target, which is a set of many old sequences, and we are interested in knowing which (if any) of the target sequences is the query related to. One of the key ideas of BLAST is that it does not require the individual alignments to be perfect; once an initial match is identified, we can fine-tune the matches later to find a good alignment which meets a threshold score. Also, BLAST exploits a distinct characteristic of database search problems: most target sequences will be completely unrelated to the query sequence, and very few sequences will match.

However, correct (near perfect) alignments will have long substrings of nucleotides that match perfectly. We base this assumption on the **Pigeonhole Principle**: if  $m$  items are put in  $n$  containers and  $m > n$ , at least 2 items must be put in the same container. Consider this scenario: you are trying to match two 9 amino-acid sequences, which have 7 identical matches and 2 mutations. You can consider a 9 amino-acid sequence to be 3 boxes with 3 amino-acids each. To introduce 2 mutations, you can randomly assign a mutation to a box. Based on the pigeonhole principle, we know that there must be a box without a mutation, and therefore we know there must be a stretch of 3 amino-acids that are perfectly conserved (Figure 3.12).

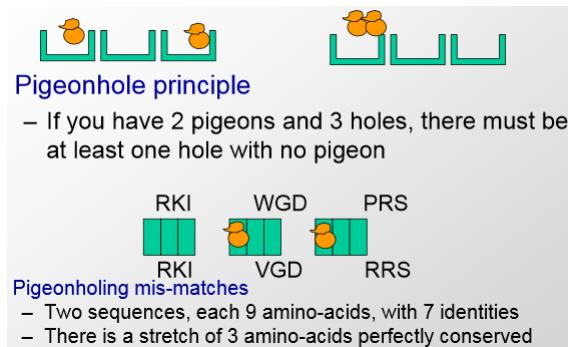


Figure 3.12: Pigeonhole Principle

In addition, in biology, functional DNA is more likely to be conserved, and therefore the mutations that we find will not actually be distributed randomly, but will be clustered in nonfunctional regions of DNA while leaving long stretches of functional DNA untouched. Therefore because of the pigeonhole principle and because highly similar sequences will have stretches of similarity, we can pre-screen the sequences for common long stretches. This idea is used in BLAST by breaking up the query sequence into  $W$ -mers, where  $W$  is an integer less than the length of the query sequence, and pre-screening the target sequences for all possible  $W$ -mers. This process will be described in more detail later.

The other aspect of BLAST that allows us to speed up repeated queries is the ability to preprocess a large database of DNA off-line. After preprocessing, searching for a sequence of length  $m$  in a database of length  $n$  will take only  $O(m)$  time. The key insights that BLAST is based on are the ideas of hashing and neighborhood search that allows one to search for  $W$ -mers, even when there are no exact-matches.

### 3.4.1 The BLAST algorithm

The steps are as follows:

1. Split query into overlapping words of length  $W$  (the  $W$ -mers)
2. Find a “neighborhood” of similar words for each word (see below)
3. Lookup each word in the neighborhood in a hash table to find the location in the database where each word occurs. Call these the *seeds*, and let  $S$  be the collection of seeds.
4. Extend the seeds in  $S$  until the score of the alignment drops off below some threshold  $X$ .
5. Report matches with overall highest scores.

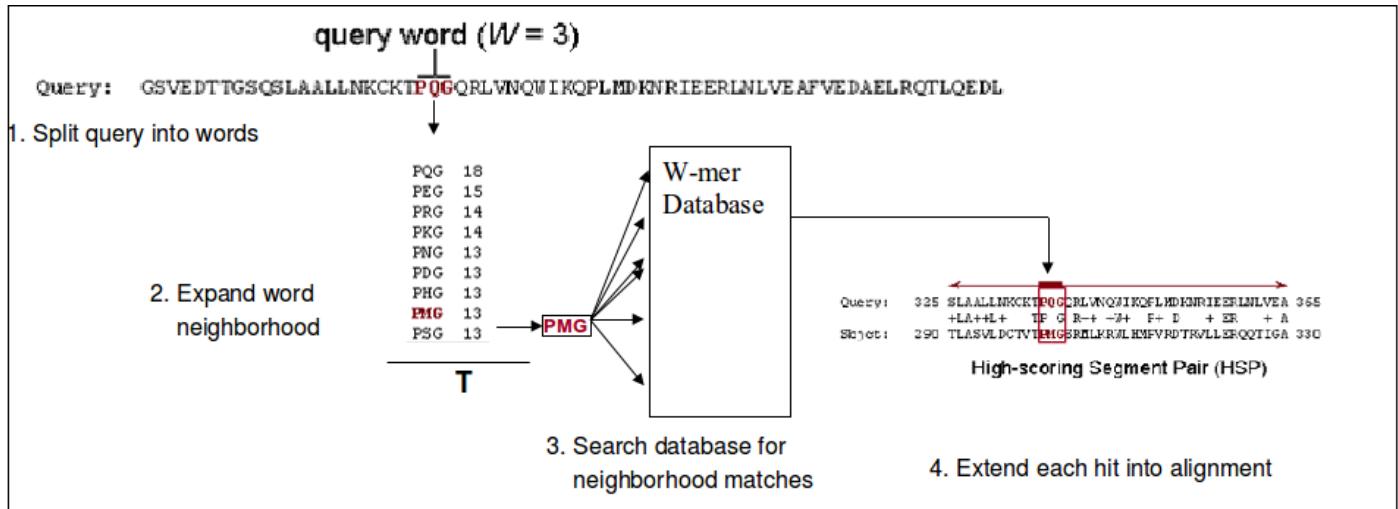


Figure 3.13: The BLAST Algorithm

In step 1, we first split the query by looking at all possible substrings of  $W$  consecutive nucleotides in the query. These are our  $W$ -mers. In step 2, to find the neighborhood of these  $W$ -mers, we then modify these sequences by changing them slightly and computing their similarity to the original sequence. We generate progressively more dissimilar words in our neighborhood until our similarity measure drops below some threshold  $T$ . This affords us flexibility to find matches that do not have exactly  $W$  consecutive matching characters in a row, but which do have enough matches to be considered similar, i.e. to meet a certain threshold score.

Then in step 3, we search for all of the words in our neighborhood in the database to find seeds of  $W$  consecutive matching nucleotides. We then extend these seeds to find our alignment using the Smith-Waterman algorithm for local alignment, until the score drops below a certain threshold  $X$ . Since the region we are considering is a much shorter segment, this will not be as slow as running the algorithm on the entire DNA database.

It is also interesting to note the influence of various parameters of BLAST on the performance of the algorithm vis-a-vis run-time and sensitivity:

- **W** Although large  $W$  would result in fewer spurious hits/collisions, thus making it faster, there are also tradeoffs associated, namely: a large neighborhood of slightly different query sequences, a large hash table, and too few hits. On the other hand, if  $W$  is too small, we may get too many hits which pushes runtime costs to the seed extension/alignment step.
- **T** If  $T$  is higher, the algorithm will be faster, but you may miss sequences that are more evolutionarily distant. If comparing two related species, you can probably set a higher  $T$  since you expect to find more matches between sequences that are quite similar.

- **X** Its influence is quite similar to  $T$  in that both will control the sensitivity of the algorithm. While  $W$  and  $T$  affect the total number of hits one gets, and hence affect the runtime of the algorithm dramatically, setting a really stringent  $X$  despite less stringent  $W$  and  $T$ , will result in runtime costs from trying unnecessary sequences that would not meet the stringency of  $X$ . So, it is important to match the stringency of  $X$  with that of  $W$  and  $T$  to avoid unnecessary computation time.

### 3.4.2 Extensions to BLAST

- **Filtering** Low complexity regions can cause spurious hits. For instance, if our query has a string of copies of the same nucleotide e.g. repeats of AC or just G, and the database has a long stretch of the same nucleotide, then there will be many many useless hits. To prevent this, we can either try to filter out low complexity portions of the query or we can ignore unreasonably over-represented portions of the database.
- **Two-hit BLAST** The idea here is to use double hashing wherein instead of hashing one long  $W$ -mer, we will hash two small W-mers. This allows us to find small regions of similarity since it is much more likely to have two smaller  $W$ -mers that match rather than one long  $W$ -mer. This allows us to get a higher sensitivity with a smaller  $W$ , while still pruning out spurious hits. This means that we'll spend less time trying to extend matches that don't actually match. Thus, this allows us to improve speed while maintaining sensitivity.

**Q:** For a long enough  $W$ , would it make sense to consider more than 2 smaller  $W$ -mers?

**A:** It would be interesting to see how the number of such  $W$ -mers influences the sensitivity of the algorithm. This is similar to using a comb, described next.

- **Combs** This is the idea of using non-consecutive  $W$ -mers for hashing. Recall from your biology classes that the third nucleotide in a triplet usually doesn't actually have an effect on which amino acid is represented. This means that each third nucleotide in a sequence is less likely to be preserved by evolution, since it often doesn't matter. Thus, we might want to look for  $W$ -mers that look similar except in every third codon. This is a particular example of a comb. A comb is simply a bit mask which represents which nucleotides we care about when trying to find matches. We explained above why 110110110 . . . (ignoring every third nucleotide) might be a good comb, and it turns out to be. However, other combs are also useful. One way to choose a comb is to just pick some nucleotides at random. Rather than picking just one comb for a projection, it is possible to randomly pick a set of such combs and project the W-mers along each of these combs to get a set of lookup databases. Then, the query string can also be projected randomly along these combs to lookup in these databases, thereby increasing the probability of finding a match. This is called Random Projection. Extending this, an interesting idea for a final project is to think of different techniques of projection or hashing that make sense biologically. One addition to this technique is to analyze false negatives and false positives, and change the comb to be more selective. Some papers that explore additions to this search include Califino-Rigoutsos'93, Buhler'01, and Indyk-Motwani'98.
- **PSI-BLAST** Position-Specific Iterative BLAST create summary profiles of related proteins using BLAST. After a round of BLAST, it updates the score matrix from the multiple alignment, and then runs subsequent rounds of BLAST, iteratively updating the score matrix. It builds a Hidden Markov Model to track conservation of specific amino acids. PSI-BLAST allows detection of distantly-related proteins.

## 3.5 Probabilistic Foundations of Sequence Alignment

As described above, the BLAST algorithm uses a scoring (substitution) matrix to expand the list of  $W$ -mers in order to look for and determine an approximately matching sequence during seed extension. Also, a scoring matrix is used in evaluating matches or mismatches in the alignment algorithms. But how do we construct this matrix in the first place? How do you determine the value of  $s(x_i, y_j)$  in global/local alignment?

	A	G	T	C
A	+1	-1/2	-1	-1
G	-1/2	+1	-1	-1
T	-1	-1	+1	-1/2
C	-1	-1	-1/2	+1

Figure 3.14: Nucleotide match scores

The idea behind the scoring matrix is that the score of alignment should reflect the probability that two similar sequences are homologous i.e. the probability that two sequences that have a bunch of nucleotides in common also share a common ancestry. For this, we look at the likelihood ratios between two hypotheses.

1. **Hypothesis 1:** – That the alignment between the two sequence is due to chance and the sequences are, in fact, unrelated.
2. **Hypothesis 2:** – That the alignment is due to common ancestry and the sequences are actually related.

Then, we calculate the probability of observing an alignment according to each hypothesis.  $Pr(x, y|U)$  is the probability of aligning  $x$  with  $y$  assuming they are unrelated, while  $Pr(x, y|R)$  is the probability of the alignment, assuming they are related. Then, we define the alignment score as the log of the likelihood ratio between the two:

$$S \equiv \log \frac{P(\mathbf{x}, \mathbf{y}|R)}{P(\mathbf{x}, \mathbf{y}|U)} \quad (3.1)$$

Since a sum of logs is a log of products, we can get the total score of the alignment by adding up the scores of the individual alignments. This gives us the probability of the whole alignment, assuming each individual alignment is independent. Thus, an additive matrix score exactly gives us the probability that the two sequences are related, and the alignment is not due to chance. More formally, considering the case of aligning proteins, for unrelated sequences, the probability of having an  $n$ -residue alignment between  $x$  and  $y$  is a simple product of the probabilities of the individual sequences since the residue pairings are independent.

That is,

$$\begin{aligned} \mathbf{x} &= \{x_1 \dots x_n\} \\ \mathbf{y} &= \{y_1 \dots y_n\} \\ q_a &= P(\text{amino acid } a) \\ P(\mathbf{x}, \mathbf{y}|U) &= \prod_{i=1}^n q_{x_i} \prod_{i=1}^n q_{y_i} \end{aligned}$$

For related sequences, the residue pairings are no longer independent so we must use a different joint probability, assuming that each pair of aligned amino acids evolved from a common ancestor:

$$\begin{aligned} p_{ab} &= P(\text{evolution gave rise to } a \text{ in } \mathbf{x} \text{ and } b \text{ in } \mathbf{y}) \\ P(\mathbf{x}, \mathbf{y}|R) &= \prod_{i=1}^n p_{x_i y_i} \end{aligned}$$

Then, the likelihood ratio between the two is given by:

$$\begin{aligned}\frac{P(\mathbf{x}, \mathbf{y}|R)}{P(\mathbf{x}, \mathbf{y}|U)} &= \frac{\prod_{i=1}^n p_{x_i y_i}}{\prod_{i=1}^n q_{x_i} \prod_{i=1}^n q_{y_i}} \\ &= \frac{\prod_{i=1}^n p_{x_i y_i}}{\prod_{i=1}^n q_{x_i} q_{y_i}}\end{aligned}$$

Since we eventually want to compute a sum of scores and probabilities require add products, we take the log of the product to get a handy summation:

$$\begin{aligned}S &\equiv \log \frac{P(\mathbf{x}, \mathbf{y}|R)}{P(\mathbf{x}, \mathbf{y}|U)} \\ v &= \sum_i \log \left( \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}} \right) \\ &\equiv \sum_i s(x_i, y_i)\end{aligned}\tag{3.2}$$

Thus, the substitution matrix score for a given pair  $a, b$  is give by

$$s(a, b) = \log \left( \frac{p_{ab}}{q_a q_b} \right)$$

The above expression is then used to crank out a substitution matrix like the BLOSUM62 for amino acids. It is interesting to note that the score of a match of an amino acid with itself depends on the amino acid itself because the frequency of random occurrence of an amino acid affects the terms used in calculating the likelihood ratio score of alignment. Hence, these matrices capture not only the sequence similarity of the alignments, but also the chemical similarity of various amino acids.

## 3.6 Deterministic linear-time exact string matching

In this section, we describe ways of pre-processing a database for rapid string lookup, each of which has both practical and theoretical importance. We also discuss algorithms that utilize these methods of pre-processing for string matching, so that in the worst case, the algorithm is guaranteed to run in linear time.

To facilitate the explanations, we will be using  $P$  to refer to the pattern being searched for, and  $T$  to refer to the complete text.

### 3.6.1 Pre-processing

The hashing technique at the core of the BLAST algorithm is a powerful way of string **pre-processing preprocessing** for rapid lookup. A substantial time is invested to process the whole genome, or a large set of genomes, *in advance* of obtaining a query sequence. Once the query sequence is obtained, it can be similarly processed and its parts searched against the indexed database in linear time. There are several different methods of pre-processing that are outline below.

#### Suffix Trees

Suffix trees provide a powerful tree representation of substrings of a target sequence  $T$ , by capturing all suffixes of  $T$  in a radix tree.

- Representation of a sequence in a suffix tree (See Figure 14)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																			C	
S	-1	4																		S	
T	-1	1	5																	T	
P	-3	-1	-1	7																P	
A	0	1	0	-1	4															A	
G	-3	0	-2	-2	0	6														G	
N	-3	1	0	-2	-2	0	6													N	
D	-3	0	-1	-1	-2	-1	1	6												D	
E	-4	0	-1	-1	-1	-2	0	2	5											E	
Q	-3	0	-1	-1	-1	-2	0	0	2	5										Q	
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8									H	
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5								R	
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5							K	
M	-1	-1	-1	-2	-1	-3	-2	-3	0	-2	-1	-1	5							M	
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	1	4					I		
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	2	2	4				L		
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	1	3	1	4			V		
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6		F		
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-1	-1	-1	-1	3	7	Y		
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Figure 3.15: BLOSUM62 matrix for amino acids

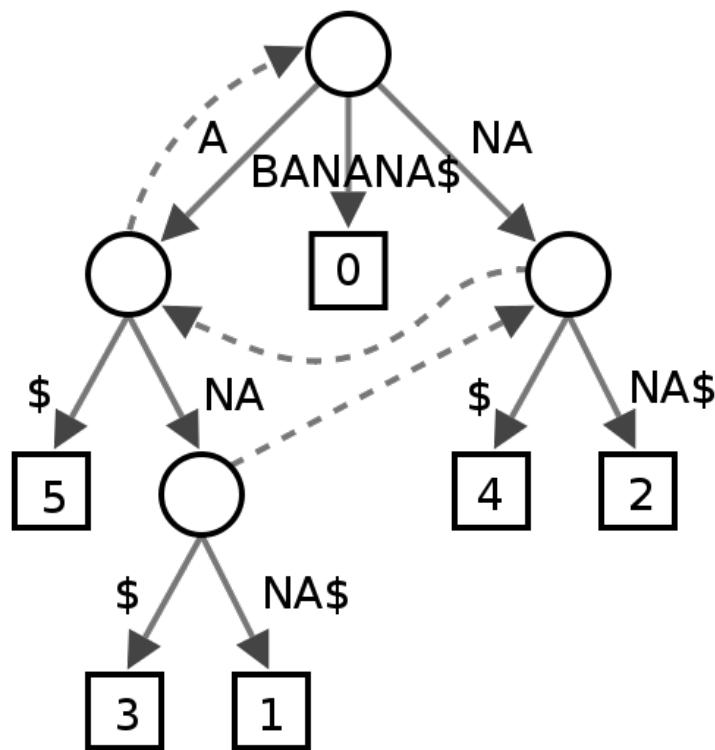


Figure 3.16: Suffix tree representation of "BANANA"

- Searching a new sequence against a suffix tree

A sequence,  $P$ , can be searched for against a suffix tree by running a depth first search algorithm starting from the root of the suffix tree. This can occur in  $O(m)$  time,  $m$  being the length of the pattern,  $P$ , under the assumption that the alphabet maintains a constant size and all inner nodes in the suffix tree are aware of the leaves beneath them.

- Linear-time construction of suffix trees
- The three key ideas for linear-time construction

### Suffix Arrays

For many genomic applications, suffix trees are too expensive to store in memory, and more efficient representations were needed. Suffix arrays were developed specifically to reduce the memory consumption of suffix trees, and achieve the same goals with a significantly reduced space need.

- Add information on suffix arrays

Using suffix arrays, any substring can be found by doing a binary search on the ordered list of suffixes. By thus exploring the prefix of every suffix, we end up searching all substrings.

### The Burrows-Wheeler Transform

An even more efficient representation than suffix trees is given by the Burrows-Wheeler Transform (BWT), which enables storing the entire hashed string in the same number of characters as the original string (and even more compactly, as it contains frequent homopolymer runs of characters that can be more easily compressed). This has helped make programs that can run even more efficiently.

We first consider the BWT matrix, which is an extension of a suffix array, in that it contains not only all suffixes in sorted (lexicographic) order, but it appends to each suffix starting at position  $i$  the prefix ending at position  $i - 1$ , each row thus containing a full rotation of the original string. This enables all the suffix-array and suffix-tree operations, of finding the position of suffixes in time linear in the query string.

The key difference from Suffix Arrays is space usage, where instead of storing all suffixes in memory, which even for suffix arrays is very expensive, only the last column of the BWT matrix is stored, based on which the original matrix can be recovered.

- Add example of compress-uncompress

An auxiliary array can be used to speed things even further and avoid having to repeat operations of finding the first occurrence of each character in the modified suffix array.

Lastly, once the positions of 100,000s of substrings are found in the modified string (the last column of the BWT matrix), these coordinates can be transformed to the original positions, saving runtime by amortizing the cost of the transformation across the many many reads.

The BWT has had a very strong impact on short-string matching algorithms, and nearly all the fastest read mappers are currently based on the Burrows-Wheeler Transform.

### Fundamental pre-processing

This is a variation of processing that has theoretical interest but has found relatively little practical use in bioinformatics. It relies on the Z vector, that contains at each position  $i$  the length of the longest prefix of a string that also matches the substring starting at  $i$ . This enables computing the  $L$  and  $R$  (Left and Right) vectors that denote the end of the longest duplicate substrings that contains the current position  $i$ .

$Z_i$  is the fundamental property of the internal redundancy structure, which is why this is termed fundamental pre-processing.

### 3.6.2 String matching algorithms

These string matching algorithms use information gathered at every comparison when matching strings to improve string matching to  $O(n)$ .

The naive algorithm is as follows: it compares its string of length  $m$  character by character to the sequence. After comparing the entire string, if there are any mismatches, it moves to the next index and tries again.

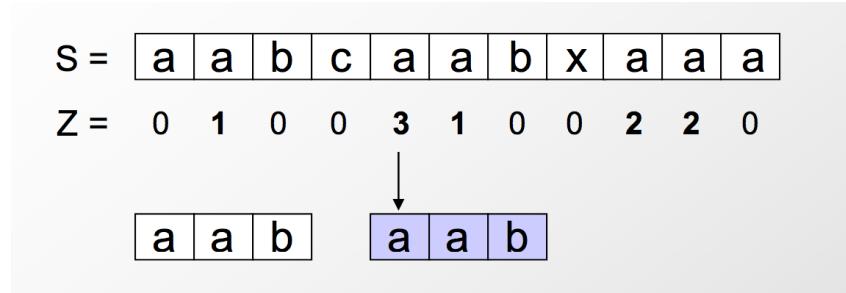


Figure 3.17: Sample Z vector

This completes in  $O(m * n)$  time.

One improvement to this algorithm is to discontinue the current comparison if a mismatch is found. However, this still completes in  $O(m * n)$  time when the string we are comparing matches the entire sequence.

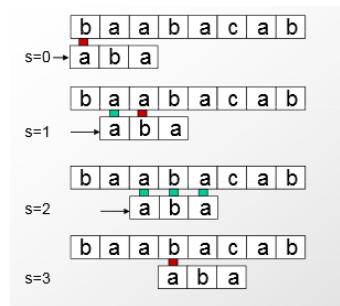


Figure 3.18: Educated String Matching

The key insight comes from learning from the internal redundancy in the string to compare, and using that to make bigger shifts down the target sequence. When a mistake is made, all bases in the current comparison can be used to move the frame considered for the next comparison further down. As seen below, this greatly reduces the number of comparisons required, decreasing runtime to  $O(n)$ .

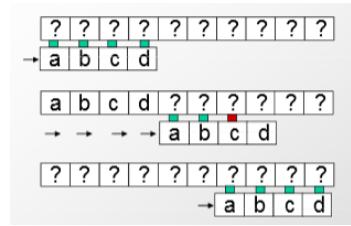


Figure 3.19: Final String Matching

In this section, we will describe three algorithms that utilize this larger shift insight as well as fundamental pre-processing.

### Knuth-Morris-Pratt Algorithm

This algorithm uses pre-processing, as well as the idea where when a mismatch occurs,  $P$  provides enough information to determine where the next match could start, thus preventing the re-examination of previously seen characters.

Pre-processing:  $Sp_i(P) =$  length of longest proper suffix of  $P[1..i]$  that matches a prefix of  $P$   
 Every comparison begins at a text position where the last comparison ended, bounded by the characters in

$T$ . Thus, every shift results in at most one extra comparison. As there are at most  $|T|$  shifts, the running times is bounded by  $2 * |T|$ , which is linear time.

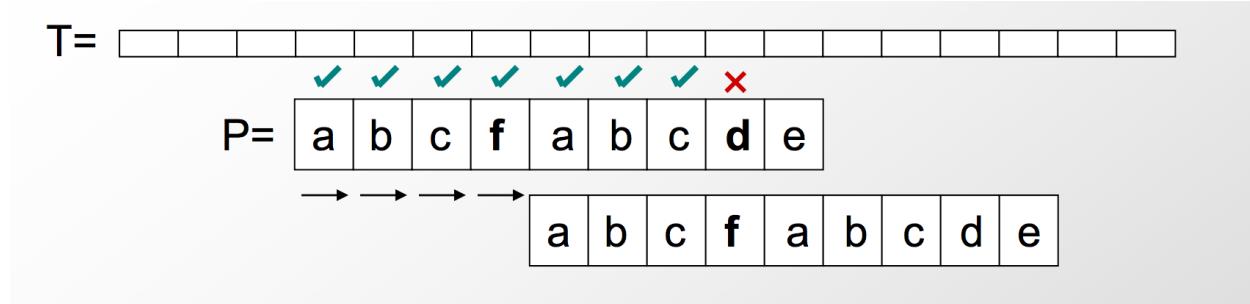


Figure 3.20: Knuth-Morris-Pratt Algorithm Step

### Boyer-Moore Algorithm

The Boyer-Moore algorithm runs on three fundamental ideas:

- Right-to-left comparison
- Alphabet-based shift rule
- Preprocessing-based shift rule

This results in a very good algorithm in practice. The second rule results in large shifts and sub-linear time, which works well with larger alphabets (as in English). The third rule ensures that it behaves linearly in the worst case, which works well with small alphabets (as in DNA).

The pattern  $P$  is what is pre-processed, which allows this algorithm to work well in cases where the pattern is much shorter than the text, or where the pattern is used for many searches.

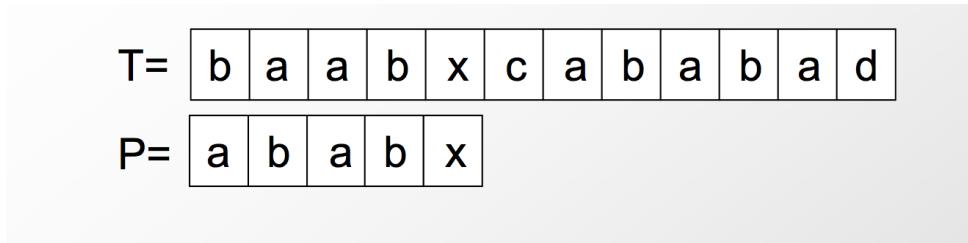


Figure 3.21: Boyer-Moore Algorithm Step

### Z Algorithm

The Z algorithm involves concatenating  $P$ , the pattern, a special character  $\$$ , and  $T$ , the text. Fundamental pre-processing is completed on the resultant string, which takes about  $O(m + n)$  time, and creates the Z vector as described above. The starting positions  $i$  for which  $Z_i = |P|$  are reported and are the matching texts.

## 3.7 Current Research Directions

## 3.8 Further Reading

- BLAST related algorithms: Califino-Rigoutsos'93, Buhler'01, and Indyk-Motwani'98

P+T= [ a | b | a | b | a | \$ | b | a | a | b | a | c | a | b | a | b | a | d ]

Figure 3.22: Z Algorithm Step

### 3.9 Tools and Techniques

### 3.10 What Have We Learned?

In this section we explored alignment algorithms beyond global alignment. We began by reviewing our use of dynamic programming to solve global alignment problems using the Needleman-Wunsch algorithm. We then explored alternatives of local (Smith-Waterman) and semi-global alignments. We then discussed using hash function to match exact strings in linear time (Karp-Rabin) as well as doing a neighborhood search, investigating similar sequences in probabilistic linear time (pigeonhole principle, combs, 2-hit blast, random projections). We have also addressed using pre-processing for linear time string matching, as well as the probabilistic background for sequence alignment.

## Bibliography



---

CHAPTER  
**FOUR**

---

## COMPARATIVE GENOMICS I: GENOME ANNOTATION

Quanquan Liu (2013),  
Mark Smith, Yarden Katz  
(Partially adapted from notes by:  
Angela Yen, Christopher Robde, Timo Somervuo and Saba Gul)

### Figures

---

4.1	Exon conservation from mammals to fish. . . . .	68
4.2	Comparative identification of functional elements in 12 <i>Drosophila</i> species. . . . .	70
4.3	A comparison between two genomic regions with different selection rates $\omega$ . . . . .	70
4.4	Unusual patterns of substitution . . . . .	71
4.5	Increase in power to detect small constrained elements . . . . .	72
4.6	HOXB5 conservation across mammalian species. . . . .	73
4.7	Modeling mutations using rate matrices. . . . .	87
4.8	Measuring genome-wide excess constraint. . . . .	88
4.9	Detecting functional elements from their evolutionary signature. <b>A</b> Distribution of constraint for the whole genome against ancestral repeats (background). <b>B</b> Difference between whole genome and background constraint. <b>C</b> Discovery of functional elements from excess constraint. Novel elements are shown in red. <b>D</b> Enrichment of elements for regions of excess constraint. . . . .	88
4.10	Coverage depth across different sets of elements. Functional elements experience greater conservation. . . . .	89
4.11	Different mutation patterns in protein-coding and non-protein-coding regions. . . . .	89
4.12	Evolutionary signatures of protein-coding genes . . . . .	89
4.13	RNA with secondary stem-loop structure . . . . .	90
4.14	Silent point mutations . . . . .	90
4.15	Amino acid circle . . . . .	90
4.16	Protein-coding vs. non-protein-coding conserved regions . . . . .	91
4.17	Reading frame conservation. . . . .	91
4.18	Rejected open reading frame. . . . .	92
4.19	Null and alternate model rate matrices. . . . .	92
4.20	Probability that a region is protein-coding . . . . .	92
4.21	Prediction of new genes and exons using evolutionary signatures. . . . .	93
4.22	OPRL1 neurotransmitter: a novel translational read-through candidate. . . . .	93
4.23	Stop codon suppression interpretations. . . . .	93
4.24	Z-curve for <i>Caki</i> . . . . .	94
4.25	Motif inside a gene . . . . .	94
4.26	miRNA hairpin structure. . . . .	94

4.27 miRNA characteristic conservation pattern.	95
4.28 Novel miRNA Evidence 1.	95
4.29 Novel miRNA Evidence 2.	95
4.30 miRNA detection decision tree.	96
4.31 CG31044 and CG33311 Transcripts.	96
4.32 TAATTA regulatory motif.	97

## 4.1 Introduction

In this chapter we will explore the emerging field of comparative genomics, primarily through examples of multiple species genome alignments (work done by the Kellis lab.) One approach to the analysis of genomes is to infer important gene functions by applying an understanding of evolution to search for expected evolutionary patterns. Another approach is to discover evolutionary trends by studying genomes themselves. Taken together, evolutionary insight and large genomic datasets offer great potential for discovery of novel biological phenomena.

A recurring theme is to take a global computational approach to analyzing elements of genes and RNAs encoded in the genome and find interesting new biological phenomena. We can do this by seeing how individual examples “divege” or differ from the average case. For example, by examining many protein-coding genes, we can identify features representative of that loci class. We can then come up with highly accurate tests for distinguishing protein-coding genes from non-protein-coding genes. Often, these computational tests, based on thousands of examples, will be far more definitive than conventional low-throughput wet lab tests. (Such tests can include mass spectrometry to detect protein products, in cases where we want to know if a particular locus is protein coding.)

### 4.1.1 Motivation and Challenge

As the cost of genome sequencing continues to drop, the availability of sequenced genome data has greatly increased. However, analysis of the data has not kept up, but there are many interesting biological phenomena lying undiscovered in the endless strings of ATGCs. The goal of comparative genomics is to leverage the vast amounts of information available and look for biological patterns.

As the name suggests, comparative genomics does not focus on one specific set of genomes. The problem with purely focusing on the single genome level is that key evolutionary signatures are missed. Comparative genomics solves this problem by comparing genomes from many species that evolved from a common ancestor. As evolution changes a species’s genome, it leaves behind traces of its presence. We will see later in this chapter that **evolution discriminates between portions of a genome on the basis of biological function.** By exploiting this correlation between evolutionary fingerprints and the biological role of a genomic subsequence, comparative genomics is able to direct wet lab research to interesting portions of the genome and discover new biological phenomena.

### FAQ

**Q:** Why do mutations only accumulate in certain regions of the genome, whereas other regions are conserved?

**A:** In non-functional regions of DNA, accumulated mutations are kept because they do not disturb the function of the DNA. In functional regions, these mutations can lead to decreased fitness; these mutations are then discarded from the species by natural selection.

We can obtain much information about evolution by studying genomics, and, similarly, we can learn about the genome through studying evolution. For example, from the principle of “survival of the fittest,” we can compare related species to discover which portions of the genome are functional elements. The evolutionary process introduces mutations into any genome. In non-functional regions of DNA, accumulated

mutations are kept because they do not disturb the function of the DNA. However, in functional regions, accumulated mutations often lead to decreased fitness. Thus, these fitness-decreasing mutations are not likely to perpetuate to future generations. As time progresses, evolutionarily unfit organisms are not likely to survive, and their genes thin out. By comparing surviving species' genomes with their ancestors' genomes, we can see which portions constitute functional elements and which constitute "junk DNA."

To date, various important biological markers and phenomena have been discovered through comparative genomics methods. For example, CRISPRs (Clustered Regularly Interspaced Short Palindromic Repeats), found in bacteria and archaea, were first discovered through comparative genomics. Follow-up experiments revealed that they provide adaptive immunity to plasmids and phages. Another example, which we will look at later in this chapter, is the phenomenon of stop-codon read-through, where stop codons are occasionally ignored during the process of translation phase of protein biosynthesis. Without comparative genomics to guide them, experimentalists might have ignored both of these features for many years.

Without a system for interpreting and identifying important features in genomes, all of the DNA sequences on earth are just a meaningless sea of data. However, we cannot ignore the importance of both computer science and biology in comparative genomics. Without knowledge of biology, one might miss the signatures of synonymous substitutions or frame shift mutations. On the other hand, ignoring computational approaches would lead to an inability to parse ever larger datasets emerging from sequencing centers. Comparative genomics require rare multidisciplinary skills and insight.

This is a particularly exciting time to enter the field of comparative genomics, because the field is mature enough that there are tools and data available to make discoveries. But it is young enough that important findings will likely continue to be made for many years.

#### 4.1.2 Importance of many closely-related genomes

In order to resolve significant biological features we need both sufficient similarity to enable comparison and sufficient divergence to identify signatures of change over evolutionary time. We improve the resolution of our analysis by extending analysis to many genomes simultaneously with some clusters of similar organisms and some dissimilar organisms. A simple analogy is one of observing an orchestra. If you place a single microphone, it will be difficult to decipher the signal coming from the entire system, because it will be overwhelmed by the local noise from the single point of observation, the nearest instrument. If you place many microphones distributed across the orchestra at reasonable distances, then you get a much better perspective not only on the overall signal, but also on the structure of the local noise. Similarly, by sequencing many genomes across the tree of life we are able to distinguish the biological signals of functional elements from the noise of neutral mutations operating at shorter time scales.

In this chapter, we will assume that we already have a complete genome-wide alignment of multiple closely-related species, spanning both coding and non-coding regions. In practice, constructing complete genome assemblies and whole-genome alignments is a very challenging problem; that will be the topic of the next chapter.

#### FAQ

**Q:** Why is there more resolving power when the evolutionary distance or branch length between species increases?

**A:** If we are comparing two species like human and chimp that are very close to each other, we expect to see little to no mutations. This gives us little discriminative power because we see no difference between the number of mutations in functional elements vs. the number of mutations in non-functional elements. However, as we increase the evolutionary time between species, we expect to see more mutations, but what we actually see are a notable decrease in the observed number of mutations in certain regions of the genome. We can conclude that these regions are functional regions. Therefore, our confidence in perceived functional elements increases as branch length increases.

## FAQ

**Q:** Why is it better to have many closely related species for the same branch length rather than one distantly related species?

**A:** As branch length increases between distantly related species, even functional elements are not conserved. Furthermore, reliably aligning genes from distantly related relatives of the same species is difficult if not impossible using current technology such as BLAST.

### 4.1.3 Comparative genomics and evolutionary signatures

Given a genome-wide alignment, we can subsequently analyze the level of conservation of functional elements in each of the genomes considered. Using the UCSC genome browser, one may see a level of conservation for every gene in the human genome derived from aligning the genomes of many other species. In Figure 4.1 below, we see a DNA sequence represented on the x-axis, while each “row” represents a different species. The y-axis within each row represents the amount of conservation for that species in that part of the chromosome (though other species that are not shown were also used to calculate conservation). Higher bars correspond with greater conservation.

From this figure, we can see that there are blocks of conservation separated by regions that are not conserved. The 12 exons (highlighted by red rectangles) are mostly conserved across species, but sometimes, certain exons are missing; for example, zebrafish is missing exon 9. However, we also see that there is a spike in some species (as circled in red) that do not correspond to a known protein coding gene. This tells us that some intronic regions have also been evolutionarily conserved, since DNA regions that do not code for proteins can still be important as functional elements, such as RNA, microRNA, and regulatory motifs. By observing how regions are conserved, instead of just looking at the amount of conservation, we can observe ‘evolutionary signatures’ of conservation for different functional elements.

The pattern of mutation/insertion/deletion can help us distinguish different types of functional elements in the genome. Different functional elements are under different selective pressures and by considering which selective pressures each element is under, we can develop evolutionary signatures characteristic of each function. For example, we see the difference in evolutionary signatures as exhibited by protein-coding genes as opposed to regulatory motifs...etc.

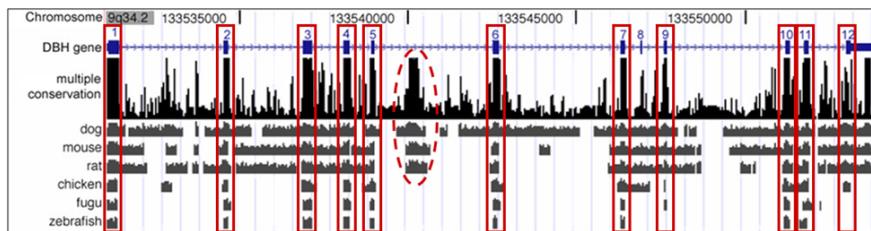


Figure 4.1: Exons (boxed in red) are deeply conserved from mammals to fish, suggesting regulatory elements. Other less universally conserved elements (the circled peak) is conserved only among mammals, suggesting a function that is present in animals but not aves or fish.

**FAQ**

**Q:** Given an alignment of genes from multiple species, what can you measure to determine the level of conservation of a specific gene(s)?

**A:** One simple method is just to look at the alignment score for each gene. If one wants to distinguish between highly conserved protein coding segments from non-protein coding segments, one may also look at codon conservation. However, in both of these approaches, we have to consider the position of each species being compared in the phylogenetic tree. A pairwise comparison score that is lower between two species separated by a greater distance in the phylogenetic tree than the pairwise score between two closely related species would not necessarily imply lower conservation.

## 4.2 Conservation of genomic sequences

### 4.2.1 Functional elements in *Drosophila*

In a 2007 paper<sup>1</sup>, Stark et al. identified evolutionary signatures of different functional elements and predicted function using conserved signatures. One important finding is that across evolutionary time, genes tend to remain in a similar location. This is illustrated by Figure 4.2, which shows the result of a multiple alignment on orthologous segments of genomes from twelve *Drosophila* species. Each genome is represented by a horizontal blue line, where the top line represents the reference sequence. Grey lines connect orthologous functional elements, and it is clear that their positions are generally conserved across the different species.

**FAQ**

**Q:** Why is it significant that the position of orthologous elements is conserved?

**A:** The fact that positions are conserved is what allows us to make comparisons across species. Otherwise, we would not be able to align non-coding regions reliably.

*Drosophila* is a great species to study because, in fact, the separation of fruit flies is greater than that of mammals. This brings us to an interesting side-note, that of which species to select when looking at conservation signatures. You don't want to have very similar species (such as humans and chimpanzees, which share 98% of the genome), because it would be difficult to distinguish regions that are different from ones that are the same. When comparing species to humans, the right level of conservation to look at is the mammals. Specifically, most research done in this field is done using 29 eutherian mammals (placental mammals, no marsupials or monotremes) to study. Another things to take into account is branch-length differences between two species. Your ideal subjects of study would be a few closely related (short branch-length) species, to avoid problems of interpretation that arise with a long branch-length mutations, such as back-mutations.

### 4.2.2 Rates and patterns of selection

Now that we have established that there is structure to the evolution of genomic sequences, we can begin analyzing specific features of the conservation. For this section, let us consider genomic data at the level of individual nucleotides. Later on in this chapter we will see that we can also analyze amino acid sequences.

We may estimate the intensity of a constraint of selection  $\omega$  by making a probabilities model of the substitution rate inferred from genome alignment data. Using a Maximum Likelihood (ML) estimation of  $\omega$  can provide us with the rate of selection  $\omega$  as well as the log odds score that the rate is non-natural.

<sup>1</sup><http://www.nature.com/nature/journal/v450/n7167/abs/nature06340.html>

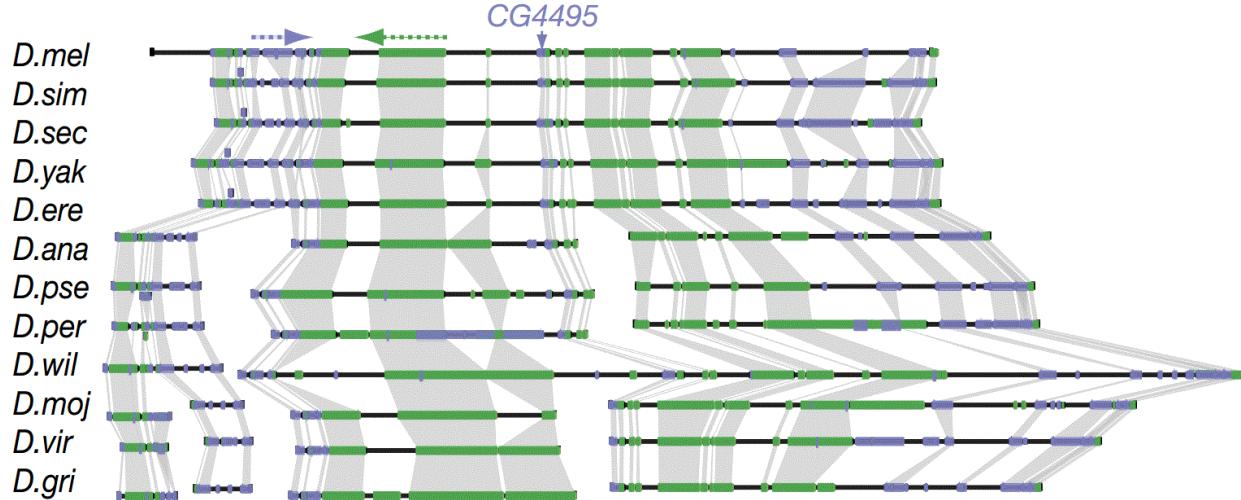


Figure 4.2: Comparative identification of functional elements in 12 *Drosophila* species. Grey lines indicate the alignment of orthologous regions. Color indicates direction of transcription.

One property that this measures that we may consider is the rate of nucleotide substitution in a genome. Figure 4.3 shows two nucleotide sequences from a collection of mammals. One of the sequences is subject to normal rates of change, while the other demonstrates a reduced rate. Hence we may hypothesize that the latter sequence is subject to a greater level of evolutionary constraint, and may represent a more biologically important section of the genome.

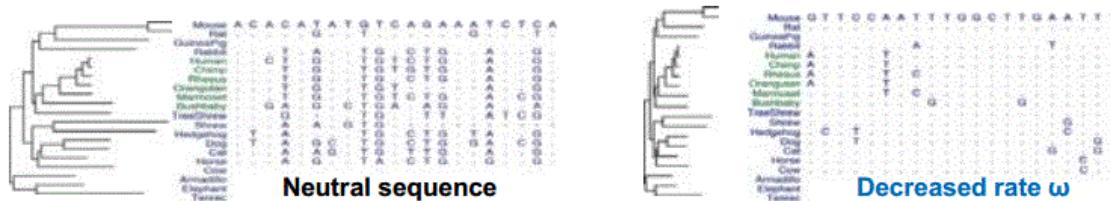


Figure 4.3: A comparison between two genomic regions with different selection rates  $\omega$ . The sequence on the left demonstrates normal rates of mutation, while the sequence on the right shows a high conservation level, as evidenced by the reduced number of mutations.

We can further detect unusual patterns of selection  $\pi$  by looking at a probabilistic model of a stationary distribution that is different from the background distribution. The ML estimation of  $\pi$  provides us with the Probability Weight Matrix (PWM) for each k-mer in the genome as well as the log odds score for substitutions that are unusual (e.g. one base changing to one and only one other base). As one may see from Figure 4.4, specific letters matter because some bases selectively change to one (or two other bases), and the specific base it changes to may suggest what the function of the sequence may be.

We can increase our detection power of constraint elements by looking at more species, as shown in Figure 4.5 where we see a dramatic increase in the power to detect small constrained elements.

### 4.3 Excess Constraint

In most regions of the genome where we see conservation across species, we expect there to be at least some amount of **synonymous substitution** - “silent” nucleotide substitutions that modify a codon in such a way that the amino acid it encodes is unchanged. In a 2011 paper<sup>2</sup>, Lindblad-Toh et al. studied evolutionary constraint in the human genome by doing comparative analysis of 29 mammalian species. They found that

<sup>2</sup><http://www.nature.com/nature/journal/v478/n7370/full/nature10530.html>

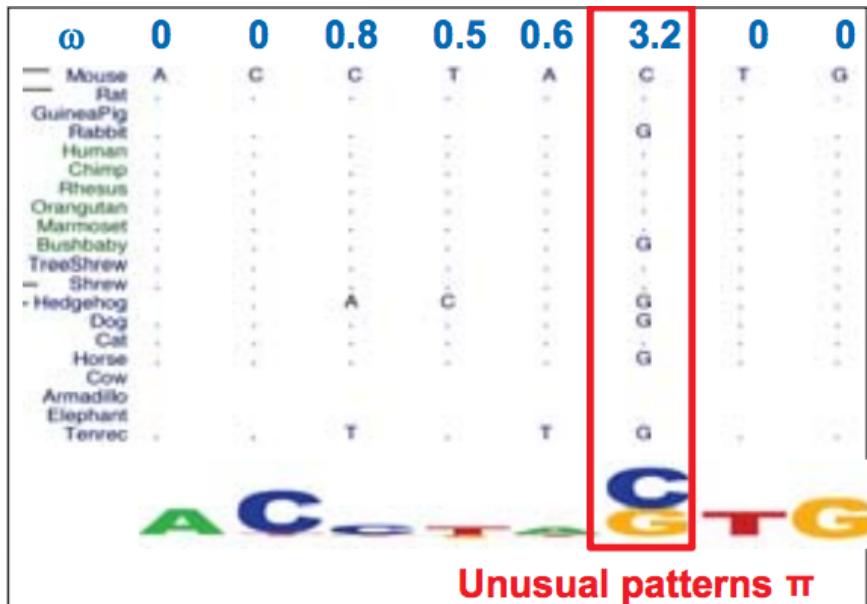


Figure 4.4: This sequence displays an unusual substitution rate of substituting C with G and vice versa.

among the 29 genomes, the average nucleotide site showed 4.5 substitutions per site (resulting in essentially a random distribution of the 4 bases over time).

Given such a high average substitution rate, we do not expect to see perfect conservation across all regions that are conserved. For example, ignoring all other effects, the probability of a 12-mer remaining fixed across all 29 species is less than  $10^{-25}$ . Thus, regions which are nearly perfectly conserved across multiple species must be under evolutionary pressure to remain the same - a principle we call **excess constraint**. One such region is shown in Figure 4.6.

#### 4.3.1 Causes of Excess Constraint

The question is what evolutionary pressures cause certain regions to be so perfectly conserved? The following were all mentioned in class as possibilities:

- Could it be that there is a special structure of DNA shielding this area from mutation?
- Is there some special error correcting machinery that sits at this spot?
- Can the cell use the methylation state of the two copies of DNA as an error correcting mechanism? This mechanism would rely on the fact that the new copy of DNA is unmethylated, and therefore the DNA replication machinery could check the new copy against the old methylated copy.
- Maybe the next generation can't survive if this region is mutated?

Another possible explanation is that selection is occurring to conserve specific codons. Some codons are more efficient than others: for example, higher abundant proteins that need rapid translation might select codons that give the most efficient translation rate, while other proteins might select for codons that give less efficient translation.

Still, these regions seem too perfectly conserved to be explained by codon usage alone. What else can explain excess constraint? There must be some degree of accuracy needed at the nucleotide level that keeps these sequences from diverging.

It could be that we are looking at the same region in two species that have only recently diverged or that there is a specific genetic mechanism protecting this area. However, it is more likely that so much conservation is a sign of protein coding regions that simultaneously encode other functional elements. For

	$\pi$ log-odds (12mers)	$\pi$ log-odds (50mers)	$\omega$ (12mers)	$\omega$ (50mers)
<b>29 mammals</b>	<b>7.1/1.5/4.6</b>	<b>6.8/1.8/4.1</b>	<b>5.7/ 1.1/3.8</b>	<b>5.7/1.8/3.0</b>
<b>(HMRD) Human Mouse Rat Dog</b>	<b>4.2/0.0/0.0</b>	<b>5.3/0.1/0.3</b>	<b>4.5/0.0/0.0</b>	<b>5.1/0.6/1.7</b>

**Estimated / kmers detectable at 5% FDR / base pairs detectable at 5% FDR**

Figure 4.5: By increasing the number of mammals studied, we see an increase in the constrained k-mers and base pairs that are detectable.

example, the HOXB5 gene shows obvious excess constraint, and there is evidence that the 5' end of the HOXB5 ORF encodes both protein and an RNA secondary structure.

Regions that encode more than one type of functional element are under overlapping selective pressures. There might be pressure in the protein coding space to keep the amino acid sequence corresponding to this region the same, combined with pressure from the RNA space to keep a nucleotide sequence that preserves the RNA's secondary structure. As a result of these two pressures to keep codons for the same amino acids and to produce the same RNA structure, the region is likely to show much less tolerance for any synonymous substitution patterns.

The process of estimating evolutionary constraint from genomic alignment data across multiple species follows the steps below:

- Count the number of edit operations (i.e. the number of substitutions and/or deletions/insertions)
- Estimate the number of mutations including back-mutations
- Incorporate information about the neighborhood elements of the conserved element by looking at "conservation windows"
- Estimate the probability of a constrained "hidden state" through using Hidden Markov Models
- Use phylogeny to estimate tree mutation rate (i.e. reject substitutions that should occur along the tree)
- Allow different portions of the tree to have different mutation rates

### 4.3.2 Modeling Excess Constraint

To better study region of excess constraint, we develop mathematical models to systematically measure the amount of synonymous and non-synonymous conservation of different regions. We will measure two rates: codon and nucleotide conservation.

To represent the null model, we can build rate matrices ( $4 \times 4$  in the nucleotide case and  $64 \times 64$  for the codon case) that give the rates of substitutions between either codons or nucleotides for a unit time. We

```

anc aa M S S Y F V N S F S G R Y P N G P D Y Q L L N T G S C S S L S G S Y R D P A A M H T G S Y G Y N Y N
ancestor ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Human ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Chimp ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Rhesus ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Tarsier ----- -----
Mouse_lemur ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Bushbaby ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Tree shrew ATG A----- -----
Mouse ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Cat ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Kangaroo ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Guinea Pig ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Squirrel ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Rabbit ----- -----
Pika ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Alpaca ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Dolphin ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Cow ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Horse ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Cat ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Dog ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Microbat ----- -----
Megabat ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Hedgehog ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Shrew ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Elephant ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Rock_hyrax ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Tenrec ATG AGC TCG TAC TTT GTAA AAC TCC TTC TCG GGG CGT TAT CCA AAA GGC CGG GAC TAT CAG TGG CTA AAT TAT GGC AGT GGC AGC AGG TCT CTC AGG GGC TGT TAC AGG GAT CCC GCT GGC ATG CAC ACC GGC TCT TAC GGC TAC AAT TAC AAT
Armadillo ----- -----
Sloth ----- -----

```

Figure 4.6: Many genomic regions, such as HOXB5, show more conservation than we would expect in normal conserved coding regions. Among the 29 species under study, all but 7 of them had the exact same nucleotide sequence. The green areas are areas that have undergone evolutionary mutations.

estimate the rates in the null model by looking at a ton of data and estimating the probabilities of each type of substitution. See Figure 4.19a in section 4.5.2 for an example of a null matrix for the codon case.

For the alternative model, we can define *scaling factors* in each region. These scaling factors measure how the rate of each type of substitution in our region of interest compares to that of the null model.

- $\lambda_s$ : the rate of synonymous substitutions
- $\lambda_n$ : the rate of nonsynonymous substitutions

For example, if  $\lambda_s = 0.5$ , then the rate of synonymous substitutions is half of what is expected from the null model in that region. We can then evaluate the statistical significance of the rate estimates we obtain, and find regions where the rate of substitution is much lower than expected.

Using a null model here helps us account for biases in alignment coverage of certain codons and also accounts for the possibility of codon degeneracy, in which case we would expect to see a much higher rate of substitutions. We will learn how to combine such models with phylogenetic methods when we talk about phylogenetic trees and evolution later on in the course.

Applying this model shows that the sequences in the first translated codons, cassette exons (exons that are present in one mRNA transcript but absent in an isoform of the transcript), and alternatively spliced regions have especially low rates of synonymous substitutions.

### 4.3.3 Excess Constraint in the Human Genome

In this section, we will examine the problem of determining the total proportion of the human genome under excess constraint. In particular, we will revisit the work of Lindblad-Toh et al. (2011), which compared 29 mammalian genomes. They measured conservation levels throughout the genome by applying the process described in the previous section to 50-mers. By considering only 50-mers which were part of ancestral repeats, it is possible to determine a background level of conservation. We can imagine that the intensities of conservation among the 50-mers are distributed according to a hidden probability distribution, as illustrated in Figure 4.8. In the figure, the background curve represents the distribution of constraint in the absence of special mechanisms for excess constraint, as determined by looking at ancestral repeats, while the signal (foreground) curve represents the actual distribution of the genome. The signal curve has more conservation overall due to the purifying effects of natural selection.

We may wish to investigate specific regions of the genome which are under excess constraint by setting a threshold level of conservation and examining regions which are more conserved. In the illustration, this corresponds to considering all 50-mers which fall to the right of one of the orange lines. We see that while this method does indeed give us regions under excess constraint, it also gives us false positives. This is

because even in the absence of purifying selection and other effects, certain regions will be heavily conserved, simply due to random chance. Setting the threshold higher, such as by using the dotted orange line as our threshold, reduces the proportion of false positives (FP) to true positives (TP), while also lowering the number of true positives detected, thus trading higher specificity for lower sensitivity.

However, not all hope is lost. It is possible to empirically measure both the background (BG) and foreground (FG) signal curves, as described above. Once that is done, the area of the region between them, which is shaded in gray in Figure 4.8, can be determined by integration. This area represents the proportion of the genome which is under excess constraint. Because the curves overlap, we cannot detect all conserved elements but we can estimate the total amount of excess constraint. This number of estimated constraint turns out to be about 5% of the human genome, depending on how large a window is used. Those regions are likely to all be functional, but since about 1.5% of the human genome is protein-coding, we can infer that the remaining 3.5% consists of functional, non-coding elements, most of which probably play regulatory roles.

We have seen that evolutionary constraint over the whole genome can be estimated by evaluating genomic constraint against a background distribution. Lindblad-Toh et al. (2011) compare genome conservation across 29 mammals against a background calculated from ancestral repeat elements to find regions with excess constraint (Figure 4.9 A and B). Annotation of evolutionarily constrained bases reveals that the majority of discovered regions are intergenic and intronic and demonstrates that going from four (HMRD) to 29 mammalian genomes increases the power of this analysis primarily in non-coding regions (Figure 4.9C). The most constrained regions in the genome are coding regions (Figure 4.9D).

As shown in Figure 4.9, the increase from HMRD to a 29 genome alignment vastly improves the power of this analysis. However, while the amount of intergenic elements detected increased significantly, detection is still limited by the fact that non-functional elements have much lower species coverage depth in multiple alignments than functional regions (Figure 4.10). For example, ancestral repeats (AR,  $\mu = 11.4$ ) have a much lower average coverage depth than exons ( $\mu = 20.9$ ). On one hand, this shows evidence of selection against insertions and deletions in functional elements, which are not examined in the analysis of base constraint. On the other, it also complicates the analysis of evolutionary constraint, as such work must then handle varying coverage across the genome.

#### 4.3.4 Examples of Excess Constraint

Examples of excess constraint have been found in the following cases:

- Most Hox genes show overlapping constraint regions. In particular, as mentioned above the first 50 amino acids of HOXB5 are almost completely conserved. In addition, HOXA2 shows overlapping regulatory modules. These two loci encode developmental enhancers, providing a mechanism for tissue specific expression.
- ADAR: the main regulator of mRNA editing, has a splice variant where a low synonymous substitution rate was found at a resolution of 9 codons.
- BRCA1: Hurst and Pal (2001) found a low rate of synonymous substitutions in certain regions of BRCA1, the main gene involved in breast cancer. They hypothesized that purifying selection is occurring in these regions. (This claim was refuted by Schmid and Yang (2008) who claim this phenomenon is the artifact of a sliding window analysis).
- THRA/NR1D1: these genes, also involved in breast cancer, are part of a dual coding region that codes for both genes and is highly conserved.
- SEPHS2: has a hairpin involved in selenocysteine recoding. Because this region must select codons to both conserve the protein's amino acid sequence and the nucleotides to keep the same RNA secondary structure, it shows excess constraint.

#### 4.3.5 Measuring constraint at individual nucleotides

By measuring evolutionary constraint at individual nucleotides instead of blocks of the sequence, we may find individual transcription factor binding sites, position-specific bias within motif instances, and reveal

motif consensus among most species. Specifically, we can detect SNPs that disrupt conserved regulatory motifs and determine the level of evolution by looking at every nucleotide in the gene.

## 4.4 Diversity of Evolutionary Signatures: An Overview of Selection Patterns

Independently of the substitution rate, we may also consider the *pattern* of substitutions in a particular nucleotide subsequence. Doing so gives us much more analytical power when searching for constrained elements, as shown in Figure 4.5; notice that the number of constrained elements substantially increases when we look for specific patterns (“ $\pi$  log-odds”) as opposed to searching by rate alone (“ $\omega$ ”). To understand how we may use patterns to aid us in the search for functional elements, such as protein-coding regions, consider a protein-coding region of the genome; then, due to tRNA wobble, a mutation in the third nucleotide of a codon is less likely to affect the final protein than a mutation in the other positions. Hence, we expect to see a pattern of increased substitutions every third base when looking at protein-coding sequences. This can indeed be verified experimentally, as shown in Figure 4.11: notice that in the protein-coding exon sequence on the left, substitutions occur predominantly every third nucleotide, while in the non-coding region on the right, the substitution pattern is much more random. In this case, therefore, we wish to search for mutation patterns where every third nucleotide may be mutated.

### Did You Know?

It is important to remember that we are focusing on nucleotide-level substitution patterns, even though selection happens on the function level. For example, it is certainly possible that an amino acid-altering mutation actually has no effect on the function of the final protein, in which case there will be no selective pressure counteracting that mutation, even though it is in a protein-coding region and not on the final nucleotide in a codon. We do not explore this distinction in detail, but it is important to keep in mind when considering evolutionary signatures, both for protein-coding regions and for other functional elements.

### FAQ

**Q:** In the protein-coding exon of Figure 4.11, we also see indels (insertions/deletions) in groups of three or six. Why is that?

**A:** Indels that are a multiple of three preserve the reading frame. If all the nucleotides of one codon are deleted, the rest of the codons are unaffected during translation. However, if we delete a number of nucleotides that is not a multiple of three (e.g. we only delete part of some codon), then the resulting shift changes all subsequent codons, since the reading frame has changed.

In Figure 4.12, we can see another feature of protein-coding sequences. Namely that their boundaries of conservation are very distinct, and lie near splice sites. Periodic, frame-conserving mutations begin to occur only after the splice site boundary.

So, we have seen the benefit of looking for not just high conservation rates, but also for specific patterns of conservation, and we have seen several such patterns for protein-coding sequences in particular. However, these **evolutionary signatures** of conservation exist for a wide range of different functional elements.

### 4.4.1 Selective Pressures On Different Functional Elements

Different functional elements bear different selective pressures due to their structures and other characteristics. Changes (insertions, deletions, mutations) that are extremely harmful to one may be innocuous to another. By figuring out what the signatures are for different elements, we can more accurately search for them.

An evolutionary signature, by definition, is a function-preserving pattern of change that is tolerated within a certain element. This is distinct from the *rate* of mutation, because the relative positions of change are taken into account. Evolutionary signatures arise because evolution and natural selection are acting on

different levels in certain functional elements. For instance, in a protein-coding gene, evolution is acting on the level of amino acids, and so natural selection will not filter out nucleotide changes that do not affect the amino acid sequence of the produced protein. By contrast, a structural RNA will be under pressure to preserve nucleotide pairs, but not necessarily individual nucleotides.

Importantly, the pattern of conservation has a distinct phylogenetic structure. More similar species – say mammals – group together with shared conserved domains that, say, fish lack, suggesting a mammalian-specific innovation, perhaps for regulatory elements not shared by fish. Meanwhile, some features are globally conserved, suggesting a universal significance, such as protein-coding. Initial approximate annotation of protein-coding regions in the human genome was possible using the simple heuristic that, if it was conserved from human to fish, it likely served as a protein-coding region.

An interesting idea for a final project would be to map divergences in the multiple alignment and call these events “births” of new coding elements. By focusing on a particular element (say microRNAs) one could identify periods of innovation and isolate portions of a phylogenetic tree enriched for certain classes of these elements.

The rest of the chapter will focus on quantifying the degree to which a sequence follows a given pattern. Kellis compared the process of evolution to exploring a fitness landscape, with the fitness score of a particular sequence constrained by the function it encodes. For example, as discussed, protein-coding genes are constrained by selection on the translated product, so synonymous substitutions in the third base pair of a codon are tolerated.

Below is a summary of the expected patterns followed by various functional elements:

- **Protein-coding genes exhibit particular frequencies of codon substitution as well as reading frame conservation.** This makes sense because the significance of the genes is the proteins they code for; therefore, changes that result in the same or similar amino acids can be easily tolerated, while a tiny change that drastically changes the resulting protein can be considered disastrous. In addition to the error correction of the mismatch repair system and DNA polymerase itself, the redundancy of the genetic code provides an additional level of intrinsic error correction/tolerance.
- Structural RNA is selected based on the secondary structure of the transcribed RNA, and thus requires compensatory changes. For example, some RNA has a secondary stem-loop structure (also known as a **hairpin**) such that one component of its sequence binds to another to form the stem, as shown in figure 4.13. Imagine taking the structured RNA shown in this figure and pulling apart the stem; we would be left with two **reverse-complementary** sequences separated by whatever nucleotides made up the loop. It is precisely this reverse-complementarity that constitutes an evolutionary signature for such RNAs.

Imagine that a nucleotide (A) and its partner (T) bind to each other in the stem, and then (A) mutates to a (C). This would ruin the secondary structure of the RNA. To correct this, either the (C) would mutate back to an (A), or the (T) would mutate to a (G). Then the (C)-(G) pair would maintain the secondary structure. This is called a compensatory mutation. Therefore, in RNA structures, the amount of change to the secondary structure (e.g. stem-loop) is more important than the amount of change in the primary structure (just the sequence). Understanding the effects of changes in RNA structure requires knowledge of the secondary structure. The likely secondary structure of an RNA can be determined by modeling the stability of many possible conformations and choosing the most likely conformation.

- MicroRNA is a molecule that is ejected from the nucleus into the cytoplasm. Their characteristic trait is that they also have the hairpin (stem-loop) structure illustrated in Figure 4.13, but a section of the stem is complementary to a portion of mRNA.

When microRNA binds its complementary sequence to the respective portion of mRNA, it degrades the mRNA. This means that it is a post-transcriptional regulator, since it's being used to limit the production of a protein (translation) after transcription. MicroRNA is conserved differently than structural RNA. Due to its binding to an mRNA target, the region of binding is much more conserved to maintain target specificity.

- Finally, **regulatory motifs are conserved in sequence** (to bind particular interacting protein partners) **but not necessarily in location**. Regulatory motifs can move around since they only need to recruit a factor to a particular region. Small changes (insertions and deletions) that preserve the consensus of the motif are tolerated, as are changes upstream and downstream that move the location of the motif.

When trying to understand the role of conservation in functional class prediction, an important question is how much of the observed conservation can be explained by known patterns. Even after accounting for “random” conservation, roughly 60% of non-random conservation in the fly genome was not accounted for – that is, we couldn’t identify it as a protein-coding gene, RNA, microRNA, or regulatory motif. The fact that they remain conserved, however, suggests a functional role. That so much conserved sequence remains poorly understood indicates that many exciting questions remain to be answered. One final project for 6.047 in the past was using clustering (unsupervised learning) to account for the other conservation. It developed into an M.Eng project, and some clusters were identified, but the function of these clusters was, and still is, unclear. It’s an open problem!

## 4.5 Protein-Coding Signatures

In slide 12, we see three examples of conservation: an intronic sequence with poor conservation, a coding region with high conservation, and a non-coding region with high conservation, meaning it is probably a functional element. As discussed, the important characteristic of protein-coding regions is that codons code for amino acids, which make up proteins. This gives rise to the following evolutionary signatures:

- Reading-frame conservation
- Codon-substitution patterns

The intuition behind each of these is relatively straightforward.

Firstly, reading frame conservation makes sense, since an insertion or deletion of one or two nucleotides will “shift” how all the following codons are read. However, if an insertion or deletion happens in a multiple of 3, the other codons will still be read in the same way, so this is a less significant change. For instance, consider the following nucleotides and their corresponding amino acid sequence:

Nucleotides:	C	G	C	U	C	G	U	A	C
	↓		↓		↓		↓		
Amino acids:	Arg		Ser			Tyr			

If the first nucleotide is deleted, the subsequent codons will code for different amino acids:

Nucleotides:	-	G	C	U	C	G	U	A	C
		↓		↓		↓			
Amino acids:		Ala			Arg				

The reading frame has “shifted”. Now, imagine that the number of deleted nucleotides was a multiple of three:

Nucleotides:	-	-	-	U	C	G	U	A	C
			↓		↓		↓		
Amino acids:				Ser		Tyr			

Despite the deletions, the reading frame was conserved, and subsequent amino acids remain unchanged. Because of this, indels that are a of a length which is a multiple of three are evolutionary signatures of protein-coding regions.

Secondly, it also makes sense that mutations in the last nucleotide of a codon are tolerated in these regions, since many of these do not actually change the final amino acid. This redundancy is shown explicitly in Figure 4.15: notice that the sectors of the circle representing a particular amino acid typically incorporate multiple nucleotides in the third ring. Returning again to the example above, imagine the second nucleotide in the second codon was mutated:

Nucleotides:	C	G	C	U	G	G	U	A	C
				↓		↓		↓	
Amino acids:	Arg			Trp			Tyr		

The resulting amino acid has changed, as dictated by the conversion chart. Now, let us imagine that instead the last nucleotide of the same codon was mutated:

Nucleotides:	C	G	C	U	C	A	U	A	C
				↓		↓		↓	
Amino acids:	Arg			Ser			Tyr		

The resulting amino acid stayed the same! This shows the redundancy in the genetic code, and the resulting evolutionary signature that arises from it, where the last nucleotide in every codon is allowed to mutate (in protein-coding regions).

These distinctive patterns allow us to “color” the genome and clearly see where the exons are, as shown in Figure 4.16.

When using these patterns in distinguishing evolutionary signatures, we have to make sure to consider the ideas below:

- Quantify the distinctiveness of all  $64^2$  possible codon substitutions by considering synonymous (frequent in protein-coding sequences) and nonsense (more frequent in non-coding than coding sequences) regions.
- Model the phylogenetic relationship among the species: multiple apparent substitutions may be explained by one evolutionary event.
- Tolerate uncertainty in the input such as unknown ancestral sequences and gaps in alignment (missing data).
- Report the certainty or uncertainty of the result: quantify the confidence that a given alignment is protein-coding using various units such as  $p$ -value, bits, decibans, etc.

#### 4.5.1 Reading-Frame Conservation (RFC)

Now that we know about the reading-frame conservation pattern in protein-coding genes, we can develop methods to determine if a given nucleotide sequence is in fact likely to be protein-coding based on it.

We can quantify this by scoring the pressure to stay in the same reading frame. We do this by having a target sequence (e.g. Scer, the genome of *S. cerevisiae*), and then aligning a selected sequence (e.g. Spar, *S. paradoxus*) to it and calculating what proportion of the time the selected sequence matches the target sequence’s reading frame.

Since we don’t know where the reading frame starts in the selected sequence, we align three times to try all possible offsets:

$$(\text{Spar}_{f1}, \text{Spar}_{f2}, \text{Spar}_{f3})$$

From these, we choose the alignment where the selected sequence is most often in sync with the target sequence. For example, we can begin numbering the nucleotides “1, 2, 3, etc.” until we reach a gap that we do not number. Or we can start numbering the nucleotides “2, 3, 1, etc.” where each triplet of “1,2,3” represents a codon. In this way, we encode each of the three possible reading frames. Finally, for the best alignment, we calculate the percentage of nucleotides that are out of frame. If this number is above a cutoff, the selected species “votes” that this region is a protein-coding region, and if it is low, this species “votes” that this is an intergenic region. The “votes” are tallied from all the species to sum to the RFC score.

This method is not robust to sequencing error. We can compensate for these errors by using a smaller scanning window and observing local reading frame conservation.

The method was shown to have 99.9% specificity and 99% sensitivity when applied to the yeast genome. When applied to 2000 hypothetical ORFs (open reading frames, or proposed genes)<sup>3</sup> in yeast, it rejected 500 of these putative protein coding genes as not being protein coding.

<sup>3</sup>Kellis M, Patterson N, Endrizzi M, Birren B, Lander E. S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. Science. 423: 241–254.

Similarly, 4000 hypothetical genes in the human genome were rejected by this method. This model created a specific hypothesis (that these DNA sequences were unlikely to code for proteins) that has subsequently been supported with experimental confirmation that the regions do not code for proteins *in vivo*.<sup>4</sup>

This represents an important step forward for genome annotation, because previously it was difficult to conclude that a DNA sequence was non-coding simply from lack of evidence. By narrowing the focus and creating a new null hypothesis (that the gene in question appears to be a non-coding gene) it became much easier to not only accept coding genes, but to reject non-coding genes with computational support. During the discussion of reading frame conservation in class, we identified an exciting idea for a final project which would be to look for the birth of new functional proteins resulting from frame shift mutations.

### 4.5.2 Codon-Substitution Frequencies (CSFs)

The second signature of protein-coding regions, the codon substitution frequencies, acts on multiple levels of conservation. To explore these frequencies, it is helpful to remember that codon evolution can be modeled by conditional probability distributions (CPDs) – the likelihood of a descendant having a codon  $b$  conditioned on an ancestor having codon  $a$ , separated by some time  $t$ .

The most conservative event is exact maintenance of the codon. A mutation that codes for the same amino acid may be conservative but not totally synonymous, because of species-specific codon usage biases. Even mutations that alter the identity of the amino acid might be conservative if they code for amino acids with similar biochemical properties.

We use a CPD in order to capture the net effect of all of these considerations. To calculate these CPDs, we need a “rate” matrix,  $Q$ , which measures the exchange rate for a unit time; that is, it indicates how often codon  $a$  in species 1 is substituted for codon  $b$  in species 2, for a unit branch length on the phylogenetic tree. Then, by using  $e^{Qt}$ , we can estimate the frequency of substitution at time  $t$ .

#### FAQ

**Q:** Why is the frequency of substitutions exponential in the rate matrix  $Q$ ?

**A:** The mutation process over time can be viewed as a continuous-time Markov chain, which gives rise to a matrix differential equation of the form  $\frac{d}{dt}P(t) = QP(t)$  for our rate matrix  $Q$ . This equation has solution  $P(t) = e^{Qt}$ .

#### FAQ

**Q:** How can a matrix be exponentiated?

**A:** For some matrix  $A$ , we can compute  $e^A$  using a Taylor expansion:  $e^A = \sum_{n=0}^{\infty} \frac{A^n}{n!}$ .

When the CPD is considered in conjunction with the topology of a network graph representing the evolutionary tree, it has approximately  $(2L - 2) \cdot 64^2$  parameters, where  $L$  is the number of leaves in the tree (species in the evolutionary phylogeny). This number of parameters is derived from the number of entries in  $Q$  and the number of independent branch lengths,  $t$ .

The CPD is defined in terms of  $e^{Qt}$  as follows:

$$\Pr(\text{child} = a | \text{parent} = b; t) = [e^{Qt}]_{a,b} \quad (4.1)$$

The intuition, is that as time increases, the probability of substitutions increase, while at the “initial” time ( $t = 0$ ),  $e^{Qt}$  is the identity matrix, since every codon is guaranteed to be itself. But, how do we estimate the rate matrix? We can use a maximum likelihood approach:

---

<sup>4</sup>Clamp M et al. 2007. Distinguishing protein-coding and noncoding genes in the human genome. PNAS. 104: 19428–19433.

- $Q$  can be “learned” from a set of known protein-coding sequences by using expectation-maximization (EM), although there are many alternative approaches as well. Similarly, the non-coding rate matrix is learned using a set of known non-coding regions, such as ancestral repeat regions.
- **E-step:** Given the parameters of the model, we can use Felsenstein’s algorithm[1] to compute the probability of any alignment, while taking into account phylogeny, given the substitution model:

$$\text{Likelihood}(\mathbf{Q}) = \Pr(\text{Training Data}; \mathbf{Q}, t). \quad (4.2)$$

- **M-step:** Then, given the alignments and phylogeny, we can choose the parameters (the rate matrix  $Q$  and branch lengths  $t$ ) that maximize the likelihood of those alignments in the. For example, to estimate  $Q$ , we can count the number of times one codon is substituted for another in the alignment. The argument space consists of thousands of possibilities for  $Q$  and  $t$ . This space is represented by  $\mathbf{Q}$ , and  $\hat{\mathbf{Q}}$  is the parameter that maximizes the likelihood:

$$\hat{\mathbf{Q}} = \underset{\mathbf{Q}}{\operatorname{argmax}} \text{Likelihood}(\mathbf{Q}) \quad (4.3)$$

Other maximization strategies include gradient ascent, simulated annealing, and spectral decomposition. Notice that through this approach, we optimize not only  $Q$  but also the branch lengths  $t$  simultaneously.

## FAQ

**Q:** How does the branch length contribute to determining the rate matrix?

**A:** The branch lengths specify how much “time” passed between any two nodes in the phylogenetic tree. The rate matrix describes the relative frequencies of codon substitutions per **unit branch length**, making it generalizable to any two nodes in the tree, if we know their branch length.

Once we have estimated the substitution models for coding ( $Q_C$ ) and non-coding ( $Q_N$ ) regions, we can compute the probability under each for a given set of alignments, and use the likelihood ratio  $\frac{\Pr(\text{Leaves}; Q_C, t)}{\Pr(\text{Leaves}; Q_N, t)}$  to ultimately classify the region we are examining.

As an example of how to use the rate matrix  $Q$ , let us consider a *nucleotide* substitution model with

$$Q = \begin{pmatrix} -4 & 2 & 1 & 1 \\ 2 & -4 & 1 & 1 \\ 1 & 1 & -4 & 2 \\ 1 & 1 & 2 & -4 \end{pmatrix},$$

where the order of nucleotides along the rows and columns is A,G,C,T. This matrix describes the rate of nucleotide substitution *per unit time*. So, for instance, we may ask: what is the probability of a G being substituted for a C after  $t_1 = 0.1$  time units? To determine this, we first exponentiate  $0.1Q$ :

$$e^{Qt_1} = \exp \left[ 0.1 \begin{pmatrix} -4 & 2 & 1 & 1 \\ 2 & -4 & 1 & 1 \\ 1 & 1 & -4 & 2 \\ 1 & 1 & 2 & -4 \end{pmatrix} \right] = \begin{pmatrix} 0.69 & 0.14 & 0.08 & 0.08 \\ 0.14 & 0.69 & 0.08 & 0.08 \\ 0.08 & 0.08 & 0.69 & 0.14 \\ 0.08 & 0.08 & 0.14 & 0.69 \end{pmatrix}.$$

Now, we can simply read the corresponding cell of the resulting matrix, and conclude that the desired probability is 8%. It is also interesting to examine what happens over very long time spans. For example, for  $t_2 = 10$  time units:

$$e^{Qt_2} = \exp \left[ 10 \begin{pmatrix} -4 & 2 & 1 & 1 \\ 2 & -4 & 1 & 1 \\ 1 & 1 & -4 & 2 \\ 1 & 1 & 2 & -4 \end{pmatrix} \right] = \begin{pmatrix} 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix}.$$

The substitution matrix is uniform! Over such long periods, all substitutions have time to accumulate even if they are very improbable, meaning the ancestral nucleotide has less of an effect on the child nucleotide. In this case, for  $t = 10$ , we see that the child nucleotide is in fact completely independent of the ancestral one. We have focused on nucleotide substitution here, but the codon substitution case is very much analogous.

Figures 4.19a and 4.19b show example rate matrices for intergenic and genic regions, respectively. A number of salient features present themselves in the codon substitution matrix (CSM) for genes. Note that the main diagonal element has been removed, because the frequency of a triplet being exchanged for itself will obviously be much higher than any other exchange. Nevertheless:

1. It is immediately obvious that there is a strong diagonal element in the protein coding regions.
2. We also note certain high-scoring off diagonal elements in the coding CSM: these are substitutions that are close in function rather than in sequence, such as 6-fold degenerate codons or very similar amino acids.
3. We also note dark vertical stripes, which indicate these substitutions are especially unlikely. These columns correspond to stop codons, since substitutions to this triplet would significantly alter protein function, and thus are strongly selected against.

On the other hand, in the matrix for intergenic regions, the exchange rates are more uniform. In these regions, what matters is the *mutational proximity*, i.e. the edit distance or number of changes from one sequence to another. Genetic regions are dictated by *selective proximity*, or the similarity in amino acid sequence of the protein resulting from the gene.

### **Did You Know?**

The evolutionary process under this model is time-reversible, so we can use the same approach to model ancestral sequences given the children.

Now that we have the two rate matrices for the two regions, we can calculate the probabilities that each matrix generated the genomes of the two species. This can be done by using Felsenstein's algorithm, and adding up the "score" for each pair of corresponding codons in the two species. Finally, we can calculate the likelihood ratio that the alignment came from a coding region to a non-coding region by dividing the two scores as indicated above – this represents our confidence in our classification of the region in question. If the ratio is greater than 1, we can guess that it is a coding region, and if it is less than 1, that it is a non-coding region. For example, in Figure 4.20, we are very confident about the respective classifications of each region, as indicated by the extreme magnitudes of the likelihood ratios.

It should be noted, however, that although the "coloring" of the sequences confirms our classifications, the likelihood ratios are calculated independently of the coloring in that they use our knowledge of synonymous or conservative substitutions. This further implies that this method automatically infers the genetic code from the pattern of substitutions that occurs, simply by looking at the high scoring substitutions. In species with a different genetic code, the patterns of codon exchange will be different; for example, in *Candida albumin*, the "CTG" codon codes for serine (polar) rather than leucine (hydrophobic), and this can be deduced from the CSMs. However, no knowledge of this is required by the method; instead, we can deduce this *a posteriori* from the CSM.

In summary, we are able to distinguish between non-coding and coding regions of the genome based on their evolutionary signatures, by creating two separate  $64 \times 64$  *rate matrices*: one measuring the rate of codon substitutions in coding regions, and the other in non-coding regions. The rate matrix gives the exchange rate of codons or nucleotides over a unit time.

We used the two matrices to calculate two probabilities for any given alignment: the likelihood that it came from a coding region and the likelihood that it came from a non-coding region. Taking the *likelihood ratio* of these two probabilities gives a measure of confidence that the alignment is protein-coding, as demonstrated in Figure 4.20. Using this method, we can pick out regions of the genome that evolve according to the protein coding signature.

We will see later how to combine this likelihood ratio approach with phylogenetic methods to find evolutionary patterns of protein coding regions.

However, this method only lets us find regions that are selected at the translational level. The key point is that here we are measuring for only protein coding selection. We will see how we can look for other conserved functional elements that exhibit their own unique signatures.

### 4.5.3 Classification of *Drosophila* Genome Sequences

We have seen that using these RFC and CSF metrics allow us to classify exons and introns with extremely high specificity and sensitivity. The classifiers that use these measures to classify sequences can be implemented using a HMM or semi-Markov conditional random field (SMCRF). CRFs allow the integration of diverse features that do not necessarily have a probabilistic nature, whereas HMMs require us to model everything as transition and emission probabilities. One might wonder why these more complex methods need to be implemented, when the simpler method of checking for conservation of the reading frame worked well. The reason is that in very short regions, insertions and deletions will be very infrequent, even by chance, so there will not be enough signal to make the distinction between protein-coding and non-protein-coding regions. In the figure below, we see a DNA sequence along the *x*-axis, with the rows representing an annotated gene, amount of conservation, amount of protein-coding evolutionary signature, and the result of Viterbi decoding using a SMCRF, respectively.

This is one example of how utilization of the protein-coding signature to classify regions has proven very successful. Identification of regions that had been thought to be genes but that did not have high protein-coding signatures allowed us to strongly reject 414 genes in the fly genome previously classified as CGid-only genes, which led FlyBase curators to delete 222 of them and flag another 73 as uncertain. In addition, there were also definite false negatives, as functional evidence existed for the genes under examination. Finally, in the data, we also see regions with both conservation, as well as a large protein-coding signature, but had not been previously marked as being parts of genes, as in Figure 4.21. Some of these have been experimentally tested and have been shown to be parts of new genes or extensions of existing genes. This underscores the utility of computational biology to leverage and direct experimental work.

### 4.5.4 Leaky Stop Codons

Stop codons (TAA, TAG, TGA in DNA and UAG, UAA, UGA in RNA) typically signal the end of a gene. They clearly reflect translation termination when found in mRNA and release the amino acid chain from the ribosome. However, in some unusual cases, translation is observed beyond the first stop codon. In instances of single read-through, there is a stop codon found within a region with a clear protein-coding signature followed by a second stop-codon a short distance away. An example of this in the human genome is given in Figure 4.22. This suggests that translation continues through the first stop codon. Instances of double read-through, where two stop codons lie within a protein coding region, have also been observed. In these instances of stop codon suppression, the stop codon is found to be highly conserved, suggesting that these skipped stop codons play an important biological role.

Translational read-through is conserved in both flies, which have 350 identified proteins exhibiting stop codon read-through, and humans, which have 4 identified instances of such proteins. They are observed mostly in neuronal proteins in adult brains and brain expressed proteins in *Drosophila*.

The kelch gene exhibits another example of stop codon suppression at work. The gene encodes two ORFs with a single UGA stop codon between them. Two proteins are translated from this sequence, one from the first ORF and one from the entire sequence. The ratio of the two proteins is regulated in a tissue-specific manner. In the case of the kelch gene, a mutation of the stop codon from UGA to UAA results in a loss of function, suggesting that tRNA suppression is the mechanism behind stop codon suppression.

An additional example of stop codon suppression is Caki, a protein active in the regulation of neurotransmitter release in *Drosophila*. Open reading frames (ORFs) are DNA sequences which contain a start and stop codon. In Caki, reading the gene in the first reading frame (Frame 0) results in significantly more ORFs than reading in Frame 1 or Frame 2 (a 440 ORF excess). Figure 4.23 lists twelve possible interpretations for the ORF excess. However, because the excess is observed only in Frame 0, only the first 4 interpretations are likely:

- Stop-codon read-through: the stop codon is suppressed when the ribosome pulls in tRNA that pairs incorrectly with the stop codon.
- Recent nonsense: Perhaps some recent nonsense mutation is causing stop codon read-through.
- A to I editing: Unlike we previously thought, RNA can still be edited after transcription. In some cases the A base is changed to an I, which can be read as a G. This could change a TGA stop codon to a TGG, which encodes an amino acid. However, this phenomenon is only found in a couple of cases.

- Selenocysteine, the “21st amino acid”: Sometimes when the TGA codon is read by a certain loop which leads to a specific fold of the RNA, it can be decoded as selenocysteine. However, this only happens in four fly proteins, so can’t explain all of stop codon suppression.

Among these four, three of them (recent nonsense, A to I editing, and selenocysteine) account for only 17 of the cases. Hence, it seems that read-through must be responsible for most if not all of the remaining cases. In addition, biased stop codon usage is observed hence ruling out other processes such as alternative splicing (where RNA exons following transcription are reconnected in multiple ways leading to multiple proteins) or independent ORFs.

Read-through regions can be determined in a single species based on their pattern of codon usage. The Z-curve as shown in Figure 4.24 measures codon usage patterns in a region of DNA. From the figure, one can observe that the read-through region matches the distribution before the regular stop codon. After the second stop however, the region matches regions found after regular stops – that is, non-coding regions.

Another suggestion, offered in class, was the possibility of ribosome slippage, where the ribosome skips some bases during translation. This might cause the ribosome to skip past a stop codon. This event occurs in bacterial and viral genomes, which have a greater pressure to keep their genomes small, and therefore can use this slipping technique to read a single transcript in each different reading frame. However, humans and flies are not under such extreme pressure to keep their genomes small. Additionally, we showed above that the excess we observe beyond the stop codon is frame specific to frame 0, suggesting that ribosome slipping is not responsible.

Cells are stochastic in general and most processes tolerate mistakes at low frequencies. The system isn’t perfect and stop codon leaks happen. However, the following evidence suggests that stop codon read-through is not random but instead subject to regulatory control:

- Perfect conservation of read-through stop codons is observed in 93% of cases, which is much higher than the 24% found in background.
- Increased conservation is observed upstream of the read-through stop codon.
- Stop codon bias is observed. TGAC is the most frequent sequence found at the stop codon in read-through and the least frequent found at normal terminated stop codons. It is known to be a “leaky” stop codon. TAAA is found almost universally only in non-read-through instances.
- Unusually high numbers of GCA repeats observed through read-through stop codons.
- Increased RNA secondary structure is observed following transcription suggesting evolutionarily conserved hairpins.

#### 4.5.5 Overlapping Selection

Another interesting case that we see in practice is overlapping selection. Consider the HOXA2 gene shown in Figure 4.25; notice that there are a couple points where the synonymous substitution rate unexpectedly drops to zero. In these cases, we instead see exact conservation, indicating that there is some regulatory element (motif) sitting at these positions.

Beyond this, we sometimes see cases where multiple genes share “real estate” on the genome, leading to much lower synonymous substitution rates throughout. In general, overlapping selection leads to interesting, sometimes subtle deviations from the expected evolutionary signatures.

### 4.6 microRNA (miRNA) Gene Signatures

One example of functional genomic regions subject to high levels of conservation are sequences encoding microRNAs (miRNAs). miRNAs are RNA molecules that bind to complementary sequences in the 3' untranslated region of targeted mRNA molecules, causing gene silencing.

How do we find evolutionary signatures for miRNA genes and their targets, and can we use these to gain new insights on their biological functions? We will see that this is a challenging task, as miRNAs leave a highly conserved but very subtle evolutionary signal.

### 4.6.1 Computational Challenge

Predicting the location of miRNA genes and their targets is a computationally challenging problem. We can look for “hairpin” regions, where we find nucleotide sequences that are complementary to each other and predict a hairpin structure. But out of 760,355 miRNA-like hairpins found in the cell, only 60–100 were true miRNAs. So to make any test that will give us regions statistically likely to be miRNAs, we need a test with 99.99% specificity.

Figure 4.27 is an example of the conservation pattern for miRNA genes. You can see the two hairpin structures conserved in the red and blue regions, with a region of low conservation in the middle. This pattern is characteristic of miRNAs.

By analyzing evolutionary and structural features specific to miRNA, we can use combinations of these features to pick out regions of miRNAs with >4,500-fold enrichment compared to random hairpins. The following are examples of features that help pick out miRNAs:

- miRNAs bind to highly conserved target motifs in the 3' UTR
- miRNAs can be found in introns of known genes
- miRNAs have a preference for the positive strand of DNA and for transcription factors
- miRNAs are typically not found in exonic and repetitive elements of the genome (counter-example in Figure 4.31).
- Novel miRNAs may cluster with known miRNAs, especially if they are in the same family or have a common origin

These features of miRNA-coding regions can be grouped into structural families, enabling classifiers to be built based on known RNAs in each family. Energy considerations for RNA structure can be used to support this classification into families. Within each family, orthologous conservation(genes in different species for same function with common ancestral gene) and paralogous conservation (duplicated genes within same species that evolved to serve different functions) occurs.

Evolutionary	Structural
Correlation with conservation profile	Hairpin stability (MFE z-score)
MFE of the consensus fold	Number of asymmetric loops
Structure conservation index	Number of symmetric loops

We can combine several features into one test by using a decision tree, as illustrated in Figure 4.30. At each node of the tree, a test is applied which determines which branch will be followed next. The tree is traversed starting from the root until a terminal node is reached, at which point the tree will output a classification. A decision tree can be trained using a body of classified genome subsequences, after which it can be used to predict whether new subsequences are miRNAs or not. In addition, many decision trees can be combined into a “random forest,” where several decision trees are trained. When a new nucleotide sequence needs to be classified, each tree votes on whether or not it is an miRNA, and then the votes are aggregated to determine the final classification.

Applying this technique to the fly genome showed 101 hairpins above the 0.95 cutoff, rediscovering 60 of 74 of known miRNAs, predicting 24 novel miRNAs that were experimentally validated, and finding an additional 17 candidates that showed evidence of diverse function.

### 4.6.2 Unusual miRNA Genes

The following four “surprises” were found when looking at specific miRNA genes:

Surprise 1 Both strands might be expressed and functional. For instance, in the miR-iab-4 gene, expression of the sense and antisense strands are seen in distinct embryonic domains. Both strands score > 0.95 for miRNA prediction.

Surprise 2 Some miRNAs might have multiple 5' ends for a single miRNA arm, giving evidence for an imprecise start site. This could give rise to multiple mature products, each potentially with its own functional targets.

Surprise 3 High scoring miRNA\* regions (the star arm is complementary to the actual miRNA sequence) are very highly expressed, giving rise to regions of the genome that are both highly expressed and contain functional elements.

Surprise 4 Both miR-10 and miR-10\* have been shown to be very important Hox regulators, leading to the prediction that miRNAs could be “master Hox regulators”. Pages 10 and 11 of the first set of lecture 5 slides show the importance of miRNAs that form a network of regulation for different Hox genes.

#### 4.6.3 Example: Re-examining ‘dubious’ protein-coding genes

Two genes, CG31044 and CG33311 were independently rejected because their conservation patterns did not match those characteristic of a protein evolutionary signatures (see Section 4.5). They were identified as precursor miRNA based on genomic properties and high expression levels (Lin et al.). This is a rare example of miRNA being found in previously exonic sequences and illustrates the challenge of identifying miRNA evolutionary signatures.

### 4.7 Regulatory Motifs

Another class of functional element that is highly conserved across many genomes contains *regulatory motifs*. A regulatory motif is a highly conserved sequence of nucleotides that occurs many times throughout the genome and serves some regulatory function. For instance, these motifs might characterize enhancers, promoters, or other genomic elements.

#### 4.7.1 Computationally Detecting Regulatory Motifs

Computational methods have been developed to measure conservation of regulatory motifs across the genome, and to find new unannotated motifs *de novo*. Known motifs are often found in regions with high conservation, so we can increase our testing power by testing for conservation, and then finding signatures for regulatory motifs.

Evaluating the pattern of conservation for known motifs versus the “null model” of regions without motifs gives the following signature:

Conservation within:	Gal4 (known motif region)	Controls
All intergenic regions	13%	2%
Intergenic: coding	13%: 3%	2%: 7%
Upstream: downstream	12: 0	1:1

So as we can see, regions with regulatory motifs show a much higher degree of conservation in intergenic regions and upstream of the gene of interest.

To discover novel motifs, we can use the following pipeline:

- Pick a motif “seed” consisting of two groups of three non-degenerate characters with a variable size gap in the middle.
- Use a conservation ratio to rank the seed motifs
- Expand the seed motifs to fill in the bases around the seeds using a hill climbing algorithm.
- Cluster to remove redundancy.

Discovering motifs and performing clustering has led to the discovery of many motif classes, such as tissue specific motifs, function specific motifs, and modules of cooperating motifs.

### 4.7.2 Individual Instances of Regulatory Motifs

To look for expected motif regions, we can first calculate a *branch-length score* for a region suspected to be a regulatory motif, and then use this score to give us a confidence level of how likely something is to be a real motif.

The branch length score (BLS) sums evidence for a given motif over branches of a phylogenetic tree. Given the pattern of presence or absence of a motif in each species in the tree, this score evaluates the total branch length of the sub-tree connecting the species that contain the motif. If all species have the motif, the BLS is 100%. Note more distantly related species are given higher scores, since they span a longer evolutionary distance. If a predicted motif has spanned such a long evolutionary time frame, it is likely it is a functional element rather than just a region conserved by random chance.

To create a null model, we choose control motifs. The null model motifs should be chosen to have the same composition as the original motif, to not be too similar to each other, and to be dissimilar from known motifs. We can get a confidence score by comparing the fraction of motif instances to control motifs at a given BLS score.

## 4.8 Further Reading

1. For more on constraint calculations and identification, refer to Lindblad-Toh's et. al.'s "A high-resolution map of human evolutionary constraint using 29 mammals".
2. For more on translational read-through and evolutionary signature, refer to Lin et. al.'s "Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes".

## 4.9 Tools and Techniques

1. For sequence alignment of proteins, see <http://mafft.cbrc.jp/alignment/software/>.
2. For prediction of genes through frameshifts in prokaryotes, see GeneTack.

## 4.10 Bibliography

### Bibliography

- [1] Joseph Felsenstein. Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981. 10.1007/BF01734359.

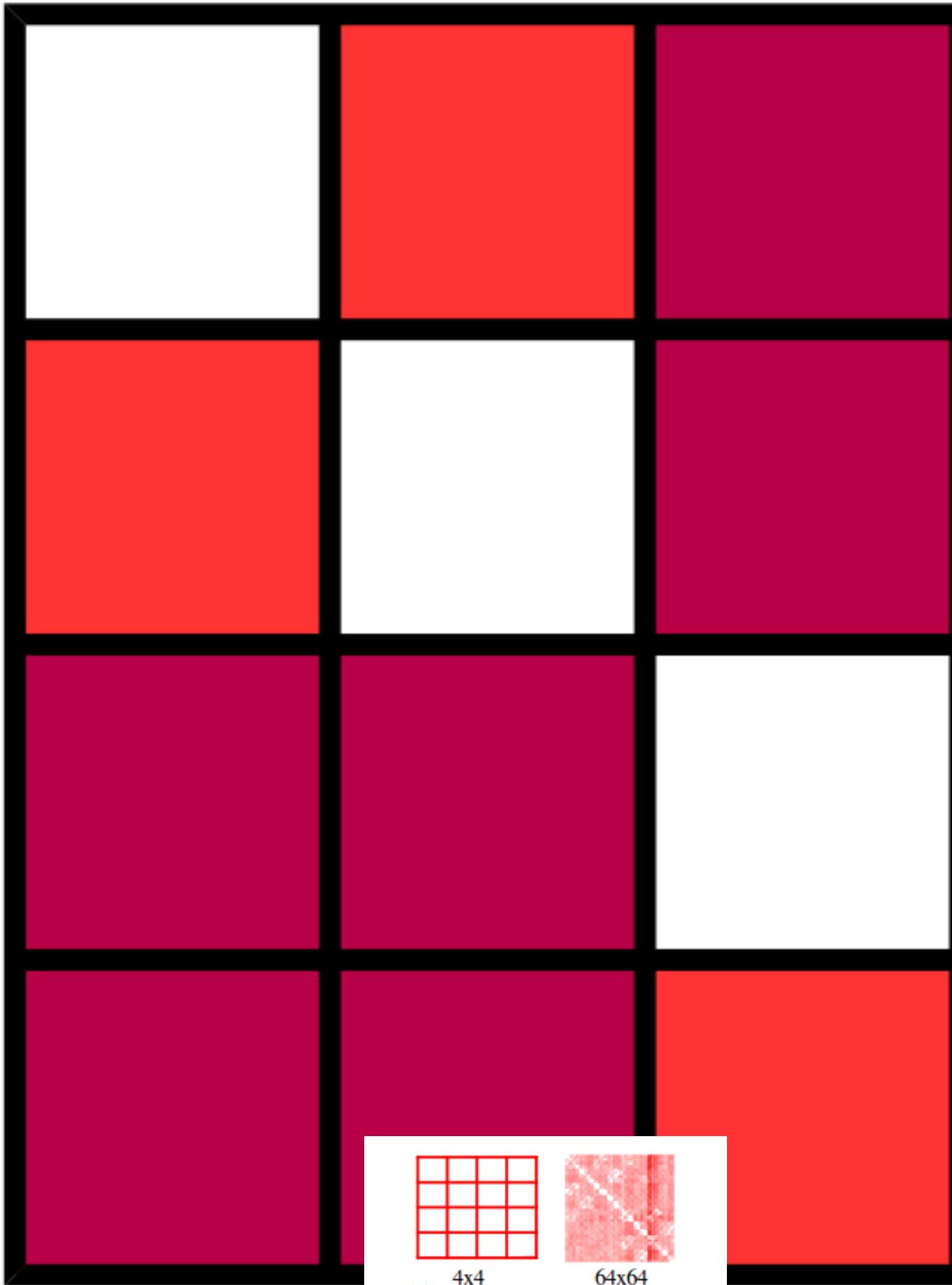


Figure 4.7: We can model mutations using rate matrices<sup>87</sup>, as shown here for nucleotide substitutions on the left and codon substitutions on the right. In each matrix, the cell in the  $m$ th row and  $n$ th column represents the likelihood that the  $m$ th symbol will mutate into the  $n$ th symbol. The darker the color, the less likely the mutation.

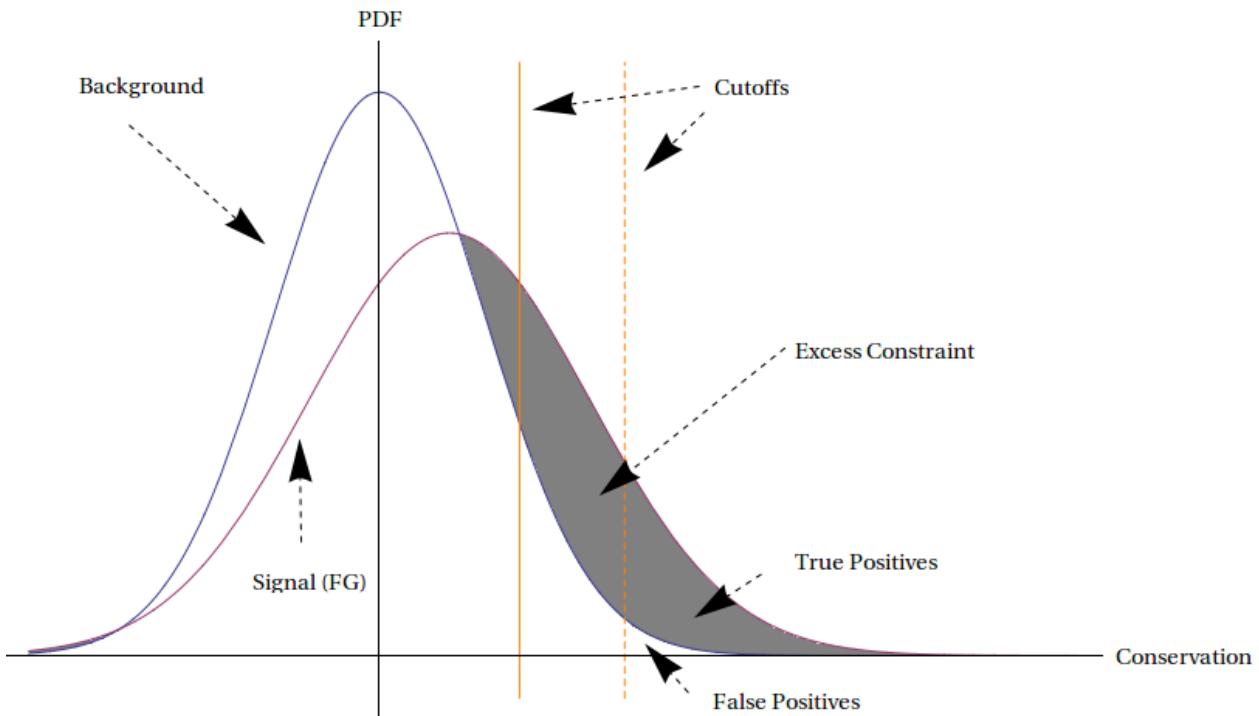


Figure 4.8: Measuring genome-wide excess constraint. See accompanying text for explanation.

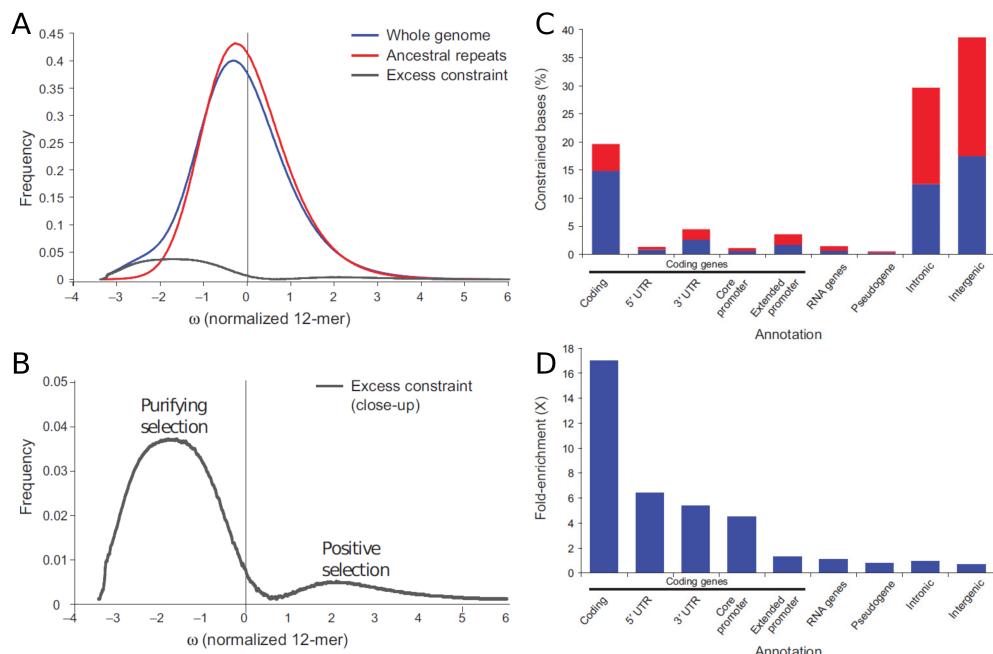


Figure 4.9: Detecting functional elements from their evolutionary signature. **A** Distribution of constraint for the whole genome against ancestral repeats (background). **B** Difference between whole genome and background constraint. **C** Discovery of functional elements from excess constraint. Novel elements are shown in red. **D** Enrichment of elements for regions of excess constraint.

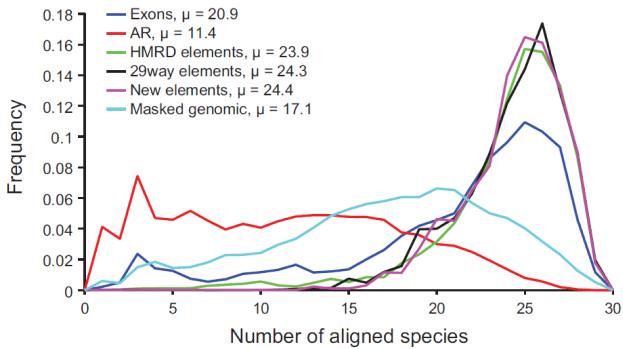


Figure 4.10: Coverage depth across different sets of elements. Functional elements experience greater conservation.

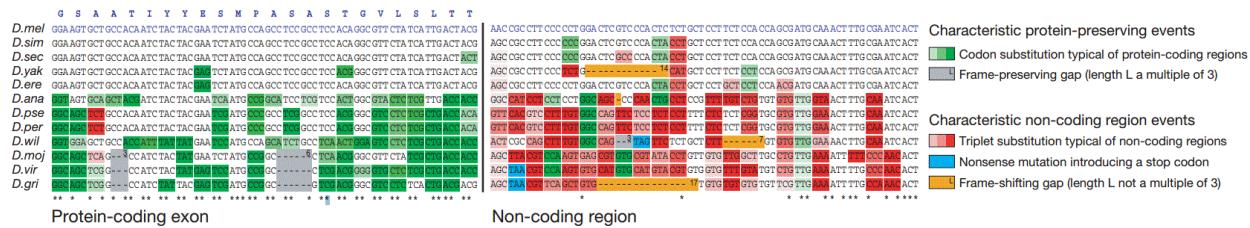


Figure 4.11: Different mutation patterns in protein-coding and non-protein-coding regions. Asterisks indicate that the nucleotide was conserved across all species. Note that within the protein-coding exon, nucleotides 1 and 2 of each codon tend to be conserved, while codon 3 is allowed to vary more, which is consistent with the phenomenon of wobble.

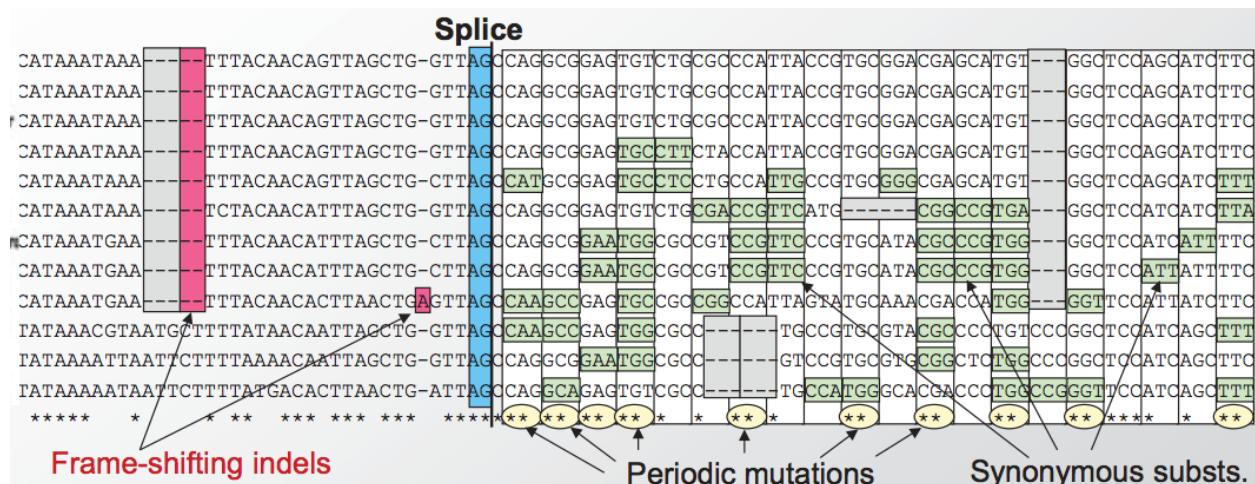


Figure 4.12: In addition to reading frame conservation and substitutions every third nucleotide, we also see sharp conservation boundaries that pinpoint splice sites.

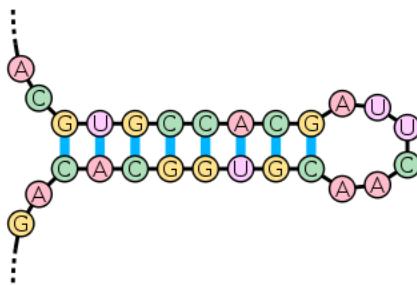


Figure 4.13: RNA with secondary stem-loop structure

				Second Letter	
				T      C      A      G	T C A G
				TTC      TCC      TAC      TGT	T C A G
T	TTC      TTC      TTA      TTG	TCC      TCA      TAA      TAG	TAC      TGA      Stop      Stop	TGT      TGC      TGA      TGG	Phe      Ser      Tyr      Trp
C	CTT      CTC      CTA      CTG	CCT      CCC      CCA      CCG	CAC      CAA      CAG	CGT      CGC      CGA      CGG	Leu      Pro      His      Arg
A	ATT      ATC      ATA      ATG	ACC      ACA      ACG	AAT      AAC      AAA      AAG	AGT      AGC      AGA      AGG	Ile      Thr      Asn      Lys
G	GTT      GTC      GTA      GTG	GCT      GCC      GCA      GCG	GAC      GAA      GAG	GGT      GGC      GGA      GGG	Val      Ala      Asp      Glu      Gly

Figure 4.14: Codon-to-amino acid translation chart

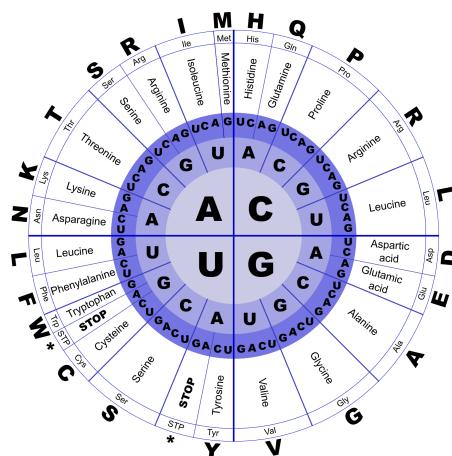
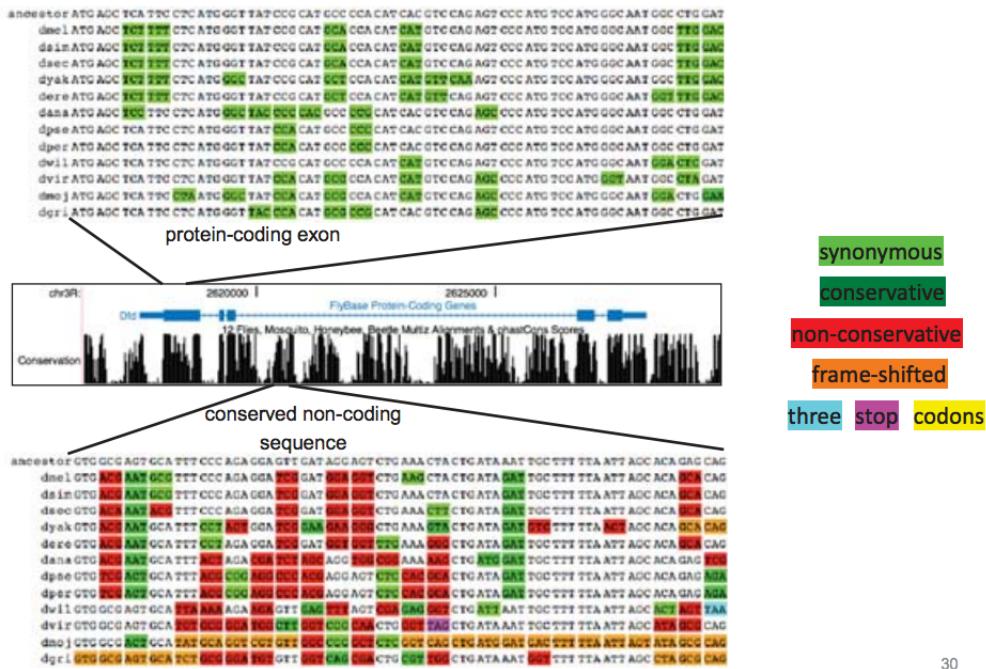


Figure 4.15: Codon-to-amino acid translation chart, showing tolerance of mutations in last nucleotide.



30

Figure 4.16: By coloring the types of insertions/deletions/substitutions that occur on a sequence, we can see patterns or evolutionary signatures that distinguish a protein-coding conserved region from a non-protein-coding conserved region.



Figure 4.17: Two alignments showing conservation pattern differences between gene and intergenic sequences. Red boxes represent gaps that shift the coding frame, and gray boxes are non-frame-shifting gaps (in multiples of three). Green regions are conserved, and yellow ones are mutated. Note the pattern of “match, match, mismatch” in the protein-coding sequence that indicates synonymous mutations.



Figure 4.18: Red boxes represent frame-shifting gaps, and gaps in multiples of three are uncolored. Conserved and mutated regions are green and yellow, respectively.

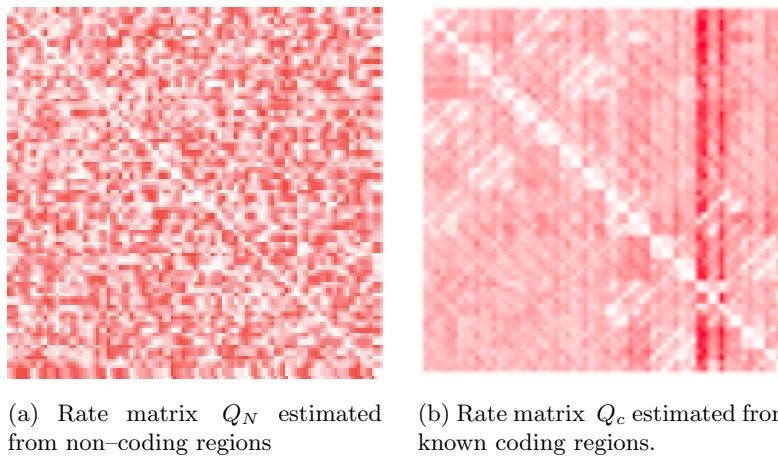


Figure 4.19: Rate matrices for the null and alternate models. A lighter color means substitution is more likely.

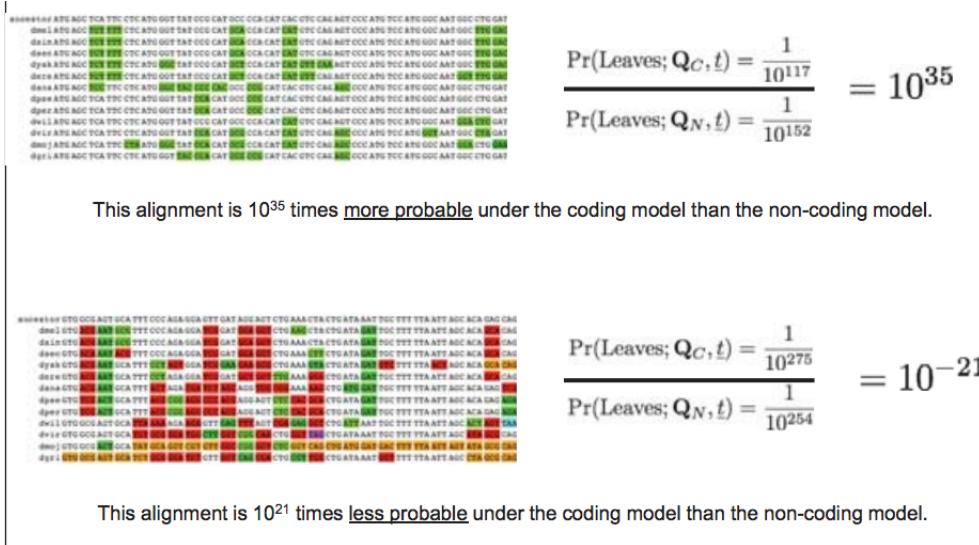


Figure 4.20: As we can see in the figure that the likelihood ratio is positive for sequences that are likely to be protein coding and negative for sequences that are not likely to be protein coding.

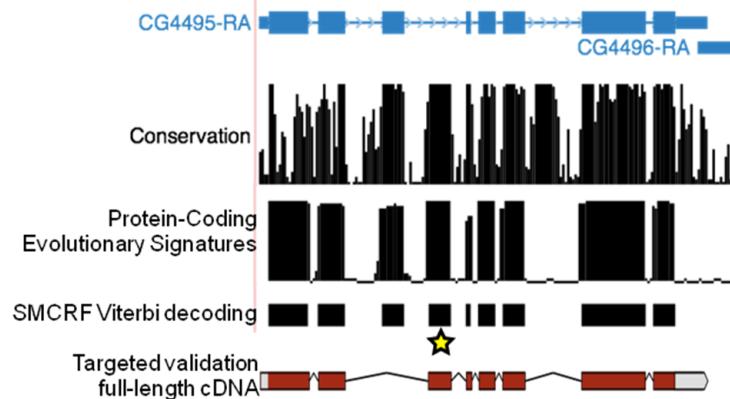


Figure 4.21: Evolutionary signatures can predict new genes and exons. The star denotes a new exon, which was predicted using the three comparative genomics tests, and later verified using cDNA sequencing.

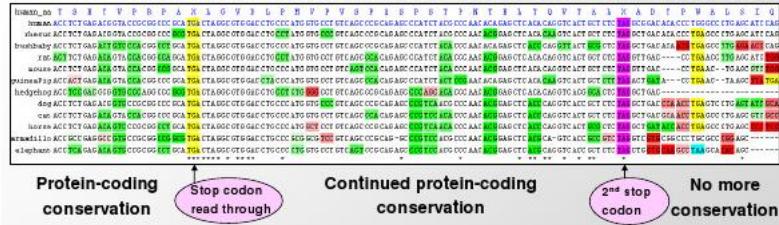


Figure 4.22: OPRL1 neurotransmitter: one of the four novel translational read-through candidates in the human genome. Note that the region after the first stop codon exhibits an evolutionary signature similar to that of the coding region before the stop codon, indicating that the stop codon is “suppressed”.

	Frame 0	Frame 1	Frame 2
Interpretation			
Readthrough	✓		
Recent nonsense	✓		
A → I editing	✓		
Selenocysteine	✓		
Alternative splicing	✓	✓	✓
Dicistronic	✓	✓	✓
Cryptic promotor	✓	✓	✓
Hopping	✓	✓	✓
Antisense	✓	✓	✓
Chance	✓	✓	✓
Frame shift		✓	✓
CDS overlaps stop		✓	✓
# with PCSF > 0	662	219	225

Figure 4.23: Various interpretations of stop codon suppression. See text for explanation.

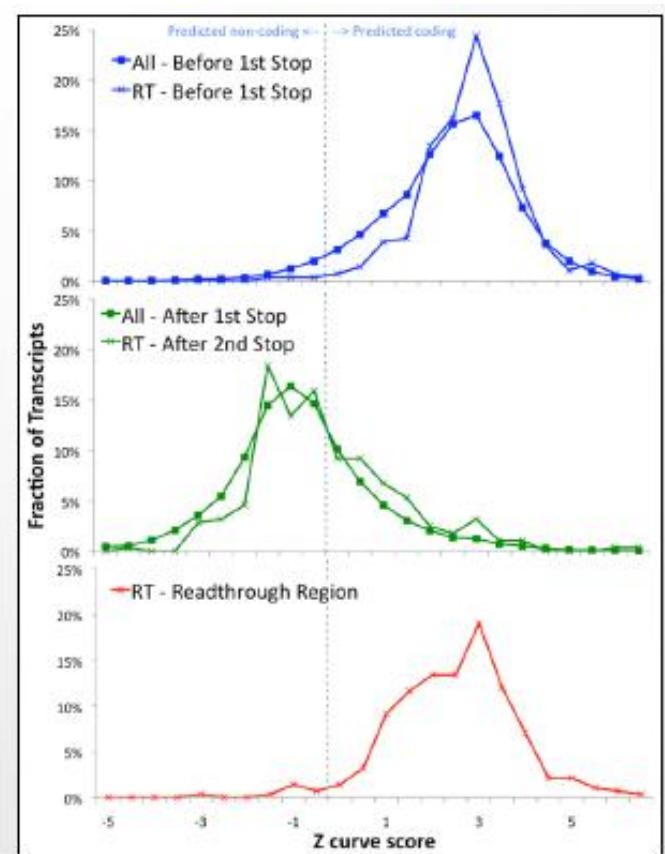


Figure 4.24: Z-curve for *Caki*. Note that the codon usage in the read through region is similar to that in the region before the first stop codon.



Figure 4.25: Example of overlapping signatures, where a regulatory motif is found within a gene.

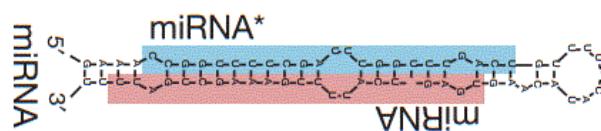


Figure 4.26: The hairpin structure of a microRNA. Note that miRNA\* denotes the strand on the opposite side of the hairpin, which has the same sequence as the mRNA molecules that are suppressed by the miRNA.

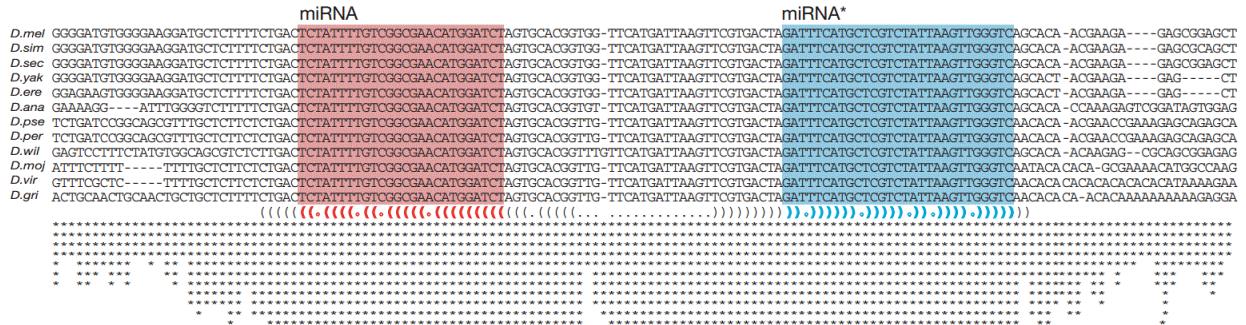


Figure 4.27: Characteristic conservation pattern of miRNAs. The number of asterisks below a nucleotide indicates the number of species where it is conserved. The blue and red highly conserved regions represent the complementary strands of the miRNA, as in figure 4.26.

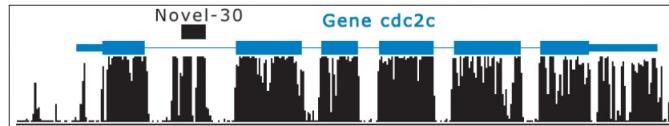


Figure 4.28: Novel miRNA in intron

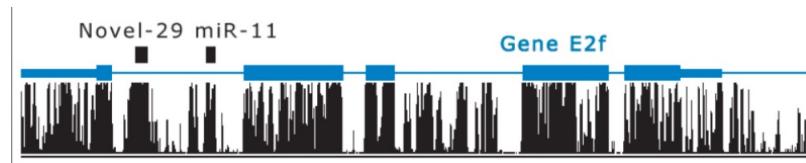


Figure 4.29: Novel and Known miRNA clustered.

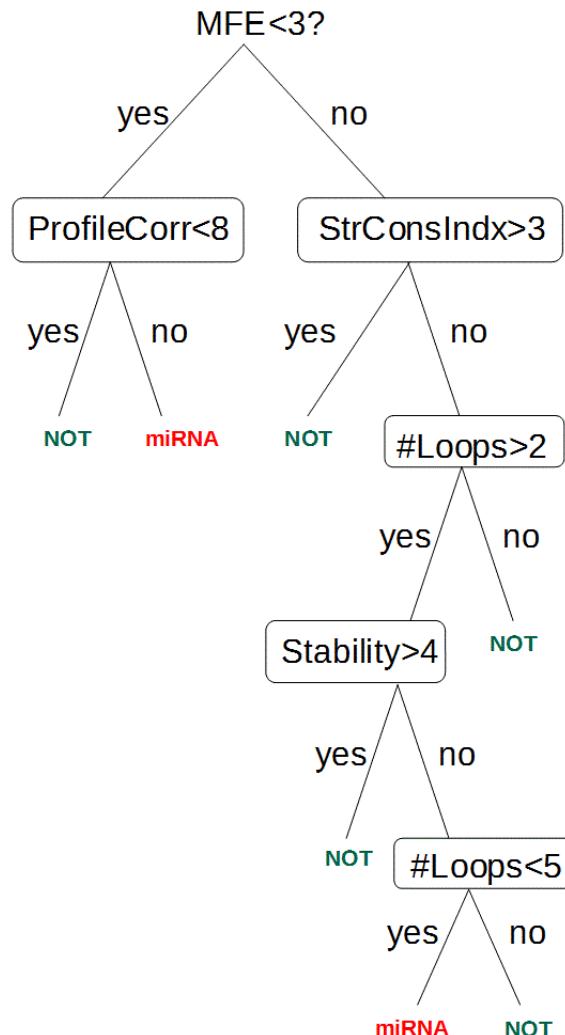


Figure 4.30: A possible decision tree for miRNA detection. The features used in this tree are minimum free energy, conservation profile correlation, structure conservation index, number of loops, and stability.

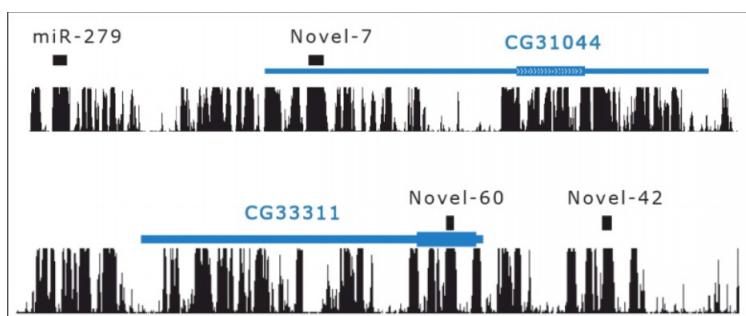


Figure 4.31: Existing annotations and transcription levels in 'dubious' protein-coding regions.

Figure 4.32: TAATTA is a hexamer that appears as a conserved element throughout the genome in many different functional elements, including here. It is an example of a regulatory motif.



---

CHAPTER  
**FIVE**

---

## GENOME ASSEMBLY AND ALIGNMENT

Melissa Gymrek, Liz Tsai, Rebecca Taft (2012), Keshav Dhandhania (2012), Joe Vitti (2013), Matt Fox (2014), Dina Levy-Lambert (2016), Sitara Persad (2016), Samuel Hendel (2017)

### Figures

---

5.1	We can use evolutionary signatures to find genomic functional elements, and in turn can study mechanisms of evolution by looking at patterns of genomic variation and change. . . . .	101
5.2	Here is a quick look at a few platforms that can be used to read genomes. . . . .	102
5.3	Shotgun sequencing involves randomly shearing a genome into small fragments so they can be sequenced, and then computationally reassembling them into a continuous sequence. . . . .	103
5.4	We can use Illumina Sequencing, a type of Next Generation Sequencing, to sequence fragments of DNA. . . . .	103
5.5	Constructing a sequence from read overlap . . . . .	104
5.6	We can visualize the process of merging fragments into contigs by letting the nodes in a graph represent reads and edges represent overlaps. By removing the transitively inferable edges (the pink edges in this image), we are left with chains of reads ordered to form contigs. . . . .	105
5.7	Overcollapsed contigs are caused by repetitive regions of the genome which cannot be distinguished from one another during sequencing. Branching patterns of alignment that arise during the process of merging fragments into contigs are a strong indication that one of the regions may be overcollapsed. . . . .	106
5.8	In this graph connecting contigs, repeated region X has indegree and outdegree equal to 2. The target seqence shown at the top can be inferred from the links in the graph. . . . .	106
5.9	Mate pairs help us determine the relative order of contigs in order to link them into into supercontigs. . . . .	107
5.10	We derive the multiple alignment consensus sequence by weighted voting at each base. . . . .	107
5.11	Constructing a string graph. . . . .	108
5.12	Constructing a string graph . . . . .	108
5.13	Example of string graph undergoing removal of transitive edges. . . . .	109
5.14	Example of string graph undergoing chain collapsing. . . . .	109
5.15	<i>Left:</i> Flow resolution concept. <i>Right:</i> Flow resolution example. . . . .	110
5.16	The Needleman-Wunsch algorithm for alignments of 2 and 3 genomes. . . . .	112
5.17	We can save time when performing a global alignment by first finding all the local alignments and then chaining them together along the diagonal with restricted dynamic programming. . . . .	112
5.18	Glocal alignment allows for the possibility of duplications, inversion, and translocations. . . . .	113
5.19	The steps to run the SLAGAN algorithm are A. Find all the local alignments, B. Build a rough homology map, and C. globally align the consistent parts using the regular LAGAN algorithm . . . . .	114
5.20	Using the concepts of glocal alignment, we can discover inversions, translocations, and other homologous relations between different species such as human and mouse. . . . .	114

5.21 Graph of <i>S. cerevisiae</i> and <i>S. bayanus</i> gene correspondence. . . . .	115
5.22 Illustration of gene correspondence for <i>S.cerevisiae</i> Chromosome VI (250-300bp). . . . .	116
5.23 Dynamic view of a changing gene. . . . .	117
5.24 Six types of genome rearrangement . . . . .	118
5.25 Moving further back in evolutionary time for Saccharomyces. . . . .	119
5.26 <i>K. waltii</i> Scaffolds . . . . .	119
5.27 Gene Correspondence for <i>S.cerevisiae</i> chromosomes and <i>K.waltii</i> scaffolds. . . . .	120
5.28 Gene interleaving shown by sister regions in <i>K.waltii</i> and <i>S.cerevisiae</i> . . . . .	120
5.29 Whole-genome duplication resolved . . . . .	121
5.30 S-LAGAN results. . . . .	122
5.31 S-LAGAN results for IGF locus. . . . .	122
5.32 S-LAGAN results for IGF locus. . . . .	122

## 5.1 Introduction

In the previous chapter, we saw the importance of comparative genomics analysis for discovering functional elements. In “part IV” of this book, we will see how we can use comparative genomics for studying gene evolution across species and individuals. In both cases however, we assumed that we had access to complete and aligned genomes across multiple species.

In this chapter, we will study the challenges of genome assembly and whole-genome alignment that are the foundations of whole-genome comparative genomics methodologies. First, we will study the core algorithmic principles underlying many of the most popular genome assembly methods available today. Second, we will study the problem of whole-genome alignment, which requires understanding mechanisms of genome rearrangement (e.g. segmental duplication and other translocations). The two problems of genome assembly and whole-genome alignment are similar in nature. We will close by discussing some of the parallels between them.

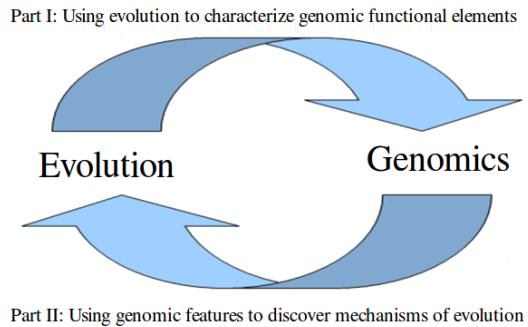


Figure 5.1: We can use evolutionary signatures to find genomic functional elements, and in turn can study mechanisms of evolution by looking at patterns of genomic variation and change.

## 5.2 Genome Assembly I: Overlap-Layout-Consensus Approach

Many areas of research in computational biology rely on the availability of complete whole-genome sequence data. Yet the process to sequence a whole genome is itself non-trivial and an area of active research. The problem lies in the fact that current genome-sequencing technologies cannot continuously read from one end of a long genome sequence to the other; they can only accurately sequence small sections of base pairs (ranging from 100 to a few thousand, depending on the method), called *reads*. Therefore, in order to construct a sequence of millions or billions of base pairs (such as the human genome), computational biologists must find ways to combine smaller reads into larger, continuous DNA sequences. The fundamental approach in sequencing DNA lies in cutting up DNA segments at random and sequencing them.

In this section, we will first examine aspects of the experimental setup for the overlap-layout-consensus approach, and then we will move forward to learning about how to combine reads and learn information from them.

### 5.2.1 Setting up the experiment

The first challenge that must be tackled when setting up this experiment is that we need to start with many copies of each chromosome in order to use this approach. This number is on the order of  $10^5$ . It is important to note that the way we obtain these copies is very important and will affect our outcomes later on as it many of the comparisons we make will depend on consistent data. The first way that we may think

to get this much data is to amplify a given genome. However, amplification does damage which will throw off our algorithms in later steps and cause worse results. Another possible method would be to inbreed the genome to get many copies of each chromosome. If you are looking to get rid of polymorphism, this may be a good technique, but we also lose valuable data from the polymorphic sites when we inbreed. A suggested method for obtaining this data is to use one individual, though the organism would need to be rather large. We could also use techniques such as progeny of one or progeny of two to get as few versions of each chromosome as possible. This will get high sequencing depth on each chromosome, which is the reason we want all chromosomes to be as similar as possible.

Next, let's look at how we could decide on our read lengths given current technology. Looking at (Figure 5.2), we can see that a cost-benefit analysis must be done to decide which platform to use on a given project. With current technology, we commonly use HiSeq2500 with a read length of about 250, though this is rapidly changing.

<b>capability</b>	<b>platform</b>	<b>read length</b>	<b>scale</b>	<b>cost</b>	<b>other</b>
short reads	HiSeq 2500	125	high	medium	high output mode
longer reads	HiSeq 2500	250	high	medium	rapid run mode
	MiSeq	300	low	high	
cheap human reads	HiSeq X	150	high	low	human only min 10 purchase

Figure 5.2: Here is a quick look at a few platforms that can be used to read genomes.

Finally, let's look at a few sequences that cause trouble when using platforms with short reads. Sequences with high GC content (e.g. GGCGGCGATC), low GC content (e.g. AAATAATCAA), or low complexity (e.g. ATATATATA) can cause trouble with short reads. This is still an active area of research, but some possible explanations include Polymerase slippage and DNA denaturing too easily or not easily enough.

This section will examine one of the most successful early methods for computationally assembling a genome from a set of DNA reads, called ?shotgun sequencing? (Figure 5.3). Shotgun sequencing involves randomly shearing multiple copies of the same genome into many small fragments, as if the DNA were shot with a shotgun. Typically, the DNA is actually fragmented using either sonication (brief bursts from an ultrasound) or a targeted enzyme designed to cleave the genome at specific sequence motifs. Both of these methods can be tuned to create fragments of varying sizes.

After the DNA has been amplified and fragmented, the technique developed by Frederick Sanger in 1977 called chain-termination sequencing (also called Sanger sequencing) is used to sequence the fragments. In brief, fragments are extended by DNA polymerase until a dideoxynucleotriphosphate is incorporated; these special nucleotides cause the termination of a fragment's extension as they lack an available hydroxide for chain elongation. The length of the fragment therefore becomes a proxy for where a given ddNTP was added in the sequence. One combines the fragmented DNA with a mix of DNA polymerase, dNTPs, and a

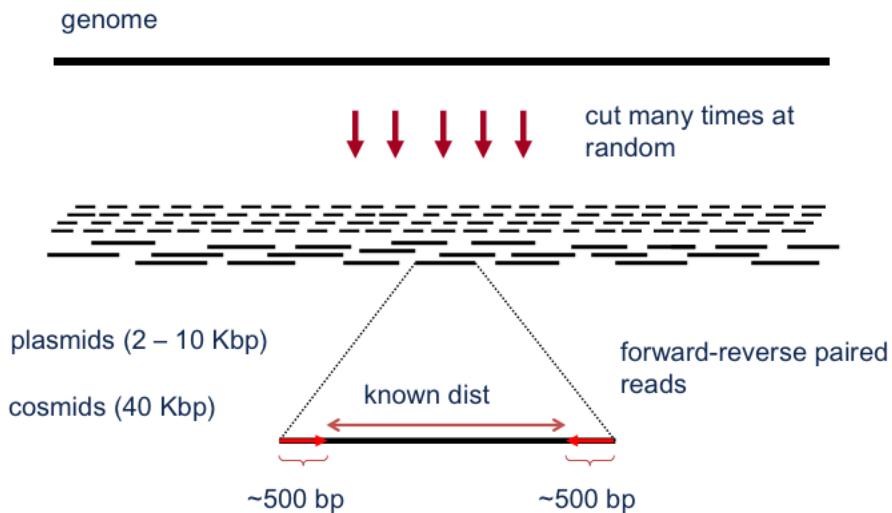


Figure 5.3: Shotgun sequencing involves randomly shearing a genome into small fragments so they can be sequenced, and then computationally reassembling them into a continuous sequence.

small fraction of radioactively or fluorescently labeled ddNTPs (A, G, C, T) and then run out the results on a gel in order to determine the relative ordering of bases. The result is many sequences of bases with corresponding per-base quality scores, indicating the probability that each base was called correctly. The shorter fragments can be fully sequenced, but the longer fragments can only be sequenced at each of their ends since the quality diminishes significantly after about 500-900 base pairs. While this method is still used on smaller scale project, its fundamental principles have been leveraged and expanded in designing Next Generation Sequencing (NGS), commonly used for large scale projects since the early 2000s. Illumina sequencing, a type of NGS, is a popular choice for sequencing which uses 100-150 base pair reads. Fragments are ligated to generic adaptors, which are in turn used to anneal the fragments to a slide. PCR is carried out to amplify each read and the double-stranded products are then separated into individual strands. The slide is then flooded with nucleotides (containing a terminator and fluorescently labelled based on the base), DNA polymerase. After each nucleotide is added, an image of the slide is taken, the terminator is removed, and the next one is added. This process continues until the entire strand is sequenced. Computers are then used to reconstruct the sequence based on the imaging sequence. These paired-end reads are called *mate pairs*. In the rest of this section, we discuss how to use the reads to construct much longer sequences, up to the size of entire chromosomes.

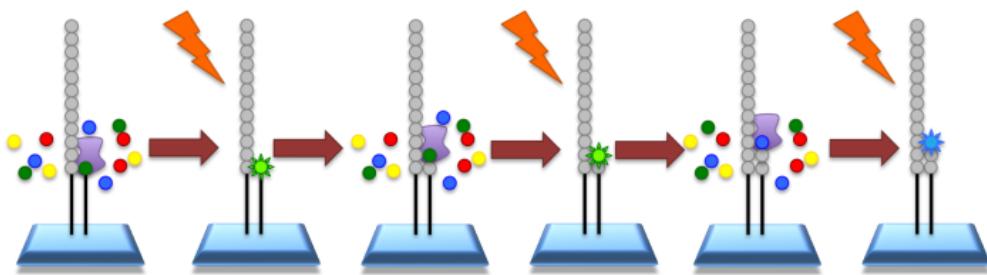


Figure 5.4: We can use Illumina Sequencing, a type of Next Generation Sequencing, to sequence fragments of DNA.

### 5.2.2 Finding overlapping reads

Using the described sequencing methods, DNA fragments can be sequenced into specific segment lengths (determined by the size of the primers, such as plasmids, used in replication, which predetermines the spaces between reads). As the DNA fragments were cut at random, they are probabilistically very likely to overlap, which can be exploited to produce larger reads. To combine the DNA fragments into larger segments, we must find places where two or more reads overlap, i.e. where the beginning sequence of one fragment matches the end sequence of another fragment. By having sequences of length  $y$ , with an overlap of length  $x$  (and therefore an overhang of  $y-x$  on each side), we can determine a sequence of length  $x+2*(y-x)$ . For example, given two fragments such as ACGTTGACCGCATTGCCATA and GACCGCATTGCCATACGGCATT, we can construct a larger sequence based on the overlap: ACGTTGACCGCATTGCCATACGGCATT (Figure 5.5).



Figure 5.5: Constructing a sequence from read overlap

One method for finding matching sequences is the Needleman-Wunsch dynamic programming algorithm, which was discussed in chapter 2. The Needleman-Wunsch method is impractical for genome assembly, however, since we would need to perform millions of pairwise-alignments, each taking  $O(n^2)$  time, in order to construct an entire genome from the DNA fragments.

A better approach is to use the BLAST algorithm (discussed in chapter 3) to hash all the  $k$ -mers (unique sequences of length  $k$ ) in the reads and find all the locations where two or more reads have one of the  $k$ -mers in common. This allows us to achieve  $O(k^n)$  efficiency rather than  $O(n^2)$  pairwise comparisons.  $k$  can be any number smaller than the size of the reads, but varies depending on the desired sensitivity and specificity. By adjusting the read length to span the repetitive regions of the genome, we can correctly resolve these regions and come very close to the ideal of a complete, continuous genome. One popular overlap-layout-consensus assembler developed by Serafim Batzoglou called ARACHNE uses  $k = 24$  [2]. It uses kmers of size 24, because that is a unique length kmer in the human genome. ARACHNE has several key features: an efficient and sensitive procedure for finding overlaps; a way to score these overlaps that corrects errors before assembly; read merger; and detection of repeat contigs. ARACHNE has been used to sequence the human genome.

Given the matching  $k$ -mers, we can align each of the corresponding reads and discard any matches that are less than 97% similar. We do not require that the reads be identical since we allow for the possibility of sequencing errors (i.e. an adenine is mistakenly replaced by a guanine, which may happen as both are purines) and heterozygosity (i.e., a diploid organism like a human may have two different variants at a polymorphic site).

### 5.2.3 Merging reads into contigs

Using the techniques described above to find overlaps between DNA fragments, we can piece together larger segments of continuous sequences called *contigs*. One way to visualize this process is to create a graph in which all the nodes represent reads, and the edges represent overlaps between the reads (Figure 5.6). Our graph will have *transitive overlap*; that is, some edges will connect disparate nodes that are already connected by intermediate nodes. By removing the transitively inferable overlaps, we can create a chain of reads that have been ordered to form a larger contig. These graph transformations are discussed in greater

depth in section 5.3.1 below. In order to get a better understanding of the size of contigs, we calculate something known as  $N50$ . Because measures of contig length tend to be highly sensitive to the smallest contig cutoff,  $N50$  is calculated as the length-weighted median. For a human,  $N50$  is usually close to 125 kb.

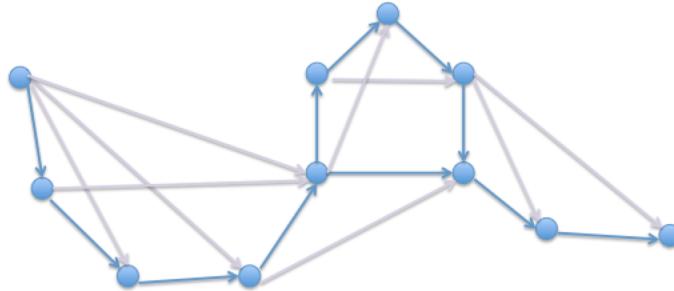


Figure 5.6: We can visualize the process of merging fragments into contigs by letting the nodes in a graph represent reads and edges represent overlaps. By removing the transitively inferable edges (the pink edges in this image), we are left with chains of reads ordered to form contigs.

In theory, we should be able to use the above approach to create large contigs from our reads as long as we have adequate coverage of the given region. In practice, we often encounter large sections of the genome that are extremely repetitive and as a result are difficult to assemble. For example, it is unclear exactly how to align the following two sequences: ATATATAT and ATATATATAT. Due to the extremely low information content in the sequence pattern, they could overlap in any number of ways. Furthermore, these repetitive regions may appear in multiple locations in the genome, and it is difficult to determine which reads come from which locations. One has to recognize the boundaries of the repetitive regions to identify where the sequences diverge. If there are two very large identical segments flanked by different regions, there is no technology to say how they come together. The contig can only be built out to a certain place, at which point it diverges. Contigs made up of these ambiguous, repetitive reads are called *overcollapsed contigs*.

In order to determine which sections are overcollapsed, it is often possible to quantify the depth of coverage of fragments making up each contig. If one contig has significantly more coverage than the others, it is a likely candidate for an overcollapsed region. If one sequence is 10 times the size of the genome, the regions that appear more than 10 times indicate the divergent points. For example, if a sequence appears 20 times, there must be two identical sequences present, and if it appears 30 times, there must be three. Additionally, several unique contigs may overlap one contig in the same location, which is another indication that the contig may be overcollapsed.

(Figure 5.7).

After fragments have been assembled into contigs up to the point of a possible repeated section, the result is a graph in which the nodes are contigs, and the edges are links between unique contigs and overcollapsed contigs (Figure 5.8).

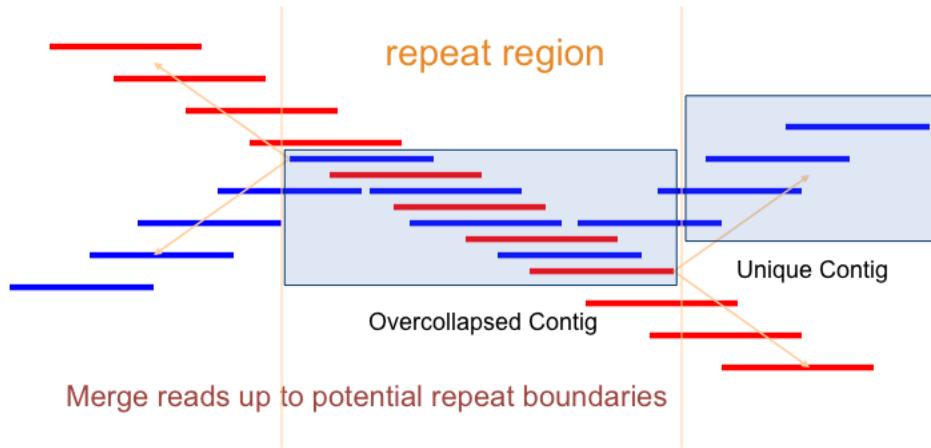


Figure 5.7: Overcollapsed contigs are caused by repetitive regions of the genome which cannot be distinguished from one another during sequencing. Branching patterns of alignment that arise during the process of merging fragments into contigs are a strong indication that one of the regions may be overcollapsed.

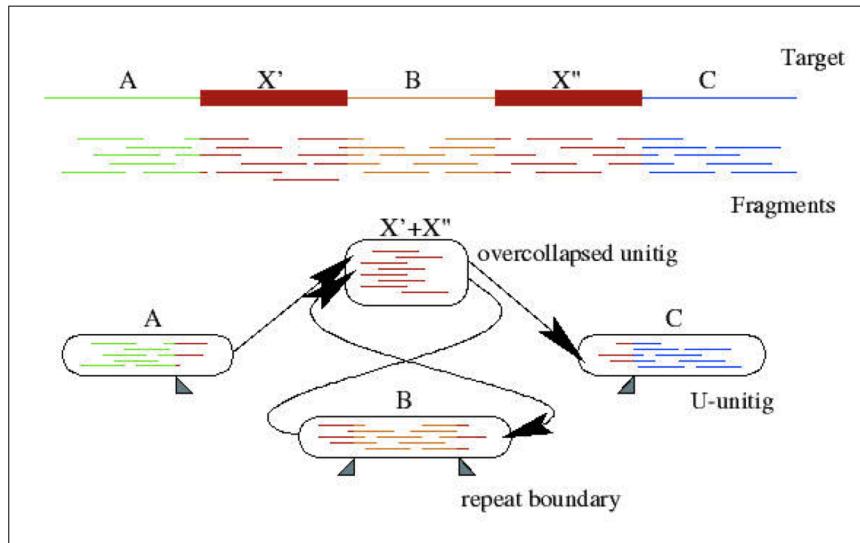


Figure 5.8: In this graph connecting contigs, repeated region X has indegree and outdegree equal to 2. The target seqence shown at the top can be inferred from the links in the graph.

#### 5.2.4 Laying out contig graph into scaffolds

Once our fragments are assembled into contigs and contig graphs, we can use the larger mate pairs to link contigs into *supercontigs* or *scaffolds*. Mate pairs are useful both to orient the contigs and to place them in the correct order. If the mate pairs are long enough, they can often span repetitive regions and help resolve the ambiguities described in the previous section (Figure 5.9).

Unlike contigs, supercontigs may contain some gaps in the sequence due to the fact that the mate pairs connecting the contigs are only sequenced at the ends. Since we generally know how long a given mate pair is we can estimate how many base pairs are missing, but due to the randomness of the cuts in shotgun sequencing, we may not have the data available to fill in the exact sequence. Filling in every single gap can

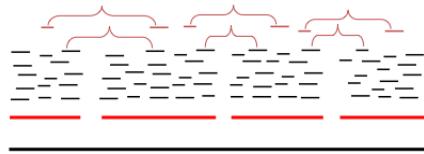


Figure 5.9: Mate pairs help us determine the relative order of contigs in order to link them into supercontigs.

be extremely expensive, so even the most completely assembled genomes usually contain some gaps.

### 5.2.5 Deriving consensus sequence

The goal of genome assembly is to create one continuous sequence, so after the reads have been aligned into contigs, we need to resolve any differences between them. As mentioned above, some of the overlapping reads may not be identical due to sequencing errors or polymorphism. We can often determine when there has been a sequencing error when one base disagrees with all the other bases aligned to it. Taking into account the quality scores on each of the bases, we can usually resolve these conflicts fairly easily. This method of conflict resolution is called ?weighted voting? (Figure 5.10). Another alternative is to ignore the frequencies of each base and take the maximum quality letter as the consensus. Sometimes, you will want to keep all of the bases that form a polymorphic set because it can be important information. In this case, we would be unable to use these methods to derive a consensus sequence.

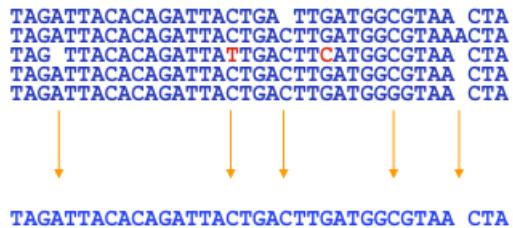


Figure 5.10: We derive the multiple alignment consensus sequence by weighted voting at each base.

In some cases, it is not possible to derive a consensus if, for example, the genome is heterozygous and there are equal numbers of two different bases at one location. In this case, the assembler must choose a representative.

#### **Did You Know?**

Since polymorphism can significantly complicate the assembly of diploid genomes, some researchers induce several generations of inbreeding in the selected species to reduce the amount of heterozygosity before attempting to sequence the genome.

In this section, we saw an algorithm to do genome assembly given reads. However, this algorithm works well when the reads are 500 - 900 bases long or more, which is typical of Sanger sequencing. Alternate genome assembly algorithms are required if the reads we get from our sequencing methods are much shorter.

## 5.3 Genome Assembly II: String graph methods

Shotgun sequencing, which is a more modern and economic method of sequencing, gives reads that around 100 bases in length. The shorter length of the reads results in a lot more repeats of length greater than that of the reads. Hence, we need new and more sophisticated algorithms to do genome assembly correctly.

### 5.3.1 String graph definition and construction

The idea behind string graph assembly is similar to the graph of reads we saw in section 5.2.2. In short, we are constructing a graph in which the nodes are sequence data and the edges are overlap, and then trying to find the most robust path through all the edges to represent our underlying sequence.

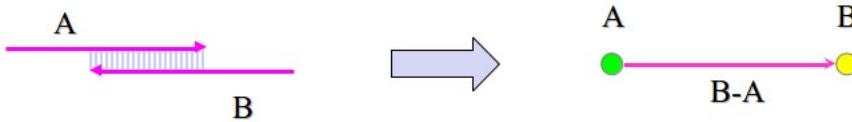


Figure 5.11: Constructing a string graph.

Starting from the reads we get from Shotgun sequencing, a string graph is constructed by adding an edge for every pair of overlapping reads. Note that the vertices of the graph denote junctions, and the edges correspond to the string of bases. A single node corresponds to each read, and reaching that node while traversing the graph is equivalent to reading all the bases upto the end of the read corresponding to the node. For example, in figure 5.11, we have two overlapping reads A and B and they are the only reads we have. The corresponding string graph has two nodes and two edges. One edge doesn't have a vertex at its tail end, and has A at its head end. This edge denotes all the bases in read A. The second edge goes from node A to node B, and only denotes the bases in B-A (the part of read B which is not overlapping with A). This way, when we traverse the edges once, we read the entire region exactly once. In particular, notice that we do not traverse the overlap of read A and read B twice.

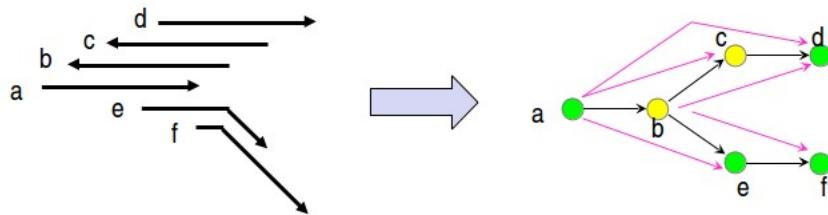


Figure 5.12: Constructing a string graph

There are a couple of subtleties in the string graph (figure 5.12) which need mentioning:

- We have two different colors for nodes since the DNA can be read in two directions. If the overlap is between the reads as is, then the nodes receive same colors. And if the overlap is between a read and the complementary bases of the other read, then they receive different colors.

- Secondly, if A and B overlap, then there is ambiguity in whether we draw an edge from A to B, or from B to A. Such ambiguity needs to be resolved in a consistent manner at junctions caused due to repeats.

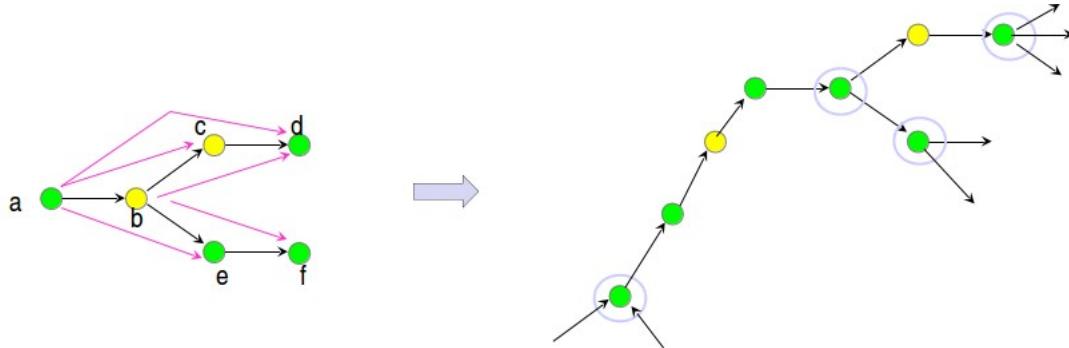


Figure 5.13: Example of string graph undergoing removal of transitive edges.

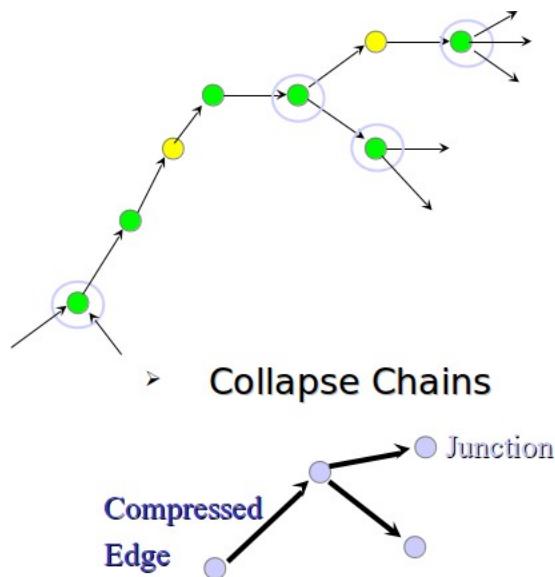


Figure 5.14: Example of string graph undergoing chain collapsing.

After constructing the string graph from overlapping reads, we:-

- *Remove transitive edges:* Transitive edges are caused by transitive overlaps, i.e. A overlaps B overlaps C in such a way that A overlaps C. There are randomized algorithms which remove transitive edges in  $O(E)$  expected runtime. In figure 5.13, you can see the an example of removing transitive edges.
- *Collapse chains:* After removing the transitive edges, the graph we build will have many chains where each node has one incoming edge and one outgoing edge. We collapse all these chains to a single edge. An example of this is shown in figure 5.14.

### 5.3.2 Flows and graph consistency

After doing everything mentioned above we will get a pretty complex graph, i.e. it will still have a number of junctions due to relatively long repeats in the genome compared to the length of the reads. We will now see how the concepts of flows can be used to deal with repeats.

First, we estimate the weight of each edge by the number of reads we get corresponds to the edge. If we have double the number of reads for some edge than the number of DNAs we sequenced, then it is fair to assume that this region of the genome gets repeated. However, this technique by itself is not accurate enough. Hence sometimes we may make estimates by saying that the weight of some edge is  $\geq 2$ , and not assign a particular number to it.

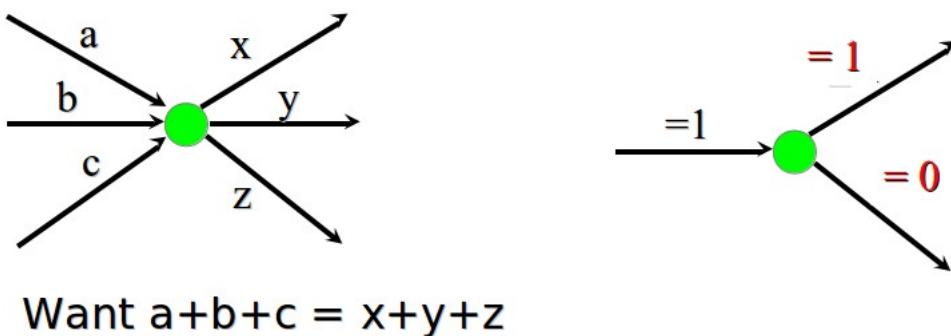


Figure 5.15: *Left:* Flow resolution concept. *Right:* Flow resolution example.

We use reasoning from flows in order to resolve such ambiguities. We need to satisfy the flow constraint at every junction, i.e. the total weight of all the incoming edges must equal the total weight of all the outgoing edges. For example, in the figure 5.15 there is a junction with an incoming edge of weight 1, and two outgoing edges of weight  $\geq 0$  and  $\geq 1$ . Hence, we can infer that the weights of the outgoing edges are exactly equal to 0 and 1 respectively. A lot of weights can be inferred this way by iteratively applying this same process throughout the entire graph.

### 5.3.3 Feasible flow

Once we have the graph and the edge weights, we run a min cost flow algorithm on the graph. Since larger genomes may not have a unique min cost flow, we iteratively do the following:

- Add  $\epsilon$  penalty to all edges in solution
- Solve flow again - if there is an alternate min cost flow it will now have a smaller cost relative to the previous flow
- Repeat until we find no new edges

After doing the above, we will be able to label each edge as one of the following

- *Required:* edges that were part of all the solutions
- *Unreliable:* edges that were part of some of the solutions
- *Not required:* edges that were not part of any solution

### 5.3.4 Dealing with sequencing errors

There are various sources of errors in the genome sequencing procedure. Errors are generally of two different kinds, local and global.

Local errors include insertions, deletions and mutations. Such local errors are dealt with when we are looking for overlapping reads. That is, while checking whether reads overlap, we check for overlaps while being tolerant towards sequencing errors. Once we have computed overlaps, we can derive a consensus by mechanisms such as removing indels and mutations that are not supported by any other read and are contradicted by at least 2.

Global errors are caused by other mechanisms such as two different sequences combining together before being read, and hence we get a read which is from different places in the genome. Such reads are called chimeras. These errors are resolved while looking for a feasible flow in the network. When the edge corresponding to the chimer is in use, the amount of flow going through this edge is smaller compared to the flow capacity. Hence, the edge can be detected and then ignored.

Each step of the algorithm is made as robust and resilient to sequencing errors as possible. And the number of DNAs split and sequenced is decided in a way so that we are able to construct most of the DNA (i.e. fulfill some quality assurance such as 98% or 95%).

### 5.3.5 Resources

Some popular genome assemblers using String Graphs are listed below

- Euler (Pevzner, 2001/06) : Indexing → deBruijn graphs → picking paths → consensus
- Valvel (Birney, 2010) : Short reads → small genomes → simplification → error correction
- ALLPATHS (Gnerre, 2011) : Short reads → large genomes → jumping data → uncertainty

## 5.4 Whole-Genome Alignment

Once we have access to whole-genome sequences for several different species, we can attempt to align them in order to infer the path that evolution took to differentiate these species. In this section we discuss some of the methods for performing whole-genome alignments between multiple species.

### 5.4.1 Global, local, and 'glocal' alignment

The Needleman-Wunsch algorithm discussed in chapter 2 is the best way to generate an optimal alignment between two or more genome sequences of limited size. At the level of whole genomes, however, the  $O(n^2)$  time bound is impractical.

When aligning the genomes of two species, we use the Needleman-Wunsch algorithm and traverse a rectangle in two dimensions, yielding this  $O(n^2)$  runtime. Similarly, when we align three species, we traverse a cube, yielding an  $O(n^3)$  runtime. In order to find an optimal alignment between  $k$  different species, the time for the Needleman-Wunsch algorithm is extended to  $O(n^k)$ , which becomes even more impractical.

For genomes that are millions of bases long, this run time is prohibitive (Figure 5.16).

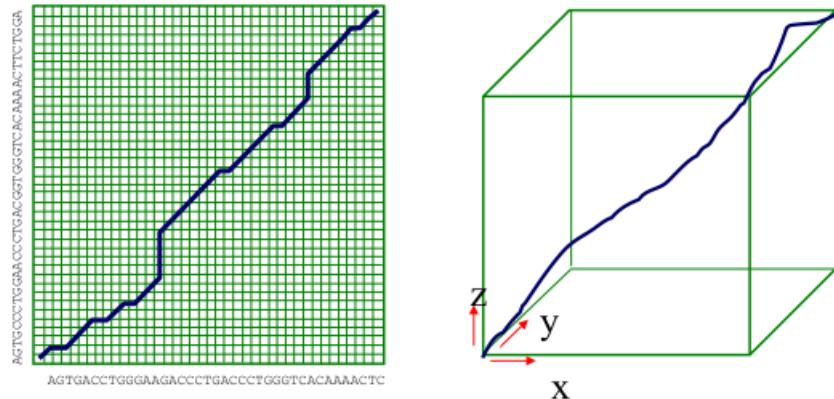


Figure 5.16: The Needleman-Wunsch algorithm for alignments of 2 and 3 genomes.

If even the  $O(n^2)$  runtime is impractical for comparing the genomes of two species, then we need to find a way to reduce the complexity of our algorithm, which retaining a close to optimal solution. We previously discussed the concept of bounded-DP, where we restrict our DP matrix to some subset of the overall space to reduce runtime.

Motivated by this, we consider the alternative: local alignment. We use an efficient local alignment tool such as BLAST to find all of the local alignments, and then chain them together along the diagonal to form global alignments. This approach can save a significant amount of time, since the process of finding local alignments is very efficient, and then we only need to perform the time-consuming Needleman-Wunsch algorithm in the small rectangles between local alignments (Figure 5.17).

This is based off of the intuition that going off the diagonal is costly, and is only worth it we are moving to a very good alignment. This restriction reduces our bounded DP for two species to  $O(m^2)$ , where  $m \ll n$ .

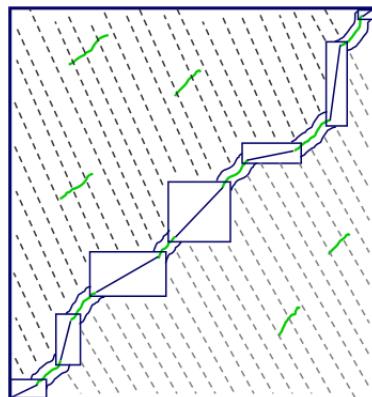


Figure 5.17: We can save time when performing a global alignment by first finding all the local alignments and then chaining them together along the diagonal with restricted dynamic programming.

This local alignment technique has the added advantage of allowing us to detect inversions, duplications and translocations. Then we can chain these elements together using the least-cost transformations between sequences. This approach is commonly called ?global? alignment, since it seeks to combine the best of local

and global alignment to create the most accurate picture of how genomes evolve over time. In Figure 5.18, we observe the detection of an inverse and translocation (from bottom left to top right).

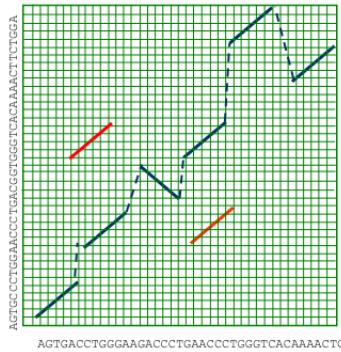


Figure 5.18: Glocal alignment allows for the possibility of duplications, inversion, and translocations.

### 5.4.2 Lagan: Chaining local alignments

LAGAN is a popular software toolkit that incorporates many of the above ideas and can be used for local, global, glocal, and multiple alignments between species.

The regular LAGAN algorithm consists of finding local alignments, chaining local alignments along the diagonal, and then performing restricted dynamic programming to find the optimal path between local alignments.

Multi-LAGAN uses the same approach as regular LAGAN but generalizes it to multiple species alignment. In this algorithm, the user must provide a set of genomes and a corresponding phylogenetic tree. Multi-LAGAN performs pairwise alignment guided by the phylogenetic tree. It first compares highly related species, and then iteratively compares more and more distant species.

Shuffle-LAGAN is a glocal alignment tool that finds local alignments, builds a rough homology map, and then globally aligns each of the consistent parts (Figure 5.19). In order to build a homology map, the algorithm chooses the maximum scoring subset of local alignments based on certain gap and transformation penalties, which form a non-decreasing chain in at least one of the two sequences. Unlike regular LAGAN, all possible local alignment sequences are considered as steps in the glocal alignment, since they could represent translocations, inversions and inverted translocations as well as regular untransformed sequences. Once the rough homology map has been built, the algorithm breaks the homologous regions into chunks of local alignments that are roughly along the same continuous path. Finally, the LAGAN algorithm is applied to each chunk to link the local alignments using restricted dynamic programming.

By running Shuffle-LAGAN or other glocal alignment tools, we can discover inversions, translocations, and other homologous relations between different species. By mapping the connections between these rearrangements, we can gain insight into how each species evolved from the common ancestor (Figure 5.20).

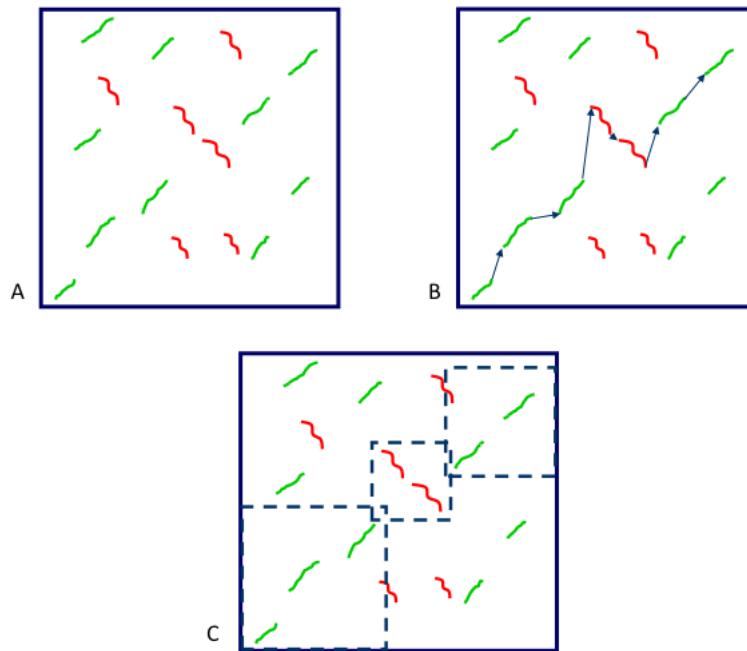


Figure 5.19: The steps to run the SLAGAN algorithm are A. Find all the local alignments, B. Build a rough homology map, and C. globally align the consistent parts using the regular LAGAN algorithm

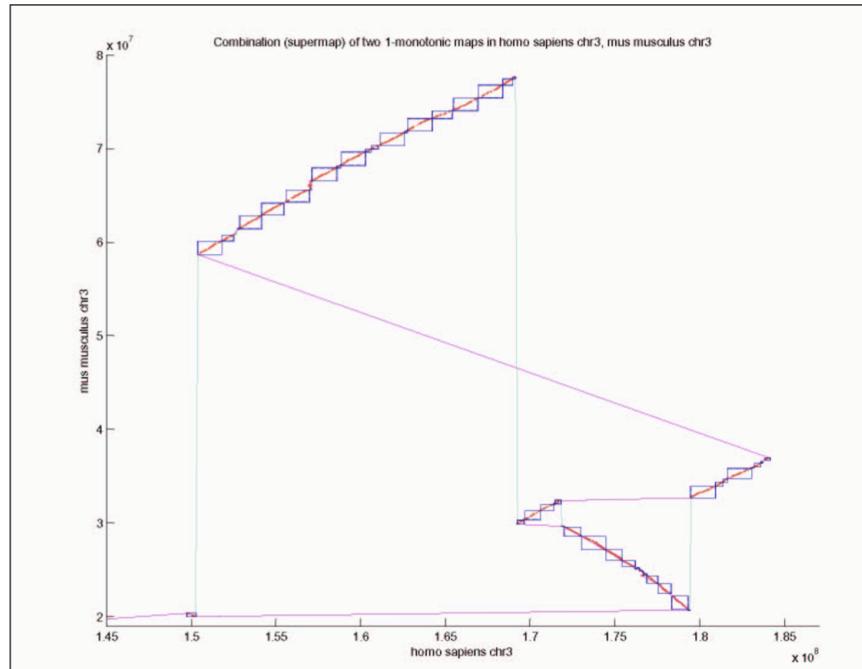


Figure 5.20: Using the concepts of glocal alignment, we can discover inversions, translocations, and other homologous relations between different species such as human and mouse.

## 5.5 Gene-based region alignment

( 58)

An alternative way for aligning multiple genomes anchors genomic segments based on the genes that they contain, and uses the correspondence of genes to resolve corresponding regions in each pair of species. A nucleotide-level alignment is then constructed based on previously-described methods in each multiply-conserved region.

Because not all regions have one-to-one correspondence and the sequence is not static, this is more difficult: genes undergo divergence, duplication, and losses and whole genomes undergo rearrangements. To help overcome these challenges, researchers look at the amino-acid similarity of gene pairs across genomes and the locations of genes within each genome.

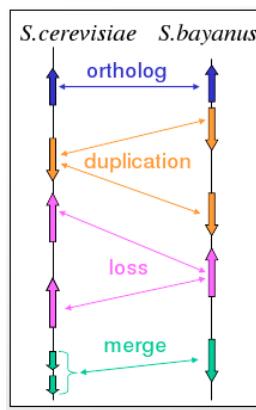


Figure 5.21: Graph of *S. cerevisiae* and *S. bayanus* gene correspondence.

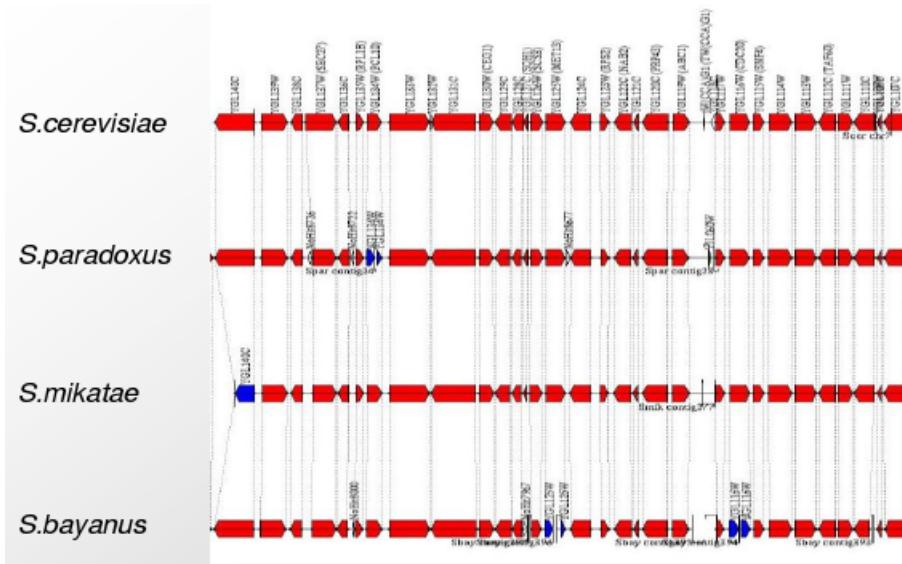
Gene correspondence can be represented by a weighted bipartite graph with nodes representing genes with coordinates and edges representing weighted sequence similarity (Figure 5.21). Orthologous relationships are one-to-one matches and paralogous relationships are one-to-many or many-to-many matches. The graph is first simplified by eliminating spurious edges and then edges are selected based on available information such as blocks of conserved gene order and protein sequence similarity.

The Best Unambiguous Subgroups (BUS) algorithm can then be used to resolve the correspondence of genes and regions. BUS extends the concept of best-bidirectional hits and uses iterative refinement with an increasing relative threshold. It uses the complete bipartite graph connectivity with integrated amino acid similarity and gene order information.

### Did You Know?

A bipartite graph is a graph whose vertices can be split into two disjoint sets U and V such that every edge connects a vertex in U to a vertex in V.

In the example of a correctly resolved gene correspondence of *S. cerevisiae* with three other related species, more than 90% of the genes had a one-to-one correspondence and regions and protein families of rapid change were identified.

Figure 5.22: Illustration of gene correspondence for *S. cerevisiae* Chromosome VI (250-300bp).

## 5.6 Mechanisms of Genome Evolution

Now that we have assembled and aligned regions of the genome, we can ask ourselves, what are the differences between genomes? How can we observe the history of evolution written in these differences? So, once we have alignments of large genomic regions (or whole genomes) across multiple related species, we can begin to make comparisons in order to infer the evolutionary histories of those regions.

Rates of evolution vary across species and across genomic regions. In *S. cerevisiae*, for example, 80% of ambiguities are found in 5% of the genome. Telomeres are repetitive DNA sequences at the end of chromosomes which protect the ends of the chromosomes from deterioration. Telomere regions are inherently unstable, tending to undergo rapid structural evolution, and the 80% of variation corresponds to 31 of the 32 telomeric regions. Gene families contained within these regions such as HXT, FLO, COS, PAU, and YRF show significant evolution in number, order, and orientation. Several novel and protein-coding sequences can be found in these regions. Since very few genomic rearrangements are found in *S. cerevisiae* aside from the telomeric regions, regions of rapid change can be identified by protein family expansions in chromosome ends.

Genes evolve at different rates. For example as illustrated in Figure 5.23, on one extreme, there is YBR184W in yeast which shows unusually low sequence conservation and exhibits numerous insertions and deletions across species. On the other extreme there is MatA2, which shows perfect amino acid and nucleotide conservation. Mutation rates often also vary by functional classification. For example, mitochondrial ribosomal proteins are less conserved than ribosomal proteins.

The fact that some genes evolve more slowly in one species versus another may be due to factors such as longer life cycles. Lack of evolutionary change in specific genes, however, suggests that there are additional biological functions which are responsible for the pressure to conserve the nucleotide sequence. Yeast can switch mating types by switching all their A and  $\alpha$  genes and MatA2 is one of the four yeast mating-type genes (MatA2, Mata2, MatA1, Mata1). Its role could potentially be revealed by nucleotide conservation analysis.

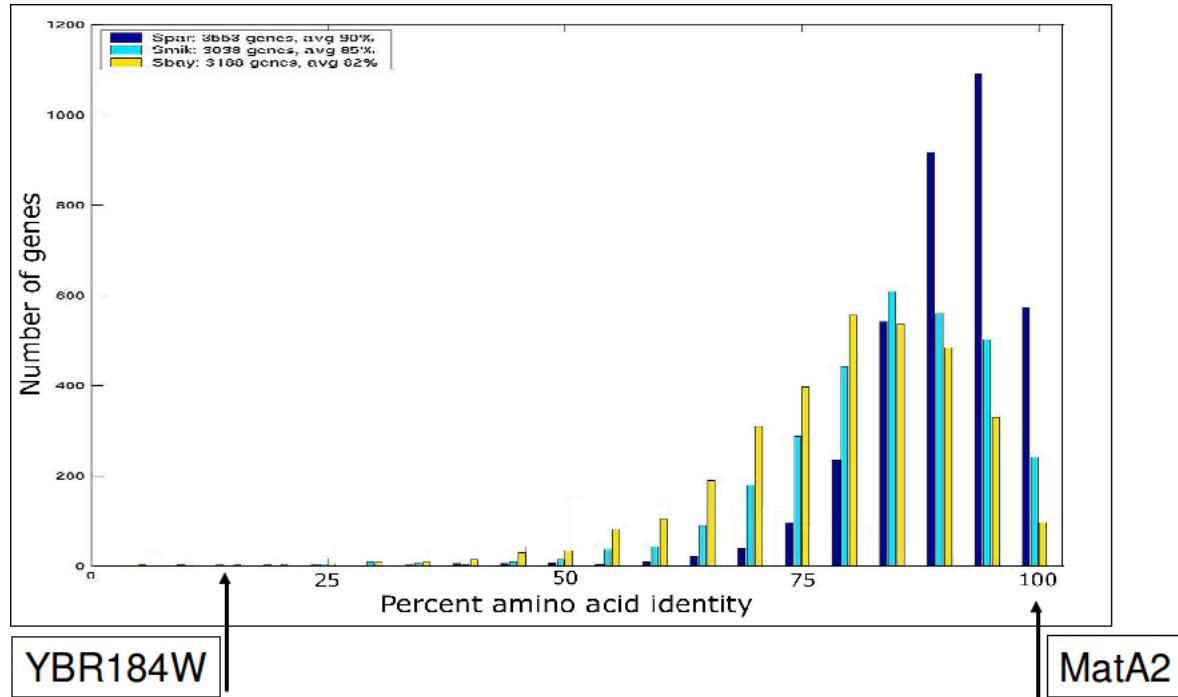


Figure 5.23: Dynamic view of a changing gene.

Fast evolving genes can also be biologically meaningful. Mechanisms of rapid protein change include:

- Protein domain creation via stretches of Glutamine (Q) and Asparagine (N) and protein-protein interactions,
- Compensatory frame-shifts which enable the exploration of new reading frames and reading/creation of RNA editing signals,
- Stop codon variations and regulated read-through where gains enable rapid changes and losses may result in new diversity
- Inteins, which are segments of proteins that can remove themselves from a protein and then rejoin the remaining protein, gain from horizontal transfers of post-translationally self-splicing inteins.

We now look at differences in gene content across different species (*S. cerevisiae*, *S. paradoxus*, *S. mikatae*, and *S. bayanus*). A lot can be revealed about gene loss and conversion by observing the positions of paralogs across related species and observing the rates of change of the paralogs. There are 8-10 genes unique to each genome which are involved mostly with metabolism, regulation and silencing, and stress response. In addition, there are changes in gene dosage with both tandem and segment duplications. Protein family expansions are also present with 211 genes with ambiguous correspondence. All in all however, there are few novel genes in the different species.

### 5.6.1 Regions of Rapid Change

When compare different species of yeast, we see massive expansions in chromosomes on the regions just prior to the telomeres. We observe from these expansions written in the genome that there are classes of proteins evolving rapidly near these telomeres.

In yeast, regions near the end of these chromosomes are involved in cell surface adhesion, and this adaptation reflects that ability of the yeast to adhere to surfaces in different environment through the rapid evolution of these protein classes.

### 5.6.2 Chromosomal Rearrangements

These are often mediated by specific mechanisms as illustrated for *Saccharomyces* in Figure 5.24.

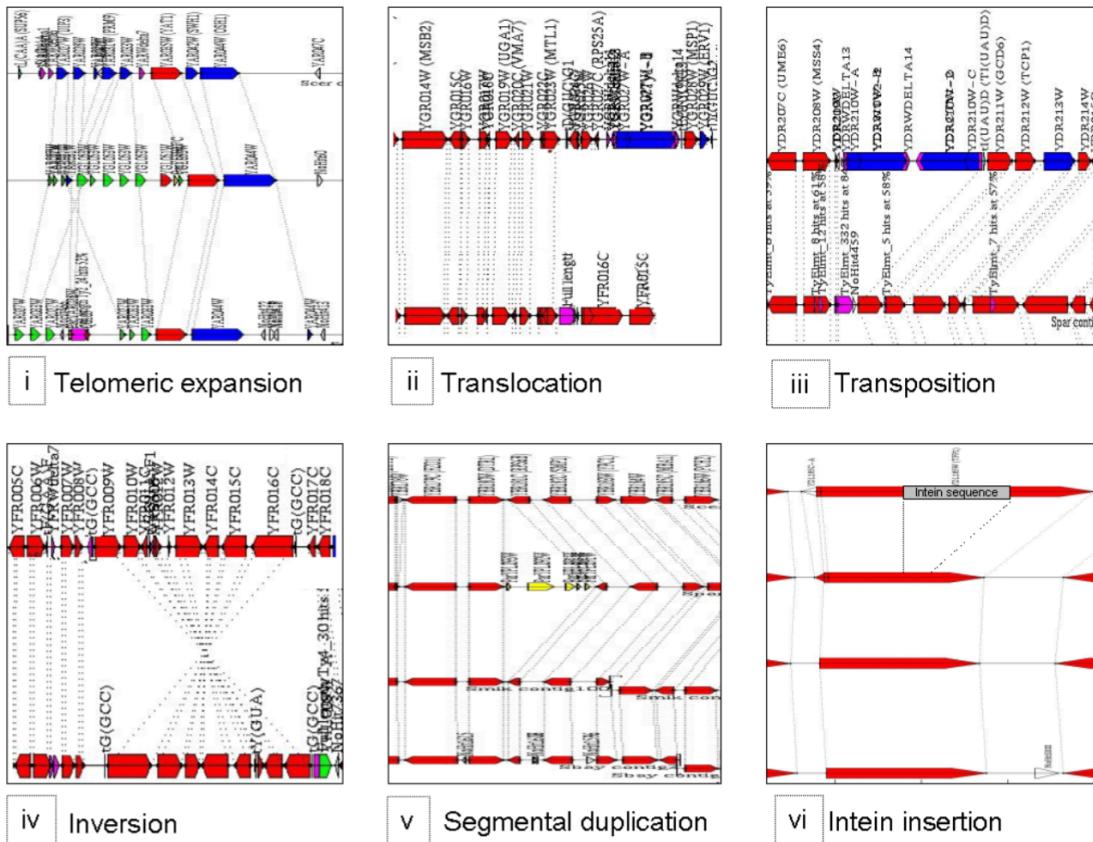


Figure 5.24: Six types of genome rearrangement

Translocations across dissimilar genes often occur across transposable genetic elements (Ty elements in yeast for example). Transposon locations are conserved with recent insertions appearing in old locations and long terminal repeat remnants found in other genomes. They are evolutionarily active however (for example with Ty elements in yeast being recent), and typically appear in only one genome. The evolutionary advantage of such locationally conserved transposons may lie in the possibility of mediating reversible arrangements. Inversions are often flanked by tRNA genes in opposite transcriptional orientation. This may suggest that they originate from recombination between tRNA genes.

### 5.6.3 Rapid Protein Change

We present another interesting mechanism of gene evolution observed in *S.cerevisiae*: intein insertion (as illustrated in Figure 5.24).

\* An intein is a sort of protein intron - a segment of a protein that is able to excise itself and join the remaining portions (the exteins) with a peptide bond in a process termed protein splicing. To transcribe this region, you translate into protein, which folds in such a way that its structure cleaves itself off, binds to the remaining RNA, picks up the RNA and inserts it into the genome. Thus, we observe gain from horizontal transfer.

## 5.7 Whole Genome Duplication

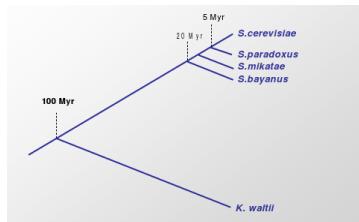


Figure 5.25: Moving further back in evolutionary time for *Saccharomyces*.

As you trace species further back in evolutionary time, you have the ability to ask different sets of questions. In class, the example used was *K. waltii*, which dates to about 95 millions years earlier than *S.cerevisiae* and 80 million years earlier than *S.bayanus*.

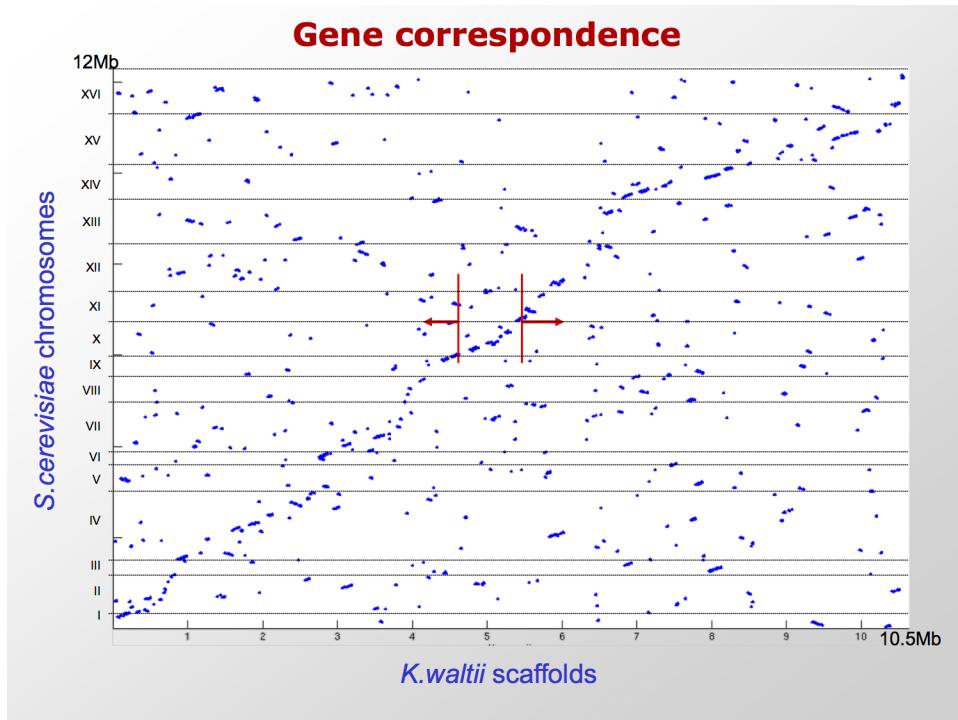


Figure 5.26: *K. waltii* Scaffolds

Looking at the dotplot of *S.cerevisiae* chromosomes and *K.waltii* scaffolds, a divergence was noted along the diagonal in the middle of the plot (Figure 5.26), whereas most pairs of conserved region exhibit a dot

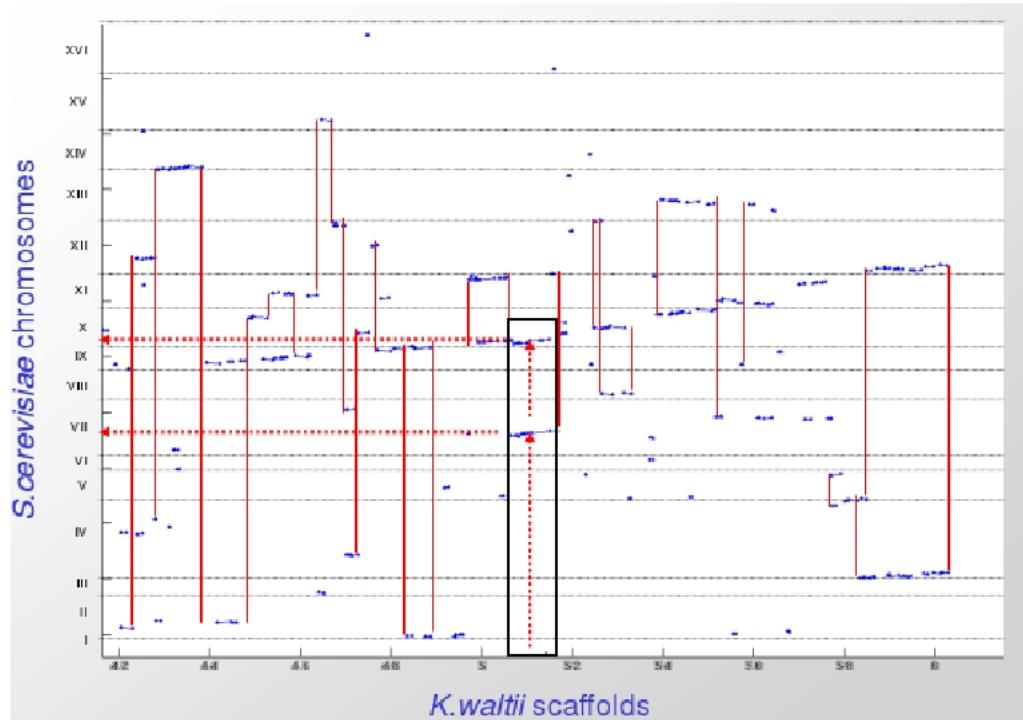


Figure 5.27: Gene Correspondence for *S. cerevisiae* chromosomes and *K. waltii* scaffolds.

plot with a clear and straight diagonal. Viewing the segment at a higher magnification (Figure 5.27), it seems that *S. cerevisiae* sister fragments all map to corresponding *K. waltii* scaffolds.

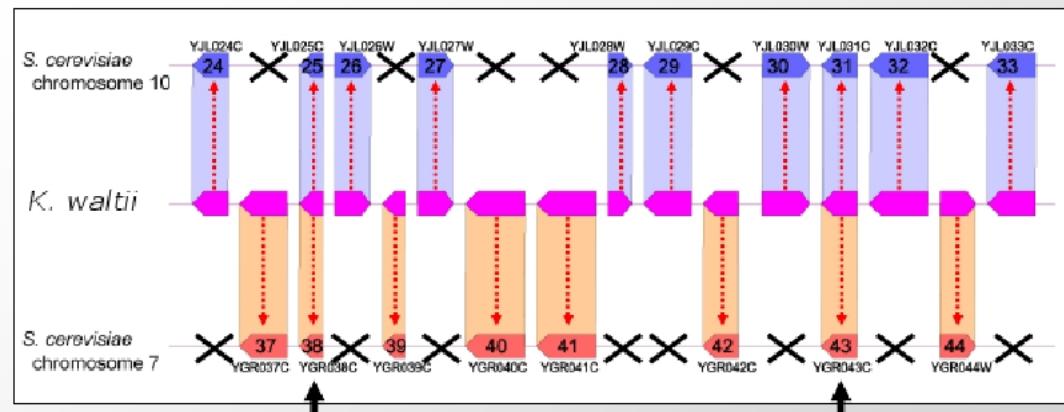


Figure 5.28: Gene interleaving shown by sister regions in *K. waltii* and *S. cerevisiae*

At one region of chromosome of this species, there are two regions that it corresponds to. For example, this newly sequence corresponds to chromosome 7 and chromosome 10. If further tracing the entire chunk of genome, you'll find it's true for almost every single location.

If plot the genes (Figure 5.32), the boundary appears to be the gene itself. The fact that the genes come from two different chromosome indicates that there was a duplication event in the ancestral sequence. The

interleaving gene is essentially from either one or the other copy of gene.

Schematically (Figure 5.28) sister regions show gene interleaving. In duplicate mapping of centromeres, sister regions can be recognized based on gene order. This observed gene interleaving provides evidence of complete genome duplication.

Given that *S. cerevisiae* has only a small, compact genome, the resolution of the duplicated segment is the loss of alternative gene, loss of intermediate gene and compensation (Figure 5.29). In total, after the divergence from *K. waltii*, *S. cerevisiae* has lost 90 percent of the duplicated genome over evolutionary time and is now recognized as a diploid organism.

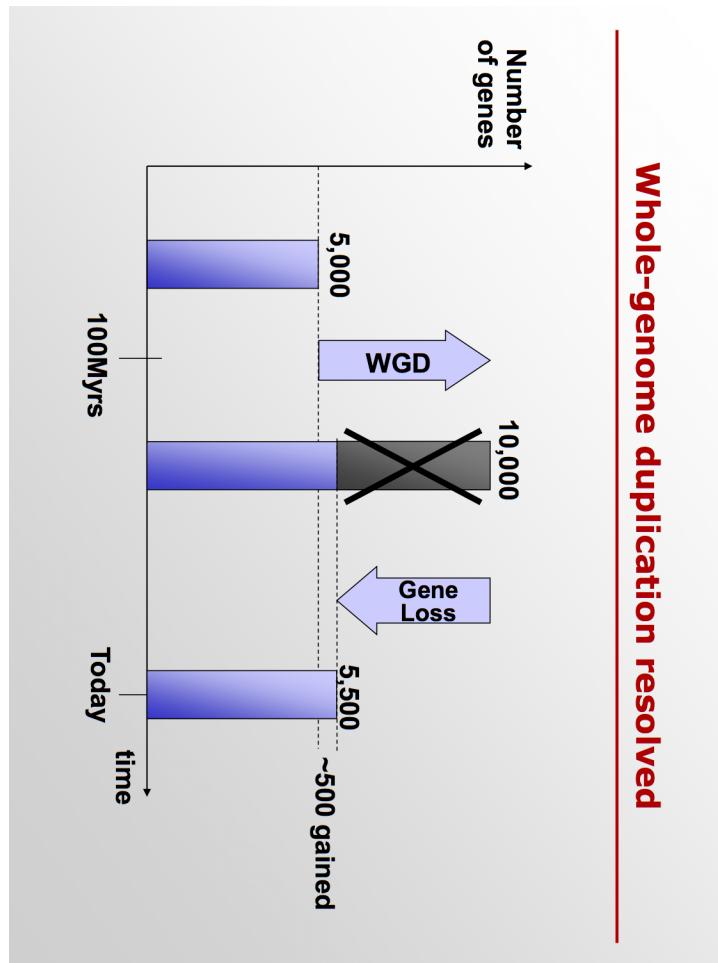


Figure 5.29: Whole-genome duplication resolved

## 5.8 Additional figures

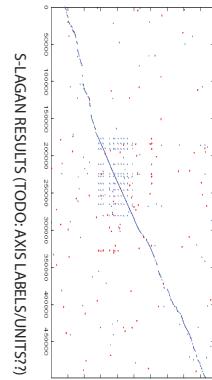


Figure 5.30: S-LAGAN results.

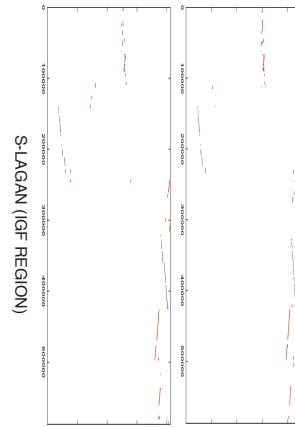


Figure 5.31: S-LAGAN results for IGF locus.

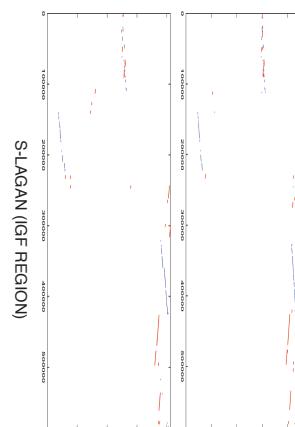


Figure 5.32: S-LAGAN results for IGF locus.

## Bibliography

- [1] Embl allelextron database - cassette exons.
- [2] Batzoglou S et al. Arachne: a whole-genome shotgun assembler. *Genome Res*, 2002.
- [3] Manolis Kellis. Lecture slides 04: Comparative genomics i. September 21,2010.
- [4] Manolis Kellis. Lecture slides 05.1: Comparative genomics ii. September 23, 2010.
- [5] Manolis Kellis. Lecture slides 05.2: Comparative genomics iii, evolution. September 25,2010.
- [6] Nikolaus Rajewsky Kevin Chen. The evolution of gene regulation by transcription factors and micrornas. *Nature Reviews Genetics*, 2007.
- [7] Douglas Robinson and Lynn Cooley. Examination of the function of two kelch proteins generated by stop codon suppression. *Development*, 1997.
- [8] Stark. Discovery of functional elements in 12 drosophila genomes using evolutionary signatures. *Nature*, 2007.
- [9] Angela Tan. Lecture 15 notes: Comparative genomics i: Genome annotation. November 4, 2009.



## BACTERIAL GENOMICS– MOLECULAR EVOLUTION AT THE LEVEL OF ECOSYSTEMS

Guest Lecture by Eric Alm  
Scribed by Deniz Yorukoglu (2011)

### Figures

---

6.1	A tree of life displaying rates of gene birth, duplication, loss, and horizontal gene transfer at each branching point. . . . .	127
6.2	Rates of new gene birth, duplication, loss and horizontal gene transfer during Archean Gene Expansion . . . . .	128
6.3	Abundance levels of different bacterial groups in control patients, Crohn's disease patients and patients with ulcerative colitis. . . . .	129
6.4	Gut bacterial abundances plotted through time for the two donors participating in HuGE project. . . . .	130
6.5	Description of how to read a horizon plot. . . . .	130
6.6	Horizon plot of Donor B in HuGE study. . . . .	131
6.7	Horizon plot of Donor A in HuGE study. . . . .	132
6.8	Day-to-day bacterial abundance correlation matrices of Donor A and Donor B. . . . .	132
6.9	Rate of horizontal gene transfer between different bacterial groups taken from non-human sites, human sites, same site within human, and different sites within human. . . . .	134
6.10	Rate of horizontal gene transfer between bacterial groups sampled from the same continent and from different continents. . . . .	134
6.11	Rate of horizontal gene transfer between different human and non-human sites (top right) and the percentage of antibiotic resistance genes among horizontal gene transfers (bottom left). . . . .	135

---

### 6.1 Introduction

With the magnitude and diversity of bacterial populations in human body, human microbiome has many common properties with natural ecosystems researched in environmental biology. As a field with a large number of quantitative problems to tackle, bacterial genomics offers an opportunity for computational biologist to be actively involved in the progress of this research area.

There are approximately  $10^{14}$  microbial cells in an average human gut, whereas there are only  $10^{13}$  human cells in a human body in total. Furthermore, there are  $10^{12}$  external microbial cells living on our skin. From a cell count perspective, this corresponds to 10 times more bacterial cells in our body than our own cells. From a gene count perspective, there are 100 times more genes belonging to the bacteria living in/on us than to our own cells. For this reason, these microbial communities living in our bodies are an integral part of

what makes us human and we should research upon these genes that are not directly encoded in our genome, but still have a significant effect on our physiology.

### 6.1.1 Evolution of microbiome research

Earlier stages of microbiome research were mostly based on data collection and analysis of surveys of bacterial groups present in a particular ecosystem. Apart from collecting data, this type of research also involved sequencing of bacterial genomes and identification of gene markers for determining different bacterial groups present in the sample. The most commonly used marker for this purpose is 16S rRNA gene, which is a section of the prokaryotic DNA that codes for ribosomal RNA. Three main features of 16S gene that makes it a very effective marker for microbiome studies are: (1) its short size (~1500 bases) that makes it cheaper to sequence and analyze, (2) high conservation due to exact folding requirements of the ribosomal RNA it encodes for, and (3) its specificity to prokaryote organisms that allows us to differentiate from contaminant protist, fungal, plant and animal DNAs.

A further direction in early microbial research was inferring rules from generated datasets upon microbial ecosystems. These studies investigated initially generated microbial data and tried to understand rules of microbial abundance in different types of ecosystems and infer networks of bacterial populations regarding their co-occurrence, correlation and causality with respect to one another.

A more recent type of microbial research takes a predictive approach and aims to model the change of bacterial populations in an ecosystem through time making use of differential equations. For example, we can model the rate of change for the population size of a particular bacterial group in human gut as an ordinary differential equation (ODE) and use this model to predict the size of the population at a future time point by integrating over the time interval.

We can further model change of bacterial populations with respect to multiple parameters, such as time and space. When we have enough data to represent microbial populations temporally and spatially, we can model them using partial differential equations (PDEs) for making predictions using multivariate functions.

### 6.1.2 Data generation for microbiome research

Data generation for microbiome research usually follows the following work-flow: (1) a sample of microbial ecosystem is taken from the particular site being studied (e.g. a patient's skin or a lake), (2) the DNAs of the bacteria living in the sample are extracted, (3) 16S rDNA genes are sequenced, (4) conserved motifs in some fraction of the 16S gene (DNA barcodes) are clustered into **operational taxonomic units** (OTUs), and (5) a vector of abundance is constructed for all species in the sample. In microbiology, bacteria are classified into OTUs according to their functional properties rather than species, due to the difficulty in applying the conventional species definition to the bacterial world.

In the remainder of the lecture, a series of recent studies that are related to the field of bacterial genomics and human microbiome studies are described.

## 6.2 Study 1: Evolution of life on earth

This study [2] is inspired from a quote by Max Delbrück: "Any living cell carries with it the experience of a billion years of experimentation by its ancestors". In this direction, it is possible to find evidence in the genomes of living organisms for ancient environmental changes with large biological impacts. For instance, the oxygen that most organisms currently use would have been extremely toxic to almost all life on earth before the accumulation of oxygen via oxygenic photosynthesis. It is known that this event happened approximately 2.4 billion years ago and it caused a dramatic transformation of life on earth.

A dynamic programming algorithm was developed in order to infer gene birth, duplication, loss and horizontal gene transfer events given the phylogeny of species and phylogeny of different genes. **Horizontal gene transfer** is the event in which bacteria transfer a portion of their genome to other bacteria from different taxonomic groups.

Figure 6.1 shows an overview of these inferred events in a phylogenetic tree focusing on prokaryote life. In each node, the size of the pie chart represents the amount of genetic change between two branches and each colored slice stands for the rate of a particular genetic modification event. Starting from the root of the

tree, we see that almost the entire pie chart is represented by newly born genes represented by red. However, around 2.5 billion years ago green and blue slices become more prevalent, which represent rate of horizontal gene transfer and gene duplication events.

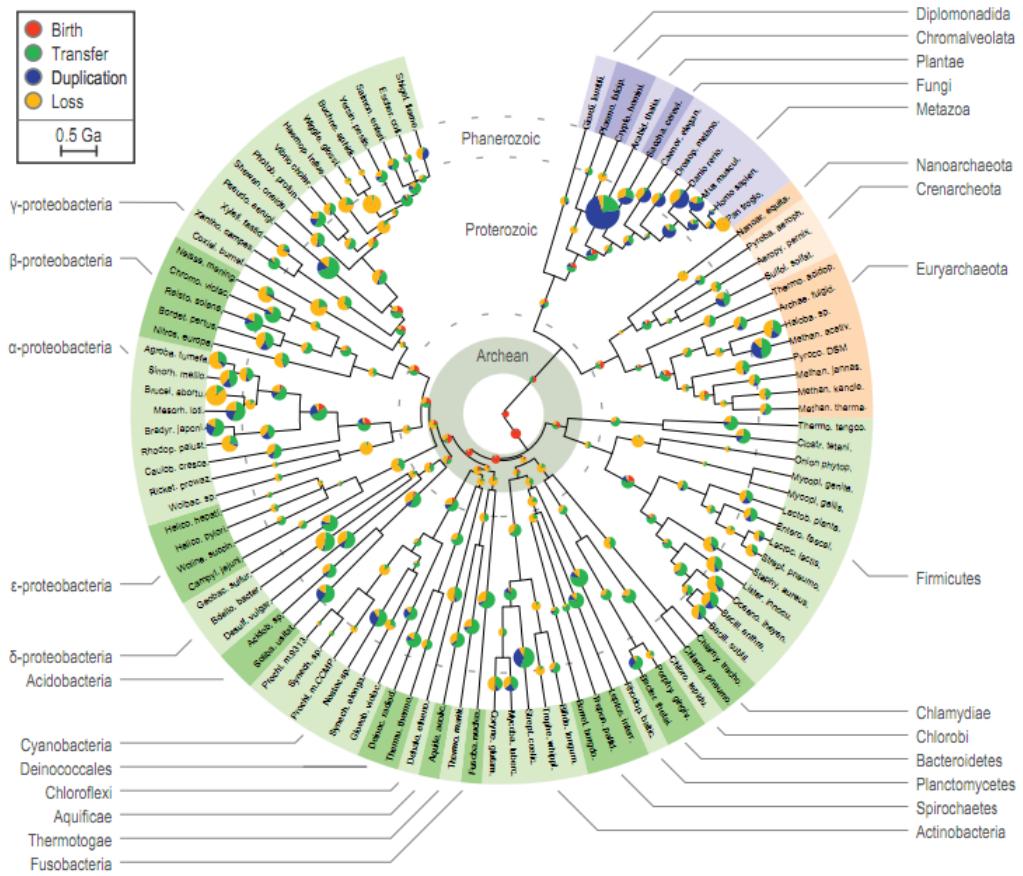


Figure 6.1: A tree of life displaying rates of gene birth, duplication, loss, and horizontal gene transfer at each branching point.

In Figure 6.2, a large spike can be seen during Archean eon representing large amount of genetic change on earth occurring during this particular time period. This study looked for enzymatic activity of genes that were born in this eon different from the genes that were already present. On the right hand side of Figure 6.2, logarithmic enrichment levels of different metabolites are displayed. Most enriched metabolites produced by these genes were discovered to be functional in oxidation reduction and electron transport. Overall, this study suggests that life invented modern electron transport chain around 3.3 billion years ago and around 2.8 billion years ago organisms evolved to use the same proteins that are used for producing oxygen also to breathe oxygen.

### 6.3 Study 2: Pediatric IBD study with Athos Boudvaros

In some diseases such as Inflammatory Bowel Disease (IBD); if the disease is not diagnosed and monitored closely, the results can be very severe, such as the removal of the patient's colon. On the other hand, currently existing most reliable diagnosis methods are very invasive (e.g. colonoscopy). An alternative approach for diagnosis can be abundance analysis of the microbial sample taken from the patients' colon. This study aims to predict the disease state of the subject from bacterial abundances in stool samples taken from the patient.

105 samples were collected for this study among the patients of Dr. Athos Boudvaros; some of them

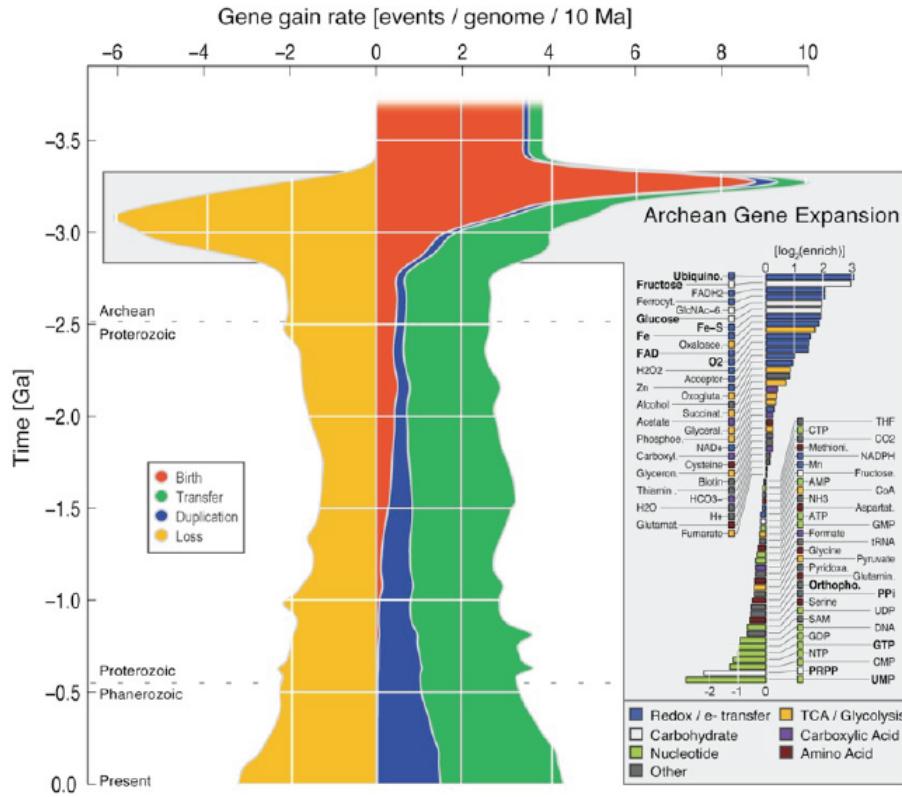


Figure 6.2: Rates of new gene birth, duplication, loss and horizontal gene transfer during Archean Gene Expansion

displaying IBD symptoms and others different diseases (control group). In Figure 6.3, each row block represents a set of bacterial groups at a taxonomic level (phylum level at the top and genus level at the bottom) and each column block represents a different patient group: control patients, Crohn's disease (CD), and ulcerative colitis (UC). The only significant single biomarker was *E. Coli*, which is not seen in control and CD patients but seen in about a third of the UC patients. There seems to be no other single bacterial group that gives significant classification between the patient groups from these abundance measures.

Since *E. Coli* abundance is not a clear-cut single bacterial biomarker, using it as a diagnostic tool would yield low accuracy classification. On the other hand, we can take the entire bacterial group abundance distribution and feed them into a random forest and estimate cross-validation accuracy. After the classification method was employed, it was able to tell with 90% accuracy if the patient is diseased or not. This suggests that it is a competitive method with respect to other non-invasive diagnostic approaches which are generally highly specific but not sensitive enough.

One key difference between control and disease groups is the decrease in the diversity of the ecosystem. This suggests that the disease status is not controlled by a single germ but the overall robustness and the resilience of the ecosystem. When diversity in the ecosystem decreases, the patient might start showing disease symptoms.

## 6.4 Study 3: Human Gut Ecology (HuGE) project

This study aims to identify more than three hundred dietary and environmental factors affecting human microbiome. The factors, which were regularly tracked by an iPhone App, were the food the subject ate, how much they slept, the mood they were in etc. Moreover, stool samples were taken from the subjects every day for a year in order to perform sequence analysis of the bacterial group abundances for a specific day

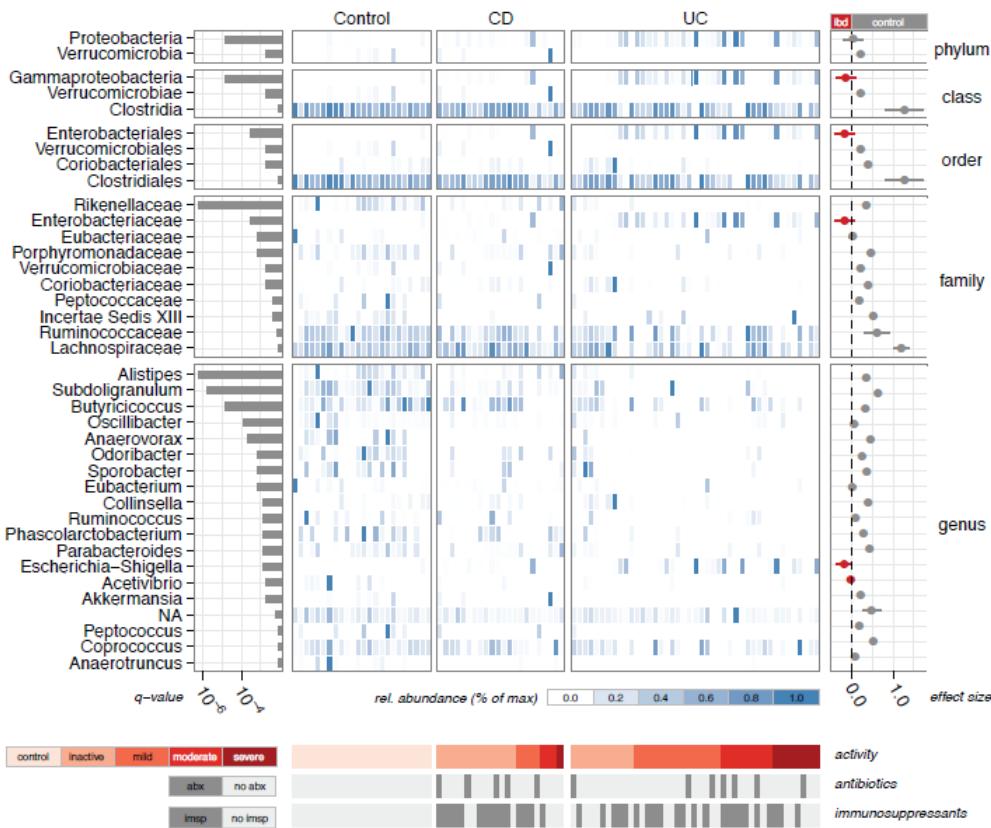


Figure 6.3: Abundance levels of different bacterial groups in control patients, Crohn's disease patients and patients with ulcerative colitis.

relevant to a particular environmental factor. The motivation behind carrying out this study is that, it is usually very hard to get a strong signal between bacterial abundances and disease status. Exploring dietary effects on human microbiome might potentially elucidate some of these confounding factors in bacterial abundance analysis. However, this study analyzed dietary and environmental factors on only two subjects' gut ecosystems; inferring statistically significant correlations with environmental factors would require large cohorts of subjects.

Figure 6.4 shows abundance levels of different bacterial groups in the gut of the two donors throughout the experiment. One key point to notice is that within an individual, the bacterial abundance is very similar through time. However, bacterial group abundances in the gut significantly differ from person to person.

One statistically significant dietary factor that was discovered as a predictive marker for bacterial population abundances is fiber consumption. It was inferred that fiber consumption is highly correlated with the abundance of bacterial groups such as Lachnospiraceae, Bifidobacteria, and Ruminococcaceae. In Donor B, 10g increase in fiber consumption increased the overall abundance of these bacterial groups by 11%.

In Figure 6.6 and Figure 6.7, a horizon plot of the two donors B and A are displayed respectively. A legend to read these horizon plots is given in Figure 6.5. For each bacterial group the abundance-time graph is displayed with different colors for different abundance layers, segments of different layers are collapsed into the height of a single layer displaying only the color with the highest absolute value difference from the normal abundance, and finally the negative peaks are switched to positive peaks preserving their original color.

In Figure 6.6, we see that during the donor's trip to Thailand, there is a significant change in his gut bacterial ecosystem. A large number of bacterial groups disappear (shown on the lower half of the horizon

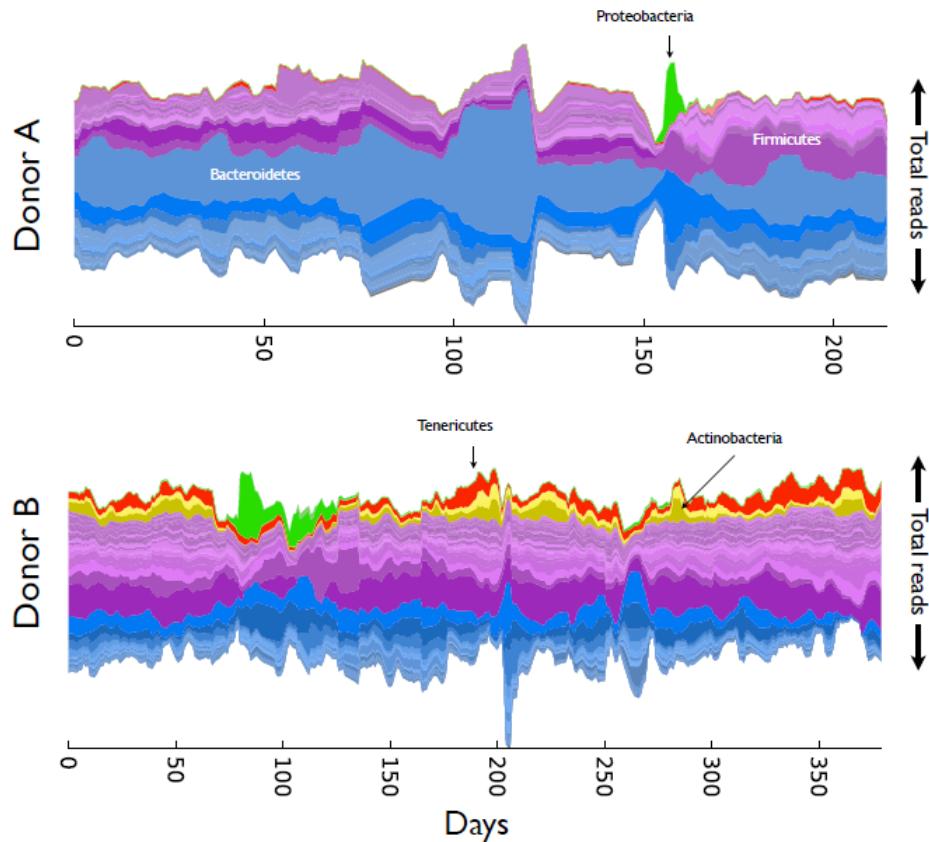


Figure 6.4: Gut bacterial abundances plotted through time for the two donors participating in HuGE project.

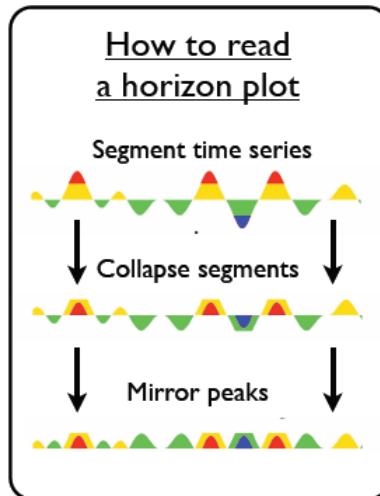


Figure 6.5: Description of how to read a horizon plot.

plot) as soon as the donor starts living in Thailand. And as soon as the donor returns to U.S., the abundance levels of these bacterial groups quickly return back to their normal levels. Moreover, some bacterial groups that are normally considered to be pathogens (first 8 groups shown on top) appears in the donor's ecosystem almost as soon as the donor moves to Thailand and mostly disappears when he returns back to United States.

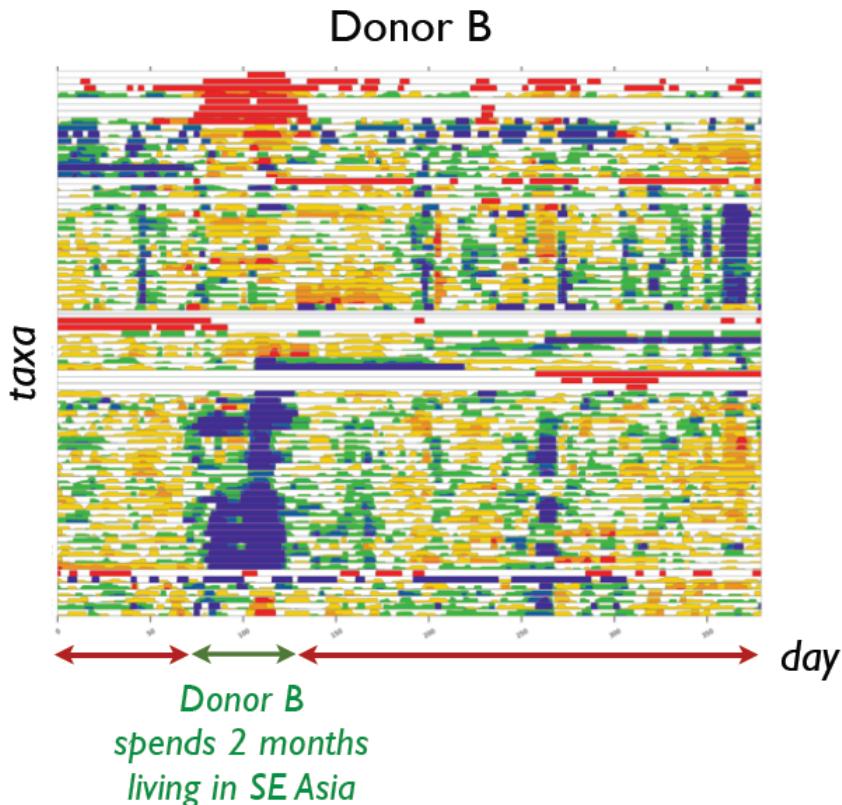


Figure 6.6: Horizon plot of Donor B in HuGE study.

This indicates that environmental factors (such as location) can cause major changes in our gut ecosystem while the environmental factor is present but can disappear after the factor is removed.

In Figure 6.7, we see that after the donor is infected with salmonella, a significant portion of his gut ecosystem is replaced by other bacterial groups. A large number of bacterial groups permanently disappear during the infection and other bacterial groups replace their ecological niches. In other words, the introduction of a new environmental factor takes the bacterial ecosystem in the donor's gut from one equilibrium point to a completely different one. Even though the bacterial population mostly consists of salmonella during the infection, before and after the infection the bacterial count stays more or less the same. The scenario that happened here is that salmonella drove some bacterial groups to extinction in the gut and similar bacterial groups took over their empty ecological niches.

In Figure 6.8, p-values are displayed for day-to-day bacterial abundance correlation levels for Donor A and B. In Donor A's correlation matrix, there is high correlation within the time interval *a* corresponding to pre-infection and within the time interval *b* corresponding to post-infection. However, between *a* and *b* there is almost no correlation at all. On the other hand, in the correlation matrix of donor B, we see that pre-Thailand and post-Thailand time intervals, *c*, have high correlation within and between themselves. However, the interval *d* that correspond to the time period of Donor B's trip to Thailand, we see relatively little correlation to *c*. This suggests that the perturbations in the bacterial ecosystem of Donor B wasn't enough to cause a permanent shift of the abundance equilibrium as in the case with Donor A due to salmonella infection.

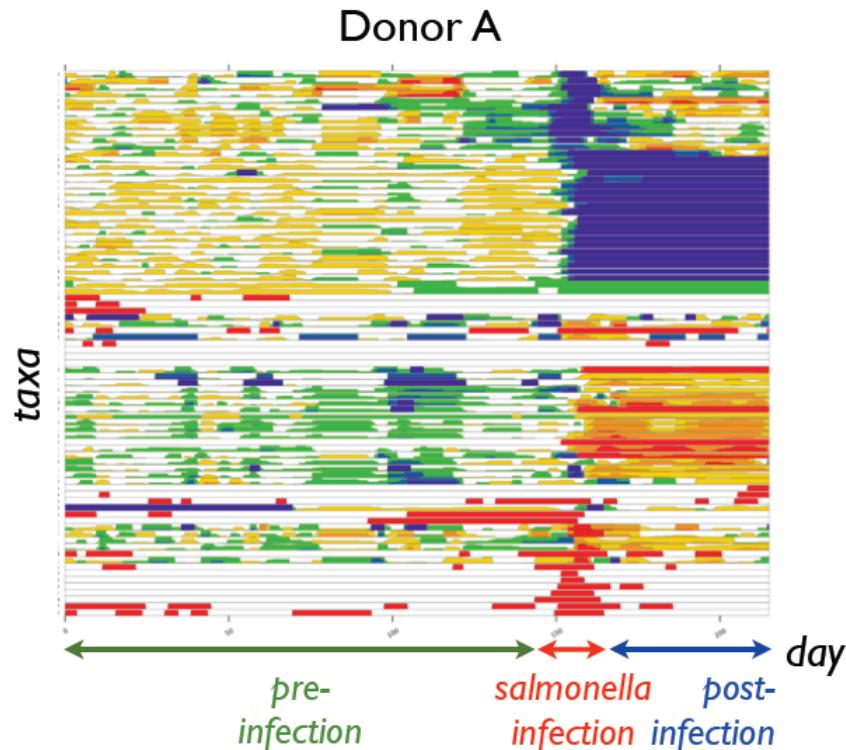


Figure 6.7: Horizon plot of Donor A in HuGE study.

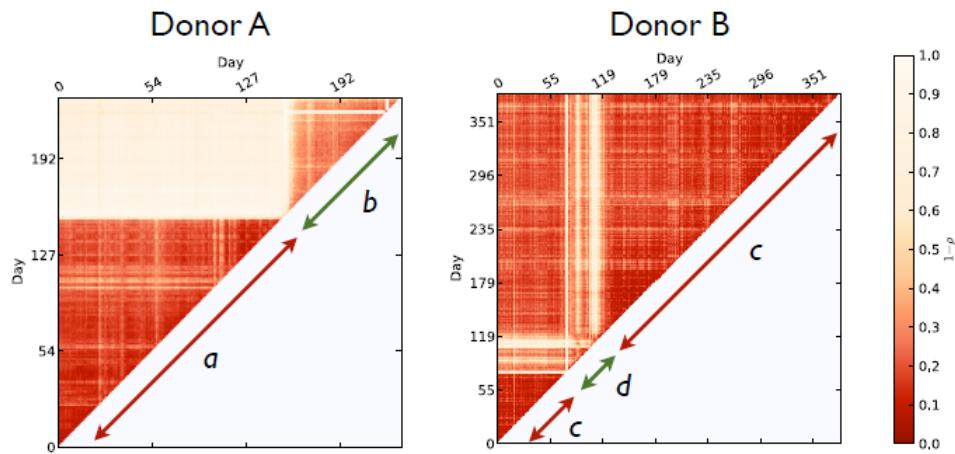


Figure 6.8: Day-to-day bacterial abundance correlation matrices of Donor A and Donor B.

## 6.5 Study 4: Microbiome as the connection between diet and phenotype

In a study by Mozaffarian et al. [4] more than a hundred thousand patients were analyzed with the goal of discovering the effect of diet and lifestyle choices on long-term weight gain and obesity. This study built a model to predict the patients' weights based on the types and amounts of food they consumed over a certain period of time. They found out that fast-food type of food (processed meats, potato chips, sugar-sweetened

beverages) were most highly correlated with obesity. On the other hand, consumption level of yogurt was inversely correlated with obesity.

Further experiments with mouse and human cohorts showed that, within both control group and fast-food group, increased consumption of yogurt leads to weight loss. In the experiment with mice, some female mice were given *Lactobacillus reuteri* (a group of bacteria found in yogurt) and allowed to eat as much regular food or fast-food they wanted to. This resulted in significant weight loss in the group of mice that were given the purified bacterial extract.

An unexpected phenotypical effect of organic yogurt consumption was discovered to be shinier coat of the mice and dogs that were given yogurt as part of their diet. A histological analysis of the skin biopsy of the control and yogurt fed mice proves that the mice that were fed the bacteria in yogurt had hair follicles that are active, leading to active development of healthier and shiny coat and hair.

## 6.6 Study 5: Horizontal Gene Transfer (HGT) between bacterial groups and its effect on antibiotic resistance

A study by Hehemann et al. [3] discovered a specific gene that digests a type of sulfonated carbohydrate that is only found in seaweed sushi wrappers. This gene is found in the gut microbes of Japanese people but not North Americans. The study concluded that this specific gene has transferred at some point in history from the algae itself to the bacteria living on it and then to the gut microbiome of a Japanese person by horizontal gene transfer. This study also suggests that, even though some bacterial group might live in our gut for our entire lives, they can gain new functionalities throughout our lives by picking up new genes depending on the type of food that we eat.

In this direction, a study in Alm's Laboratory investigated around 2000 bacterial genomes published in [1] with the aim of detecting genes that are 100% similar but belong to bacteria in different taxonomic groups. Any gene that is exactly the same between different bacterial groups would indicate a horizontal gene transfer event. In this study, around 100000 such instances were discovered.

When looked at specific environments, it was discovered that the bacteria isolated from humans share genes mostly with other bacteria isolated from human sites. If we focus on more specific sites; we see that bacterial genomes isolated from human gut share genes mostly with other bacteria that are isolated from gut, and bacterial genomes isolated from human skin shared gene mostly with other isolated from human skin. This finding suggests that independent from the phylogeny of the bacterial groups, ecology is the most important factor determining the amount of gene transfer instances between bacterial groups.

In Figure 6.9, we see that between different bacterial groups taken from human that has at least 3% 16S gene distance, there is around 23% chance that they will share an identical gene in their genome. Furthermore, there is more than 40% chance that they share an identical gene if they are sampled from the same site as well.

On the other hand, Figure 6.10 shows that geography is a weak influence on horizontal gene transfer. Bacterial populations sampled from the same continent and different continents had little difference in terms of the amount of horizontal gene transfer detected.

Figure 6.11 shows a color coded matrix of the HGT levels between various human and non-human environments; top-right triangle representing the amount of horizontal gene transfers and the bottom-left triangle showing the percentage of antibiotic resistance (AR) genes among the transferred genes. In the top-right corner, we see that there is a slight excess of HGT instances between human microbiome and bacterial samples taken from farm animals. And when we look at the corresponding percentages of antibiotic resistance genes, we see that more than 60% of the transfers are AR genes. This result shows the direct effect of feeding subtherapeutic antibiotics to livestock on the emergence of antibiotic resistance genes in the bacterial populations living in human gut.

## 6.7 Study 6: Identifying virulence factors in Meningitis

Bacterial meningitis is a disease that is caused by very diverse bacteria that are able to get into the blood stream and cross the blood-brain barrier. This study aimed to investigate the virulence factors that can turn

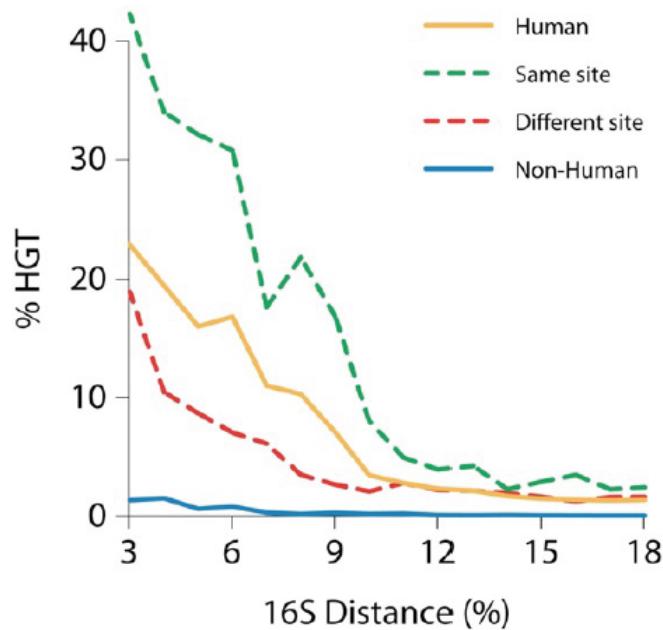


Figure 6.9: Rate of horizontal gene transfer between different bacterial groups taken from non-human sites, human sites, same site within human, and different sites within human.

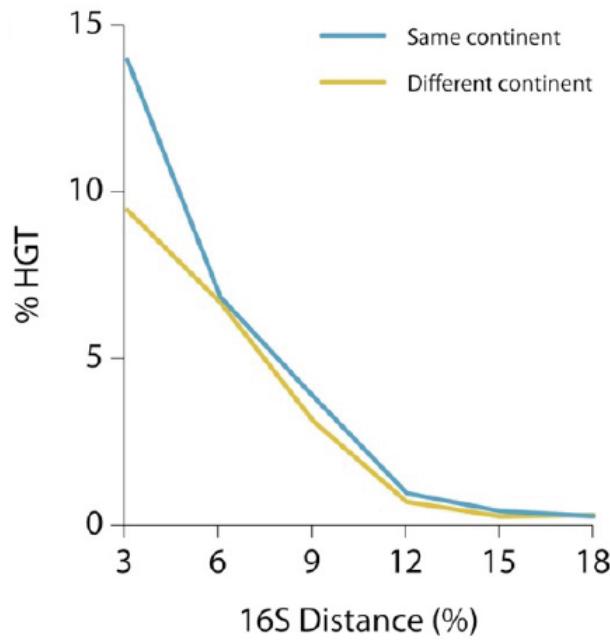


Figure 6.10: Rate of horizontal gene transfer between bacterial groups sampled from the same continent and from different continents.

bacteria into a type that can cause meningitis.

The study involved 70 bacterial strains isolated from meningitis patients, comprising 175172 genes in total. About 24000 of these genes had no known function. There could be some genes among these 24000

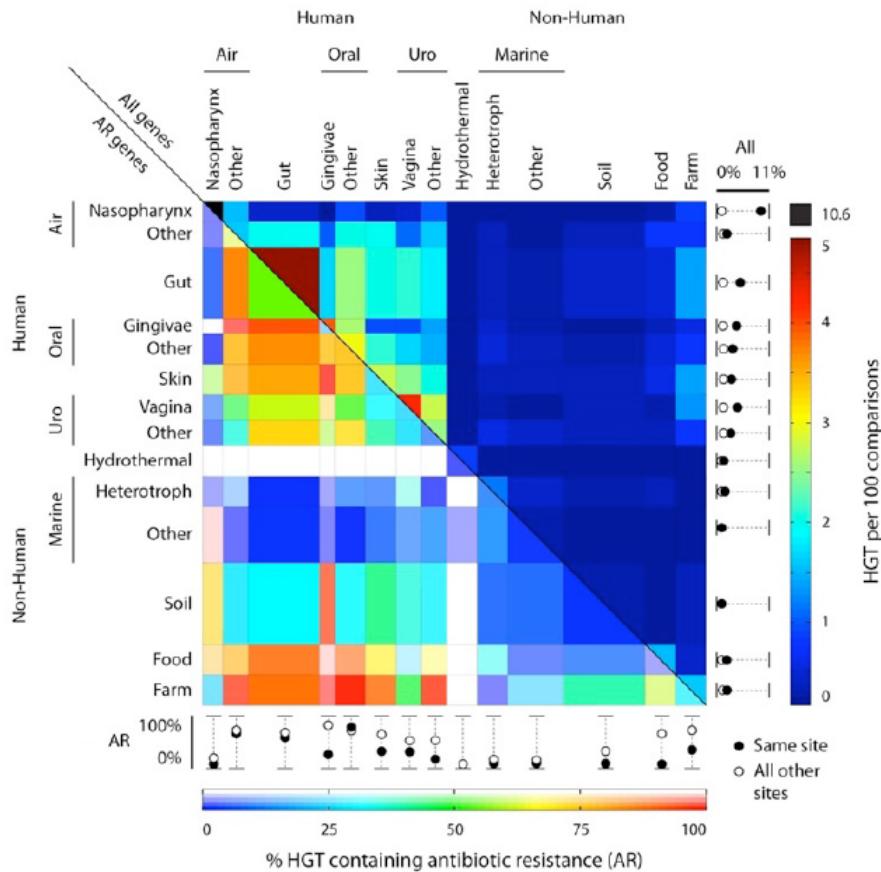


Figure 6.11: Rate of horizontal gene transfer between different human and non-human sites (top right) and the percentage of antibiotic resistance genes among horizontal gene transfers (bottom left).

that might be leading to meningitis causing bacteria and might be good drug targets. Moreover, 82 genes were discovered to be involved in horizontal gene transfer. 69 of these had known functions and 13 of them belonged to the 24000 genes that we do not have any functional information. Among the genes with known function, some of them were related to AR, detoxification, and also some were related to known virulence factors such as hemolysin that lets the bacteria live in the blood stream and adhesin that helps the bacteria latch onto the vein and potentially cross blood brain barrier.

## 6.8 Q/A

Q: Do you think after some time Donor A in Study 3 will have its bacterial ecosystem return back to its original pre-infection state?

A: The salmonella infection caused certain niches to be wiped out from the bacterial ecosystem of Donor A which were then filled in by similar type of bacteria and reached to a different ecosystem at a new equilibrium. Since these niches are dominated by the new groups of bacteria, it would not be possible for the previous bacterial groups to replace them without a large-scale change in his gut ecosystem.

Q: Is the death of certain bacterial groups in the gut during salmonella infection caused directly by the infection or is it an immune response to cure the disease?

A: It can be both, but it is very hard to tell from the data in Study 3 since it is only a data point that corresponds to the event that we can observe. A future study that tries to figure out what is happening in our immune system during the infection can be observed by drawing blood from the patients during the infection.

Q: Is there a particular connection between an individual's genome and the dominant bacterial groups in the bacterial ecosystem? Would twins show more similar bacterial ecosystems?

A: Twins in general have similar bacterial ecosystems independent from whether they live together or are separated. Even though this seems to be a genetic factor at first, monozygotic and dizygotic twins have the exact same effect, as well as displaying similarity to their mothers' bacterial ecosystem. The reason for this is that starting from birth there is a period of time in which the bacterial ecosystem is programmed. The similarity effect between twins is based on this more than genetic factors.

## 6.9 Current research directions

A further extension to HuGE study could observe mice gut microbiome during a salmonella infection and observe the process of some bacterial groups being driven to extinction and other types of bacteria replacing the ecological niches that are emptied by them. A higher resolution observation of this phenomenon in mice could illuminate how bacterial ecosystems shift from one equilibrium to another.

## 6.10 Further Reading

- Overview of Human Microbiome Project: <http://commonfund.nih.gov/hmp/overview.aspx>
- Lawrence A. David and Eric J. Alm. (2011). Rapid evolutionary innovation during an Archaean genetic expansion. *Nature*, 469(7328):93-96.
- A tutorial on 16S rRNA gene and its use in microbiome research: [http://greengenes.lbl.gov/cgi-bin/JD\\_Tutorial/nph-Tutorial\\_2Main2.cgi](http://greengenes.lbl.gov/cgi-bin/JD_Tutorial/nph-Tutorial_2Main2.cgi)
- Dariush Mozaffarian, Tao Hao, Eric B. Rimm, Walter C. Willett, and Frank B. Hu. (2011). Changes in diet and lifestyle and long-term weight gain in women and men. *The New England journal of medicine*, 364(25):2392-2404.
- JH Hehemann, G Correc, T Barbevron, W Helbert, M Czjzek, and G Michel. (2010). Transfer of carbohydrate- active enzymes from marine bacteria to japanese gut microbiota. *Nature*, 464(5):908-12.
- The Human Microbiome Jumpstart Reference Strains Consortium. (2010). A Catalog of Reference Genomes from the Human Microbiome. *Science*, 328(5981):994-999

## 6.11 Tools and techniques

## 6.12 What have we learned?

In this lecture, we learned about the field of bacterial genomics in general and how bacterial ecosystems can be used to verify major environmental changes at early stages of evolution (Study 1), can act as a noninvasive diagnostic tool (Study 2), are temporarily or permanently affected by different environmental and dietary factors (Study 3), can act as the link between diet and phenotype (Study 4), can cause antibiotic resistance genes to be carried between different species' microbiome through horizontal gene transfer (Study 5), and can be used to identify significant virulence factors in disease states (Study 6).

## Bibliography

- [1] The Human Microbiome Jumpstart Reference Strains Consortium. A Catalog of Reference Genomes from the Human Microbiome. *Science*, 328(5981):994–999, May 2010.
- [2] Lawrence A. David and Eric J. Alm. Rapid evolutionary innovation during an Archaean genetic expansion. *Nature*, 469(7328):93–96, January 2011.

- [3] JH Hehemann, G Correc, T Barbeyron, W Helbert, M Czjzek, and G Michel. Transfer of carbohydrate-active enzymes from marine bacteria to japanese gut microbiota. *Nature*, 464(5):908–12, 2010 Apr 8.
- [4] Dariush Mozaffarian, Tao Hao, Eric B. Rimm, Walter C. Willett, and Frank B. Hu. Changes in diet and lifestyle and long-term weight gain in women and men. *The New England journal of medicine*, 364(25):2392–2404, June 2011.



## **Part II**

### **Coding and Non-Coding Genes**



## HIDDEN MARKOV MODELS I

Ruth Park (September 24, 2016)  
Anastasiya Belyaeva, Justin Gullingsrud (Sep 26, 2015)  
William Leiserson, Adam Sealfon (Sep 22, 2014)  
Haoyang Zeng (Sep 28, 2013)  
Sumaiya Nazeen (Sep 25, 2012)  
Chrisantha Perera (Sep 27, 2011)  
Gleb Kuznetsov, Sheida Nabavi (Sep 28, 2010)  
Elham Azizi (Sep 29, 2009)

### Figures

---

7.1	Modeling biological sequences with HMMs (aka: what to do with big unlabelled chunks of DNA) . . . . .	143
7.2	Model Schematics of a Markov Chain (left) and HMM(right) . . . . .	144
7.3	Parameterization of HMM Prediction Model . . . . .	144
7.4	State of the casino die represented by a Hidden Markov model . . . . .	146
7.5	Potential DNA sources: viral injection vs. normal production . . . . .	146
7.6	A possible sequence of observed die rolls. . . . .	146
7.7	Modeling the probability of the sequence given that the path consists of only fair dice . . . . .	147
7.8	Modeling the probability of the sequence given that the path consists of only loaded dice . . . . .	147
7.9	Partial runs and die switching . . . . .	148
7.10	HMMs as a generative model for finding GC-rich regions. . . . .	149
7.11	Probability of sequence if all promoter . . . . .	150
7.12	Probability of sequence if all background . . . . .	150
7.13	Probability of sequence if mixed . . . . .	151
7.14	Some biological applications of HMM . . . . .	151
7.15	The six algorithmic settings for HMMs. 1: path known. 2: path unknown, so search through all paths to find most likely path. $\pi$ = path of hidden states, $x$ = sequence, $k$ = state, $i$ = position . . . . .	152
7.16	The Viterbi algorithm . . . . .	154
7.17	The Forward Algorithm . . . . .	155
7.18	CpG Islands - Incorporating Memory . . . . .	156

---

### 7.1 Introduction

Hidden Markov Models (HMMs) are a fundamental tool from machine learning that is widely used in computational biology. Using HMMs, we can explore the underlying structure of DNA or polypeptide

sequences, detecting regions of special interest. For instance, we can identify conserved subsequences or uncover regions with statistically different nucleotide or amino acid distributions, such as promoter regions or CpG islands. This probabilistic model can be used to illuminate the properties and structural components of sequences, allowing us locate genes and other functional elements.

This is the first of two lectures on HMMs. In this lecture we will define Markov Chains and HMMs, and provide a series of motivating examples. In the second half of this lecture, we will discuss the scoring and decoding of HMMs. We will learn how to compute the probability of a particular combination of observations and states. We will introduce the Forward Algorithm for computing the probability of a given sequence of observations across all possible sequences of states. Finally, we will discuss the Viterbi algorithm, which provides a method of determining the most likely path of states that resulted in the given observations.

In the second lecture on HMMs, we will continue our discussion of decoding by exploring posterior decoding, which allows us to compute the most likely state at each point in the sequence. We will then explore how to learn a Hidden Markov Model. We cover both supervised and unsupervised learning, explaining how each can be used to learn the model's parameters. In supervised learning, we have training data available that labels sequences with particular models. In unsupervised learning, we do not have labels so we must seek to partition the data into discrete categories based on discovered probabilistic similarities. In our discussion of unsupervised learning we will introduce the widely applicable and generalizable Expectation Maximization (EM) algorithm.

## 7.2 Motivation:

### 7.2.1 We have a new sequence of DNA, now what?

#### 1. Align it:

- to things we know about (database search).
- to unknown things (assemble/clustering) For example piece together entire genome by matching and assembling specific genes.

#### 2. Visualize it: “Genomics rule #1”: *Look at your data!*

- Look for nonstandard nucleotide compositions (e.g. promoter regions, CpG islands).
- Look for k-mer frequencies that are associated with protein coding regions, recurrent data, high GC content, etc.
- Look for motifs, evolutionary signatures.
- Translate and look for open reading frames, stop codons, etc.
- Look for patterns, then develop machine learning tools to determine reasonable probabilistic models. For example by looking at a number of quadruples we decide to color code them to see where they most frequently occur.

#### 3. Model it:

- Make hypotheses.
- Build a generative model to describe the hypotheses. For example, exploring what makes a human “human” (eyes, nose, mouth) and building a model to generate a new human based on the features determined to be human characteristics.
- Use that model to find sequences of similar **type**.

We're not looking for sequences that necessarily have common ancestors. Rather, we're interested in sequences with similar properties. We actually don't know how to model whole genomes, but we can model small aspects of genomes. The task requires an understanding of all the properties of genome regions and computationally building generative models to represent hypotheses. For a given sequence, we want to annotate regions such as introns, exons, intergenic, promoter, or otherwise classifiable regions.

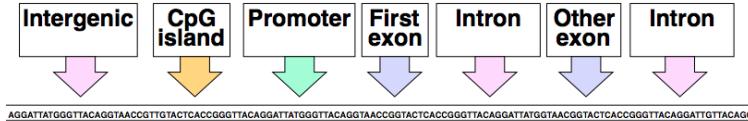


Figure 7.1: Modeling biological sequences with HMMs (aka: what to do with big unlabelled chunks of DNA)

Building this framework will give us the ability to:

- Emit (generate) sequences of similar type according to the generative model.
- Recognize the hidden state that most likely generated the observation
- Learn (train) large datasets and apply to previously labeled data (supervised learning) and/or unlabeled data (unsupervised learning). We can then classify future data according to the learned model parameters.

In this lecture we discuss algorithms for emission and recognition.

### 7.2.2 Why probabilistic sequence modeling?

- Biological data is noisy, and we want to capture the variability in the noise.
- It can be used to update previous knowledge about biological sequences.
- Probability provides a calculus for manipulating models, because we can use it to infer hidden states and generate data from model.
- We are not limited to yes/no answers, so probabilistic models can provide degrees of belief. (e.g. "there is a 75% chance this is a promoter region" instead of "this is/is not a promoter region")

For these reasons, many common computational tools are based on probabilistic models. For our purposes, these tools consist of Markov Chains and HMMs.

## 7.3 Markov Chains and HMMs

### 7.3.1 Motivating Example: Weather Prediction

Weather prediction has always been difficult, especially when we would like to forecast the weather many days, weeks, or months in advance. However, if we only need to predict the weather of the next day, we can achieve decent precision using simple models such as Markov Chains and Hidden Markov Models, as shown by the graphical models shown in Figure 7.2.

On the left is a Markov Chain model that consists of four types of weather (Sun, Rain, Clouds, and Snow) that can directly transition between each other. In this case, "what you see is what you get" in that the next state only depends on the current state; there is no memory of any previous state. On the right is a HMM, in which the types of weather are modeled as emissions (or outcomes) of unknown, or "hidden", seasons (Summer, Fall, Winter, and Spring). The key insight behind the HMM is that the hidden states of the model (e.g. seasons) determine the emission probabilities, while the transitions between hidden states are governed by a Markov Chain.

### 7.3.2 Formalizing the Markov Chain and HMM

To fully understand Hidden Markov Models, we must define the key parameters, as shown in Figure 7.3. The vector  $x$  represents the sequence of observations (emissions). The vector  $\pi$  represents the hidden path, which is the sequence of hidden states. Each entry  $a_{kl}$  of the transition matrix  $A$  denotes the probability of transition

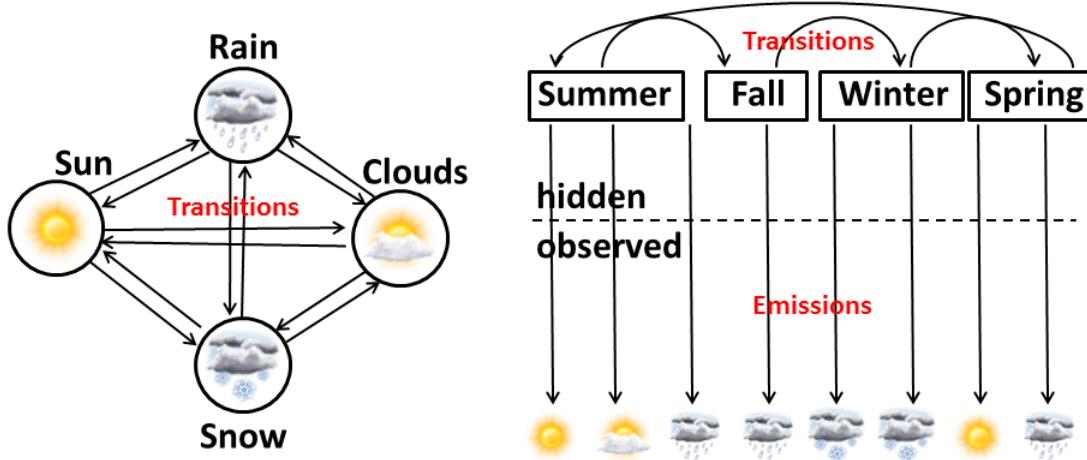


Figure 7.2: Model Schematics of a Markov Chain (left) and HMM(right)

from state  $k$  to state  $l$ . Each entry  $e_k(x_i)$  of the emission vector denotes the probability of observing  $x_i$  from state  $k$ . Finally, with these parameters and Bayes' rule, we can use the known  $p(x_i|\pi_i = k)$  to estimate the unknown probability  $p(\pi_i = k|x_i)$ .

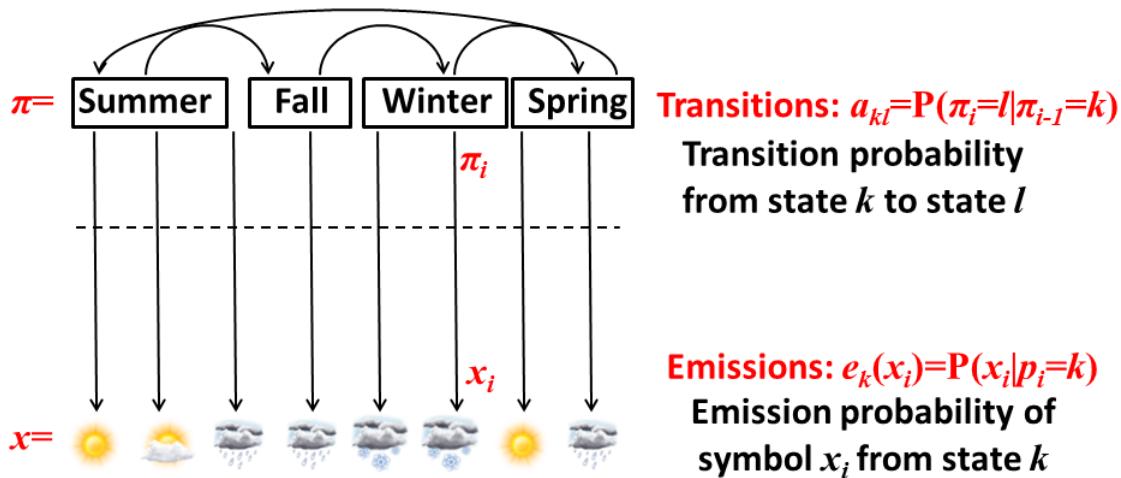


Figure 7.3: Parameterization of HMM Prediction Model

### Markov Chains

A Markov Chain is defined by a finite set of states and the transition probabilities between these states. At every time step, the Markov Chain is in a particular state and undergoes a transition to another state. The probability of transitioning to each other state depends only on the current state, independent of how the current state was reached. More formally, a Markov Chain is a triplet  $(Q, p, A)$  which consists of:

- A set of states  $Q$ .
- A transition matrix  $A$  whose elements  $A_{ij}$  correspond to the probability of transitioning from state  $i$  to state  $j$ .

- A vector  $p$  of initial state probabilities.

The key property of Markov Chains is that they are memoryless, i.e., each state depends only on the previous state. So we can immediately define the probability of the next state given the current state:

$$P(x_i|x_{i-1}, \dots, x_1) = P(x_i|x_{i-1})$$

In this way, the probability of the entire sequence can be decomposed as follows:

$$P(x) = P(x_L, x_{L-1}, \dots, x_1) = P(x_L|x_{L-1})P(x_{L-1}|x_{L-2})\dots P(x_2|x_1)P(x_1)$$

$P(x_L)$  can also be calculated from the transition probabilities. If we multiply the initial state probabilities at time  $t = 0$  by the transition matrix  $A$ , we get the probabilities of states at time  $t = 1$ . Therefore, multiplying them by the transition matrix taken to the appropriate power  $A^L$  provides the state probabilities at time  $t = L$ .

### Hidden Markov Models

Hidden Markov Models are used to represent a problem space in which observations come about as a result of states of the system that we are unable to observe directly. These observations, or emissions, result from the hidden states based on a set of probabilities associated with those states. These probabilities are known as emission probabilities.

Formally, a Hidden Markov Model is a 5-tuple  $(Q, A, p, V, E)$  which consists of the following parameters:

- A series of states,  $Q$ .
- A transition matrix,  $A$
- A vector of initial state probabilities ,  $p$ .
- A set of observation symbols,  $V$ , for example {A, T, C, G} or the set of amino acids or words in an English dictionary.
- A matrix of emission probabilities,  $E$ : For each  $s, t$ , in  $Q$ , the emission probability is

$$e_{sk} = P(v_k \text{ at time } t | q_t = s)$$

The key property of memorylessness is inherited from Markov Models, so that the emissions and transitions depend only on the current state and not on the past history.

## 7.4 Applications of HMMs: From the Casino to Biology

### 7.4.1 The Dishonest Casino

#### The Scenario

Imagine the following scenario: You enter a casino that offers a dice-rolling game. You bet \$1 and you and a dealer both roll a die. If you roll a higher number you win \$2. Now there's a twist to this seemingly simple game. You are aware that the casino has two types of dice:

1. Fair die:  $P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = 1/6$
2. Loaded die:  $P(1) = P(2) = P(3) = P(4) = P(5) = 1/10$  and  $P(6) = 1/2$

The dealer can switch between these two dice at any time without you knowing it; the only information that you have are the rolls that you observe. On average, the dealer switches the dice every 20 turns. We can represent the state of the casino die with a simple Markov model:

The model shows the two possible states, their emissions, and the transition and emission probabilities.

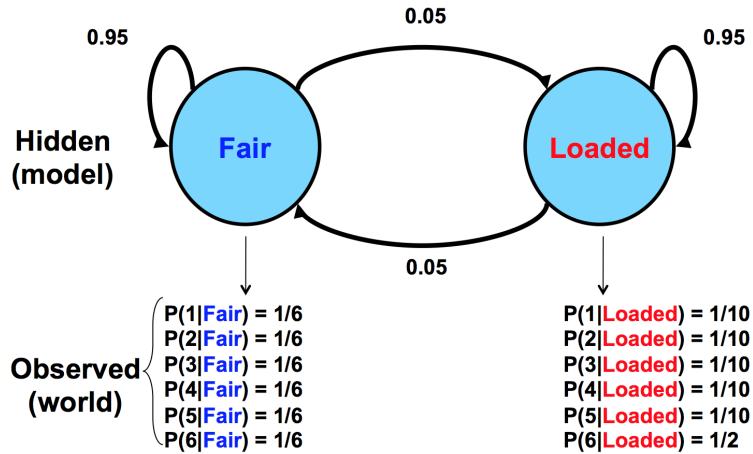


Figure 7.4: State of the casino die represented by a Hidden Markov model

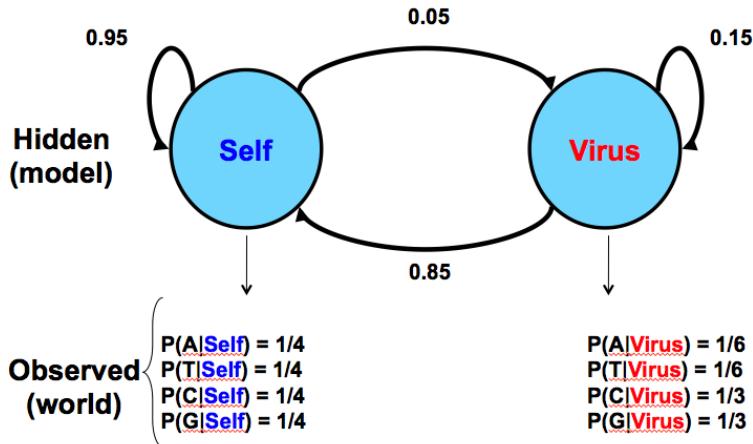


Figure 7.5: Potential DNA sources: viral injection vs. normal production

### A biological analogy: the dishonest genome

Figure 7.5 below gives a similar model for a biological situation in which a sequence of DNA has two potential sources: injection by a virus or normal production by the organism itself.

This model assumes that viral inserts will have higher CpG prevalence, which leads to higher probabilities of C and G emission. In addition, it is assumed *a priori* that the sequence is very likely to result from its own genome, and that transitions to viral DNA are unlikely, resulting in the transition probabilities shown in the model. Given this model as a hypothesis, we would observe the frequencies of C and G to give us clues as to the source of the sequence in question.

### Running the Model

Say we are at the casino and observe the sequence of rolls given in Figure 7.6. We would like to know whether it is more likely that the casino is using the fair die or the loaded die.



Figure 7.6: A possible sequence of observed die rolls.

We will consider two possible sequences of states to better understand the behavior of the underlying HMM: one in which the dealer is always using a fair die, and the other in which the dealer is always using a loaded die. For each case, we want to compute the joint probability of the observed outcome given that sequence of hidden states.

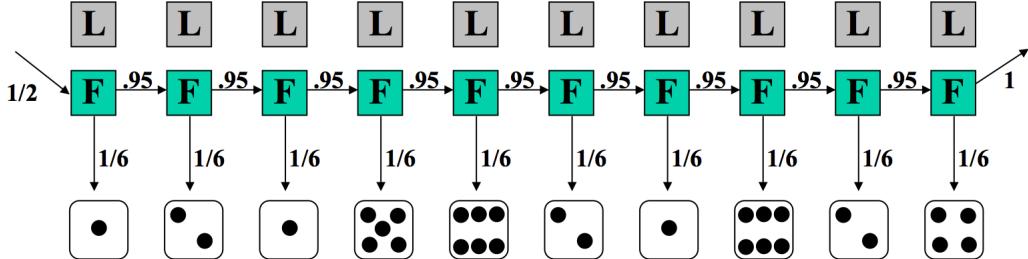


Figure 7.7: Modeling the probability of the sequence given that the path consists of only fair dice

In the case of the dealer using only the fair die, the transition and emission probabilities are shown in Figure 7.7. We assume  $\pi = \{F, F, F, F, F, F, F, F, F, F\}$ , and we observe  $x = \{1, 2, 1, 5, 6, 2, 1, 6, 2, 4\}$ . The probability of this sequence of states producing the observed emissions is a product of three terms:  $1/2$ , the probability of starting with the fair die;  $(1/6)^{10}$ , the probability of the sequence being produced by the fair die; and  $(0.95)^9$ , the probability that we always continue to use the fair die.

Therefore:

$$\begin{aligned}
 P(x, \pi) &= P(x|\pi)P(\pi) \\
 &= \frac{1}{2} \times P(1|F) \times P(F|F) \times P(2|F) \dots \\
 &= \frac{1}{2} \times \left(\frac{1}{6}\right)^{10} \times (0.95)^9 \\
 &= 5.2 \times 10^{-9}
 \end{aligned}$$

Since the probability is so small, this might appear to be an extremely unlikely case. In actuality, the probability is low simply because there are a large number of possible sequences, so no one outcome is *a priori* likely. The question is not whether this sequence of hidden states is absolutely likely, but whether it is *more* likely than the alternatives. This can be easily determined by looking at ratios rather than absolute probabilities.

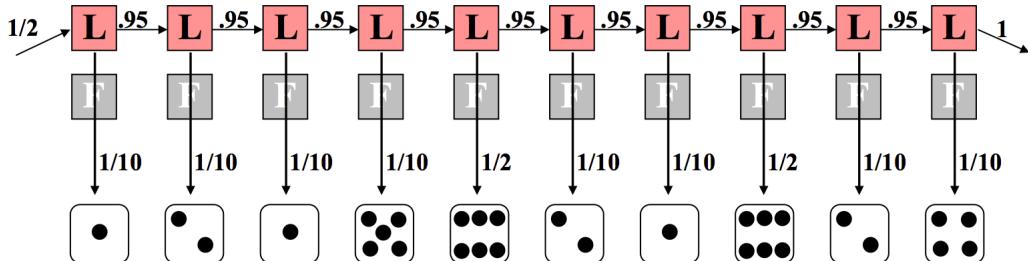


Figure 7.8: Modeling the probability of the sequence given that the path consists of only loaded dice

Let us consider the opposite extreme where the dealer always uses the loaded die, as depicted in Figure 7.8. The calculation here is similar, except for a difference in the emission probability term. This time, 8 of the 10 rolls carry a probability of  $1/10$  because the loaded die favors sixes. The two rolls of six have each a probability of  $1/2$  of occurring. Again we multiply all of these probabilities together according to principles of independence and conditioning, as follows:

$$\begin{aligned}
P(x, \pi) &= \frac{1}{2} \times P(1|L) \times P(L|L) \times P(2|L) \cdots \\
&= \frac{1}{2} \times \left(\frac{1}{10}\right)^8 \times \left(\frac{1}{2}\right)^2 \times (0.95)^9 \\
&= 7.9 \times 10^{-10}
\end{aligned}$$

Note the difference in exponents: the situation in which only a fair die is used is  $52 \times 10^{-10}$ , compared with  $7.9 \times 10^{-10}$  for the loaded die. Therefore, it is six times more likely that only the fair die was used than it is that only the loaded die was used. This is not too surprising—two rolls out of ten yielding a 6 is not very far from the expected number of 1.7 for the fair die, but much further from the expected number of 5 for the loaded die.

### Adding Complexity

Now imagine the more complex, and interesting, case where the dealer switches the die at some point during the sequence. Figure 7.9 represents a guess at the underlying model.

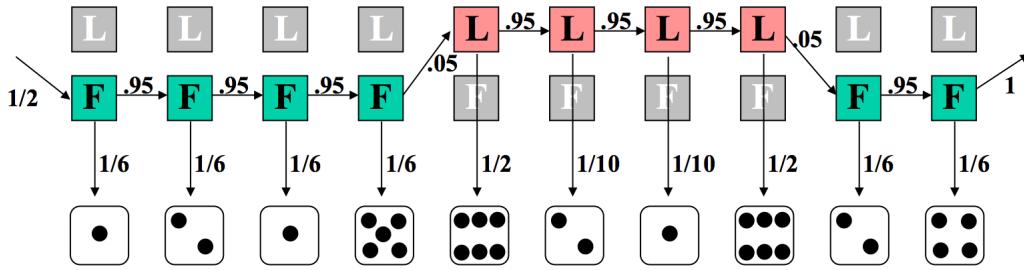


Figure 7.9: Partial runs and die switching

Again, we can calculate the likelihood of the joint probability of this sequence of states and observations. Here, six of the rolls are calculated with the fair die, and four with the loaded one. Additionally, not all of the transition probabilities are 95% anymore; the two die swaps each have a probability of 5%.

$$\begin{aligned}
P(x, \pi) &= \frac{1}{2} \times P(1|L) \times P(L|L) \times P(2|L) \cdots \\
&= \frac{1}{2} \times \left(\frac{1}{10}\right)^2 \times \left(\frac{1}{2}\right)^2 \times \left(\frac{1}{6}\right)^6 \times (0.95)^7 \times (0.05)^2 \\
&= 4.67 \times 10^{-11}
\end{aligned}$$

Clearly, our guess is far less likely than either of the previous two cases. But we cannot possibly calculate *all* possible scenarios in this way to find the most likely, so we need new techniques for inferring the underlying model. In the above cases we more-or-less just guessed at the model, but what we want is a way to systematically derive likely models. This must be done using HMM decoding algorithms, which will be described later.

### 7.4.2 Back to Biology

Now that we have formalized HMMs, we want to use them to solve some real biological problems. In fact, HMMs are a great tool for gene sequence analysis, because we can look at a sequence of DNA as being emitted by a mixture of models. These may include introns, exons, transcription factors, etc. While we may have some sample data that matches models to DNA sequences, in the case where we start with a new sequence of DNA, we can use HMMs to ascribe some potential models to the sequence. We will first

introduce and discuss a simple example. Then, we will review some of the interesting biological applications of HMMs, before finally describing the HMM techniques that can be used solve the problems that arise in such a first-attempt/native analysis.

### A simple example: Finding GC-rich regions

In this scenario, we want to find GC-rich regions by modeling nucleotide sequences drawn from two different distributions: background and promoter. Background regions have a uniform distribution of A, T, G, C, with a probability of 0.25 for each. Promoter regions, on the other hand, have non-uniform probabilities (A: 0.15, T: 0.13, G: 0.30, C: 0.42). We can learn the initial state distributions based on steady state probabilities. Given a single nucleotide observation, we cannot say anything about the region from which it originated because either will emit each nucleotide at some probability. However, our goal is to look at a sequence and identify which regions originate from the background distribution (B) and which regions are from the promoter distribution (P).

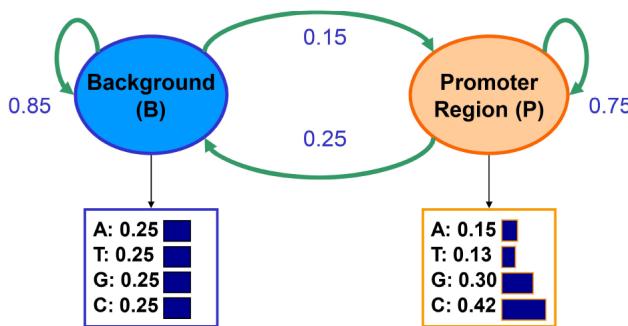


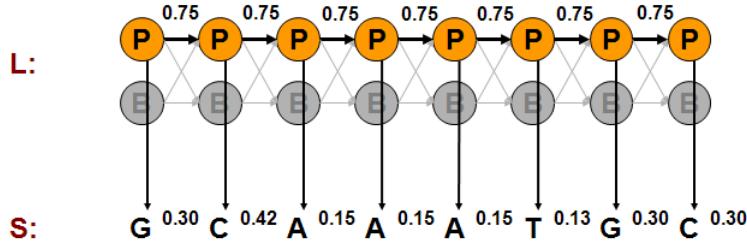
Figure 7.10: HMMs as a generative model for finding GC-rich regions.

The transition and emission probabilities are derived from the relevant distributions and the average length of regions of each state, where  $x$  = vector of observable emissions consisting of {A,T,G,C};  $\pi$  = vector of states in a path (e.g. BPPBP); and  $\pi^*$  = maximum likelihood path. In our interpretation, the maximum likelihood path will be found by dynamic programming.

HMMs are generative models, in that an HMM gives the probability of emission given a state, essentially telling you how likely certain states are to generate the given sequence. Therefore, we can start anywhere and always be able to run a generative model for transitions between states. In an HMM, you can pick an initial state based on the initial probability vector. In the example above, we will start in state B with high probability since most locations do not correspond to promoter regions. An emission is then drawn from the  $P(X|B)$  distribution. The probability distribution of the subsequent state then depends only on the fact that we are in the background state and is independent of the current emission, so that the probability of remaining in state B is 0.85 and the probability of transitioning to state P is 0.15.

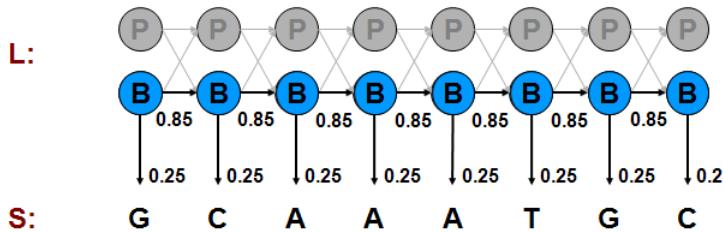
We can compute the probability of one such generation by multiplying the probabilities that the model makes exactly the choices we assumed. Consider the examples shown in Figures 7.11, 7.12, and 7.13, which are analogous to the dice examples given earlier.

$$\begin{aligned}
 P(x, \pi_P) &= a_P \times e_P(G) \times a_{PP} \times e_P(G) \times \dots \\
 &= a_P \times (0.75)^7 \times (0.15)^3 \times (0.13) \times (0.3)^2 \times (0.42)^2 \\
 &= 9.3 \times 10^{-7} \\
 P(x, \pi_B) &= (0.85)^7 \times (0.25)^8 \\
 &= 4.9 \times 10^{-6} \\
 P(x, \pi_{mixed}) &= (0.85)^3 \times (0.25)^6 \times (0.75)^2 \times (0.42)^2 \times 0.3 \times 0.15 \\
 &= 6.7 \times 10^{-7}
 \end{aligned}$$



$$\begin{aligned}
P(x, \pi) &= a_p * e_p(G) * a_{pp} * e_p(G) * a_{pp} * e_p(C) * a_{pp} * e_p(A) * a_{pp} * \dots \\
&= a_p * (0.75)^7 * (0.15)^3 * (0.13)^1 * (0.30)^2 * (0.42)^2 \\
&= 9.3 * 10^{-7}
\end{aligned}$$

Figure 7.11: Probability of sequence if all promoter



$$\begin{aligned}
P &= P(G | B)P(B_1 | B_0)P(C | B)P(B_2 | B_1)P(A | B)P(B_3 | B_2) \dots P(C | B_7) \\
&= (0.85)^7 * (0.25)^8 \\
&= 4.9 * 10^{-6}
\end{aligned}$$

A: 0.25	[blue square]
T: 0.25	[blue square]
G: 0.25	[blue square]
C: 0.25	[blue square]

Figure 7.12: Probability of sequence if all background

The most likely option of these three was that in which all emissions were drawn from the background distribution. But how do we find the most likely of all possible paths of states?

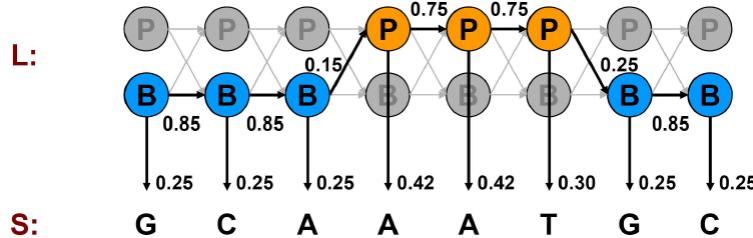
The brute force approach is to calculate the probability of all possibilities as above. However, there are an exponential number of possible paths so this solution is computationally infeasible. In the following sections, we will develop algorithms to solve this problem.

### Other Applications of HMMs in Biology

In addition to the one described above, HMMs can be used to answer many biological questions, some of which are summarized in Figure 7.14.

## 7.5 Algorithmic Settings for HMMs

We use HMMs for three types of operations: **scoring**, **decoding**, and **learning**. We will talk about scoring and decoding in this lecture. Each of these operations can be carried out for a single path or over all possible paths. The single path operations are focused on discovering the path with maximum probability,



$$\begin{aligned}
 P &= P(G|B)P(B_1|B_0)P(C|B)P(B_2|B_1)P(A|B)P(P_3|B_2)\dots P(C|B_7) \\
 &= (0.85)^3 \times (0.25)^6 \times (0.75)^2 \times (0.42)^2 \times 0.30 \times 0.15 \\
 &= 6.7 \times 10^{-7}
 \end{aligned}$$

Figure 7.13: Probability of sequence if mixed

Application	Detection of GC-rich region	Detection of Conserved region	Detection of Protein coding exons	Detection of Protein coding conservation	Detection of Protein coding gene structures	Detection of chromatin states
Topology / Transitions	2 states, different nucleotide composition	2 states, different conservation levels	2 states, different tri-nucleotide composition	2 states, different evolutionary signatures	~20 states, different composition / conservation, specific structure	40 states, different chromatin mark combinations
Hidden States / Annotation	GC-rich / AT-rich	Conserved / non-Conserved	Coding (exon) / non-Coding (intron or intergenic)	Coding (exon) / non-Coding (intron or intergenic)	First / last / middle coding exon, UTRs, intron 1/2/3, intergenic, *(+,-) strand	Enhancer / Promoter / Transcribed / Repressed / Repetitive
Emissions / Observations	Nucleotides	Level of conservation	Triplets of nucleotides	64 x 64 matrix of codon substitution frequencies	Codons, nucleotides, splice sites, start/stop codons	Vector of chromatin mark frequencies

Figure 7.14: Some biological applications of HMM

while for all paths we are interested in the sequence of observations or emissions regardless of the specific corresponding paths.

### 7.5.1 Scoring

#### Scoring over a single path

The Dishonest Casino and the prediction of GC-rich regions problems described in section 7.4 are both examples of finding the probability of a single path. For this case we define the scoring problem as follows:

- **Input:** A sequence of observations  $x = x_1 x_2 \dots x_n$  generated by an HMM  $M(Q, A, p, V, E)$  and a path of states  $\pi = \pi_1 \pi_2 \dots \pi_n$ .
- **Output:** Joint probability  $P(x, \pi)$  of observing  $x$  if the hidden state sequence is  $\pi$ . The single path calculation essentially uses the following formula:

$$P(x, \pi) = P(x|\pi)P(\pi)$$

	<b>One path</b>	<b>All paths</b>
<b>Scoring</b>	1. Scoring $x$ , one path $P(x, \pi)$ Prob of a path, emissions	2. Scoring $x$ , all paths $P(x) = \sum_{\pi} P(x, \pi)$ Prob of emissions, over all paths
<b>Decoding</b>	3. Viterbi decoding $\pi^* = \operatorname{argmax}_{\pi} P(x, \pi)$ Most likely path	4. Posterior decoding $\pi^A = \{\pi_i \mid \pi_i = \operatorname{argmax}_k \sum_{\pi} P(\pi_i=k x)\}$ Path containing the most likely state at any time point.
<b>Learning</b>	5. Supervised learning, given $\pi$ $\Lambda^* = \operatorname{argmax}_{\Lambda} P(x, \pi   \Lambda)$ 6. Unsupervised learning. $\Lambda^* = \operatorname{argmax}_{\Lambda} \max_{\pi} P(x, \pi   \Lambda)$ Viterbi training, best path	6. Unsupervised learning $\Lambda^* = \operatorname{argmax}_{\Lambda} \sum_{\pi} P(x, \pi   \Lambda)$ Baum-Welch training, over all paths

Figure 7.15: The six algorithmic settings for HMMs. 1: path known. 2: path unknown, so search through all paths to find most likely path.  $\pi$  = path of hidden states,  $x$  = sequence,  $k$  = state,  $i$  = position

### Scoring over all paths

We define the all paths version of the scoring problem as follows:

- **Input:** A sequence of observations  $x = x_1 x_2 \dots x_n$  generated by an HMM  $M(Q, A, p, V, E)$ .
- **Output:** The joint probability  $P(x, \pi)$  of observing  $x$  over all possible sequences of hidden states  $\pi$ .

The probability over all hidden state paths  $\pi$  of the given sequence of observations is given by the following formula:

$$P(x) = \sum_{\pi} P(x, \pi)$$

We use this score when we are interested in the likelihood of a particular sequence for a given HMM. However, naively computing this sum requires calculating an exponential number of possible paths. Later in the lecture we will see how to compute this quantity in polynomial time.

### 7.5.2 Decoding

The problem of decoding is to determine, given some observed sequence, which path gives us the maximum likelihood of observing that sequence? Formally we define the problem as follows:

- **Decoding over a single path:**
  - Input: A sequence of observations  $x = x_1 x_2 \dots x_N$  generated by an HMM  $M(Q, A, p, V, E)$ .
  - Output: The most probable path of states,  $\pi^* = \pi_1^* \pi_2^* \dots \pi_N^*$

- **Decoding over all paths:**

- Input: A sequence of observations  $x = x_1 x_2 \dots x_N$  generated by an HMM  $M(Q, A, p, V, E)$ .
- Output: The path of states,  $\pi^* = \pi_1^* \pi_2^* \dots \pi_N^*$  that contains the most likely state at each time point.

In this lecture, we will look only at the problem of decoding over a single path. The problem of decoding over all paths will be discussed in the next lecture.

As mentioned previously, a brute force approach to the single path problem involves calculating the joint probabilities of the given emission sequence and all possible paths. The exponential number of possibilities makes this impractical. **Dynamic Programming** can be used to solve this problem.

We would like to find the most likely sequence of states based on the observation. As inputs, we are given the model parameters  $e_i(s)$ , the emission probabilities for each state, and  $a_{ij}s$ , the transition probabilities. The sequence of emissions  $x$  is also given. The goal is to find the sequence of hidden states,  $\pi^*$ , which maximizes the joint probability for the given sequence of emissions. That is,

$$\pi^* = \arg \max_{\pi} P(x, \pi)$$

Given the emitted sequence  $x$  we can evaluate any path through hidden states. However, we are looking for the best path. We start by looking for the optimal substructure of this problem.

We can say that the best path through a given state must contain the following:

- The best path to the previous state
- The best transition from the previous state to the current state
- The best path to the end state

Therefore the best path can be obtained based on the best path of the previous states, i.e., we can recursively find the best path. The **Viterbi algorithm** is a dynamic programming approach that is commonly used to solve this problem.

### The Viterbi algorithm

Suppose  $v_k(i)$  is the known probability that the most likely path ends at position (or time instance)  $i$  in state  $k$ , for each  $k$ . Then we can compute the corresponding probabilities at time  $i + 1$  by means of the following recursion.

$$v_l(i+1) = e_l(x_{i+1}) \max_k (a_{kl} v_k(i))$$

The most probable path  $\pi^*$ , or the maximum  $P(x, \pi)$ , can be found recursively, and the new most likely path for each state depends on

- The maximum score of the previous states
- The transition probability
- The emission probability.

In other words, the new maximum score for a particular state at time  $i$  is the one that maximizes the following: the penalty of transition from each previous state multiplied by their maximum previous scores multiplied by the emission probability at the current time.

All sequences have to start in state 0 (the initial state). By tracking pointers, the actual state sequence can then be found by backtracking. The solution of this Dynamic Programming problem is very similar to that of the alignment algorithms that were presented in previous lectures.

The steps of the Viterbi algorithm [? ] are summarized below:

1. Initialization ( $i = 0$ ):  $v_0(0) = 1$ ,  $v_k(0) = 0$  for  $k > 0$ .
2. Recursion ( $i = 1 \dots N$ ):  $v_k(i) = e_k(x_i) \max_j (a_{jk} v_j(i-1))$ ;  $\text{ptr}_i(l) = \arg \max_j (a_{jk} v_j(i-1))$ .

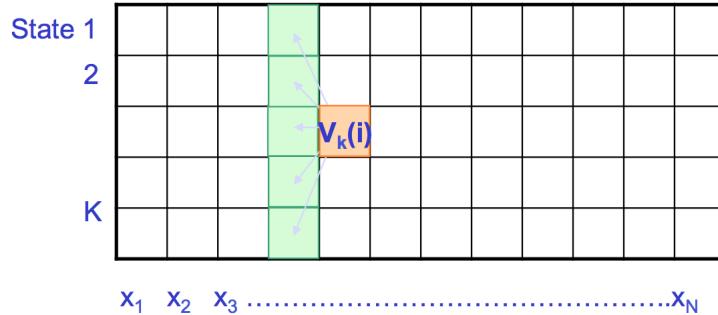


Figure 7.16: The Viterbi algorithm

3. Termination:  $P(x, \pi^*) = \max_k v_k(N); \pi_N^* = \arg \max_k v_k(N).$

4. Traceback ( $i = N \dots 1$ ):  $\pi_{i-1}^* = \text{ptr}_i(\pi_i^*).$

As shown in Figure 7.16, we fill the matrix from left to right and trace back. Each position in the matrix has  $K$  states to consider and there are  $KN$  cells in the matrix, so the required computation time is  $O(K^2N)$  and the required space is  $O(KN)$  (to remember the pointers). Note that the running time has been reduced from exponential to polynomial. In practice, we use log probabilities for the computation.

### 7.5.3 Evaluation

The goal of evaluation is to determine how well our model of the data captures the actual data, regardless of particular path. Many paths can generate any given sequence; the question is then how likely is the sequence to have been produced by this model, or how well does the model capture the sequence's characteristics? In other words, is this a good model? Additionally, evaluation can be used to compare different models.

A formal definition of the evaluation problem is as follows:

- Input: A sequence of observations  $x = x_1 x_2 \dots x_N$  and an HMM  $M(Q, A, p, V, E)$ .
- Output: The probability that  $x$  was generated by  $M$ , summed over all paths.

We know that if we are given an HMM we can generate a sequence of length  $n$  using the following steps:

- Start at state  $\pi_1$  according to probability  $a_{0\pi_1}$  (obtained using vector,  $p$ ).
- Emit letter  $x_1$  according to emission probability  $e_{\pi_1}(x_1)$ .
- Go to state  $\pi_2$  according to the transition probability  $a_{\pi_1|\pi_2}$
- Keep doing this until we emit  $x_N$ .

Thus we can emit any sequence and calculate its likelihood. However, many state sequences can emit the same  $x$ , so how do we calculate the total probability of generating a given  $x$  over all paths? In other words, our goal is to obtain the following probability:

$$P(x|M) = P(x) = \sum_{\pi} P(x, \pi) = \sum_{\pi} P(x|\pi)P(\pi)$$

Once again, the challenge is the exponential number of possible paths. One approach would be to use just the Viterbi path and ignore the others, since we already know how to obtain this path. But its probability is very small as it is only one of the many possible paths. It is a good approximation only if it has high probability density. The correct approach for calculating the exact sum iteratively is, as before, through dynamic programming. The algorithm that does this is known as the **Forward Algorithm**.

### The Forward Algorithm

First we define the forward probability  $f(i)$ .

$$f_l(i) = P(x_1 \dots x_i | \pi = l) \quad (7.1)$$

$$= \sum_{\pi_1 \dots \pi_{i-1}} P(x_1 \dots x_{i-1} | (\pi_1, \dots, \pi_{i-2}, \pi_{i-1}), \pi_i = l) e_l(x_i) \quad (7.2)$$

$$= \sum_k \sum_{\pi_1 \dots \pi_{i-2}} P(x_1 \dots x_{i-1} | (\pi_1, \dots, \pi_{i-2}), \pi_{i-1} = k) a_{kl} e_l(x_i) \quad (7.3)$$

$$= \sum_k f_k(i-1) a_{kl} e_l(x_i) \quad (7.4)$$

$$= e_l(x_i) \sum_k f_k(i-1) a_{kl} \quad (7.5)$$

$$(7.6)$$

The full algorithm[? ] is summarized below:

- Initialization ( $i = 0$ ):  $f_0(0) = 1$ ,  $f_k(0) = 0$  for  $k > 0$ .
- Iteration ( $i = 1 \dots N$ ):  $f_k(i) = e_k(x_i) \sum_j f_j(i-1) a_{jk}$ .
- Termination:  $P(x, \pi^*) = \sum_k f_k(N)$

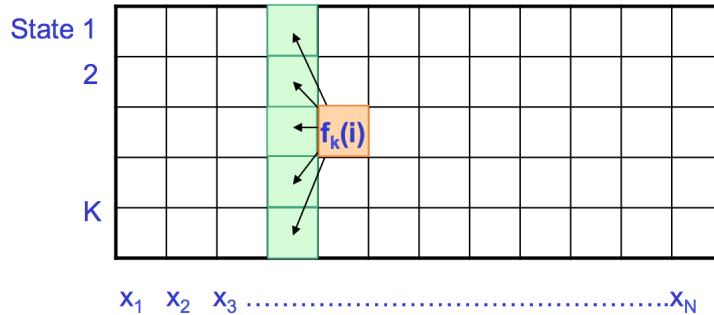


Figure 7.17: The Forward Algorithm

From Figure 7.17, it can be seen that the Forward Algorithm is very similar to the Viterbi algorithm, except that in the Forward Algorithm, a summation is used instead of maximization. Thus we can reuse the computations of the previous problem including penalty of emissions, penalty of transitions and sums of previous states. The required computation time is  $O(K^2N)$  and the required space is  $O(KN)$ . The drawback of this algorithm is that in practice, taking the sum of logs is difficult; therefore, approximations and scaling of probabilities are used instead.

## 7.6 An Interesting Question: Can We Incorporate Memory in Our Model?

The answer to this question is - Yes, we can! But how? Recall that, Markov models are memoryless. In other words, all memory of the model is enclosed in states. So, in order to store additional information, we must increase the number of states. Now, look back to the biological example we gave in Section 7.4.2. In our model, state emissions were dependent only on the current state. And, the current state encoded only one nucleotide. But, what if we want our model to count di-nucleotide frequencies (for CpG islands<sup>1</sup>), or,

<sup>1</sup>CpG stands for C-phosphate-G. So, CpG island refers to a region where GC di-nucleotide appear on the same strand.

tri-nucleotide frequencies (for codons), or di-codon frequencies involving six-nucleotide? We need to expand number of states.

For example, the last-seen nucleotide can be incorporated into the HMM's "memory" by splitting the plus and minus states from our High-GC/Low-GC HMM into multiple states: one for each nucleotide/region combination, as in Figure 7.18.

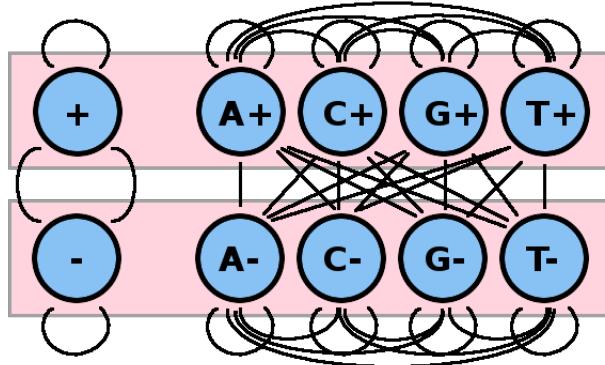


Figure 7.18: CpG Islands - Incorporating Memory

Moving from two to eight states allows us to retain memory of the last nucleotide observed, while also distinguishing between two distinct regions. Four new states now correspond to each of the original two states in the High/Low-GC HMM. Whereas the transition weights in the smaller HMM were based purely on the frequencies of individual nucleotides, now in the larger one, they are based on di-nucleotide frequencies.

With this added power, certain di-nucleotide sequences, such as CpG islands, can be modeled specifically: the transition from C+ to G+ can be assigned greater weight than the transition from A+ to G+. Further, transitions between + and - can be modeled more specifically to reflect the frequency (or infrequency) of particular di-nucleotide sequences within one or the other.

The process of adding memory to an HMM can be generalized and more memory can be added to allow the recognition of sequences of greater length. For instance, we can detect codon triplets with 32 states, or di-codon sextuplets with 2048 states. Memory within the HMM allows for increasingly tailored specificity in scanning.

## 7.7 Further Reading

### 7.7.1 Length Distributions of States and Generalized Hidden Markov Models

Given a Markov chain with the transition from any state to the end state having probability  $\tau$ , the probability of generating a sequence of length  $L$  (and then finishing with a transition to the end state) is given by:

$$\tau(1 - \tau)^{L-1}$$

Similarly, in the HMMs that we have been examining, the length of states will be exponentially distributed, which is not appropriate for many purposes. (For example, in a genomic sequence, an exponential distribution does not accurately capture the lengths of genes, exons, introns, etc). How can we construct a model that does not output state sequences with an exponential distribution of lengths? Suppose we want to make sure that our sequence has length exactly 5. We might construct a sequence of five states with only a single path permitted by the transition probabilities. If we include a self loop in one of the states, we will output sequences of minimum length 5, with longer sequences exponentially distributed. Suppose we have a chain of  $n$  states, with all chains starting with state  $\pi_1$  and transitioning to an end state after  $\pi_n$ . Also assume that the transition probability between state  $\pi_i$  and  $\pi_{i+1}$  is  $1 - p$ , while the self transition probability of state  $\pi_i$  is  $p$ . The probability that a sequence generated by this Markov chain has length  $L$  is given by:

$$\binom{L-1}{n-1} p^{L-n} (1-p)^n$$

This is called the negative binomial distribution.

More generally, we can adapt HMMs to produce output sequences of arbitrary length. In a *Generalized Hidden Markov Model* [? ] (also known as a *hidden semi-Markov model*), the output of each state is a string of symbols, rather than an individual symbol. The length as well as content of this output string can be chosen based on a probability distribution. Many gene finding tools are based on generalized hidden Markov models.

### 7.7.2 Conditional random fields

The conditional random field model a discriminative undirected probabilistic graphical model that is used alternatively to HMMs. It is used to encode known relationships between observations and construct consistent interpretations. It is often used for labeling or parsing of sequential data. It is widely used in gene finding. The following resources can be helpful in order to learn more about CRFs:

- Lecture on Conditional Random Fields from Probabilistic Graphical Models course: <https://class.coursera.org/pgm/lecture/preview/33>. For background, you might also want to watch the two previous segments, on pairwise Markov networks and general Gibbs distributions.
- Conditional random fields in biology: <http://www.cis.upenn.edu/~pereira/papers/crf.pdf>
- Conditional Random Fields tutorial: <http://people.cs.umass.edu/~mccallum/papers/crf-tutorial.pdf>

## 7.8 Current Research Directions

For the reasons outlined in this lecture, HMMs are a very useful tool in modeling biological sequences of many types. Some of the current applications of HMMs in biology are described below; see Figure 14 in Section 4.2.2 for a more compressed summary.

- **Modeling protein-coding regions:** Because the distributions

## 7.9 Tools for implementing HMMs

Hidden Markov Models and the associated algorithms introduced in this section have been implemented in several open-source packages that make it simple to build, train, and analyze HMMs. Below is a sample of the tools that are available.

- The R package 'HMM' enables you to initialize an HMM and run several types of analyses. For example, functions are implemented for calculating forward, backward, and posterior probabilities, as well as the Dishonest Casino model as presented in this section. The package is available in CRAN; see the documentation for more information: <https://cran.r-project.org/web/packages/HMM/HMM.pdf>
- In Python, HMMs are implemented in the hmmlearn package, which includes all of the algorithms discussed in this and the next lecture. Code and documentation is available here: <https://github.com/hmmlearn/hmmlearn>

## 7.10 What Have We Learned?

In this section, the main contents we covered are as following:

- We understand how to model sequential data by recognizing type of sequence, genomic, oral, verbal visual etc (not just whether they're similar but of the same class)

- We introduced the motivation behind adopting Hidden Markov Models in our analysis of genome annotation.
- We formalized Markov Chains and HMM under the light of weather prediction example.
- We got a sense of how to apply HMM in real world data by looking at Dishonest Casino and CG-rich region problems.
- We systematically introduced algorithmic settings of HMM and went into detail of three of them:
  - Scoring: scoring over single path
  - Scoring: scoring over all paths
  - Decoding: Viterbi coding in determining most likely path
- More specifically we know that given  $x$  we can calculate  $y$ :
  - Running the model: given the model we can generate sequence of a 'type'
  - Evaluation: given the model, emissions and states we can figure out the probability
  - Viterbi: given the model and emissions we can determine the optimal path
  - Forward: give the model and emissions we can find the total probability over all paths
- Finally, we discussed the possibility of introducing memory in the analysis of HMM and provided further readings for interested readers.

## Bibliography

---

CHAPTER  
EIGHT

---

## HIDDEN MARKOV MODELS II - POSTERIOR DECODING AND LEARNING

Charalampos Mavroforakis and Chidube Ezeozue (2012)  
Thomas Willems (2011)  
Amer Fejzic (2010)  
Elham Azizi (2009)

### Figures

---

8.1	Modeling Weather with HMM	160
8.2	Genomic applications of HMMs	161
8.3	The Forward Algorithm	162
8.4	The Backward Algorithm	166
8.5	HMM for CpG Islands	168
8.6	Supervised Learning of CpG islands	170
8.7	HMM model for alignment with affine gap penalties	173
8.8	State Space Diagram used in GENSCAN	175

---

## 8.1 Review of previous lecture

### 8.1.1 Introduction to Hidden Markov Models

In the last lecture, we familiarized ourselves with the concept of discrete-time Markov chains and Hidden Markov Models (HMMs). In particular, a Markov chain is a discrete random process that abides by the Markov property, i.e. that the probability of the next state depends only on the current state; this property is also frequently called "memorylessness." To model how states change from step to step, the Markov chain uses a matrix of transition probabilities. In addition, it is characterized by a one-to-one correspondence between the states and observed symbols; that is, the state fully determines all relevant observables. More formally, a Markov chain is fully defined by the following variables:

- $\pi_i \in Q$ , the state at the  $i^{th}$  step in a sequence of finite states  $Q$  of length  $N$  that can hold a value from a finite alphabet  $\Sigma$  of length  $K$
- $a_{jk}$ , the transition probability of moving from state  $j$  to state  $k$ ,  $P(\pi_i = k | \pi_{i-1} = j)$ , for each  $j, k$  in  $Q$
- $a_{0j} \in P$ , the probability that the initial state will be  $j$

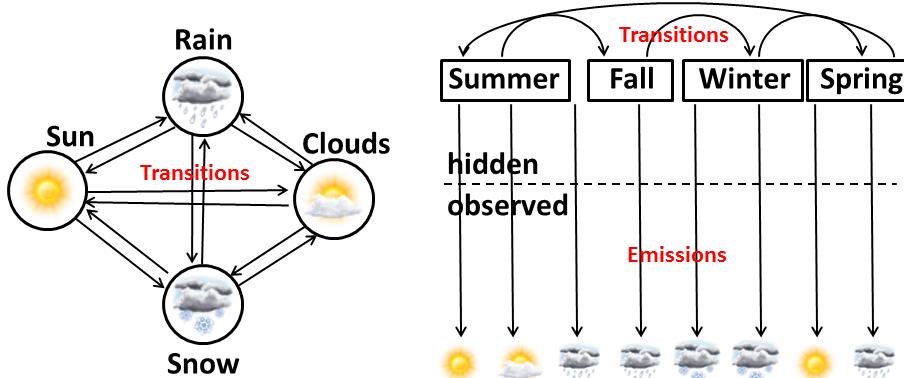


Figure 8.1: Modeling Weather with HMM

Examples of Markov chains are abundant in everyday life. In the last lecture, we considered the canonical example of a weather system in which each state is either rain, snow, sun or clouds and the observables of the system correspond exactly to the underlying state: there is nothing that we don't know upon making an observation, as the observation, i.e. whether it is sunny or raining, fully determines the underlying state, i.e. whether it is sunny or raining. Suppose, however, that we are considering the weather as it is probabilistically determined by the seasons - for example, it snows more often in the winter than in the spring - and suppose further that we are in ancient times and did not yet have access to knowledge about what the current season is. Now consider the problem of trying to infer the season (the hidden state) from the weather (the observable). There is some relationship between season and weather such that we can use information about the weather to make inferences about what season it is (if it snows a lot, it's probably not summer); this is the task that HMMs seek to undertake. Thus, in this situation, the states, the seasons, are considered "hidden" and no longer share a one-to-one correspondence with the observables, the weather. These types of situations require a generalization of Markov chains known as Hidden Markov Models (HMMs).

### Did You Know?

Markov Chains may be thought of as WYSIWYG - What You See Is What You Get

HMMs incorporate additional elements to model the disconnect between the observables of a system and the hidden states. For a sequence of length  $N$ , each observable state is instead replaced by a hidden state (the season) and a character emitted from that state (the weather). It is important to note that characters from each state are emitted according to a series of emission probabilities (say there is a 50% chance of snow, 30% chance of sun, and 20% chance of rain during winter). More formally, the two additional descriptors of an HMM are:

- $x_i \in X$ , the emission at the  $i^{th}$  step in a sequence of finite characters  $X$  of length  $N$  that can hold a character from a finite set of observation symbols  $v_l \in V$
- $e_k(v_l) \in E$ , the emission probability of emitting character  $v_l$  when the state is  $k$ ,  $P(x_i = v_l | \pi_i = k)$

In summary, an HMM is defined by the following variables:

- $a_{jk}$ ,  $e_k(v_l)$ , and  $a_{0j}$  that model the discrete random process
- $\pi_i$ , the sequence of hidden states
- $x_i$ , the sequence of observed emissions

### 8.1.2 Genomic Applications of HMMs

The figure below shows some genomic applications of HMMs

Application	Detection of GC-rich region	Detection of Conserved region	Detection of Protein coding exons	Detection of Protein coding conservation	Detection of Protein coding gene structures	Detection of chromatin states
<b>Topology / Transitions</b>	2 states, different nucleotide composition	2 states, difference conservation levels	2 states, different tri-nucleotide composition	2 states, different evolutionary signatures	~20 states, different composition / conservation, specific structure	40 states, different chromatin mark combinations
<b>Hidden States / Annotation</b>	GC-rich / AT-rich	Conserved/ non-Conserved	Coding (exon) / non-Coding (intron or intergenic)	Coding (exon) / non-Coding (intron or intergenic)	First / last / middle coding exon, UTRs, intron 1/2/3, intergenic, *(+,-) strand	Enhancer / Promoter / Transcribed / Repressed / Repetitive
<b>Emissions / Observations</b>	Nucleotides	Level of conservation	Triplets of nucleotides	64 x 64 matrix of codon substitution frequencies	Codons, nucleotides, splice sites, start/stop codons	Vector of chromatin mark frequencies

Figure 8.2: Genomic applications of HMMs

Niceties of some of the applications shown in figure 8.2 include:

- **Detection of Protein coding conservation**

This is similar to the application of detecting protein coding exons because the emissions are also not nucleotides but different in the sense that, instead of emitting codons, substitution frequencies of the codons are emitted.

- **Detection of Protein coding gene structures**

Here, it is important for different states to model first, last and middle exons independently, because they have distinct relevant structural features: for example, the first exon in a transcript goes through a start codon, the last exon goes through a stop codon, etc., and to make the best predictions, our model should encode these features. This differs from the application of detecting protein coding exons because in this case, the position of the exon is unimportant.

It is also important to differentiate between introns 1,2 and 3 so that the reading frame between one exon and the next exon can be remembered e.g. if one exon stops at the second codon position, the next one has to start at the third codon position. Therefore, the additional intron states encode the codon position.

- **Detection of chromatin states**

Chromatin state models are dynamic and vary from cell type to cell type so every cell type will have its own annotation. They will be discussed in fuller detail in the genomics lecture including strategies for stacking/concatenating cell types.

### 8.1.3 Viterbi decoding

Previously, we demonstrated that when given a full HMM  $(Q, A, X, E, P)$ , the likelihood that the discrete random process produced the provided series of hidden states and emissions is given by:

$$P(x_1, \dots, x_N, \pi_1, \dots, \pi_N) = a_{0\pi_1} \prod_i e_{\pi_i}(x_i) a_{\pi_i \pi_{i+1}} \quad (8.1)$$

This corresponds to the total joint probability,  $P(x, \pi)$ . Usually, however, the hidden states are not given and must be inferred; we're not interested in knowing the probability of the observed sequence given an underlying model of hidden states, but rather want to us the observed sequence to infer the hidden

states, such as when we use an organism's genomic sequence to infer the locations of its genes. One solution to this *decoding* problem is known as the **Viterbi decoding** algorithm. Running in  $O(K^2N)$  time and  $O(KN)$  space, where  $K$  is the number of states and  $N$  is the length of the observed sequence, this algorithm determines the sequence of hidden states (the path  $\pi^*$ ) that maximizes the joint probability of the observables and states, i.e.  $P(x, \pi)$ . Essentially, this algorithm defines  $V_k(i)$  to be the probability of the most likely path ending at state  $\pi_i = k$ , and it utilizes the optimal substructure argument that we saw in the sequence alignment module of the course to recursively compute  $V_k(i) = e_k(x_i) \times \max_j(V_j(i-1)a_{jk})$  in a dynamic programming algorithm.

#### 8.1.4 Forward Algorithm

Returning for a moment to the problem of 'scoring' rather than 'decoding,' another problem that we might want to tackle is that of, instead of computing the probability of a single path of hidden state emitting the observed sequence, calculating the *total* probability of the sequence being produced by all possible paths. For example, in the casino example, if the sequence of rolls is long enough, the probability of any *single* observed sequence and underlying path is very low, even if it is the single most likely sequence-path combination. We may instead want to take an agnostic attitude toward the path and assess the total probability of the observed sequence arising in any way.

In order to do that, we proposed the **Forward algorithm**, which is described in Figure 8.3

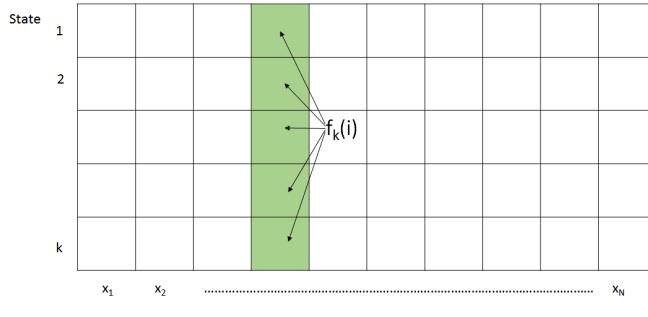


Figure 8.3: The Forward Algorithm

The forward algorithm first calculates the joint probability of observing the first  $t$  emitted characters and being in state  $k$  at time  $t$ . More formally,

$$f_k(t) = P(\pi_t = k, x_1, \dots, x_t) \quad (8.2)$$

Given that the number of paths is exponential in  $t$ , dynamic programming must be employed to solve this problem. We can develop a simple recursion for the forward algorithm by employing the Markov property as follows:

$$f_k(t) = \sum_l P(x_1, \dots, x_t, \pi_t = k, \pi_{t-1} = l) = \sum_l P(x_1, \dots, x_{t-1}, \pi_{t-1} = l) * P(x_t, \pi_t | \pi_{t-1}) \quad (8.3)$$

Recognizing that the first term corresponds to  $f_l(t-1)$  and that the second term can be expressed in terms of transition and emission probabilities, this leads to the final recursion:

$$f_k(t) = e_k(x_t) \sum_l f_l(t-1) * a_{lk} \quad (8.4)$$

Intuitively, one can understand this recursion as follows: Any path that is in state  $k$  at time  $t$  must have come from a path that was in state  $l$  at time  $t-1$ . The contribution of each of these sets of paths is then weighted by the cost of transitioning from state  $l$  to state  $k$ . It is also important to note that the Viterbi algorithm and forward algorithm largely share the same recursion. The only difference between the two algorithms lies in the fact that the Viterbi algorithm, seeking to find only the single most likely path, uses a

maximization function, whereas the forward algorithm, seeking to find the total probability of the sequence over all paths, uses a sum.

We can now compute  $f_k(t)$  based on a weighted sum of all the forward algorithm results tabulated during the previous time step. As shown in Figure 8.3, the forward algorithm can be easily implemented in a KxN dynamic programming table. The first column of the table is initialized according to the initial state probabilities  $a_{i0}$  and the algorithm then proceeds to process each column from left to right. Because there are KN entries and each entry examines a total of K other entries, this leads to  $O(K^2N)$  time complexity and  $O(KN)$  space.

In order now to calculate the total probability of a sequence of observed characters under the current HMM, we need to express this probability in terms of the forward algorithm gives in the following way:

$$P(x_1, \dots, x_n) = \sum_l P(x_1, \dots, x_n, \pi_N = l) = \sum_l f_l(N) \quad (8.5)$$

Hence, the sum of the elements in the last column of the dynamic programming table provides the total probability of an observed sequence of characters. In practice, given a sufficiently long sequence of emitted characters, the forward probabilities decrease very rapidly. To circumvent issues associated with storing small floating point numbers, logs-probabilities are used in the calculations instead of the probabilities themselves. This alteration requires a slight adjustment to the algorithm and the use of a Taylor series expansion for the exponential function.

### 8.1.5 This lecture

- This lecture will discuss **posterior decoding**, an algorithm which again will infer the hidden state sequence  $\pi$  that maximizes a different metric. In particular, it finds the most likely state at every position over all possible paths and does so using both the **forward** and **backward** algorithm.
- Afterwards, we will show how to encode “memory” in a Markov chain by adding more states to search a genome for dinucleotide storing stateands.
- We will then discuss how to use Maximum Likelihood parameter estimation for supervised learning with a labelled dataset
- We will also briefly see how to use Viterbi learning for unsupervised estimation of the parameters of an unlabelled dataset
- Finally, we will learn how to use Expectation Maximization (EM) for unsupervised estimation of parameters of an unlabelled dataset where the specific algorithm for HMMs is known as the Baum-Welch algorithm.

## 8.2 Posterior Decoding

### 8.2.1 Motivation

Although the **Viterbi decoding** algorithm provides one means of estimating the hidden states underlying a sequence of observed characters, another valid means of inference is provided by **posterior decoding**.

Posterior decoding provides the most likely state at any point in time. To gain some intuition for posterior decoding, let's see how it applies to the situation in which a dishonest casino alternates between a fair and loaded die. Suppose we enter the casino knowing that the unfair die is used 60 percent of the time. With this knowledge and no die rolls, our best guess for the current die is obviously the loaded one. After one roll, the probability that the loaded die was used is given by

$$P(\text{die} = \text{loaded} | \text{roll} = k) = \frac{P(\text{die} = \text{loaded}) * P(\text{roll} = k | \text{die} = \text{loaded})}{P(\text{roll} = k)}. \quad (8.6)$$

If we instead observed a sequence of  $N$  die rolls, how do perform a similar sort of inference? By allowing information to flow between the  $N$  rolls and influence the probability of each state, posterior decoding is a natural extension of the above inference to a sequence of arbitrary length. More formally, instead of identifying a single path of maximum likelihood, posterior decoding considers the probability of any path lying in state  $k$  at time  $t$  given all of the observed characters, i.e.  $P(\pi_t = k | x_1, \dots, x_n)$ . The state that maximizes this probability for a given time is then considered as the most likely state at that point.

It is important to note that in addition to information flowing forward to determine the most likely state at a point, information may also flow backward from the end of the sequence to that state to augment or reduce the likelihood of each state at that point. This is partly a natural consequence of the reversibility of Bayes' rule: our probabilities change from prior probabilities into posterior probabilities upon observing more data. To elucidate this, imagine the casino example again. As stated earlier, without observing any rolls, the  $\text{state}_0$  is most likely to be unfair: this is our prior probability. If the first roll is a 6, our belief that  $\text{state}_1$  is unfair is reinforced (if rolling sixes is more likely in an unfair die). If a 6 is rolled again, information flows backwards from the second die roll and reinforces our  $\text{state}_1$  belief of an unfair die even more. The more rolls we have, the more information that flows backwards and reinforces or contrasts our beliefs about the state thus illustrating the way information flows backward and forward to affect our belief about the states in Posterior Decoding.

Using some elementary manipulations, we can rearrange this probability into the following form using Bayes' rule:

$$\pi_t^* = \operatorname{argmax}_k P(\pi_t = k | x_1, \dots, x_n) = \operatorname{argmax}_k \frac{P(\pi_t = k, x_1, \dots, x_n)}{P(x_1, \dots, x_n)} \quad (8.7)$$

Because  $P(x)$  is a constant, we can neglect it when maximizing the function. Therefore,

$$\pi_t^* = \operatorname{argmax}_k P(\pi_t = k, x_1, \dots, x_t) * P(x_{t+1}, \dots, x_n | \pi_t = k, x_1, \dots, x_t) \quad (8.8)$$

Using the Markov property, we can simply write this expression as follows:

$$\pi_t^* = \operatorname{argmax}_k P(\pi_t = k, x_1, \dots, x_t) * P(x_{t+1}, \dots, x_n | \pi_t = k) = \operatorname{argmax}_k f_k(t) * b_k(t) \quad (8.9)$$

Here, we've defined  $f_k(t) = P(\pi_t = k, x_1, \dots, x_t)$  and  $b_k(t) = P(x_{t+1}, \dots, x_n | \pi_t = k)$ . As we will shortly see, these parameters are calculated using the **forward algorithm** and the **backward algorithm** respectively. To solve the posterior decoding problem, we merely need to solve each of these subproblems. The forward algorithm has been illustrated in the previous chapter and in the review at the start of this chapter and the backward algorithm will be explained in the next section.

### 8.2.2 Backward Algorithm

As previously described, the backward algorithm is used to calculate the following probability:

$$b_k(t) = P(x_{t+1}, \dots, x_n | \pi_t = k) \quad (8.10)$$

We can begin to develop a recursion by expanding into the following form:

$$b_k(t) = \sum_l P(x_{t+1}, \dots, x_n, \pi_{t+1} = l | \pi_t = k) \quad (8.11)$$

From the Markov property, we then obtain:

$$b_k(t) = \sum_l P(x_{t+2}, \dots, x_n | \pi_{t+1} = l) * P(\pi_{t+1} = l | \pi_t = k) * P(x_{t+1} | \pi_{t+1} = k) \quad (8.12)$$

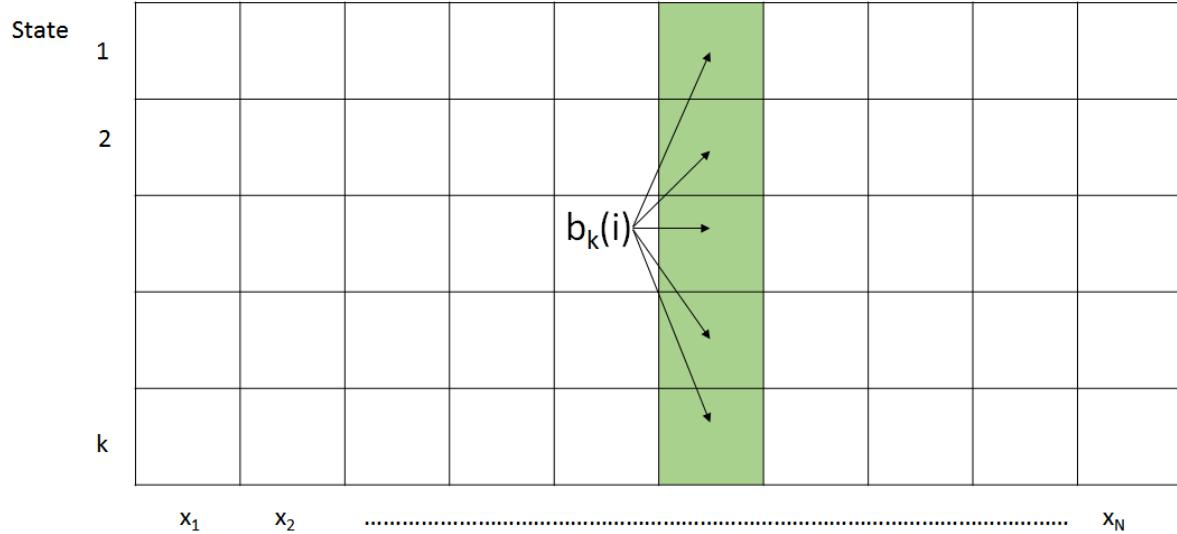
The first term merely corresponds to  $b_l(t+1)$ . Expressing in terms of emission and transition probabilities gives the final recursion:

$$b_k(t) = \sum_l b_l(i+1) * a_{kl} * e_l(x_{t+1}) \quad (8.13)$$

Comparison of the forward and backward recursions leads to some interesting insight. Whereas the forward algorithm uses the results at  $t - 1$  to calculate the result for  $t$ , the backward algorithm uses the results from  $t + 1$ , leading naturally to their respective names. Another significant difference lies in the emission probabilities; while the emissions for the forward algorithm occur from the current state and can therefore be excluded from the summation, the emissions for the backward algorithm occur at time  $t + 1$  and therefore must be included within the summation.

Given their similarities, it is not surprising that the backward algorithm is also implemented using a KxN dynamic programming table. The algorithm, as depicted in Figure 8.4, begins by initializing the rightmost column of the table to unity. Proceeding from right to left, each column is then calculated by taking a weighted sum of the values in the column to the right according to the recursion outlined above. After calculating the leftmost column, all of the backward probabilities have been calculated and the algorithm terminates. Because there are KN entries and each entry examines a total of K other entries, this leads to  $O(K^2 N)$  time complexity and  $O(KN)$  space, bounds identical to those of the forward algorithm.

Just as  $P(X)$  was calculated by summing the rightmost column of the forward algorithm's DP table,  $P(X)$  can also be calculated from the sum of the leftmost column of the backward algorithm's DP table. Therefore, these methods are virtually interchangeable for this particular calculation.



- Input:  $x = x_1 \dots x_N$
- Initialization:  $b_k(N) = a_{k0}$ , for all  $k$
- Iteration:  $b_k(i) = \sum_l e_l(x_{i+1}) a_{kl} b_l(i+1)$
- Termination:  $P(x) = \sum_l a_{0l} e_l(x_1) b_l(1)$

Figure 8.4: The Backward Algorithm

### Did You Know?

Note that even when executing the backward algorithm, forward transition probabilities are used i.e if moving in the backward direction involves a transition from state B  $\rightarrow$  A, the probability of transitioning from state A  $\rightarrow$  B is used. This is because moving backward from state B to state A implies that state B follows state A in our normal, forward order, thus calling for the same transition probability.

### 8.2.3 The Big Picture

Why do we have to make both forward and backward calculations for posterior decoding, while the algorithms that we have discussed previously call for only one direction? The difference lies in the fact that posterior decoding seeks to produce probabilities for the underlying states of **individual positions** rather than whole sequences of positions. In seeking to find the most likely underlying state of a given position, we need to take into account the entire sequence in which that position exists, both before and after it, as befits a Bayesian approach - and to do this in a dynamic programming algorithm, in which we compute recursively and end with a maximizing function, we must approach our position of interest from both sides.

Given that we can calculate both  $f_k(t)$  and  $b_k(t)$  in  $\theta(K^2N)$  time and  $\theta(KN)$  space for all  $t = 1 \dots n$ , we can use posterior decoding to determine the most likely state  $\pi_t^*$  for  $t = 1 \dots n$ . The relevant expression is given by

$$\pi_t^* = \operatorname{argmax}_k P(\pi_i = k | x) = \frac{f_k(i) * b_k(i)}{P(x)} \quad (8.14)$$

With two methods (Viterbi and posterior) to decode, which is more appropriate? When trying to classify each hidden state, the Posterior decoding method is more informative because it takes into account all possible paths when determining the most likely state. In contrast, the Viterbi method only takes into account one path, which may end up representing a minimal fraction of the total probability. At the same time, however, posterior decoding may give an invalid sequence of states! By selecting for the maximum probability state of each position independently, we're not considering how likely the transitions between these states are. For example, the states identified at time points  $t$  and  $t + 1$  might have zero transition probability between them. As a result, selecting a decoding method is highly dependent on the application of interest.

## FAQ

**Q:** What does it imply when the Viterbi algorithm and Posterior decoding disagree on the path?

**A:** In a sense, it is simply a reminder that our model gives us what it's selecting for. When we seek the maximum probability state of each independent position and disregard transitions between these max probability states, we may get something different than when we seek to find the most likely total path. Biology is complicated; it is important to think about what metric is most relevant to the biological situation at hand. In the genomic context, a disagreement might be a result of some 'funky' biology; alternative splicing, for instance. In some cases, the Viterbi algorithm will be close to the Posterior decoding while in some others they may disagree.

## 8.3 Encoding Memory in a HMM: Detection of CpG islands

CpG islands are defined as regions within a genome that are enriched with pairs of C and G nucleotides on the same strand. Typically, when this dinucleotide is present within a genome, it becomes methylated, and when deamination of the cytosine occurs, as it does at some base frequency, it becomes a thymine, another natural nucleotide, and thus cannot as easily be recognized by the cell as a mutation, causing a C to T mutation. This increased mutation frequency at CpG islands depletes CpG islands over evolutionary time and renders them relatively rare. Because the methylation can occur on either strand, CpGs usually mutate into a TpG or a CpA. However, when situated within an active promoter, methylation is suppressed, and CpG dinucleotides are able to persist. Similarly, CpGs in regions important to cell function are conserved due to evolutionary pressure. As a result, detecting CpG islands can highlight promoter regions, other transcriptionally active regions, or sites of purifying selection within a genome.

### Did You Know?

CpG stands for [C]ytosine - [p]hosphate backbone - [G]uanine. The 'p' implies that we are referring to the same strand of the double helix, rather than a G-C base pair occurring across the helix.

Given their biological significance, CpG islands are prime candidates for modelling. Initially, one may attempt to identify these islands by scanning the genome for fixed intervals rich in GC. This approach's efficacy is undermined by the selection of an appropriate window size; while too small of a window may not capture all of a particular CpG island, too large of a window would result in missing many smaller but bona fide CpG islands. Examining the genome on a per codon basis also leads to difficulties because CpG pairs do not necessarily code for amino acids and thus may not lie within a single codon. Instead, HMMs are much better suited to modelling this scenario because, as we shall shortly see in the section on unsupervised learning, HMMs can adapt their underlying parameters to maximize their likelihood.

Not all HMMs, however, are well suited to this particular task. An HMM model that only considers the single nucleotide frequencies of C's and G's will fail to capture the nature of CpG islands. Consider one such HMM with the two following hidden states :

- '+' state representing CpG islands
- '-' state: representing non-islands

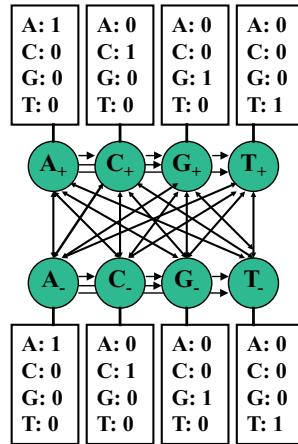


Figure 8.5: HMM for CpG Islands

Each of these two states then emits A, C, G and T bases with a certain probability. Although the CpG islands in this model can be enriched with C's and G's by increasing their respective emission probabilities, this model will fail to capture the fact that the C's and G's predominantly occur in pairs.

Because of the Markov property that governs HMM's, the only information available at each time step must be contained within the current state. One way to expand the number of states encoded by the model is to change the emission matrix to emit not a single base, but two bases at once. Another way, which we will focus on is to augment the state space. To do so, the individual '+' and '-' states can be replaced with 4 '+' states and 4 '-' states:  $A_+$ ,  $C_+$ ,  $G_+$ ,  $T_+$ ,  $A_-$ ,  $C_-$ ,  $G_-$ ,  $T_-$  (Figure 8.5). Specifically, there are 2 ways to model this, and this choice will result in different emission probabilities:

- One model suggests that the state  $A_+$ , for instance, implies that we are currently in a CpG island and the *previous* character was an A. The emission probabilities here will carry most of the information and the transitions will be fairly degenerate.
- Another model suggests that the state  $A_+$ , for instance, implies that we are currently in a CpG island and the *current* character is an A. The emission probability here will be 1 for A and 0 for all other letters and the transition probabilities will bear most of the information in the model and the emissions will be fairly degenerate. We will assume this model from now on.

### Did You Know?

The number of transitions is the square of the number of states. This gives a rough idea of how increasing HMM "memory" (and hence states) scale.

The memory of this system derives from the fact that each state can only emit one character and therefore "remembers" its emitted character. Furthermore, the dinucleotide nature of the CpG islands is incorporated within the transition matrices. In particular, the transition frequency from  $C_+$  to  $G_+$  states is significantly higher than from  $C_-$  to a  $G_-$  states, demonstrating that these pairs occur more often within the islands.

### FAQ

**Q:** Since each state emits only one character, can we then say this reduces to a Markov Chain instead of a HMM?

**A:** No. Even though the emissions indicate the letter of the hidden state, they do not indicate if the state is a CpG island or not: both an  $A_-$  and an  $A_+$  state emit only the observable A.

**FAQ**

**Q:** How do we incorporate our knowledge about the system while training HMM models eg. some emission probabilities of 0 in the CpG island detection case?

**A:** We could either force our knowledge on the model by setting some parameters and leaving others to vary or we could let the HMM loose on the model and let it discover those relationships. As a matter of fact, there are even methods that simplify the model by forcing a subset of parameters to be 0 but allowing the HMM to choose which subset.

Given the above framework, we can use posterior decoding to analyze each base within a genome and determine whether it is most likely a constituent of a CpG island or not. But having constructed the expanded HMM model, how can we verify that it is in fact better than the single nucleotide model? We previously demonstrated that the forward or backward algorithm can be used to calculate  $P(x)$  for a given model. If the likelihood of our dataset is higher given the second model than the first model, it most likely captures the underlying behavior more effectively.

However, there is one risk in complicating the model, which is overfitting. Increasing the number of parameters for an HMM makes the HMM more likely to overfit the data and be less accurate in capturing the underlying behavior. A common solution to this in machine learning is to use regularization, which is essentially using fewer parameters. In this case, it is possible to reduce number of parameters to learn by constraining all +/- transition probabilities to be the same value and all -/+ transition probabilities to be the same value, as the transitions back and forth from the + and - states are what we are interested in modeling, and the actual bases where the transition occurred are not that important to our model. Thus for this constrained model we have to learn fewer parameters which leads to a simpler model and can help to avoid overfitting.

**FAQ**

**Q:** Are there other ways to encode the memory for CpG island detection?

**A:** Other ideas that may be experimented with include

- Emit dinucleotides and figure out a way to deal with overlap.
- Add a special state that goes from C to G.

## 8.4 Learning

We saw how to score and decode an HMM-generated sequence in two different ways. However, these methods assumed that we already knew the emission and transition probabilities. While we are always free to hazard a guess at these, we may sometimes want to use a more data-driven, empirical approach to deriving these parameters. Fortunately, the HMM framework enables the learning of these probabilities when provided a set of training data and a set architecture for the model.

When the training data is labelled, estimation of the probabilities is a form of supervised learning. One such instance would occur if we were given a DNA sequence of one million nucleotides in which the CpG islands had all been experimentally annotated and were asked to use this to estimate our model parameters.

In contrast, when the training data is unlabelled, the estimation problem is a form of unsupervised learning. Continuing with the CpG island example, this situation would occur if the provided DNA sequence contained no island annotation whatsoever and we needed to both estimate model parameters and identify the islands.

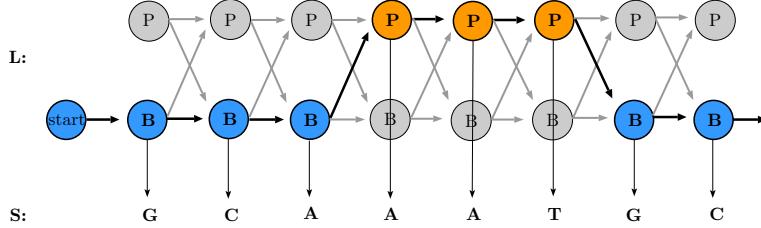


Figure 8.6: Supervised Learning of CpG islands

### 8.4.1 Supervised Learning

When provided with labelled data, the idea of estimating model parameters is straightforward. Suppose that you are given a labelled sequence  $x_1, \dots, x_N$  as well as the true hidden state sequence  $\pi_1, \dots, \pi_N$ . Intuitively, one might expect that the probabilities that maximize the data's likelihood are the actual probabilities that one observes within the data. This is indeed the case and can be formalized by defining  $A_{kl}$  to be the number of times hidden state  $k$  transitions to  $l$  and  $E_k(b)$  to be the number of times  $b$  is emitted from hidden state  $k$ . The parameters  $\theta$  that maximize  $P(x|\theta)$  are simply obtained by counting as follows:

$$a_{kl} = \frac{A_{kl}}{\sum_i A_{ki}} \quad (8.15a)$$

$$e_k(b) = \frac{E_k(b)}{\sum_c E_k(c)} \quad (8.15b)$$

One example training set is shown in Figure 8.6. In this example, it is obvious that the probability of transitioning from B to P is  $\frac{1}{3+1} = \frac{1}{4}$  (there are 3 B to B transitions and 1 B to P transitions) and the probability of emitting a G from the B state is  $\frac{2}{2+2+1} = \frac{2}{5}$  (there are 2 G's emitted from the B state, 2 C's and 1 A).

Notice, however, that in the above example the emission probability of character T from state B is 0 because no such emissions were encountered in the training set. A zero probability, either for transitioning or emitting, is particularly problematic because it leads to an infinite log penalty. In reality, however, the zero probability may merely have arisen due to over-fitting or a small sample size. To rectify this issue and maintain flexibility within our model, we can collect more data on which to train, reducing the possibility that the zero probability is due to a small sample size. Another possibility is to use 'pseudocounts' instead of absolute counts: artificially adding some number of counts to our training data which we think more accurately represent the actual parameters and help counteract sample size errors.

- $A_{kl}^* = A_{kl} + r_{kl}$
- $E_k(b)^* = E_k(b) + r_k(b)$

Larger pseudocount parameters correspond to a strong prior belief about the parameters, reflected in the fact that these pseudocounts, derived from your priors, are comparatively overwhelming the observations, your training data. Likewise, small pseudocount parameters ( $r \ll 1$ ) are more often used when our priors are relatively weak and we are aiming not to overwhelm the empirical data but only to avoid excessively harsh probabilities of 0.

### 8.4.2 Unsupervised Learning

Unsupervised learning involves estimating parameters based on unlabelled data. This may seem impossible - how can we take data about which we know nothing and use it to "learn"? - but an iterative approach can yield surprisingly good results, and is the typical choice in these cases. This can be thought of loosely as an evolutionary algorithm: from some initial choice of parameters, the algorithm assesses how well the

parameters explain or relate to the data, uses some step in that assessment to make improvements on the parameters, and then assesses the new parameters, producing incremental improvements in the parameters at every step just as the fitness or lack thereof of a particular organism in its environment produces incremental increases over evolutionary time as advantageous alleles are passed on preferentially.

Suppose we have some sort of prior belief about what each emission and transition probability should be. Given these parameters, we can use a decoding method to infer the hidden states underlying the provided data sequence. Using this particular decoding parse, we can then re-estimate the transition and emission counts and probabilities in a process similar to that used for supervised learning. If we repeat this procedure until the improvement in the data's likelihood remains relatively stable, the data sequence should ultimately drive the parameters to their appropriate values.

## FAQ

**Q:** Why does unsupervised learning even work? Or is it magic?

**A:** Unsupervised learning works because we have the sequence (input data) and this guides every step of the iteration; to go from a labelled sequence to a set of parameters, the later are guided by the input and its annotation, while to annotate the input data, the parameters and the sequence guide the procedure.

For HMMs in particular, two main methods of unsupervised learning are useful.

### Expectation Maximization using Viterbi training

The first method, **Viterbi training**, is relatively simple but not entirely rigorous. After picking some initial best-guess model parameters, it proceeds as follows:

**E step:** Perform Viterbi decoding to find  $\pi^*$ , the most likely path of states

Calculate  $A_{kl}^*, E_k(b)^*$  using pseudocounts based on the transitions and emissions observed in  $\pi^*$  states given the latest parameters and observed sequence (Expectation step)

**M step:** Calculate the new parameters  $a_{kl}, e_k(b)$  using the simple counting formalism in supervised learning (Maximization step)

**Iteration:** Repeat the E and M steps until the likelihood  $P(x|\theta)$  converges

Although Viterbi training converges rapidly, its resulting parameter estimations are usually inferior to those of the Baum-Welch Algorithm. This result stems from the fact that Viterbi training only considers the most probable hidden path instead of the collection of all possible hidden paths.

### Expectation Maximization: The Baum-Welch Algorithm

The more rigorous approach to unsupervised learning involves an application of Expectation Maximization to HMM's. In general, EM proceeds in the following manner:

**Init:** Initialize the parameters to some best-guess state

**E step:** Estimate the expected probability of hidden states given the latest parameters and observed sequence (Expectation step)

**M step:** Choose new maximum likelihood parameters using the probability distribution of hidden states (Maximization step)

**Iteration:** Repeat the E and M steps until the likelihood of the data given the parameters converges

The power of EM lies in the fact that  $P(x|\theta)$  is guaranteed to increase with each iteration of the algorithm. Therefore, when this probability converges, a local maximum has been reached. As a result, if we utilize a variety of initialization states, we will most likely be able to identify the global maximum, i.e. the best parameters  $\theta$ . The Baum-Welch algorithm generalizes EM to HMM's. In particular, it uses the forward and backward algorithms to calculate  $P(x|\theta)$  and to estimate  $A_{kl}$  and  $E_k(b)$ . The algorithm proceeds as follows:

**Initialization** 1. Initialize the parameters to some best-guess state

- Iteration**
1. Run the forward algorithm
  2. Run the backward algorithm
  3. Calculate the new log-likelihood  $P(x|\theta)$
  4. Calculate  $A_{kl}$  and  $E_k(b)$
  5. Calculate  $a_{kl}$  and  $e_k(b)$  using the pseudocount formulas
  6. Repeat until  $P(x|\theta)$  converges

Previously, we discussed how to compute  $P(x|\theta)$  using either the forward or backward algorithm's final results. But how do we estimate  $A_{kl}$  and  $E_k(b)$ ? Let's consider the expected number of transitions from state  $k$  to state  $l$  given a current set of parameters  $\theta$ . We can express this expectation as

$$A_{kl} = \sum_t P(\pi_t = k, \pi_{t+1} = l | x, \theta) = \sum_t \frac{P(\pi_t = k, \pi_{t+1} = l, x | \theta)}{P(x | \theta)} \quad (8.16)$$

Exploiting the Markov property and the definitions of the emission and transition probabilities leads to the following derivation:

$$A_{kl} = \sum_t \frac{P(x_1 \dots x_t, \pi_t = k, \pi_{t+1} = l, x_{t+1} \dots x_N | \theta)}{P(x | \theta)} \quad (8.17a)$$

$$= \sum_t \frac{P(x_1 \dots x_t, \pi_t = k) * P(\pi_{t+1} = l, x_{t+1} \dots x_N | \pi_t, \theta)}{P(x | \theta)} \quad (8.17b)$$

$$= \sum_t \frac{f_k(t) * P(\pi_{t+1} = l | \pi_t = k) * P(x_{t+1} | \pi_{t+1} = l) * P(x_{t+2} \dots x_N | \pi_{t+1} = l, \theta)}{P(x | \theta)} \quad (8.17c)$$

$$\Rightarrow A_{kl} = \sum_t \frac{f_k(t) * a_{kl} * e_l(x_{t+1}) * b_l(t+1)}{P(x | \theta)} \quad (8.17d)$$

A similar derivation leads to the following expression for  $E_k(b)$ :

$$E_k(b) = \sum_{i|x_i=b} \frac{f_k(t) * b_k(t)}{P(x | \theta)} \quad (8.18)$$

Therefore, by running the forward and backward algorithms, we have all of the information necessary to calculate  $P(x|\theta)$  and to update the emission and transition probabilities during each iteration. Because these updates are constant time operations once  $P(x|\theta)$ ,  $f_k(t)$  and  $b_k(t)$  have been computed, the total time complexity for this version of unsupervised learning is  $\theta(K^2 NS)$ , where  $S$  is the total number of iterations.

## FAQ

**Q:** How do you encode your prior beliefs when learning with Baum-Welch?

**A:** Those prior beliefs are encoded in the initializations of the forward and backward algorithms

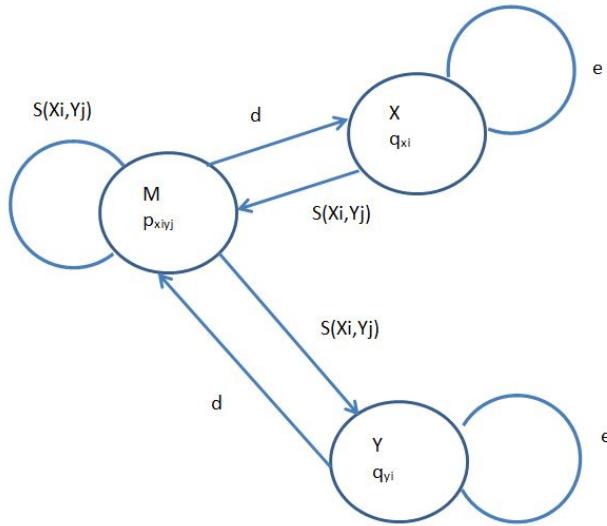


Figure 8.7: HMM model for alignment with affine gap penalties

## 8.5 Using HMMs to align sequences with affine gap penalties

We can use HMM to align sequences with affine gap penalties. Recall that affine gap penalties penalizes more to open/ start the gap than to extend it, thus the penalty of a gap of length  $g$  is  $r(g) = -d - (g-1)*e$ , where  $d$  is the penalty to open the gap and  $e$  is the penalty to extend an already open gap.

We will look into aligning two sequences with the affine gap penalty. We are given two sequences are  $X$  and  $Y$ , the scoring matrix  $S$  ( $S(x_i, y_j)$ ) = score of matching  $x_i$  with  $y_j$ , gap opening penalty of  $d$  and gap extension penalty of  $e$ . We can map this problem into an HMM problem by using the following states, transition probabilities and emission probabilities.

### States:

There are three states involves:  $M$  (matching  $x_i$  with  $y_j$ ),  $X$  (aligning  $x_i$  with a gap),  $Y$  (aligning  $y_j$  with a gap). Also, alongside each transition, there's an update of the  $i, j$  indices. Whenever we are in state  $M$ ,  $(i, j) = (i, j) + (1, 1)$ . In state  $X$ ,  $(i, j) = (i, j) + (1, 0)$ . In state  $Y$ ,  $(i, j) = (i, j) + (0, 1)$ .

### Transition probabilities:

There are 7 transition probabilities to consider as shown in figure 6.

$$P(\text{next State} = M \mid \text{current} = M) = S(x_i, y_j)$$

$$P(\text{next State} = X \mid \text{current} = M) = d$$

$$P(\text{next State} = Y \mid \text{current} = M) = d$$

$$P(\text{next State} = X \mid \text{current} = X) = e$$

$$P(\text{next State} = M \mid \text{current} = X) = S(x_i, y_j)$$

$$P(\text{next State} = Y \mid \text{current} = X) = e$$

$$P(\text{next State} = M \mid \text{current} = Y) = S(x_i, y_j)$$

We can also save the transition probabilities in a transition matrix  $A = [a_{ij}]$ , where  $a_{ij} = P(\text{next State} = j \mid \text{current} = i)$  and  $\sum_j A_{ij} = 1$

### Emission probabilities:

The emission probabilities are:

From state  $M$ :  $p_{x_i y_j} = p(x_i \text{ aligned to } y_j)$

From state  $X$ :  $q_{x_i} = p(x_i \text{ aligned to gap})$

From state  $Y$ :  $q_{y_i} = p(y_j \text{ aligned to gap})$

Example:

$X = 'VLSPADK'$

$Y = 'HLAESK'$

The alignment generated by the model is:

MMXXMYM

Which corresponds to:

$X = 'VLSPAD.K'$

$Y = 'HL\_AESK'$

### Did You Know?

For classification purposes, Posterior decoding 'path' is more informative than Viterbi path as it is a more refined measure of which hidden states generated x. However, it may give an invalid sequence of states, for example when not all  $j \rightarrow k$  transitions may be possible, it might have state( $i$ ) =  $j$  and state( $i+1$ ) =  $k$

## 8.6 Current Research Directions

- HMM's have been used extensively in various fields of computational biology. One of the first such applications was in a gene-finding algorithm known as GENSCAN written by Chris Burge and Samuel Karlin [1]. Because the geometric length distribution of HMM's does not model exonic regions well, Burge et. al used an adaptation of HMM's known as hidden semi-Markov models (HSMM's). These types of models differ in that whenever a hidden state is reached, the length of duration of that state ( $d_i$ ) is chosen from a distribution and the state then emits exactly  $d_i$  characters. The transition from this hidden state to the next is then analogous to the HMM procedure except that  $a_{kk} = 0$  for all  $k$ , thereby preventing self-transitioning. Many of the same algorithms that were previously developed for HMM's can be modified for HSMM's. Although the details won't be discussed here, the forward and backward algorithms can be modified to run in  $O(K^2N^3)$  time, where  $N$  is the number of observed characters. This time complexity assumes that there is no upper bound on the length of a state's duration, but imposing such a bound reduces the complexity to  $O(K^2ND^2)$ , where  $D$  is the maximum possible duration of a state.

The basic state diagram underlying Burge's model is depicted in Figure 8.8. The included diagram only lists the states on the forward strand of DNA, but in reality a mirror image of these states is also included for the reverse strand, resulting in a total of 27 hidden states. As the diagram illustrates, the model incorporates many of the major functional units of genes, including exons, introns, promoters, UTR's and poly-A tails. In addition, three different intronic and exonic states are used to ensure that the total length of all exons in a gene is a multiple of three. Similar to the CpG island example, this expanded state-space enabled the encoding of memory within the model.

- A recent effort has been made to make an HMM-based approach to homology searches, called HMMER, a viable alternative to BLAST in terms of computational efficiency. Unlike most other homology search algorithms, HMMER, written by Sean Eddy, uses the Forward algorithm's average over alignment uncertainty, rather than only reporting the maximum likelihood alignment (a la Viterbi); this approach is often better for detecting more remote homologues, as as divergence times increase, there may become more viable ways of aligning sequences, each of them individually not sufficiently strong to be differentiated from noise but together giving evidence for homology. A particularly exciting recent development is that HMMER is now available as a web server; it can be found at <http://www.ebi.ac.uk/Tools/hmmer/>.
- An interesting subject that may be explored also concerns the agreement of Viterbi and Posterior decoding paths; not just for CpG island detection but even for chromatin state detection. One may look at multiple paths by sampling, asking questions such as:
  - What is the maximum a posteriori vs viterbi path? Where do they differ?
  - Can complete but maximally disjoint (from Viterbi) paths be found?

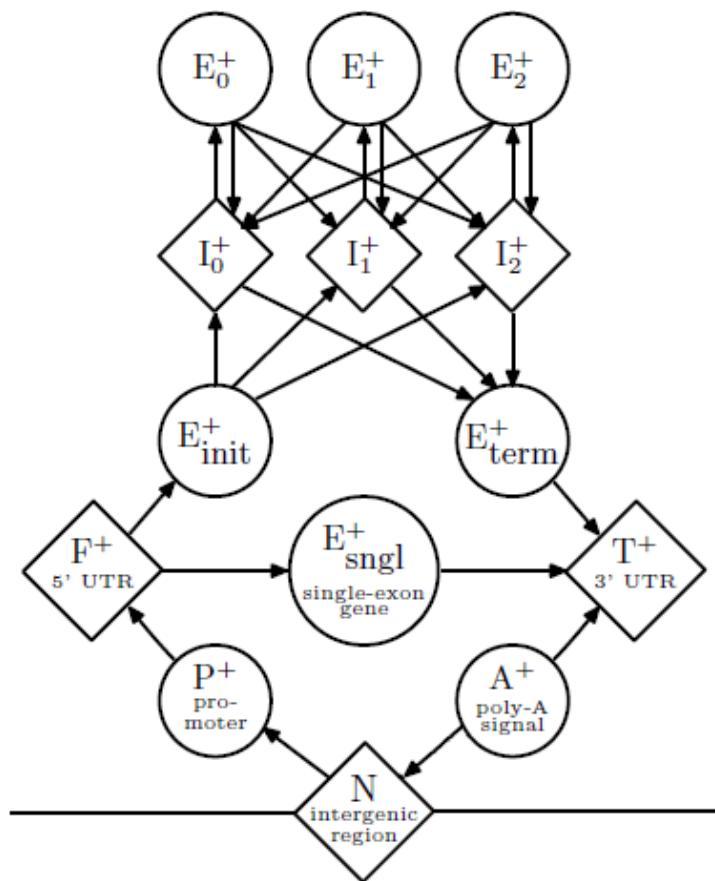


Figure 8.8: State Space Diagram used in GENSCAN

## 8.7 Further Reading

## 8.8 Tools and Techniques

## 8.9 What Have We Learned?

Using the basic computational framework provided by Hidden Markov Models, we've learned how to infer the most likely set of hidden states underlying a sequence of observed characters. In particular, a combination of the forward and backward algorithms enabled one form of this inference, i.e. posterior decoding, in  $O(KN^2)$  time. We also learned how either unsupervised or supervised learning can be used to identify the best parameters for an HMM when provided with an unlabelled or labelled dataset. The combination of these decoding and parameter estimation methods enable the application of HMM's to a wide variety of problems in computational biology, of which CpG island and gene identification form a small subset. Given the flexibility and analytical power provided by HMM's, these methods will play an important role in computational biology for the foreseeable future.

## Bibliography

- [1] Christopher B Burge and Samuel Karlin. Finding the genes in genomic dna. *Current Opinion in Structural Biology*, 8(3):346 – 354, 1998.

## GENE IDENTIFICATION: GENE STRUCTURE, SEMI-MARKOV, CRFS

Tim Helbig (2011)  
Jenny Cheng (2010)

### Figures

---

9.1	Intergenic DNA . . . . .	178
9.2	Intron/Exon Splicing . . . . .	178
9.3	Delineation of Exons and Open Reading Frames . . . . .	179
9.4	Hidden Markov Model Utilizing GT Donor Assumption . . . . .	179
9.5	Multiple lines of evidence for gene identification . . . . .	179
9.6	HMMs with composite emissions . . . . .	180
9.7	State diagram that considers direction of RNA translation . . . . .	180
9.8	Conditional random fields: a discriminative approach conditioned on the input sequence . . . . .	181
9.9	Examples of feature functions . . . . .	181
9.10	Conditional probability score of an emitted sequence . . . . .	181
9.11	A comparison of HMMs and CRFs . . . . .	182

---

### 9.1 Introduction

After a genome has been sequenced, a common next step is to attempt to infer the functional potential of the organism or cell encoded through careful analysis of that sequence. This mainly takes the form of identifying the protein coding genes within the sequence as they are thought to be the primary units of function within living systems; this is not to say that they are the only functional units within genomes as things such as regulatory motifs and non-coding RNAs are also imperative elements.

This annotation of the protein coding regions is too laborious to do by hand, so it is automated in a process known as computational gene identification. The algorithms underlying this process are often based on Hidden Markov Models (HMMs), a concept discussed in previous chapters to solve simple problems such as knowing whether a casino is rolling a fair versus a loaded die. Genomes, however, are very complicated sets of data, replete with long repeats, overlapping genes (where one or more nucleotides are part of two or more distinct genes) and pseudogenes (non-transcribed regions that look very similar to genes) among many other obfuscations. Thus, experimental and evolutionary data often needs to be included into HMMs for greater annotational accuracy, which can result in a loss of scalability or a reliance on incorrect assumptions of independence. Alternative algorithms have been utilized to address the problems of HMMs including those based on Conditional Random Fields (CRFs), which rely on creating a distribution of the hidden states of the genomic sequence in question conditioned on known data. Use of CRFs has not phased out HMMs as both are used with varying degrees of success in practice.<sup>1</sup>

---

<sup>1</sup> R. Guigo (1997). “Computational gene identification: an open problem.” Computers Chem. Vol. 21.

## 9.2 Overview of Chapter Contents

This chapter will begin with a discussion of the complexities of the Eukaryotic gene. It will then describe how HMMs can be used as a model to parse Eukaryotic genomes into protein coding genes and regions that are not; this will include reference to the strengths and weaknesses of an HMM approach. Finally, the use of CRFs to annotate protein coding regions will be described as an alternative.

## 9.3 Eukaryotic Genes: An Introduction

Within eukaryotic genomes, only a small fraction of the nucleotide content actually consists of protein coding genes (in humans, protein coding regions make up about 1%-1.5% of the entire genome). The rest of the DNA is classified as intergenic regions (See Figure 9.1) and contains things such as regulatory motifs, transposons, integrins and non-protein coding genes.<sup>2</sup>

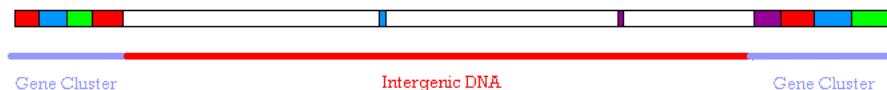


Figure 9.1: Intergenic DNA

Further, of the small fraction of the DNA that is transcribed into mRNA, not all of it is translated into protein. Certain regions known as introns, are removed or “spliced” out of the precursor mRNA. This now processed mRNA, containing only “exons” and some other additional modifications discussed in previous chapters, is translated into protein. (See Figure 9.2) The goal of computational gene identification is thus not only to pick out the few regions of the entire Eukaryotic genome that encode for proteins but also to parse those protein coding regions into identities of exon or intron so that the sequence of the synthesized protein can be known.

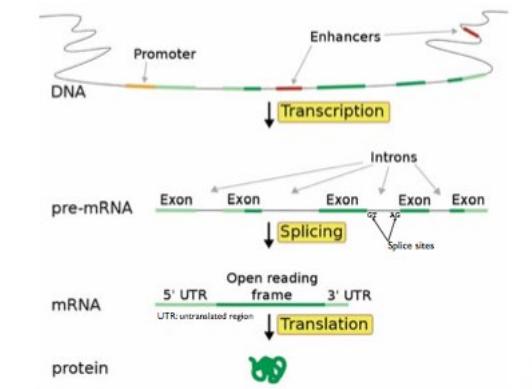


Figure 9.2: Intron/Exon Splicing

## 9.4 Assumptions for Computational Gene Identification

The general assumptions for computational gene identification are that exons are delineated by a sequence AG at the start of the exon and a sequence of GT at the end of the exon. For protein-coding genes, the start codon (ATG) and the end codons (TAA, TGA, TAG) delineate the open reading frame. (Most of these ideas can be seen in Figure 9.3) These assumptions will be incorporated into more complex HMMs described below.

<sup>2</sup>“Intergenic region.” [http://en.wikipedia.org/wiki/Intergenic\\_region](http://en.wikipedia.org/wiki/Intergenic_region)

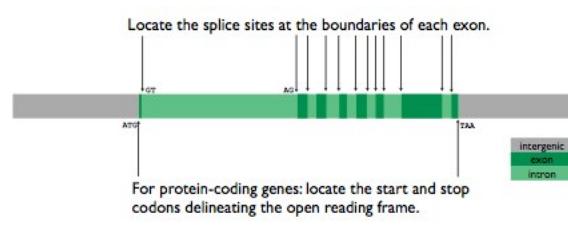


Figure 9.3: Delineation of Exons and Open Reading Frames

## 9.5 Hidden Markov Models

A toy Hidden Markov Model is a generative approach to model this behavior. Each emission of the HMM is one DNA base/letter. The hidden states of the model are intergenic, exon, intron. Improving upon this model would involve including hidden states DonorG and DonorT. The DonorG and DonorT states utilize the information that exons are delineated by GT at the end of the sequence before the start of an intron. (See Figure 9.4 for inclusion of DonorG and DonorT into the model)

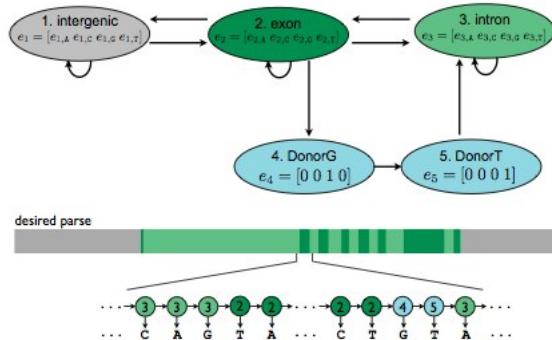


Figure 9.4: Hidden Markov Model Utilizing GT Donor Assumption

The  $e$  in each state represents emission probabilities and the arrows indicate the transition probabilities. Aside from the initial assumptions, additional evidence such as evolutionary conservation and experimental mRNA data can help create an HMM to better model the behavior. (See Figure 9.5)

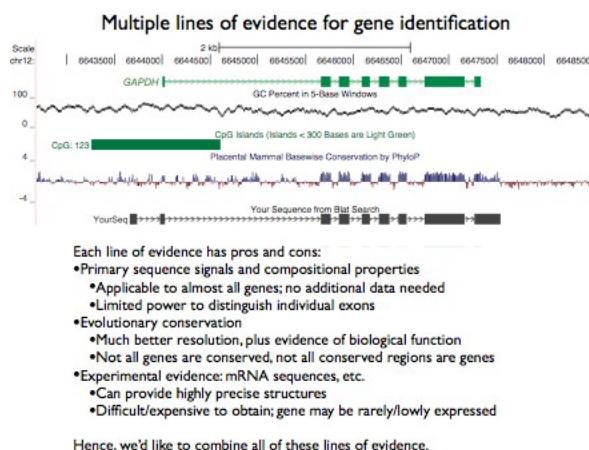


Figure 9.5: Multiple lines of evidence for gene identification

Combining all the lines of evidence discussed above, we can create an HMM with composite emissions in that each emitted value is a “tuple” of collected values. (See Figure 9.6)

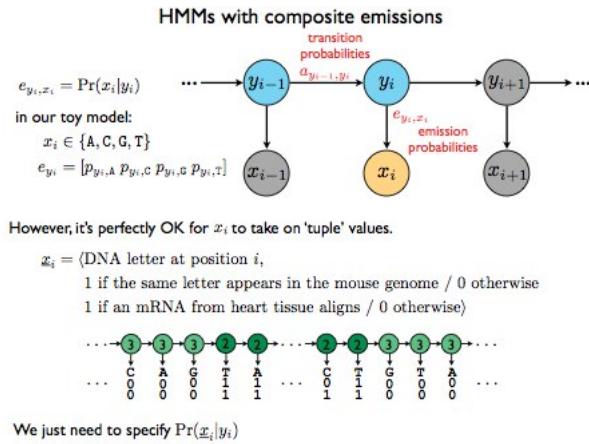


Figure 9.6: HMMs with composite emissions

A few assumptions of this composite model are that each new emission “feature” is independent of the rest. However, this creates the problem that with each new feature, the tuple increases in length, and the number of states of the HMM increases exponentially, leading to a combinatorial explosion, which thus means poor scaling. (Examples of more complex HMMs that can result in poor scaling can be found in Figure 9.7)

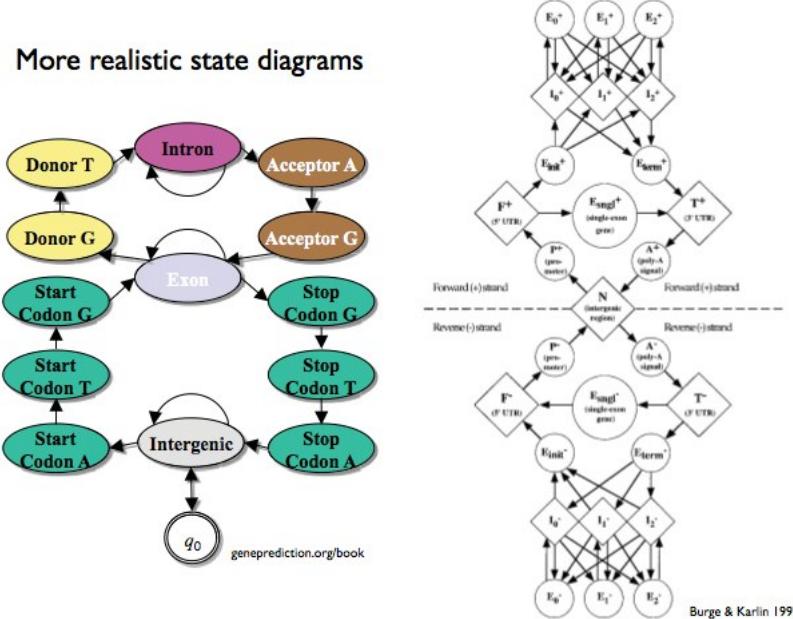


Figure 9.7: State diagram that considers direction of RNA translation

## 9.6 Conditional Random Fields

Conditional Random Fields, CRFs, are an alternative to HMMs. Being a discriminative approach, this type of model doesn't take into account the joint distribution of everything, as does a poorly scaling HMM. The hidden states in a CRF are conditioned on the input sequence. (See Figure 9.8)<sup>3</sup>

<sup>3</sup>Conditional Random Field. Wikipedia. [http://en.wikipedia.org/wiki/Conditional\\_random\\_field](http://en.wikipedia.org/wiki/Conditional_random_field)

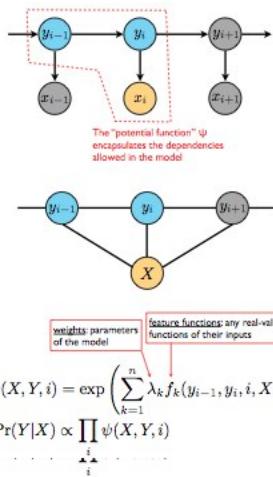


Figure 9.8: Conditional random fields: a discriminative approach conditioned on the input sequence

A feature function is like a score, returning a real-valued number as a function of its inputs that reflects the evidence for a label at a particular position. (See Figure 9.9) The conditional probability of the emitted sequence is its score divided by the total score of the hidden state. (See Figure 9.10)

$$f_1(y_{i-1}, y_i, i, X) = \begin{cases} 1 & \text{if } y_i = \text{exon and position } i \text{ is conserved in mouse} \\ 0 & \text{otherwise} \end{cases}$$

$$f_2(y_{i-1}, y_i, i, X) = \begin{cases} 1 & \text{if } y_i = \text{exon and position } i \text{ is conserved in rat} \\ 0 & \text{otherwise} \end{cases}$$

$$f_3(y_{i-1}, y_i, i, X) = \# \text{ mRNA sequences aligning to position } i \text{ (if } y_i = \text{exon; 0 otherwise)}$$

Figure 9.9: Examples of feature functions

$$\Pr(Y|X) = \frac{1}{Z(X)} \prod_i \psi(X, Y, i) \quad \text{where} \quad Z(X) = \sum_{Y'} \prod_i \psi(X, Y', i)$$

Figure 9.10: Conditional probability score of an emitted sequence

Each feature function is weighted, so that during the training, the weights can be set accordingly.

The feature functions can incorporate vast amounts of evidence without the Naive Bayes assumption of independence, making them both scalable and accurate. However, training is much more difficult with CRFs than HMMs.

## 9.7 Other Methods

Besides HMMs and CRFs, other methods do exist for computational gene identification. Semi-markov models generate variable sequence length emissions, meaning that the transitions are not entirely memory-less on the hidden states.

Max-min models are adaptations of support vector machines. These methods have not yet been applied to mammalian genomes.<sup>4</sup>

<sup>4</sup>For better understanding of SVM: <http://dspace.mit.edu/bitstream/handle/1721.1/39663/6-034Fall-2002/OcwWeb/Electrical-Engineering-and-Computer-Science/6-034Artificial-IntelligenceFall2002/Tools/detail/svmachine.htm>

## 9.8 Conclusion

Computational gene identification, because it entails finding the functional elements encoded within a genome, has a lot of practical significance as well as theoretical significance for the advancement of biological fields.

The two approaches described above are summarized below in Figure 9.11:

A comparison	
HMM	CRF
$\psi(X, Y, i) = a_{y_{i-1}, y_i} \cdot e_{y_i, x_i}$	$\psi(X, Y, i) = \exp \left( \sum_{k=1}^n \lambda_k f_k(y_{i-1}, y_i, i, X) \right)$
$\Pr(X, Y) = \prod_i \psi(X, Y, i)$	
$\Pr(Y X) = \frac{\Pr(X, Y)}{\Pr(X)}$ <small>(Bayes' law)</small>	$\Pr(Y X) = \frac{1}{Z(X)} \prod_i \psi(X, Y, i)$
$\Pr(Y X) = \frac{1}{\Pr(X)} \prod_i \psi(X, Y, i)$	
$\Pr(X) = \sum_Y \prod_i \psi(X, Y, i)$	$Z(X) = \sum_Y \prod_i \psi(X, Y, i)$
<p style="text-align: center;"><small>Q: How do we compute this efficiently? A: Forward algorithm. CRFs have a direct analog (Viterbi too)</small></p>	
$\lambda_1 = 1, \quad f_1(y_{i-1}, y_i, i, x) = \log(a_{y_{i-1}} \cdot e_{y_i, x}) \quad \Rightarrow \quad \{\text{HMM}\} \subset \{\text{CRF}\}$	

Figure 9.11: A comparison of HMMs and CRFs

### 9.8.1 HMM

- generative model
- randomly generates observable data, usually with a hidden state
- specifies a joint probability distribution
- $P(x, y) = P(x|y)P(y)$
- sometimes hard to model dependencies correctly
- hidden states are the labels for each DNA base/letter
- composite emissions are a combination of the DNA base/letter being emitted with additional evidence

### 9.8.2 CRF

- discriminative model
- models dependence of unobserved variable  $y$  on an observed variable  $x$
- $P(y|x)$
- hard to train without supervision
- more effective for when the model doesn't require joint distribution

In practice, the resulting gene specification using CONTRAST, a CRF implementation, is about 46.2% at its maximum. This is because in biology, there are a lot of exceptions to the standard model, such as overlapping genes, nested genes, and alternative splicing. Having models include all of those exceptions sometimes yields worse predictions; this is a non-trivial tradeoff. However, technology is improving and within the next five years, there will be more experimental data to fuel the development of computational gene identification, which in turn will help generate a better understanding of the syntax of DNA.

## 9.9 Current Research Directions

## 9.10 Further Reading

## 9.11 Tools and Techniques

## 9.12 What Have We Learned?

## Bibliography

- 1.R. Guigo (1997). “Computational gene identification: an open problem.”
- 2.“Intergenic region.” [http://en.wikipedia.org/wiki/Intergenic\\_region](http://en.wikipedia.org/wiki/Intergenic_region)
- 3.Conditional Random Field. Wikipedia. [http://en.wikipedia.org/wiki/Conditional\\_random\\_field](http://en.wikipedia.org/wiki/Conditional_random_field)
- 4.<http://dspace.mit.edu/bitstream/handle/1721.1/39663/6-034Fall-2002/OcwWeb/Electrical-Engineering-and-Computer-Science/6-034Artificial-IntelligenceFall2002/Tools/detail/svmachine.htm>



---

CHAPTER  
**TEN**

---

## RNA MODIFICATIONS

add author

missing

### Figures

---

10.1 mRNA is not always an appropriate proxy for protein levels. . . . .	186
10.2 Discrepancy between mRNA levels and protein abundance. . . . .	187
10.3 The genetic signals which amino acids are mapped to specific three nucleotide sequences. . . . .	187
10.4 Depiction of ribosome profiles when Cyclohexamide (elongation freeze) or Harringtonine are used (initiation freeze). . . . .	189
10.5 Ribosome profile when harringtonine is used vs. no drug. The red peaks show the different places initiation of translation can start, depicting the different possible isoforms. . . . .	189
10.6 Ribosome profile when harringtonine is used vs. no drug. The red peaks previously unidentified ORFs. . . . .	190
10.7 Ribosome profile when during rich conditions and starvation conditions. This images shows the dramatic decrease in translation of proteins during starvation. The mRNA profile is not indicative of this. . . . .	190

---

### 10.1 Introduction

Many ideas in biology rely on knowing the protein levels in a cell. Protein abundance is often extrapolated from corresponding mRNA levels. This extrapolation is made as it is relatively easy to measure mRNA levels. In addition, for a long time, it was thought that all of the regulation of expression occurred prior to mRNA formation. Now, it is known that expression continues to be regulated at the translation stage. Figure 1 shows that the data available for post-transcriptional regulation is minimal and illustrates an example of how mRNA levels are not indicative of protein abundance.

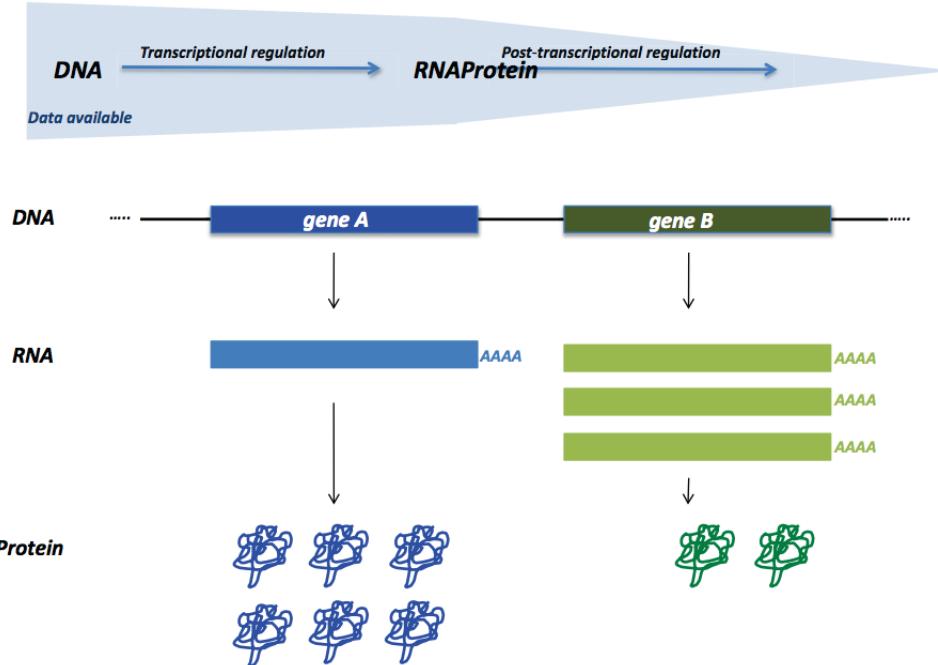


Figure 10.1: mRNA is not always an appropriate proxy for protein levels.

There are many factors that may be affecting how mRNA is translated, causing mRNA level to not be directly related to protein levels. These factors include:

1. **Translation elongation rates**
  - depends on codon usage bias, tRNA adaptation, and RNA editing
2. **Translation initiation rates**
  - depends on AUG frequency, TOP presence, type of initiation (cap-dependent/IRES), and secondary structures
3. **Translation termination rates**
  - depends on termination codon identity
4. **mRNA degradation rates**
  - depends on polyA tail length, capping, mRNA editing, and secondary structure
5. **Protein degradation rates**
  - depends on PEST sequences, protein stability, unstructured regions, and the presence of polar amino acids
6. **Cis and Trans regulatory elements**
  - depends on AU-rich elements, miRNAs, ncRNAs, and RNA-binding proteins

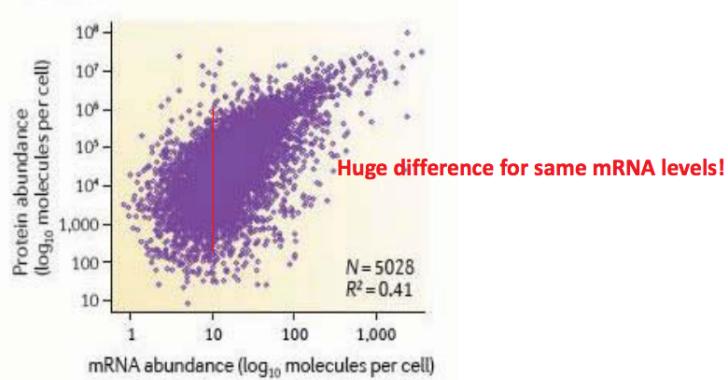


Figure 10.2: Discrepancy between mRNA levels and protein abundance.

## 10.2 Post-Transcriptional Regulation

### 10.2.1 Basics of Protein Translation

For the basics of transcription and translation, refer to Lecture 1, sections 4.3 - 4.5.

		Second letter					
		U	C	A	G		
First letter	U	UUU UUC UUA UUG	UCU UCC UCA UCG	UAU UAC UAA UAG	UGU UGC UGA UGG	U C A G	
	C	CUU CUC CUA CUG	CCU CCC CCA CCG	CAU CAC CAA CAG	CGU CGC CGA CGG	U C A G	
	A	AUU AUC AUA AUG	ACU ACC ACA ACG	AAU AAC AAA AAG	AGU AGC AGA AGG	U C A G	
	G	GUU GUC GUA GUG	GCU GCC GCA GCG	GAU GAC GAA GAG	GGU GGC GGA GGG	U C A G	
		Third letter					

Figure 10.3: The genetic signals which amino acids are mapped to specific three nucleotide sequences.

The genetic code is almost universal.

Add reference figure.

### FAQ

**Q:** Why is genetic code so similar across organisms?

**A:** Genomic material is not only transmitted vertically (from parents) but also horizontally between organisms. This gene interaction creates an evolutionary pressure for an universal genetic code.

**FAQ**

**Q:** What accounts for the slight differences in the genetic code across organisms?

**A:** Late/early evolutionary arrival of amino acids can account for the differences. Also, certain species (e.g. bacteria in deep sea vents) have more resources to synthesize specific amino acids, thus they will favor those in the genetic code.

**Did You Know?**

Threonine and Alanine are often accidentally interchanged by tRNA synthetase because they originated from one amino acid.

### 10.2.2 Measuring Translation

Translation efficiency is defined as,

$$T_{eff} = \frac{[\text{mRNA}]}{[\text{protein}]}$$

We are interested in seeing just how much of our mRNA is translated to protein, i.e. the efficiency. However, specifically measuring how much mRNA becomes protein is a difficult task, one that requires a bit of creativity. There are a variety of ways to tackle this problem, but each has its own downfalls:

1. **Measure mRNA and protein levels directly**

Pitfall: Does not consider rates of synthesis and degradation. This method measures the protein levels for the 'old' mRNA since there is a time lag from mRNA to protein.

2. **Use drugs to inhibit transcription and translation**

Pitfall: Drugs have side effects altering translation

3. **Artificial fusion of proteins with tags**

Pitfall: Protein tags can affect protein stability

4. **Pulse label with radioactive nucleosides or amino acids (SILAC)** \*\*in use today\*\*

Pitfall: Offers no information on dynamic changes: it is simply a snapshot of the resulting mRNA and protein levels after X hours

Another common technique is using 'ribosome profiling' to measure protein translation at subcodon resolution. This is done by freezing ribosomes in the process of translation and degrading the non-ribosome protected sequences. At this point, the sequences can be pieced back together and the frequency with which a region is translated can be interpolated. The disadvantage to using these ribosome footprints, to see which regions are being translated, is that regions in between ribosomes are lost. This technique requires an RNA-seq in parallel.

The question remains, why is Ribosome profiling advantageous? This technique is a better approach to measuring protein abundance as it:

1. Is a **better measure of protein abundance**

2. Is **independent of protein degradation** (compared to the protein abundance/mRNA ratio)

3. Allows us to **measure codon-specific translation rates**

of ribosome  
protein

Using ribosome profiling, it is possible to see which codon is being decoded: this is done by mapping ribosome footprints and then deciphering the translating codon based on footprint length. We can verify our prediction by mapping translated codon profiles based on periodicity (three bases in a codon). The technique can be improved even further by using anti-translation drugs such as *harringtonine* and *cyclohexamide*. Cyclohexamide blocks elongation and Harringtonine inhibits initiation. The later can be used to find the starting points (which genes are about to be translated). Figure 4 shows the effects of the drugs on the ribosome profiles.

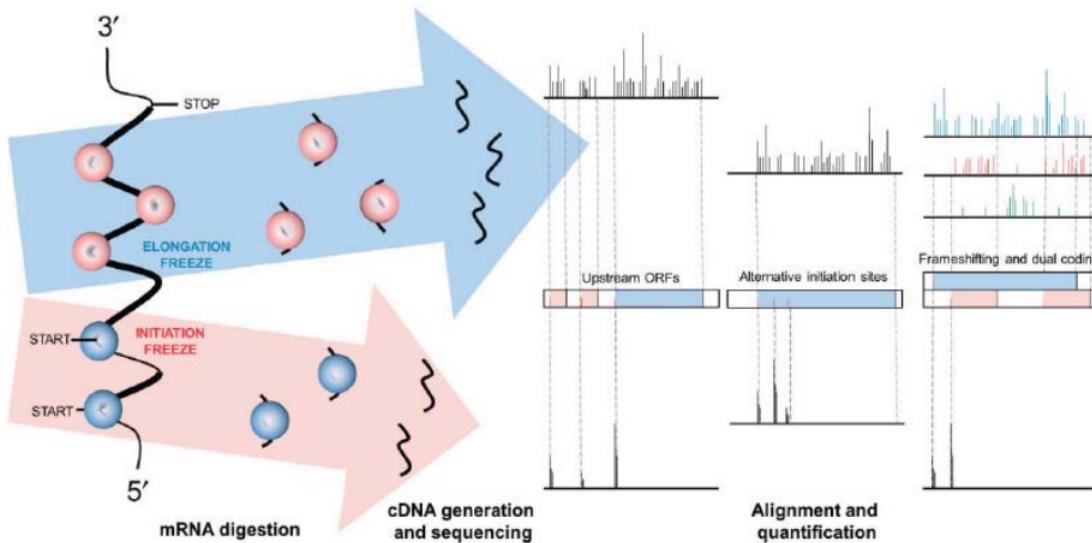


Figure 10.4: Depiction of ribosome profiles when Cyclohexamide (elongation freeze) or Harringtonine are used (initiation freeze).

This technique has much more to offer than simply quantifying translation. Ribosome profiling allows for:

1. *Prediction of alternative isoforms* (different places where translation can start)

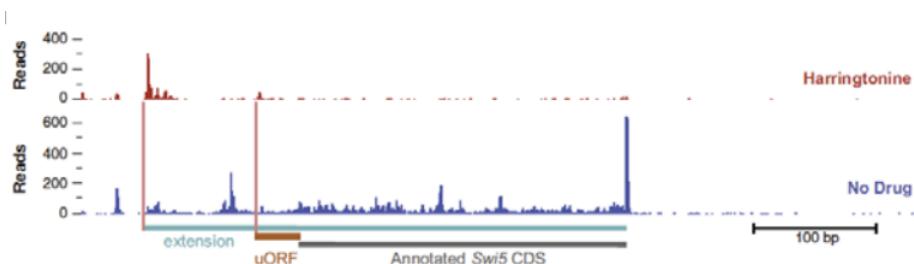


Figure 10.5: Ribosome profile when harringtonine is used vs. no drug. The red peaks show the different places initiation of translation can start, depicting the different possible isoforms.

2. *Prediction of un-identified ORFs* (open reading frames)

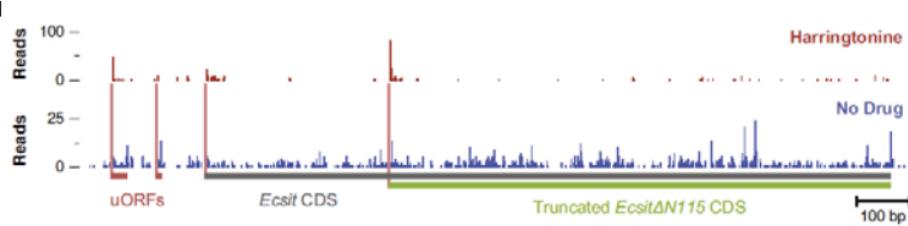


Figure 10.6: Ribosome profile when harringtonine is used vs. no drug. The red peaks previously un-identified ORFs.

### 3. Comparing translation across different environmental conditions

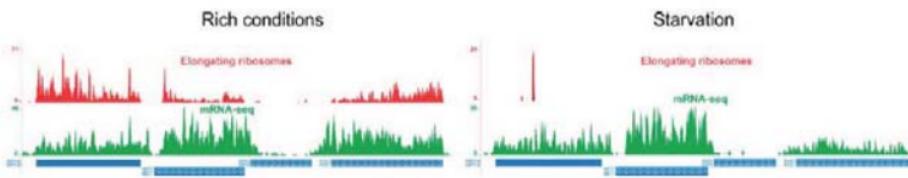


Figure 10.7: Ribosome profile when during rich conditions and starvation conditions. This images shows the dramatic decrease in translation of proteins during starvation. The mRNA profile is not indicative of this.

### 4. Comparing translation across life stages

→ to Figures

Thus, we see that ribosome profiling is a very powerful tool with lots of potential to reveal previously elusive information about the translation of a genome.

#### 10.2.3 Codon Evolution

##### Basic concepts

Something to make clear is that codons are *not* used with equal frequencies. In fact, which codons can be considered optimal differs across different species based on RNA stability, strand-specific mutation bias, transcriptional efficacy, GC composition, protein hydrophathy, and translational efficiency. Likewise, tRNA isoacceptors are not used with equal frequencies within and across species. The motivation for the next section is to determine how we may measure this codon bias.

##### Measures of Codon Bias

There are a few methods to accomplish this task:

- Calculate the frequency of optimal codons, which is defined as “optimal” codons/ sum of “optimal” and “non-optimal” codons. The limitations to this method are that this requires knowing which codon is recognized by each tRNA and it assumes that tRNA abundance is highly correlated with tRNA gene copy number.
- Calculate a codon bias index. This measures the rate of optimal codons with respect to the total codons encoding for that same amino acid. However, in this case the number of optimal codons are normalized with respect to the expected random usage.  $CBI = (o_{opt} - e_{rand}) / (o_{tot} - e_{rand})$ . The limitation of this method is that it requires a reference set of proteins, such as highly expressed ribosomal proteins.

- c) Calculate a codon adaptation index. This measures the relative adaptiveness or deviation of the codon usage of a gene towards the codon usage of a reference set of proteins, i.e. highly expressed genes. It is defined as the geometric mean of the relative adaptiveness values, measured as weights associated to each codon over the length of the gene sequence (measured in codons). Each weight is computed as the ratio between the observed frequency of a given codon and the frequency of its corresponding amino acid. The limitation to this approach is that it requires the definition of a reference set of proteins, just as the last method did.
- d) Calculate the effective number of codons. This measures the total number of different codons used in a sequence, which measures the bias toward the use of a smaller subset of codons, away from equal use of synonymous codons.  $N_c = 20$  if only one codon is used per amino acid, and  $N_c = 61$  when all possible synonymous codons are used equally. The steps to the process are to compute the homozgosity for each amino acid as estimated from the squared codon frequencies, obtain effective number of codons per amino acid, and compute the overall number of effective codons. This method is advantageous because it does not require any knowledge of tRNA-codon pairing, and it does not require any reference set. However, it is limited in that it does not take into account the tRNA pool.

- e) Calculate the tRNA adaptation index. Assume that tRNA gene copy number has a high positive correlation with tRNA abundance within the cell. This then measures how well a gene is adapted to the tRNA pool.

It is important to distinguish among when to use each index. The situation in which a certain index is favorable is very context-based, and thus it is often preferable to use one index above all others when the situation calls for it. By carefully choosing an index, one can uncover information about the frequency by which a codon is translated to an amino acid.

## RNA Modifications

The story becomes more complicated when we consider modifications that can occur to RNA. For instance, some modifications can expand or restrict the wobbling capacity of the tRNA. Examples include inosine modifications and  $\text{x}^5\text{U}$  modifications. These modifications allow tRNAs to decode a codon that they could not read before. One might ask why RNA modification was positively selected in the context of evolution, and the rationale is that this allows for the increase in the probability that a matching tRNA exists to decode a codon in a given environment.

### Examples of applications

There are a few natural applications that result from our understanding of codon evolution.

- a) Codon optimization for heterologous protein expression
- b) Predicting coding and non-coding regions of a genome
- c) Predicting codon read-through
- d) Understanding how genes are decoded - studying patterns of codon usage bias along genes

### 10.2.4 Translational Regulation

There are many known means of regulation at the post-transcriptional level. These include modulation of tRNA availability, changes in mRNA, and cis -and trans-regulatory elements. First, tRNA modulation has a large impact. Changes in tRNA isoacceptors, changes in tRNA modifications, and regulation at tRNA aminoacylation levels. Changes in mRNA that affect translation include changes in mRNA modification, polyA tail, splicing, capping, and the localization of mRNA (importing to and exporting from nucleus). Cis- and trans- regulatory elements include RNA interference (i.e. siRNA and miRNA), frameshift events, and riboswitches. Additionally, many regulatory elements are still yet to be discovered!

### 10.3 Current Research Directions

### 10.4 Further Reading

### 10.5 Tools and Techniques

### 10.6 What Have We Learned?

Hopefully at the end of this chapter we have come to realize the importance in transcriptional regulation. We see that mRNA levels are not 1:1 with protein levels. Additionally, we saw that the genetic code is not universal, and what are considered preferred tRNA-codon pairs are dynamic. Likewise, synonymous mutations are not equivalent across species. We have seen how powerful the technique of ribosome profiling is, as it allows us to measure translation with subcodon resolution. Despite all this, it is possible to model translation and codon evolutions using tools to help increase translation efficiency/folding of proteins in heterologous systems, predict coding regions, understand cell type-specific translation patterns, and compare translation between healthy and disease states. Finally, by analyzing translational regulation, we see how protein levels are tuned, and we see that there are many different ways to achieve post-transcriptional regulation. Perhaps we may come to realize that there is more interconnection between these different regulation strategies than we originally thought.

## Bibliography

---

CHAPTER  
**ELEVEN**

---

## LARGE INTERGENIC NON-CODING RNAs

Guest lecture by John Rinn  
Scribed by Eli Stickgold (2010)

### Figures

---

11.1 Tuxedo Tools . . . . .	195
11.2 How spaced seeds indexing works . . . . .	196
11.3 How Burrows-Wheeler indexing works . . . . .	196
11.4 An example of a gap in alignment . . . . .	197
11.5 An example of how to use the graph to find transcripts . . . . .	197
11.6 $F_1$ . . . . .	198
11.7 $F_2$ . . . . .	198
11.8 Technical variability follows a Poisson distribution . . . . .	199
11.9 Human fibroblasts specialize via epigenetic regulation to form different skin types based on their location within the body. Research has found that the type of skin in the hands shares a remarkably similar epigenetic signature to the skin in the feet, which is also distally located. . . . .	200
11.10 Two skin cell types are analyzed for their chromatin domains. There exists a clear boundary between the lung cell type which is proximal to the body, and the foot cell type which is distal to the body. . . . .	200
11.11 Polycomb, a protein that can remodel chromatin so that epigenetic silencing of genes can take place, may be regulated by non-coding RNA such as HOTAIR. . . . .	200
11.12 lincRNAs neighbor developmental regulators . . . . .	202

---

### 11.1 Introduction

Epigenetics is the study of heritable changes in genetic expression and phenotype that do not result from a sequence of DNA. Each cell, despite having an identical copy of the genome, is able to differentiate into a specialized type. There are many biological devices for accomplishing these including DNA methylation, histone modification, and various types of RNA.

DNA methylation is a binary code that is effectively equivalent to turning a gene "on" or "off". However, often times a gene might need to be more highly expressed as opposed to just being turned on. For this, histones have tails that are subject to modification. The unique combination of these two elements on a stretch of DNA can be thought of as a barcode for cell type. Even more important is the method of their preservation during replication. In the case of DNA methylation, one appropriately methylated strand is allocated to each mother or daughter cell. By leaving one trail behind, the cell is able to fill in the gaps and appropriately methylate the other cell.

As the intermediary between DNA sequences and proteins, RNA is arguably the most versatile means of regulation. As such, they will be the focus of this chapter.

### Did You Know?

Cell types can be determined by histone modification or DNA methylation (a binary code, which relies on a euchromatic and heterochromatic state). These histone modifications can be thought of as a type of epigenetic barcode that allows cell DNA to be scanned for types. Non-coding RNAs called Large Intergenic Non-Coding RNAs (lincRNAs) are heavily involved in this process.

A quick history of RNA:

- **1975:** A lab testing relative levels of RNA and DNA in bull sperm discovers twice as much RNA as DNA.
- **1987:** After automated sequencing developed, weird non-coding RNAs are first found.
- **1988:** RNA is proved to be important for maintaining chromosome structures, via chromatin architecture
- **1990s:** A large number of experiments start to research
- **2000s:** Study shows Histone-methyl transferases depend on RNA, as RNAase causes the proteins to delocalize.

Transcription is a good proxy of what's active in the cell and what will turn into protein. Microarrays led to the discovery of twice as many non-coding genes as coding genes initially; now we know the ratio is even far higher than this.

## 11.2 Noncoding RNAs from Plants to Mammals

Basic Cycle: large RNA gets chopped up into small RNAs (siRNAs) RNA use by category:

**Protists:** RNA is used as a template to splice out DNA (RNA-dependent DNA elimination and splicing)

**mRNA and DNA in nucleus:** DNA chopped and recombined based on gaps in mRNA (“quirky phenomena”)

**Plants:** RNA-dependent RNA polymerase, where the polymerase takes template of RNA and make a copy of it, is available in plants but not humans, and can make small RNAs. Mammals have at most one copies. Very different than RNA polymerase and DNA polymerase in structure. From this, we know that plants do DNA methylation with noncoding RNA.

**Flies:** use RNAs for an RNA switch; coordinated regulation of hox gene requires noncoding RNA.

**Mammals:** Non-coding RNAs can form triple helices, guide proteins to them; chromatin-modifying complexes; involved in germ line; guide behaviour of transcription factors.

For the rest of this talk, we focus on specifically lincRNA, which we will define as RNA larger than 200 nucleotides.

### 11.2.1 Long non-coding RNAs

There are a number of different mechanisms and biological devices by which epigenetic regulation occurs. One of these is long non-coding RNAs which can be thought of as fulfilling an air traffic control function within the cell.

Long non-coding RNAs share many similar characteristics with microRNAs. They are spliced, contain multiple exons, are capped, and poly-adenylated. However, they do not have open reading frames. They look just like protein coding genes, but cannot.

They are better classified by their anatomical position:

**Antisense:** These are encoded on the opposite strand of a protein coding gene.

**Intronic:** Entirely contained within an intron of a protein coding gene.

**Bidirectional:** These share the same promoter as a protein coding gene, but are on the opposite side.

**Intergenic:** These do not overlap with any protein coding genes. Think of them as sitting blindly out in the open. They are much easier targets and will be the focus of this chapter.

## 11.3 Practical topic: RNAseq

RNA-seq is a method that utilizes next-generation sequencing technology to sequence cDNA allowing us to gain insight into the contents of RNA. The two main problems that RNA-seq addresses are (1) discover new genes such as splice isoforms of previously discovered genes and (2) uncover the expression levels of genes and transcripts from the sequencing data. Additionally, RNA-seq is also beginning to replace many traditional sequencing techniques allowing labs to perform experiments more efficiently.

### 11.3.1 How it works

The RNA-Seq machine grabs a transcript and breaks it into different fragments, where the fragments are normally distributed. With the speed that the RNA-seq can sequence these transcript fragments (or reads), there are an abundant number of reads allowing us to extract expression levels. The basic idea behind this method relies on the fact that the more abundant a transcript is, the more fragments we'll sequence from it.

The tools used to analyze RNA-Seq data are collectively known as the “Tuxedo Tools”

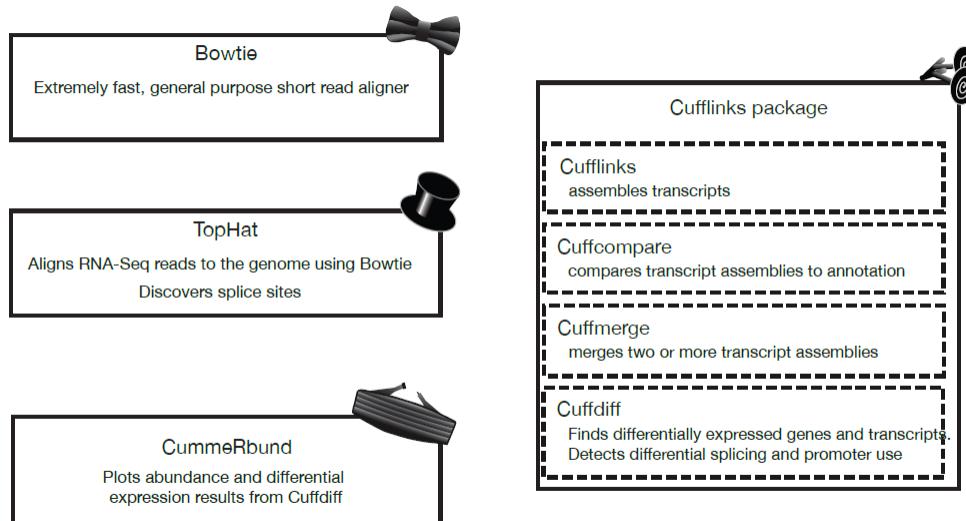


Figure 11.1: Tuxedo Tools

### 11.3.2 Aligning RNA-Seq reads to genomes and transcriptomes

Since RNA-Seq produces so many reads, the alignment algorithm must have a fast runtime, approximately of the order of  $O(n)$ . There are two main strategies for aligning short reads, which require that we already have the transcripts.

1. Spaced seeds indexing

Spaced seeds indexing involves taking each read and breaking it into fragments, or “seeds”. We take every combination of two fragments (“seed pairs”) and compare them to an index of seeds (which will

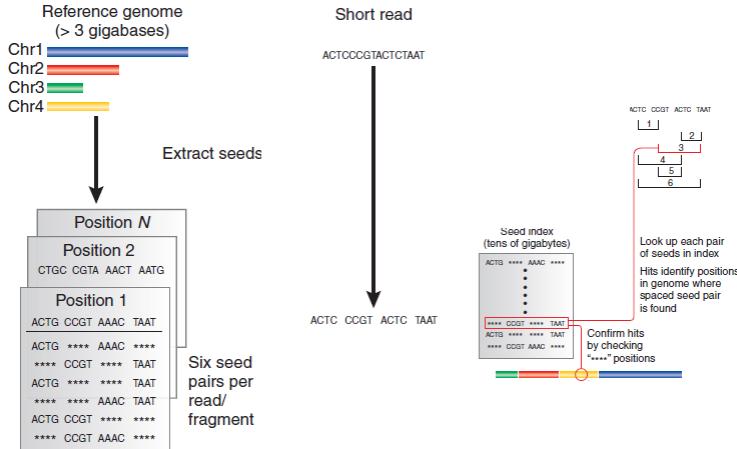


Figure 11.2: How spaced seeds indexing works

take tens of gigabytes of space) for potential hits. Compare the other seeds to the index to make sure we have a hit.

## 2. Burrows-Wheeler indexing

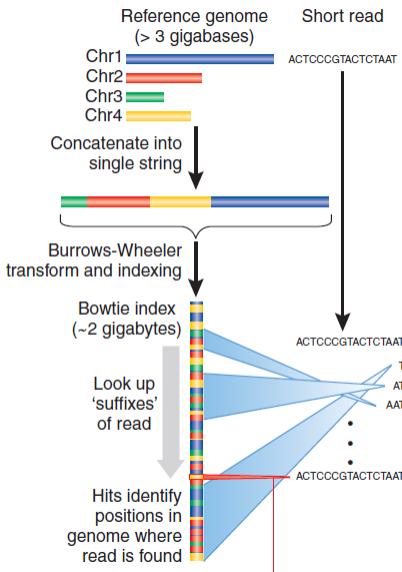


Figure 11.3: How Burrows-Wheeler indexing works

Burrows-Wheeler indexing takes the genome and scrambles it up in such a way such that you can look at the read one character at a time and throw out a huge chunk of the genome as possible alignment positions very quickly.

One major problem with these two general purpose alignment strategies is that they don't account for large gaps in alignment.

To get around this, TopHat breaks the reads into smaller pieces. These pieces are aligned and reads with pieces that are mapped far apart are flagged for possible intron sites. The pieces that weren't able to be aligned are used to confirm the splice sites. The reads are then stitched back together to make full read alignments.

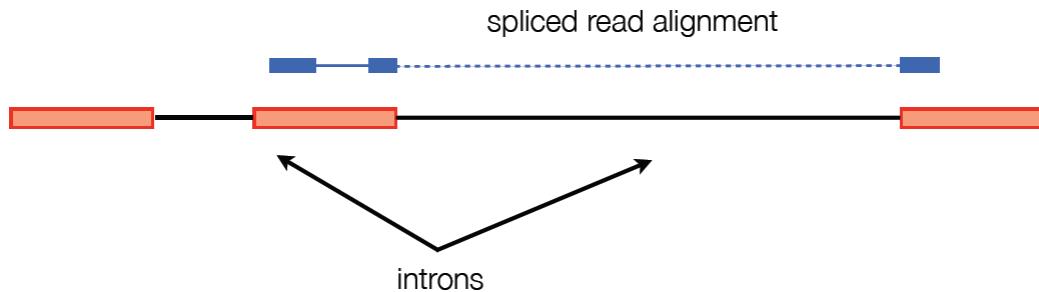


Figure 11.4: An example of a gap in alignment

There are two strategies for assembling transcripts based on RNA-Seq reads.

1. Genome-guided approach (used in software such as Cufflinks)

The idea behind this approach is that we don't necessarily know if two reads come from the same transcript, but we will know if they come from different transcripts. The algorithm is as follows: take the alignments and put them in a graph. Add an edge from  $x \rightarrow y$  if  $x$  is to the left of  $y$  in the genome,  $x$  and  $y$  overlap consistently, and  $y$  is not contained in  $x$ . So we have an edge from  $x \rightarrow y$  if they might come from the same transcript.

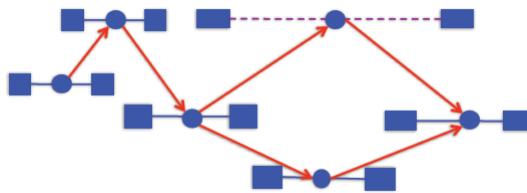


Figure 11.5: An example of how to use the graph to find transcripts

If we walk across this graph from left to right, we get a potential transcript. Applying Dilworth's theorem to read partial orders, we can see that the size of the largest antichain in the graph is the minimum number of transcripts needed to explain the alignment. An antichain is a set of alignments with the property that no two are compatible (i.e. could arise from the same transcript)

2. Genome-independent approach (used in software such as trinity)

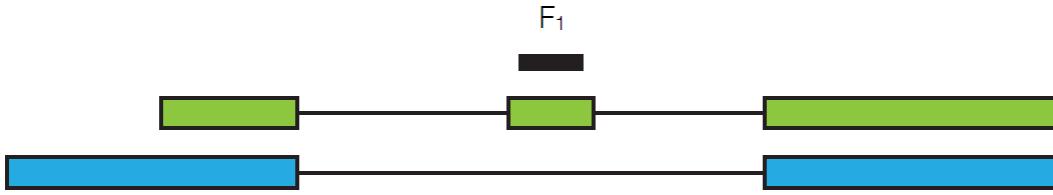
The genome-independent approach attempts to piece together the transcripts directly from the reads using classical methods for overlap based read assembly, similar to the genome assembly methods.

### 11.3.3 Calculating expression of genes and transcripts

We want to count the number of reads from each transcript to find the expression level of the transcript. However, since we divide transcripts into equally-sized fragments, we run into the problem that longer transcripts will naturally produce more reads than a shorter transcript. To account for this, we compute expression levels in FPKM, fragments per kilobase per million fragments mapped.

#### Likelihood function for a gene

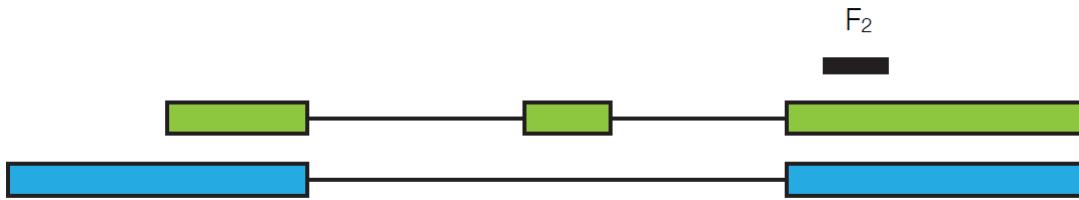
Suppose we sequence a particular read, call it  $F_1$ .

Figure 11.6:  $F_1$ 

In order to get this particular read, we need to pick the particular transcript it's in and then we need to pick this particular read out from the whole transcript. If we define  $\gamma_{\text{green}}$  to be the relative abundance of the green transcript, then we have

$$P(F_1|\gamma_{\text{green}}) = \frac{\gamma_{\text{green}}}{l_{\text{green}}}$$

where  $l_{\text{green}}$  is the length of the green transcript. Now suppose we look at a different read,  $F_2$ .

Figure 11.7:  $F_2$ 

It could have come from either the green transcript or the blue transcript, so:

$$P(F_2|\gamma) = \frac{\gamma_{\text{green}}}{l_{\text{green}}} + \frac{\gamma_{\text{blue}}}{l_{\text{blue}}}$$

We can see that the probability of getting both  $F_1$  and  $F_2$  is just the product of the individual probabilities:

$$P(F|\gamma) = \frac{\gamma_{\text{green}}}{l_{\text{green}}} \cdot \left( \frac{\gamma_{\text{green}}}{l_{\text{green}}} + \frac{\gamma_{\text{blue}}}{l_{\text{blue}}} \right)$$

We define this as our likelihood function,  $L(F|\gamma)$ . Given an input of abundances, we get a probability of how likely our sequence of reads is. So from a set of reads and transcripts, we can build a likelihood function and calculate the values for gamma that will maximize this function. Cufflinks achieves this using hill climbing or EM on the log-likelihood function.

### 11.3.4 Differential analysis with RNA-Seq

Suppose we perform an RNA-Seq analysis for a gene under two different conditions. How can we tell if there is a significant difference in the fragment counts? We calculate expression by estimating the expected number of fragments that come from each transcript. To test for significance, we need to know the variance of that estimate. We model the variance as:

$$\text{Var(expression)} = \text{Technical variability} + \text{Biological variability}$$

Technical variability, which is variability from uncertainty in mapping reads, can be modeled well with a Poisson distribution (see figure below). However, using Poisson to model biological variability, or variability across replicates, results in overdispersion.

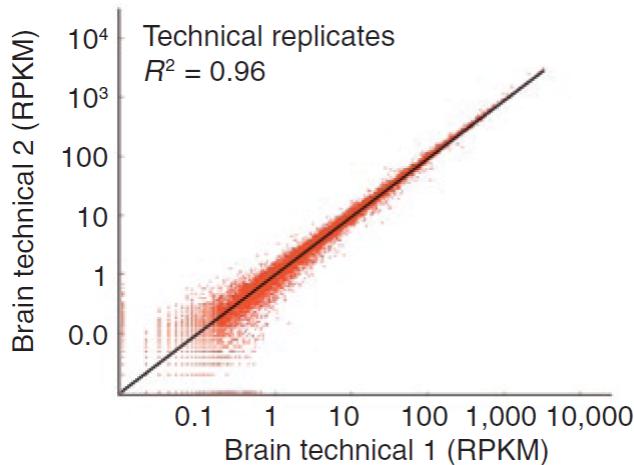


Figure 11.8: Technical variability follows a Poisson distribution

In the simple case where we have variability across replicates, but no uncertainty, we can mix the Poisson distributions from each replicate into a new distribution to model biological variability. We can treat the lambda parameter of the Poisson distribution as a random variable that follows a gamma distribution:

$$X \sim Poisson(\Gamma(r, p))$$

The counts from this model follow a negative binomial distribution. To figure out the parameters for the negative binomial for each gene, we can fit a gamma function through a scatter plot of the mean count vs. count variance across replicates.

In the simple case where there is read mapping uncertainty, but not biological variability, we need to include the mapping uncertainty in our variance estimate. Since we assign reads to transcripts probabilistically, we need to calculate the variance in that assignment.

The two threads of RNA-Seq expression analysis research focus on the problems in these two simple cases. One of the threads focuses on inferring the abundances of individual isoforms to learn about differential splicing and promoter use, while the other thread focuses on modeling variability across replicates to create more robust differential gene expression analysis. Cuffdiff unites these two separate threads to study the case where we have biological variability and read mapping ambiguity. Since overdispersion can be modeled with a negative binomial distribution and mapping uncertainty can be modeled with a Beta distribution, we combine these two to model this case with a beta negative binomial distribution.

## 11.4 Long non-coding RNAs in Epigenetic Regulation

Let's examine human skin as an example of long non-coding RNAs being used in epigenetic regulation. Human skin is huge, in fact it is the largest organ by weight in the body. It is intricate, with specialized features, and it is constantly regenerating to replace old dead cells with new ones. The skin must be controlled so hair only grows on the back of your hand rather than on your palm. Moreover, these boundaries cannot change and are maintained ever since birth.

The skin in all parts of the body is composed of an epithelial layer and a layer of connective tissue made up of cells called fibroblasts. These fibroblasts secrete cytokine signals that control the outer layer, determining properties such as the presence or absence of hair. Fibroblasts all around the body are identical except for the specific epigenetic folding that dictates what type of skin will be formed in a given location. Based on whether the skin is distal or proximal, interior or exterior, posterior or anterior, a different set of epigenetic folds will determine the type of skin that forms.

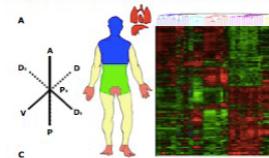


Figure 11.9: Human fibroblasts specialize via epigenetic regulation to form different skin types based on their location within the body. Research has found that the type of skin in the hands shares a remarkably similar epigenetic signature to the skin in the feet, which is also distally located.

It has been found that specific HOX genes delineate these anatomical boundaries during development. Just by looking at the human HOX genetic code, one can predict where a cell will be located. Using ChIP-on-chip (chromatin immunoprecipitation microarrays) diametric chromatin domains have been found among these HOX genes. In the figure below, we can see a clear boundary between the chromatin domains of a cell type located proximally and another located distally. Not only is this boundary precise, but it is maintained across trillions of skin cells.

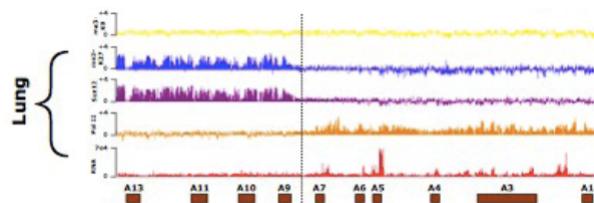


Figure 11.10: Two skin cell types are analyzed for their chromatin domains. There exists a clear boundary between the lung cell type which is proximal to the body, and the foot cell type which is distal to the body.

HOTAIR or HOX transcript antisense intergenic RNA has been investigated as possible RNA regulator that keeps these boundary between the diametric domains in chromatin. When HOTAIR was knocked out in the HOXC locus, it was hypothesized that the chromatin domains might slip through into one another. While it was found that this HOTAIR did not directly affect the epigenetic boundary, researchers did find evidence of RNA based genomic cross talk. The HOTAIR gene affected a different locus called HOXD.

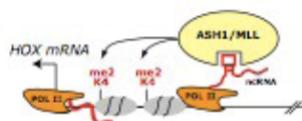


Figure 11.11: Polycomb, a protein that can remodel chromatin so that epigenetic silencing of genes can take place, may be regulated by non-coding RNA such as HOTAIR.

Through a process of ncRNA dependent Polycomb repression, the HOTAIR sequence can control epigenetic regulation. Polycomb is a protein that puts stop marks on the tails of histones so that they can cause specific folds in the genetic material. On their own histones, are undirected, so it is necessary for some mechanism to dictate how they attach to the genome. This process of discovery has led to great interest in the power of long intergenic non-coding RNAs to affect epigenetic regulation.

## 11.5 Intergenic Non-coding RNAs: missing lincs in Stem/Cancer cells?

### 11.5.1 An example: XIST

XIST was one of the first lncRNAs to be characterized. It is directly involved in deactivation of one of the female X chromosomes during embryonic development. It has been described as having the ability to "crumple an entire chromosome". This is important because deactivation prevents lethal overexpression of genes found on the X chromosome.

RNA is important for getting polychrome complex to chromosome ncRNAs can activate downstream genes in Cis, opposite in trans; Xist does the same thing.

## 11.6 Technologies: in the wet lab, how can we find these?

How would we find ncRNAs? We have about 20-30 examples of ncRNAs with evidence of importance, but more are out there. Chromatin state maps (from ENCODE, chip-seq) can be used to find transcriptional units that do not overlap proteins. We can walk along map and look for genes (look by eye at chromatin map to find ncRNAs). Nearly 90% of time such a signature is found, RNA will be transcribed from it. We can validate this through northern blot

When looking at a chromatin map to find ncRNAs, we are essentially looking through the map with a window of a given size and seeing how much signal vs. noise we are getting, compared to what we might expect from a random-chance hypothesis. As both large and small windows have benefits, both should be used on each map section. Larger windows encapsulate more information; smaller windows are more sensitive.

After finding intergenic regions, we find conserved regions.

We check if new regions are under selective pressure; fewer mutations in conserved regions. If a nucleotide never has a mutation between species, it's highly conserved.

linc-RNAs are more conserved than introns, but less conserved than protein-coding introns, possibly due to non-conserved sequences in loop regions of lincRNAs.

Finding what lncRNAs' functions are: "Guilt by association": We can find proteins that correlate with particular lncRNA in terms of expression; lncRNAs are probably correlated to a particular pathway. In this way, we acquire a 'multidimensional barcode' for each lncRNA (what it is and is not related to). We Can cluster lncRNA signatures and identify common patterns. Lots have to do with cell cycle genes. (This approach works 60-70% of the time)

As most lncRNAs are over 3000 bases, many contain sequences for 100 amino acid open reading frames, simply by chance. This results in many false negatives during detection.

It has been found that many lncRNAs tend to neighbor developmental regions of the genome. They also tend to be lowly expressed compared to protein coding genes.

### 11.6.1 Example: p53

Independent validation: we use animal models, where one is a wild-type p53, and one is a knockout. We induce p53, then ask if lncRNAs turn on. 32 of 39 lncRNAs found associated with p53 were temporally induced upon turning on p53.

One RNA in particular sat next to a protein-coding gene in the p53 pathway. We tried to figure out if p53 bound to promoter and turned it on. To do this, we cloned the promoter of lncRNA, and asked 'does p53 turn it on'? We IPed the p53 protein, to see if it associated with the lncRNA of the promoter. It turned out that lncRNA is directly related to p53 - p53 turns it on. P53 also turns genes off - certain lncRNAs act as a repressor.

From this example (and others), we start to see that RNAs usually have a protein partner

RNA can bring myriad of different proteins together, allowing the cell lots of diversity. In this way it's similar to phosphorylation. RNAs bind to important chromatin complexes, and is required for reprogramming skin cells into stem cells.

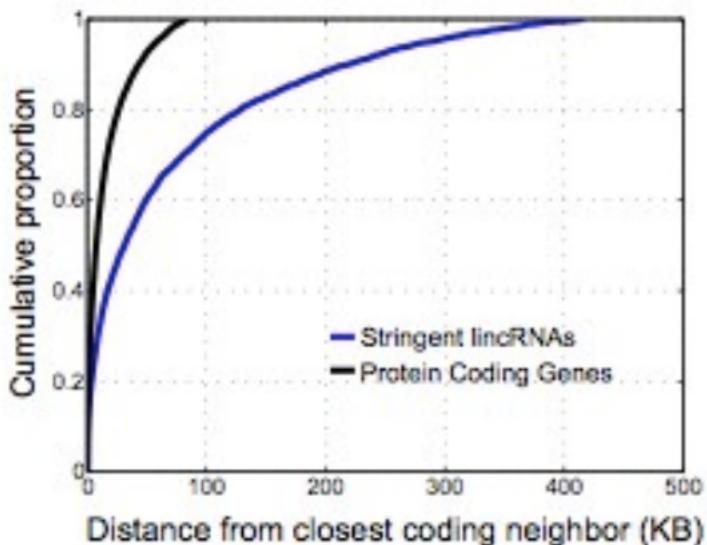


Figure 11.12: lincRNAs neighbor developmental regulators

## 11.7 Current Research Directions

## 11.8 Further Reading

## 11.9 Tools and Techniques

## 11.10 What Have We Learned?

## Bibliography

- [1] R.P. Dilworth. A decomposition theorem for partially ordered sets. *Annal of Mathematics*, 1950.
- [2] Mitchell Guttman, Manuel Garber, Joshua Z Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J Koziol, Andreas Gnirke, Chad Nusbaum, and et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nature Biotechnology*, 28(5):503–510, 2010.
- [3] C. Trapnell.





## **Part III**

# **Gene and Genome Regulation**



---

CHAPTER  
**TWELVE**

---

## MRNA SEQUENCING FOR EXPRESSION ANALYSIS AND TRANSCRIPT DISCOVERY

Guest lecture by Manuel Garber, scribed by Asa Adadey (2011)

Guest lecture by Cole Trapnell, scribed by Nirvan Tyagi (2013)

Scribed by Ang Cui (2016)

Scribed by Fangyuan Hong (2017)

### Figures

---

12.1 Expression microarray process . . . . .	208
12.2 The data generation pipeline of a RNA-sequencing experiment . . . . .	209
12.3 A high-throughput method for Illumina RNA-seq library preparation. . . . .	209
12.4 Two ways of transcript reconstruction from RNA-Seq reads . . . . .	210
12.5 Spaced k-mer method of mapping reads to reference genome . . . . .	210
12.6 Two ways of transcript reconstruction from RNA-Seq reads . . . . .	211
12.7 TopHat aligns RNA-Seq reads . . . . .	212
12.8 Reconstruction works by determining, for a particular window, the probability of observing that number of reads (top left) given the uniform distribution of the total reads (bottom left). This probability follows the Poisson distribution. . . . .	212
12.9 Process for reconstructing genome based on reads, using the scan distribution . . . . .	212
12.10 Alternative isoforms present a challenge for reconstruction, which must depend on exon junction spanning reads . . . . .	213
12.11 The Scripture method for identifying isoforms . . . . .	214
12.12 The FPKM normalizing unit representing reads per unit transcript. . . . .	214
12.13 To infer which transcript each fragment came from. Some fragments could have come from: any transcript in black, only one in blue or yellow, or subset only in purple to red or blue.	214
12.14 Overview of the de novo transcriptome assembly strategy . . . . .	215
12.15 Trinity as one representative of several transcriptome assemblers. The figure is from Lyer et al, NatBT 2011 . . . . .	216
12.16 Drosophila S2 RNA-Seq biological variability across replicates . . . . .	217
12.17 Technical variability follows a Poisson distribution . . . . .	217

---

### 12.1 Introduction

The purpose of mRNA sequencing is to measure the levels of mRNA transcripts for every gene in a given cell. mRNA sequencing was a daunting task, and requires approximately 40 million aligned reads in order to accurately measure mRNA transcripts. This did not become possible until 2009, when next-generation sequencing technologies became more advanced and efficient.

In this chapter, we will explore the different techniques for using mRNA sequencing data to aid in gene and transcript discovery as well as in expression analysis.

## 12.2 Functional Diversity of RNA

## 12.3 Microarrays for Gene Expression

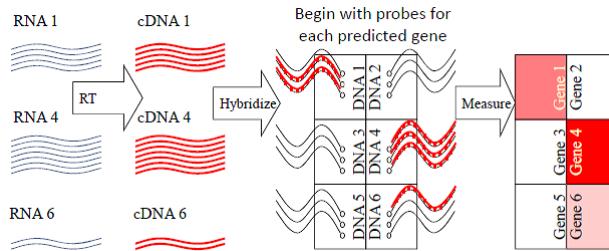


Figure 12.1: Expression microarray process

Changes in gene expression are important in biological contexts such as development and diseases. Prior to the development of mRNA sequencing technology, mRNA levels were measured using DNA microarrays, a high throughput tool to study changes in gene expression. These microarrays function by inserting a DNA probe on a slide and measuring the levels transcripts that undergo complimentary hybridization with the DNA. This process outputs expression on a gene by gene basis (Figure 12.1), thus allowing researchers to focus on small regions even with few molecules or cells.

However, microarray technology has several limitations: there could be chip flaws, RNA degradation, contamination, reverse transcription efficiency; it cannot distinguish mRNA isoforms; it does not allow for sequence analysis; it can only measure known transcripts; and the expression measurements become less reliable for highly saturated transcript levels.

## 12.4 mRNA Sequencing

### 12.4.1 RNA-seq

RNA-Seq technology sequence short reads from mRNA and map them to genome. With the advent of Illumina (Figure 12.3), high-throughput DNA sequencing is made possible. A typical RNA-sequencing experiment is illustrated in Figure 12.2. The first step in mRNA sequencing is to lyse the cells of interest. This creates a mass of proteins, nucleotides, and other molecules which are then filtered through so that only RNA (or specifically mRNA) molecules remain. The resulting transcripts are then fragmented into reads 200-1000 base pairs long and undergo a reverse transcription reaction to build a strand-specific DNA library. Finally, both ends of these DNA fragments are barcoded and sequenced.

After establishing these sequenced reads, the computational part of RNA-Seq can be divided into three parts: read mapping, reconstruction, and quantification. There are two methods to perform transcript reconstruction as illustrated in Figure 12.6. The first one is to align the reads to a reference genome. The second is to assemble transcripts *de novo*, which is useful when the reference genome is not available.

### 12.4.2 Read Mapping

The idea behind read mapping is to align the sequenced reads to a reference genome. Sequence alignment algorithms discussed in earlier chapters will not work for this case due to the scale of the problem. The goal is to align millions of reads to the genome and would take too long if each was aligned individually. Instead, we will introduce the Spaced Seed Alignment approach.

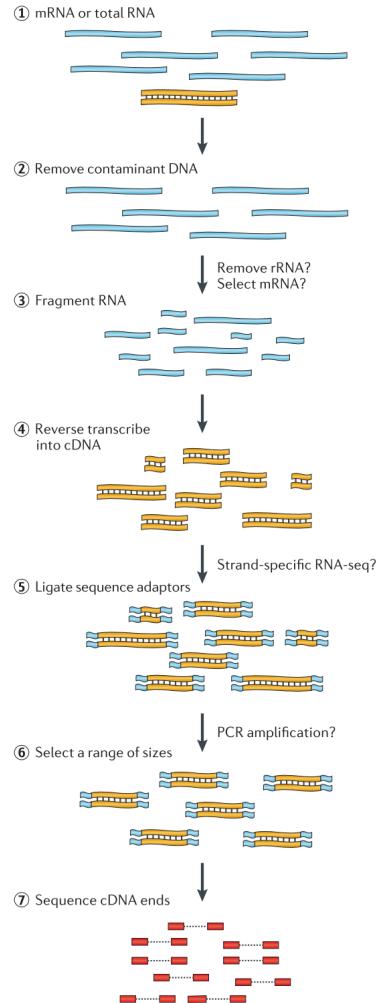


Figure 12.2: The data generation pipeline of a RNA-sequencing experiment

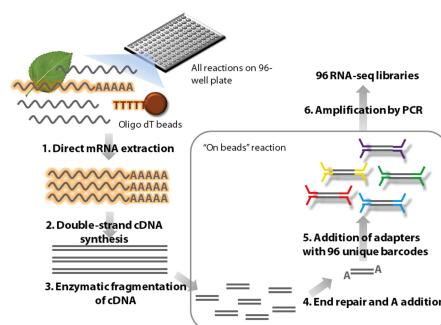


Figure 12.3: A high-throughput method for Illumina RNA-seq library preparation.

This process begins by using the reference genome to creating a hash table of 8-mers, which do not have to be contiguous. The positions of these stored spaced seeds are mapped to the hash table. Using these spaced 8-mers, each read is then compared with each possible position in the reference genome and scored based on the number of base pair matches (Figure 12.5).

More accurately, for each position, it is possible to calculate the score using the equation  $q_{MS} = -10 \log_{10}(1 - P(i|G, q))$ , where  $P(i|G, q)$  represents the probability that the read,  $q$ , is mapped to position

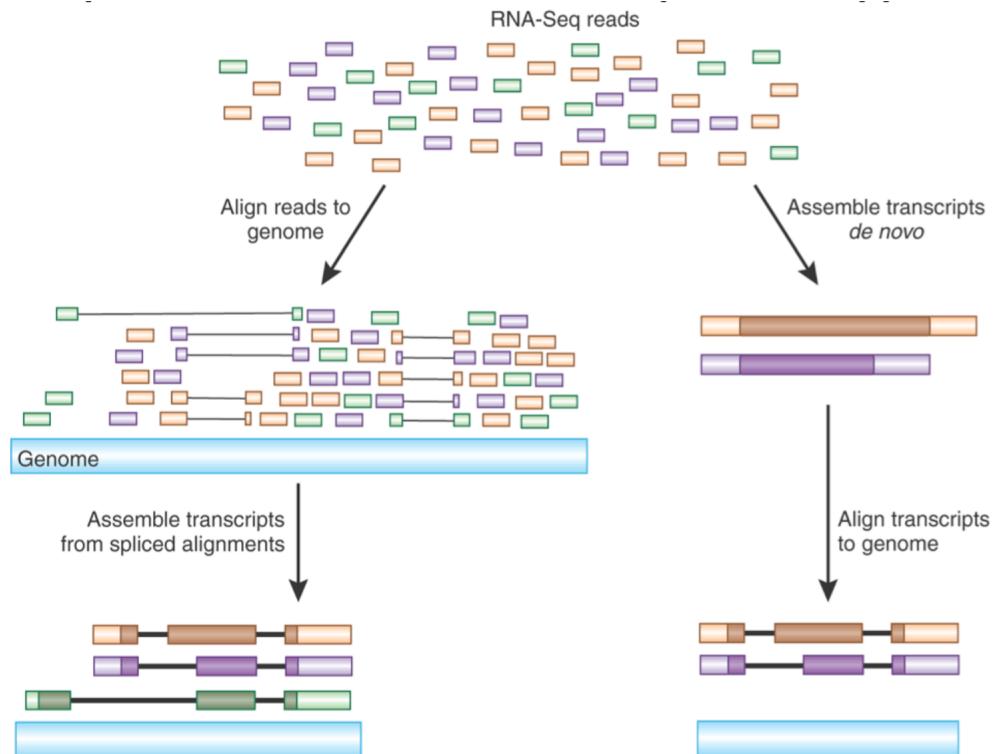


Figure 12.4: Two ways of transcript reconstruction from RNA-Seq reads

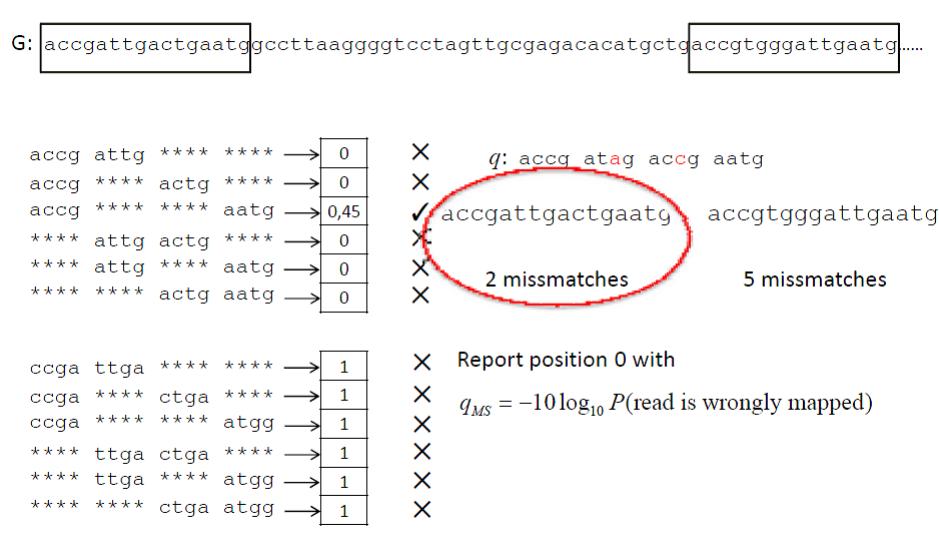


Figure 12.5: Spaced k-mer method of mapping reads to reference genome

i of reference genome G. More details on deriving this score can be found in Figure 13.2.

It is possible to adjust the parameters of this method in order to alter the sensitivity, speed, and memory of the algorithm. Using smaller k-mer seeds allows for less precise base pair matching (greater sensitivity), but requires more matches to be attempted. Smaller seeds take up less memory, while larger seeds run faster.

There exist methods other than the one described above to perform this alignment. The most popular of which is the Burrows-Wheeler approach. The Burrows-Wheeler transform is an even more efficient algorithm

for mapping reads and will be discussed in a later chapter. It is able to speed up the process of finding matches in the large genome by reordering the genome in a very specific permutation. This allows reads to be matched solely as a function of the length of the read and not the genome. As better sequencing technology allows for larger read lengths, more algorithms will need to be developed to handle the extra processing.

Unlike ChIP-Seq, RNA-Seq is more complex. This is because the read mapper needs to worry about small exons interspersed between large introns and be able to find both sides of an exon. This complexity can be overcome by using the above mentioned spaced seed matching technique, and detecting when two k-mers from the same read are separated by a long distance. This would signal a possible intron and can be fixed by then extending the k-mers to fill in gaps (SNO methods). Another method is to base the alignment on contiguous reads, which are further fragmented into 20-30 bp regions. These regions are remapped, and the positions with two or more different alignments are marked as splice junctions. Exon-first aligners are faster than the previous methods, but come at a cost: they fail to differentiate pseudogenes, prespliced genes, and transposed genes.

#### 12.4.3 RNA Transcript Reconstruction - Map to Genome Approach

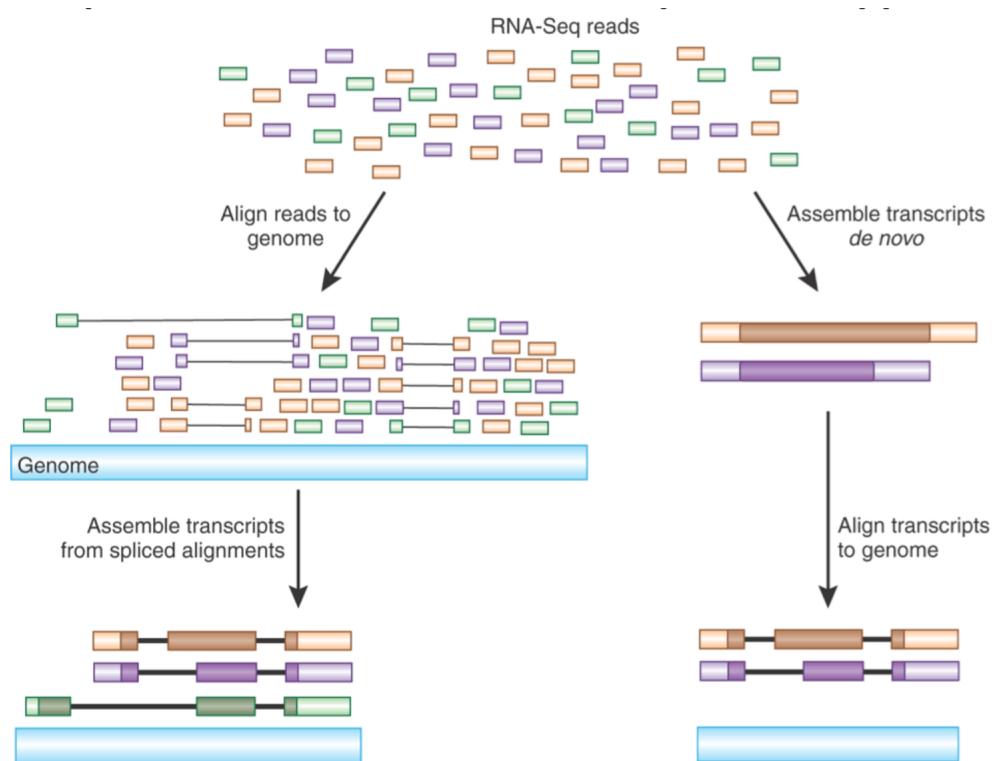


Figure 12.6: Two ways of transcript reconstruction from RNA-Seq reads

A popular map-to-genome method is the Tuxedo Suite, which consists of Bowtie, TopHat, Cufflinks, Cuffdiff and CummeRbund that perform alignment of unspliced reads, alignment of spliced reads, transcript construction, differential expression analysis and visualization, respectively.

After read mapping, we need to align RNA-Seq reads to the genome. The difficulty in aligning spliced reads arises from the fact that general alignment algorithms do not allow large gaps. To overcome this issue, special splicing-aware tools, such as TopHat illustrated in Figure 12.7, were developed. TopHat first fragments reads into little pieces and aligning these to the genome. The reads whose pieces map far apart are used to tag possible splice sites, and the pieces that did not align initially are used to confirm splice sites. Partially aligned reads are then stitched back together to make full read alignments.

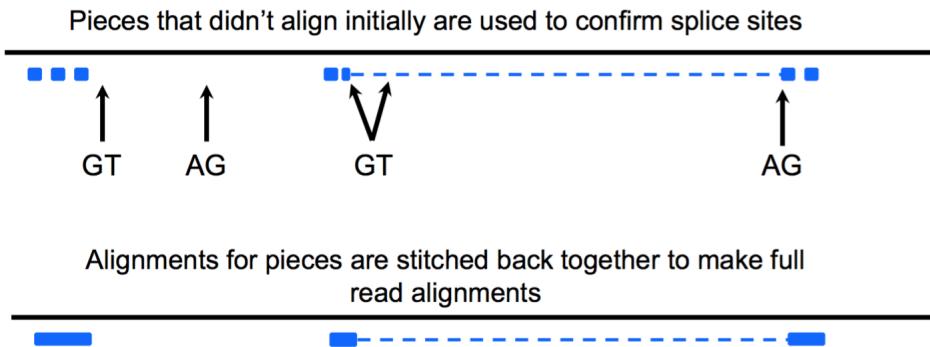


Figure 12.7: TopHat aligns RNA-Seq reads

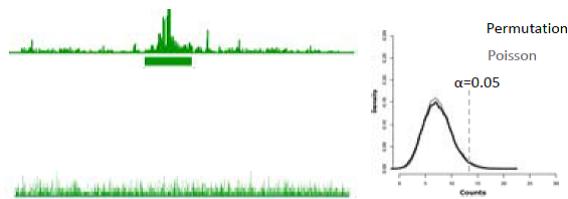


Figure 12.8: Reconstruction works by determining, for a particular window, the probability of observing that number of reads (top left) given the uniform distribution of the total reads (bottom left). This probability follows the Poisson distribution.

Reconstruction of reads is a largely statistical problem. The goal is to determine a score for each fixed-sized window in the genome. This score represents the probability of seeing the observed number of reads given the window size. In other words, is the number of reads in a particular window unlikely given the genome? The expected number of reads per window is derived from a uniform distribution based on the total number of reads (Figure 12.8). This score is modeled by a Poisson distribution.

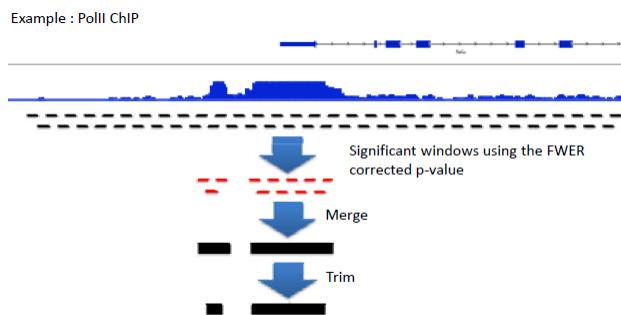


Figure 12.9: Process for reconstructing genome based on reads, using the scan distribution

However, this score must account for the problem of multiple testing hypotheses, due to the approximately 150 million expected bases. One option for dealing with this is the Bonferroni correction, where the nominal p-value =  $n * p$ -value. This method leads to low sensitivity, due to its very conservative nature. Another option is to permute the reads observed in the genome, and find the maximum number of reads seen on a single base. This allows for a max count distribution model, but the process is very slow. The scan distribution speeds up this process by computing a closed form for max count distribution to account for dependency of overlapping windows (Figure 12.9). The probability of observing k reads on a window of size w in a genome of size L given a total of N reads can be approximated by [slide is not clear].

Choosing a window size is also an important decision, as genes exist at different expression levels and span different orders of magnitude. Small windows are better at detecting punctuate regions, while larger windows can detect longer spans of moderate enhancement. In most cases, windows of different sizes are used to pick up signals of varying size.

Transcript reconstruction can be seen as a segmentation problem, with several challenges. As mentioned above, genes are expressed at different levels, over several orders of magnitude. In addition, the reads used for reconstruction are obtained from both mature and immature mRNA, the latter still containing introns. Finally, many genes have multiple isoforms, and the short nature of reads makes it difficult to differentiate between these different transcripts. A computational tool called Scripture uses a priori knowledge of fragment connectivity to detect transcripts.

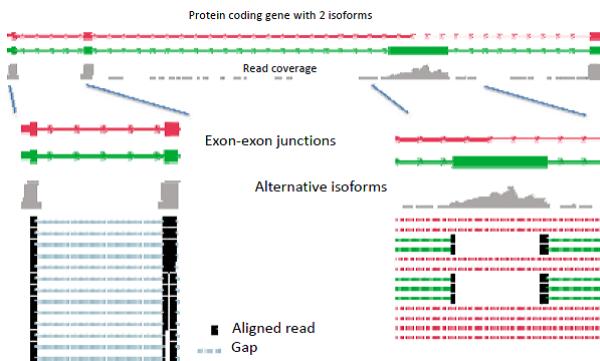


Figure 12.10: Alternative isoforms present a challenge for reconstruction, which must depend on exon junction spanning reads

Alternative isoforms can only be detected via exon junction spanning reads, which contain the ends of an exon. Longer reads have a greater chance of spanning these junctions (Figure 12.10). Scripture and Cufflinks work by modeling the reads using graph structure, where bases are connected to neighbor bases, as well as splice neighbors. This process differs from the string graph technique, because it focuses on whole genome, and does not map overlapping sequences directly. When sliding the window, these methods can jump across splice junctions yet still examine alternative isoforms. From this oriented connectivity graph, the program identifies segments across the graph, and looks for significant segments (Figure 12.11).

#### 12.4.4 Quantification

The goal of the quantification step is to score regions in the genome based on the number of reads. Recall that each transcript is fragmented into many smaller reads. Therefore, it is insufficient to simply count the number of reads per region, as this value would be influenced by (1) expression rates and (2) length of transcript. The higher the expression rate of a transcript the more reads we will have for it. Similarly, the longer a transcript is, the more reads we will have. This issue can be solved by normalizing the number of reads by the length of the transcript and the total number of reads in the experiment. This provides the FPKM/RPKM value, or reads per kilobase of exonic sequence per million mapped reads (Figure 12.15).

This method is robust for genes with only one isoform. However, there is the possibility of overlap between conflicting variants of a transcript. When multiple transcript variants are involved, this problem is known as differential expression analysis, which we will discuss in greater detail in this chapter. There are a few different methods for handling this complexity. The exon intersection model scores only the constituent exons. The exon union model simply scores based on a merged transcript, but can easily be biased based on the relative ratios of each isoform.

After aligning spliced reads, we have to resolve individual transcripts and their expression levels. Gene expression level is proportional to number of reads, but we need to correct for the total gene length. The task is to infer which splice variant/fragment each transcript is from, as illustrated in Figure 12.13. The conditional probability that a fragment came from a given isoform is essentially a function of that isoform's abundance. For fragments that could have come from multiple isoforms, we need to add up the probability

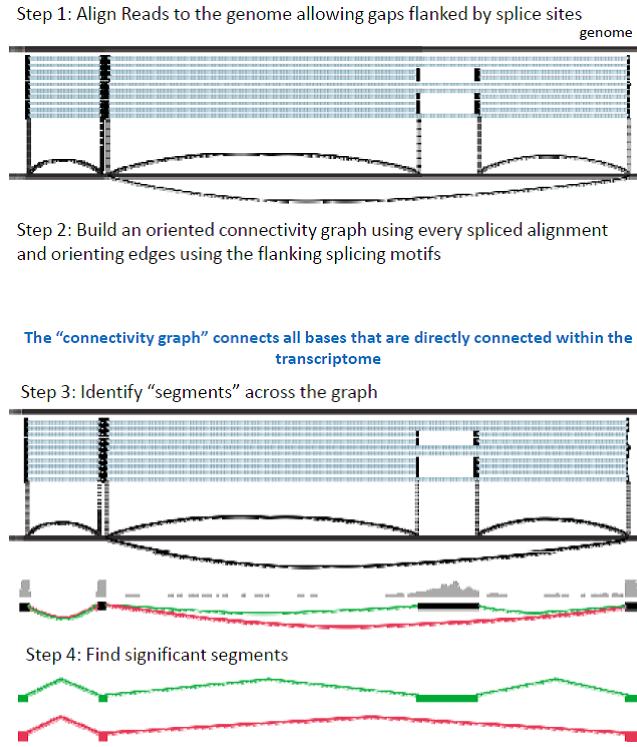


Figure 12.11: The Scripture method for identifying isoforms

$$FPKM = \frac{\text{Counts of mapped fragments}}{\text{Total mapped fragments (million)} \times \text{Exon length of transcript (KB)}}$$

Figure 12.12: The FPKM normalizing unit representing reads per unit transcript.

of getting it from each isoform. The likelihood of a given set of reads coming from a given set of isoforms is calculated by multiplying the probability of each read. Then we can use Expectation-Maximization (EM) algorithm repartition reads to find the most likely assignment of reads to transcripts. This algorithm is performed by both Cufflinks and Cuffdiff (Tuxedo), as well as RSEM in R, and eXpress.

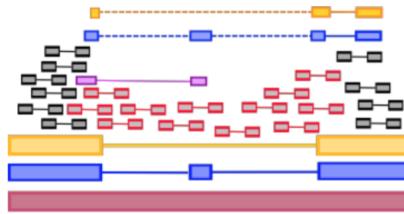


Figure 12.13: To infer which transcript each fragment came from. Some fragments could have come from: any transcript in black, only one in blue or yellow, or subset only in purple to red or blue.

### 12.4.5 The Align-de-novo Approach

In addition to map-to-genome approach, *de novo* transcript assembly is another method of reconstruction. While genome-guided approaches are ideal for annotating high quality genomes and expanding the catalog of

expressed transcripts, direct transcript assembly methods are able to reconstruct transcripts from organisms without a reference sequence, which empower the study of non-model organisms.

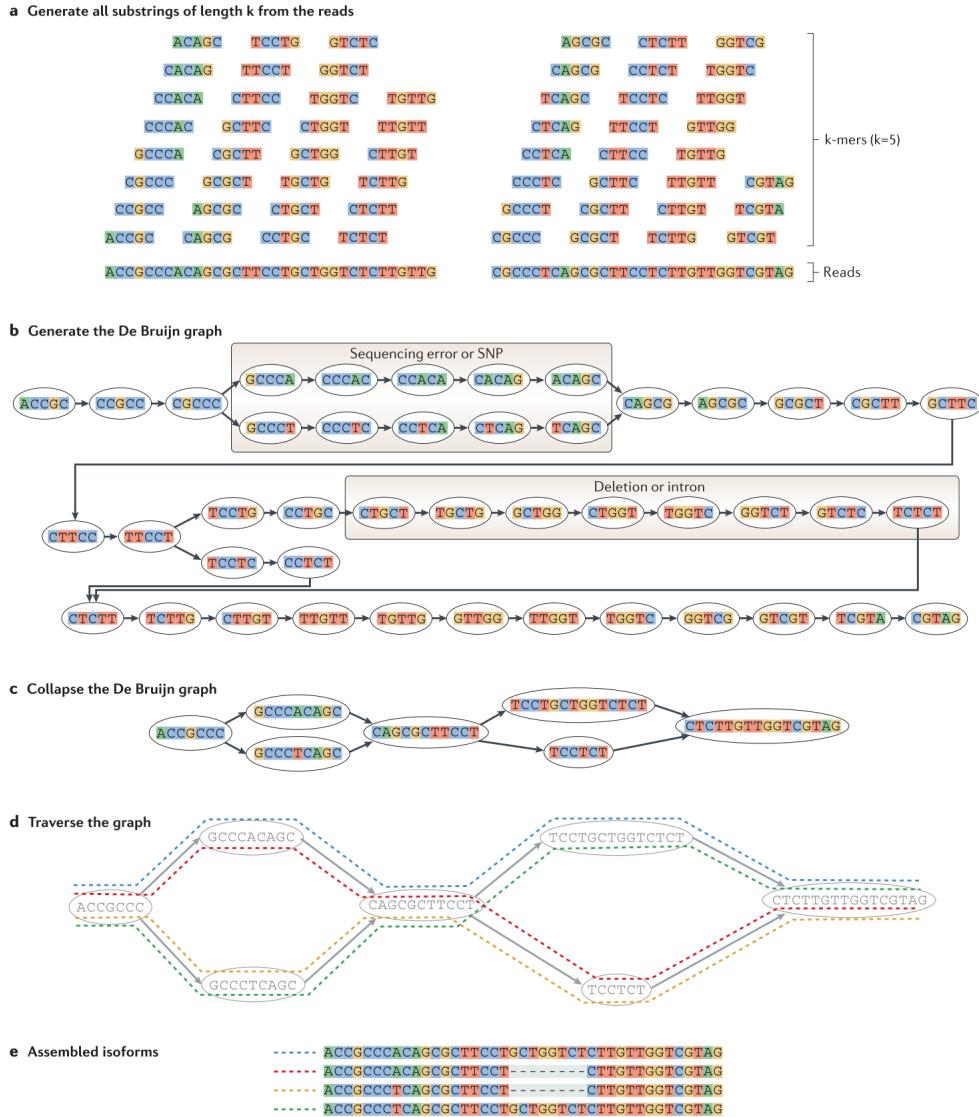


Figure 12.14: Overview of the de novo transcriptome assembly strategy

The de-novo assembly method involves generating all k-mers from the reads, linking them together using the De Bruijn graph, and collapsing the graph to obtain the isoforms (Figure 12.14).

The Trinity program (Figure 12.15) consists of three steps: (1) Inchworm assembles reads into linear contigs. It starts with the seed k-mer (usually the most abundant k-mer), then extends the 3' end one nucleotide at a time. It repeats until all k-mers are assembled. Low-coverage k-mers are thought as sequencing artifacts and are removed. (2) Chrysalis groups related contigs into De Bruijn graphs. Ideally, there will be one graph corresponding to each gene. (3) Butterfly re-groups related Inchworm contigs into isoforms or paralog genes. (4) Lastly treat this as a new reference and map reads back to quantify.

Hybrid approaches are used for lesser quality transcripts or transcriptomes that have undergone major rearrangements, such as those of cancer cells. Popular transcript assembly tools include Oasis, Trans-ABYSS, and Trinity. Regardless of methodology or software type, any sequencing experiment that produces more genome coverage will have better transcript reconstruction.

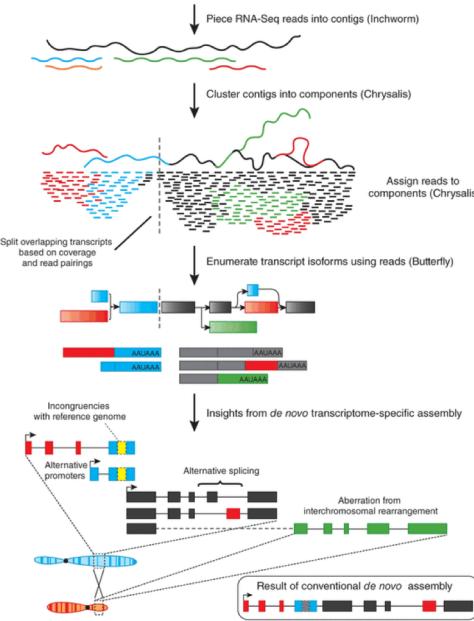


Figure 12.15: Trinity as one representative of several transcriptome assemblers. The figure is from Lyer et al, NatBT 2011

## 12.5 Differential Expression Analysis

Suppose we want to perform an RNA-Seq analysis for two different genes in a given sample or a gene under two different conditions. How can we tell if there is a significant difference in the fragment counts? We calculate expression by estimating the expected number of fragments that come from each transcript. To test for significance, we need to estimate the variance of expectation. The challenge here is to account for both technical and biological uncertainties.

For modeling variance across technical replicates, we use the Poisson distribution with the assumption that variance of FPKM is equal to expectation of FPKM. Note that this assumption works only for technical variations across replicates, but not for biological variations. Using Poisson to model biological variability, or variability across replicates, results in overdispersion. The solution is to mix the Poisson into a new kind of distribution in order to model biological variability:

$$X \sim \text{Poisson}(\Gamma(r, p))$$

This is a Negative Binomial distribution, which is used by the DESeq package in R. To learn parameters of the negative binomial for each gene, we can just fit a gamma function through a scatter plot of mean count versus count variance across replicates (Figure 12.16).

The difficulty arrises in normalizing for input material. This is generally straightforward when just considering read depth as one can simply divide the counts by a factor of differential depth. However, when multiple genes are being considered, this can get more complicated since counts are relative abundance and not absolute, though simplifications can be made if the majority (95%) of genes are constant. One such formulation of a normalization constant for experiment  $j$  would be:

$$s_j = \text{median}_i \frac{k_{ij}}{(\prod_{v=1}^m k_{iv})^{1/m}}$$

Where  $i$  iterates through all  $n$  genes,  $j$  iterates through all  $m$  samples, and  $k_{ij}$  is the observed counts for gene  $i$  in sample  $j$ .

If many of the genes are fluctuating in a given dataset, mathematical methods of normalizing data for comparisons will be uninformed. In this case, the only solution is to use spike-ins. Spike-ins are pieces

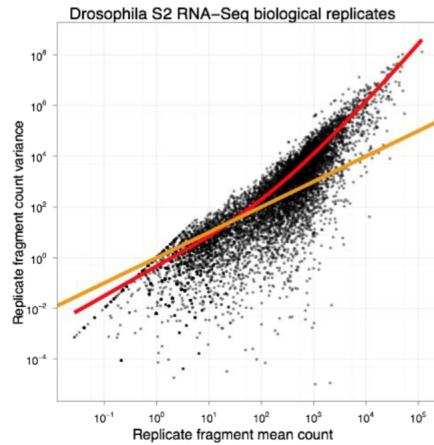


Figure 12.16: Drosophila S2 RNA-Seq biological variability across replicates

of RNA that are added after RNA isolation but before reverse transcription. They are added in a known quantity and by using them, the reverse transcription efficiency as well as the amount of input can be measured material.

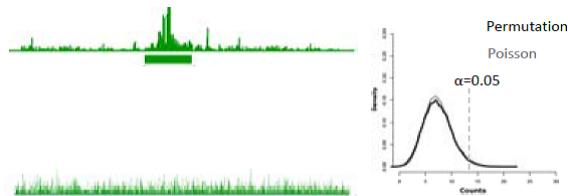


Figure 12.17: Technical variability follows a Poisson distribution

The two threads of RNA-Seq expression analysis research focus on the above problems. One of the threads focuses on inferring the abundances of individual isoforms to learn about differential splicing and promoter use, while the other thread focuses on modeling variability across replicates to create more robust differential gene expression analysis. Cuffdiff unites these two separate threads to study the case where we have biological variability and read mapping ambiguity. Since overdispersion can be modeled with a negative binomial distribution and mapping uncertainty can be modeled with a Beta distribution, we combine these two to model this case with a beta negative binomial distribution.



---

CHAPTER  
THIRTEEN

---

## GENE REGULATION 1 – GENE EXPRESSION CLUSTERING

Melissa Slaughter and Larry Zhang (2016)  
Ge Liu and Shau-Chieh Hsu (2015)  
Franck Dernoncourt (2012)  
Arvind Thiagarajan (2011)  
Tahin Syed (2010)  
Barrett Steinberg and Brianna Petrone (2009)  
Mia Y. Qia (2008)

### Figures

---

13.1 Clustering compared to classification. In clustering we group observations into clusters based on how near they are to one another. In classification we want a rule that will accurately assign labels to new points. . . . .	220
13.2 Gene expression values from microarray experiments can be represented as heat maps to visualize the result of data analysis. . . . .	221
13.3 RNA-Seq reads mapping to a gene ( <i>c-fos</i> ) and its splice junctions. Densities along exon represent read densities mapping to exons (in $\log_{10}$ ), arcs correspond to junction reads, where arc width is drawn in proportion to number of reads in that junction. The gene is downregulated in Sample 2 compared to Sample 1. . . . .	222
13.4 Transforming Figure 4 to a heatmap . . . . .	223
13.5 Gene expression level in log value comparison with reference sample . . . . .	223
13.6 A sample matrix of gene expression values, represented as a heatmap and with hierarchical clusters. [1] . . . . .	223
13.7 Using gene expression matrix to infer more about a disease and gene segment . . . . .	224
13.8 The k-means clustering algorithm . . . . .	225
13.9 Example of final cluster assignments using fuzzy $k$ -means ( $k = 4$ ) with centroids, actual clusters, and assigned clusters marked as crosses, shapes, and colors respectively. Note that the original data set is non-Gaussian. . . . .	226
13.10 $K$ -Means as a Generative Model. Samples were drawn from normal distributions. . . . .	226
13.11 $K$ -Means as an expectation maximization (EM) algorithm. . . . .	227
13.12 Comparison of clustering, HMM and motif discovery with respect to expectation minimization (EM) algorithm. . . . .	228
13.13 Hierarchical Clustering . . . . .	228
13.14 Distance Metrics for Hierarchical Clustering. Clockwise from top left: minimum, maximum, average distance and centroid distance. . . . .	229
13.15 Calculation of probability that you have more than $r$ +'s in a randomly selected cluster. .	230

---

## 13.1 Introduction

In this chapter, we consider the problem of discerning similarities or patterns within large datasets. Finding structure in such data sets allows us to draw conclusions about the process as well as the structure underlying the observations. We approach this problem through the application of clustering techniques. The following chapter will focus on classification techniques.

### 13.1.1 Clustering vs Classification

One important distinction to be made early on is the difference between classification and clustering. **Classification** is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations or instances whose category membership is known. The training set is used to learn rules that will accurately assign labels to new observations. The difficulty is to find the most important features (feature selection).

In the terminology of machine learning, classification is considered an instance of supervised learning, i.e. learning where a training set of correctly-identified observations is available. The corresponding unsupervised procedure is known as **clustering** or cluster analysis, and involves grouping data into categories based on some measure of inherent similarity, such as the distance between instances, considered as vectors in a multi-dimensional vector space. The difficulty is to identify the structure of the data. Figure 13.1 illustrates the difference between clustering and classification.

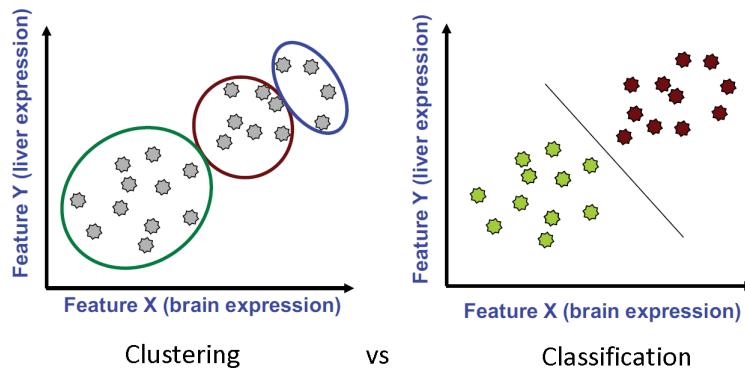


Figure 13.1: Clustering compared to classification. In clustering we group observations into clusters based on how near they are to one another. In classification we want a rule that will accurately assign labels to new points.

### 13.1.2 Applications

Clustering was originally developed within the field of artificial intelligence. Being able to group similar objects, with full implications of generality implied, is indeed a fairly desirable attribute for an artificial intelligence, and one that humans perform routinely throughout life. As the development of clustering algorithms proceeded apace, it quickly becomes clear that there was no intrinsic barrier involved in applying these algorithms to larger and larger datasets. This realization led to the rapid introduction of clustering to computational biology and other fields dealing with large datasets.

Clustering has many applications to computational biology. For example, let's consider expression profiles of many genes taken at various developmental stages. Clustering may show that certain sets of genes line up (i.e. show the same expression levels) at various stages. This may indicate that this set of genes has common expression or regulation and we can use this to infer similar function. Furthermore, if we find a uncharacterized gene in such a set of genes, we can reason that the uncharacterized gene also has a similar function through guilt by association.

Chromatin marks and regulatory motifs can be used to predict logical relationships between regulators and target genes in a similar manner. This sort of analysis enables the construction of models that allow us

to predict gene expression. These models can be used to modify the regulatory properties of a particular gene, predict how a disease state arose, or aid in targeting genes to particular organs based on regulatory circuits in the cells of the relevant organ.

Computational biology deals with increasingly large and open-access datasets. One such example is the ENCODE project [2]. Launched in 2003, the goal of ENCODE is to build a comprehensive list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active. ENCODE data are now freely and immediately available for the entire human genome: <http://genome.ucsc.edu/ENCODE/>. Using all of this data, it is possible to make functional predictions about genes through the use of clustering.

## 13.2 Methods for Measuring Gene Expression

The most intuitive way to investigate a certain phenotype is to measure the expression levels of functional proteins present at a given time in the cell. However, measuring the concentration of proteins can be difficult, due to their varying locations, modifications, and contexts in which they are found, as well as due to the incompleteness of the proteome. mRNA expression levels, however, are easier to measure, and are often a good approximation. By measuring the mRNA, we analyze regulation at the transcription level, without the added complications of translational regulation and active protein degradation, which simplifies the analysis at the cost of losing information. In this chapter, we will consider two techniques for generating gene expression data: microarrays and RNA-seq.

### 13.2.1 Microarrays

Microarrays allow the analysis of the expression levels of thousands of preselected genes in one experiment. The basic principle behind microarrays is the hybridization of complementary DNA fragments. To begin, short segments of DNA, known as probes, are attached to a solid surface, commonly known as a gene chip. Then, the RNA population of interest, which has been taken from a cell, is reverse transcribed to cDNA (complementary DNA) via reverse transcriptase, which synthesizes DNA from RNA using the poly-A tail as a primer. For intergenic sequences which have no poly-A tail, a standard primer can be ligated to the ends of the mRNA. The resulting DNA has more complementarity to the DNA on the slide than the RNA. The cDNA is then washed over the chip and the resulting hybridization triggers the probes to fluoresce. This can be detected to determine the relative abundance of the mRNA in the target, as illustrated in figure 13.2.

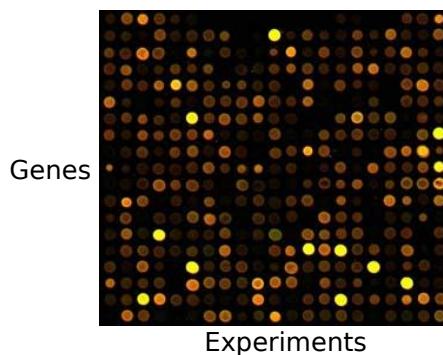


Figure 13.2: Gene expression values from microarray experiments can be represented as heat maps to visualize the result of data analysis.

Two basic types of microarrays are currently used. Affymetrix gene chips have one spot for every gene and have longer probes on the order of 100s of nucleotides. On the other hand, spotted oligonucleotide arrays tile genes and have shorter probes around the tens of bases.

There are numerous sources of error in the current methods and future methods seek to remove steps in the process. For instance, reverse transcriptase may introduce mismatches, which weaken interaction with the correct probe or cause cross hybridization, or binding to multiple probes. One solution to this has

been to use multiple probes per gene, as cross hybridization will be different for each gene. Still, reverse transcription is necessary due to the secondary structure of RNA. The structural stability of DNA makes it less probable to bend and not hybridize to the probe. The next generation of technologies, such as RNA-Seq, sequences the RNA as it comes out of the cell, essentially probing every base of the genome.

### 13.2.2 RNA-seq

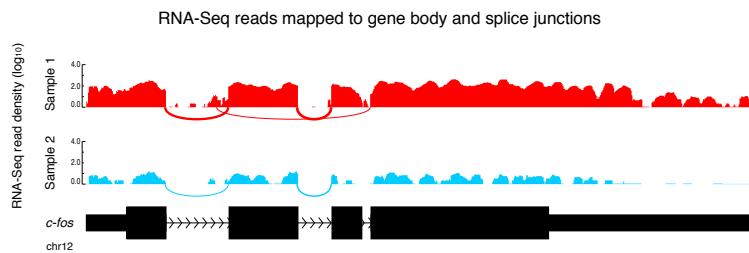


Figure 13.3: RNA-Seq reads mapping to a gene (*c-fos*) and its splice junctions. Densities along exon represent read densities mapping to exons (in  $\log_{10}$ ), arcs correspond to junction reads, where arc width is drawn in proportion to number of reads in that junction. The gene is downregulated in Sample 2 compared to Sample 1.

RNA-Seq, also known as whole transcriptome shotgun sequencing, attempts to perform the same function that DNA microarrays have been used to perform in the past, but with greater resolution. In particular, DNA microarrays utilize specific probes, and creation of these probes necessarily depends on prior knowledge of the genome and the size of the array being produced. RNA-seq removes these limitations by simply sequencing all of the cDNA produced in microarray experiments. This is made possible by next-generation sequencing and alignment technology. The technique has been rapidly adopted in studies of diseases like cancer [4]. The data from RNA-seq is then analyzed by clustering in the same manner as data from microarrays would normally be analyzed.

### 13.2.3 Gene Expression Matrices

Microarrays and RNA-seq are frequently used to compare the gene expression profiles of cells under various conditions. The amount of data generated from these experiments is enormous. Microarrays can analyze thousands of genes, and RNA-seq can, in principle, analyze every gene that is actively expressed. The expression level of each of those genes is measured across a variety of conditions, including time courses, stages of development, phenotypes, healthy vs. sick, and other factors.

To understand what the heatmap of a gene expression matrix (Figure 13.4) convey, we have to first understand what the expression data matrix tells us. By using microarrays and RNA-seq, we can obtain gene expression level in quantitative form in an experiment. If we have multiple experiments, we can construct a value matrix (Figure 13.5) representing a log value of  $(T/R)$ , where  $T$  is the gene expression level in test sample and  $R$  is the gene expression level in reference sample.

If we visualize the matrix as a heatmap, then we obtain the following new colored-matrix:

These matrices can be clustered hierarchically showing the relation between pairs of genes, pairs of pairs, and so on, creating a dendrogram in which the rows and columns can be ordered using optimal leaf ordering algorithms.

The Expression Matrix is a representation of data from multiple microarray experiments.

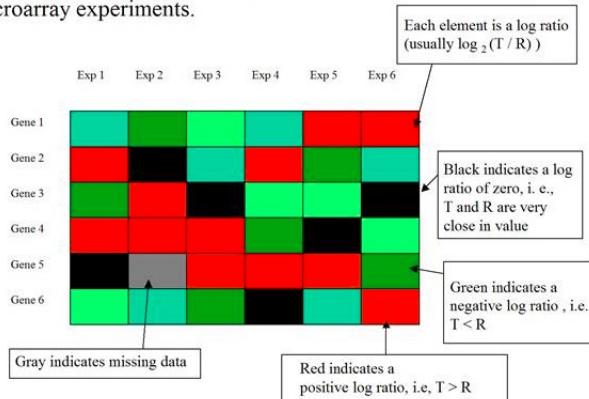


Figure 13.4: Transforming Figure 4 to a heatmap

	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Exp 6
Gene 1	-1.2	-2.1	-3	-1.5	1.8	2.9
Gene 2	2.7	0.2	-1.1	1.6	-2.2	-1.7
Gene 3	-2.5	1.5	-0.1	-1.1	-1	0.1
Gene 4	2.9	2.6	2.5	-2.3	-0.1	-2.3
Gene 5	0.1		2.6	2.2	2.7	-2.1
Gene 6	-2.9	-1.9	-2.4	-0.1	-1.9	2.9

Figure 13.5: Gene expression level in log value comparison with reference sample

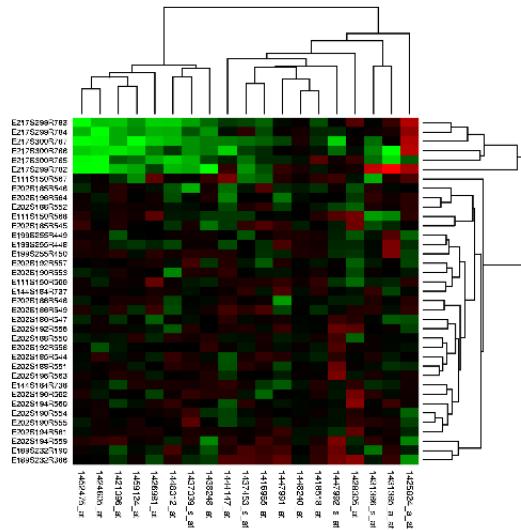


Figure 13.6: A sample matrix of gene expression values, represented as a heatmap and with hierarchical clusters. [1]

By revealing the hidden structure of a long segment of genome, we obtain great insight of what a fragment of gene does, and subsequently understand more about the root cause of an unknown disease.

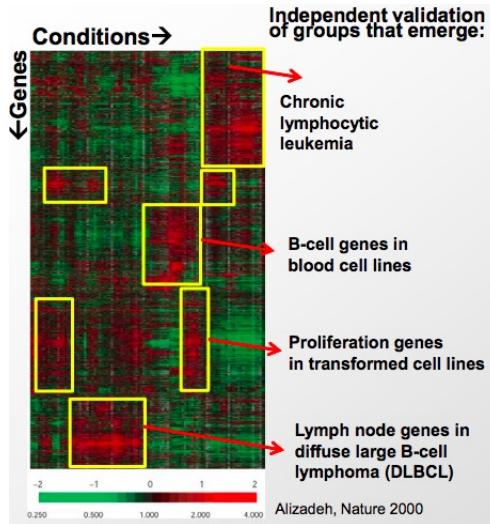


Figure 13.7: Using gene expression matrix to infer more about a disease and gene segment

This predictive and analytical power is increased due to the ability of biclustering the data; that is, clustering along both dimensions of the matrix. The matrix allows for the comparison of expression profiles of genes, as well as comparing the similarity of different conditions such as diseases. A challenge, though, is the curse of dimensionality. As the space of the data increases, the clustering of the points diminishes. Sometimes, the data can be reduced to lower dimensional spaces to find structure in the data using clustering to infer which points belong together based on proximity.

Interpreting the data can also be a challenge, since there may be other biological phenomena in play. For example, protein-coding exons have higher intensity, due to the fact that introns are rapidly degraded. At the same time, not all introns are junk and there may be ambiguities in alternative splicing. There are also cellular mechanisms that degrade aberrant transcripts through non-sense mediated decay.

### 13.3 Clustering Algorithms

To analyze the gene expression data, it is common to perform clustering analysis. There are two types of clustering algorithms: partitioning and agglomerative. Partitional clustering divides objects into non-overlapping clusters so that each data object is in one subset. Alternatively, agglomerative clustering methods yield a set of nested clusters organized as a hierarchy representing structures from broader to finer levels of detail.

#### 13.3.1 K-Means Clustering (Clustering by Partitioning)

The  $k$ -means algorithm clusters  $n$  objects based on their attributes into  $k$  partitions. This is an example of partitioning, where each point is assigned to exactly one cluster such that the sum of distances from each point to its correspondingly labeled center is minimized. The motivation underlying this process is to make the most compact clusters possible, usually in terms of a Euclidean distance metric.

The  $k$ -means algorithm, as illustrated in figure 13.8, is implemented as follows:

1. Assume a fixed number of clusters,  $k$
2. *Initialization:* Randomly initialize the  $k$  means  $\mu_k$  associated with the clusters and assign each data point  $x_i$  to the nearest cluster, where the distance between  $x_i$  and  $\mu_k$  is given by  $d_{i,k} = (x_i - \mu_k)^2$ .
3. *Iteration:* Recalculate the centroid of the cluster given the points assigned to it:  $\mu_k(n+1) = \sum_{x_i \in k} \frac{x_i}{|x_k|}$  where  $|x_k|$  is the number of points with label  $k$ . Reassign data points to the  $k$  new centroids by

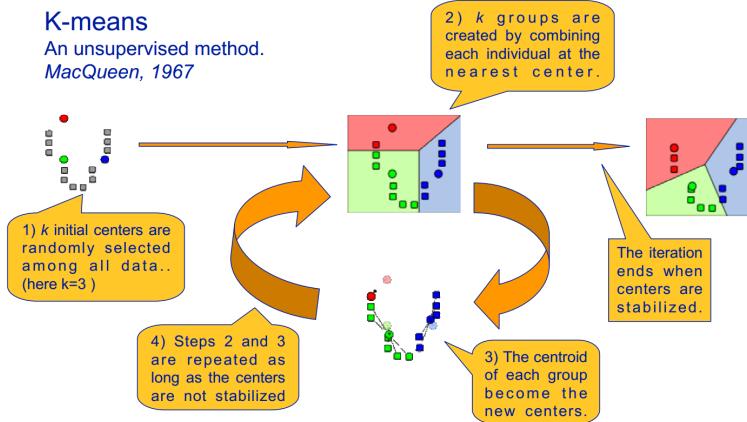


Figure 13.8: The k-means clustering algorithm

the given distance metric. The new centers are effectively calculated to be the average of the points assigned to each cluster.

4. *Termination*: Iterate until convergence or until a user-specified number of iterations has been reached. Note that the iteration may be trapped at some local optima.

There are several methods for choosing *k*: simply looking at the data to identify potential clusters or iteratively trying values for *k*, while penalizing model complexity. We can always make “better” clusters by increasing *k*, but at some point we begin overfitting the data.

We can also think of *k*-means as trying to minimize a cost criterion associated with the size of each cluster, where the cost increases as the clusters get less compact. However, some points can be almost halfway between two centers, which doesn’t fit well with the binary belonging *k*-means clustering.

### 13.3.2 Fuzzy K-Means Clustering

In fuzzy clustering, each point has a probability of belonging to each cluster, rather than completely belonging to just one cluster. Fuzzy *k*-means specifically tries to deal with the problem where points are somewhat in between centers or otherwise ambiguous by replacing distance with probability, which of course could be some function of distance, such as having probability relative to the inverse of the distance. Fuzzy *k*-means uses a weighted centroid based on those probabilities. Processes of initialization, iteration, and termination are the same as the ones used in *k*-means. The resulting clusters are best analyzed as probabilistic distributions rather than a hard assignment of labels. One should realize that *k*-means is a special case of fuzzy *k*-means when the probability function used is simply 1 if the data point is closest to a centroid and 0 otherwise.

The fuzzy *k*-means algorithm is the following:

1. Assume a fixed number of clusters *k*
2. *Initialization*: Randomly initialize the *k* means  $\mu_k$  associated with the clusters and compute the probability that each data point  $x_i$  is a member of a given cluster *k*,  $P(\text{point } x_i \text{ has label } k|x_i, \mu_k)$ .
3. *Iteration*: Update the centroid  $\mu_k$  of the cluster as the weighted mean of all data points  $x_i$  assigned to it:

$$\mu_k(n+1) = \frac{\sum_{x_i \in k} x_i P(\mu_k|x_i)^b}{\sum_{x_i \in k} P(\mu_k|x_i)^b}$$

Reassign each point  $x_i$  to all centers  $\mu_k$ . One such way of doing this is to calculate the probability of membership weighted by distance:  $P(\text{point } x_i \text{ has label } k|x_i, \mu_k)$ .

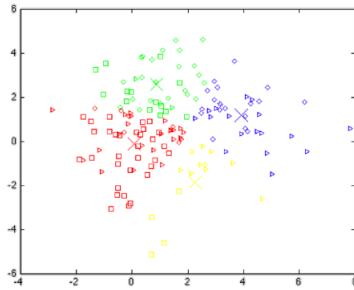


Figure 13.9: Example of final cluster assignments using fuzzy  $k$ -means ( $k = 4$ ) with centroids, actual clusters, and assigned clusters marked as crosses, shapes, and colors respectively. Note that the original data set is non-Gaussian.

4. *Termination:* Iterate until membership matrix converges or until a user-specified number of iterations has been reached. Note that the iteration may be trapped at some local optima.

In the above equation,  $b$  represents the “degree of fuzziness.” A large  $b$  results in smaller probabilities of membership, and hence fuzzier clusters. Additionally, the value of  $b$  will determine the speed of convergence.

### 13.3.3 $K$ -Means as a Generative Model

A generative model is a model for randomly generating observable-data values, given some hidden parameters. While a generative model is a probability model of all variables, a discriminative model provides a conditional model only of the target variable(s) using the observed variables.

In order to make  $k$ -means a generative model, we now look at it in a probabilistic manner, where we assume that data points in cluster  $k$  are generated using a probabilistic distribution. Here, we will use a Gaussian Mixture Model with the mean on the center of cluster and a variance of 1, which gives  $P(x_i|\mu_k) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x_i - \mu_k)^2}{2}\right\}$ . This gives a stochastic representation of the data, as shown in figure 13.10. Now this turns into a maximum likelihood problem, which, we will show below, is exactly equivalent to the original  $k$ -means algorithm mentioned above.

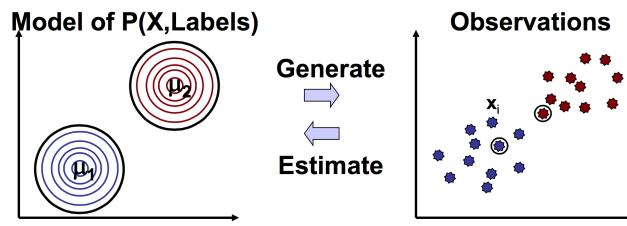


Figure 13.10:  $K$ -Means as a Generative Model. Samples were drawn from normal distributions.

In the generating step, we want to find a most likely partition, or assignment of label, for each  $x_i$  given the mean  $\mu_k$ . With the assumption that each point is drawn independently, we could look for the maximum likelihood label for each point separately:

$$\arg \max_k P(x_i|\mu_k) = \arg \max_k \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x_i - \mu_k)^2}{2}\right\} = \arg \min_k (x_i - \mu_k)^2$$

This is totally equivalent to finding the nearest cluster center in the original  $k$ -means algorithm.

In the Estimation step, we look for the maximum likelihood estimate of the cluster mean  $\mu_k$ , given the

partitions (labels):

$$\arg \max_{\mu} \left\{ \log \prod_i P(x_i | \mu) \right\} = \arg \max_{\mu} \sum_i \left\{ -\frac{1}{2}(x_i - \mu)^2 + \log\left(\frac{1}{\sqrt{2\pi}}\right) \right\} = \arg \min_{\mu} \sum_i (x_i - \mu)^2$$

Note that the solution of this problem is exactly the centroid of the  $x_i$ , which is the same procedure as the original  $k$ -means algorithm.

Unfortunately, since  $k$ -means assumes independence between the axes, covariance and variance are not accounted for using  $k$ -means, so models such as oblong distributions are not possible. However, this issue can be resolved when generalize this problem into expectation maximization problem.

### 13.3.4 Expectation Maximization

$K$ -means can be seen as an example of EM (expectation maximization algorithms), as shown in figure 13.11 where expectation consists of estimation of hidden labels,  $Q$ , and maximizing of expected likelihood occurs given data and  $Q$ . Assigning each point the label of the nearest center corresponds to the E step of estimating the most likely label given the previous parameter. Then, using the data produced in the E step as observation, moving the centroid to the average of the labels assigned to that center corresponds to the  $M$  step of maximizing the likelihood of the center given the labels. This case is analogous to Viterbi learning. A similar comparison can be drawn for fuzzy  $k$ -means, which is analogous to Baum-Welch from HMMs. Figure 13.12 compares clustering, HMM and motif discovery with respect to expectation minimization algorithm.

It should be noted that using the EM framework, the  $k$  means approach can be generalized to clusters of oblong shape and varying sizes. With  $k$  means, data points are always assigned to the nearest cluster center. By introducing a covariance matrix to the Gaussian probability function, we can allow for clusters of different sizes. By setting the variance to be different along different axes, we can even create oblong distributions.

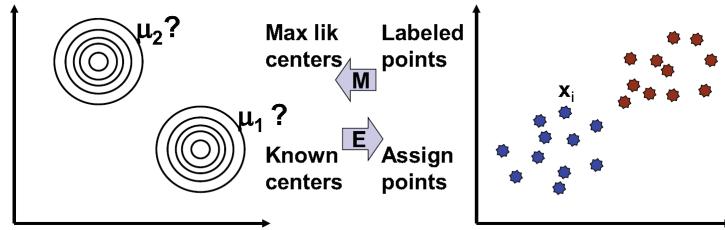


Figure 13.11:  $K$ -Means as an expectation maximization (EM) algorithm.

EM is guaranteed to converge and guaranteed to find the best possible answer, at least from an algorithmic point of view. The notable problem with this solution is that the existence of local maxima of probability density can prevent the algorithm from converging to the global maximum. One approach that may avoid this complication is to attempt multiple initializations to better determine the landscape of probabilities.

### 13.3.5 The limitations of the $K$ -Means algorithm

The  $k$ -means algorithm has a few limitations which are important to keep in mind when using it. First of all, it requires a metric.

The second main limitation of the  $k$ -means algorithm is its sensitivity to noise. One way to try to reduce the noise is to run a principle component analysis beforehand. Another way is to weight each variable in order to give less weight to the variables affected by significant noise: the weights will be calculated dynamically at each iteration of the k-means algorithm [3].

The third limitation is that the choice of initial centers can influence the results. There exist heuristics to select the initial cluster centers, but none of them are perfect.

Lastly, we need to know a priori the number of classes. As we have seen, there are ways to circumvent this problem, essentially by running the algorithm several times while varying  $k$  or using the rule of thumb

Update rule	Update assignments (E step) → Estimate hidden labels	Algorithm implementing E step in each of the three settings			Update model parameters (M step) → max likelihood
		Expression clustering	HMM learning	Motif discovery	
The hidden label is:	Cluster labels	State path $\pi$	Motif positions		
Pick a best	Assign each point to best label	<b>K-means:</b> Assign each point to nearest cluster	<b>Viterbi training:</b> label sequence with best path	<b>Greedy:</b> Find best motif match in each sequence	Average of those points assigned to label
Average all	Assign each point to all labels, probabilistically	<b>Fuzzy K-means:</b> Assign to all clusters, weighted by proximity	<b>Baum-Welch training:</b> label sequence w all paths (posterior decoding)	<b>MEME:</b> Use all positions as a motif occurrence weighed by motif match score	Average of all points, weighted by membership
Sample one	Pick one label at random, based on their relative probability	<b>N/A:</b> Assign to a random cluster, sample by proximity	<b>N/A:</b> Sample a single label for each position, according to posterior prob.	<b>Gibbs sampling:</b> Use one position for the motif, by sampling from the match scores	Average of those points assigned to label(a sample)

Figure 13.12: Comparison of clustering, HMM and motif discovery with respect to expectation minimization (EM) algorithm.

$k \approx \sqrt{n/2}$  if we are short on the computational side. [http://en.wikipedia.org/wiki/Determining\\_the\\_number\\_of\\_clusters\\_in\\_a\\_data\\_set](http://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set) summarizes well the different techniques to select the number of clusters. Hierarchical clustering provides a handy approach to choosing the number of cluster.

### 13.3.6 Hierarchical Clustering (Clustering by Agglomeration)

While the clustering discussed thus far often provides valuable insight into the nature of various data, they generally overlook an essential component of biological data, namely the idea that similarities might exist on multiple levels. To be more precise, similarity is an intrinsically hierarchical property, and this aspect is not addressed in the clustering algorithms discussed thus far. Hierarchical clustering specifically addresses this in a very simple manner, and is perhaps the most widely used algorithm for expression data. As illustrated in figure 13.13, it is implemented as follows:

1. *Initialization:* Initialize a list containing each point as an independent cluster.
2. *Iteration:* Create a new cluster containing the two closest clusters in the list. Add this new cluster to the list and remove the two constituent clusters from the list.

One key benefit of using hierarchical clustering and keeping track of the times at which we merge certain clusters is that we can create a tree structure that details the times at which we joined every cluster, as can be seen in figure 13.13. Thus, to get a number of clusters that fits your problem, you simply cut at a cut-level of your choice and that gives you the number of clusters corresponding to that cut-level. However, be aware that one potential pitfall with this approach is that at certain cut-levels, elements that are fairly close in space (such as e and b in figure 13.13), might not be in the same cluster.

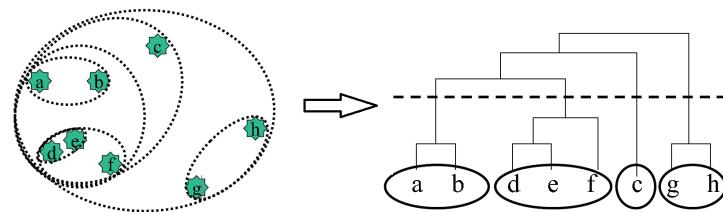


Figure 13.13: Hierarchical Clustering

Of course, a method for determining distances between clusters is required. The particular metric used varies with context, but (as can be seen in figure 13.14) some common implementations include the maximum, minimum, and average distances between constituent clusters, and the distance between the centroids of the clusters.

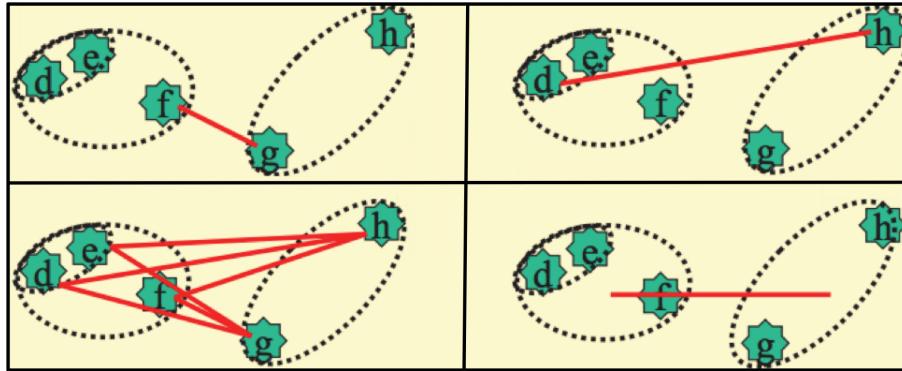


Figure 13.14: Distance Metrics for Hierarchical Clustering. Clockwise from top left: minimum, maximum, average distance and centroid distance.

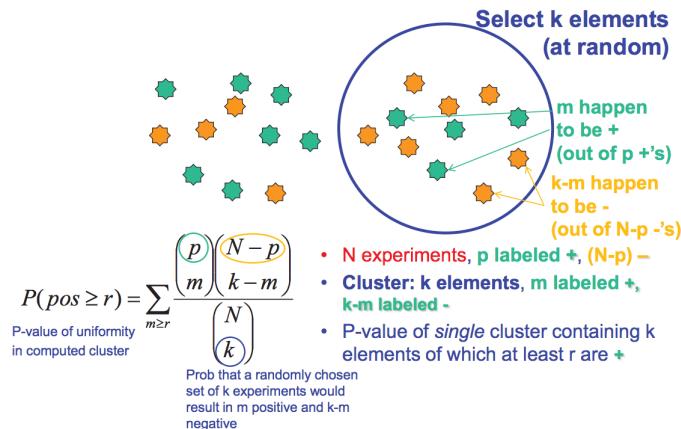
Noted that when choosing the closest clusters, calculating all pair-wise distances is very time and space consuming, therefore a better scheme is needed. One possible way of doing this is as follows: 1) Define some bounding boxes that divide the feature space into several subspaces. 2) Calculate pair-wise distances within each box. 3) Shift the boundaries of the boxes in different directions and recalculate pair-wise distances. 4) Choose the closest pair based on the results in all iterations.

### 13.3.7 Evaluating Cluster Performance

The validity of a particular clustering can be evaluated in a number of different ways. The overrepresentation of a known group of genes in a cluster, or, more generally, correlation between the clustering and confirmed biological associations, are good indicators of validity and significance. If biological data is not yet available, however, there are ways to assess validity using statistics. For instance, robust clusters will appear from clustering even when only subsets of the total available data are used to generate clusters. In addition, the statistical significance of a clustering can be determined by calculating the probability of a particular distribution having been obtained randomly for each cluster. This calculation utilizes variations on the hypergeometric distribution. As can be seen from figure 13.15, we can do this by calculating the probability that we have more than  $r$  '+'s when we pick  $k$  elements from a total of  $N$  elements. [http://en.wikipedia.org/wiki/Cluster\\_analysis#Evaluation\\_of\\_clustering\\_results](http://en.wikipedia.org/wiki/Cluster_analysis#Evaluation_of_clustering_results) gives several formula to assess the quality of the clustering.

## 13.4 Current Research Directions

The most significant problems associated with clustering now are associated with scaling existing algorithms cleanly with two attributes: size and dimensionality. To deal with larger and larger datasets, algorithms such as canopy clustering have been developed, in which datasets are coarsely clustered in a manner intended to pre-process the data, following which standard clustering algorithms (e.g.  $k$ -means) are applied to subdivide the various clusters. Increase in dimensionality is a much more frustrating problem, and attempt to remedy this usually involve a two stage process in which appropriate relevant subspaces are first identified by appropriate transformations on the original space and then subjected to standard clustering algorithms.

Figure 13.15: Calculation of probability that you have more than  $r$  +'s in a randomly selected cluster.

## 13.5 Further Reading

- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition, February 2009. Found online at <http://www-stat.stanford.edu/~tibs/ElemStatLearn/download.html>
- Numerical Recipes: The Art of Scientific Computing (3rd ed.). New York: Cambridge University Press.
- McLachlan, G.J. and Basford, K.E. (1988) "Mixture Models: Inference and Applications to Clustering", Marcel Dekker.
- Bezdek, J. C., Ehrlich, R., Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. Computers and Geosciences, 10(2), 191-203.
- <http://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-clustering-1.html>
- <http://compbio.uthsc.edu/microarray/lecture1.html>

## 13.6 Resources

- Cluster 3.0: open source clustering software that implements the most commonly used clustering methods for gene expression data analysis.
- MATLAB: K-means clustering: <http://www.mathworks.com/help/stats/kmeans.html>; Fuzzy C-means clustering: <http://www.mathworks.com/help/fuzzy/fcm.html>; Hierarchical Clustering: <http://www.mathworks.com/help/stats/linkage.html>
- Orange is a free data mining software suite (see module orngClustering for scripting in Python): <http://bonsai.hgc.jp/~mdehoon/software/cluster/software.htm>
- R (see Cluster Analysis and Finite Mixture Models)
- SAS CLUSTER

## 13.7 What Have We Learned?

To summarize, in this chapter we have seen that:

- In *clustering*, we identify structure in unlabeled data. For example, we might use clustering to identify groups of genes that display similar expression profiles.
  - Partitioning clustering algorithms, construct non-overlapping clusters such that each item is assigned to exactly one cluster. Example: k-means
  - Agglomerative clustering algorithms construct a hierarchical set of nested clusters, indicating the relatedness between clusters. Example: hierarchical clustering
  - By using clustering algorithms, we can reveal hidden structure of a gene expression matrix, which gives us valuable clues for understanding the mechanism of complicated diseases and categorizing different diseases
- In *classification*, we partition data into known labels. For example, we might construct a classifier to partition a set of tumor samples into those likely to respond to a given drug and those unlikely to respond to a given drug based on their gene expression profiles. We will focus on classification in the next chapter.

## Bibliography

- [1] <http://en.wikipedia.org/wiki/File:Heatmap.png>.
- [2] <http://genome.ucsc.edu/ENCODE/>.
- [3] J.Z. Huang, M.K. Ng, Hongqiang Rong, and Zichen Li. Automated variable weighting in k-means type clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(5):657 –668, may 2005.
- [4] Christopher A. Maher, Chandan Kumar-Sinha, Xuhong Cao, Shanker Kalyana-Sundaram, Bo Han, Xiaojun Jing, Lee Sam, Terrence Barrette, Nallasivam Palanisamy, and Arul M. Chinnaiyan. Transcriptome sequencing to detect gene fusions in cancer. *Nature*, 458(7234):97–101, Mar 05 2009.



---

CHAPTER  
**FOURTEEN**

---

## GENE REGULATION 2 –CLASSIFICATION

Melissa Slaughter and Larry Zhang (2016)  
Arvind Thiagarajan  
Fulton Wang  
Salil Desai  
David Charlton  
Kevin Modzelewski  
Robert Toscano

### Figures

---

14.1 Support Vector Machines . . . . .	238
--	-----

---

## 14.1 Introduction

In the previous chapter we looked at *clustering*, which provides a tool for analyzing data without any prior knowledge of the underlying structure. As we mentioned before, this is an example of “unsupervised” learning. This chapter deals with **supervised learning**, in which we are able to use pre-classified data to construct a model by which to classify more datapoints. In this way, we will use existing, known structure to develop rules for identifying and grouping further information.

There are two ways to do classification. The two ways are analogous to the two ways in which we perform motif discovery: HMM, which is a generative model that allows us to actually describe the probability of a particular designation being valid, and CRF, which is a discriminative method that allows us to distinguish between objects in a specific context. There is a dichotomy between generative and discriminative approaches.

In this lecture we will look at two new algorithms: a generative classifier, Naïve Bayes, and a discriminative classifier, Support Vector Machines (SVMs). We will discuss biological applications of each of these models, specifically in the use of Naïve Baye’s classifiers to predict mitochondrial proteins across the genome and the use of SVMs for the classification of cancer based on gene expression monitoring by DNA microarrays. The salient features of both techniques and caveats of using each technique will also be discussed.

Like with clustering, classification (and more generally supervised learning) arose from efforts in Artificial Intelligence and Machine Learning. Furthermore, much of the motivating infrastructure for classification had already been developed by probability theorists prior to the advent of either AI or ML.

## 14.2 Classification - Bayesian Techniques

Consider the problem of identifying mitochondrial proteins. If we look at the human genome, how do we determine which proteins are involved in mitochondrial processes, or more generally which proteins are

targeted to the mitochondria?<sup>1</sup> This is particularly useful because if we know the mitochondrial proteins, we can study how these proteins mediate disease processes and metabolic functions. The classification method we will look at considers 7 features for all human proteins:

1. targeting signal
2. protein domains
3. co-expression
4. mass spectrometry
5. sequence homology
6. induction
7. motifs

Our overall approach will be to determine how these features are distributed for both mitochondrial and non-mitochondrial proteins. Then, given a new protein, we can apply probabilistic analysis to these seven features to decide which class it most likely falls into.

#### 14.2.1 Single Feature and Bayes? Rule

To begin, we will focus on one feature. We must first assume that there is a class dependent distribution for the features, i.e. the likelihood derived from real data. Second, we need a prior chance of drawing a sample of a particular class before looking at the data. The chance of getting a particular class is simply the relative size of the class. Once we have these probabilities, we can use Bayes? rule to get the probability a sample is in a particular class given the data (this is called the posterior). We have forward generative probabilities, and use Bayes? rules to perform the backward inference. Note that it is not enough to just consider the probability the feature was drawn from each class dependent distribution, because if we knew a prior that one class (say class A) is much more common than the other, then it should take overwhelming evidence that the feature was drawn from class B's distribution for us to believe the feature was indeed from class B. The correct way to find what we need based on both evidence and prior knowledge is to use Bayes? Rule:

$$P(\text{Class}|\text{feature}) = \left( \frac{P(\text{feature}|\text{Class})P(\text{Class})}{P(\text{feature})} \right)$$

- **Posterior** :  $P(\text{Class}|\text{feature})$
- **Prior** :  $P(\text{Class})$
- **Likelihood** :  $P(\text{feature}|\text{Class})$
- **Evidence** :  $P(\text{feature})$

This formula gives us exactly the connection we need to flip known feature probabilities into class probabilities for our classifying algorithm. It lets us integrate both the likelihood we derive from our observations and our prior knowledge about how common something is. Note that the evidence  $P(\text{feature})$  does not depend on class nor involve in the classification process; we can simply view it as a normalizing constant. In the case of mtDNA, for example, we can estimate that mitochondrial DNA makes up something like 1500/21000 (i.e. less than 10%) of the human genome. Therefore, applying Bayes? rule, our classifier should only classify a gene as mitochondrial if there is a **very strong likelihood based on the observed features**, since the prior probability that any gene is mitochondrial is so low.

<sup>1</sup>Mitochondria is the energy producing machinery of cell. Very early in life, the mitochondria was engulfed by the predecessor to modern day eukaryotes, and now, we have different compartments in our cells. So the mitochondria has its own genome, but it is very depleted from its own ancestral genome - only about 11 genes remain. But there are hundreds of genes that make the mitochondria work, and these proteins are encoded by genes transcribed in the nucleus, and then transported to the mitochondria. So the goal is to figure out which proteins encoded in the genome are targeted to the mitochondria. This is important because there are many diseases associated with the mitochondria, such as aging.

With this rule, we can now form a maximum likelihood rule for predicting an object's class based on an observed feature. We want to choose the class that has the highest probability given the observed feature, so we will choose Class1 instead of Class2 if:  $P(\text{feature}|\text{Class1})P(\text{Class1}) > P(\text{feature}|\text{Class2})P(\text{Class2})$

Another way of looking at this is as a discriminant function: by rearranging the formulas above and taking the logarithm ratio, we should select Class1 instead of Class2 precisely when  $G(X) = \log\left(\frac{P(\text{feature}|\text{Class1})P(\text{Class1})}{P(\text{feature}|\text{Class2})P(\text{Class2})}\right) > 0$

In this case the use of logarithms provide distinct advantages:

1. Numerical stability
2. Easier math (it's easier to add the expanded terms than multiply them)
3. Monotonically increasing discriminators.

To find point or interval estimates of a class using the posterior distribution, i.e. choosing the class that maximizes the posterior probability, we can use the **maximum a posteriori (MAP) estimation**:

$$\begin{aligned} \text{BestClass} &= \operatorname{argmax}_C P(\text{Class}|\text{Feature}) \\ &= \operatorname{argmax}_C P(\text{Feature}|\text{Class})P(\text{Class}). \end{aligned}$$

This discriminant function does not capture the penalties associated with misclassification (in other words, is one classification more detrimental than other). In this case, we are essentially minimizing the number of misclassifications we make overall, but not assigning penalties to individual misclassifications. From examples discussed in class and in the problem set - if we are trying to classify a patient as having cancer or not, it could be argued that it is far more harmful to misclassify a patient as being healthy if they have cancer than to misclassify a patient as having cancer if they are healthy. In the first case, the patient will not be treated and would be more likely to die, whereas the second mistake involves emotional grief but no greater chance of loss of life. To formalize the penalty of misclassification we define something called a loss function,  $L_{kf}$ , which assigns a loss to the misclassification of an object as class j when the true class is class k (a specific example of a loss function was seen in Problem Set 2).

### 14.2.2 Multiple features and Naïve Bayes

In classifying mitochondrial DNA, we were looking at 7 features and not just one. In order to use the preceding methods with multiple features, we would need not just one bin for each individual feature range, but one for each combination of features ? if we look at two features with five ranges each, that's already 25 bins. All seven features gives us almost 80,000 bins ? and we can expect that most of those bins will be empty simply because we don't have enough training data to fill them all. This would cause problems because zeroes cause infinite changes in the probabilities of being in one class. Clearly this approach won't scale well as we add more features, so we need to estimate combined probabilities in a better way.

The solution we will use is to **assume the features are independent given class**. This is the Naïve Bayes Assumption, and it is almost always false, but it is often used anyway for the combined reasons that it is very easy to manipulate mathematically and it is often close enough to the truth that it gives a reasonable approximation. Note that this assumption does not say that all features are independent; instead the assumption says that those connections are divided by the different classes, and that within each individual class there are no further correlations. Also, if you know that some features are coupled, you could learn the joint conditional distribution in those pairs of the features.

Once we assume independence, the probability of combined features is simply the product of the individual probabilities associated with each feature. Using  $f_i, i = 1, \dots, N$  to denote different features, we have:  $P(f_1, f_2, \dots, f_N|\text{Class}) = P(f_1|\text{Class})P(f_2|\text{Class})\dots P(f_N|\text{Class})$

where  $f_1$  represents feature 1. Similar to single feature, we use discriminant function with multiple features to compare posterior probabilities of class 1 and class 2:  $G(f_1, \dots, f_N) = \log\left(\frac{\prod_{i=1}^N P(f_i|\text{Class1})P(\text{Class1})}{\prod_{j=1}^N P(f_j|\text{Class2})P(\text{Class2})}\right)$ .

### 14.2.3 Collecting Data

The preceding tells us how to handle predictions if we already know the exact probabilities corresponding to each class. If we want to classify mitochondrial proteins based on feature  $X$ , we still need ways of determining the probabilities  $P(\text{mito})$ ,  $P(\text{not mito})$ ,  $P(X|\text{mito})$  and  $P(X|\text{not mito})$ . To do this, we need a training set: a set of data that is already classified that our algorithm can use to ?learn? the distributions corresponding to each class. A **high-quality training set** (one that is both large and unbiased) is the most important part of any classifier. An important question at this point is, how much data do we need about known genes in order to build a good classifier for unknown genes? This is a hard question whose answer is not fully known. However, there are some simple methods that can give us a good estimate: when we have a fixed set of training data, we can keep a holdout set that we don?t use for our algorithm, and instead use those (known) data points to test the accuracy of our algorithm when we try to classify them. By trying different sizes of training versus *holdout* set, we can check the accuracy curve of our algorithm. Generally speaking, we have ?enough? training data when we see the accuracy curve flatten out as we increase the amount of training data (this indicates that additional data is likely to give only a slight marginal improvement). The holdout set is also called the test set, because it allows us to test the generalization power of our classifier.

Supposing we have already collected our training data, however, how should we model  $P(X|\text{Class})$ ? There are many possibilities. One is to use the same approach we did with clustering in the last lecture and model the feature as a Gaussian ? then we can follow the maximum likelihood principle to find the best center and variance. The one used in the mitochondrial study is a simple density estimate: for each feature, divide the range of possibilities into a set of bins (say, five bins per feature). Then we use the given data to estimate the probability of a feature falling into each bin for a given class. The principle behind this is again maximum likelihood, but for a multinomial distribution rather than a Gaussian. We may choose to discretize a otherwise continuous distribution because estimating a continuous distribution can be complex.

There is one issue with this strategy: what if one of the bins has zero samples in it? A probability of zero will override everything else in our formulas, so that instead of thinking this bin is merely unlikely, our classifier will believe it is *impossible*. There are many possible solutions, but the one taken here is to apply the *Laplace Correction*: add some small amount (say, one element) into each bin, to draw probability estimates slightly towards uniform and account for the fact that (in most cases) none of the bins are truly impossible. Another way to avoid having to apply the correction is to choose bins that are not too small so that bins will not have zero samples in them in practice. If you have many many points, you can have more bins, but run the risk of overfitting your training data.

### 14.2.4 Estimating Priors

We now have a method for approximating the feature distribution for a given class, but we still need to know the relative probability of the classes themselves. There are **three general approaches**:

1. Estimate the priors by counting the relative frequency of each class in the training data. This is prone to bias, however, since data available is often skewed disproportionately towards less common classes (since those are often targeted for special study). If we have a high-quality (representative) sample for our training data, however, this works very well.
2. Estimate from expert knowledge—there may be previous estimates obtained by other methods independent of our training data, which we can then use as a first approximation in our own predictions. In other words, you might ask experts what the percentage of mitochondrial proteins are.
3. Choose a noninformative prior by assuming all classes are equally likely ? we would typically do this if we have no information at all about the true frequencies. This is effectively what we do when we use the maximum likelihood principle: our clustering algorithm was essentially using Bayesian analysis under the assumption that all priors are equal. This is actually a strong assumption, but when you have no other data, this is the best you can do.

For classifying mitochondrial DNA, we use method (2), since some estimates on the proportions of mtDNA were already known.

### 14.2.5 Testing a classifier

A classifier should always be tested on data not contained in its training set. We can imagine in the worst case an algorithm that just memorized its training data and behaved randomly on anything else ? a classifier that did this would perform perfectly on its training data, but that indicates nothing about its real performance on new inputs. This is why it's important to use a test, or holdout, set as mentioned earlier. However, a simple error rate doesn't encapsulate all of the possible consequences of an error. For a simple binary classifier (an object is either in or not in a single target class), there are the following four types of errors:

1. True positive (TP)
2. True negative (TN)
3. False positive (FP)
4. False negative (FN)

The frequency of these errors can be encapsulated in performance metrics of a classifier which are defined as,

1. *Sensitivity (recall) =  $\frac{TP}{TP+FN}$*  ? what fraction of objects that are in a class are correctly labeled as that class? That is, what fraction have true positive results? High sensitivity means that elements of a class are very likely to be labeled as that class. Low sensitivity means there are too many false negatives.
2. *Specificity =  $\frac{TN}{TN+FP}$*  ? what fraction of objects not in a class are correctly labeled as not being in that class? That is, what fraction have true negative results? High specificity means that elements labeled as belonging to a class are very likely to actually belong to it. Low specificity means there are too many false positives.

In most algorithms there is a **tradeoff between sensitivity and specificity**. For example, we can reach a sensitivity of 100% by labeling everything as belonging to the target class, but we will have a specificity of 0%, so this is not useful. Generally, most algorithms have some probability cutoff they use to decide whether to label an object as belonging to a class (for example, our discriminant function above). Raising that threshold increases the specificity but decreases the sensitivity, and decreasing the threshold does the reverse. The MAESTRO algorithm for classifying mitochondrial proteins (described in this lecture) achieves 99% specificity and 71% sensitivity.

### 14.2.6 MAESTRO ? Mitochondrial Protein Classification

They find a class dependent distribution for each feature by creating several bins and evaluating the proportion of mitochondrial and non mitochondrial proteins in each bin. This lets you evaluate the **usefulness of each feature** in classification. You end up with a bunch of medium strength classifiers, but when you combine them together, you hopefully end up with a stronger classifier. Calvo et al. [1] sought to construct high-quality predictions of human proteins localized to the mitochondrion by generating and integrating data sets that provide complementary clues about mitochondrial localization. Specifically, for each human gene product  $p$ , they assign a score  $s_i(p)$ , using each of the following seven genome-scale data sets ? targeting signal score, protein domain score, cis-motif score, yeast homology score, ancestry score, coexpression score, and induction score (details of each of the meaning and content of each of these data sets can be found in the manuscript). Each of these scores  $s_1 - s_7$  can be used individually as a weak genome-wide predictor of mitochondrial localization. Each method's performance was assessed using large ?gold standard? curated training sets - 654 mitochondrial proteins  $T_{\text{mito}}$  maintained by the MitoP2 database<sup>1</sup> and 2,847 nonmitochondrial proteins  $T_{\text{mito}}$  annotated to localize to other cellular compartments. To improve prediction accuracy, the authors integrated these eight approaches using a naïve Bayes classifier that was implemented as a program called MAESTRO. So we can take several weak classifiers, and combine them to get a stronger classifier.

When MAESTRO was applied across the human proteome, 1451 proteins were predicted as mitochondrial proteins and 450 novel proteins predictions were made. As mentioned in the previous section The MAESTRO

algorithm achieves a 99% specificity and a 71% sensitivity for the classification of mitochondrial proteins, suggesting that even with the assumption of feature independence, Naïve Bayes classification techniques can prove extremely powerful for large-scale (i.e. genome-wide) scale classification.

## 14.3 Classification - Support Vector Machines

The previous section looked at using probabilistic (or generative) models for classification, this section looks at discriminative techniques ? in essence, can we run our data through a function to determine its structure? Such discriminative techniques avoid the inherent cost involved in generative models which might require more information than is actually necessary.

For binary classification, support vector machines involve drawing a vector that's orthogonal (normal) to the hyperplane separating the training data. There are, in general, many hyperplanes that can separate the data; we want to draw the hyperplane that separates the data the most - we wish to choose the line that maximizes the minimum distance to this hyperplane across all data samples. This optimal hyperplane is the decision boundary, and the minimum distance is the margin. Once the decision boundary and margins are determined, data points right on the boundary of margins that keep us from expanding the margin any further are called support vectors. If we add new data points outside the margin or remove points that are not support vectors, the maximum margin would not be changed.

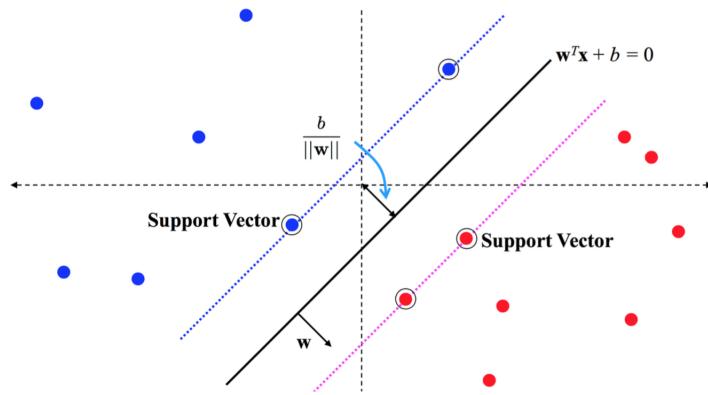


Figure 14.1: Support Vector Machines

Consider the simplest case where we have a binary classification with linear separator. Suppose that the vector orthogonal to the decision boundary is  $\mathbf{w}$ , and that the decision boundary is expressed as  $\mathbf{w}^T \mathbf{x}_i - b = 0$ . Then a set of points  $\mathbf{x}_i$  are classified as +1 if  $\mathbf{w} \cdot \mathbf{x}_i - b$  is greater than 0, and -1 otherwise. It can be shown that the optimal  $\mathbf{w}$  can be written as a linear combination of the data vectors  $\sum_i a_i \mathbf{x}_i$ . To classify new data points  $\mathbf{x}_{\text{new}}$ , we need to take the dot product of  $\mathbf{w}$  with  $\mathbf{x}_i$  to arrive at a scalar. Notice that this scalar,  $\sum_i a_i (\mathbf{x}_i \cdot \mathbf{x}_{\text{new}})$  only depends on the dot product of input data. Furthermore, it can be shown that finding the maximum margin hyperplane for a set of (training) points amounts to maximizing a linear program where the objective function only depends on the dot product of the training points with each other. This is good because it tells us that the complexity of solving that linear program is independent of the dimension of data points. If we pre-compute the pairwise dot products of the training vectors, then it makes no difference what the dimensionality of the data is in regards to the running time of solving the linear program.

### 14.3.1 Kernels

We covered the case when data are linearly separable in low dimensions. Here we use kernel to transform data to higher dimensional space  $H$  using function  $\Phi(\mathbf{x}) : \mathbb{R}^d \rightarrow H$ . Since SVMs are dependent only on the dot product of the vectors, for two-dimensional input vectors  $\mathbf{x}_i \cdot \mathbf{x}_j$  are transformed into  $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$ .

Note that the dot product of  $\Phi$  creates a scalar  $K$  in high dimensions, i.e.  $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$ . We can thus just use  $K$  without knowing  $\Phi(\mathbf{x})$  explicitly.

Some common kernels are :

1. Linear kernel:  $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$  which represents the trivial mapping of  $\Phi(x) = x$ .
2. Polynomial kernel:  $K(\mathbf{x}_i, \mathbf{x}_j) = (r + \gamma \mathbf{x}_i^T \mathbf{x}_j)^d$ .
3. Radial basis kernel:  $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$ . This transformation is actually from a point  $x$  to a function (which can be thought of as being a point in Hilbert space) in an infinite-dimensional space. So we are transforming our training set into functions, and combining them to get a decision boundary. The functions are Gaussians centered at the input points.
4. Sigmoid kernel:  $K(\mathbf{x}_i, \mathbf{x}_j) = \tanh[\gamma(\mathbf{x}_i^T \mathbf{x}_j + r)]$ . Sigmoid kernels have been popular for use in SVMs due to their origin in neural networks (e.g. sigmoid kernel functions are equivalent to two-level, perceptron neural networks). It has been pointed out in previous work (Vapnik 1995) that the kernel matrix may not be positive semi-definite for certain values of the parameters  $\mu$  and  $r$ . The sigmoid kernel has nevertheless been used in practical applications [2].

Here is a specific example of a kernel function. Consider the two classes of one-dimensional data:

$$\{-5, -4, -3, 3, 4, 5\} \text{ and } \{-2, -1, 0, 1, 2\}$$

This data is clearly not linearly separable, and the best separation boundary we can find might be  $x > -2.5$ . Now consider applying the transformation . The data can now be written as new pairs,

$$\{-5, -4, -3, 3, 4, 5\} \rightarrow \{(-5, 25), (-4, 16), (-3, 9), (3, 9), (4, 16), (5, 25)\} \text{ and } \{-2, -1, 0, 1, 2\} \rightarrow \{(-2, -4), (-1, 1), (0, 0), (1, 2)\}$$

This data is separable by the rule  $y > 6.5$ , and in general the more dimensions we transform data to the more separable it becomes.

An alternate way of thinking of this problem is to transform the classifier back in to the original low-dimensional space. In this particular example, we would get the rule  $x^2 < 6.5$  , which would bisect the number line at two points. In general, the higher dimensionality of the space that we transform to, the more complicated a classifier we get when we transform back to the original space.

One of the caveats of transforming the input data using a kernel is the risk of overfitting (or over-classifying) the data. More generally, the SVM may generate so many feature vector dimensions that it does not generalize well to other data. To avoid overfitting, cross-validation is typically used to evaluate the fitting provided by each parameter set tried during the grid or pattern search process. In the radial-basis kernel, you can essentially increase the value of  $\beta$  until each point is within its own classification region (thereby defeating the classification process altogether). SVMs generally avoid this problem of over-fitting due to the fact that they maximize margins between data points.

When using difficult-to-separate training sets, SVMs can incorporate a cost parameter  $C$ , to allow some flexibility in separating the categories. This parameter controls the trade-off between allowing training errors and forcing rigid margins. It can thereby create a *soft* margin that permits some misclassifications. Increasing the value of  $C$  increases the cost of misclassifying points and forces the creation of a more accurate model that may not generalize well.

Can we use just any function as our kernel? The answer to this is provided by Mercer's Condition which provides us an analytic criterion for choosing an acceptable kernel[3]. Mercer's Condition states that a kernel  $K(x, y)$  is a valid kernel if and only if the following holds. For any  $g(x)$  such that  $\int g(x)^2 dx$  is finite, we have:

$$\int \int K(x, y)g(x)g(y)dxdy \geq 0$$

In all, we have defined SVM discriminators and shown how to perform classification with appropriate kernel mapping functions that allow performing computations on lower dimension while being to capture all the information available at higher dimensions. The next section describes the application of SVMs to the classification of tumors for cancer diagnostics.

### 14.3.2 Open Problems

Although the kernel trick

Genes	Rejects	Errors	Confidence level	$ d $
7129	3	0	93%	0.1
40	0	0	93%	0.1
5	3	0	92%	0.1

## 14.4 Tumor Classification with SVMs

A generic approach for classifying two types of acute leukemias (acute myeloid leukemia (AML) and acute lymphoid leukemia (ALL)) was presented by Golub et al.[4]. This approach centered on effectively addressing three main issues:

1. Whether there were genes whose expression pattern to be predicted was strongly correlated with the class distinction (i.e. can ALL and AML be distinguished)
2. How to use a collection of known samples to create a “class predictor” capable of assigning a new sample to one of two classes
3. How to test the validity of their class predictors

They addressed (1) by using a “neighbourhood analysis” technique to establish whether the observed correlations were stronger than would be expected by chance. This analysis showed that roughly 1100 genes were more highly correlated with the AML-ALL class distinction than would be expected by chance. To address (2) they developed a procedure that uses a fixed subset of “informative genes” (chosen based on their correlation with the class distinction of AML and ALL) and makes a prediction based on the expression level of these genes in a new sample. Each informative gene casts a “weighted vote” for one of the classes, with the weight of each vote dependent on the expression level in the new sample and the degree of that gene’s correlation with the class distinction. The votes are summed to determine the winning class. To address (3) and effectively test their predictor by first testing by cross-validation on the initial data set and then assessing its accuracy on an independent set of samples. Based on their tests, they were able to identify 36 of the 38 samples (which were part of their training set!) and all 36 predictions were clinically correct. On the independent test set 29 of 34 samples were strongly predicted with 100% accuracy and 5 were not predicted.

A SVM approach to this same classification problem was implemented by Mukherjee et al.[5]. The output of classical SVM is a binary class designation. In this particular application it is particularly important to be able to reject points for which the classifier is not confident enough. Therefore, the authors introduced a confidence interval on the output of the SVM that allows for rejection of points with low confidence values. As in the case of Golub et al.[4] it was important for the authors to infer which genes are important for the classification. The SVM was trained on the 38 samples in the training set and tested on the 34 samples in the independent test set (exactly in the case of Golub et al.). The authors’ results are summarized in the following table (where  $|d|$  corresponds to the cutoff for rejection).

These results a significant improvement over previously reported techniques, suggesting that SVMs play an important role in classification of large data sets (as those generated by DNA microarray experiments).

## 14.5 Semi-Supervised Learning

In some scenarios we have a data set with only a few labeled data points, a large number of unlabeled data points and inherent structure in the data. This type of scenario both clustering and classification do not perform well and a hybrid approach is required. This semi-supervised approach could involve the clustering of data first followed by the classification of the generated clusters.

## 14.6 Current Research Directions

### 14.7 Further Reading

- Bishop, Christopher M. (2006) Pattern Recognition and Machine Learning, Springer.
- Richard O. Duda, Peter E. Hart, David G. Stork (2001) Pattern classification (2nd edition), Wiley, New York
- See previous chapter for more books and articles.

### 14.8 Resources

- Statistical Pattern Recognition Toolbox for Matlab.
- See previous chapter for more tools

## Bibliography

- [1] Calvo, S., Jain, M., Xie, X., Sheth, S.A., Chang, B., Goldberger, O.A., Spinaz- zola, A., Zeviani, M., Carr, S.A., and Mootha, V.K. (2006). Systematic identifi- cation of human mitochondrial disease genes through integrative genomics. *Nat. Genet.* 38, 576–582.
- [2] Scholokopf, B., et al., 1997. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing*.
- [3] Christopher J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.
- [4] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, and C. D. Bloomfield. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.
- [5] S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, J. P. Mesirov, and T. Poggio. Support vector machine classification of microarray data. Technical report, AI Memo 1677, Massachusetts Institute of Technology, 1998.



---

CHAPTER  
**FIFTEEN**

---

## REGULATORY MOTIFS, GIBBS SAMPLING, AND EM

Erin Hong and Rushina Shah (2016)  
Jenny Lin (2014)  
Maria Alexis (2013)  
James Yeung (2012)  
Yinqing Li and Arvind Thiagarajan (2011)  
Bianca Dumitrascu and Neal Wadhwa (2010)  
Joseph Lane (2009)  
Brad Cater and Ethan Heilman (2008)

### Figures

---

15.1 Transcription factors binding to DNA at a motif site . . . . .	245
15.2 Example Profile Matrix . . . . .	245
15.3 Iteration of E and M Steps . . . . .	246
15.4 Examples of the Z matrix computed . . . . .	247
15.5 Selecting motif location: the greedy algorithm will always pick the most probable location for the motif. The EM algorithm will take an average while Gibbs Sampling will actually use the probability distribution given by $Z$ to sample a motif in each step . . . . .	247
15.6 Sample position weight matrix . . . . .	247
15.7 Sample Background Matrix . . . . .	248
15.8 MEME Landscape . . . . .	249
15.9 Gibbs Sampling . . . . .	250
15.10 Discovering motifs by testing for conservations in many regions . . . . .	252
15.11 Examples of the Z matrix computed via EM, Gibbs Sampling, and the Greedy Algorithm . . . . .	252
15.12 Selecting motif location: the greedy algorithm will always pick the most probable location for the motif. The EM algorithm will take an average while Gibbs Sampling will actually use the probability distribution given by $Z$ to sample a motif in each step . . . . .	253
15.13 Sequences with zero, one or two motifs . . . . .	254
15.14 Entropy is maximized when both heads and tails have an equal chance of occurring . . . . .	255
15.15 The height of each stack represents the number of bits of information that Gibbs sampling or EM told us about the position in the motif . . . . .	255
15.16 lexA binding site assuming low G-C content and using K-L distance . . . . .	256

---

### 15.1 Introduction to regulatory motifs and gene regulation

We have already explored the areas of dynamic programming, sequence alignment, sequence classification and modeling, hidden Markov models, and expectation maximization. In the following chapter, we will look at how these techniques are also useful in identifying novel motifs and elucidating their functions.

### 15.1.1 The regulatory code: Transcription Factors and Motifs

Motifs are short (6-8 bases long), recurring patterns that have well-defined biological functions. Motifs include DNA patterns in enhancer regions or promoter motifs, as well as motifs in RNA sequences such as splicing signals. As we have discussed, genetic activity is regulated in response to environmental variations. Motifs are responsible for recruiting transcription factors, or regulatory proteins, to the appropriate target gene. Motifs can also be recognized by microRNAs, which bind to motifs given through complementarity; nucleosomes, which recognize motifs based on their GC content; and other RNAs, which use a combination of DNA sequence and structure. Once bound, they can activate or repress the expression of the associated gene.

Transcription factors (TFs) can use several mechanisms in order to control gene expression, including acetylation and deacetylation of histone proteins, recruitment of cofactor molecules to the TF-DNA complex, and stabilization or disruption of RNA-DNA interfaces during transcription. They often regulate a group of genes that are involved in similar cellular processes. Thus, genes that contain the same motif in their upstream regions are likely to be related in their functions. In fact, many regulatory motifs are identified by analyzing the regions upstream of genes known to have similar functions.

Motifs have become exceedingly useful for defining genetic regulatory networks and deciphering the functions of individual genes. Given their important role in gene regulation, a single mutation in a motif can lead to changes in phenotype and the onset of disease. For instance, a single motif polymorphism prevents the binding of a repressor whose gene helps decide between burning and storing energy, leading to obesity. Similarly, a single mutation can lead to diabetes in which a C → T substitution mutates the ARID5B motif. Thus, deciphering these relationships can help us better understand the causes for human disease, motivating the discovery of motifs. With our current computational abilities, regulatory motif discovery and analysis has progressed considerably and remains at the forefront of genomic studies.

### 15.1.2 Challenges of motif discovery

Before we can get into algorithms for motif discovery, we must first understand the characteristics of motifs, especially those that make motifs somewhat difficult to find. As mentioned above, motifs are generally very short, usually only 6-8 base pairs long. Additionally, motifs can contain any set of nucleotides (except for the starting codon ATG and other rules). However, when several proteins are necessary to regulate gene expression, the search base is significantly reduced since it is less likely for a series of approximately 6 base pairs to happen by random chance. Over time, motifs can become degenerate, where only the nucleotides at certain locations within the motif affect the motif's function. This degeneracy arises because transcription factors are free to interact with their corresponding motifs in manners more complex than a simple complementarity relation. As seen in 15.1, many proteins interact with the motif not by opening up the DNA to check for base complementarity, but instead by scanning the spaces, or grooves, between the two sugar phosphate backbones. Depending on the physical structure of the transcription factor, the protein may only be sensitive to the difference between purines and pyrimidines or weak and strong bases, as opposed to identifying specific base pairs. The topology of the transcription factor may even make it such that certain nucleotides aren't interacted with at all, allowing those bases to act as wildcards.

This issue of degeneracy within a motif poses a challenging problem. If we were only looking for a fixed k-mer, we could simply search for the k-mer in all the sequences we are looking at using local alignment tools. However, the motif may vary from sequence to sequence. Because of this, a string of nucleotides that is known to be a regulatory motif is said to be an *instance* of a motif because it represents one of potentially many different combinations of nucleotides that fulfill the function of the motif.

In our approaches, we make two assumptions about the data. First, we assume that there are no pairwise correlations between bases, i.e. that each base is independent of every other base. While such correlations do exist in real life, considering them in our analysis would lead to an exponential growth of the parameter space being considered, and consequently we would run the risk of overfitting our data. The second assumption we make is that all motifs have fixed lengths; indeed, this approximation simplifies the problem greatly. Even with these two assumptions, however, motif finding is still a very challenging problem. The relatively small size of motifs, along with their great variety, makes it fairly difficult to locate them. When the length of the motif is unknown, sequences bound by the same regulator are often aligned with allowed gaps in order to see if there is a visible commonality. In addition, a motif's location relative to the corresponding gene is far

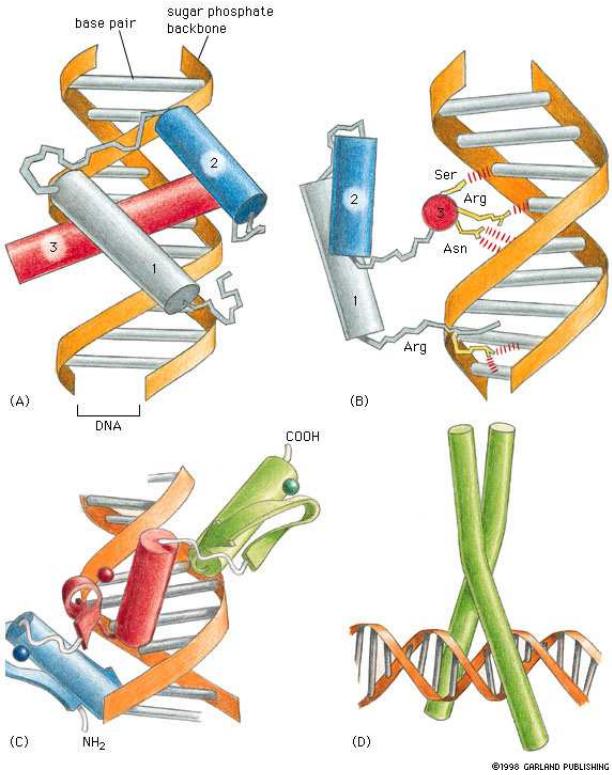


Figure 15.1: Transcription factors binding to DNA at a motif site

from fixed; the motif can be upstream or downstream, and the distance between the gene and the motif also varies. Indeed, sometimes the motif is as far as  $10k$  to  $10M$  base pairs from the gene.

### 15.1.3 Motifs summarize TF sequence specificity

Because motif instances exhibit great variety, we generally use a Position Weight Matrix (PWM) to characterize the motif. This matrix gives the frequency of each base at each location in the motif. The figure below shows an example PWM, where  $p_{ck}$  corresponds to the frequency of base  $c$  in position  $k$  within the motif, with  $p_{c0}$  denoting the distribution of bases in non-motif regions.

	sequence positions							
	1	2	3	4	5	6	7	8
A	0.1	0.3	0.1	0.2	0.2	0.4	0.3	0.1
C	0.5	0.2	0.1	0.1	0.6	0.1	0.2	0.7
G	0.2	0.2	0.6	0.5	0.1	0.2	0.2	0.1
T	0.2	0.3	0.2	0.2	0.1	0.3	0.3	0.1

Figure 15.2: Example Profile Matrix

We now define the problem of motif finding more rigorously. We assume that we are given a set of co-regulated and functionally related genes. Many motifs were previously discovered by doing footprint

experiments, which isolate sequences bound by specific transcription factors, and therefore more likely to correspond to motifs. There are several computational methods that can be used to locate motifs:

1. Perform a local alignment across the set of sequences and explore the alignments that resulted in a very high alignment score.
2. Model the promoter regions using a Hidden Markov Model and then use a generative model to find non-random sequences.
3. Reduce the search space by applying prior knowledge for what motifs should look like.
4. Search for conserved blocks between different sequences.
5. Examine the frequency of kmers across regions highly likely to contain a motif.
6. Use probabilistic methods, such as EM, Gibbs Sampling, or a greedy algorithm

Method 5, using relative kmer frequencies to discover motifs, presents a few challenges to consider. For example, there could be many common words that occur in these regions that are in fact not regulatory motifs but instead different sets of instructions. Furthermore, given a list of words that could be a motif, it is not certain that the most likely motif is the most common word; for instance, while motifs are generally overrepresented in promoter regions, transcription factors may be unable to bind if an excess of motifs are present. One possible solution to this problem might be to find kmers with maximum relative frequency in promoter regions as compared to background regions. This strategy is commonly performed as a post processing step to narrow down the number of possible motifs.

In the next section, we will talk more about these probabilistic algorithms as well as methods to use kmer frequency for motif discovery. We will also come back to the idea of using kmers to find motifs in the context of using evolutionary conservation for motif discovery.

## 15.2 Expectation maximization

### 15.2.1 The key idea behind EM

We are given a set of co-regulated/functionally related sequences with the assumption that motifs are enriched in them. The task is to find the common motif in those sequences. The key idea behind the following probabilistic algorithms is that if we were given motif starting positions in each sequence, finding the motif PWM would be trivial; similarly, if we were given the PWM for a particular motif, it would be easy to find the starting positions in the input sequences. Let  $Z$  be the matrix in which  $Z_{ij}$  corresponds to the probability that a motif instance starts at position  $j$  in sequence  $i$  (a graphical of the probability distributions summarized in  $Z$  is shown in Figure 15.11). These algorithms therefore rely on a basic iterative approach: given a motif length  $L$  and an initial matrix  $Z$ , we can use the starting positions to estimate the motif, and in turn use the resulting motif to re-estimate the starting positions, iterating over these two steps until convergence on a motif as seen in Figure 15.3.

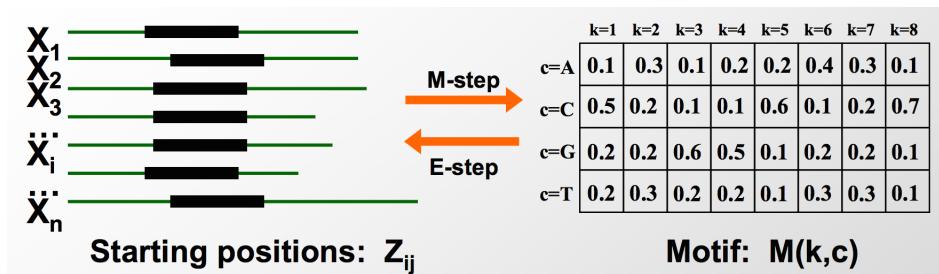


Figure 15.3: Iteration of E and M Steps

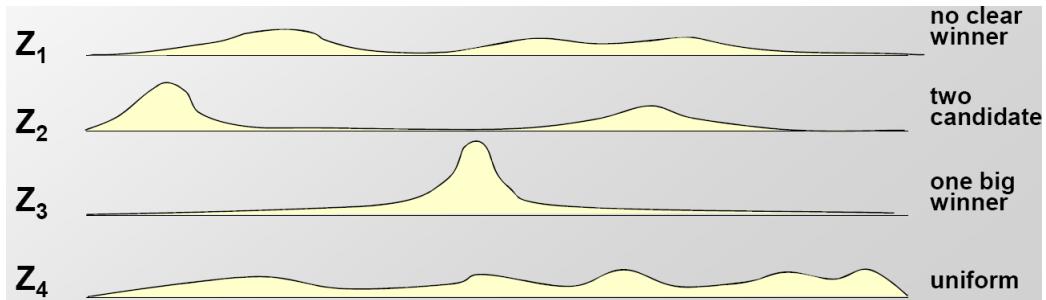


Figure 15.4: Examples of the Z matrix computed

### 15.2.2 The E step: Estimating $Z_{ij}$ from the PWM

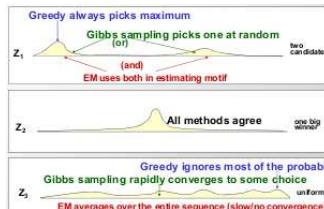


Figure 15.5: Selecting motif location: the greedy algorithm will always pick the most probable location for the motif. The EM algorithm will take an average while Gibbs Sampling will actually use the probability distribution given by  $Z$  to sample a motif in each step

**Step 1: Initialization** The first step in EM is to generate an initial probability weight matrix (PWM). The PWM describes the frequency of each nucleotide at each location in the motif. In 15.6, there is an example of a PWM. In this example, we assume that the motif is eight bases long.

If you are given a set of aligned sequences and the location of suspected motifs within them, then finding the PWM is accomplished by computing the frequency of each base in each position of the suspected motif. We can initialize the PWM by choosing starting locations randomly.

We refer to the PWM as  $p_{ck}$ , where  $p_{ck}$  is the probability of base  $c$  occurring in position  $k$  of the motif. Note: if there is 0 probability, it is generally a good idea to insert pseudo- counts into your probabilities. The PWM is also called the profile matrix. In addition to the PWM, we also keep a background distribution  $p_{ck,k=0}$ , a distribution of the bases not in the motif as seen in 15.7.

	sequence positions							
	1	2	3	4	5	6	7	8
A	0.1	0.3	0.1	0.2	0.2	0.4	0.3	0.1
C	0.5	0.2	0.1	0.1	0.6	0.1	0.2	0.7
G	0.2	0.2	0.6	0.5	0.1	0.2	0.2	0.1
T	0.2	0.3	0.2	0.2	0.1	0.3	0.3	0.1

Figure 15.6: Sample position weight matrix

<b>B =</b>	<b>A</b>	<b>0.26</b>
	<b>C</b>	<b>0.24</b>
	<b>G</b>	<b>0.23</b>
	<b>T</b>	<b>0.27</b>

Figure 15.7: Sample Background Matrix

**Step 2: Expectation** In the expectation step, we generate a vector  $Z_{ij}$  which contains the probability of the motif starting in position  $j$  in sequence  $i$ . In EM, the  $Z$  vector gives us a way of classifying all of the nucleotides in the sequences and tell us whether they are part of the motif or not. We can calculate  $Z_{ij}$  using Bayes' Rule. This simplifies to:

$$Z_{ij}^t = \frac{Pr^t(X_i|Z_{ij})Pr^t(Z_{ij}=1)}{\sum_{k=1}^{L-W+1} Pr^t(X_i|Z_{ik}=1)Pr^t(Z_{ik}=1)}$$

where  $Pr^t(X_i|Z_{ij}=1) = Pr(X_i|Z_{ij}=1, p)$  is defined as

$$\Pr(X_i | Z_{ij} = 1, p) = \underbrace{\prod_{k=1}^{j-1} p_{c_k, 0}}_{\text{before motif}} \underbrace{\prod_{k=j}^{j+W-1} p_{c_k, k-j+1}}_{\text{motif}} \underbrace{\prod_{k=j+W}^L p_{c_k, 0}}_{\text{after motif}}$$

This is the probability of sequence  $i$  given that the motif starts at position  $j$ . The first and last products correspond to the probability that the sequences preceding and following the candidate motif come from some background probability distribution whereas the middle product corresponds to the probability that the candidate motif instance came from a motif probability distribution. In this equation, we assume that the sequence has length  $L$  and the motif has length  $W$ .

### 15.2.3 M step: Finding the maximum likelihood motif from starting positions $Z_{ij}$

**Step 3: Maximization** Once we have calculated  $Z^t$ , we can use the results to update both the PWM and the background probability distribution. We can update the PWM using the following equation

Remember that  $M(k,c)$  represents the probability of nucleotide  $c$  at position  $k$

$$M^{(t+1)}(k, c) = \frac{n_{k,c} + d}{\sum_c (n_{k,c} + d)}$$

**where**  $n_{c,k} = \sum_i \sum_{\{j|X_{i,j+k-1}=c\}} Z_{ij}$

pseudo-counts

**Step 4: Repeat** Repeat steps 2 and 3 until convergence.

One possible way to test whether the profile matrix has converged is to measure how much each element in the PWM changes after step maximization. If the change is below a chosen threshold, then we can terminate the algorithm. EM is a deterministic algorithm and is entirely dependent on the initial starting points because it uses an average over the full probability distribution. Another good thing to note is that EM often converges after a small number of iterations when local maximum has been

reached.

It is therefore advisable to rerun the algorithm with different initial starting positions to try reduce the chance of converging on a local maximum that is not the global maximum and to get a good sense of the solution space. One possible solution is MEME (Multiple EM for Motif Elicitation) which uses the idea of minimizing the effects of local maxima by starting at several points and running each for just one iteration. In figure 15.8, MEME starts at the shown green nodes, runs for one iteration, and then selects the starting point that got the "highest" and continues to run.

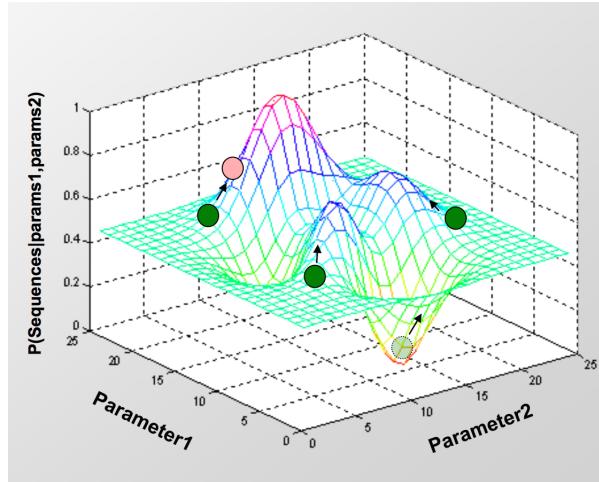


Figure 15.8: MEME Landscape

## 15.3 Gibbs Sampling: Sample from joint (M,Z<sub>ij</sub>) distribution

### 15.3.1 Sampling motif positions based on the Z vector

Gibbs sampling is a variation of generalized EM algorithm in that it is a stochastic process, while EM is deterministic. In the expectation step, we only consider nucleotides within the motif window in Gibbs sampling. In the maximization step, we sample from  $Z_{ij}$  and use the result to update the PWM instead of averaging over all values as in EM.

**Step 1: Initialization** As with EM, you generate your initial PWM with a random sampling of initial starting positions. The main difference lies in the Maximization step. During EM, the algorithm creates the sequence motif by considering all possible starting points of the motif. During Gibbs, the algorithm picks a single starting point of the motif with the probability of the starting point's  $Z$ .

**Step 2: Remove** Remove one sequence,  $X_i$ , from your set of sequences. You will change the starting location of for this particular sequence.

**Step 3: Update** Using the remaining set of sequences, update the PWM by counting how often each base occurs in each position, adding pseudocounts as necessary.

**Step 4: Sample** Using the newly updated PWM, compute the score of each starting point in the sequence  $X_i$ . To generate each score,  $Z_{ij}$ , the following formula is used:

$$A_{ij} = \frac{\prod_{k=j}^{j+W-1} p_{ck}, k - j + 1}{\prod_{k=j}^{j+W-1} p_{ck}, 0}$$

This is simply the probability that the sequence was generated using the motif PWM divided by the probability that the sequence was generated using the background PWM.

Select a new starting position for  $X_i$  by randomly choosing a position based on its  $Z_{ij}$ .

Note that Gibbs sampler takes a stochastic approach, and is not guaranteed to converge to the same alignment with different random seeds. One solution is to run the Gibbs sampler a couple of times to reach a satisfactory alignment.

**Step 5: Iterate** Loop back to Step 2 and iterate the algorithm until convergence.

### 15.3.2 More likely to find global maximum, easy to implement

Because Gibbs updates its sequence motif during Maximization based of a single sample of the Motif rather than every sample weighted by their scores, Gibbs is less dependent on the starting PWM. EM is much more likely to get stuck on a local maximum than Gibbs because of this fact. However, this does not mean that Gibbs will always return the global maximum. Gibbs must be run multiple times to ensure that you have found the global maximum and not the local maximum. Two popular implementations of Gibbs Sampling applied to this problem are AlignACE and BioProspector. A more general Gibbs Sampler can be found in the program WinBUGS. Both AlignACE and BioProspector use the aforementioned algorithm for several choices of initial values and then report common motifs. Gibbs sampling is easier to implement than E-M, and in theory, it converges quickly and is less likely to get stuck at a local optimum. However, the search is less systematic.

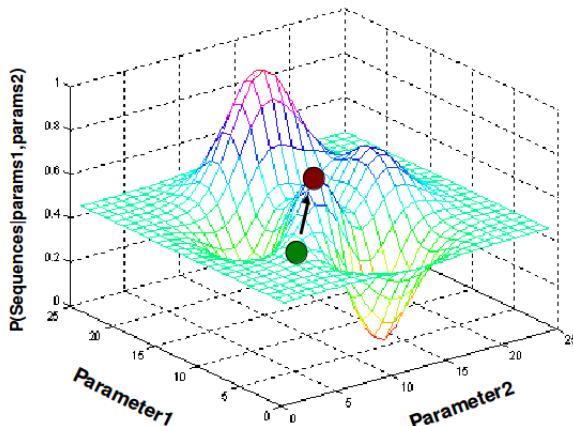


Figure 15.9: Gibbs Sampling

## 15.4 De novo motif discovery

As discussed in beginning of this chapter, the core problem for motif finding is to define the criteria for what is a valid motif and where they are located. Since most motifs are linked to important biological functions, one could subject the organism to a variety of conditions in hope of triggering these biological functions. One could then search for differentially expressed genes, and then use those genes as a basis for which genes are functionally related and thus likely to be controlled by the same motif instance. However, this technique not only relies on prior knowledge of interesting biological functions to probe for, but is also subject to biases in the experimental procedure. Alternatively, one could use ChIP-seq to search for motifs, but this method relies on not only having a known Transcription Factor of interest, but also requires developing antibodies to recognize said Transcription Factor, which can be costly and time consuming.

Ideally one would be able to discover motifs de novo, or without relying on an already known gene set or Transcription Factor. While this seems like a difficult problem, it can in fact be accomplished by taking advantage of genome-wide conservation. Because biological functions are usually conserved across species and have distinct evolutionary signatures, one can align sequences from close species and search specifically

in conserved regions (also known as Island of Conservation) in order to increase the rate of finding functional motifs.

### 15.4.1 Motif discovery using genome-wide conservation

Conservation islands often overlap known motifs, so doing genome-wide scans through evolutionary conserved regions can help us discover motifs, de novo. However, not all conserved regions will be motifs; for instance, nucleotides surrounding motifs may also be conserved even though they are not themselves part of a motif. Distinguishing motifs from background conserved regions can be done by looking for enrichments which will select more specifically for kmers involved in regulatory motifs. For instance, one can find regulatory motifs by searching for conserved sequences enriched in intergenic regions upstream of genes as compared to control regions such as coding sequences, since one would expect motifs to be enriched in or around promoters of genes. One can also expand this model to find degenerate motifs: we can look for conservation of smaller, non-degenerate motifs separated by a gap of variable length, as shown in the figure below. We can also extend this motif through a greedy search in order to get closer to find the local maximum likelihood motif. Finally, evolution of motifs can also reveal which motifs are degenerate; since a particular motif is more likely to be degenerate if it is often replaced by another motif throughout evolution, motif clustering can reveal which kmers are likely to correspond to the same motif.

In fact, the strategy has its biological relevance. In 2003, Professor Kellis argued that there must be some selective pressure to cause a particular sequence to occur on specific places. His PhD. thesis on the topic can be found at the following location: [http://web.mit.edu/www/publications/Kellis\\_Thesis\\_03.pdf](http://web.mit.edu/www/publications/Kellis_Thesis_03.pdf).

### 15.4.2 Validation of discovered motifs with functional datasets

These predicted motifs can then be validated with functional datasets. Predicted motifs with at least one of the following features are more likely to be real motifs: -enrichment in co-regulated genes. One can extend this further to larger gene groups; for instance, motifs have been found to be enriched in genes expressed in specific tissues -overlap with TF binding experiments -enrichment in genes from the same complex -positional biases with respect to the transcription start site (TSS): motifs are enriched in gene TSS's -upstream vs. downstream of genes, inter- vs. intra-genic positional biases: motifs are generally depleted in coding sequences -similarity to known transcription factor motifs: some, but not all, discovered motifs may match known motifs (however, not all motifs are conserved and *known* motifs may not be exactly correct)

Fill in or remove as wanted

expand

## 15.5 Evolutionary signatures for regulatory motifs

We can start by looking at known motif instances across various species. Using evolutionary evidence, the islands of conservation can be motifs with added confidence by testing many regions for conserved islands across various species. Taking any conservation island as a motif would be incorrect as conservation islands can arise by chance or even non-motif sequences.

## 15.6 Phylogenies, Branch length score as Confidence score

### 15.6.1 Foreground vs. background. Real vs. control motifs.

## 15.7 Possibly deprecated stuff below:

### 15.7.1 Greedy

While the greedy algorithm is not used very much in practice, it is important know how it functions and mainly its advantages and disadvantages compared to EM and Gibbs sampling. The Greedy algorithm works just like Gibbs sampling except for a main difference in Step 4. Instead of randomly choosing selecting a new starting location, it always picks the starting location with the highest probability.

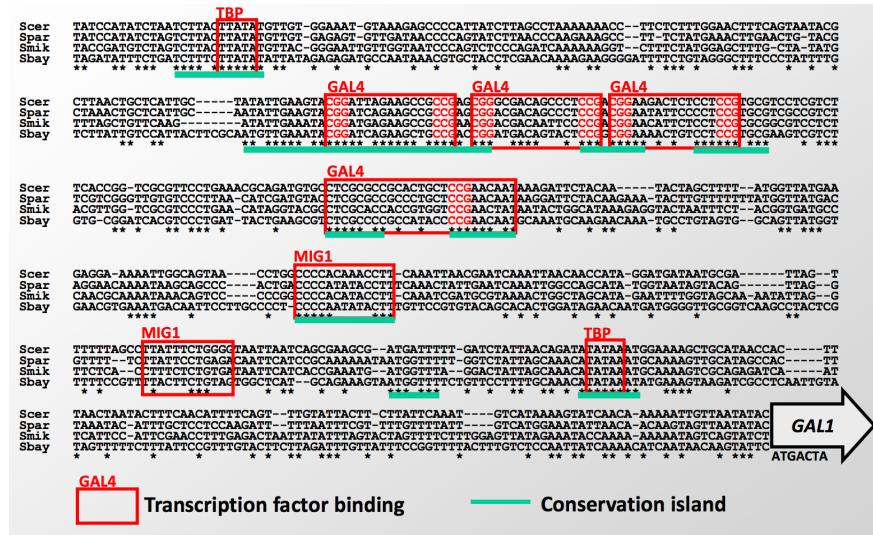


Figure 15.10: Discovering motifs by testing for conservations in many regions

This makes the Greedy algorithm slightly faster than Gibbs sampling but reduces its chances of finding a global maximum considerably. In cases where the starting location probability distribution is fairly evenly distributed, the greedy algorithm ignores the weights of every other starting position other than the most likely.

## 15.8 Comparing different Methods

The main difference between Gibbs, EM, and the Greedy algorithm lies in their maximization step after computing their  $Z$  matrix. Examples of the  $Z$  matrix are graphically represented below. This  $Z$  matrix is then used to recompute the original profile matrix until convergence. Some examples of this matrix are graphically represented by 15.11

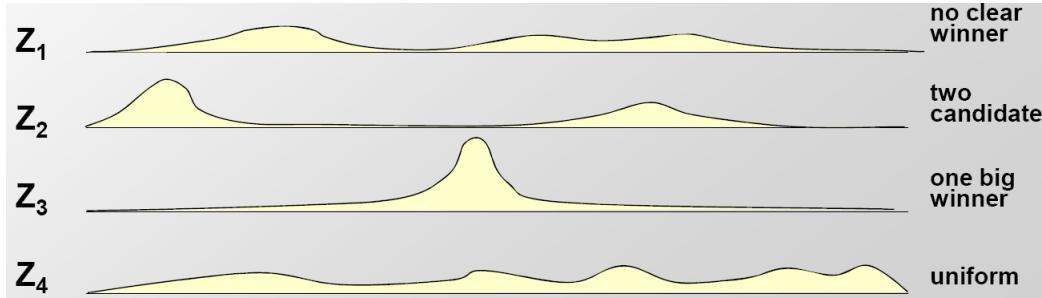


Figure 15.11: Examples of the  $Z$  matrix computed via EM, Gibbs Sampling, and the Greedy Algorithm

Intuitively, the greedy algorithm will always pick the most probable location for the motif. The EM algorithm will take an average of all values while Gibbs Sampling will actually use the probability distribution given by  $Z$  to sample a motif in a step.

## 15.9 OOPS,ZOOPS,TCM

The different types of sequence model make differing assumptions about how and where motif occurrences appear in the dataset. The simplest model type is OOPS (One-Occurrence-Per-Sequence) since it assumes

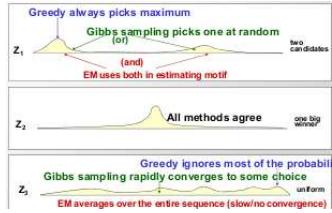


Figure 15.12: Selecting motif location: the greedy algorithm will always pick the most probable location for the motif. The EM algorithm will take an average while Gibbs Sampling will actually use the probability distribution given by  $Z$  to sample a motif in each step

that there is exactly one occurrence per sequence of the motif in the dataset. This is the case we have analyzed in the Gibbs sampling section. This type of model was introduced by Lawrence & Reilly (1990) [2], when they describe for the first time a generalization of OOPS, called ZOOPS (Zero-or-One-Occurrence-Per-Sequence), which assumes zero or one motif occurrences per dataset sequence. Finally, TCM (Two-Component Mixture) models assume that there are zero or more non-overlapping occurrences of the motif in each sequence in the dataset, as described by Baily & Elkan (1994). [1] Each of these types of sequence model consists of two components, which model, respectively, the motif and non-motif (background) positions in sequences. A motif is modelled by a sequence of discrete random variables whose parameters give the probabilities of each of the different letters (4 in the case of DNA, 20 in the case of proteins) occurring in each of the different positions in an occurrence of the motif. The background positions in the sequence are modelled by a single discrete random variable.

## 15.10 Extension of the EM Approach

### 15.10.1 ZOOPS Model

The approach presented before (OOPS) relies on the assumption that every sequence is characterized by only one motif (e.g., there is exactly one motif occurrence in a given sequence). The ZOOPS model takes into consideration the possibility of sequences not containing motifs.

In this case let  $i$  be a sequence that does not contain a motif. This extra information is added to our previous model using another parameter  $\lambda$  to denote the prior probability that any position in a sequence is the start of a motif. Next, the probability of the entire sequence to contain a motif is  $\lambda = (L - W + 1) * \lambda$

#### The E-Step

The E-step of the ZOOPS model calculates the expected value of the missing information—the probability that a motif occurrence starts in position  $j$  of sequence  $X_i$ . The formulas used for the three types of model are given below.

$$Z_{ij}^t = \frac{\Pr^{(t)}(X_i | Z_{ij} = 1)\lambda^{(t)}}{\Pr^{(t)}(X_i | Q_i = 0)(1 - \lambda^{(t)}) + \sum_{k=1}^{L-W+1} \Pr^{(t)}(X_i | Z_{ik} = 1)\lambda^{(t)}}$$

where  $\lambda^t$  is the probability that sequence  $i$  has a motif,  $\Pr^t(X_i | Q_i = 0)$  is the probability that  $X_i$  is generated from a sequence  $i$  that does not contain a motif

#### The M-Step

The M-step of EM in MEME re-estimates the values for  $\lambda$  using the preceding formulas. The math remains the same as for OOPS, we just update the values for  $\lambda$  and  $\gamma$

$$\lambda^{(t+1)} = \frac{\gamma^{(t+1)}}{(L-W+1)} = \frac{1}{n(L-W+1)} \sum_{i=1}^n \sum_{j=1}^m Z_{i,j}^{(t)}$$

The model above takes into consideration sequences that do not have any motifs. The challenge is to also take into consideration the situation in which there is more than one motif per sequence. This can be accomplished with the more general model TCM. TCM (two-component mixture model) is based on the assumption that there can be zero, one, or even two motif occurrences per sequence.



Figure 15.13: Sequences with zero, one or two motifs.

### 15.10.2 Finding Multiple Motifs

All the above sequence model types model sequences containing a single motif (notice that TCM model can describe sequences with multiple occurrences of the same motif). To find multiple, non-overlapping, different motifs in a single dataset, one incorporates information about the motifs already discovered into the current model to avoid rediscovering the same motif. The three sequence model types assume that motif occurrences are equally likely at each position  $j$  in sequences  $x_i$ . This translates into a uniform prior probability distribution on the missing data variables  $Z_{ij}$ . A new prior on each  $Z_{ij}$  had to be used during the E-step that takes into account the probability that a new width- $W$  motif occurrence starting at position  $X_{ij}$  might overlap occurrences of the motifs previously found. To help compute the new prior on  $Z_{ij}$  we introduce variables  $V_{ij}$  where  $V_{ij} = 1$  if a width- $W$  motif occurrence could start at position  $j$  in the sequence  $X_i$  without overlapping an occurrence of a motif found on a previous pass. Otherwise  $V_{ij} = 0$ .

$$V_{ij} = \begin{cases} 1, & \text{no previous motifs in } [X_{i,j}, \dots, X_{i,j+w-1}] \\ 0, & \text{otherwise} \end{cases}$$

## 15.11 Motif Representation and Information Content

Instead of a Profile Matrix, we can also represent Motifs using information theory. In information theory, information about a certain event is communicated through a message. The amount of information carried by a message is measured in bits. We can determine the bits of information carried by a message by observing the probability distribution of the event described in the message. Basically, if we don't know anything about the outcome of the event, the message will contain a lot of bits. However, if we are pretty sure how the event is going to play out, and the message only confirms our suspicions, the message carries very few bits of information. For example, The sentence "the sun will rise tomorrow" is not very surprising, so the information of that sentence is quite low.. However, the sentence "the sun will not rise tomorrow" is very surprising and it has high information content. We can calculate the specific amount of information in a given message with the equation:  $-\log p$ .

Shannon Entropy is a measure of the expected amount of information contained in a message. In other words, it is the information contained by a message of every event that could possibly occur weighted by each event's probability. The Shannon entropy is given by the equation:

$$H(X) = - \sum_i p_i \log_2 p_i$$

Entropy is maximum when all events have an equal probability of occurring. This is because Entropy tells us the expected amount of information we will learn. If each even has the same chance of occurring we know as little as possible about the event, so the expected amount of information we will learn is maximized. For example, a coin flip has maximal entropy only when the coin is fair. If the coin is not fair, then we know more about the event of the coin flip, and the expected message of the outcome of the coin flip will contain less information.

### Example: Coin Toss

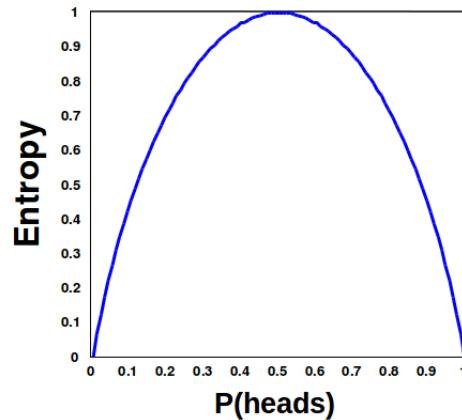


Figure 15.14: Entropy is maximized when both heads and tails have an equal chance of occurring

We can model a motif by how much information we have of each position after applying Gibbs Sampling or EM. In the following figure, the height of each letter represents the number of bits of information we have learned about that base. Higher stacks correspond to greater certainty about what the base is at that position of the motif while lower stacks correspond to a higher degree of uncertainty. With four codons to choice from, the Shannon Entropy of each position is 2 bits. Another way to look at this figure is that the height of a letter is proportional to the frequency of the base at that position.

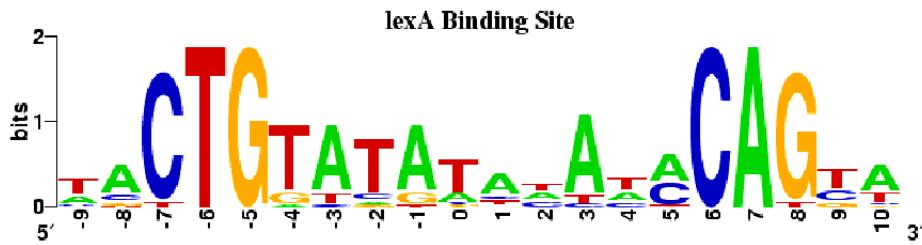


Figure 15.15: The height of each stack represents the number of bits of information that Gibbs sampling or EM told us about the postion in the motif

There is a distance metric on probability distributions known as the Kullback-Leibler distance. This allows us to compare the divergence of the motif distribution to some true distribution. The K-L distance is given by

$$D_{KL}(P_{motif} | P_{background}) = \sum_{A,T,G,C} P_{motif}(i) \log \frac{P_{motif}(i)}{P_{background}(i)}$$

In Plasmodium, there is a lower G-C content. If we assume a G-C content of 20%, then we get the following representation for the above motif. C and G bases are much more unusual, so their prevalence is highly unusual. Note that in this representation, we used the K-L distance, so that it is possible for the stack to be higher than 2.

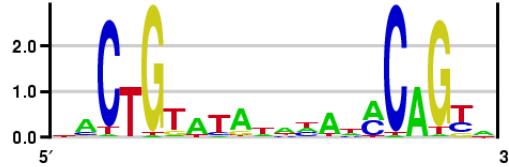


Figure 15.16: *lexA* binding site assuming low G-C content and using K-L distance

## Bibliography

- [1] Timothy L. Bailey. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 28–36. AAAI Press, 1994.
  - [2] C E Lawrence and A A Reilly. An expectation maximization (em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, 7(1):41–51, 1990.

---

CHAPTER  
**SIXTEEN**

---

## REGULATORY GENOMICS

Lecturer: Pouya Kheradpour  
Scribe: Maria Frendberg

### Figures

16.1 Challenges in Regulatory Genomics . . . . .	258
--	-----

---

## 16.1 Introduction

Every cell has the same DNA, but they all have different expression patterns due to the temporal and spatial regulation of genes. Regulatory genomics explains these complex gene expression patterns. The regulators we will be discussing are:

- *Transcription Factor (TF)*- Regulates transcription of DNA to mRNA. TFs are proteins which bind to DNA before transcription and either increase or decrease transcription. We can determine the specificity of a TF through experimental methods using protein or antibodies. We can find the genes by their similarity to known TFs.
- *Micro RNA (miRNA)*- Regulates translation of mRNA to Proteins. miRNAs are RNA molecules which bind to mRNA after transcription and can reduce translation. We can determine the specificity of a miRNA through experimental methods, such as cloning, or computational methods, using conservation and structure.

### 16.1.1 History of the Field

### 16.1.2 Open Problems

Both TFs and miRNAs are regulators and we can find them through both experimental and computational methods. We will discuss some of these computational methods, specifically the use of evolutionary signatures. These regulators bind to specific patterns, called motifs. We can predict the motifs to which a regulator will bind using both experimental and computational methods. We will be discussing identification of miRNAs through evolutionary and structural signatures and the identification of both TFs and miRNAs through de novo comparative discovery, which theoretically can find all motifs. Given a motif, it is difficult to find the regulator which binds to it.

A target is a place where a factor binds. There are many sequence motifs, however many will not bind; only a subset will be targets. Targets for a specified regulator can be determined using experimental methods. In Lecture 11, methods for finding a motif given a target were discussed. We will also discuss finding targets given a motif.

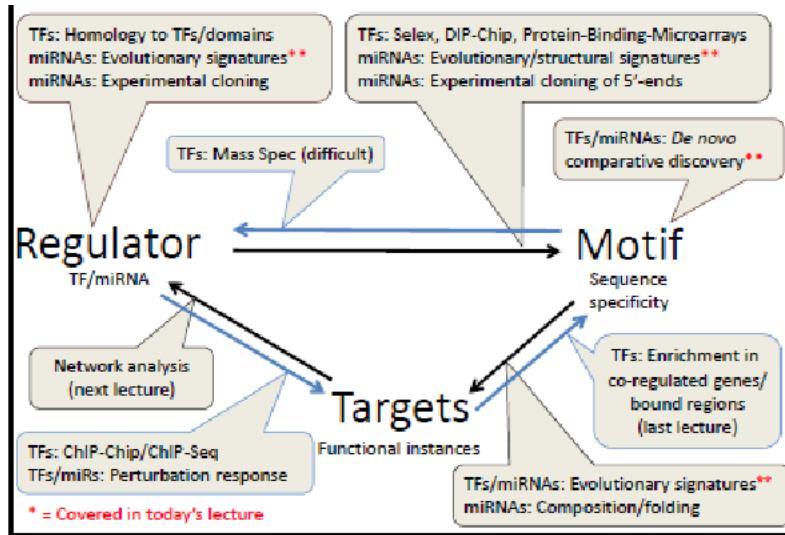


Figure 16.1: Challenges in Regulatory Genomics

## 16.2 *De Novo* Motif Discovery

### 16.2.1 TF Motif Discovery

Transcription Factors influence the expression of target genes as either activators or repressors by binding to the DNA near genes. This binding is guided by TF sequence specificity. The closer the DNA is to the base preference, the more likely it is that the factor will bind. These motifs can be found both computationally and experimentally. There are three main approaches for discovering these motifs.

- *Co-Regulation-* In Lecture 11, we discussed a co-regulation type of discovery of motifs by finding sequences which are likely to have the motif bound. We can then use enumerative approaches or alignment methods to find these motifs in the upstream regions. We can apply similar techniques to experimental data where you know where motif is bound.
- *Factor Centric-* There are also factor centric methods for discovering motifs. These are mostly experimental methods which require a protein or antibody. Examples include SELEX, DIP-Chip, and PBMs. All of these methods are *in vitro*.
- *Evolutionary-* Instead of focusing on only one factor, evolutionary methods focus on all factors. We can begin by looking at a single factor and determining which properties we can exploit. There are certain sequences which are preferentially conserved (conservation islands). However, these are not always motifs and instead can be due to chance or non-motif conservation. We can then look at many regions, find more conserved motifs, and determine which ones are more conserved overall. By testing conservation in many regions across many genomes, we increase the power. These motifs have certain evolutionary signatures that help us to identify them: motifs are more conserved in intergenic regions than in coding regions, motifs are more likely to be upstream from a gene than downstream. This is a method for taking a known motif and testing if it is conserved.

We now want to find everything that is more conserved than expected. This can be done using a hill climbing approach. We begin by enumerating the motif seeds, which are typically in 3-gap-3 form. Then, each of these seeds is scored and ranked using a conservation ratio corrected for composition and small counts. These seeds are then expanded to fill unspecified bases around the seed using hill climbing. Through these methods, it is possible to arrive at the same, or very similar seeds in different manners. Thus, our final step consists of clustering the seeds using sequence similarity to remove redundancy.

A final method that we can use is recording the frequency with which one sequence is “replaced” by another in evolution. This produces clusters of k-mers that correspond to a single motif.

### 16.2.2 Validating Discovered Motifs

There are many ways that we can validate discovered motifs. Firstly, we expect them to match real motifs, which does happen significantly more often than with random motifs. However, this is not a perfect agreement, possibly due to the fact that many known motifs are not conserved and that known motifs are biased and may have missed real motifs. Positional bias. Biased towards TSS,

Motifs also have functional enrichments. If a specific TF is expressed in a tissue, then we expect the upstream region will have that factor's motif. This also reveals modules of cooperating motifs. We also see that most motifs are avoided in ubiquitously expressed genes, so that they are not randomly turned on and off.

### 16.2.3 Summary

There are disadvantages to all of these approaches. Both TF and region-centric approaches are not comprehensive and are biased. TF centric approaches require a transcription factor or antibody, take lots of time and money, and also have computational challenges. De novo discovery using conservation is unbiased, but it can't match motifs to factors and requires multiple genomes.

## 16.3 Predicting Regular Targets

### 16.3.1 Motif Instance Identification

Once potential motifs are discovered, the next step is to discover which motif matches are real. This can be done by both experimental and computational methods.

- *Experimental* - Instances can be identified experimentally using ChIP-Chip and ChIP-Deq methods. Both of these are *in vivo* methods. This is done by cross linking cells. DNA is first broken into sections. Then the protein and its antibody or tagged protein is added, which binds to various sequences. These bound sequences are now pulled out and cross linking is reversed. This allows us to determine where in the genome the factor was bound. This has a high false positive rate because there are many instances where a factor binds, but is not functional. This is a very popular experimental methods, but it is limited by the availability of antibodies, which are difficult to get for many factors.
- *Computational*- Computation approaches. There are also many computational approaches to identify instances. Single genome approaches use motif clustering. They look for many matches to increase power and are able to find regulatory regions (CRMs). However, they miss instances of motifs that occur alone and require a set of specific factors that act together. Multi-genome approaches, known as phylogenetic footprinting, face many challenges. They begin by aligning many sequences, but even in functional motifs, sequences can move, mutate, or be missing. The approach taken by Kheradpour handles this by not requiring perfect conservation (by using a branch length score) and by not requiring an exact alignment (by searching within a window).

Branch Length Scores (BLS) are computed by taking a motif match and searching for it in other species. Then, the smallest subtree containing all species with a motif match is found. The percentage of total tree is the BLS. Calculating the BLS in this way allows for mutations permitted by motif degeneracy, misalignment and movement within a window, and missing motifs in dense species trees.

This BLS is then translated into a confidence score. This enables us to evaluate the likelihood of a given score and to account for differences in motif composition and length. We calculate this confidence score by counting all motif instances and control motifs at each BLS. We then want to see which fraction of the motif instances seem to be real. The confidence score is then  $\text{signal}/(\text{signal}+\text{noise})$ . The control motifs used in this calculation are produced by producing 100 shuffles of the original motif, and filtering the results by requiring that they match the genome with  $+/- 20\%$  of the original motif. These are then sorted based on their similarity to known motifs and clustered. At most one motif is taken from each cluster, in increasing order of similarity, to produce our control motifs.

### 16.3.2 Validating Targets

Similar to motif discovery, we can validate targets by seeing where they fall in the genome. Confidence selects for TF motif instances in promoters and miRNA motifs in 3' UTRs, which is what we expect. TFs can occur on either strand, whereas miRNA must fall on only one strand. Thus, although there is no preference for TFs, miRNA are found preferentially on the plus strand.

Another method of validating targets is by computing enrichments. This requires having a background and foreground set of regions. These could be a promoter of co-regulated genes vs all genes or regions bound by a factor vs other intergenic regions. Enrichment is computed by taking the fraction of motif instances inside the foreground vs the fraction of bases in the foreground. Composition and conservation level are corrected for with control motifs. These fractions can be made more conservative using a binomial confidence interval.

Targets can then be validated by comparing to experimental instances found using ChIP-Seq. This shows the conserved CTCF motif instances are highly enriched in ChIP-Seq sites. Increasing confidence also increases enrichment. Using this, many motif instances are verified. ChIP-Seq does not always find functional motifs, so these results can further be verified by comparing to conserved bound regions. This finds that enrichment in intersections is dramatically higher. This shows where factors are binding that have an effect worthwhile conserving in evolution. These two approaches are complementary and are even more effective when used together.

## 16.4 MicroRNA Genes and Targets

### 16.4.1 MiRNA Gene Discovery

MiRNAs are post-transcriptional regulators that bind to mRNAs to silence a gene. They are an extremely important regulator in development. These are formed when a miRNA gene is transcribed from the genome. The resulting strand forms a hairpin at some point. This is processed, trimmed and exported to the cytoplasm. Then, another protein trims the hairpin and one half is incorporated into a RISK complex. By doing this, it is able to tell the RISK complex where to bind, which determines which gene is turned off. The second strand is usually discarded. It is a computational problem to determine which strand is which. The computational problem here is how to find the genes which correspond to these miRNAs.

The first problem is finding hairpins. Simply folding the genome produces approximately 760,000 hairpins, but there are only 60 to 200 true miRNAs. Thus we need methods to help improve specificity. Structural features, including folding energy, loops (number, symmetry), hairpin length and symmetry, substructures and pairings, can be considered, however, this only increases specificity by a factor of 40. Thus structure alone cannot predict miRNAs. Evolutionary signatures can also be considered. MiRNA show characteristic conservation properties. Hairpins consist of a loop, two arms and flanking regions. In most RNA, the loop is the most well conserved due to the fact that it is used in binding. In miRNA, however, the arms are more conserved because they determine where the RISK complex will bind. This increases specificity by a factor of 300. Both these structural features and conservation properties can be combined to better predict potential miRNAs.

These features are combined using machine learning, specifically random forests. This produces many weak classifiers (decision trees) on subsets of positives and negatives. Each tree then votes on the final classification of a given miRNA. Using this technique allows us to reach the desired sensitivity (increased by 4,500 fold).

### 16.4.2 Validating Discovered MiRNAs

Discovered miRNAs can be validated by comparing to known miRNAs. An example given in class shows that 81% of discovered miRNAs were already known to exist, which shows that these methods perform well. The putative miRNAs have yet to be tested, however this can be difficult to do as testing is done by cloning.

Region specificity is another method for validating miRNAs. In the background, hairpins are fairly evenly distributed between introns, exons, intergenic regions, and repeats and transposons. Increasing confidence in

predictions causes almost all miRNAs to fall in introns and intergenic regions, as expected. These predictions also match sequencing reads.

This also produced some genomic properties typical of miRNAs. They have a preference for transcribed strand. This allows them to piggyback in intron of real gene, and thus not require a separate transcription. They also cluster with known and predicted miRNAs. This indicates that they are in the same family and have a common origin.

### **16.4.3 MiRNA's 5' End Identification**

The first seven bases determine where an miRNA binds, thus it is important to know exactly where cleavage occurs. If this cleavage point is wrong by even two bases, the miRNA will be predicted to bind to a completely different gene. These cleavage points can be discovered computationally by searching for highly conserved 7-mers which could be targets. These 7-mers also correlate to a lack of anti-targets in ubiquitously expressed genes. Using these features, structural features and conservational features, it is possible to take a machine learning approach (SVMs) to predict cleavage site. Some miRNAs have no single high scoring position, and these also show imprecise processing in the cell. If the star sequence is highly scored, then it tends to be more expressed in the cell also.

### **16.4.4 Functional Motifs in Coding Regions**

Each motif type has distinct signatures. DNA is strand symmetric, RNA is strand-specific and frame-invariant, and Protein is strand-specific and frame-biased. This frame-invariance can be used as a signature. Each frame can then be evaluated separately. Motifs due to di-codon usage biases are conserved in only one frame offset while motifs due to RNA-level regulation are conserved in all three frame offsets. This allows the ability to distinguish overlapping pressures.

## **16.5 Current Research Directions**

### **16.6 Further Reading**

### **16.7 Tools and Techniques**

### **16.8 What Have We Learned?**

## **Bibliography**



---

CHAPTER  
SEVENTEEN

---

EPIGENOMICS/CHROMATIN STATES

**Figures**

---

17.1 A. There is a wide diversity of modifications in the epigenome. Some regions of DNA are compactly wound around histones, making the DNA inaccessible and the genes inactive. Other regions have more accessible DNA and thus active genes. Epigenetic factors can bind to the tails of these histones to modify these properties. B. Histone modifications provide information about what types of proteins are bound to the DNA and what the function of the region is. In this example, The histone modifications allow for an enhancer region (potentially over 100 kilo bases away) to interact with the promoter region. [7] . . . . .	267
17.2 The method of chromatin immunoprecipitation [5]. . . . .	268
17.3 (Top) In the Burrows-Wheeler forward transformation rotations are generated and sorted. The last column of the sorted list (bolded) consists of the transformed string. (Bottom)In the Burrows-Wheeler reverse transformation the transformed string is sorted, and two columns are generated: one consisting of the original string and the other consisting of the sorted. These effectively form two columns from the rotations in the forward transformation. This process is repeated until the complete rotations are generated. . . . .	269
17.4 To use input DNA as a control, one can run the ChIP experiment as normal while simultaneously running the same experiment (with same DNA) without an antibody. This generates a background signal for which we can correct. . . . .	271
17.5 In the figure above each column is a color-coded histogram that encodes the fraction of all mapped reads that have base score Q (y-axis) at each position(x-axis). A low average per base score implies greater probability of mismappings. We typically reject reads whose average score Q is less than 10. . . . .	271
17.6 A high quality ChIP-seq data set will have the highest cross-correlation between forward and reverse reads at the fragment length of about 150 base pairs (left). If fragments have not been properly pulled down in the experiment, there will be a peak in cross correlation at the read length of about 35 base pairs (middle and right). . . . .	273
17.7 A sample signal track. Here, the red signal is derived from the number of reads that mapped to the genome at each position for a ChIP-seq experiment with the target H3K36me3. The signal gives a level of enrichment of the mark. . . . .	276
17.8 Sample signal tracks for both the true experiment and the background (control). Regions are considered to have statistically significant enrichment when the true experiment signal values are well above the background signal values. . . . .	276
17.9 Example of the data and the annotation from the HMM model. The bottom section shows the raw number of reads mapped to the genome. The top section shows the annotation from the HMM model. . . . .	277
17.10 Emission probabilities for the final model with 51 states. The cell corresponding to mark $i$ and state $k$ represents the probability that mark $i$ is observed in state $k$ . . . . .	279

17.11 Transition probabilities for the final model with 51 states. The transition probability increases from green to red. Spatial relationships between neighboring chromatin states and distinct sub-groups of states are revealed by clustering the transition matrix. Notably, the matrix is sparse, so indicating that most are not possible. . . . .	280
17.12 Chromatin state definition and functional interpretation. [8] a. Chromatin mark combinations associated with each state. Each row shows the specific combination of marks associated with each chromatin state and the frequencies between 0 and 1 with which they occur in color scale. These correspond to the emission probability parameters of the HMM learned across the genome during model training. b. Genomic and functional enrichments of chromatin states, including fold enrichment in different part of the genome (e.g. transcribed regions, TSS, RefSeq 5' end or 3'end of the gene etc), in addition to fold enrichment for evolutionarily conserved elements, DNaseI hypersensitive sites, CpG islands, etc. All enrichments are based on the posterior probability assignments. c. Brief description of biological state function and interpretation (chr, chromatin; enh, enhancer). . . . .	282
17.13 ChromImpute: Training and Prediction Strategy. First, we assume there is no training data for the target mark in the target epigenome. We can generate separate regression trees for each of the epigenomes for the mark is found, and restrict features to only common marks between the target mark and tissue. Finally, we apply the tree to the target epigenome and mark and take the average of the predictions to be the expression level. . . . .	284
17.14 Epigenome Roadmap Project [8] Pink: observed and imputed. Red: imputed only. Blue: observed only. Even with experiments done in many of the different cell types, there is so much information that is missing for other cell types, and we would like to impute the data. Through epigenetic imputation, 4000 datasets were able to be generated from 1000 present datasets. . . . .	286
17.15 Comparing Imputed vs Observed Data [8] The imputed data makes very high confidence predictions. Each region noted here is 2 Mb with 1 tissue per mark. . . . .	287

---

## 17.1 Introduction

The human body contains approximately 210 different cell types. Despite having the same genetic code, cells not only develop into distinct types from this same sequence, but they also maintain the same cell type over time and across divisions. This information about the cell type and the state of the cell is called *epigenomic* information. The epigenome enables a single genome to give rise to diverse cell types serving diverse functions. The epigenome (“*epi*” meaning above in Greek) is the set of chemical modifications that influence gene expression and are transferred across cell divisions and, in some limited cases, across generations of organisms.

As shown in Figure 17.1, epigenomic information in a cell is encoded in diverse ways. For example, methylation of DNA (e.g. at CpG dinucleotides) can alter gene expression. Similarly, the positioning of nucleosomes (a unit of packing for DNA) determines which parts of DNA are accessible for transcription factors and other enzymes to bind. Almost two decades of work have revealed hundreds of post translational modifications of the tails of histone proteins found in the nucleosome. Since an extremely large number of histone modification states are possible for any given histone tail, the “histone code hypothesis” has been proposed. This hypothesis states that particular combinations of histone modifications encode information. Although a controversial hypothesis, it has guided the field of epigenomics. The core of epigenomics is understanding how chemical modifications to chromatin (e.g. DNA methylation, histone modifications or other chromatin architecture modifications) are established and how the cell “interprets” this information to establish and maintain gene expression states.

In this chapter we will explore the experimental and computational techniques used to uncover chromatin states within a cell type. We will learn how chromatin immunoprecipitation (CHIP) can be used to infer the regions of the genome bound by a protein of interest, and how a common algorithm (the Burrows-Wheeler) transform can be used to rapidly map large numbers of short sequencing reads to a reference genome. From this, we introduce how a hidden Markov model (HMM) may be used to segment the genome into regions which share similar chromatin states. We will close by showing how these comprehensive maps of chromatin

states can be compared across cell types and can be used to provide information on how cell states are established and maintained, and the impact of genetic variation on gene expression.

## 17.2 Epigenetic Information in Nucleosomes

In order to fit two meters of DNA into a 5-20  $\mu\text{m}$  diameter cell nucleus, while being arranged in such a way to give easy access to transcriptional machinery, DNA is packaged into chromatin. Nucleosomes form the unit of this packaging. A nucleosome is composed of DNA approximately 150-200 bp long wrapped around an octamer complex consisting of two copies each of histone proteins H2A, H2B, H3, and H4 (and occasionally a linker histone H1 or H5). While the structure and importance of higher-level packaging of nucleosomes is less known, the lower-level arrangement and modification of nucleosomes has been established to be very important to transcriptional regulation and the development of different cell types. Histone proteins H3 and H4 are the most highly conserved proteins in the eukaryotic domain of life.

Nucleosomes encode epigenetic information in two main ways: chromatin accessibility and histone modifications.

First, the nucleosomes' positions on the DNA determine which parts of DNA are accessible. Nucleosomes are often positioned at the promoters of inactive genes. To initiate transcription of a gene, transcription factors (TFs) and the RNA polymerase complex have to bind to its promoter. Therefore, when a gene becomes active, the nucleosomes located at its promoter are often removed from the promoter to allow RNA polymerase to initiate transcription. Hence, nucleosome positioning on the DNA is stable, yet mutable. This property of stability and mutability is a prerequisite for any form of epigenetic information because cells need to maintain the identity of a particular cell type, yet still be able to change their epigenetic state to respond to environmental circumstances.

Chromatin accessibility can also be modulated by transcribed RNA (specifically, “enhancer RNA,” or eRNA) floating around the nucleus. In particular, Mousavi et al. found in 2013 that eRNAs, which are transcribed at extragenic enhancer regions, enhance RNA pol II occupancy (which is rate-limited by chromatin accessibility) and deployment of other transcriptional machinery, leading to enhanced expression of distal target genes [10].

Second, histones contain unstructured tails protruding from the globular core domains that comprise the nucleosome octamer. These tails can undergo post-translational modification such as methylation, acetylation and phosphorylation, each of which affect gene expression. Some proteins involved in transcriptional regulation bind specifically to particular histone modifications or combinations of modifications, and recruit yet more transcription factors which enhance or repress expression of nearby genes. Thus, the “histone code hypothesis” posits that different combinations of histone modifications at specific genomic loci encode biological function via differential transcriptional regulation. In this model, histone modifications are analogous to different readers marking sections of a book with different-colored post-it notes – histone modifications allow the same genome to be interpreted (i.e., transcribed) differently at different times and in different tissues.

There are over 100 distinct histone modifications that have been found experimentally. Six of the most well-characterized histone modifications, along with the typical signature widths of their appearances in the genome and their putative associated regulatory elements, are listed in Table 17.1. Note that all of these modifications are on lysines in H3 and H4. Modifications of H3 and H4 are most well-characterized because H3 and H4 are the most highly conserved histones (making modifications of those histones more likely to have conserved regulatory function) and because good antibodies exist for all of the commonly-observed modifications of those histones.

Compression of chromosomes is another way of encoding epigenetic information. An average chromosome is 50,000 times shorter than extended DNA. Tightening this structure reduces accessibility, while loosening it makes the DNA more accessible. Acetyl groups loosen DNA, allowing the genes in that section of DNA to be more heavily expressed. Methyl groups, on the other hand, silence genes by tightening the chromosome structure. Combinations of methyl groups silence genes cooperatively.

Histone modifications are so commonly-referenced that a shorthand has been developed to identify them. This shorthand consists of the name of the histone protein, the amino acid residue on its tail that has been modified and the type of modification made to this residue. To illustrate, the fourth residue from the

N-terminus of histone H3, lysine, is often methylated at the promoters of active genes. This modification is described as H3K4me3. The first part of the shorthand corresponds to the histone protein, in this case H3; K4 corresponds to the 4th residue from the end, in this case a lysine, and me3 corresponds to the actual modification, the addition of 3 methyl groups in this case.

Histone Modification	Signature	Associated Regulatory Element
H3K4me1	(wide) focal	active promoters/enhancers
H3K4me3	(wide) focal	active promoters/enhancers
H3K9me3	wide	repressed regions
H3K27ac	focal	active promoters/enhancers
H3K27me3	wide	repressed regions
H3K36me3	wide	transcribed regions

Table 17.1: Six of the most well-characterized histone modifications along with their typical signature widths and putative associated regulatory elements. “Focal” indicates that each instance of the histone modification has a relatively narrow signature in the genome (peak width < 5kb) whereas “wide” indicates wide signatures.

One example of epigenetic modifications influencing biological function is often seen in enhancer regions of the genome. Often these enhancer regions are far away from the genes and promoters that they regulate. The enhancer is able to come into contact with a specific promoter, however, by histone modification (acetylation and methylation). This causes the DNA to fold upon itself to bring the promoter, enhancer, and recruited transcription factors into contact, activating the previously repressed promoter. This system can be very dynamic such that less than a minute after histone modification the cell will show signs of epigenetic influence. Other histone modifications, and in particular those during development, occur on a slower timescale. This is also an example of how certain types of modifications of the histones can help us to predict enhancer regions.

It is possible for more than one histone modification to be present at a given genomic locus, and histone modifications thereby can act cooperatively or competitively. It is even possible for the two copies of a given histone protein within the same nucleosome to have different modifications (though usually the histone modification “writers” will localize together, thereby creating the same modification on both copies within the nucleosome). Thus, it is necessary to simultaneously take into account all histone modifications in a genomic region in order to accurately call the chromatin state of that region. As we will see in Section , with the completion of the Roadmap Epigenome Project in 2015, a robust hidden Markov model developed with histone modifications as emissions and chromatin states as hidden states can be used to classify a section of DNA as an enhancer, promoter, transcribed, or repressed section.

### Did You Know?

The simplest organisms that have epigenetic modifications are yeasts. Yeast is a single celled organism; thus, epigenetic modifications are not responsible for cell differentiation. As organisms become more complex they tend to have more epigenetic modifications.

#### 17.2.1 Epigenetic Inheritance

The extent to which epigenetic/epigenomic features are heritable is poorly understood and is therefore the subject of much debate and ongoing investigation. In organisms that reproduce sexually, most epigenetic modifications are lost during meiosis and/or at fertilization, but some modifications are sometimes maintained. Additionally, biases exist in the ways in which paternal versus maternal epigenetic marks are removed or remodeled during this process. In particular, maternal DNA methylation is often retained at fertilization, whereas paternal DNA is almost always completely demethylated. Furthermore, for unknown reasons, some genomic elements, such as centromeric satellites, are more likely to evade epigenetic reset. In cases where epigenetic erasure does not occur completely at meiosis and fertilization, trans-generational epigenetic inheritance can occur. See [4] for further review.

Another mechanism likely to be governed by epigenetic inheritance is the phenomenon of parental imprinting. In parental imprinting, certain autosomal genes are expressed if and only if they are inherited from an individual’s mother, and other autosomal genes are expressed if and only if they are inherited

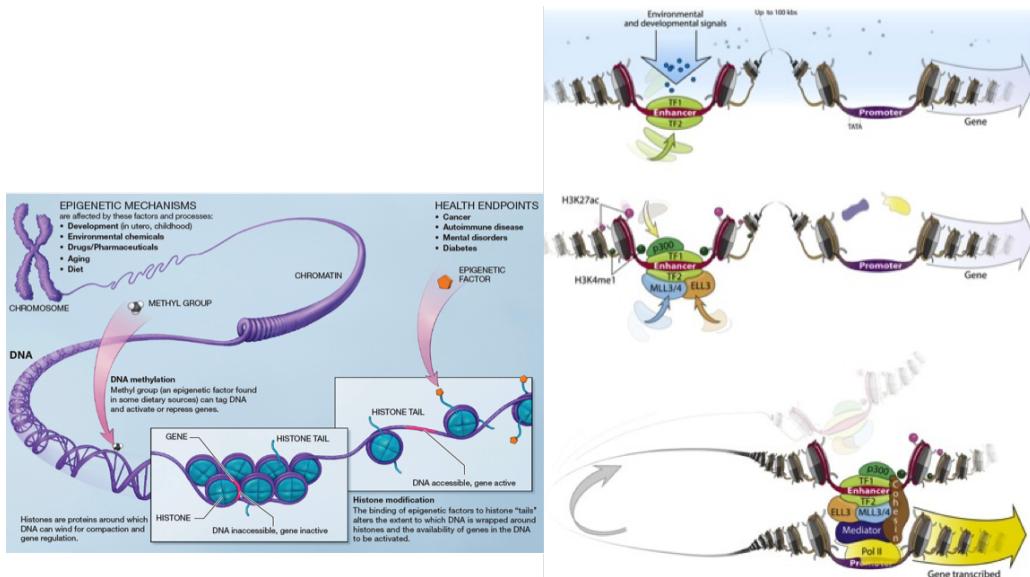


Figure 17.1: A. There is a wide diversity of modifications in the epigenome. Some regions of DNA are compactly wound around histones, making the DNA inaccessible and the genes inactive. Other regions have more accessible DNA and thus active genes. Epigenetic factors can bind to the tails of these histones to modify these properties. B. Histone modifications provide information about what types of proteins are bound to the DNA and what the function of the region is. In this example, The histone modifications allow for an enhancer region (potentially over 100 kilo bases away) to interact with the promoter region. [7]

from an individual's father. Examples are the Igf2 gene in mice (only expressed if inherited from the father) and the H19 gene in mice (only expressed if inherited from the mother). There are no changes in the DNA sequence of these genes, but extra methyl groups are observed on certain nucleotides within the inactivated copy of the gene. The mechanisms and causality of this imprinting are poorly understood.

## 17.3 Epigenomic Assays

### 17.3.1 ChIP: a method for determining where proteins bind to DNA or where histones are modified

Given the importance of epigenomic information in biology, significant efforts have been made to study signals that quantify this information. One common method for epigenomic mark measurement is called chromatin immunoprecipitation (ChIP). **ChIP technology** yields fragments of DNA whose location in the genome denote the positions of a particular histone modification or transcription factor binding. The procedures of ChIP are described as follows and are depicted in Figure 17.2:

1. Antibodies are grown to recognize a specific protein, such as those involved in histone modification. The antibodies are grown by exposing the proteins of interest to mammals, such as goats or rats, whose immune response then causes the production of the desired antibodies.
2. Cells are exposed to a cross-linking agent such as formaldehyde, which causes covalent bonds to form between DNA and its bound proteins (e.g., histones with specific modifications).
3. Genomic DNA is isolated from the cell nucleus.
4. Isolated DNA is sheared by sonication or enzymes.
5. Antibodies are added to the solution to immunoprecipitate and purify the complexes.

6. The cross-linking between the protein and DNA is reversed and the DNA fragments specific to the epigenetic marks are purified.

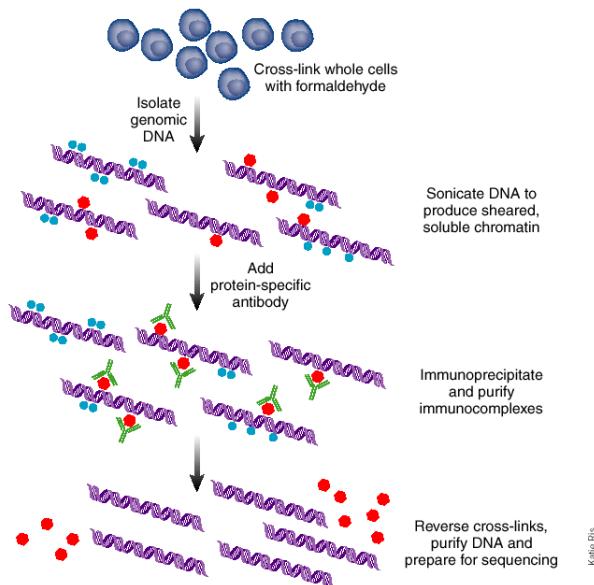


Figure 17.2: The method of chromatin immunoprecipitation [5].

After a ChIP experiment, we have short sequences of DNA that correspond to places where histones were bound to the DNA. To identify the location of these DNA fragments in the genome, one can hybridize them to known DNA segments on an array or gene chip and visualize them with fluorescent marks; this method is known as ChIP-chip. Alternatively, one can do massive parallel next-generation sequencing of these fragments; this is known as **ChIP-seq**. ChIP-seq is a newer approach that is preferred because it has a wider dynamic range of detection and avoids problems like cross-hybridization in ChIP-chip.

Each sequence tag is 30 base pairs long. These tags are mapped to unique positions in the reference genome of 3 billion bases. The number of reads depending on sequencing depth, but typically there are on the order of 10 million mapped reads for each ChIP-seq experiment.

There is a fairly standard pipeline used to infer the enrichment of the protein of interest at each site in the genome given a set of short sequencing reads from a ChIP-seq experiment. First, the DNA fragments must be mapped to the DNA (called read mapping). Next, we must determine which regions of the genome have statistically significant enrichment of the protein of interest (called peak calling). From this data set, we can build different supervised and unsupervised models to study chromatin states and their relation to biological function.

### 17.3.2 Bisulfite Sequencing: a method for determining where DNA is methylated

DNA methylation was the first epigenomic modification to be discovered and is an important transcriptional regulator. Methylation of cytosine residues in CpG dinucleotides results in “silencing,” or repression, of transcription. Bisulfite sequencing is a method by which DNA is treated with bisulfite before sequencing, allowing the precise determination of the nucleotides at which the DNA had been methylated. Bisulfite treatment converts unmethylated cytosine residues to uracil, but does not affect methylated cytosines. Thus, genomic DNA can be sequenced with or without bisulfite treatment, and the sequences can be compared. Any difference between the treated and nontreated sequences are sites at which cytosine was methylated. However, this analysis assumes complete conversion of unmethylated cytosine residues to uracil, so incomplete conversion can result in false positives [12].

Figure 17.3: (Top) In the Burrows-Wheeler forward transformation rotations are generated and sorted. The last column of the sorted list (**bolded**) consists of the transformed string. (Bottom) In the Burrows-Wheeler reverse transformation the transformed string is sorted, and two columns are generated: one consisting of the original string and the other consisting of the sorted. These effectively form two columns from the rotations in the forward transformation. This process is repeated until the complete rotations are generated.

## 17.4 Primary data processing of ChIP data

#### 17.4.1 Read mapping

The problem of **read mapping** seeks to assign a given read to the best matching location in the reference genome. Given the large number of reads and the size of human genome, one common requirement of all read mapping algorithms is that they be efficient in both space and time. Furthermore, they must allow mismatches due to sequencing errors and SNPs.

Based on previous lectures, we know various ways to perform mapping of reads: naive sequence alignment ( $O(mn)$  time for mapping a sequence of length  $m$  to a sequence of length  $n$ ) and hash-based approaches such as BLAST. Other approaches exist as well: linear time string matching ( $O(m + n)$  time) and suffix trees and suffix arrays ( $O(m)$  time). However, a problem with all these techniques is that they have a large memory requirement (often  $O(mn)$ ). Instead, state-of-the-art techniques based on the Burrows-Wheeler transformation [1] are used. These run in  $O(m)$  time and require just  $O(n)$  space where  $m$  is the sequence to map and  $n$  is the size of the reference genome.

The **Burrows-Wheeler transform** originally arose from the need to compress information. It takes a long string and rearranges it in a way that has adjacent repeating letters. This string can be compressed because, for example, instead of writing 100 A's the computer can now just indicate that there are 100 A's in a row. The Burrows-Wheeler transform also has some other special properties that we will exploit to search in sublinear time.

The Burrows-Wheeler transform creates a unique transformed string that is shorter than the original string. It also can be reversed easily to generate the original string, so no information is lost. The transformed string is in sorted order, which allows for easy searching. The details of Burrows-Wheeler transformation are described below and are illustrated in Figure 17.3.

First, we produce a transform from an original string by the following steps. In particular, we produce a transform of the reference genome.

- For a given reference genome, add a special character at the beginning and end of the string (e.g.,

“BANANA” becomes `^BANANA@`). Then generate all the rotations of this string (e.g., one such rotation would be `NANA@^BA`).

2. Sort the rotations lexicographically — i.e., in alphabetical order — with special characters sorted last.
3. Only keep the last column of the sorted list of rotations. This column contains the transformed string

Once a Burrows-Wheeler transform has been computed, it is possible to reverse the transform to compute the original string. This can be done with the procedure in Figure ???. Briefly, the reverse transformation works as follows: given the transformed string, sort the string characters in alphabetical order; this gives the first column in the transform. Combine the last column with the first to get pairs of characters from the original rotations. Sort the pairs and repeat.

By using sorting pointers rather than full strings, it is possible to generate this transform of the reference genome using a space that is linear in its size. Furthermore, even with a very large number of reads, it is only necessary to do the transform one in a forward direction. After counting the reads in the transformed space, it is then only necessary to do the reverse transform once to map the counts to genome coordinates.

In particular, from the Burrows-Wheeler transform we observe that all occurrences of the same suffix are effectively next to each other rather than scattered throughout the genome. Moreover, the  $i^{th}$  occurrence of a character in the first column corresponds to the  $i^{th}$  occurrence in the last column. Searching for substrings using the transform is also easy. Suppose we are looking for the substring “ANA” in the given string. Then the problem of search is reduced to searching for a prefix “ANA” among all possible sorted suffixes (generated by rotations). The last letter of the substring (“A”) is first searched for in the first letters of the sorted rotations. Then, the one-letter rotations of these matches are considered; the last two letters of the substring (“NA”) are searched for among the first two letters of these one-letter rotations. This process can be continued with increasing length suffixes to find the substring as a prefix of a rotation. Specifically, each read is searched for and is found as a prefix of a rotation of the reference genome; this gives the position of the read in the genome. By doing a reverse transform, it is possible to find the genomic coordinates of the mapped reads.

Note that this idea is no faster in theory than hashing, but it can be faster in practice because it uses a smaller memory footprint.

### 17.4.2 Quality control metrics

As with all experimental data, ChIP methods contain biases and their output may be of varied quality. As a result, before processing the data, it is necessary to control for these biases, to determine which reads in the data achieve a certain level of quality, and to set target thresholds on the quality of the data set as a whole. In this section we will describe these **quality control** problems and metrics associated with them.

#### QC1: Use of input DNA as control

First, the reads given by ChIP are *not* uniformly scattered in the genome. For example, accessible regions of the genome can be fragmented more easily, leading to non-uniform fragmentation. To control for this bias, we can run the ChIP experiment on the same portion of DNA without using an antibody. This yields input DNA, which can then be fragmented and mapped to give a signal track that can be thought of as a background — i.e., reads we would expect by chance. (Indeed, even in the background we do not see uniformity.) Additionally, we have a signal track for the true experiment, which comes from the chromo-immunoprecipitated DNA. Shown in Figure 17.4

#### QC2: Read-level sequencing quality score threshold

When sequencing DNA, each base pair is associated with a quality score. Thus, the reads given by ChIP-seq contain quality scores on the base pair level, where lower quality scores imply a greater probability of mismappings. The read quality then tends to be lower towards the ends of reads, where there are more mismappings. We can easily use this information in a preprocessing step by simply rejecting any reads whose average quality score falls below some threshold, which is typically 10 (e.g., only use reads where  $Q$ , the average quality score, is greater than 10). Shown in Figure 17.5.

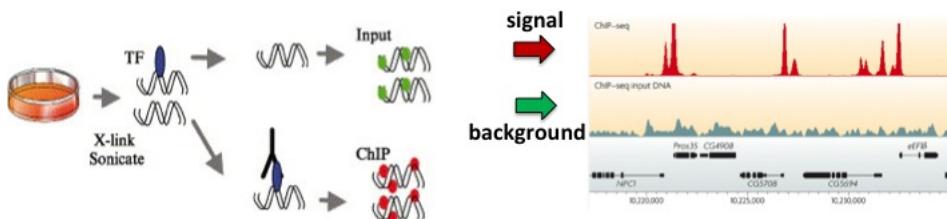


Figure 17.4: To use input DNA as a control, one can run the ChIP experiment as normal while simultaneously running the same experiment (with same DNA) without an antibody. This generates a background signal for which we can correct.

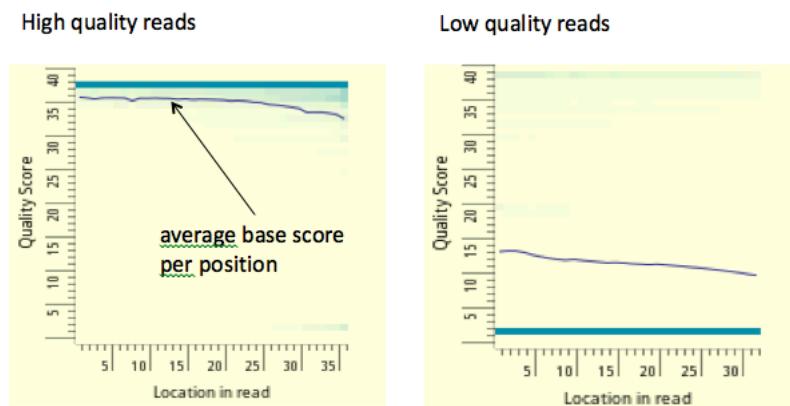


Figure 17.5: In the figure above each column is a color-coded histogram that encodes the fraction of all mapped reads that have base score Q (y-axis) at each position(x-axis). A low average per base score implies greater probability of mismappings. We typically reject reads whose average score Q is less than 10.

### QC3: Fraction of short reads mapped

Each read that passes the above quality metric may map to exactly one location in the genome, to multiple locations, or to no locations at all. When reads map to multiple locations, there are a number of approaches for handling this:

- A conservative approach: We do not assign the reads to any location because we are so uncertain.  
Con: we can lose signal
- A probabilistic approach: We fractionally assign the reads to all locations. Con: can add artifacts (unreal peaks)
- A sampling approach: We only select one location at random for a read. Chances are, across many reads, we will assign them uniformly. Con: can add artifacts (unreal peaks)
- An EM approach: We can map reads based on the density of unambiguous reads. That is, many unique reads that map to a region give a high prior probability that a read maps to that region. Note: we must make the assumption that the densities are constant within each region
- A paired-end approach: Because we sequence both ends of a DNA fragment, if we know the mapping of the read from one end, we can determine the mapping of the read at the other end even if it is ambiguous.

Either way, there will likely be reads that do not map to the genome. One quality control metric would be considering the fraction of reads that map; we may set a target of 50%, for instance. Similarly, there may be regions to which no reads map. This may be due to a lack of assembly coverage, too many reads mapping to the region, or simply that there is no activity observed in this location. To deal with the mapability biases, we can label the region as black-listed (promiscuous across many datasets) or white-listed (at least some dataset has unique reads), and treat the unmappable regions as missing data.

### QC4: Library Complexity

As a final quality control metric, we can consider the complexity of the library, or the fraction of reads that are non-redundant. In a region with signal, we might expect reads to come from all positions in that region; however, we sometimes see that only a small number of positions in a region have reads mapping to them. This may be the result of an amplification artifact in which a single read amplifies much more than it should or when the sample does not contain sufficient DNA. Consequently, we consider the non-redundant fraction of a library:

$$\text{NRF} = \frac{\text{No. of distinct unique-mapping reads}}{\text{No. of unique mapping reads}}$$

This value measures the complexity of the library. Low values indicate low complexity, which may occur, for example, when there is insufficient DNA or one DNA fragment is over-sequenced. When working with 10 million to 80 million uniquely mapped reads, we typically set a target of at least 0.8 for the NRF.

### QC5: Cross-correlation analysis

In addition to other quality control metrics, cross-correlation analysis exploits forward and reverse reads to determine the quality of a dataset. If single-end reads are employed, the a DNA binding protein will generate a peak of reads mapping to the forward strand offset a distance roughly equal to the DNA fragment length from a peak of reads mapping to the reverse strand. A similar pattern is generated from paired end reads, in which read ends fall into two groups with a given offset, one read end will map to the forward strand and the other to the reverse strand. The average fragment length can be inferred by computing the correlation between the number of reads mapping to the forward strand and number of reads mapping to the reverse strand as a function of distance between the forward and reverse reads. The correlation will peak at the mean fragment length.

The cross-correlation analysis also provides information on the quality of the ChIP-seq data set. Input DNA should not contain any real peaks, but often shows a strong read cross-correlation at a distance equal to

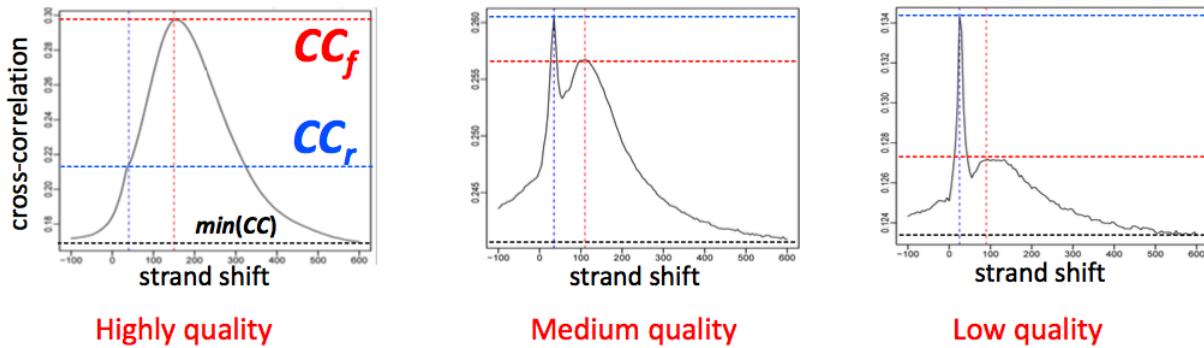


Figure 17.6: A high quality ChIP-seq data set will have the highest cross-correlation between forward and reverse reads at the fragment length of about 150 base pairs (left). If fragments have not been properly pulled down in the experiment, there will be a peak in cross correlation at the read length of about 35 base pairs (middle and right).

the read length (called a phantom peak). This occurs because some reads map uniquely in between regions that are unmappable. If a read can map uniquely at position  $x$  in between two unmappable regions on the forward strand, then a read can also map uniquely to the reverse strand at position  $x + r - 1$ , where  $r$  is the read length. Reads that map in this manner generate the strong cross- correlation at distance equal to the read length in the input DNA. If a ChIP-seq experiment was unsuccessful and did not significantly enrich for the protein of interest, then a large component of the reads will be similar to the unenriched input, which will produce a peak in the cross-correlation at read length (Figure 17.6).

Thus, the strength of the cross-correlation at read length relative to the strength at fragment length can be used to evaluate the quality of the ChIP-seq data set. Acceptable ChIP-seq libraries should have a cross-correlation at fragment length at least as high as at read-length, and the higher the ratio between the fragment-length cross-correlation and the read-length cross-correlation, the better. A good data set then will have high absolute cross-correlation at fragment length (NSC) and high fragment length cross-correlation relative to read length cross-correlation (RSC). Here, both types of cross-correlation must be taken into account. NSC (normalized strand cross-correlation) and RSC (relative strand cross-correlation) can be calculated like below, with  $CC_f$  and  $CC_r$  indicating the cross-correlation at fragment length and read length, respectively.

$$\text{NSC} = \frac{CC_f}{\min(CC)}$$

$$\text{RSC} = \frac{CC_f - \min(CC)}{CC_r - \min(CC)}$$

The peak at the fragment length should beat the phantom peak, which is reflected through an RSC value of at least 1.

### 17.4.3 Peak Calling and Selection

After reads are aligned, signal tracks as shown in Figure 17.7 can be generated. This data can be ordered into a long histogram spanning the length of the genome, which corresponds to the number of reads (or degree of fluorescence in the case of ChIP-chip) found at each position in the genome. More reads (or fluorescence) suggests a stronger presence of the epigenetic marker of interest at this particular location.

In particular, to generate these signal tracks we transform the read counts into a normalized intensity signal. First, we can use the strand cross-correlation analysis to estimate the fragment length distribution  $f$ . Since we now know  $f$ , as well as the length of each read, we can extend each read (typically just 36 bp) from the 5' to 3' direction so that its length equals the average fragment length. Then, rather than just summing the intensity of each base in the original reads, we can sum the intensity of each base in the extended reads from both strands. In other words, even though we only sequence a small read, we are able

to use information about an entire segment of which that read is a part. We can do this same operation on the control data. This yields signal tracks for both the true experiment and the control, as shown in Figure 17.8.

To process the data, we are first interested in using these signal tracks to discover regions (i.e., discrete intervals) of enrichment. This is the goal of **peak calling**. There are many programs that perform peak calling with different approaches. For example, MACS uses a local Poisson distribution as its statistical model, whereas PeakSeq uses a conditional binomial model.

One way to model the read count distribution is with a Poisson distribution. We can estimate the expected read count,  $\lambda_{\text{local}}$  from the control data. Then,

$$\Pr(\text{count} = x) = \frac{\lambda_{\text{local}}^x e^{-\lambda_{\text{local}}}}{x!}$$

Thus, the Poisson  $p$ -value for a read count  $x$  is given by  $\Pr(\text{count} \geq x)$ . We specify a threshold  $p$ -value (e.g., 0.00001) below which genomic regions are considered peaks.

We can transform this  $p$ -value into an empirical false discovery rate, or eFDR, by swapping the ChIP (true) experiment data with the input DNA (control) tracks. This would yield the locations in the genome where the background signal is higher than the ChIP signal. For each  $p$ -value, we can find from both the ChIP data and the control data. Then, for each  $p$ -value, the eFDR is simply the number of control peaks divided by the number of ChIP peaks. With this, we can then choose which peaks to call based on an eFDR threshold.

A major problem that arises is that no single universal eFDR or  $p$ -value threshold can be used. Ideal thresholds depend on a range of factors, including the ChIP, the sequencing depth, and the ubiquity of the target factor. Furthermore, small changes in the eFDR threshold can yield very large changes in the peaks that are discovered. We want to be able to combine two replicates, since replicates can show small difference in peak heights, and it's difficult to distinguish the unique peaks. Though this could be done with simple solutions, these result in other problems as well. An alternative measure is the Irreproducible Discovery Rate, or IDR, and this measure avoids these FDR-specific issues.

### Irreducible Discovery Rate (IDR)

A major drawback of using traditional statistical methods to evaluate the significance of ChIP-seq peaks is that FDR and  $p$ -value-based approaches make particular assumptions regarding the relationship between enrichment and significance. Evaluating the significance of ChIP peaks using IDR rather than a  $p$ -value or FDR is advantageous because it allows us to leverage the information present in biological replicates to call peaks without setting a threshold for significance. IDR-based approaches rely upon the idea that real signal is likely to be reproducible between replicates, whereas noise should not be reproducible. Using IDR to call significant peaks returns peaks that satisfy a given threshold for significance. To determine which peaks are significant via IDR, the peaks in each biological replicate are ranked based on their enrichment in descending order. The top  $N$  peaks in each replicate are then compared against each other, and the IDR for a given replicate is the fraction of peaks present in the top  $N$  peaks in the replicate that are not present in the other replicates (i.e., the fraction of peaks that are not reproducible between replicates). To develop more mathematical intuition, the following (entirely optional) subsection will rigorously introduce the concept of the IDR.

### Mathematical Derivation of the IDR

Since the IDR utilizes ranks, this means that the marginal distributions are uniform, and the information is mostly encoded in the joint distributions of the ranks across biological replicates. Specifically, when the marginal distributions are uniform, we can model the joint distributions through a **copula** model. Simply put, a copula is a multivariate probability distribution in which the marginal probability of each variable is uniform. **Skar's Theorem** states that there exists at least one copula function which allows us to express the joint in terms of the dependence of the marginal distributions.

$$F_k(x_1, x_2, \dots, x_k) = C_x(F_{X_1}(x_1), \dots, F_{X_k}(x_k))$$

Where  $C_x$  is the copula function and the  $F(x)$  is the cumulative distribution for a variable  $x$ . Given this information, we can set a Bernoulli distribution  $K_i \sim \text{Bern}(\pi_i)$  that denotes whether the  $i$ th peak is from the consistent set or the spurious set. We can derive  $z_1 = (z_{1,1}, z_{1,2})$  if  $K_i = 1$  or  $z_0 = (z_{0,1}, z_{0,2})$  if  $K_i = 0$  (where  $z_{0,i}$  means that it's from the spurious set in biological replicate  $i$ ). Using this, we can model the  $z_{1,1}$  and  $z_{0,1}$  models as the following:

$$\begin{pmatrix} z_{i,1} \\ z_{i,2} \end{pmatrix} | K_i = k \sim N \left( \begin{pmatrix} \mu_k \\ \mu_k \end{pmatrix}, \begin{pmatrix} \sigma_k^2 & \rho_k \sigma_k^2 \\ \rho_k \sigma_k^2 & \sigma_k^2 \end{pmatrix} \right)$$

We can utilize two different models to model whether it comes from the spurious set (denoted by 0), or the real set (1). If the real set, we have  $\mu_1 > 0$  and  $0 < \rho_1 < 1$ , whereas in the null set we have  $\mu_0 = 0$ , and  $\sigma_0^2 = 1$ . We can model a variable  $u_{i,1}$  and  $u_{i,2}$  with the following formulas:

$$\begin{aligned} u_{i,1} &= G(z_{i,1}) = \pi_1 \Phi \left( \frac{z_{i,1} - \mu_1}{\sigma_1} \right) + \pi_0 \Phi(z_{i,1}) \\ u_{i,2} &= G(z_{i,2}) = \pi_1 \Phi \left( \frac{z_{i,2} - \mu_1}{\sigma_1} \right) + \pi_0 \Phi(z_{i,2}) \end{aligned}$$

Where  $\Phi$  is the normal cumulative distribution function. Then, let the observed  $x_{i,1} = F^{-1}(u_{i,1})$  and  $x_{i,2} = F^{-1}(u_{i,2})$ , where  $F_1$  and  $F_2$  are the marginal distributions of the two coordinates. Thus, for a signal  $i$ , we have:

$$P(X_{i,1} \leq x_1, X_{i,2} \leq x_2) = \pi_0 h_0(G^{-1}(F_1(x_{i,1})), G^{-1}(F_2(x_{i,2}))) + \pi_1 h_1(G^{-1}(F_1(x_{i,1})), G^{-1}(F_2(x_{i,2})))$$

We can express  $h_0$  and  $h_1$  with the following normal distributions, similar to the  $z_1$  and  $z_2$  that were defined above:

$$\begin{aligned} h_0 &\sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \\ h_1 &\sim N \left( \begin{pmatrix} \mu_1 \\ \mu_1 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_1 \sigma_1^2 \\ \rho_1 \sigma_1^2 & \sigma_1^2 \end{pmatrix} \right) \end{aligned}$$

We can now infer the parameters  $\theta = (\mu_1, \rho_1, \sigma_1, \pi_0)$ , using a EM algorithm, where the inference is based on  $P(K_i = 1 | (x_{i,1}, x_{i,2}); \hat{\theta})$ . Thus, we can define the local irreproducible discovery rate as:

$$idr(x_{i,1}, x_{i,2}) = P(K_i = 0 | (x_{i,1}, x_{i,2}); \hat{\theta})$$

So to control the IDR at some level  $\alpha$ , we can rank  $(x_{i,1}, x_{i,2})$  by their IDR values. We can then select  $(x_{(i),1}, x_{(i),2}), i = 1 \dots l$ , where

$$I = \operatorname{argmax}_i \frac{1}{i} \sum_{j=1}^i idr_j \leq \alpha$$

IDR is analogous to a FDR control in this copula mixture model. This subsection summarizes the information provided in this lecture: [https://www.biostat.wisc.edu/~kendzior/STAT877/SK\\_2.pdf](https://www.biostat.wisc.edu/~kendzior/STAT877/SK_2.pdf). The original paper, along with an even more detailed formulation of IDR, can be found in Li et al. [11]

### Advantages and use cases of the IDR

IDR analysis can be performed with increasing  $N$ , until the desired IDR is reached (for example,  $N$  is increased until IDR=0.05, meaning that 5% of the top  $N$  peaks are not reproducible). Note that  $N$  can be different for different replicates of the same experiment, as some replicates may be more reproducible than others due to either technical or biological artifacts.

IDR is also superior to simpler approaches to use the reproducibility between experiments to define significance. One approach might be to take the union of all peaks in both replicates as significant, however; this method will accept both real peaks and the noise in each data set. Another approach is to take the intersection of peaks in both replicates, that is, only count peaks present in both data sets as significant.

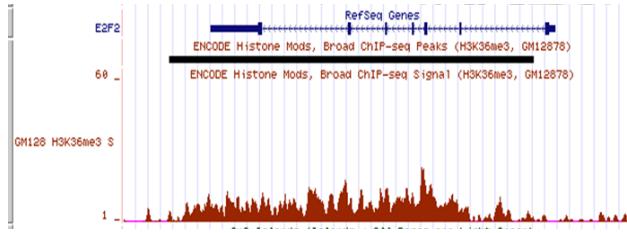


Figure 17.7: A sample signal track. Here, the red signal is derived from the number of reads that mapped to the genome at each position for a ChIP-seq experiment with the target H3K36me3. The signal gives a level of enrichment of the mark.

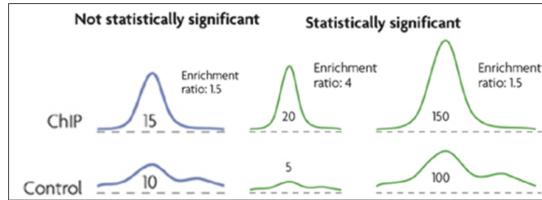


Figure 17.8: Sample signal tracks for both the true experiment and the background (control). Regions are considered to have statistically significant enrichment when the true experiment signal values are well above the background signal values.

While this method will very effectively eliminate spurious peaks, it is likely to miss many genuine peaks. A third approach would be to simply combine both datasets. However, any information gain from the independence between the two replicates is lost with this approach. IDR can be thought of as combining all these approaches, as it accepts all peaks, regardless as to whether they are reproducible, so long as the peaks have sufficient enrichment to fall within the segment of the data with an overall irreproducibility rate above a given threshold. Another advantage to IDR is that it can still be performed even if biological replicates are not available, which can often be the case for ChIP experiments performed in rare cell types. Psuedo-replicates can be generated from a single data set by randomly assigning half the reads to one pseudo-replicate and half to another pseudo-replicate.

### Interpreting Chromatin Marks

We now move onto techniques for interpreting chromatin marks. There are many ways to analyze epigenomic marks, such as aggregating chromatin signals (e.g., H3K4me3) on known feature types (e.g., promoters of genes with high or low expression levels) and performing supervised or unsupervised machine learning methods to derive epigenomic features that are predictive of different types of genomics elements such as promoters, enhancers or large intergenic non-coding RNAs. In particular, in this lecture, we examine in detail the analysis of chromatin marks as done in [8].

## 17.5 Annotating the Genome Using Chromatin Signatures

The histone code hypothesis suggests that chromatin-DNA interactions are guided by **combinatorial histone modifications**. These combinatorial modifications, when taken together, can in part determine how a region of DNA is interpreted by the cell (i.e. as a transcription factor binding domain, a splice site, an enhancer region, an actively expressed gene, a repressed gene, or a non functional region). We are interested in interpreting this “code” (i.e. determining from histone marks at a region whether the region is a transcription start site, enhancer, promoter, etc.). With an understanding of the combinatorial histone marks, we can annotate the genome into functional regions and predict novel enhancers, promoters, genes, etc. The challenge is that there are dozens of marks and they exhibit complex combinatorial effects.

Stated another way, DNA can take on a series of (hidden) states (coding, noncoding, etc). Each of these

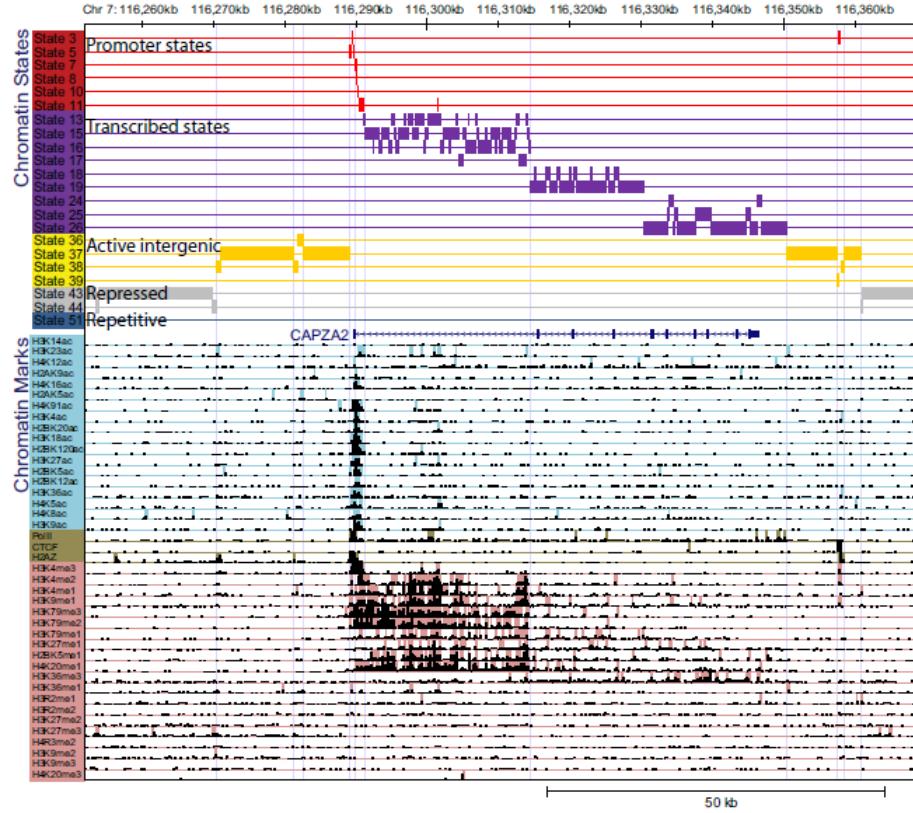


Figure 17.9: Example of the data and the annotation from the HMM model. The bottom section shows the raw number of reads mapped to the genome. The top section shows the annotation from the HMM model.

states emits a specific combination of epigenetic modifications (H3K4me3, H3K36me3, etc) that the cell recognizes. We want to be able to predict these hidden, biologically relevant states from observed epigenetic modifications.

In this section, we explore a technique for interpreting the “code” and its application to a specific dataset [8], which measured 41 chromatin marks across the human genome.

### 17.5.1 Data Binarization

Data for this analysis consisted of 41 chromatin marks including acetylations, methylations, H2AZ, CTCF and PolII in CD4 T cells. First, the genome was divided into 200 bp non-overlapping bins in which the binary absence or presence of each of the 41 chromatin marks was determined. This data was processed using **data binarization**, in which each mark in each interval is assigned a value of 0 or 1 depending on whether the enrichment of the mark’s signal in that interval exceeds a threshold. Specifically, let  $C_{ij}$  be the number of reads detected by ChIP-seq for mark  $i$ , mapping to the 200bp bin  $j$ . Let  $\lambda_i$  be the average number of reads mapping to a bin for mark  $i$ . The mark  $i$  is determined to be present in bin  $j$  if  $P(X > C_{ij})$  is less than the accepted threshold of  $10^{-4}$  where  $X$  is a Poisson random variable with mean  $\lambda_i$  and absent otherwise. The threshold is user defined, similar to a Poisson p-value. In other words, the read enrichment for a specific bin has to be significantly greater than a random process of putting reads into bins. An example for chromatin states around the CAPZA2 gene on chromosome 7 is shown in Figure 17.9. So in this way, for each mark  $i$ , we can label each bin  $j$  with a 1 if the mark is present and a 0 if it isn’t. Looking at the data as a whole, we can think of it as large binary matrix, where each row corresponds to a mark and each column corresponds to a bin (which is simply a 200bp region of the genome).

Additional data used for analysis included gene ontology data, SNP data, expression data, and others.

### 17.5.2 HMMs for Chromatin State Annotation

Our goal is to identify biologically meaningful and spatially coherent combinations of chromatin marks. Remember that we broke the genome up into 200bp blocks, so by spatially coherent we mean that if we have a genomic element that is longer than 200bps, we expect the combination of chromatin marks to be consistent on each 200bp bin in the region. We'll call these biologically meaningful and spatially coherent combinations of chromatin marks **chromatin states**. In previous lectures, we've seen HMMs applied to genome annotation for genes and CpG islands. We would like to apply the same ideas to this situation, but in this case, we don't know the hidden states *a priori* (e.g. CpG island region or not), we'd like to learn them *de novo*. This model can capture both the functional ordering of different states (e.g from promoter to transcribed regions) and the spreading of certain chromatin domains across the genomes. To summarize, we want to learn an HMM where the hidden states of the HMM are chromatin states.

As we learned previously, even if we don't know the emission probabilities and transition probabilities of an HMM, we can use the Baum-Welch training algorithm to learn the maximum likelihood values for those parameters. In our case, we have an added difficulty, we don't even know how many chromatin states exist! In the following subsections, we'll expand on how the data is modeled and how we can choose the number of states for the HMM.

#### Emission of a Vector

In HMMs from previous lectures, each state emitted either a single nucleotide or a single string of nucleotides at a time. In the HMM for this problem, each state emits a combination of epigenetic marks. Each combination can be represented as an  $n$ -dimensional vector where  $n$  is the number of chromatin marks being analyzed ( $n = 41$  for our data). The values in the emission vector can be real (Segway) or be binary, indicating presence or absence (ChromHMM). For example, assuming you have four possible epigenetic modifications: H3K4me3, H2BK5ac, Methyl-C, and Methyl-A, a sequence containing H3K4me3 and Methyl-C could be presented as the vector  $(1, 0, 1, 0)$ . One could imagine many different probability distributions on binary  $n$ -vectors and for simplicity, we assume that the marks are independent and modeled as Bernoulli random variables. So we are assuming the marks are independent given the hidden state of the HMM (note that this is not the same as assuming the marks are independent).

If there are  $n$  input marks, each state  $k$  has a vector  $(p_{k1}, \dots, p_{kn})$  of probabilities of observing marks 1 to  $n$ . Since the probability is modeled as a set of independent Bernoulli random variables, the probability of observing a set of marks given that we are in the hidden state  $k$  equals the product of the probabilities of observing individual marks. For example if  $n = 4$ , the observed marks at bin  $j$  were  $(1, 0, 1, 0)$  and we were in state  $k$ , then the likelihood of that data is  $p_{k1}(1 - p_{k2})p_{k3}(1 - p_{k4})$ .

The learned emission probabilities for the data are shown in Figure 17.10.

#### Transition Probabilities

Recall that the transition probabilities represent the frequency of transitioning from one hidden state to another hidden state. In this case, our hidden states are chromatin states. The transition matrix for our data is shown in Figure 17.11. As seen from the figure, the matrix is sparse, indicating that only a few of the possible transitions actually occur. The transition matrix reveals the spatial relationships between neighboring states. Blocks of states in the matrix reveal sub-groups of states and from these higher level blocks, we can see transitions between these meta-states.

### 17.5.3 Model Complexity: Choosing the Number of States to Model

As with most machine learning algorithms, increasing the complexity of the model (e.g. the number of hidden states) will allow it to better fit training data. However, the training data is only a limited sample of the true population. As we add more complexity, at some point we are fitting patterns in the training data that only exist due to limited sampling, so that the model will not generalize to the true population. This is called **over-fitting** training data; we should stop adding complexity to the model before it fits the noise in the training data.

Figure 17.10: Emission probabilities for the final model with 51 states. The cell corresponding to mark  $i$  and state  $k$  represents the probability that mark  $i$  is observed in state  $k$ .

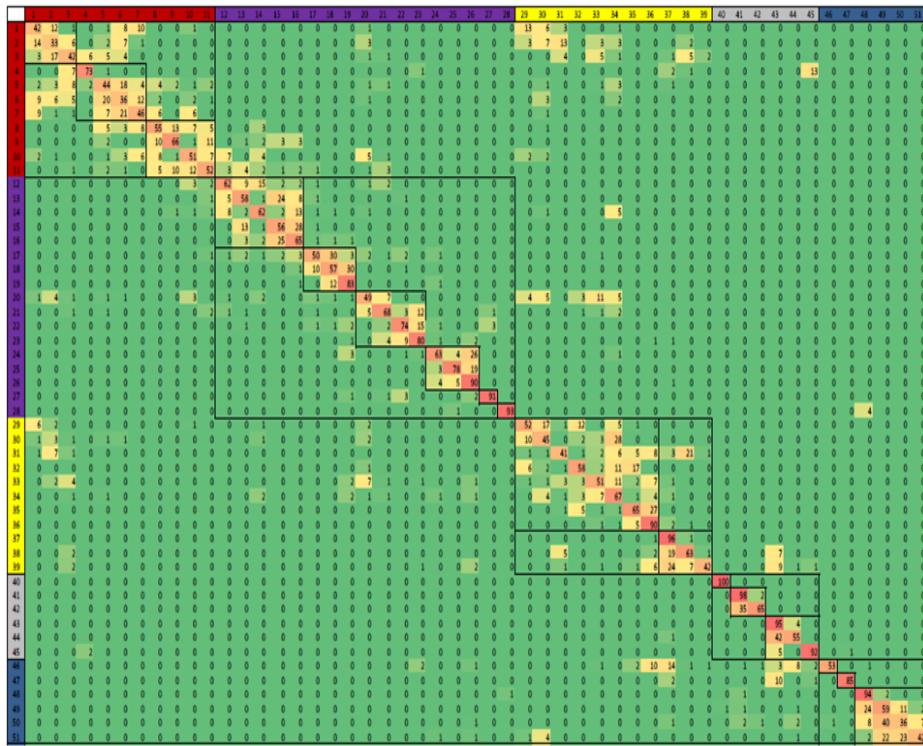


Figure 17.11: Transition probabilities for the final model with 51 states. The transition probability increases from green to red. Spatial relationships between neighboring chromatin states and distinct sub-groups of states are revealed by clustering the transition matrix. Notably, the matrix is sparse, so indicating that most are not possible.

**Bayesian Information Criterion (BIC)** is a common technique for optimizing the complexity of a model that balances increased fit to the data with complexity of the model. Using BIC, we can visualize the increasing power of the HMM as a function of the number of states. Generally, one will choose a value for  $k$  (the number of states) such that the addition of more states has relatively little benefit in terms of predictive power gain. However, there is a tradeoff between model complexity and model interpretability that BIC cannot help with. The optimal model according to BIC is likely to have more states than an ideal model because we are willing to trade some predictive power for a model with fewer states that can be interpreted biologically. The human genome is so big and the chromatin marks so complex that statistically significant differences are easy to find, yet many of these differences are not biologically significant.

To solve this problem, we start with a model with more hidden states than we believe are necessary and prune hidden states as long as all states of interest in the larger model are adequately captured. The Baum-Welch algorithm (and EM in general) is sensitive to the initial conditions, so we try several random initializations in our learning. For each number of hidden states from 2 - 80, we generate three random initializations of the parameters and train the model using Baum-Welch. The best model according to BIC had 79 states and states were then iteratively removed from this set of 79 states.

As we mentioned earlier, Baum-Welch is sensitive to the initial parameters, so when we pruned states, we used a nested initialization rather than a randomized initialized for the pruned model. Specifically, states were greedily removed from the BIC-optimal 79 state model. The state to be removed was the state that such that all states from the 237 randomly initialized models were well captured. When removing a state, the emission probabilities would be removed and any state transitioning to the removed state would have that transition probability uniformly redistributed to the remaining states. This was used as the initialization to the Baum-Welch training. The number of states for a model to analyze can then be selected by choosing the model trained from such nested initialization with the smallest number of states that sufficiently captures all states offering distinct biological interpretations. The resulting final model had 51 states.

We can also check model fit by looking at how the data violates model assumptions. Given the hidden state, the HMM assumes that each mark is independent. We can test how well the data conforms to this assumption by plotting the dependence between marks. This can reveal states that fit well and those that do not. In particular, repetitive states reveal a case where the model does not fit well. As we add more states, the model is better able to fit the data and hence fit the dependencies. By monitoring the fit on individual states that we are interested in, we can control the complexity of the model.

#### 17.5.4 Results

This multivariate HMM model resulted in a set of 51 biologically relevant ‘chromatin states’. However, there were no one-to-one relationship between each state and known classes of genomic elements (e.g. introns, exons, promoters, enhancers, etc) Instead, multiple chromatin states were often associated with one genomic element. Each chromatin state encoded specific biological relevant information about its associated genomic element. For instance, three different chromatin states were associated with transcription start site (TSS), but one was associated with TSS of highly expressed genes, while the other two were associated with TSS of medium and lowly expressed genes respectively. Such use of epigenetic markers greatly improved genome annotation, particularly when combined with evolutionary signals discussed in previous lectures. The 51 chromatin states can be divided in five large groups. The properties of these groups are described as follows and further illustrated in [17.12](#):

##### 1. Promoter-Associated States (1-11):

These chromatin states all had high enrichment for promoter regions. 40-89% of each state was within 2 kb of a RefSeq TSS. compared to 2.7% genome-wide. These states all had a high frequency of H3K4me3, significant enrichments for DNase I hypersensitive sites, CpG islands, evolutionarily conserved motifs and bound transcription factors. However, these states differed in the levels of associated marks such as H3K79me2/3, H4K20me1, acetylations etc. These states also differed in their functional enrichment based on Gene Ontology (GO). For instance, genes associated with T cell activation were enriched in state 8 while genes associated with embryonic development were enriched in state 4. Additionally, among these promoter states there were distinct positional enrichments. States 1-3 peaked both



and H3K36me3. These states have subgroups associated with 5'-proximal or 5'-distal locations. Some of these states were associated with spliced exons, transcription start sites or end sites. Of interest, state 28, which was characterized by high frequency for H3K9me3, H4K20me3, and H3K36me3, showed a high enrichment in zinc-finger genes. This specific combination of marks was previously reported as marking regions of KAP1 binding, a zinc-finger specific co-repressor.

### 3. Active Intergenic States (29-39):

These states were associated with several classes of candidate enhancer regions and insulator regions and were associated with higher frequencies for H3K4me1, H2AZ, several acetylation marks but lower frequencies of methylation marks. Moreover, the chromatin marks could be used to distinguish active from less active enhancers. These regions were usually away from promoters and were outside of transcribed genes. Interestingly, several active intergenic states showed a significant enrichment for disease SNPs, or single nucleotide polymorphism in genome-wide association study (GWAS). For instance, a SNP (rs12619285) associated with plasma eosinophil count levels in inflammatory diseases was found to be located in the chromatin state 33, which was enriched for GWAS hits. In contrast, the surrounding region of this SNP was assigned to other chromatin states with no significant GWAS association. This can shed light on the possible functional significance of disease SNPs based on its distinct chromatin states.

### 4. Large-Scale Repressed States (40-45):

These states marked large-scale repressed and heterochromatic regions, representing 64% of the genome. H3K27me3 and H3K9me3 were two most frequently detected marks in this group.

### 5. Repetitive States (46-51):

These states showed strong and distinct enrichments for specific repetitive elements. For instance, state 46 had a strong sequence signature of low-complexity repeats such as (CA) $n$ , (TG) $n$ , and (CATG) $n$ . States 48-51 showed seemingly high frequencies for many modification but also enrichment in reads from non-specific antibody control. The model was thus able to also capture artifacts resulting from lack of coverage for additional copies of repeat elements.

Since many of the chromatin states were described by multiple marks, the contribution of each mark to a state was quantified. Varying subsets of chromatin marks were tested to evaluate their potential for distinguishing between chromatin states. In general, increasing subsets of marks were found to converge to an accurate chromatin state when marks were chosen greedily.

The predictive power of chromatin states for discovery of functional elements consistently outperformed predictions based on individual marks. Such unsupervised model using epigenomic mark combination and spatial genomic information performed as well as many supervised models in genome annotation. It was shown that this HMM model based on chromatin states was able to reveal previously unannotated promoters and transcribed regions that were supported by independent experimental evidence. When chromatin marks were analyzed across the whole genome, some of the properties observed were satellite enriched states (47-51) enriched in centromere, the zinc-finger enriched state (state 28) enriched on chromosome 19 etc. Thus, such genome-wide annotation based on chromatin states can help better interpret biological data and potentially discover new classes of functional elements in the genome.

#### 17.5.5 Learning Chromatin States Jointly Across Multiple Cell Types

All of the above work was done in a single cell type (CD4+ T cells). Since epigenomic markers vary over time, across cell types, and environmental circumstances, it is important to consider the dynamics of the chromatin states across different cell types and experimental conditions. The ENCODE project [3] in the Brad Bernstein Chromatin Group has measured 9 different chromatin marks in nine human cell lines. In this case, we want to learn a single set of chromatin marks for all of the data. There are three approaches to this problem: independent learning and clustering, concatenation, and stacking. With clustering, we can first train a k-state model in each cell type independently, cluster the models (using the emission probability matrix or genome annotation), and then merge the clusters and re-apply the approach to each cell type. This could find corresponding states by matching emission vectors that are similar or by matching states

that appear at the same places in the genome. For concatenation, we could combine all of the 9 cell lines as if they were a single cell line. By concatenating the different cell lines, we ensure that a common set of state definitions are learned. We can do this here because the profiled marks were the same in each experiment. However, if we profiled different marks for different cell lines, we need to use another approach. Alternatively, we can align the 9 cell lines and treat all of the marks as a super-vector. This allows us to learn cell line specific activity states, for example there might be a state for ES-specific enhancers (in that state there would be enhancer marks in ES, but no marks in other cell types). Unfortunately, this greatly increases the dimension of the vectors emitted by the HMM, which translates to an increase in the model complexity needed to adequately fit the data.

Suppose we had multiple cell types where we profiled different marks and we wanted to concatenate them. One approach is to treat the missing marks as missing data. The EM framework allows for unspecified data points, so as long as pairwise relationships are observed between marks in some cell type, we can use EM. Otherwise, we can predict the missing chromatin marks based on the observed marks using maximum-likelihood as in the Viterbi algorithm. This is a less powerful approach if the ultimate goal is chromatin state learning because we are only looking at the most likely state instead of averaging over all possibilities as in the second approach. The more elegant approach then, is to try to impute and predict the maximum likelihood chromatin track for the missing datasets. This is explained more in depth in the next section.

In the case with 9 marks in 9 human cell lines, the cell lines were concatenated and a model with 15 states was learned [9]. Each cell type was analyzed for class enrichment. It was shown that some chromatin states, such as those encoding active promoters were highly stable across all cell types. Other states, such as those encoding strong enhancers, were highly enriched in a cell-type specific manner, suggesting their roles in tissue specific gene expression. Finally, it was shown that there was significant correlation between the epigenetic marks on enhancers and the epigenetic marks on the genes they regulate, even though these can be thousands of base pairs away. Such chromatin state model has proven useful in matching enhancers to their respective genes, a problem that has been largely unsolved in modern biology. Thus, chromatin states provide a means to study the dynamic nature of chromatin across many cell types. In particular, we can see the activity of a particular region of the genome based on the chromatin annotation. It also allows us to summarize important information contained in 2.4 billion reads in just 15 chromatin states.

A 2015 *Nature* publication by the Epigenome Roadmap Project has shown produced an unparalleled reference for human epigenomics signatures across over a hundred different tissues [2]. In their analysis, they make use of several of the concepts we have discussed in-depth in this chapter, such as a 15-state or 18-state ChromHMM model to annotate the epigenome. Training over a 111 data sets allowed for greater robustness to the HMM models discussed earlier. The Roadmap project explored many interesting directions in their paper, and interested readers are strongly encouraged to read over this publication. Interesting conclusions include that H3K4-me1 associated states are the most tissue-specific chromatin marks, and that bivalent promoters and repressed states were also the most highly variable annotations across different tissue types. For enhancers, the Roadmap project found that a significant amount of disease-related SNPs are associated with annotated enhancer regions. Active exploration of this connection is ongoing in the Computational Biology Group at MIT.

### 17.5.6 Epigenomic Imputation

When there isn't any available data for the dataset we are trying to predict, we want to complete the very large matrix of possible tissue/mark combinations instead of just treating the marks as missing. In order to fill in these missing holes in the data matrix, we can exploit the correlation between marks across the different cell types. For example, if a mark is highly correlated across different cell types, we can predict the mark for cell types that we don't have information for. Similarly, we can look at correlation between different marks in the same cell type.

ChromImpute is a machine learning model with two classes of features aimed at predicting missing tissue/mark combinations [6]. The first class looks at different marks in the same cell type. We would look at every pair of marks in all the other cell types to learn the correlation, and then apply the correlation to the cell type that we are interested in. The second class examines the same mark but across different cell types. Here, we would look at correlation of marks between different cell types, and then apply that to our target mark. Then, a regression tree can be used to predict the expression using the two types of correlation

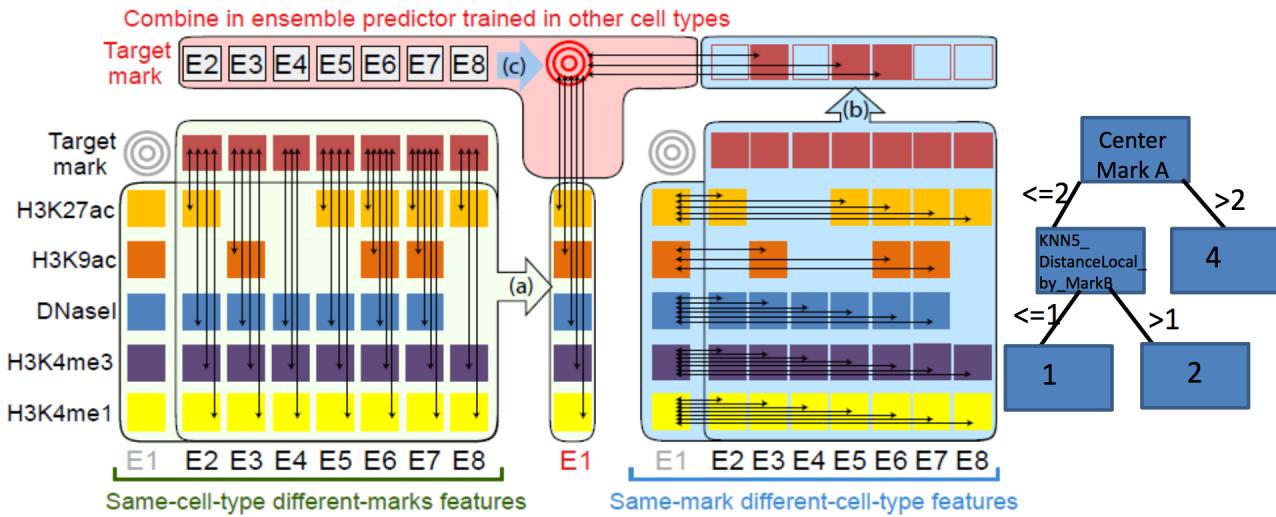


Figure 17.13: ChromImpute: Training and Prediction Strategy. First, we assume there is no training data for the target mark in the target epigenome. We can generate separate regression trees for each of the epigenomes for the mark is found, and restrict features to only common marks between the target mark and tissue. Finally, we apply the tree to the target epigenome and mark and take the average of the predictions to be the expression level.

instead of using training data. See Figure 17.13.

We can look at the applications of epigenomic imputation further with the Epigenome Roadmap Project. Looking at Figure 17.14, we see that it can be difficult to fill in the entire matrix even after many experiments. However, through imputation, we can predict marks across any epigenome and any mark with high confidence, as seen in Figure 17.15, with regions of 2Mb. Even at multiple resolutions (200kb and 10kb), the imputed data is a close match to the observed data. When the observed and imputed data disagree, is it usually the observed data that is wrong, due to the numerous datasets of marks that the predictive model uses. With the imputed data outperforming the observed data, we can use the imputed data now as a quality control metric as a measure for what the observed data should look like. Knowing this, we can flag the low-quality datasets where there are more discrepancies between the imputed and observed data because the imputed data is more likely to be correct. We can also rank the marks based on their usefulness in predicting other marks across the genome.

## 17.6 Current Research Directions

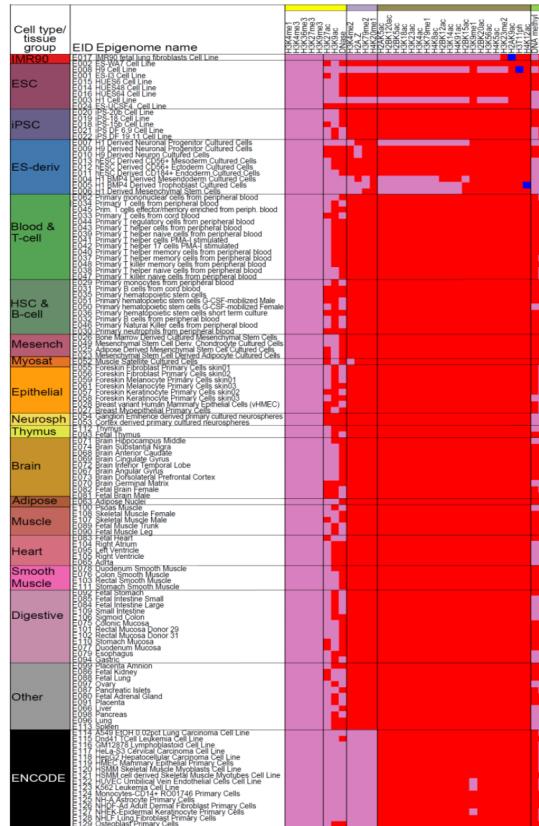
Several large-scale data production efforts such as ENCODE, modENCODE and Epigenome Roadmap projects are currently in progress and therefore there are several opportunities to computationally analyze this new data. Epigenomic data is also being used to study how behavior can alter your genome. There are studies being done that look at diet and exercise and their effects on disease susceptibility.

Another interesting area of research is the analysis of epigenetic changes in disease. Current research in the Computational Biology Group at MIT is looking at the link between chromatin states and Alzheimer's disease. A selection of papers in epigenetics-disease linkage has been provided below.

## 17.7 Further Reading

There are several interesting papers that are looking at chromatin states and epigenetics in general. Several urls are listed below to begin your exploration:

1. <http://www.nature.com/nmeth/journal/v8/n9/full/nmeth.1673.html>



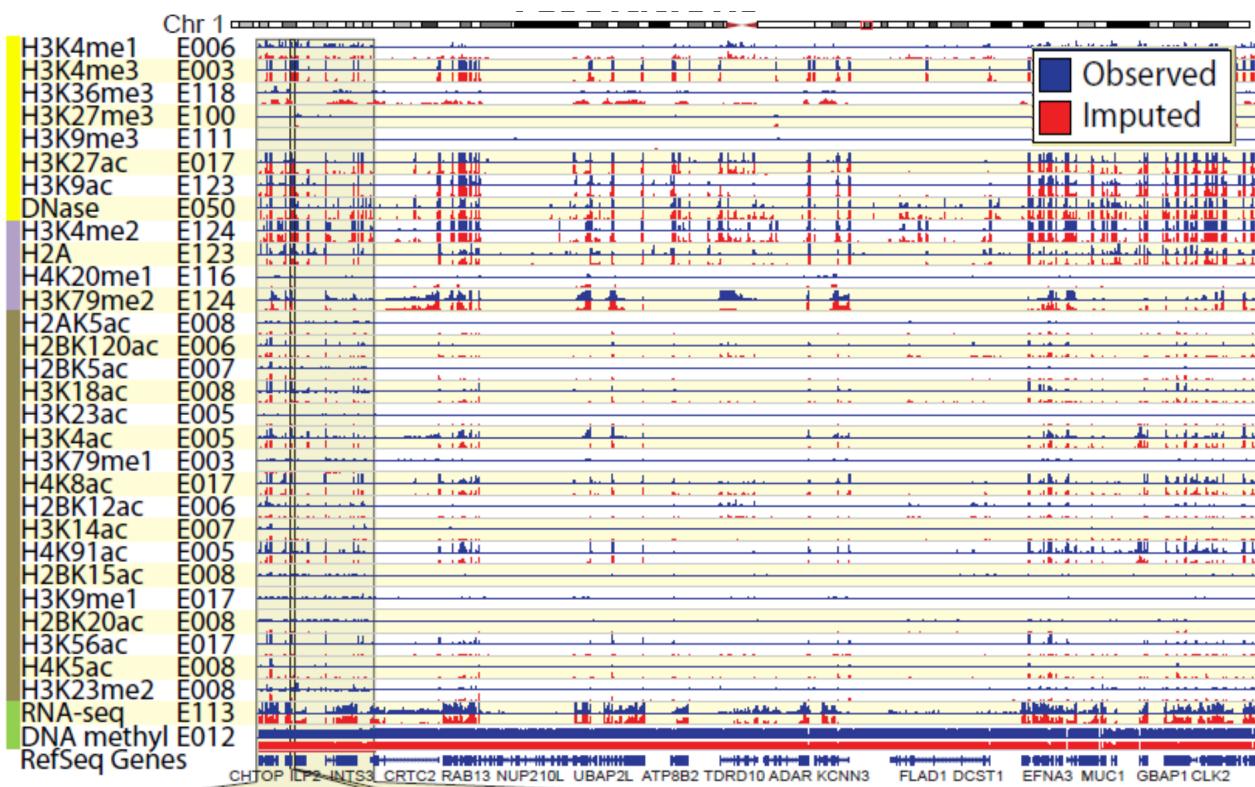


Figure 17.15: Comparing Imputed vs Observed Data [8] The imputed data makes very high confidence predictions. Each region noted here is 2 Mb with 1 tissue per mark.

2. <http://www.nature.com/nature/journal/v473/n7345/full/nature09906.html>
3. <http://www.nature.com/nbt/journal/v28/n8/abs/nbt.1662.html>
4. [http://www.nytimes.com/2012/09/09/opinion/sunday/why-fathers-really-matter.html?\\_r=1](http://www.nytimes.com/2012/09/09/opinion/sunday/why-fathers-really-matter.html?_r=1)
5. <http://www.nature.com/doifinder/10.1038/nature14248>

These are a few selected publications that deal with epigenetics and disease.

1. <http://www.nature.com/nature/journal/v429/n6990/full/nature02625.html>
2. <http://www.sciencedirect.com/science/article/pii/S2211124712003725>
3. <http://www.nature.com/nbt/journal/v28/n10/pdf/nbt.1685.pdf>

## 17.8 Tools and Techniques

ChromHMM is the HMM described in the text. It is available free for download with instructions and examples at: <http://compbio.mit.edu/ChromHMM/>.

Segway is another method for analyzing multiple tracks of functional genomics data. It uses a dynamic Bayesian network (HMMs are a particular type of dynamic Bayesian network) which enables it to analyze the entire genome at 1-bp resolution. The downside is that it is much slower than ChromHMM. It is available free for download here: <http://noble.gs.washington.edu/proj/segway/>.

## 17.9 What Have We Learned?

In this lecture, we learned how chromatin marks can be used to infer biologically relevant states. The analysis in [8] presents a sophisticated method to apply previously learned techniques such as HMMs to a complex problem. The lecture also introduced the powerful Burrows-Wheeler transform that has enabled efficient read mapping.

## Bibliography

- [1] Langmead B, Trapnell C, Pop M, and Salzberg S. Ultrafast, memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3), 2009.
- [2] Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.
- [3] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012.
- [4] Heard E and Martienssen RA. Transgenerational epigenetic inheritance: Myths and mechanisms. *Cell*, 157(1):95–109, 2014.
- [5] Mardis ER. ChIP-seq: welcome to the new frontier. *Nature Methods*, 4(8):614–614, 2007.
- [6] Jason Ernst and Manolis Kellis. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature Biotechnology*, 33(4):364–376, 2015.
- [7] Herz H-M, Hu D, and Shilatifard A. Enhancer malfunction in cancer. *Molecular Cell*, 53(6):859–866, 2014.
- [8] Ernst J and Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, 28:817–825, 2010.

- [9] Ernst J, Kheradpour P, Mikkelsen TS, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, 2011.
- [10] Mousavi K, Zare H, Dell’orso S, Grontved L, et al. eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. *Molecular Cell*, 51(5):606–17, 2013.
- [11] Qunhua Li, James B. Brown, Haiyan Huang, and Peter J. Bickel. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5(3):1752–1779, 2011.
- [12] Li Y and Tollefsbol TO. DNA methylation detection: Bisulfite genomic sequencing analysis. *Methods Molecular Biology*, 791:11–21, 2011.



---

CHAPTER  
**EIGHTEEN**

---

## REGULATORY NETWORKS: INFERENCE, ANALYSIS, APPLICATION

Guest Lecture by  
Sushmita Roy (2010) / Soheil Feizi (2012)  
Scribed by Ben Holmes (2010) / Hamza Fawzi and Sara Brockmueller (2012)

### Figures

---

18.3 The solid symbols give the in-degree distribution of genes in the regulatory network of <i>S. cerevisiae</i> (the in-degree of a gene is the number of transcription factors that bind to the promoter of this gene). The open symbols give the in-degree distribution in the comparable random network. Figure taken from [4]. . . . .	299
18.4 Scale-free vs. random Erdős-Renyi networks . . . . .	300
18.5 Network motifs in regulatory networks: Feed-forward loops involved in speeding-up response of target gene. Regulators are represented by blue circles and gene promoters are represented by red rectangles (figure taken from [4]) . . . . .	301
18.6 . . . . .	301
18.7 A network with 8 nodes. . . . .	305

---

### 18.1 Introduction

Living systems are composed of multiple layers that encode information about the system. The primary layers are:

1. Epigenome: Defined by chromatin configuration. The structure of chromatin is based on the way that histones organize DNA. DNA is divided into nucleosome and nucleosome-free regions, forming its final shape and influencing gene expression.<sup>1</sup>
2. Genome: Includes coding and non-coding DNA. Genes defined by coding DNA are used to build RNA, and Cis-regulatory elements regulate the expression of these genes.
3. Transcriptome – RNAs (ex. mRNA, miRNA, ncRNA, piRNA) are transcribed from DNA. They have regulatory functions and manufacture proteins.
4. Proteome – Composed of proteins. This includes transcription factors, signaling proteins, and metabolic enzymes.

Interactions between these components are all different, but understanding them can put particular parts of the system into the context of the whole. To discover relationships and interactions within and between layers, we can use networks.

---

<sup>1</sup>More in the epigenetics lecture.

MAKE A NE  
This intro is u  
introduce netw  
the concep of :  
Please refer to  
12B: Networks

### 18.1.1 Introducing Biological Networks

Biological networks are composed as follows:

**Regulatory Net** – set of regulatory interactions in an organism.

- Nodes are regulators (ex. transcription factors) and associated targets.
- Edges correspond to regulatory interaction, directed from the regulatory factor to its target. They are signed according to the positive or negative effect and weighted according to the strength of the reaction.

**Metabolic Net** – connects metabolic processes. There is some flexibility in the representation, but an example is a graph displaying shared metabolic products between enzymes.

- Nodes are enzymes.
- Edges correspond to regulatory reactions, and are weighted according to the strength of the reaction.

**Signaling Net** – represents paths of biological signals.

- Nodes are proteins called signaling receptors.
- Edges are transmitted and received biological signals, directed from transmitter to receiver.

**Protein Net** – displays physical interactions between proteins.

- Nodes are individual proteins.
- Edges are physical interactions between proteins.

**Co-Expression Net** – describes co-expression functions between genes. Quite general; represents functional rather than physical interaction networks, unlike the other types of nets. Powerful tool in computational analysis of biological data.

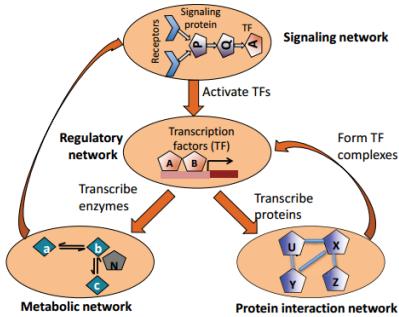
- Nodes are individual genes.
- Edges are co-expression relationships.

Today, we will focus exclusively on regulatory networks. Regulatory networks control context-specific gene expression, and thus have a great deal of control over development. They are worth studying because they are prone to malfunction and causing disease.

### 18.1.2 Interactions Between Biological Networks

Individual biological networks (that is, layers) can themselves be considered nodes in a larger network representing the entire biological system. We can, for example, have a signaling network sensing the environment governing the expression of transcription factors. In this example, the network would display that TFs govern the expression of proteins, proteins can play roles as enzymes in metabolic pathways, and so on.

The general paths of information exchange between these networks are shown in figure 18.4.



(a) Interactions between biological networks.

### 18.1.3 Studying Regulatory Networks

In general, networks are used to represent dependencies among variables. Structural dependencies can be represented by the presence of an edge between nodes - as such, unconnected nodes are conditionally independent. Probabilistically, edges can be assigned a "weight" that represents the strength or the likelihood of the interaction. Networks can also be viewed as matrices, allowing mathematical operations. These frameworks provides an effective way to represent and study biological systems.

These networks are particularly interesting to study because malfunctions can have a large effect. Many diseases are caused by rewirings of regulatory networks. They control context specific expression in development. Because of this, they can be used in systems biology to predict development, cell state, system state, and more. In addition, they encapsulate much of the evolutionary difference between organisms that are genetically similar.

To describe regulatory networks, there are several challenging questions to answer.

**Element Identification** What are the elements of a network? Elements constituting regulatory networks were identified last lecture. These include upstream motifs and their associated factors.

**Network Structure Analysis** How are the elements of a network connected? Given a network, structure analysis consists of examination and characterization of important properties. It can be done biological networks but is not restricted to them.

**Network Inference** How do regulators interact and turn on genes? This is the task of identifying gene edges and characterizing their actions.

**Network Applications** What can we do with networks once we have them? Applications include predicting function of regulating genes and predicting expression levels of regulated genes.

## 18.2 Structure Inference

### 18.2.1 Key Questions in Structure Inference

**How to choose network models?** A number of models exist for representing networks, a key problem is choosing between them based on data and predicted dynamics.

**How to choose learning methods?** Two broad methods exist for learning networks. Unsupervised methods attempt to infer relationships for unlabeled datapoints and will be described in sections to come. Supervised methods take a subset of network edges known to be regulatory, and learn a classifier to predict new ones.<sup>2</sup>

**How to incorporate data?** A variety of data sources can be used to learn and build networks including Motifs, ChIP binding assays, and expression. Data sources are always expanding; expanding availability of data is at the heart of the current revolution in analyzing biological networks.

<sup>2</sup>Supervised methods will not be addressed today.

### 18.2.2 Abstract Mathematical Representations for Networks

Think of a network as a function, a black box. Regulatory networks for example, take input expressions of regulators and spit out output expression of targets. Models differ in choosing the nature of functions and assigning meaning to nodes and edges.

**Boolean Network** This model discretizes node expression levels and interactions. Functions represented by edges are logic gates.

**Differential Equation Model** These models capture network dynamics. Expression rate changes are function of expression levels and rates of change of regulators. For these it can be very difficult to estimate parameters. Where do you find data for systems out of equilibrium?

**Probabilistic Graphical Model** These systems model networks as a joint probability distribution over random variables. Edges represent conditional dependencies. Probabilistic graphical models (PGMs) are focused on in the lecture.

#### Probabilistic Graphical Models

Probabilistic graphical models (PGMs) are trainable and able to deal with noise and thus they are good tools for working with biological data.<sup>3</sup> In PGMs, nodes can be transcription factors or genes and they are modeled by random variables. If you know the joint distribution over these random variables, you can build the network as a PGMs. Since this graph structure is a compact representation of the network, we can work with it easily and accomplish learning tasks. Examples of PGMs include:

**Bayesian Network** Directed graphical technique. Every node is either a parent or a child. Parents fully determine the state of children but their states may not be available to the experimenter. The network structure describes the full joint probability distribution of the network as a product of individual distributions for the nodes. By breaking up the network into local potentials, computational complexity is drastically reduced.

**Dynamic Bayesian Network** Directed graphical technique. Static bayesian networks do not allow cyclic dependencies but we can try to model them with bayesian networks allowing arbitrary dependencies between nodes at different time points. Thus cyclic dependencies are allowed as the network progresses through time and the network joint probability itself can be described as a joint over all times.

**Markov Random Field** Undirected graphical technique. Models potentials in terms of cliques. Allows modelling of general graphs including cyclic ones with higher order than pairwise dependencies.

**Factor Graph** Undirected graphical technique. Factor graphs introduce “factor” nodes specifying interaction potentials along edges. Factor nodes can also be introduced to model higher order potentials than pairwise.

It is easiest to learn networks for Bayesian models. Markov random fields and factor graphs require determination of a tricky partition function. To encode network structure, it is only necessary to assign random variables to TFs and genes and then model the joint probability distribution.

Bayesian networks provide compact representations of JPD

The main strength of Bayesian networks comes from the simplicity of their decomposition into parents and children. Because the networks are directed, the full joint probability distribution decomposes into a product of conditional distributions, one for each node in the network.<sup>4</sup>

---

<sup>3</sup>These are Dr. Roy's models of choice for dealing with biological nets.

<sup>4</sup>Bayesian networks are parametrized by  $\theta$  according to our specific choice of network model. With different choices of random variables, we will have different options for parametrizations,  $\theta$  and therefore different learning tasks:

**Discrete** Random variables suggest simple  $\theta$  corresponding to parameter choices for a multinomial distribution.

**Continuous** Random variables may be modelled with  $\theta$  corresponding to means and covariances of gaussians or other continuous distribution.

## Network Inference from Expression Data

Using expression data and prior knowledge, the goal of network inference is to produce a network graph. Graphs will be undirected or directed. Regulatory networks for example will often be directed while expression nets for example will be undirected.

## 18.3 Overview of the PGM Learning Task

We have to learn parameters from the data we have. Once we have a set of parameters, we have to use parametrizations to learn structure. We will focus on score based approaches to network building, defining a score to be optimized as a metric for network construction.

### 18.3.1 Parameter Learning for Bayesian Networks

**Maximum Likelihood** Chooses parameters to maximize the likelihood of the available data given the model.

In maximum likelihood, compute data likelihood as scores of each random variable given parents and note that scores can be optimized independently. Depending on the choice of a model, scores will be maximized in different manners. For gaussian distribution it is possible to simply compute parameters optimizing score. For more complicated model choices it may be necessary to do gradient descent.

**Bayesian Parameter Estimation** Treats  $\theta$  itself as a random variable and chooses the parameters maximizing the posterior probability. These methods require a fixed structure and seek to choose internal parameters maximizing score.

### Structure Learning

We can compute best guess parametrizations of structured networks. How do we find structures themselves?

Structure learning proceeds by comparing likelihood of ML parametrizations across different graph structures and in order to seek those structures realizing optimal of ML score.

A Bayesian framework can incorporate prior probabilities over graph structures if given some reason to believe a-priori that some structures are more likely than others.

To perform search in structure learning, we will inevitably have to use a greedy approach because the space of structures is too large to enumerate. Such methods will proceed by an incremental search analogous to gradient descent optimization to find ML parametrizations.

A set of graphs are considered and evaluated according to ML score. Since local optima can exist, it is good to seed graph searches from multiple starting points.

Besides being unable to capture cyclic dependencies as mentioned above, Bayesian networks have certain other limitations.

**Indirect Links** Since Bayesian networks simply look at statistical dependencies between nodes, it is easy for them to be tricked into putting edges where only indirect relations are in fact present.

**Neglected Interactions** Especially when structural scores are locally optimized, it is possible that significant biological interactions will be missed entirely. Coexpressed genes may not share proper regulators.

**Slow Speed** Bayesian methods so far discussed are too slow to work effectively whole-genome data.

### Excluding Indirect Links

**How to eliminate indirect links?** Information theoretic approaches can be used to remove extraneous links by pruning network structures to remove redundant information. Two methods are described.

**ARACNE** For every triplet of edges, a mutual information score is computed and the ARACNE algorithm excludes edges with the least information subject to certain thresholds above which minimal edges are kept.

**MRNET** Maximizes dependence between regulators and targets while minimizing the amount of redundant information shared between regulators by stripping edges corresponding to regulators with low variance.

Alternately it is possible to simply look at regulatory motifs and eliminate regulation edges not predicted by common motifs.

### 18.3.2 Learning Regulatory Programs for Modules

**How to fix omissions for coregulated genes?** By learning parameters for regulatory models instead of individual genes, it is possible to exploit the tendency of coexpressed genes to be regulated similarly. Similar to the method of using regulatory motifs to prune redundant edges, by modeling modules at once, we reduce network edge counts while increasing data volume to work with.

With extensions, it is possible to model cyclic dependencies as well. Module networks allow clustering revisitation where genes are reassigned to clusters based on how well they are predicted by a regulatory program for a module.

Modules however cannot accommodate genes sharing module membership. divide and conquer for speeding up learning

**How to speed up learning?** Dr. Roy has developed a method to break the large learning problem into smaller tasks using a divide and conquer technique for undirected graphs. By starting with clusters it is possible to infer regulatory networks for individual clusters then cross edges, reassign genes, and iterate.

### 18.3.3 Conclusions in Network Inference

Regulatory networks are important but hard to construct in general. By exploiting modularity, it is often possible to find reliable structures for graphs and subgraphs.<sup>5</sup>

Many extensions are on the horizon for regulatory networks. These include inferring causal edges from expression correlations, learning how to share genes between clusters, and others.

## 18.4 Applications of Networks

Using linear regression and regression trees, we will try to predict expression from networks. Using collective classification and relaxation labeling, we will try to assign function to unknown network elements.

We would like to use networks to:

1. predict the expression of genes from regulators.

In expression prediction, the goal is to parametrize a relationship giving gene expression levels from regulator expression levels. It can be solved in various manners including regression and is related to the problem of finding functional networks.

2. predict functions for unknown genes.

### 18.4.1 Overview of Functional Models

One model for prediction is a conditional gaussian: a simple model trained by linear regression. A more complex prediction model is a regression tree trained by nonlinear regression.

---

<sup>5</sup>Dr. Roy notes that many algorithms are available for running module network inference with various distributions. Neural net packages and Bayesian packages among others are available.

### Conditional Gaussian Models

Conditional gaussian models predict over a continuous space and are trained by a simple linear regression to maximize likelihood of data. They predict targets whose expression levels are means of gaussians over regulators.

Conditional gaussian learning takes a structured, directed net with targets and regulating transcription factors. You can estimate gaussian parameters,  $\mu$ ,  $\sigma$  from the data by finding parameters maximizing likelihood - after a derivative, the ML approach reduces to solving a linear equation.

From a functional regulatory network derived from multiple data sources<sup>6</sup>, Dr. Roy trained a gaussian model for prediction using time course expression data and tested it on a hold-out testing set. In comparisons to predictions by a model trained from a random network, found out that the network predicted substantially better than random.

The linear model used makes a strong assumption on linearity of interaction. This is probably not a very accurate assumption to make but it appears to work to some extent with the dataset tested.

### Regression Tree Models

Regression tree models allow the modeler to use a multimodal distribution incorporating nonlinear dependencies between regulator and target gene expression. The final structure of a regression tree describes expression grammar in terms of a series of choices made at regression tree nodes. Because targets can share regulatory programs, notions of recurring motifs may be incorporated. Regression trees are rich models but tricky to learn. regression trees in predicting expression

In practice, prediction works its way down a regression tree given regulator expression levels. Upon reaching the leaf nodes of the regression tree, a prediction for gene expression is made.

#### 18.4.2 Functional Prediction for Unannotated Nodes

Given a network with an incomplete set of labels, the goal of function annotation is to predict labels for unknown genes. We will use methods falling under the broad category of “guilt by association”. If we know nothing about a node but that its neighbors are involved in a function, assign that function to the unknown node.

Association can include any notion of network relatedness discussed above such as co-expression, protein-protein interactions and co-regulation. Many methods work, two will be discussed: collective classification and relaxation classification; both of which work for regulatory networks encoded as undirected graphs.

### Collective Classification

View functional prediction as a classification problem: “Given a node, what is its regulatory class?”.

In order to use the graph structure in the prediction problem, we capture properties of the neighborhood of a gene in “relational attribute”. Since all points are connected in a network, data points are no longer independently distributed - the prediction problem becomes substantially harder than a standard classification problem.

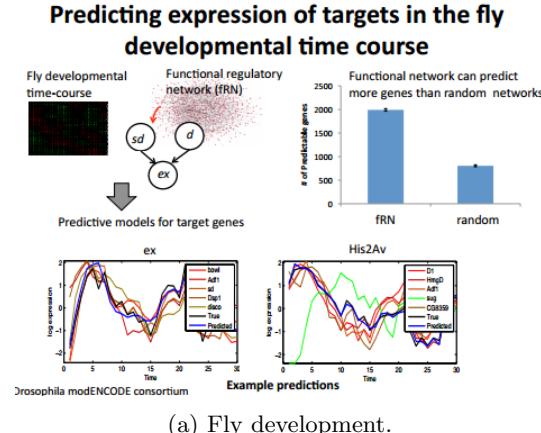
Iterative classification is a simple method with which to solve the classification problem. Starting with an initial guess for unlabeled genes it infers labels iteratively, allowing changed labels to influence node label predictions in a manner similar to gibbs sampling<sup>7</sup>

Relaxation labeling is another approach originally developed to track terrorist networks. The model uses a “suspicion score” where nodes are labeled with a suspiciousness according to the suspiciousness of its neighbors. The method is called relaxation labeling because it gradually settles on to a solution according to a learning parameter. It is another instance of iterative learning where genes are assigned probabilities of having a given function.

---

<sup>6</sup>data sources included chromatin, physical binding, expression, motif

<sup>7</sup>see the previous lecture by Manolis describing motif discovery



(a) Fly development.

## Regulatory Networks for Function Prediction

For pairs of nodes, compute a regulatory similarity – the interaction quantity – equal to the size of the intersection of their regulators divided by the size of their union. Having this interaction similarity in the form of an undirected graph over netowrk targets, can use clusters derived from a network in final functional classification.

The model is successful in predicting invaginal disk and neural system development. The blue line in Fig. 18.2a shows the score of every gene predicting its participation in neural system development.

Co-expression an co-regulation can be used side by side to augment the set of genes known to participate in neural system development.

## 18.5 Structural Properties of Networks

Much of the early work on networks was done by scientists outside of biology. Physicists looked at internet and social networks and described their properties. Biologists observed that the same properties were also present in biological networks and the field of biological networks was born. In this section we look at some of these structural properties shared by the different biological networks, as well as the networks that arise in other disciplines as well.

### 18.5.1 Degree distribution

In a network, the *degree* of a node is the number of neighbors it has, i.e., the number of nodes it is connected to by an edge. The *degree distribution* of the network gives the number of nodes having degree  $d$  for each possible value of  $d = 1, 2, 3, \dots$ . For example figure 18.3 gives the degree distribution of the *S. cerevisiae* gene regulatory network. It was observed that the degree distribution of biological networks follow a power law, i.e., the number of nodes in the network having degree  $d$  is approximately  $cd^{-\gamma}$  where  $c$  is a normalization constant and  $\gamma$  is a positive coefficient. In such networks, most nodes have a small number of connections, except for a few nodes which have very high connectivity.

This property –of power law degree distribution– was actually observed in many different networks across different disciplines (e.g., social networks, the World Wide Web, etc.) and indicates that those networks are not “random”: indeed random networks (constructed from the Erdős-Renyi model) have a degree distribution that follows a Poisson distribution where almost all nodes have approximately the same degree and nodes with higher or smaller degree are very rare [6] (see figure 18.4).

Networks that follow a power law degree distribution are known as **scale-free networks**. The few nodes in a scale-free network that have very large degree are called *hubs* and have very important interpretations. For example in gene regulatory networks, hubs represent transcription factors that regulate a very large number of genes. Scale-free networks have the property of being highly resilient to failures of “random”

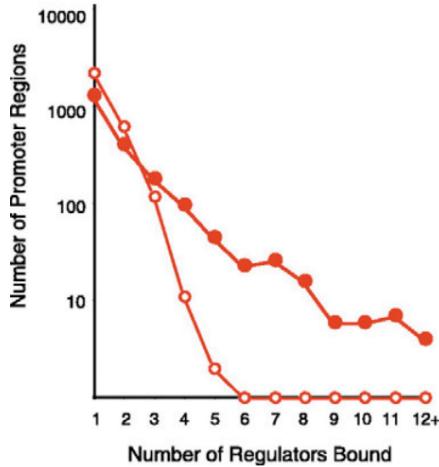


Figure 18.3: The solid symbols give the in-degree distribution of genes in the regulatory network of *S. cerevisiae* (the in-degree of a gene is the number of transcription factors that bind to the promoter of this gene). The open symbols give the in-degree distribution in the comparable random network. Figure taken from [4].

nodes, however they are very vulnerable to coordinated failures (i.e., the network fails if one of the hub nodes fails, see [1] for more information).

In a regulatory network, one can identify four levels of nodes:

1. Influential, master regulating nodes on top. These are hubs that each indirectly control many targets.
2. Bottleneck regulators. Nodes in the middle are important because they have a maximal number of direct targets.
3. Regulators at the bottom tend to have fewer targets but nonetheless they are often biologically essential!
4. Targets.

### 18.5.2 Network motifs

Network motifs are subgraphs of the network that occur significantly more than random. Some will have interesting functional properties and are presumably of biological interest.

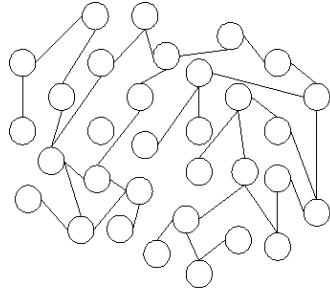
Figure 18.5 shows regulatory motifs from the yeast regulatory network. Feedback loops allow control of regulator levels and feedforward loops allow acceleration of response times among other things.

## 18.6 Network clustering

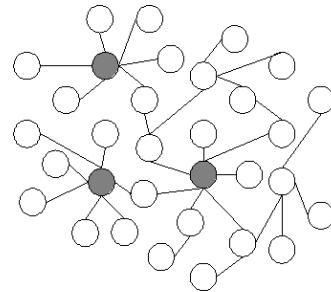
An important problem in network analysis is to be able to **cluster** or **modularize** the network in order to identify subgraphs that are densely connected (see e.g., figure 18.6a). In the context of gene interaction networks, these clusters could correspond to genes that are involved in similar functions and that are co-regulated.

There are several known algorithms to achieve this task. These algorithms are usually called *graph partitioning algorithms* since they partition the graph into separate modules. Some of the well-known algorithms include:

- Markov clustering algorithm [5]: The Markov Clustering Algorithm (MCL) works by doing a random walk in the graph and looking at the steady-state distribution of this walk. This steady-state distribution allows to cluster the graph into densely connected subgraphs.

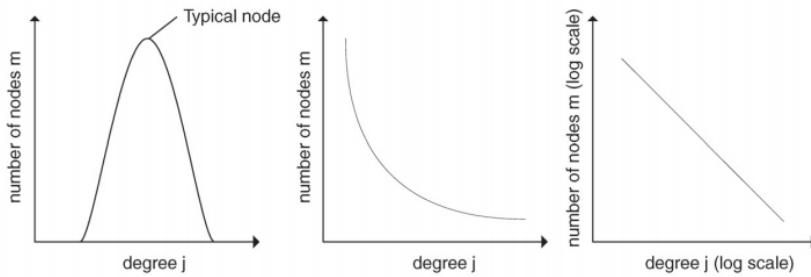


(a) Random network



(b) Scale-free network

(a) Scale-free graph vs. a random graph (figure taken from [10]).



(b) Degree distribution of scale-free network vs. random network (figure taken from [3]).

Figure 18.4: Scale-free vs. random Erdős-Renyi networks

- Girvan-Newman algorithm [2]: The Girvan-Newman algorithm uses the number of shortest paths going through a node to compute the *essentiality* of an edge which can then be used to cluster the network.
- Spectral partitioning algorithm

In this section we will look in detail at the spectral partitioning algorithm. We refer the reader to the references [2, 5] for a description of the other algorithms.

The spectral partitioning algorithm relies on a certain way of representing a network using a matrix. Before presenting the algorithm we will thus review how to represent a network using a matrix, and how to extract information about the network using matrix operations.

### 18.6.1 An algebraic view to networks

is in chapter

**Adjacency matrix** One way to represent a network is using the so-called *adjacency matrix*. The adjacency matrix of a network with  $n$  nodes is an  $n \times n$  matrix  $A$  where  $A_{i,j}$  is equal to one if there is an edge between nodes  $i$  and  $j$ , and 0 otherwise. For example, the adjacency matrix of the graph represented in figure 18.6b is given by:

$$A = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \quad (18.1)$$

If the network is weighted (i.e., if the edges of the network each have an associated weight), the definition

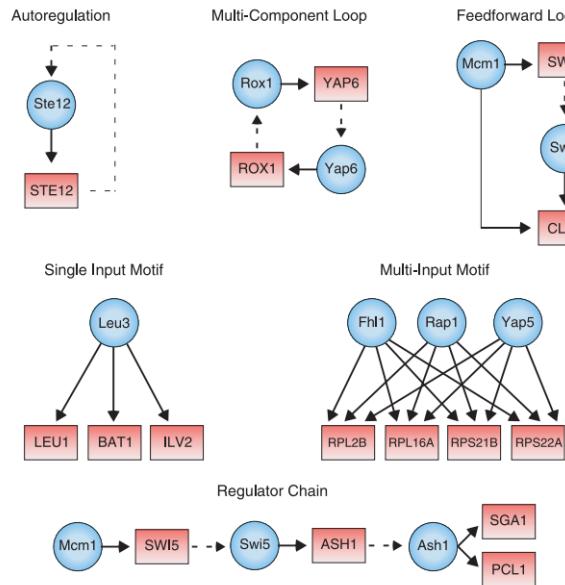


Figure 18.5: Network motifs in regulatory networks: Feed-forward loops involved in speeding-up response of target gene. Regulators are represented by blue circles and gene promoters are represented by red rectangles (figure taken from [4])

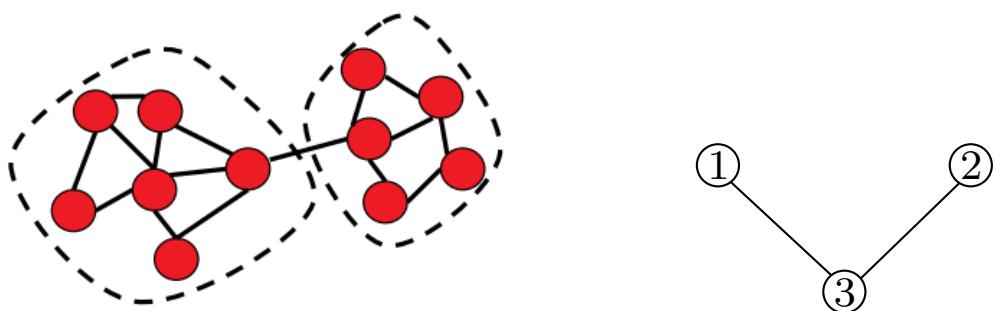


Figure 18.6

of the adjacency matrix is modified so that  $A_{i,j}$  holds the weight of the edge between  $i$  and  $j$  if the edge exists, and zero otherwise.

**Laplacian matrix** For the clustering algorithm that we will present later in this section, we will need to count the number of edges between the two different groups in a partitioning of the network. For example, in Figure 18.6a, the number of edges between the two groups is 1. The *Laplacian matrix* which we will introduce now comes in handy to represent this quantity algebraically. The Laplacian matrix  $L$  of a network on  $n$  nodes is a  $n \times n$  matrix  $L$  that is very similar to the adjacency matrix  $A$  except for sign changes and for the diagonal elements. Whereas the diagonal elements of the adjacency matrix are always equal to zero (since we do not have self-loops), the diagonal elements of the Laplacian matrix hold the *degree* of each node (where the degree of a node is defined as the number of edges incident to it). Also the off-diagonal elements of the Laplacian matrix are set to be  $-1$  in the presence of an edge, and zero otherwise. In other words, we have:

$$L_{i,j} = \begin{cases} \text{degree}(i) & \text{if } i = j \\ -1 & \text{if } i \neq j \text{ and there is an edge between } i \text{ and } j \\ 0 & \text{if } i \neq j \text{ and there is no edge between } i \text{ and } j \end{cases} \quad (18.2)$$

For example the Laplacian matrix of the graph of figure 18.6b is given by (we emphasized the diagonal elements in bold):

$$L = \begin{bmatrix} \mathbf{1} & 0 & -1 \\ 0 & \mathbf{1} & -1 \\ -1 & -1 & \mathbf{2} \end{bmatrix}$$

**Some properties of the Laplacian matrix** The Laplacian matrix of any network enjoys some nice properties that will be important later when we look at the clustering algorithm. We briefly review these here.

The Laplacian matrix  $L$  is always **symmetric**, i.e.,  $L_{i,j} = L_{j,i}$  for any  $i, j$ . An important consequence of this observation is that all the eigenvalues of  $L$  are real (i.e., they have no complex imaginary part). In fact one can even show that the eigenvalues of  $L$  are all nonnegative<sup>8</sup> The final property that we mention about  $L$  is that all the rows and columns of  $L$  sum to zero (this is easy to verify using the definition of  $L$ ). This means that the smallest eigenvalue of  $L$  is always equal to zero, and the corresponding eigenvector is  $s = (1, 1, \dots, 1)$ .

**Counting the number of edges between groups using the Laplacian matrix** Using the Laplacian matrix we can now easily count the number of edges that separate two disjoint parts of the graph using simple matrix operations. Indeed, assume that we partitioned our graph into two groups, and that we define a vector  $s$  of size  $n$  which tells us which group each node  $i$  belongs to:

$$s_i = \begin{cases} 1 & \text{if node } i \text{ is in group 1} \\ -1 & \text{if node } i \text{ is in group 2} \end{cases}$$

Then one can easily show that the total number of edges between group 1 and group 2 is given by the quantity  $\frac{1}{4}s^T L s$  where  $L$  is the Laplacian of the network.

To see why this is case, let us first compute the matrix-vector product  $Ls$ . In particular let us fix a node  $i$  say in group 1 (i.e.,  $s_i = +1$ ) and let us look at the  $i$ 'th component of the matrix-vector product  $Ls$ . By definition of the matrix-vector product we have:

$$(Ls)_i = \sum_{j=1}^n L_{i,j} s_j.$$

---

<sup>8</sup>One way of seeing this is to notice that  $L$  is diagonally dominant and the diagonal elements are strictly positive (for more details the reader can look up “diagonally dominant” and “Gershgorin circle theorem” on the Internet).

We can decompose this sum into three summands as follows:

$$(Ls)_i = \sum_{j=1}^n L_{i,j}s_j = L_{i,i}s_i + \sum_{j \text{ in group 1}} L_{i,j}s_j + \sum_{j \text{ in group 2}} L_{i,j}s_j$$

Using the definition of the Laplacian matrix we easily see that the first term corresponds to the degree of  $i$ , i.e., the number of edges incident to  $i$ ; the second term is equal to the negative of the number of edges connecting  $i$  to some other node in group 1, and the third term is equal to the number of edges connecting  $i$  to some node in group 2. Hence we have:

$$(Ls)_i = \text{degree}(i) - (\# \text{ edges from } i \text{ to group 1}) + (\# \text{ edges from } i \text{ to group 2})$$

Now since any edge from  $i$  must either go to group 1 or to group 2 we have

$$\text{degree}(i) = (\# \text{ edges from } i \text{ to group 1}) + (\# \text{ edges from } i \text{ to group 2}).$$

Thus combining the two equations above we get:

$$(Ls)_i = 2 \times (\# \text{ edges from } i \text{ to group 2}).$$

Now to get the total number of edges between group 1 and group 2, we simply sum the quantity above over all nodes  $i$  in group 1:

$$(\# \text{ edges between group 1 and group 2}) = \frac{1}{2} \sum_{i \text{ in group 1}} (Ls)_i$$

We can also look at nodes in group 2 to compute the same quantity and we have:

$$(\# \text{ edges between group 1 and group 2}) = -\frac{1}{2} \sum_{i \text{ in group 2}} (Ls)_i$$

Now averaging the two equations above we get the desired result:

$$\begin{aligned} (\# \text{ edges between group 1 and group 2}) &= \frac{1}{4} \sum_{i \text{ in group 1}} (Ls)_i - \frac{1}{4} \sum_{i \text{ in group 2}} (Ls)_i \\ &= \frac{1}{4} \sum_i s_i (Ls)_i \\ &= \frac{1}{4} s^T L s \end{aligned}$$

where  $s^T$  is the row vector obtained by transposing the column vector  $s$ .

### 18.6.2 The spectral clustering algorithm

We will now see how the linear algebra view of networks given in the previous section can be used to produce a “good” partitioning of the graph. In any good partitioning of a graph the number of edges between the two groups must be relatively small compared to the number of edges within each group. Thus one way of addressing the problem is to look for a partition so that the number of edges between the two groups is minimal. Using the tools introduced in the previous section, this problem is thus equivalent to finding a vector  $s \in \{-1, +1\}^n$  taking only values  $-1$  or  $+1$  such that  $\frac{1}{4}s^T L s$  is minimal, where  $L$  is the Laplacian matrix of the graph. In other words, we want to solve the minimization problem:

$$\underset{s \in \{-1, +1\}^n}{\text{minimize}} \quad \frac{1}{4} s^T L s$$

If  $s^*$  is the optimal solution, then the optimal partitioning is to assign node  $i$  to group 1 if  $s_i = +1$  or else to group 2.

This formulation seems to make sense but there is a small glitch unfortunately: the solution to this problem will always end up being  $s = (+1, \dots, +1)$  which corresponds to putting all the nodes of the network in group 1, and no node in group 2! The number of edges between group 1 and group 2 is then simply zero and is indeed minimal!

To obtain a meaningful partition we thus have to consider partitions of the graph that are nontrivial. Recall that the Laplacian matrix  $L$  is always symmetric, and thus it admits an eigendecomposition:

$$L = U\Sigma U^T = \sum_{i=1}^n \lambda_i u_i u_i^T$$

where  $\Sigma$  is a diagonal matrix holding the nonnegative eigenvalues  $\lambda_1, \dots, \lambda_n$  of  $L$  and  $U$  is the matrix of eigenvectors and it satisfies  $U^T = U^{-1}$ .

The cost of a partitioning  $s \in \{-1, +1\}^n$  is given by

$$\frac{1}{4} s^T L s = \frac{1}{4} s^T U \Sigma U^T s = \frac{1}{4} \sum_{i=1}^n \lambda_i \alpha_i^2$$

where  $\alpha = U^T s$  give the decomposition of  $s$  as a linear combination of the eigenvectors of  $L$ :  $s = \sum_{i=1}^n \alpha_i u_i$ .

Recall also that  $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ . Thus one way to make the quantity above as small as possible (without picking the trivial partitioning) is to concentrate all the weight on  $\lambda_2$  which is the smallest nonzero eigenvalue of  $L$ . To achieve this we simply pick  $s$  so that  $\alpha_2 = 1$  and  $\alpha_k = 0$  for all  $k \neq 2$ . In other words, this corresponds to taking  $s$  to be equal to  $u_2$  the second eigenvector of  $L$ . Since in general the eigenvector  $u_2$  is not integer-valued (i.e., the components of  $u_2$  can be different than  $-1$  or  $+1$ ), we have to convert first the vector  $u_2$  into a vector of  $+1$ 's or  $-1$ 's. A simple way of doing this is just to look at the signs of the components of  $u_2$  instead of the values themselves. Our partition is thus given by:

$$s = \text{sign}(u_2) = \begin{cases} 1 & \text{if } (u_2)_i \geq 0 \\ -1 & \text{if } (u_2)_i < 0 \end{cases}$$

To recap, the spectral clustering algorithm works as follows:

### Spectral partitioning algorithm

- Input: a network
  - Output: a partitioning of the network where each node is assigned either to group 1 or group 2 so that the number of edges between the two groups is small
1. Compute the Laplacian matrix  $L$  of the graph given by:

$$L_{i,j} = \begin{cases} \text{degree}(i) & \text{if } i = j \\ -1 & \text{if } i \neq j \text{ and there is an edge between } i \text{ and } j \\ 0 & \text{if } i \neq j \text{ and there is no edge between } i \text{ and } j \end{cases}$$

2. Compute the eigenvector  $u_2$  for the second smallest eigenvalue of  $L$ .
3. Output the following partition: Assign node  $i$  to group 1 if  $(u_2)_i \geq 0$ , otherwise assign node  $i$  to group 2.

We next give an example where we apply the spectral clustering algorithm to a network with 8 nodes.

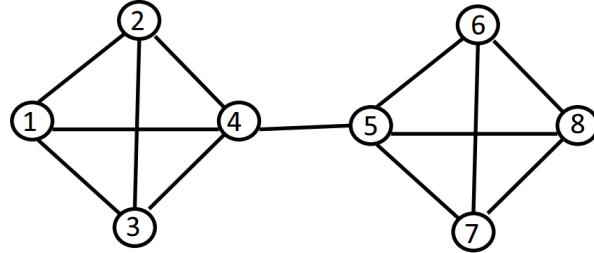


Figure 18.7: A network with 8 nodes.

**Example** We illustrate here the partitioning algorithm described above on a simple network of 8 nodes given in figure 18.7. The adjacency matrix and the Laplacian matrix of this graph are given below:

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{bmatrix} \quad L = \begin{bmatrix} 3 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 3 & -1 & -1 & 0 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 & 0 & 0 \\ -1 & -1 & -1 & 4 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 4 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & 0 & -1 & -1 & 3 & -1 \\ 0 & 0 & 0 & 0 & -1 & -1 & -1 & 3 \end{bmatrix}$$

Using the `eig` command of Matlab we can compute the eigendecomposition  $L = U\Sigma U^T$  of the Laplacian matrix and we obtain:

$$U = \begin{bmatrix} 0.3536 & \mathbf{-0.3825} & 0.2714 & -0.1628 & -0.7783 & 0.0495 & -0.0064 & -0.1426 \\ 0.3536 & \mathbf{-0.3825} & 0.5580 & -0.1628 & 0.6066 & 0.0495 & -0.0064 & -0.1426 \\ 0.3536 & \mathbf{-0.3825} & -0.4495 & 0.6251 & 0.0930 & 0.0495 & -0.3231 & -0.1426 \\ 0.3536 & \mathbf{-0.2470} & -0.3799 & -0.2995 & 0.0786 & -0.1485 & 0.3358 & 0.6626 \\ 0.3536 & \mathbf{0.2470} & -0.3799 & -0.2995 & 0.0786 & -0.1485 & 0.3358 & -0.6626 \\ 0.3536 & \mathbf{0.3825} & 0.3514 & 0.5572 & -0.0727 & -0.3466 & 0.3860 & 0.1426 \\ 0.3536 & \mathbf{0.3825} & 0.0284 & -0.2577 & -0.0059 & -0.3466 & -0.7218 & 0.1426 \\ 0.3536 & \mathbf{0.3825} & 0.0000 & 0.0000 & 0.0000 & 0.8416 & -0.0000 & 0.1426 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathbf{0.3542} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4.0000 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4.0000 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4.0000 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4.0000 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4.0000 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 5.6458 \end{bmatrix}$$

We have highlighted in bold the second smallest eigenvalue of  $L$  and the associated eigenvector. To cluster the network we look at the sign of the components of this eigenvector. We see that the first 4 components are negative, and the last 4 components are positive. We will thus cluster the nodes 1 to 4 together in the same group, and nodes 5 to 8 in another group. This looks like a good clustering and in fact this is the “natural” clustering that one considers at first sight of the graph.

### Did You Know?

The mathematical problem that we formulated as a motivation for the spectral clustering algorithm is to find a partition of the graph into two groups with a minimal number of edges between the two groups. The spectral partitioning algorithm we presented does not always give an optimal solution to this problem but it usually works well in practice.

Actually it turns out that the problem as we formulated it can be solved exactly using an efficient algorithm. The problem is sometimes called the **minimum cut** problem since we are looking to cut a minimum number of edges from the graph to make it disconnected (the edges we cut are those between group 1 and group 2). The minimum cut problem can be solved in polynomial time in general, and we refer the reader to the Wikipedia entry on *minimum cut* [9] for more information. The problem however with minimum cut partitions is that they usually lead to partitions of the graph that are not balanced (e.g., one group has only 1 node, and the remaining nodes are all in the other group). In general one would like to impose additional constraints on the clusters (e.g., lower or upper bounds on the size of clusters, etc.) to obtain more realistic clusters. With such constraints, the problem becomes harder, and we refer the reader to the Wikipedia entry on *Graph partitioning* [8] for more details.

### FAQ

**Q:** How to partition the graph into more than two groups?

**A:** In this section we only looked at the problem of partitioning the graph into two clusters. What if we want to cluster the graph into more than two clusters? There are several possible extensions of the algorithm presented here to handle  $k$  clusters instead of just two. The main idea is to look at the  $k$  eigenvectors for the  $k$  smallest nonzero eigenvalues of the Laplacian, and then to apply the  $k$ -means clustering algorithm appropriately. We refer the reader to the tutorial [7] for more information.

## Bibliography

- [1] R. Albert. Scale-free networks in cell biology. *Journal of cell science*, 118(21):4947–4957, 2005.
- [2] M. Girvan and M.E.J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.
- [3] O. Hein, M. Schwind, and W. König. Scale-free networks: The impact of fat tailed degree distribution on diffusion and communication processes. *Wirtschaftsinformatik*, 48(4):267–275, 2006.
- [4] T.I. Lee, N.J. Rinaldi, F. Robert, D.T. Odom, Z. Bar-Joseph, G.K. Gerber, N.M. Hannett, C.T. Harbison, C.M. Thompson, I. Simon, et al. Transcriptional regulatory networks in *saccharomyces cerevisiae*. *Science Signalling*, 298(5594):799, 2002.
- [5] S.M. van Dongen. *Graph clustering by flow simulation*. PhD thesis, University of Utrecht, The Netherlands, 2000.
- [6] M. Vidal, M.E. Cusick, and A.L. Barabasi. Interactome networks and human disease. *Cell*, 144(6):986–998, 2011.
- [7] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.
- [8] Wikipedia. Graph partitioning, 2012.
- [9] Wikipedia. Minimum cut, 2012.
- [10] Wikipedia. Scale-free network, 2012.

---

CHAPTER  
**NINETEEN**

---

## CHROMATIN INTERACTIONS

Silvia Canas; Vivek Dasari

### Figures

---

19.1 Chromosome Territories . . . . .	308
19.2 1) ChIP and DamID only identify regions that have come into close contact with the nuclear lamina. 2) 3C-based methods identify all DNA-DNA interactions, regardless of whether they are in the periphery of the nucleus or not . . . . .	310
19.3 ChIP Method of Measurement. Image Courtesy of Wikimedia Commons, Creative Commons Attribution-Share Alike 3.0 Unported license. . . . .	311
19.4 DamID Method of Measurement . . . . .	311
19.5 3C-based methods for identifying chromatin interactions . . . . .	312
19.6 Method for generating Hi-C data . . . . .	313
19.7 ChIA-PET protocol . . . . .	314
19.8 Lamina Associated Domains (LADs) . . . . .	315
19.9 Matrix representing Hi-C read count . . . . .	316
19.10 Image depicting sources of bias . . . . .	317
19.11 Chromosome Territories in 3D . . . . .	317
19.12 A core chromosome architecture is evident. About 70% of regions are constitutive (cLAD/ciLAD) and 30% of regions are facultative (fLAD). The LAD structure that makes up the core architecture is reminiscent of the genome isochore structure . . . . .	318
19.13 AT regions are indicators for constitutive regions . . . . .	319
19.14 LADs Through The (Single) Cell Cycle (Kind et al, Cell 2013) . . . . .	319
19.15 Steps in data-driven versus de novo modeling of chromosome structure . . . . .	320
19.16 Loops (Rao, Huntley, et al. 2014) . . . . .	320
19.17 Loops in a Hi-C map (Rao, Huntley, et al. 2014) . . . . .	321
19.18 The loop extrusion model (Sanborn, Rao et al. 2015) . . . . .	321

---

### 19.1 Introduction

In recent years, many subtle and previously disregarded mechanisms for fine genetic regulation have been discovered. Aside from direct regulation by proteins, these mechanisms include the involvement of non-protein coding regions of the genome, epigenomic factors such histone modifications, and diverse RNA switches. The spatial organization of chromatin inside the nucleus, chromatin modifier complexes and its functional consequences have also become an area of interest. In this chapter, we will delve into the study of 3D chromatin structures, starting with the state of the art in the field, relevant terminology, and current

methods. Specially we will focus on the study of DNA regions located by peripheral regions of the nucleus (thus in close contact with the nuclear lamina). Finally, we will discuss the computational methods involved in studying nuclear genome organization.

### 19.1.1 What's already known

DNA is locally compacted in nucleosomes, by wrapping around histone octamers. Each nucleosome comprises about 147 bps packed in 1.67 left-handed superhelical turns. DNA is globally compacted as chromosomes (during cell division and mitosis). Chromosomes have been dyed with different colors, and it has been shown that some chromosomes have radial preferences within the cell nucleus, even when the cell is not actively undergoing division and the chromosomes are not condensed. That is, some chromosomes prefer to stay near the center of the nucleus while others tend towards the periphery. These are known as chromosome territories (CT). The territories of homologous chromosomes usually do not lie near one another. It is also known that there is an overarching nuclear 'architecture' that is observable, conserved even between different cell types.

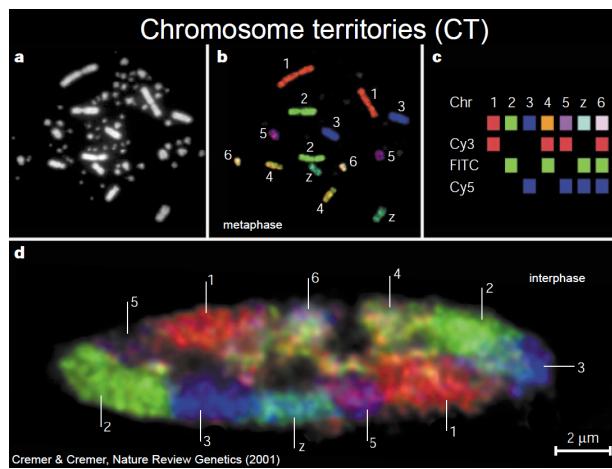


Figure 19.1: Chromosome Territories

### 19.1.2 What we don't know

While local organization (nucleosome packaging) and global organization (chromosome condensation) of DNA are somehow understood, the intermediate structures of DNA are not yet well characterized - many speculated states have only been observed *in vitro*. The positioning of genomic regions in the nucleus on a sub-chromosomal level, for example radial preference or the specific 3D conformation of a certain chromosomal region containing several genes, is also largely unknown.

While it is known that chromosomes retain some general architecture during the entire cell cycle, the mechanisms behind how that location is maintained and how different chromosomes continue to interact over the course of the entire cell cycle is unknown.

So while we understand certain aspects of the function of chromosomes, we do not have a complete mechanistic understanding of their organization and how this changes during events such as the cell cycle or differentiation.

### 19.1.3 Why do we study it?

In general, we are interested in understanding the functional characteristics of genomic regions and the molecular mechanisms encoded within, which might have implications in human diseases. Particularly, it has been shown that genes that are encoded in spatially neighboring regions are likely to be co-regulated. Also, the DNA packed inside the nucleus is the equivalent of wrapping 20 km of 20 μm thick thread in

something the size of a tennis ball, which would reach from Kendall Square to Harvard and back over 6 and a half times! Due to the potential for biomedical applications and the interesting problem it poses to tightly pack chromatin into the nucleus, chromatin organization is an exciting open area of study in computational biology.

## 19.2 Relevant terminology

### 19.2.1 Nuclear lamina

The nuclear lamina is a dense network of proteins and filaments that lies on the inner surface of the inner nuclear membrane. Its functions include: providing mechanical support, and regulating cellular events like DNA replication and cell division. Further, it plays a role in maintenance of the nuclear stability, interact with nuclear pore complexes, and organization of the chromatin by directly interacting and binding with it. The protein meshwork is predominantly made up of lamin proteins.

### 19.2.2 Lamina Associated Domains(LADs)

Lamina associated domains(LADs) are the portions of the chromatin that interact with the nuclear lamina. Mapping the interactions of the chromatin and the nuclear lamina provides insight towards mapping chromosome folding. While not much is known about LADs, it is known that these regions are closely related to both high gene expression and low gene density, an interesting combination. Additionally LADs are associated with CTCFs, promoters, and CPG islands along its borders.

### 19.2.3 Histones

Histones are highly alkaline proteins found eukaryotes that comprise the core of nucleosomes, packaging and ordering the nuclear DNA. An octamer is formed by 2 copies of the core histones H2A, H2B, H3, and H4 forms the nucleosome, which acts as a spool for DNA to wind around.

### 19.2.4 Chromatin

Chromatin is a complex form by DNA, proteins and RNA that generates the global architecture of DNA in eukaryote nuclei. Its main functions involve DNA packaging, reinforcing the DNA macromolecule to allow mitosis, preventing DNA damage, and regulating gene expression and DNA replication. Most of the mechanisms underlying the formation and regulation of the structure of chromatin are largely unknown; however, during cell division, chromatin organizes by way of chromosomes.

### 19.2.5 Chromosome territories (CT)

Chromosomes are not randomly distributed throughout the nucleus. Chromosome territories are regions of the nucleus particular chromosomes occupy preferentially.

### 19.2.6 Gross folding principles

Scribe: Expand  
talk about folding  
free energy?

## 19.3 Molecular Methods for Studying Nuclear Genome Organization

There are two main types of methods for investigating the three-dimensional structure of chromatin in the nucleus.

- The first set of methods, ChIP and DamID, are methods that measure DNA-'landmark' interactions. That is, they measure interactions of genome loci with relatively fixed nuclear landmarks, and only regions of the genome that come into contact with the nuclear lamina will be identified.

- The second set of methods, the 3C-based methods, are those that measure DNA-DNA interactions. Any two regions of DNA that interact may be identified, regardless of whether they are near the interior or periphery of the nucleus.

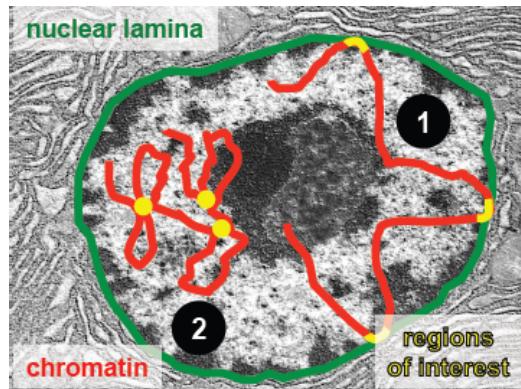


Figure 19.2: 1) ChIP and DamID only identify regions that have come into close contact with the nuclear lamina. 2) 3C-based methods identify all DNA-DNA interactions, regardless of whether they are in the periphery of the nucleus or not

### 19.3.1 Methods for measuring DNA-Nuclear Lamina interactions

The following methods, ChIP and DamID, both examine regions of DNA that specifically come in contact with the nuclear lamina.

#### ChIP: Chromatin Immuno Precipitation

ChIP is a method for detecting regions of DNA that are bound to proteins of interest. Proteins bound with DNA are cross-linked in place with formaldehyde. (Formaldehyde induces a nucleophilic attack of the amine groups of lysine and subsequent covalent bonds, allowing for cross-linking). The protein-DNA complexes are cut into smaller fragments by sonication or nuclease digestion. Then, these cross-linked fragments are pulled-down (immuno-precipitation) using affinity chromatography, mainly using specific antibodies that target the protein of interest. The recovered complexes are then dissociate, cross-links are broken, and the DNA that was bound to the proteins is fragmented and analyzed. DNA fragments can be then analyzed using sequencing (ChIP-Seq) or microarrays (ChIP-Chip). However, a big challenge associated with the various ChIP techniques is that it can be difficult to get a high-affinity antibody. To study the 3D structure of DNA within the nucleus, ChIP-Seq can be used with antibodies that target lamina proteins.

#### DamID: DNA adenine methyltransferase IDentification

DamID is used to map the binding sites of Chromatin-binding proteins. In the DamID method, DNA adenine methyltransferase (*Dam*) from *E. coli* is fused to the LaminB1 protein (the *Dam* enzyme hangs off the end of the protein and is thus in the vicinity for interactions). In *E. coli*, the *Dam* enzyme methylates the adenine in the sequence GATC; bacterial genomes contain proteins with functions like Dam to protect their own DNA from digestion by restriction enzymes, or as part of their DNA repair systems. As this process doesn't naturally occur in eukaryotes, the methylated adenines in a region can thus be attributed to an interaction with the protein fused with Dam, thereby implying that that particular region came into close contact with the nuclear lamina. As a control, unfused *Dam* can be expressed at low levels. This results in a sparse distribution of methylated adenine for which the precise position of the methylated adenines can be used to infer variation in DNA accessibility. The methylated adenine are determined using disulphide PCR assays or other PCR technique sensitive to methylations in the template DNA. In one of those assays, the genome can be digested by DpnI, which only cuts methylated GATC sequences. Adapter sequences are then

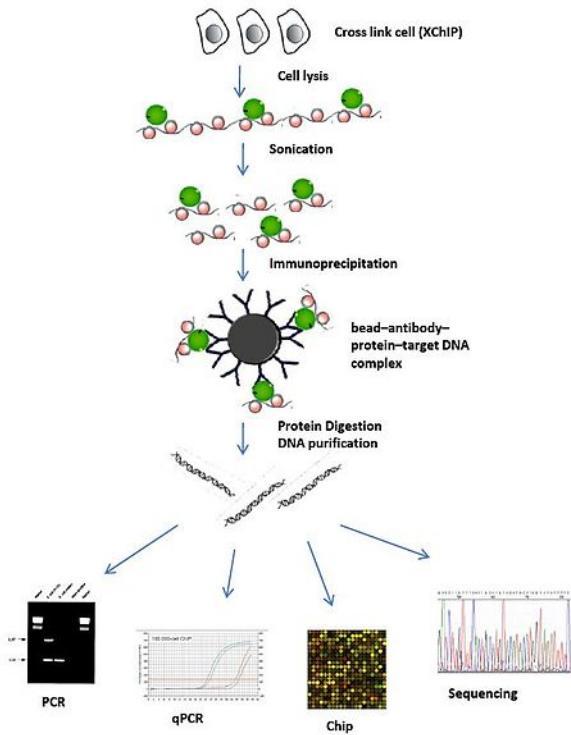


Figure 19.3: ChIP Method of Measurement. Image Courtesy of Wikimedia Commons, Creative Commons Attribution-Share Alike 3.0 Unported license.

ligated to the ends of these digested pieces, and PCR is run using primers matching the adapters. Only the regions occurring between proximal GATC positions are amplified. The final measurement is the log of the ratio between test and control lamina association: positive values are preferentially lamina-associated, and are thus identified as LADs. One advantage of using DamID over ChIP is that DamID does not require a specific antibody which may be difficult to find. However, a disadvantage of using DamID is that the fusion protein must be made and expressed.

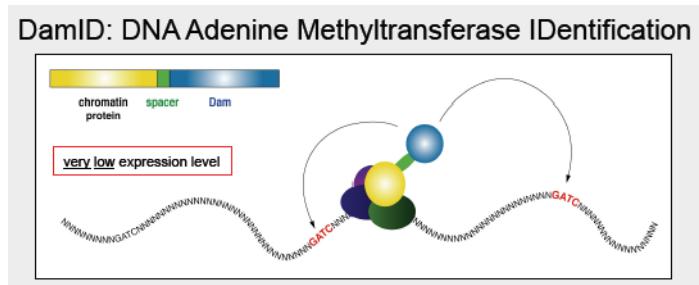


Figure 19.4: DamID Method of Measurement

## FAQ

**Q:** How close does DNA have to come to DamID to be methylated?

**A:** It doesn't have to bind directly to the lamina, but it does have to come pretty close. DamID has a range of about 1.5kb.

### 19.3.2 Measuring DNA-DNA contacts

All of the following methods are based on Chromosome Conformation Capturing (3C) with certain modifications. The methods follow a similar procedure but differ in their scope.

The techniques are used to analyze the organization of chromatin in the cell by identifying regions between loci that interact in 3D, despite being separated linearly. All methods the same general procedure: the genomic interactions between loci are first "frozen" in place via cross-linking. The genomes are cut into fragments. Then ligation is performed. The digestion then ligation of fragments is meant to create a means of measuring interactions/proximity between regions, because fragments near each other are more likely to be ligated together.

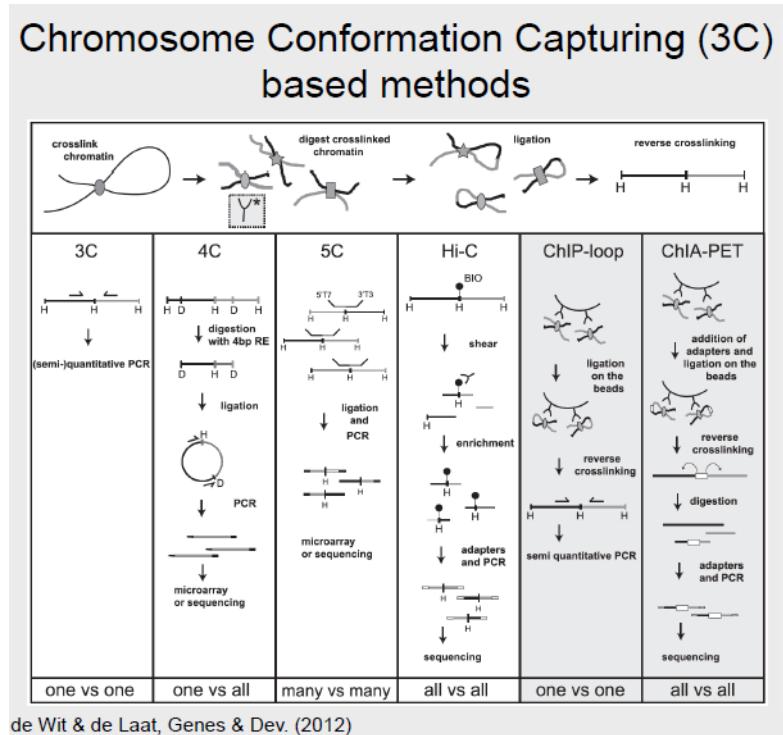


Figure 19.5: 3C-based methods for identifying chromatin interactions

#### 3C: one-vs-one

Chromosome Conformation Capturing (3C) is a method that detects which genomic loci are in close vicinity to other loci within the nucleus. Similar to the ChIP method, a cross-linking agent is used to freeze proteins bound to DNA in place and forming protein-DNA complexes. The DNA can be then be digested by a restriction enzyme after allowing the bound protein to disassociate. Typically an enzyme with a 6 bps long

recognition site that leaves sticky ends, like *HindIII*, is used. The generated fragments are then induced to self-ligate, using a very low concentration of DNA to prevent the ligation of the fragment with another random fragment. The result is a pool of linear DNA fragments, known as the 3C library, that may be analyzed via PCR by designing primers specifically for the interaction of interest. 3C can be described as a 'one vs one' method, because the primers used are specifically targeted to amplify the product of the interaction between 2 regions of interest.

### Circularized Chromatin Conformation Capture (4C): one-vs-all

4C methods can be described as a 'one vs all' because for a single region of interest, we can examine all its interactions with all other regions in the genome. 4C works similarly to 3C with the main difference being the restriction enzyme used. In 4C, a common cutter is employed to generate more and smaller fragments. These fragments are then again ligated. Some smaller fragments may be excluded, but the result is a circularized fragment of DNA. Primers can be designed to amplify the 'unknown' fragment of DNA so that all interactions with the region of interest are identified.

### Carbon-Copy Chromosome Conformation Capture (5C): many-vs-many

5C is a 'many vs many' method and allows the identification of interactions between many regions of interest and many other regions, also of interest, to be analyzed at once. 5C works similarly to 3C. However, after obtaining the 3C library, multiplex ligation-mediated amplification (LMA) is performed. LMA is a method in which multiple targets are amplified. This is done by ligating multiple primers to the different fragments. The resulting 5C library may be analyzed on a microarray or high-throughput sequencing. One drawback of 5C is that it is low coverage.

### Hi-C: all-vs-all

Hi-C can be described as an 'all vs all' method because it identifies all chromatin interactions. Hi-C works by labeling all DNA fragments with biotin before ligation, which marks all the ligation junctions. Magnetic beads are then used to purify the biotin-marked junctions. This Hi-C library may then be fed into next generation sequencing.

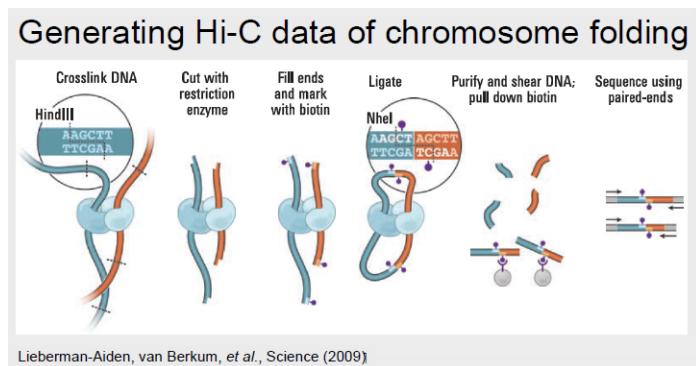


Figure 19.6: Method for generating Hi-C data

### ChIP-loop

ChIP-loop can be described as a 'one vs one' method, because similar to 3C, only an interaction between two regions of interest may be identified. ChIP-loop is a hybrid between ChIP and the 3C methods. DNA-protein complexes are first cross-linked and digested. Then, as in ChIP, the protein of interest and the DNA bound to it are pulled down using an antibody. The protocol then proceeds as in 3C: the free ends of the fragments are ligated, the cross-linking is reversed, and sequencing can proceed using primers designed specifically for a 'one vs one' interaction.

## ChIA-PET

Chromatin Interception Analysis by Paired-End Tag Sequencing, or ChIA-PET, combines the ChIP and 3C methodologies to determine long-range chromatin interactions genome-wide. It can be described as a 'all vs all' method, because although a single protein of interest must be identified, any interactions will be identified. In ChIA-PET, DNA-protein complexes are cross-linked, as in previously discussed methods. However, sonication is then used to break up chromatin, and to reduce non-specific interactions. As in the ChIP protocol, an antibody is used to pull down regions of DNA bound to a protein of interest. Two different oligonucleotide linkers are then ligated to the free ends of the DNA. These linkers both have *MmeI* cut sites. The linkers are then ligated together so that the free ends are connected, after which the fragments are digested with *MmeI*. *MmeI* cuts 20 nt downstream of its recognition sequence, so the result of the digestion is the linker bordered by the sequence of interest on either side. This is a 'tag-linker-tag' structure, and the fragments are known as PETs. The PETs may be sequenced and mapped back to the genome to determine regions of interacting DNA.

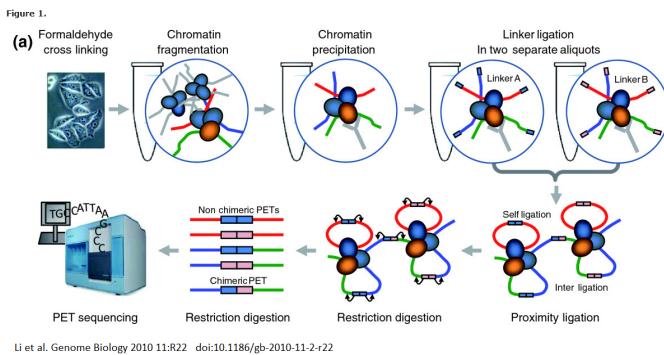


Figure 19.7: ChIA-PET protocol

## 19.4 Mapping Genome-Nuclear Lamina Interactions (LADs)

In this section, we will present how the DamID and Hi-C methods were used to map lamina-associated domains in the genome.

### 19.4.1 Interpreting DamID Data

The DamID method (described in section 3) was used to identify regions of DNA that interacted with the lamin protein at the nuclear lamina.

#### **Did You Know?**

DamID experiments typically run for 24 hours and the methylation is irreversible. The results are also the average over millions of cells. Therefore, DamID is not suitable for exact time related positioning of the genome, though single cell studies may soon make address this issue!

The results of the DamID experiment were plotted as  $\log_2 \frac{\text{Dam fusionprotein}}{\text{Dam only}}$ , as done in the figure below (black peaks). For the LaminB1 fusion experiment, positive regions (underlined in yellow in the figure below) indicate regions which preferentially associate with the nuclear lamina. These positive regions are defined as Lamina Associated Domains, or LADs. Approximately 1300 LADs were discovered in human fibroblasts. They were

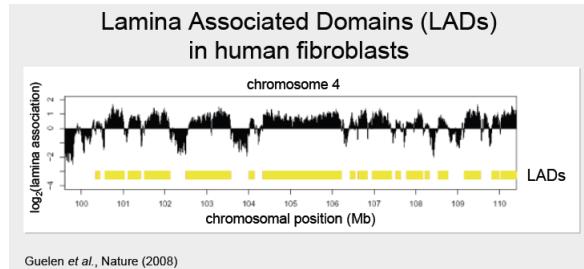


Figure 19.8: Lamina Associated Domains (LADs)

**FAQ**

surprisingly large, ranging from about 0.1Mb - 10Mb, with a median size of 5Mb.

**Q:** In this representation is a value of 0 significant?

**A:** No, there is definitely a point where we make a good estimate of where the 0 interior vs exterior of the nucleus)

After LADs have been identified, we can align their boundaries to discover various interesting features such as known gene densities or gene expression levels to the data to build our LAD model.

**FAQ**

**Q:** How does organization relate to polyclonal expression?

Experiments

**A:** Certainly something going on; however, polyclonal repression works on a smaller scale than LAD. It occurs outside of LADs, as an additional repression mechanism

have shown that LADs are characterized by low gene density and gene expression levels. It was noticed that the LAD boundaries are very sharply defined. By aligning the start positions of many LADs, it was discovered that the borders are particularly marked by CpG islands, outward pointing promoters, and CTCF binding sites.

**FAQ**

**Q:** Why CTCF binding sites? What's so important about them?

**A:** That's the question! Perhaps they help maintain LADs. They could perhaps prevent the LADs from 'spreading'. Information from experiments like Hi-C under knock-out conditions can give us more information about the role these elements play in chromatin organization.

### 19.4.2 Interpreting Hi-C Data

Hi-C data was collected, and the reads were mapped back to the genome. The read counts were compiled into a matrix  $O$  (shown below for chromosome 14) where the element  $O_{i,j}$  indicates the number of reads corresponding to an interaction between positions  $i$  and  $j$ . This matrix is visualized as a heatmap that is binned into intervals corresponding to regions of the genome. The number of contacts between loci in these bins are visualized by the intensity of the color in the contact map. A strong diagonal is clearly present, and indicates that regions that are close together in 1D are also likely to interact in 3D. Errors in Hi-C data

interpretation may occur when the assumptions of the technique are violated: for example, the assumption that the reference genome is correct, which may not be true in the case of a cancerous cell. The matrix was

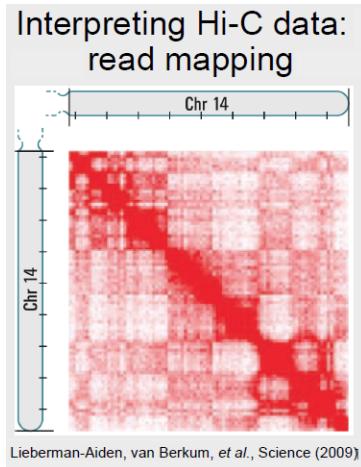


Figure 19.9: Matrix representing Hi-C read count

then normalized to account genetic distance between two regions, and a matrix indicating which interactions are enriched or depleted in the data. Principal Component Analysis (PCA) was used to understand patterns in these contact matrices. After PCA, functional characterization of the data is possible. Hi-C identified two global types of regions:

- Type A, which is characterized by open chromatin, gene richness, and active chromatin marks.
- Type B, which is characterized by closed chromatin, and is gene poor.

Both types of regions are primarily self-interacting and interactions between the two types are infrequent. Hi-C also confirmed the presence of chromosome territories, as there were far more intra-chromosomal rather than inter-chromosomal interactions.

## 19.5 Computational Methods for Studying Nuclear Genome Organization

### 19.5.1 Sources of Bias

The three steps that could potentially introduce biases include: Digestion, Ligation, and Sequencing. Digestion efficiency is a function of the restriction enzymes used and therefore some regions of the genome could be less prone to be digested as their distribution of the particular recognition site could be really sparse. Also, some regions could be enriched in the recognition site and thereby will be over-represented in the results. One solution for this is using many different restriction enzymes and compare the results. Ligation efficiency is a function of the fragment lengths. Depending on how the restriction enzymes cut the sequence, some ends may be more or less likely to ligate together. Finally, sequencing efficiency is a function of the composition of the sequence. Some DNA strands will be more difficult to sequence, based on GC richness and presence of repeats, which will introduce bias.

### 19.5.2 Bias Correction

To minimize ligation bias, non-specific ligation products are removed. Since non-specific ligation products typically have far-away restriction sites, they introduce much larger fragments. In addition, the influence of fragment size on ligation efficiency( $F_{len}(alen, b_{len})$ ), the influence of G/C content on amplification and sequencing( $F_{gc}(a_{gc}, b_{gc})$ ), and the influence of sequence uniqueness on mappability( $M(a) * M(b)$ ) can all be accounted for and corrected with the equation:

$$P(X_{a,b}) = P_{prior} * F_{len}(alen, b_{len}) * F_{gc}(a_{gc}, b_{gc}) * M(a) * M(b)$$

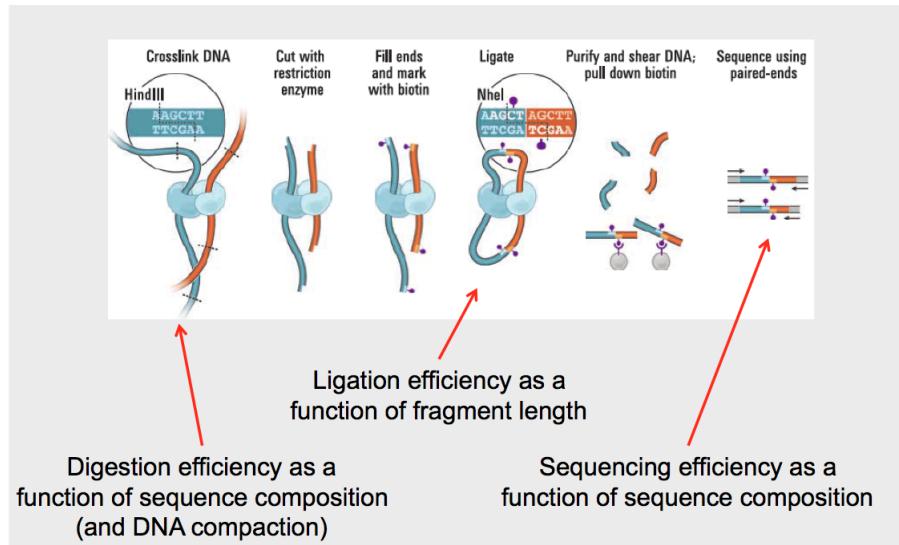


Figure 19.10: Image depicting sources of bias

Alternatively, the sources of bias can be less explicitly represented by the following equation:

$$O_{i,jj} = B_i * B_j * T_{i,j}$$

where the sum of all relative contact probabilities  $T_{i,j}$  for each bin equals 1. The biases are only assumed to be multiplicative. This is solved by matrix balancing, or proportional fitting by an iterative correction algorithm.

### 19.5.3 3D-modeling of 3C-based data

3D-modeling can reveal many general principles of genome organization. Current models are generated using a combination of inter-locus interactions and known spatial distances between nuclear landmarks. However, a lot of uncertainty remains in current 3D-models because the data is gathered from millions of cells. The practical problems affecting 3D-modeling are due to the large amount of data necessary to construct models and the different dynamics between an individual cell and a population, which lead to unstable models. Next generation modeling is trending towards using single cell genomics.

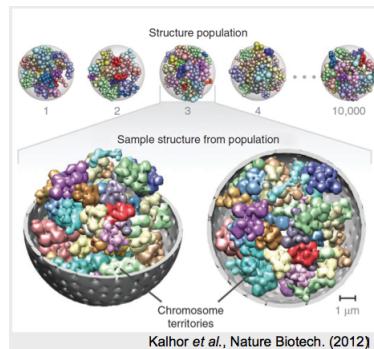


Figure 19.11: Chromosome Territories in 3D

## 19.6 Architecture of Genome Organization

### 19.6.1 Multiple cell types influence on determining architecture

Embryonic stem cells (ESC), Neural Progenitor Cells (NPC), and Astrocytes(AC) are all isogenic cell types (they all start as embryonic stem cells). Embryonic stem cells are constantly dividing and are completely undifferentiated; they generate the neural progenitor cells, which are still dividing but less so, and are only halfway differentiated. The neural progenitor cells then generate the completely differentiated astrocytes. It was discovered that during this differentiation process, some areas switched from being Lamina Associated Domains to being interior domains. In the embryonic stem cells, there is very little transcription. However, transcription goes up as the cells become more and more differentiated. This matches the localization of the domains from being primarily associated with the lamina (and thus not expressed) to being localized to the interior. Even though these cell types each have very different properties, a DamID map shows a high level of similarity between the three isogenic cells as well as an independent fibroblast cell. Hidden Markov Models were employed to identify the Lamina Associated Domains between the cells. A core chromosome architecture was found with about 70% of the chromosome being constitutive (cLad/ ciLAD) and 30% of the chromosome being facultative(fLAD).

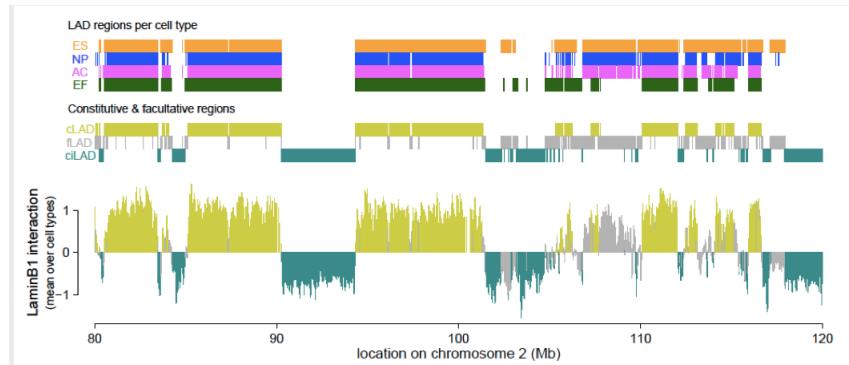


Figure 19.12: A core chromosome architecture is evident. About 70% of regions are constitutive (cLAD/ciLAD) and 30% of regions are facultative (fLAD). The LAD structure that makes up the core architecture is reminiscent of the genome isochore structure

### 19.6.2 Inter-species comparison of lamina associations

To determine lamina associations between species, a mouse and a human were used. A genome wide alignment was constructed between a mouse and a human. For each genomic region in the mouse, the best reciprocal region was matched in the human. Then the human genome was re-mapped, and used to reconstruct a mouse genome. DamID data was projected onto this map and there was 83% concordance between the two genomes (91% for the constitutive regions; 67% for the facultative regions).

### 19.6.3 A-T Content Rule

A-T content has been found to be a strong predictor for lamina association within core architecture. Additional support for this prediction is that the LAD-structure that makes up the core architecture is similar to an isochore structure (a large uniform region of DNA).

## 19.7 Mechanistic Understanding of Genome Architecture

### 19.7.1 Understanding Mitosis and LADs

Organization of chromosomes, particularly in spatial relation to other parts of the chromosome, is not well understood during the mitotic process. The conformation of cells is thought to conform to two different states.

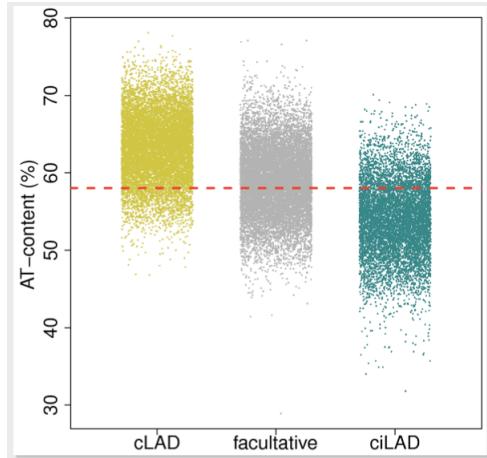


Figure 19.13: AT regions are indicators for constitutive regions

Highly compartmentalized and cell-type specific conformations are almost entirely limited to interphase. During the transition into metaphase, chromosomes enter a locus and tissue-independent folding state.

During the mitotic process, approximately 30% of LADs are positioned along the cellular periphery. This positioning, however, reflects protein-lamina contact at intermittent intervals, however, the cells are restricted to the periphery of the cells. During mitotic division, this laminar positioning is stochastically inherited by child cells.

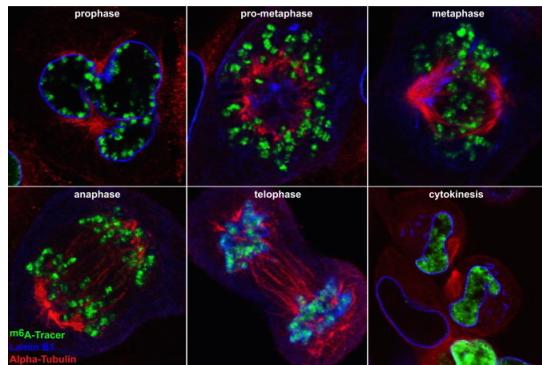


Figure 19.14: LADs Through The (Single) Cell Cycle (Kind et al, Cell 2013)

### 19.7.2 Modeling

Three dimensional modeling will be increasingly important in understanding the chromosomal interactions. Current techniques have modeled the yeast genome and the  $\alpha$ -globin locus (Duan *et al.* Nature (2010), Bau *et al.* Nature SMB (2011)). From modeling studies it has become clear that we cannot generate a direct relationship between contact probability and spatial distance (i.e. contact probability  $\neq$  spatial distance).

However, modeling is typically an inverse problem, which means it is harder to go from one representation of the information to the other than vice-versa. Specifically, it is easier to go from protein structure to a protein contact map than vice versa. Similarly, chromosomal structure is a hard problem, even if we have a contact mapping. This modeling can either be data-driven, where the spatial constraints and distances between areas of the chromosome are taken from chromosome organization data and combined to predict a consensus structure. Alternatively, a model can be built *de novo* from a hypothesis and used to make a proposed structure. The model structure can be used to simulate chromatin interaction data like we would see from Hi-C. Based on how much the model explains actual data, parameters of the model can be updated

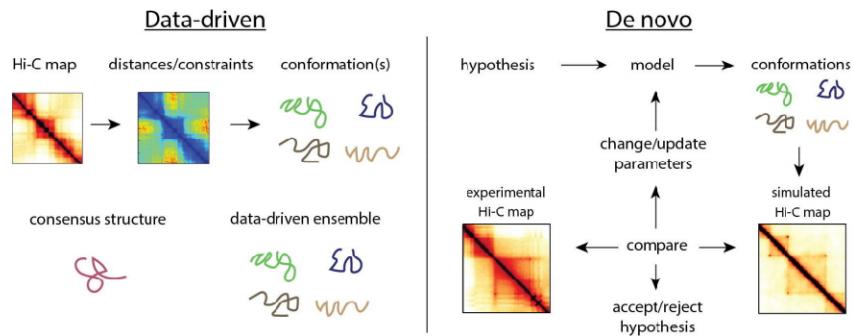


Figure 19.15: Steps in data-driven versus de novo modeling of chromosome structure

to improve results. The hypothesis the model was testing can either be accepted or rejected based on how much it reflects the data.

### 19.7.3 Mechanisms of Loop Extrusion

One of the 3-D structures that has been identified in the genome are loops. These loops define regions of the genome called 'loop domains' which show increased contact frequencies (Figure 19.16).

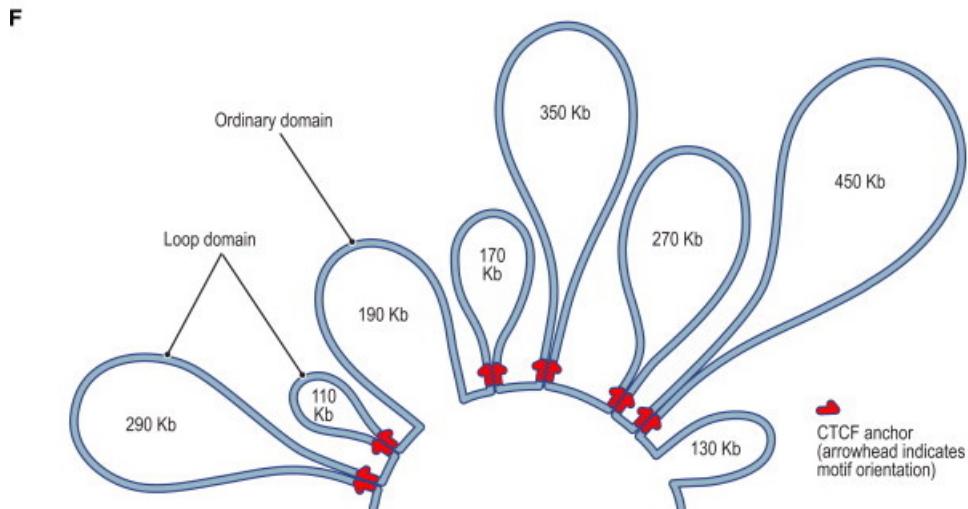


Figure 19.16: Loops (Rao, Huntley, et al. 2014)

Loop domains are recognizable on the Hi-C matrix by intensity peaks at the loop boundaries (Figure 19.17). The loci of these boundaries are associated with CTCF motifs.

The CTCF motifs found at these loop boundaries are almost always found in the convergent location (facing each other). This observation lead to the loop extrusion model for the creation of these loops. This model can be visualized in the following animation: <https://www.youtube.com/watch?v=Tn5qgEqWgW8>.

Briefly, the idea behind loop extrusion is that two rings joined together move along the genome extruding a loop. These rings are stopped at sites where CTCF are bound but only in the convergent direction (Figure 19.18). This model accounts for the observation that loops are never overlapping as well as the fact that loop boundaries are only found at CTCF motifs with the convergent orientation.

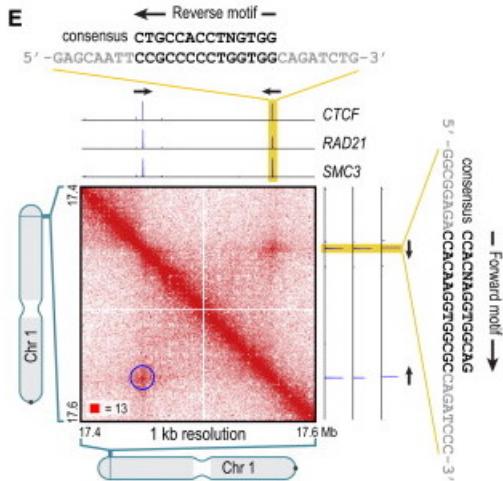


Figure 19.17: Loops in a Hi-C map (Rao, Huntley, et al. 2014)



Figure 1: Dummy

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit,  
 vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida  
 mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna.  
 Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus  
 et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra  
 metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus

Figure 19.18: The loop extrusion model (Sanborn, Rao et al. 2015)

## 19.8 Current Research Directions:

### 19.8.1 LADs

- (a) 30% of the genome is variable between cell types, how are we able to differentiate these differences?
- (b) How do lamina and LADs interact? Is there an attractions between LADs and these domains, or is it based on repulsion of the interior
- (c) Why and how are the genes along the periphery of LADs repressed?

### 19.8.2 TADs and Other Compartments:

- (a) What is the biological basis of compartments, is there some multifaceted component of the compartments?

- (b) How do cohesins work? Are cohesin-extrusion pairs enough to explain all domains?
- (c) Enhancer-promoter loops are confined to specific domains? Are these dynamic components/are they architectural loops mediated by CTCF?

### 19.8.3 Other/Miscellaneous:

- (a) How do we relate the different chromosomal components (i.e. LADs, TADs, polycomb domain, replication origins, histone modifications, gene expression)?
- (b) Evolutionary basis of genomic architecture: was there an evolutionary pressure and when did the folding principles emerge?
- (c) In chromosomal changes do localizations or changes in expression happen first?

#### **Did You Know?**

The question of localization vs expression change has (partially) been addressed. In investigating cells that go through multiple rounds of differentiation, it has been observed that some regions will localize to the lamina in the first differentiation but won't become repressed until the second differentiation!

#### **Body Guard Hypothesis**

The body guard hypothesis was proposed in 1975 by Hsu TC. It suggests that inactive DNA is localized to the periphery of the nucleus so that it can 'guard' the important, active regions of DNA from foreign dangers like viruses or free radicals. Attempts to test the hypothesis by introducing artificial DNA damage have produced circumstantial results, and the question remains open. **Single Cell Experiments**

It is known that cells retain their original organization after mitosis, as shown by chromosome staining experiments. However, recent experiments have shown that there may be a large difference in organization between the parent and daughter cells. Certain global properties, like chromosome territories, are conserved, but organization at a finer detail may greatly differ. Single cell experimentation is an emerging technique that may be able to address this open question.

#### **FAQ**

**Q:** Has anyone tried increasing expression of a gene in the middle of a LAD? What happened?

**A:** It's unclear if there is a specific example of this, however several related studies have been conducted. Researchers have tried to 'tether' a region of DNA to the nuclear lamina to see if it spontaneously becomes deactivated. However, the results were inconclusive as in half of the cases the region would become inactive and in the other half it wouldn't! So far these types of manipulations haven't yielded much, but it was found that if a protein-devoid segment of DNA was digested and mixed with highly purified lamina proteins, the bound fragments reveal a very similar pattern as the LADs. This tells us that lamina directly binds to DNA. However, this does seem to vary between species.

## 19.9 Further Reading

### 19.10 Available Tools and Techniques

- (a) Omic Tools database for many of the measurements (3C, 4C, 5C, HiC, ChIA-PET) <https://omictools.com/3c-4c-5c-hi-c-chia-pet-category>
- (b) HiC-inspector bioinformatics pipeline <https://github.com/HiC-inspector/HiC-inspector>
- (c) Hi-C data analysis library from the Mirny Lab at MIT <https://bitbucket.org/mirnylab/hiclib>
- (d) Bioconductor Hi-C data analysis package for R <http://www.bioconductor.org/packages/release/bioc/html/HiTC.html>
- (e) Juicebox: software to visualize Hi-C data with unlimited zoom <http://aidenlab.org/juicebox/>

### 19.11 What Have We Learned?

New technologies have given us more information about the organization of chromosomes and the biological mechanisms which control this organization. Current research efforts are in place to predict 3-dimensional chromosome packing within the nucleus and how this contribute efforts in cellular reprogramming as well as our understanding of cancer and other diseases.

## Bibliography



---

CHAPTER  
**TWENTY**

---

## INTRODUCTION TO STEADY STATE METABOLIC MODELING

Guest Lecture by James Galagan

Scribed by Meriem Sefta (2011)

Jake Shapiro, Andrew Shum, and Ashutosh Singhal (2010)

Molly Ford Dacus and Anand Oza (2009)

Christopher Garay (2008)

### Figures

---

20.1 The process leading to and including the citric acid cycle.	326
20.2 Adding constraints to extreme pathways.	329
20.3 Maximizing two functions with linear programming.	330
20.4 Removing a reaction is the same as removing a gene from the stoichiometric matrix.	331
20.5 Constraining the feasible solution space may create a new optimal flux.	331
20.6 Model of Coljin et. al [3]	336
20.7 Basic flow in predicting state of a metabolic system under varying drug treatments	337
20.8 Applying expression data set allows constraining of cone shape.	337
20.9 Results of nutrient source prediction experiment.	338

---

### 20.1 Introduction

Metabolic modeling allows us to use mathematical models to represent complex biological systems. This lecture discusses the role of modeling the steady state of biological systems in understanding the metabolic capabilities of organisms. We also briefly discuss how well steady state models are able to replicate in-vitro experiments.

Should this ch included?

#### 20.1.1 What is Metabolism?

According to Matthews and van Holde, metabolism is the totality of all chemical reactions that occur in living matter. This includes catabolic reactions, which are reactions that lead to the breakdown of molecules into smaller components, and anabolic reactions, which are responsible for the creation of more complex molecules (e.g. proteins, lipids, carbohydrates, and nucleic acids) from smaller components. These reactions are responsible for the release of energy from chemical bonds and the storage of this energy. Metabolic reactions are also responsible for the transduction and transmission of information (for example, via the generation of cGMP as a secondary messenger or mRNA as a substrate for protein translation).

### 20.1.2 Why Model Metabolism?

An important application of metabolic modeling is in the prediction of drug effects. An important subject of modeling is the organism *Mycobacterium tuberculosis* [15]. The disruption of the mycolic acid synthesis pathways of this organism can help control TB infection. Computational modeling gives us a platform for identifying the best drug targets in this system. Gene knockout studies in *Escherichia coli* have allowed scientists to determine which genes and gene combinations affect the growth of this important model organism [6]. Both agreements and disagreements between models and experimental data can help us assess our knowledge of biological systems and help us improve our predictions about metabolic capabilities. In the next lecture, we will learn the importance of incorporating expression data into metabolic models. In addition, a variety of infectious disease processes involve metabolic changes at the microbial level.

## 20.2 Model Building

An overarching goal of metabolic modeling is the ability to take a schematic representation of a pathway and change that into a mathematical formula modeling the pathway. For example, converting the following pathway into a mathematical model would be incredibly useful.

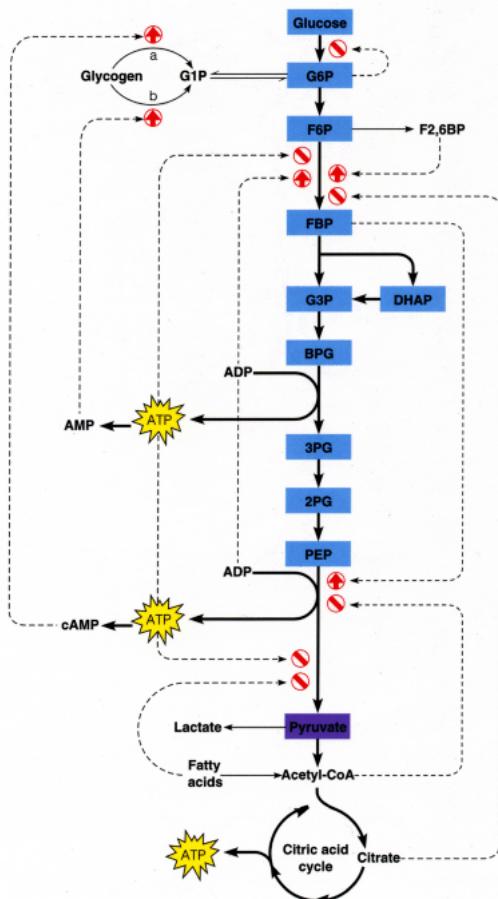


Figure 20.1: The process leading to and including the citric acid cycle.

### 20.2.1 Chemical Reactions

In metabolic models, we are concerned with modeling chemical reactions that are catalyzed by enzymes. Enzymes work by acting on a transition state of the enzyme-substrate complex that lowers the activation

energy of a chemical reaction. The diagram on slide 5 of page 1 of the lecture slides demonstrates this phenomenon. A typical rate equation (which describes the conversion of the substrates S of the enzyme reaction into its products P) can be described by a Michaelis-Menten rate law:

$$\frac{V}{V_{max}} = \frac{[S]}{K_m + [S]}$$

In this equation, V is the rate of the equation as a function of substrate concentration [S]. It is clear that the parameters  $K_m$  and  $V_{max}$  are necessary to characterize the equation.

The inclusion of multiple substrates, products, and regulatory relationships quickly increases the number of parameters necessary to characterize such equations. The figures on slides 1, 2, and 3 of page 2 of the lecture notes demonstrate the complexity of biochemical pathways. Kinetic modeling quickly becomes infeasible: the necessary parameters are difficult to measure, and also vary across organisms [10]. Thus, we are interested in a modeling method that would allow us to use a small number of precisely determined parameters. To this end, we recall the basic machinery of stoichiometry from general chemistry. Consider the chemical equation  $A + 2B \rightarrow 3C$ , which says that one unit of reactant A combines with 2 units of reactant B to form 3 units of reactant C. The rate of formation of the compound X is given by the time derivative of  $[X]$ . Note that C forms three times as fast as A. Therefore, due to the stoichiometry of the reaction, we see that the reaction rate (or reaction flux) is given by

$$\text{flux} = \frac{d[A]}{dt} = \frac{1}{2} \frac{d[B]}{dt} = \frac{1}{3} \frac{d[C]}{dt}$$

This will be useful in the subsequent sections. We must now state the simplifying assumptions that make our model tractable.

### 20.2.2 Steady-State Assumption

The steady state assumption assumes that there is no accumulation of any metabolite in the system. This allows us to represent reactions entirely in terms of their chemistry (i.e. the stoichiometric relationships between the components of the enzymatic reaction). Note that this does not imply the absence of flux through any given reaction. Rather, steady-state actually implies two assumptions that are critical to simplify metabolic modeling. The first is that the internal metabolite concentrations are constant, and the second is that fluxes, ie input and output fluxes, are also constant.

An analogy is a series of waterfalls that contribute water to pools. As the water falls from one pool to another, the water levels do not change even though water continues to flow (see page 2 slide 5). This framework prevents us from being hindered by the overly complicated transient kinetics that can result from perturbations of the system. Since we are usually interested in long-term metabolic capabilities (functions on a scale longer than milliseconds or seconds), the steady state dynamics may give us all the information that we need.

The steady-state assumption makes the ability to generalize across species and reuse conserved pathways in models much more feasible. Reaction stoichiometries are often conserved across species, since they involve only conservation of mass. The biology of enzyme catalysis, and the parameters that characterize it, are not similarly conserved. These include species-dependent parameters such as the activation energy of a reaction, substrate affinity of an enzyme, and the rate constants for various reactions. However, none of these are required for steady-state modeling.

It is also of interest to note that, since time constants for metabolic reactions are usually in the order of milliseconds, most measurement technologies used today are not able to capture these extremely fast dynamics. This is the case of metabolomics mass spectrometry based measurements for example. In this method, the amounts of all the internal metabolites in a system are measured at a given point in time, but measurements can be taken at best every hour. In the majority of circumstances, all that is ever measured is steady state.

### 20.2.3 Reconstructing Metabolic Pathways

There are several databases that can provide the information necessary to reconstruct metabolic pathways in silico. These databases allow reaction stoichiometry to be accessed using Enzyme Commission numbers.

Reaction stoichiometries are the same in all the organisms that utilize a given enzyme. Among the databases of interest are ExPASy [5], MetaCyc [16], and KEGG [14]. These databases often contain pathways organized by function that can be downloaded in SBML format, making pathway reconstruction very easy for well-characterized pathways.

## 20.3 Metabolic Flux Analysis

Metabolic flux analysis (MFA) is a way of computing the distribution of reaction fluxes that is possible in a given metabolic network at steady state. We can place constraints on certain fluxes in order to limit the space described by the distribution of possible fluxes. In this section, we will develop a mathematical formulation for MFA. Once again, this analysis is independent of the particular biology of the system; rather, it will only depend on the (universal) stoichiometries of the reactions in question.

### 20.3.1 Mathematical Representation

Consider a system with  $m$  metabolites and  $n$  reactions. Let  $x_i$  be the concentration of substrate  $i$ , so that the rate of change of the substrate concentration is given by the time derivative of  $x_i$ . Let  $x$  be the column vector (with  $m$  components) with elements  $x_i$ . For simplicity, we consider a system with  $m = 4$  metabolites A, B, C, and D. This system will consist of many reactions between these metabolites, resulting in a complicated balance between these compounds.

Once again, consider the simple reaction  $A + 2B \rightarrow 3C$ . We can represent this reaction in vector form as  $(-1 \ -2 \ 3 \ 0)$ . Note that the first two metabolites (A and B) have negative signs, since they are consumed in the reaction. Moreover, the elements of the vector are determined by the stoichiometry of the reaction, as in Section 2.1. We repeat this procedure for each reaction in the system. These vectors become the columns of the stoichiometric matrix  $S$ . If the system has  $m$  metabolites and  $n$  reactions,  $S$  will be a  $m \times n$  matrix. Therefore, if we define  $v$  to be the  $n$ -component column vector of fluxes in each reaction, the vector  $Sv$  describes the rate of change of the concentration of each metabolite. Mathematically, this can be represented as the fundamental equation of metabolic flux analysis:

$$\frac{dx}{dt} = Sv$$

The matrix  $S$  is an extraordinarily powerful data structure that can represent a variety of possible scenarios in biological systems. For example, if two columns  $c$  and  $d$  of  $S$  have the property that  $c = -d$ , the columns represent a reversible reaction. Moreover, if a column has the property that only one component is nonzero, it represents an exchange reaction, in which there is a flux into (or from) a supposedly infinite sink (or source), depending on the sign of the nonzero component.

We now impose the steady state assumption, which says that the left side of the above equation is identically zero. Therefore, we need to find vectors  $v$  that satisfy the criterion  $Sv = 0$ . Solutions to this equation will determine feasible fluxes for this system.

### 20.3.2 Null Space of $S$

The feasible flux space of the reactions in the model system is defined by the null space of  $S$ , as seen above. Recall from elementary linear algebra that the null space of a matrix is a vector space; that is, given two vectors  $y$  and  $z$  in the nullspace, the vector  $ay + bz$  (for real numbers  $a, b$ ) is also in the null space. Since the null space is a vector space, there exists a basis  $b_i$ , a set of vectors that is linearly independent and spans the null space. The basis has the property that for any flux  $v$  in the null space of  $S$ , there exist real numbers  $\alpha_i$  such that

$$v = \sum_i \alpha_i b_i$$

How do we find a basis for the null space of a matrix? A useful tool is the singular value decomposition (SVD) [4]. The singular value decomposition of a matrix  $S$  is defined as a representation  $S = UEV^*$ , where  $U$  is a unitary matrix of size  $m$ ,  $V$  is a unitary matrix of size  $n$ , and  $E$  is a  $m \times n$  diagonal matrix, with the (necessarily positive) singular values of  $S$  in descending order. (Recall that a unitary matrix is a matrix with

orthonormal columns and rows, i.e.  $U^*U = UU^* = I$  the identity matrix). It can be shown that any matrix has an SVD. Note that the SVD can be rearranged into the equation  $Sv = \sigma u$ , where  $u$  and  $v$  are columns of the matrices  $U$  and  $V$  and  $\sigma$  is a singular value. Therefore, if  $\sigma = 0$ ,  $v$  belongs to the null space of  $S$ . Indeed, the columns of  $V$  that correspond to the zero singular values form an orthonormal basis for the null space of  $S$ . In this manner, the SVD allows us to completely characterize the possible fluxes for the system.

### 20.3.3 Constraining the Flux Space

The first constraint mentioned above is that all steady-state flux vectors must be in the null space. Also negative fluxes are not thermodynamically possible. Therefore a fundamental constraint is that all fluxes must be positive. (Within this framework we represent reversible reactions as separate reactions in the stoichiometric matrix  $S$  having two unidirectional fluxes.)

These two key constraints form a system that can be solved by convex analysis. The solution region can be described by a unique set of Extreme Pathways. In this region, steady state flux vectors  $v$  can be described as a positive linear combination of these extreme pathways. The Extreme Pathways, represented in slide 25 as vectors  $b_i$ , circumscribe a convex flux cone. Each dimension is a rate for some reaction. In slide 25, the z-dimension represents the rate of reaction for  $v_3$ . We can recognize that at any point in time, the organism is living at a point in the flux cone, i.e. is demonstrating one particular flux distribution. Every point in the flux cone can be described by a possible steady state flux vector, while points outside the cone cannot.

One problem is that the flux cone goes out to infinity, while infinite fluxes are not physically possible. Therefore an additional constraint is capping the flux cone by determining the maximum fluxes of any of our reactions (these values correspond to our  $V_{max}$  parameters). Since many metabolic reactions are interior to the cell, there is no need to set a cap for every flux. These caps can be determined experimentally by measuring maximal fluxes, or calculated using mathematical tools such as diffusivity rules.

We can also add input and output fluxes that represent transport into and out of our cells ( $V_{in}$  and  $V_{out}$ ). These are often much easier to measure than internal fluxes and can thus serve to help us to generate a more biologically relevant flux space. An example of an algorithm for solving this problem is the simplex algorithm [1]. Slides 24-27 demonstrate how constraints on the fluxes change the geometry of the flux cone. In reality, we are dealing with problems in higher dimensional spaces.

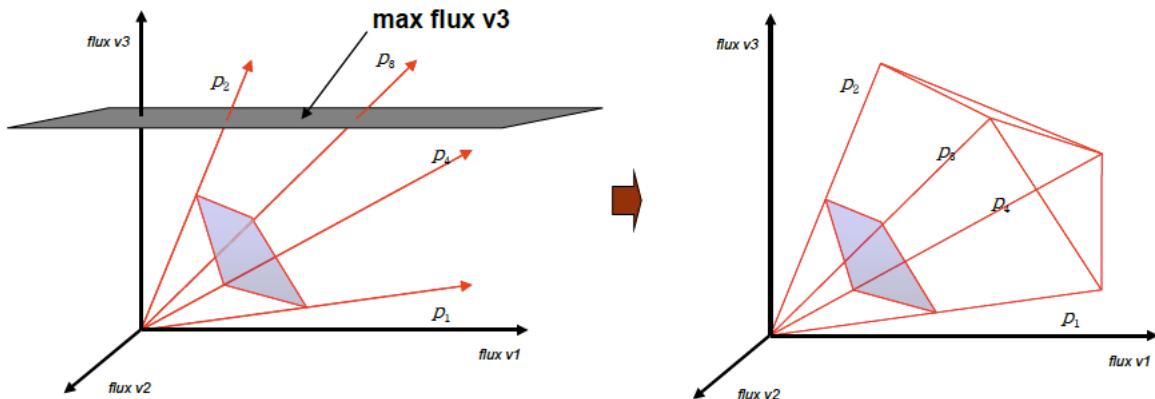


Figure 20.2: Adding constraints to extreme pathways.

### 20.3.4 Linear Programming

Linear programming is a generic solution that is capable of solving optimization problems given linear constraints. These can be represented in a few different forms.

**Canonical Form :**

- Maximize:  $c^T x$
- Subject to:  $Ax \leq b$

**Standard Form :**

- Maximize  $\Sigma c_i * x_i$
- Subject to  $a_{ij}X_i \leq b_i$  for all  $i, j$
- Non-negativity constraint:  $X_i \geq 0$

A concise and clear introduction to Linear Programming is available here: <http://www.purplemath.com/modules/linprog.htm> The constraints described throughout section 3 give us the linear programming problem described in lecture. Linear programming can be considered a first approximation and is a classic problem in optimization. In order to try and narrow down our feasible flux, we assume that there exists a fitness function which is a linear combination of any number of the fluxes in the system. Linear programming (or linear optimization) involves maximizing or minimizing a linear function over a convex polyhedron specified by linear and non-negativity constraints.

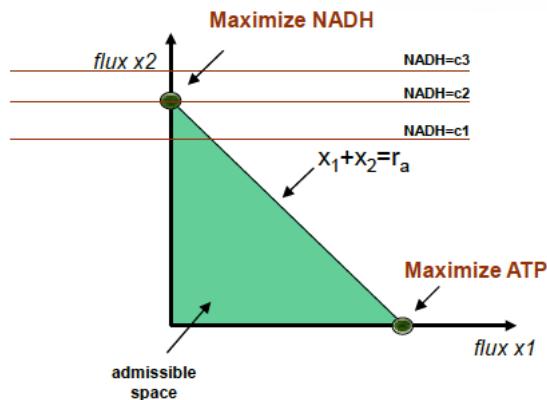


Figure 20.3: Maximizing two functions with linear programming.

We solve this problem by identifying the flux distribution that maximizes an objective function: The key point in linear programming is that our solutions lie at the boundaries of the permissible flux space and can be on points, edges, or both. By definition however, an optimal solution (if one exists) will lie at a point of the permissible flux space. This concept is demonstrated on slide 30. In that slide,  $A$  is the stoichiometric matrix,  $x$  is the vector of fluxes, and  $b$  is a vector of maximal permissible fluxes.

Linear programs, when solved by hand, are generally done by the Simplex method. The simplex method sets up the problem in a matrix and performs a series of pivots, based on the basic variables of the problem statement. In worst case, however, this can run in exponential time. Luckily, if a computer is available, two other algorithms are available. The ellipsoid algorithm and Interior Point methods are both capable of solving any linear program in polynomial time. It is interesting to note, that many seemingly difficult problems can be modeled as linear programs and solved efficiently (or as efficiently as a generic solution can solve a specific problem).

In microbes such as *E. coli*, this objective function is often a combination of fluxes that contributes to biomass, as seen in slide 31. However, this function need not be completely biologically meaningful. For example, we might simulate the maximization of mycolates in *M. tuberculosis*, even though this isn't happening biologically. It would give us meaningful predictions about what perturbations could be performed in vitro that would perturb mycolate synthesis even in the absence of the maximization of the production of those metabolites. Flux balance analysis (FBA) was pioneered by Palsson's group at UCSD and has since been applied to *E. coli*, *M. tuberculosis*, and the human red blood cell [? ].

## 20.4 Applications

### 20.4.1 *In Silico* Detection Analysis

With the availability of such a powerful tool like FBA, more questions naturally arise. For example, are we able to predict gene knockout phenotype based on their simulated effects on metabolism? Also, why would we try to do this, even though other methods, like protein interaction map connective, exist? Such analysis is actually necessary, since other methods do not take into direct consideration the metabolic flux or other specific metabolic conditions.

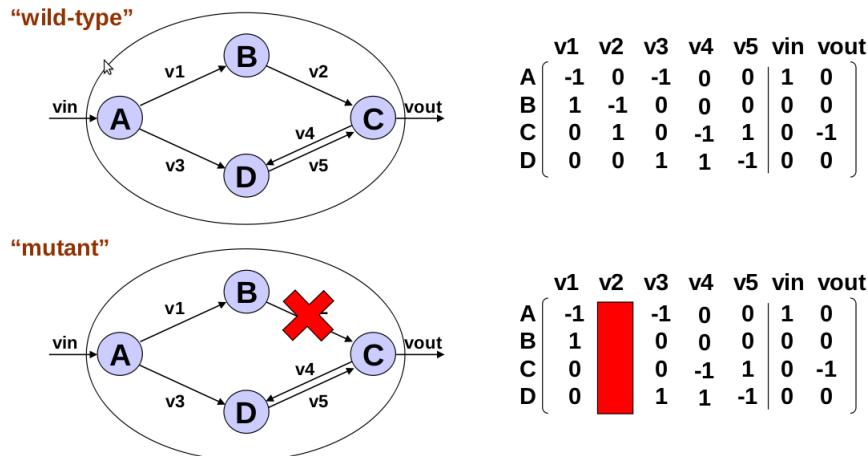


Figure 20.4: Removing a reaction is the same as removing a gene from the stoichiometric matrix.

Knocking out a gene in an experiment is simply modeled by removing one of the columns (reactions) from the stoichiometric matrix. (A question during class clarified that a single gene can knock out multiple columns/reactions.) Thereby, these knockout mutations will further constrain the feasible solution space by removing fluxes and their related extreme pathways. If the original optimal flux was outside the new space, then new optimal flux is created. Thus the FBA analysis will produce different solutions. The solution is a maximal growth rate, which may be confirmed or disproven experimentally. The growth rate at the new solution provides a measure of the knockout phenotype. If these gene knockouts are in fact lethal, then the optimal solution will be a growth rate of zero.

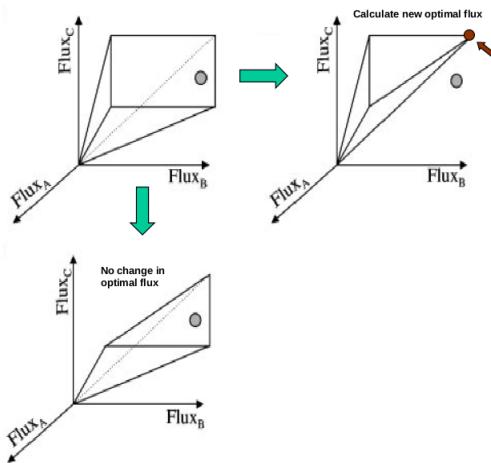


Figure 20.5: Constraining the feasible solution space may create a new optimal flux.

Studies by Edwards, Palsson (1990) explore knockout phenotype prediction use to predict metabolic changes in response to knocking out enzymes in *E. coli*, a prokaryote [? ]. In other words, an *in silico* metabolic model of *E. coli* was constructed to simulate mutations affecting the glycolysis, pentose phosphate, TCA, and electron transport pathways (436 metabolites and 719 reactions included). For each specific condition, the optimal growth of mutants was compared to non-mutants. The *in vivo* and *in silico* results were then compared, with 86% agreement. The errors in the model indicate an underdeveloped model (lack of knowledge). The authors discuss 7 errors not modeled by FBA, including mutants inhibiting stable RNA synthesis and producing toxic intermediates.

### 20.4.2 Quantitative Flux *In Silico* Model Predictions

Can models quantitatively predict fluxes, growth rate? We demonstrate the ability of FBA to give quantitative predictions about growth rate and reaction fluxes given different environmental conditions. More specifically, prediction refers to externally measurable fluxes as a function of controlled uptake rates and environmental conditions. Since FBA maximizes an objective function, resulting in a specific value for this function, we should in theory be able to extract quantitative information from the model.

An early example by Edwards, Ibarra, and Palsson (1990), predicted the growth rate of *E. coli* in culture given a range of fixed uptake rates of oxygen and two carbon sources (acetate and succinate), which they could control in a batch reactor [6]. They assumed that *E. coli* cells adjust their metabolism to maximize growth (using a growth objective function) under given environmental conditions and used FBA to model the metabolic pathways in the bacterium. The input to this particular model is acetate and oxygen, which is labeled as  $V_{IN}$ .

The controlled uptake rates fixed the values of the oxygen and acetate/succinate input fluxes into the network, but the other fluxes were calculated to maximize the value of the growth objective.

The growth rate is still treated as the solution to the FBA analysis. In sum, optimal growth rate is predicted as a function of uptake constraints on oxygen versus acetate and oxygen versus succinate. The basic model is a predictive line and may be confirmed in a bioreactor experimentally by measuring the uptake and growth from batch reactors (note: experimental uptake was not constrained, only measured).

This model by Palsson was the first good proof of principle *in silico* model. The authors quantitative growth rate predictions under the different conditions matched very closely to the experimentally observed growth rates, implying that *E. coli* do have a metabolic network that is designed to maximize growth. It had good true positive and true negative rates. The agreement between the predictions and experimental results is very impressive for a model that does not include any kinetic information, only stoichiometry. Prof. Galagan cautioned, however, that it is often difficult to know what “good” agreement is, because we don't know the significance of the size of the residuals. The organism was grown on a number of different nutrients. Therefore, the investigators were able to predict condition specific growth. Keep in mind this worked, since only certain genes are necessary for some nutrients, like fbp for gluconeogenesis. Therefore, knocking out fbp will only be lethal when there is no glucose in the environment, a specific condition that resulted in a growth solution when analyzed by FBA.

### 20.4.3 Quasi Steady State Modeling (QSSM)

We're now able describe how to use FBA to predict time-dependent changes in growth rates and metabolite concentrations using quasi steady state modeling. The previous example used FBA to make quantitative growth predictions under specific environmental conditions (point predictions). Now, after growth and uptake fluxes, we move on to another assumption and type of model.

Can we use a steady state model of metabolism to predict the time-dependent changes in the cell or environments? We do have to make a number of quasi steady state assumptions (QSSA):

1. The metabolism adjusts to the environmental/cellular changes more rapidly than the changes themselves
2. The cellular and environmental concentrations are dynamic, but metabolism operates on the condition that the concentration is static at each time point (steady state model).

Is it possible to use QSSM to predict metabolic dynamics over time? For example, if there is less acetate being taken in on a per cell basis as the culture grows, then the growth rate must slow. But now, QSSA assumptions are applied. That is, in effect, at any given point in time, the organism is in steady state.

What values does one get as a solution to the FBA problem? There are fluxes the growth rate. We are predicting rate and fluxes (solution) where VIN/OUT included. Up to now we assumed that the input and output are infinite sinks and sources. To model substrate/growth dynamics, the analysis is performed a bit differently from prior quantitative flux analysis. We first divide time into slices  $\delta t$ . At each time point  $t$ , we use FBA to predict cellular substrate uptake ( $S_u$ ) and growth ( $g$ ) during interval  $\delta t$ . The QSSA means these predictions are constant over  $\delta t$ . Then we integrate to get the biomass (B) and substrate concentration (Sc) at the next time point  $t + \delta t$ . Therefore, the new VIN is calculated each time based on points  $\delta t$  in-between time. Thus we can predict the growth rate and glucose and acetate uptake (nutrients available in the environment). The four step analysis is:

1. The concentration at time  $t$  is given by the substrate concentration from the last step plus any additional substrate provided to the cell culture by an inflow, such as in a fed batch.
2. The substrate concentration is scaled for time and biomass (X) to determine the substrate availability to the cells. This can exceed the maximum uptake rate of the cells or be less than that number.
3. Use the flux balance model to evaluate the actual substrate uptake rate, which may be more or less than the substrate available as determined by step 2.
4. The concentration for the next time step is then calculated by integrating the standard differential equations:

$$\frac{dB}{dt} = gB \rightarrow B = B_o e^{gt}$$

$$\frac{dSc}{dt} = -S_u B \rightarrow Sc = Sc_o \frac{X}{g} (e^{gt} - 1)$$

The additional work by Varma et al. (1994) specifies the glucose uptake rate a priori [17]. The model simulations work to predict time-dependent changes in growth, oxygen uptake, and acetate secretion. This converse model plots uptake rates versus growth, while still achieving comparable results *in vivo* and *in silico*. The researchers used quasi steady state modeling to predict the time-dependent profiles of cell growth and metabolite concentrations in batch cultures of *E. coli* that had either a limited initial supply of glucose (left) or a slow continuous glucose supply (right diagram). A great fit is evident.

The diagrams above show the results of the model predictions (solid lines) and compare it to the experimental results (individual points). Thus, in *E. coli*, quasi steady state predictions are impressively accurate even with a model that does not account for any changes in enzyme expression levels over time. However, this model would not be adequate to describe behavior that is known to involve gene regulation. For example, if the cells had been grown on half-glucose/half-lactose medium, the model would not have been able to predict the switch in consumption from one carbon source to another. (This does occur experimentally when *E. coli* activates alternate carbon utilization pathways only in the absence of glucose.)

#### 20.4.4 Regulation via Boolean Logic

There is a number of levels of regulation through which metabolic flux is controlled at the metabolite, transcriptional, translational, post-translational levels. FBA associated errors may be explained by incorporation of gene regulatory information into the models. One way to do this is Boolean logic. The following table describes if genes for associated enzymes are on or off in presence of certain nutrients (an example of incorporating *E. coli* preferences mentioned above):

ON no glucose(0)	ON acetate present(1)
ON glucose present(1)	OFF acetate present(1)

Therefore, one may think that the next step to take is to incorporate this fact into the models. For example, if we have glucose in the environment, the acetate processing related genes are off and therefore absent from the S matrix which now becomes dynamic as a result of incorporation of regulation into our model. In the end, our model is not quantitative. The basic regulation then describes that if one nutrient-processing enzyme is on, the other is off. Basically it is a bunch of Boolean logic, based on presence of enzymes, metabolites, genes, etc. These Boolean style assumptions are then used at every small change in time  $dt$  to evaluate the growth rate, the fluxes, and such variables. Then, given the predicted fluxes, the  $V_{IN}$ , the  $V_{OUT}$ , and the system states, one can use logic to turn genes off and on, effectively a  $\delta S$  per  $\delta t$ . We can start putting together all of the above analyses and come up with a general approach in metabolic modeling. We can tell that if glycolysis is on, then gluconeogenesis must be off.

The first attempt to include regulation in an FBA model was published by Covert, Schilling, and Palsson in 1901 [7]. The researchers incorporated a set of known transcriptional regulatory events into their analysis of a metabolic regulatory network by approximating gene regulation as a Boolean process. A reaction was said to occur or not depending on the presence of both the enzyme and the substrate(s): if either the enzyme that catalyzes the reaction (E) is not expressed or a substrate (A) is not available, the reaction flux will be zero:

$$\text{rxn} = \text{IF (A AND E)}$$

Similar Boolean logic determined whether enzymes were expressed or not, depending on the currently expressed genes and the current environmental conditions. For example, transcription of the enzyme (E) occurs only if the appropriate gene (G) is available for transcription and no repressor (B) is present:

$$\text{trans} = \text{IF (G AND NOT B)}$$

The authors used these principles to design a Boolean network that inputs the current state of all relevant genes (on or off) and the current state of all metabolites (present or not present), and outputs a binary vector containing the new state of each of these genes and metabolites. The rules of the Boolean network were constructed based on experimentally determined cellular regulatory events. Treating reactions and enzyme/metabolite concentrations as binary variables does not allow for quantitative analysis, but this method can predict qualitative shifts in metabolic fluxes when merged with FBA. Whenever an enzyme is absent, the corresponding column is removed from the FBA reaction matrix, as was described above for knockout phenotype prediction. This leads to an iterative process:

1. Given the initial states of all genes and metabolites, calculate the new states using the Boolean network;
2. perform FBA with appropriate columns deleted from the matrix, based on the states of the enzymes, to determine the new metabolite concentrations;
3. repeat the Boolean network calculation with the new metabolite concentrations; etc. The above model is not quantitative, but rather a pure simulation of turning genes on and off at any particular time instant.

On a few metabolic reactions, there are rules about allowing organism to shift carbon sources (C1, C2).

An application of this method from the study by Covert et al.[7] was to simulate diauxic shift, a shift from metabolizing a preferred carbon source to another carbon source when the preferred source is not available. The modeled process includes two gene products, a regulatory protein RPc1, which senses (is activated by) Carbon 1, and a transport protein Tc2, which transports Carbon 2. If RPc1 is activated by Carbon 1, Tc2 will not be transcribed, since the cell preferentially uses Carbon 1 as a carbon source. If Carbon 1 is not available, the cell will switch to metabolic pathways based on Carbon 2 and will turn on expression of Tc2. The Booleans can represent this information:

$$\text{RPc1} = \text{IF(Carbon1)} \quad \text{Tc2} = \text{IF NOT(RPc1)}$$

Covert et al. found that this approach gave predictions about metabolism that matched results from experimentally induced diauxic shift. This diauxic shift is well modeled by the in silico analysis see above figure. In segment A, C1 is used up as a nutrient and there is growth. In segment B, there is no growth as

C1 has run out and C2 processing enzymes are not yet made, since genes have not been turned on (or are in the process), thus the delay of constant amount of biomass. In segment C, enzymes for C2 turned on and the biomass increases as growth continues with a new nutrient source. Therefore, if there is no C1, C2 is used up. As C1 runs out, the organism shifts metabolic activity via genetic regulation and begins to take up C2. Regulation predicts diauxie, the use of C1 before C2. Without regulation, the system would grow on both C1 and C2 together to max biomass.

So far we have discussed using this combined FBA-Boolean network approach to model regulation at the transcriptional/translational level, and it will also work for other types of regulation. The main limitation is for slow forms of regulation, since this method assumes that regulatory steps are completed within a single time interval (because the Boolean calculation is done at each FBA time step and does not take into account previous states of the system). This is fine for any forms of regulation that act at least as fast as transcription/translation. For example, phosphorylation of enzymes (an enzyme activation process) is very fast and can be modeled by including the presence of a phosphorylase enzyme in the Boolean network. However, regulation that occurs over longer time scales, such as sequestration of mRNA, is not taken into account by this model. This approach also has a fundamental problem in that it does not allow actual experimental measurements of gene expression levels to be inputted at relevant time points.

We do not need our simulations to artificially predict whether certain genes are on or off. Microarray expression data allows us to determine which genes are being expressed, and this information can be incorporated into our models.

#### 20.4.5 Coupling Gene Expression with Metabolism

In practice, we do not need to artificially model gene levels, we can measure them. As discussed previously, it is possible to measure the expression levels of all the mRNAs in a given sample. Since mRNA expression data correlates with protein expression data, it would be extremely useful to incorporate it into the FBA. Usually, data from microarray experiments is clustered, and unknown genes are hypothesized to have function similar to the function of those known genes with which they cluster. This analysis can be faulty, however, as genes with similar actions may not always cluster together. Incorporating microarray expression data into FBA could allow an alternate method of interpretation of the data. Here arises a question, what is the relationship between gene level and flux through a reaction?

Say the reaction  $A \rightarrow B$  is catalyzed by an enzyme. If a lot of A present, increased expression of the gene for the enzyme causes increased reaction rate. Otherwise, increasing gene expression level will not increase reaction rate. However, the enzyme concentration can be treated as a constraint on the maximum possible flux, given that the substrate also has a reasonable physiological limit.

The next step, then, is to relate the mRNA expression level to the enzyme concentration. This is more difficult, since cells have a number of regulatory mechanisms to control protein concentrations independently of mRNA concentrations. For example, translated proteins may require an additional activation step (e.g. phosphorylation), each mRNA molecule may be translated into a variable number of proteins before it is degraded (e.g. by antisense RNAs), the rate of translation from mRNA into protein may be slower than the time intervals considered in each step of FBA, and the protein degradation rate may also be slow. Despite these complications, the mRNA expression levels from microarray experiments are usually taken as upper bounds on the possible enzyme concentrations at each measured time point. Given the above relationship between enzyme concentration and flux, this means that the mRNA expression levels are also upper bounds on the maximum possible fluxes through the reactions catalyzed by their encoded proteins. The validity of this assumption is still being debated, but it has already performed well in FBA analyses and is consistent with recent evidence that cells do control metabolic enzyme levels primarily by adjusting mRNA levels. (In 1907, Professor Galagan discussed a study by Zaslaver et al. (1904) that found that genes required in an amino acid biosynthesis pathway are transcribed sequentially as needed [2]). This is a particularly useful assumption for including microarray expression data in FBA, since FBA makes use of maximum flux values to constrain the flux balance cone.

Colijn et al. address the question of algorithmic integration of expression data and metabolic networks [3]. They apply FBA to model the maximum flux through each reaction in a metabolic network. For example, if microarray data is available from an organism growing on glucose and from an organism growing on acetate, significant regulatory differences will likely be observed between the two datasets.  $V_{max}$  tells us what the

maximum we can reach. Microarray detects the level of transcripts, and it gives an upper boundary of  $V_{max}$ .

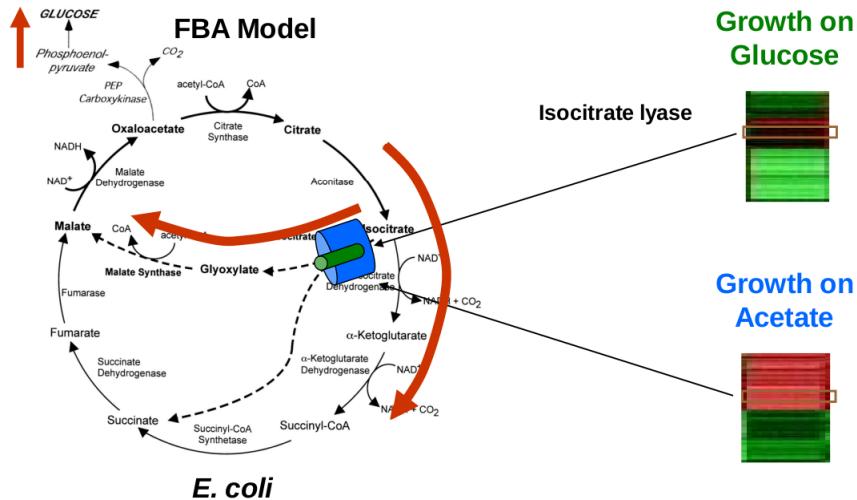


Figure 20.6: Model of Coljin et. al [3]

In addition to predicting metabolic pathways under different environmental conditions, FBA and microarray experiments can be combined to predict the state of a metabolic system under varying drug treatments. For example, several TB drugs target mycolic acid biosynthesis. Mycolic acid is a major cell wall constituent. In a 1904 paper by Boshoff et al., researchers tested 75 drugs, drug combinations, and growth conditions to see what effect different treatments had on mycolic acid synthesis [9]. In 1905, Raman et al. published an FBA model of mycolic acid biosynthesis, consisting of 197 metabolites and 219 reactions [13].

The basic flow of the prediction was to take a control expression value and a treatment expression value for a particular set of genes, then feed this information into the FBA and measure the final effect on the treatment on the production of mycolic acid. To examine predicted inhibitors and enhancers, they examined “significance,” which examines whether the effect is due to noise, and “specificity,” which examines whether the effect is due to mycolic acid or overall suppression/enhancement of metabolism. The results were fairly encouraging. Several known mycolic acid inhibitors were identified by the FBA. Interesting results were also found among drugs not specifically known to inhibit mycolic acid synthesis. 4 novel inhibitors and 2 novel enhancers of mycolic acid synthesis were predicted. One particular drug, Triclosan, appears to be an enhancer according to the FBA model, whereas it is currently “known” as an inhibitor. Further study of this particular drug would be interesting. Experimental testing and validation are currently in progress.

Clustering may also be ineffective in identifying function of various treatments. Predicted inhibitors, and predicted enhancers of mycolic acid synthesis are not clustered together. In addition, no labeled training set is required for FBA-based algorithmic classification, whereas it is necessary for supervised clustering algorithms.

#### 20.4.6 Predicting Nutrient Source

Now, we get the idea of predicting the nutrient source that an organism may be using in an environment, by looking at expression data and looking for associated nutrient processing gene expression. This is easier, since we can't go into the environment and measure all chemical levels, but we can get expression data rather easily. That is, we try to predict a nutrient source through predictions of metabolic state from expression data, based on the assumption that organisms are likely to adjust metabolic state to available nutrients. The nutrients may then be ranked by how well they match the metabolic states.

The other way around could work too. Can I predict a nutrient given a state? Such predictions could be useful for determining the nutrient requirements of an organism with an unknown natural environment, or for determining how an organism changes its environment. (TB, for example, is able to live within the environment of a macrophage phagolysosome, presumably by altering the environmental conditions in the

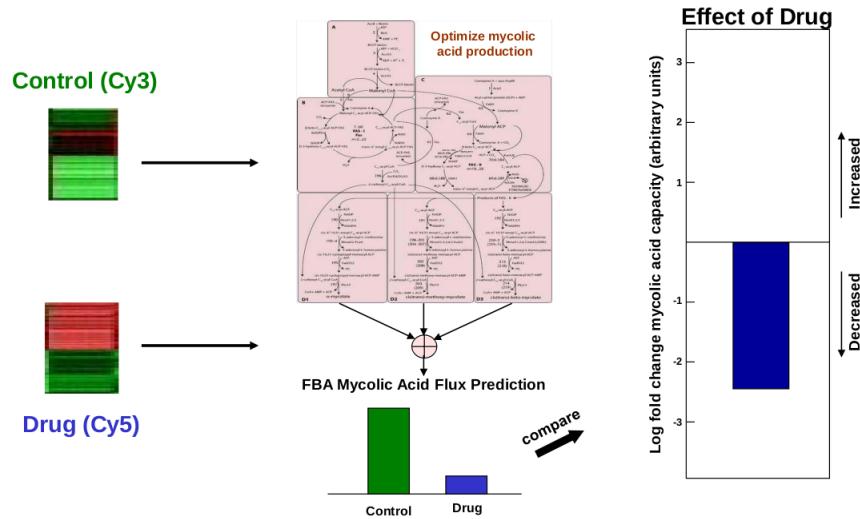


Figure 20.7: Basic flow in predicting state of a metabolic system under varying drug treatments

phagolysosome and preventing its maturation.)

We can use FBA to define a space of possible metabolic states and choose one. The basic steps are to:

- Start with max flux cone (representing best growth with all nutrients available in environment). Find optimal flux for each nutrient.
- Apply expression data set (still not knowing nutrient). This will allow you to constrain the cone shape and figure out the nutrient, which is represented as one with the closest distance to optimal solution.

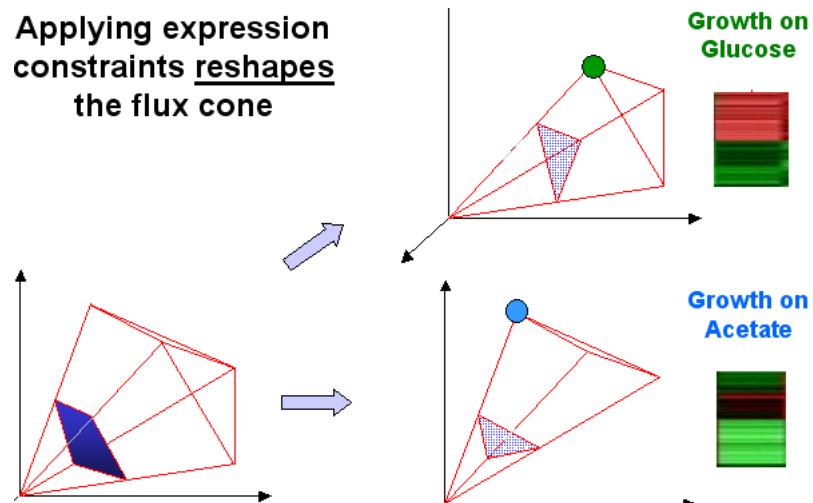


Figure 20.8: Applying expression data set allows constraining the cone shape.

In Figure 8, you may see that the first cone has a number of optimals, so the real nutrient is unknown. However, after expression data is applied, the cone is reshaped. It has only one optimal, which is still in feasible space and thereby must be that nutrient you are looking for.

As before, the measured expression levels provide constraints on the reaction fluxes, altering the shape of the flux-balance cone (now the expression-constrained flux balance cone). FBA can be used to determine the optimal set of fluxes that maximize growth within these expression constraints, and this set of fluxes can be compared to experimentally-determined optimal growth patterns under each environmental condition

of interest. The difference between the calculated state of the organism and the optimal state under each condition is a measure of how sub-optimal the current metabolic state of the organism would be if it were in fact growing under that condition.

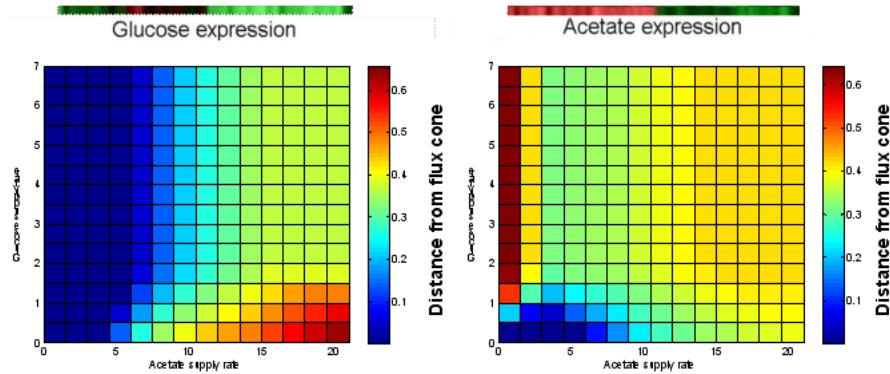


Figure 20.9: Results of nutrient source prediction experiment.

Expression data from growth and metabolism may then be applied to predict the carbon source being used. For example, consider E. coli nutrient product. We can simulate this system for glucose versus acetate. The color indicates the distance from the constrained flux cone to the optimal flux solution for that nutrient combo (same procedure described above). Then, multiple nutrients may be ranked, prioritized according to expression data. Unpublished data from Desmond Lun and Aaron Brandes provide an example of this approach.

They used FBA to predict which nutrient source E. coli cultures were growing on, based on gene expression data. They compared the known optimal fluxes (the optimal point in flux space) for each nutrient condition to the allowed optimal flux values within the expression-constrained flux-balance cone. Those nutrient conditions with optimal fluxes that remained within (or closest to) the expression-constrained cone were the most likely possibilities for the actual environment of the culture.

Results of the experiment are shown in Figure 9, where each square in the results matrices is colored based on the distance between the optimal fluxes for that nutrient condition and the calculated optimal fluxes based on the expression data. Red values indicate large distances from the expression-constrained flux cone and blue values indicate short distances from the cone. In the glucose-acetate experiments, for example, the results of the experiment on the left indicate that low acetate conditions are the most likely (and glucose was the nutrient in the culture) and the results of the experiment on the right indicate that low glucose/medium acetate conditions are the most likely (and acetate was the nutrient in the culture). When 6 possible nutrients were considered, the model always predicted the correct one, and when 18 possible nutrients were considered, the correct one was always one of the top 4 ranking predictions. These results suggest that it is possible to use expression data and FBA modeling to predict environmental conditions from information about the metabolic state of an organism.

This is important because TB uses fatty acids in macrophages in immune systems. We do not know which ones exactly are utilized. We can figure out what the TB sees in its environment as a food source and proliferation factor by analyzing what related nutrient processing genes are turned on at growth phases and such. Thereby we can figure out the nutrients it needs to grow, allowing for a potential way to kill it off by not supplying such nutrients or knocking out those particular genes.

It is easier to get expression data to see flux activity than see what's being used up in the environment by analyzing the chemistry on such a tiny level. Also, we might not be able to grow some bacteria in lab, but we can solve the problem by getting the expression data from the bacteria growing in a natural environment and then seeing what it is using to grow. Then, we can add it to the laboratory medium to grow the bacteria successfully.

## 20.5 Current Research Directions

### 20.6 Further Reading

- Becker, S. A. and B. O. Palsson (1908). “Context-Specific Metabolic Networks Are Consistent with Experiments.” *PLoS Computational Biology* 4(5): e1000082.
  - If gene expression lower than some threshold, turn the gene off in the model.
- Shlomi, T., M. N. Cabili, et al. (1908). “Network-based prediction of human tissue-specific metabolism.” *Nat Biotech* 26(9): 1003-1010.
  - Nested optimization problem.
  - First, standard FBA
  - Second, maximize the number of enzymes whose predicted flux activity is *consistent with their measured expression level*

### 20.7 Tools and Techniques

- Kegg
- BioCyc
- Pathway Explorer ([pathwayexplorer.genome.tugraz.at](http://pathwayexplorer.genome.tugraz.at))
- Palssons group at UCSD (<http://gcrg.ucsd.edu/>)
- [www.systems-biology.org](http://www.systems-biology.org)
- Biomodels database ([www.ebi.ac.uk/biomodels/](http://www.ebi.ac.uk/biomodels/))
- JWS Model Database ([jws.biochem.sun.ac.za/database/index.html](http://jws.biochem.sun.ac.za/database/index.html))

### 20.8 What Have We Learned?

## Bibliography

- [1]
- [2] Zaslaver A, Mayo AE, Rosenberg R, Bashkin P, Sberro H, Tsalyuk M, Surette MG, and Alon U. Just-in-time transcription program in metabolic pathways. *Nat. Gen.*, 36:486–491, 2004.
- [3] Caroline Coljin. Interpreting expression data with metabolic flux models: Predicting mycobacterium tuberculosis mycolic acid production. *PLoS Computational Biology*, 5(8), Aug 2009.
- [4] Price N. D., Reed J. L., Papin J.A, Famili I., and Palsson B.O. Analysis of metabolic capabilities using singular value decomposition of extreme pathway matrices. *Biophys J.*, 84(2):794–804, Feb 2003.
- [5] Gasteiger E., Gattiker A., Hoogland C. andIvanyi I., Appel R.D., , and Bairoch A. Expasy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res*, 31(13):3784–3788.
- [6] J.S. Edwards, R. U. Ibarra, and B.O. Palsson. In silico predictions of e coli metabolic capabilities are consistent with experimental data. *Nat Biotechnology*, 19:125–130, 2001.
- [7] Covert M et al. Regulation of gene expression in flux balance models of metabolism. *Journal of Theoretical Biology*, 213:73–88, Nov 2001.

- [8] J. Forster, I. Famili, B.O. Palsson, and J. Nielsen. Large-scale evaluation of in silico gene deletions in *saccharomyces cerevisiae*. *OMICS*, 7(2):193–202, 2003. PMID: 14506848.
- [9] Boshoff H.I., Myers T.G., Copp B.R., McNeil M.R., Wilson M.A., and Bary C.E. The transcriptional response of mycobacterium tuberculosis to inhibitors of metabolism: novel insights into drug mechanisms of action. *J Biol Chem*, 279:40174–40184, Sep 2004.
- [10] Holmberg. On the practical identifiability of microbial-growth models incorporating michaelis-menten type nonlinearities. *Mathematical Biosciences*, 62(1):23–43, 1982.
- [11] Edwards J.S. and Palsson B.O. volume 97, pages 5528–5533. Proceedings of the National Academy of Sciences of the United States of America, May 2000. PMC25862.
- [12] Edwards J.S., Covert M., , and Palsson B. Metabolic modeling of microbes: the flux balance approach. *Environmental Microbiology*, 4(3):133–140, 2002.
- [13] Raman Karthik, Preethi Rajagopalan, and Nagasuma Chandra. Flux balance analysis of mycolic acid pathway: Targets for anti-tubercular drugs. *PLoS Computational Biology*, 1, Oct 2005.
- [14] Kanehisa M., Goto S., Kawashima S., and Nakaya. From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res.*, 34, 2006.
- [15] Jamshidi N. and Palsson B. Investigating the metabolic capabilities of mycobacterium tuberculosis h37rv using the in silico strain inj661 and proposing alternative drug targets. *BMC Systems Biology*, 26, 2007.
- [16] Caspi R., Foerster H., Fulcher C.A., Kaipa P., Krummenacker M., Latendresse M., Paley S., Rhee S.Y., Shearer A.G., Tissier C., Walk T.C. ZhangP., and Karp P. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Res*, 36(Suppl), 2008.
- [17] A. Varma and B. O. Palsson. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type escherichia coli w3110. *Applied and Environmental Microbiology*, 60:3724–3731, Oct 1994.

---

CHAPTER  
TWENTYONE

---

## THE ENCODE PROJECT: SYSTEMATIC EXPERIMENTATION AND INTEGRATIVE GENOMICS

[add author](#)

missing

### Figures

---

21.1 Snapshot of ENCODE project experiment matrix.	342
21.2 Overview of Chip-seq [3]	342
21.3 ENCODE uniform Processing Pipeline	343
21.4 Shifting forward and reverse strand signals, Cross Correlation plot	344
21.5 IDR to assess reproducibility of CHIP-seq datasets. Scatter plots display signal scores of peaks that overlap in each replicate pair. (A,B) results for high quality replicate. (C) Estimated IDR for varying rank thresholds. [1]	344

---

### 21.1 Introduction

The human genome was sequenced in 2003, an important step in understanding the blueprint of life. However, before this information can be fully utilized, the location, identity, and function of all protein-encoding and non-protein-encoding genes must be determined. Moreover, the human genome has many other functional elements, ranging from promoters, regulatory sequences, and other factors that determine chromatin structure. These must also be determined to fully understand the human genome.

The ENCODE (Encyclopedia of DNA Elements) project aims to solve these problems by delineating all functional elements of the human genome [cite website](#). To accomplish this goal, a [consortium was formed](#) to guide the project. The consortium aimed to advance and develop technologies for annotating the human genome with higher accuracy, completeness, and cost-effectiveness, along with more standardization. They also aimed to develop a series of computational techniques to parse and analyze the data obtained.

To accomplish this goal, a pilot project was launched. The ENCODE pilot project aimed to study 1% of the human genome in depth, roughly from 2003 to 2007. From 2007 to 2012, the ENCODE project ramped up to annotate the entire genome. Finally, from 2012 onwards, the ENCODE project aims further increases in all dimensions: deeper sequencing, more assays, more transcription factors, etc.

This chapter will describe some of the experimental and computational techniques used in the ENCODE project.

### 21.2 Experimental Techniques

The ENCODE project used a wide range of experimental techniques, ranging from RNA-seq, CAGE-seq, Exon Arrays, MAINE-seq, Chromatin ChIP-seq, DNase-seq, and many more.

		Assays																									
		Methyl Array	Methyl RRBS	Open Chromatin	DNase-Seq	FAIRE-seq	RNA Binding Proteins	RIP Gene ST	RIP Tiling Array	RIP-seq	RNA Profiling	CAGE	Exon Array	RNA-chip	RNA-PET	RNA-seq	Small RNA-seq	TFBSS & Histones	ChIP-seq	Other	SC	ChIA-PET	Combined	DNA-PET	Genotype	Nucleosome	Proteogenomics
		Cell Types																									
<b>Tier 1</b>		GM12878	1 1	2 1	7 4	4 4	6 2	6 4	12 10	5 3	133	2 1	2 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	
H1-hESC		H1-hESC	1 1	2 1	3 1	4 4	9 7	9 6	17 17	7 7	91	1 1	2 2	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	
K562		K562	1 1	3 16	3 3	6 4	4 4	9 7	9 6	17 17	7 7	224	2 2	2 2	3 3	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	1 1	
<b>Tier 2</b>		A549	1 1	1 2	1 1			3 1	2 1	3 2	10 9	87								1 1							
CD20+		CD20+		1 1	1 1			1 1		2 1		4								1 1							
CD20+_R001778		CD20+_R001778			1 1							5								2 2							
CD20+_R001794		CD20+_R001794				1 1						4								4 4							
H1-neurons		H1-neurons				3 3	3 3	4 4				93	1 1	1 1	2 2	1 1	1 1	1 1	1 1	1 1							
HeLa-S3		HeLa-S3			1 1	1 2	1 1	4 4			114	6 2	5 2	2 2	8 8	3 3	36	11									
HepG2		HepG2			1 1			5 2		2 2	8 8	36								1 1	2 2	1 1					
HUVEC		HUVEC			1 1			3 3		4 4	9 9	11															
IMR90		IMR90			1 1																						

Figure 21.1: Snapshot of ENCODE project experiment matrix.

One of the most important techniques used was ChIP-seq (chromatin immunoprecipitation followed by sequencing). The first step in a ChIP experiment is to target DNA fragments associated with a specific protein. This is done by using an anti-body that targets the specific protein and is used to immunoprecipitate the DNA-protein complex. The final step is to assay the DNA. This will determine the sequences bound to the proteins.

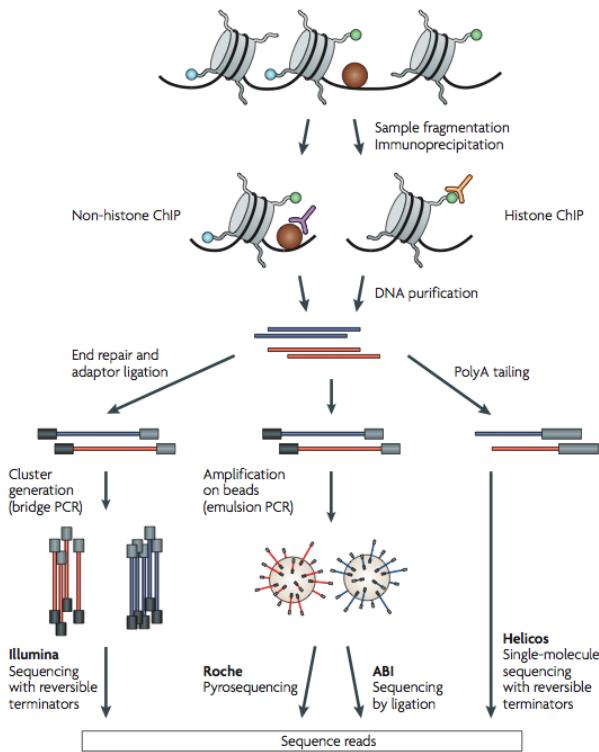


Figure 21.2: Overview of Chip-seq [3]

ChIP-seq has several advantages over previous techniques (e.g. ChIP-chip). For example, ChIP-seq has single nucleotide resolution and its alignability increases with read length. However, ChIP-seq has several disadvantages. Sequencing errors tend to increase substantially near the end of reads. Also, with low number of reads, sensitivity and specificity tend to decrease when detecting enriched regions. Both of these problems arise when processing the data and many of the computational techniques seek to rectify this.

## 21.3 Computational Techniques

This section will focus on techniques on processing raw data from the ENCODE project. Before ENCODE data can be analyzed (e.g. for motif discovery, co-association analysis, signal aggregation over elements, etc), the raw data must be processed.

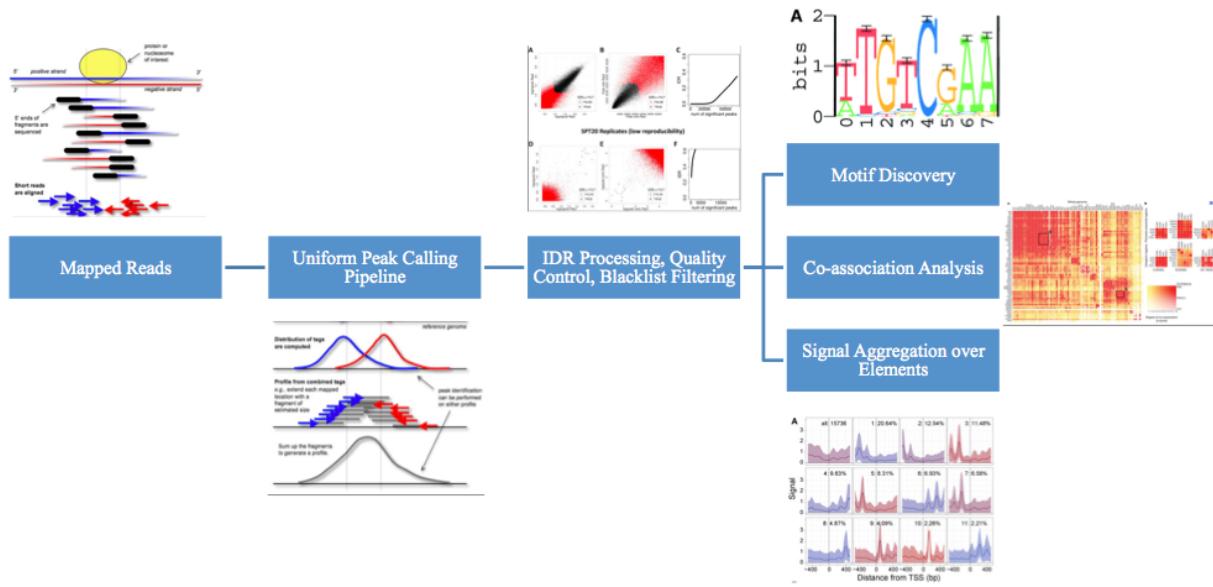


Figure 21.3: ENCODE uniform Processing Pipeline

Even before the data is processed, some quality control is applied. Quality control is needed for several reasons. Even without anti-bodies, reads are not uniformly-scattered. The biological reasons include non-uniform fragmentation of the genome, open chromatin regions fragmenting easier, and repetitive sequences over-collapsed in assembled genomes. The ENCODE project corrected for these biases in several ways. Portions of the DNA were removed before the ChIP step, removing large portions of unwanted data. Control experiments were also conducted without the use of anti-bodies. Finally, fragment input DNA sequence reads were used as a background.

Because of inherent noise in the ChIP-seq process, some reads will be of lower quality. Using a read quality metric, reads below a threshold were thrown out.

Shorter reads (and to a lesser extent, longer reads) can map to exactly one location (uniquely mapping), multiple locations (repetitive mapping), or no locations at all (unmappable) in the genome. There are many potential ways to deal with repetitive mapping, ranging from probabilistically spreading the read to use an EM approach. However, since the ENCODE project aims to be as correct as possible, it does not assign repetitive reads to any location.

If a sample does not contain sufficient DNA and/or if it is over-sequenced, you will simply be repeatedly sequencing PCR duplicates of a restricted pool of distinct DNA fragments. This is known a low-complexity library and is not desirable. To solve this problem, a histogram with the number of duplicates is created and samples with a low non-redundant fraction (NRF) are thrown out.

ChIP-seq randomly sequences from one end of each fragment, so to determine which reads came from which segment, typically strand cross-correlation analysis is used [Fig. 04]. To accomplish this, the forward and reverse strand signals are calculated. Then, they are sequentially shifted towards each other. At every step, the correlation is calculated. At the fragment length offset  $f$ , the correlation peaks.  $f$  is the length at which ChIP DNA is fragmented. Using further analysis, we can determine that we should have a high absolute cross-correlation at fragment length, and high fragment length cross-correlation relative to read-length cross-correlation. The RSC (Relative Strand Correlation) should be greater than 1.

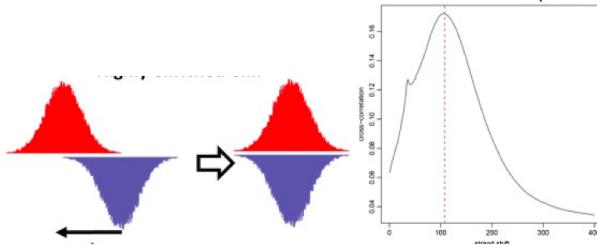


Figure 21.4: Shifting forward and reverse strand signals, Cross Correlation plot

$$RSC = \frac{CC_{fragment} - \min(CC)}{CC_{readlength} - \min(CC)} \quad (21.1)$$

Once quality control is applied, the data is further processed to determine actual areas of enrichment. To accomplish this, the ENCODE project used a modified version of peak calling. There are many existing peak calling algorithms, but the ENCODE project used MACS and PeakSeq, as they are deterministic. However, it is not possible to set a uniform p-value or false discovery rate (FDR) constant. The FDR and p-value depends on ChIP and input sequencing depth, the binding ubiquity of the factor, and is highly unstable. Moreover, different tools require different values.

The ENCODE project uses replicates (of the same experiment) and combines the data to find more meaningful results. Simple solutions have major issues: taking the union of the peaks keeps garbage from both, the intersection is too stringent and throws away good peaks, and taking the sum of the data does not exploit the independence of the datasets. Instead, the ENCODE project uses the independent discovery rate (IDR). The key idea is that true peaks will be highly ranked in both replicates. Thus, to find significant peaks, the peaks are considered in rank order, until ranks are no longer correlated.

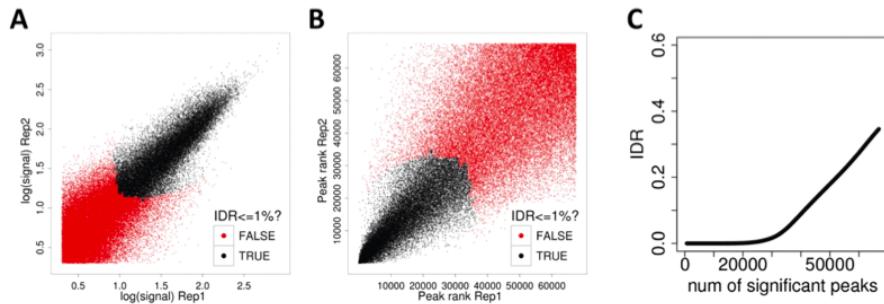


Figure 21.5: IDR to assess reproducibility of ChIP-seq datasets. Scatter plots display signal scores of peaks that overlap in each replicate pair. (A,B) results for high quality replicate. (C) Estimated IDR for varying rank thresholds. [1]

The cutoff could be different for the two replicates and actual peaks included may differ between replicates. It is modeled as a Gaussian mixture model, which can be fit via an EM-like algorithm. Using IDR leads to higher consistency between peak callers. This is because FDR only relies on enrichment over input, IDR exploits replicates. Also, using sampling methods, if there is only one replicate, the IDR pipeline can still be used with pseudo-replicates.

Expand on the following topics; Aggregation Analysis, Predictive models of TF co-association and gene expression, Scaling and Saturaion Analysis

## 21.4 Current Research Directions

The ENCODE project is still ongoing. Using saturation techniques, we believe we only have discovered a maximum 50% of elements. This number is likely to be lower due to inaccessible cell types and other factors. Also, several cell types are extremely rare and difficult to access, so sequencing data from these cell types is another challenge.

In computational frontiers, the ENCODE project has produced an enormous amount of raw data. Similar to how the full sequence of the human genome unleashed a series of computational projects, the ENCODE data can be used for a variety of computational projects.

## 21.5 Further Reading

The Nature site with ENCODE papers is available at <http://www.nature.com/encode/>.

The official ENCODE portal is <http://encodeproject.org/ENCODE/>.

To browse ENCODE data, visit <http://encodeproject.org/cgi-bin/hgHubConnect>.

Data processing tools for ENCODE data are available at <http://encodeproject.org/ENCODE/analysis.html>.

## 21.6 Tools and Techniques

ENCODE data mining, [http://genome.ucsc.edu/cgi-bin/hgTables?db=hg18&hgta\\_group=regulation&hgta\\_track=wgEncodeHudsonalphaChipSeq](http://genome.ucsc.edu/cgi-bin/hgTables?db=hg18&hgta_group=regulation&hgta_track=wgEncodeHudsonalphaChipSeq)

ENCODE data visualization, [http://genome.ucsc.edu/cgi-bin/hgTracks?hgS\\_doOtherUser=submit&hgS\\_otherUserName=Kate&hgS\\_otherUserSessionName=encodePortalSession](http://genome.ucsc.edu/cgi-bin/hgTracks?hgS_doOtherUser=submit&hgS_otherUserName=Kate&hgS_otherUserSessionName=encodePortalSession)

Software and resources for rnalysing ENCODE data, <http://genome.ucsc.edu/ENCODE/analysisTools.html>

Software tools used to create the ENCODE resource, <http://genome.ucsc.edu/ENCODE/encodeTools.html>

## 21.7 What Have We Learned?

This chapter provides an overview the ENCODE project which aims to annotate the entire human genome. It collects DNA sequences using various experimental techniques such as CHIP-seq, RNA-seq, and CAGE-seq. After the data has been obtained it needs to be processed before attempting analysis. The data goes through a number of steps; quality control, peak calling, IDR processing, and blacklist filtering. Once the accuracy of the data has been ensured other analysis can be done in the form of motif discovery, co-association analysis, and signal aggregation over elements.

there are extra figures in 'images' folder.

figure

## Bibliography

- [1] S G Landt, G K Marinov, A Kundaje, P Kheradpour, F Pauli, S Batzoglou, B E Bernstein, P Bickel, J B Brown, P Cayting, Y Chen, G DeSalvo, C Epstein, K I Fisher-Aylor, G Euskirchen, M Gerstein, J Gertz, A J Hartemink, M M Hoffman, V R Iyer, Y L Jung, S Karmakar, M Kellis, P V Kharchenko, Q Li, T Liu, X S Liu, L Ma, A Milosavljevic, R M Myers, P J Park, M J Pazin, M D Perry, D Raha, T E Reddy, J Rozowsky, N Shores, A Sidow, M Slattery, J A Stamatoyannopoulos, M Y Tolstorukov,

- K P White, S Xi, P J Farnham, J D Lieb, B J Wold, and M Snyder. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, 22(9):1813–1831, 2012.
- [2] Philippe Lefrançois, Ghia M Euskirchen, Raymond K Auerbach, Joel Rozowsky, Theodore Gibson, Christopher M Yellman, Mark Gerstein, and Michael Snyder. Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing. *BMC Genomics*, 10(1):37, 2009.
- [3] P J Park. ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics*, 10(10):669–680, October 2009.

---

CHAPTER  
**TWENTYTWO**

---

## SYNTHETIC BIOLOGY

add author

missing

### Figures

22.1 The layers of abstraction in robotics compared with those in biology (credit to Ron Weiss)	348
22.2 The repressilator genetic regulator network.	348
22.3 Fluorescence of a single cell with the repressilator circuit over a period of 10 hours.	349
22.4 Cost of synthesizing a base pair versus US dollar	349
22.5 An example of a BioCompiler program and the process of actualizing it (credit to Ron Weiss)	350
22.6 An example of combining BioBrick Pieces taken from <a href="http://2006.igem.org/wiki/index.php/Standard_Assembly">http://2006.igem.org/wiki/index.php/Standard_Assembly</a>	350

---

### 22.1 Introduction

A cell is like robot in that it needs to be able to sense its surroundings and internal state, perform computations and make judgments, and complete a task or function. The emerging discipline of synthetic biology aims to make control of biological entities such as cells and proteins similar to designing a robot. Synthetic biology combines technology, science, and engineering to construct biological devices and systems for useful purposes including solutions to world problems in health, energy, environment and, security.

Synthetic biology involves every level of biology, from DNA to tissues. Synthetic biologist aims to create layers of biological abstraction like those in digital computers in order to create biological circuits and programs efficiently. One of the major goals in synthetic biology is development of a standard and well-defined set of tools for building biological systems that allows the level of abstraction available to electrical engineers building complex circuits to be available to synthetic biologists.

Synthetic biology is a relatively new field. The size and complexity of synthetic genetic circuits has so far been small, on the order of six to eleven promoters. Synthetic genetic circuits remain small in total size ( $10^3$  -  $10^5$  base pairs) compared to size of the typical genome in a mammal or other animal ( $10^5$  -  $10^7$  base pairs) as well.

One of the first milestones in synthetic biology occurred in 2000 with the repressilator. The repressilator [2] is a synthetic genetic regulatory network which acts like an electrical oscillator system with fixed time periods. A green fluorescent protein was expressed within *E. coli* and the fluorescence was measured over time. Three genes in a feedback loop were set up so that each gene repressed the next gene in the loop and was repressed by the previous gene.

The repressilator managed to produce periodic fluctuations in fluorescence. It served as one of the first triumphs in synthetic biology. Other achievements in the past decade include programmed bacterial population control, programmed pattern formation, artificial cell-cell communication in yeast, logic gate creation by chemical complementation with transcription factors, and the complete synthesis, cloning, and assembly of a bacterial genome.

Figure 22.1: The layers of abstraction in robotics compared with those in biology (credit to Ron Weiss).

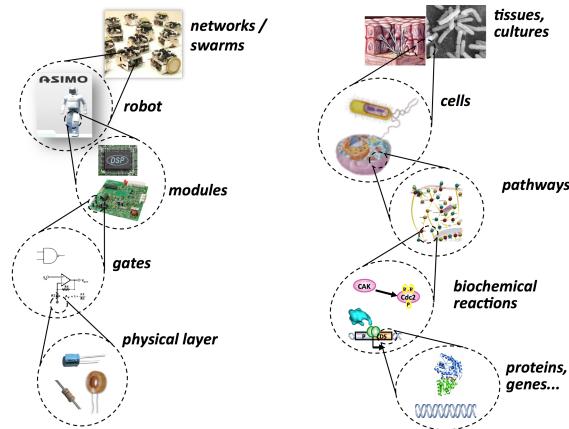
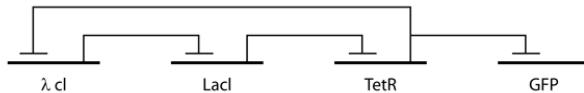


Figure 22.2: The repressilator genetic regulator network.



## 22.2 Current Research Directions

Encoding functionality in DNA is one way synthetic biologists program cells. As the price of sequencing and synthesis of DNA continues to decrease, coding DNA strands has become more feasible. In fact, the number of base pairs that can be synthesized per US\$ has increased exponentially, akin to Moore's Law.

This has made the process of designing, building, and testing biological circuits much faster and cheaper. One of the major research areas in synthetic biology is the creation of fast, automated synthesis of DNA molecules and the creation of cells with the desired DNA sequence. The goal of creating such a system is speeding up the design and debugging of making a biological system so that synthetic biological systems can be prototyped and tested in a quick, iterative process.

Synthetic biology also aims to develop abstract biological components that have standard and well-defined behavior like a part an electrical engineer might order from a catalogue. To accomplish this, the Registry of Standard Biological Parts (<http://partsregistry.org>) [4] was created in 2003 and currently contains over 7000 available parts for users. The research portion of creating such a registry includes the classification and description of biological parts. The goal is to find parts that have desirable characteristics such as:

**Orthogonality** Regulators should not interfere with each other. They should be independent.

**Composability** Regulators can be fused to give composite function.

**Connectivity** Regulators can be chained together to allow cascades and feedback.

**Homogeneity** Regulators should obey very similar physics. This allows for predictability and efficiency.

Synthetic biology is still developing, and research can still be done by people with little background in the field. The International Genetically Modified Machine (iGEM) Foundation (<http://igem.org>) [3] created the iGEM competition where undergraduate and high school students compete to design and build biological systems that operate within living cells. The student teams are given a kit of biological parts at the beginning of the summer and work at their own institutions to create biological system. Some interesting projects include:

**Arsenic Biodetector** The aim was to develop a bacterial biosensor that responds to a range of arsenic concentrations and produces a change in pH that can be calibrated in relation to arsenic concentration.

Figure 22.3: Fluorescence of a single cell with the repressilator circuit over a period of 10 hours.

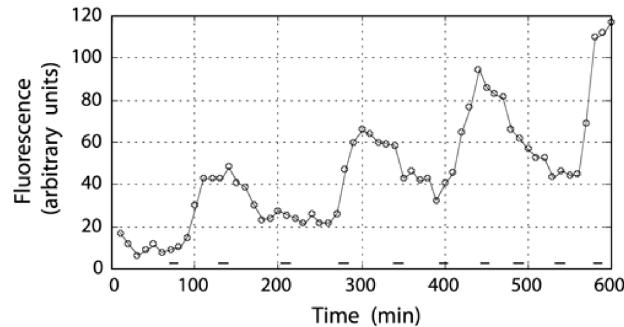
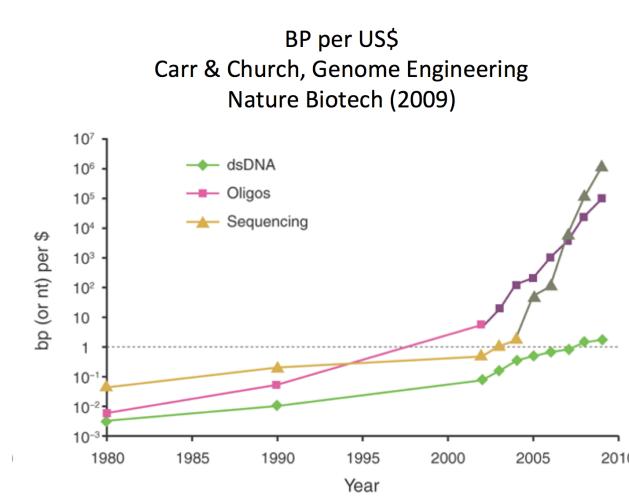


Figure 22.4: Cost of synthesizing a base pair versus US dollar



The team's goal was to help many under-developed countries, in particular Bangladesh, to detect arsenic contamination in water. The proposed device was intended be more economical, portable and easier to use in comparison with other detectors.

**BactoBlood** The UC Berkeley team worked to develop a cost-effective red blood cell substitute constructed from engineered *E. coli* bacteria. The system is designed to safely transport oxygen in the bloodstream without inducing sepsis, and to be stored for prolonged periods in a freeze-dried state.

**E. Chromi** The Cambridge team project strived to facilitate biosensor design and construction. They designed and characterised two types of parts - Sensitivity Tuners and Colour Generators – *E. coli* engineered to produce different pigments in response to different concentrations of an inducer. The availability of these parts revolutionized the path of future biosensor design.

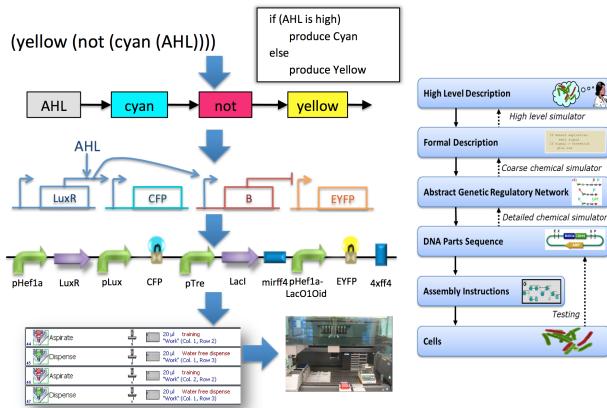
## 22.3 Further Reading

## 22.4 Tools and Techniques

Synthetic biology combines many fields, and the techniques used are not particular to synthetic biology. Much like the process of solving other engineering problems, the process of creating a useful biological system has designing, building, testing, and improving phases. Once a design or statement of the desired properties of a biological system are created, the problem becomes finding the proper biological components to build such a system.

BioCompiler [1] is a tool developed to allow the programming of biological circuits using a high-level programming language. One can write programs in a language similar to LISP and compile their program into a biological circuit. BioCompiler uses a process similar to that of a compiler for a programming language. It uses a human-written program as a high-level description of the genetic circuit, then generates a formal description of the program. From there, it looks up abstract genetic regulatory network pieces that can be combined to create the genetic circuit and goes through its library of DNA parts to find appropriate sequences to match the functionality of the abstract genetic regulatory network pieces. Assembly instructions can then be generated for creating cells with the appropriate genetic regulatory network.

Figure 22.5: An example of a BioCompiler program and the process of actualizing it (credit to Ron Weiss)



BioBrick standard biologic parts ([biobricks.org](http://biobricks.org)) are another tool used in synthetic biology. Similar to the parts in the Registry of Standard Biological Parts, BioBrick standard biological parts are DNA sequences of defined structure and function. Each BioBrick part is a DNA sequence held together in a circular plasmid. At either end of the BioBrick contains a known and well-defined sequence with restriction enzymes that can cut open the plasmid at known positions. This allows for the creation of larger BioBrick parts by chaining together smaller ones. Some competitors in the iGEM competition used BioBrick systems to develop an *E. coli* line that produced scents such as banana or mint.

Figure 22.6: An example of combining BioBrick Pieces taken from [http://2006.igem.org/wiki/index.php/Standard\\_Assembly](http://2006.igem.org/wiki/index.php/Standard_Assembly)



Figure 1: Dummy

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus com. Nulla et lectus vestibulum urna fringilla ultricies Phasellus

## 22.5 What Have We Learned?

Synthetic biology is an emerging disciplines that aims to create useful biological systems to solve problems in energy, medicine, environment, and many more fields. Synthetic biologists attempt to use abstraction to enable them to build more complex systems from simpler ones in a similar way to how a software engineer or an electrical engineer would make a computer program or a complex circuit. The Registry of Standard Biological Parts and BioBrick standard biological parts aim to characterize and standardize biological pieces just as one would a transistor or logic gate to enable abstraction. Tools such as BioCompiler allow people to describe a genetic circuit using a high-level language and actually build a genetic circuit with the described functionality. Synthetic biology is still new, and research can be done by those unfamiliar with the field, as demonstrated by the iGEM competition.

## Bibliography

- [1] J. Beal and J. Bachrach. Cells are plausible targets for high-level spatial languages, 2008.
- [2] M. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403:335–338, 2000.
- [3] iGEM. igem: Synthetic biology based on standard parts, December 2012.
- [4] Registry of Standard Biological Parts. Registry of standard biological parts, December 2012.



## **Part IV**

# **Phylogenomics and Population Genomics**



---

CHAPTER  
**TWENTYTHREE**

---

## MOLECULAR EVOLUTION AND PHYLOGENETICS

Scribed by Benjamin Bauchwitz (2014)  
Andrew Cooper, Stephanie Chang, and Stephen Serene (2012)  
Akashnil Dutta (2011)  
Albert Wang and Mashaal Sohail (2010)  
Guo-Liang Chew and Sara Baldwin (2009)

### Figures

---

23.1 Evolutionary History of Life . . . . .	357
23.2 Diagram of common phylogenetic tree terminology. . . . .	358
23.3 Three types of trees. . . . .	359
23.4 The two steps of distance based phylogenetic reconstruction. . . . .	361
23.5 Markov chain accounting for back mutations . . . . .	362
23.6 The y axis denotes probability of observing the bases - A(red), others(green). x axis denotes time. . . . .	362
23.7 Summary of the state transitions. $\alpha$ indicates the rate of substitution. . . . .	363
23.8 Fraction of altered bases (x axis) versus the Jukes-Cantor distance(y axis). Black line denotes the curve, green is the trend line for small values of f while the red line denotes the asymptotic boundary. . . . .	364
23.9 Distance models of varying levels of complexity(parameters). . . . .	365
23.10 Mapping from a tree to a distance matrix and vice versa . . . . .	366
23.11 Ultrametric distances. . . . .	367
23.12 Example of a minimum spanning tree generated from a graph . . . . .	367
23.13 Additive distances. . . . .	368
23.14 UPGMA / Hierarchical Clustering . . . . .	369
23.15 UPGMA fails to find the correct tree in this case . . . . .	369
23.16 Categories of different algorithms used for phylogenetics. . . . .	370
23.17 Since most of the real phylogenetic trees are far from being ultrametric, molecular clock is substantially dispersed. . . . .	371
23.18 Illustrates the difference between the molecular clock and the universal pacemaker model (UPM) [10]. . . . .	372
23.19 An overview of the character based methods . . . . .	373
23.20 Parsimony scoring: union and intersection . . . . .	374
23.21 Parsimony traceback to find ancestral nucleotides . . . . .	375
23.22 Parsimony scoring by dynamic programming . . . . .	375
23.23 A tree to be scored using the peeling algorithm. n=4 . . . . .	376
23.24 The recurrence . . . . .	376
23.25 An unit step using Nearest Neighbor Interchange scheme . . . . .	378

23.26Network representation of the forest of life, with NUTs in the middle . . . . .	380
23.27Forest generated from multiple trees . . . . .	380

---

## 23.1 Introduction

Phylogenetics tries to solve the general problem of inferring a complete ancestry of a set of **objects** based on the knowledge of their **traits**. These objects are anything where individual elements change gradually with time and diverge from a common origin, such as cell types, diseases, cancers, or even cars and languages. Most often, however, phylogenetics is used to study the relationships between different species and genes. Information on morphological traits as well as genetic data is used to answer questions about common ancestry and the evolution of genetic sequences.

Phylogenetics is particularly relevant in biology due to all species descending from a common ancestor that existed approximately 3.5 to 3.8 billion years ago. Genetic variation, isolation, and selection have created the great variety of species observed today. The ancestry between different species can shed light on different biological functions, such as genetic mechanisms and the process of evolution itself.

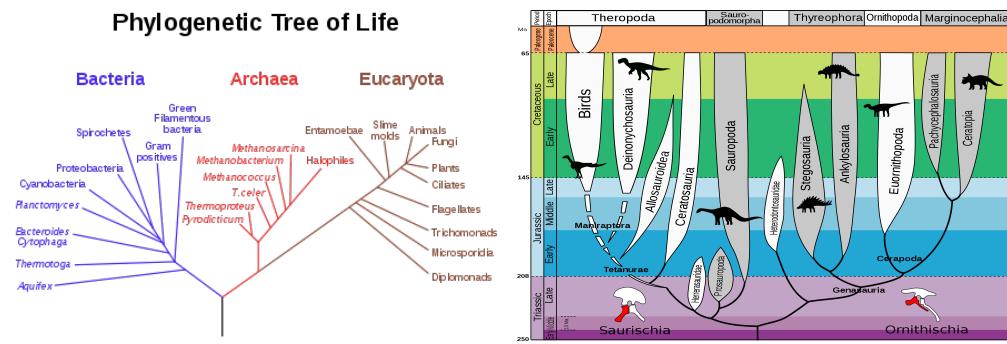


Figure 23.1: Evolutionary History of Life

## 23.2 Basics of Phylogeny

### 23.2.1 Traits

A trait is any characteristic that an object or species possesses. In the study of phylogenetics, these characteristics can be physical in nature, or **morphological data**, or they can genetic sequences or other **molecular data**. Examples of human morphological traits may be bipedalism (the ability to walk upright) or the opposable thumb, while an example of a human molecular trait is the specific genetic code that gives us color vision. In traditional phylogenetics the morphological features of different species were used to construct evolutionary trees. In modern methods, genetic sequence data or other molecular features are used instead.

**Morphological Data:** Arise from empirical evaluation of physical traits. This can be advantageous because physical characteristics are very easy to quantify and understand. Furthermore, unlike molecular data, morphological data is available for ancestral samples in the fossil record. However, morphological characters can be hard to classify. Unlike DNA, which has only four states, morphological traits can have a much larger range of states (coat color, bone structure, etc.). Quantification of these states can be extremely subjective and can affect the phylogenetic output. Convergent evolution of morphological characters can also mask true species relationships. Species that diverged millions of years ago may converge again on the few traits that are observable to scientists, giving a false representation of how closely related the species are.

**Molecular Data:** Discovered by studying the genomes or other molecular features of different species. This approach can be advantageous because it allows scientists to study the biological mechanism underlying each observable morphological trait. A sequence based approach also removes the subjectivity associated with morphological data. The biggest downside of using genomes in particular is that DNA is only built from 4 bases, so back mutations are frequent. Scientists must reconcile the signals of a large number of

ill-behaved traits instead of those of the small number of mostly well-behaved traits of morphological features.

Since sequence phylogenetics depends on the comparison between pairs of genes, it is useful to understand concepts related to **homology**, or shared ancestry between different genes.

**Paralogs:** Genes can diverge in a duplication event within the same individual. Evolution can then lead to these two genes acquiring separate functions but coexisting within members of the same species.

**Orthologs:** Speciation results in a copy of a gene in different species. These copies then evolve separately in their separate species, resulting in distinct genes across species. Orthologs generally retain their function across species.

## FAQ

**Q:** Would it be possible to use extinct species' DNA sequences?

**A:** Although some work has been done with the DNA of frozen woolly mammoths, most of the work with ancient DNA has been focused on ancient human and Neanderthal populations. It is becoming increasingly common to extract DNA from preserved human remains to infer major selective events and migrations in the early history of humans. See the following paper for a recent example:

Skoglund P, Mallick S, Bortolini MC, Chennagiri N, H?nemeier T, Petzl-Erler ML, Salzano FM, Patterson N, Reich D (2015) Genetic evidence for two founding populations of the Americas. Nature 525, 104-8. (Supplementary Materials)

### 23.2.2 Trees

A tree is a mathematical representation of relationships between objects. A general tree is built from nodes and edges. Each node represents an object, and each edge represents a relationship between two nodes. In phylogenetics, we can use trees to represent evolution. The common ancestor is represented by the root node of the tree, the terminal or leaf nodes represent current species, and the branches and nodes between the root and the leaves represent the lineages and hypothetical intermediaries.



Figure 1: Dummy

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat se, adipiscing vitae, felis. Curabitur dictum gravidae musris. Nam arcu libero, nonummy eget, consetetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus egestas augue. Nunc ac tortor condidit enim fermentum ultricies. Phasellus

Figure 23.2: Diagram of common phylogenetic tree terminology.

Biological phylogenetic trees may be classified into three categories, based on the information conveyed by their edge lengths:

**Cladogram:** This type of graph only provides information about the relationships of the samples; as the lengths of the branches have no meaning, nothing besides the topology of the branching matters.

**Phylogram:** Phylogenograms contain information about topology, divergence times *and* divergence rates. Branch lengths are directly related to the amount of **genetic change**; the longer the branch of a tree, the greater the amount of phylogenetic change that has taken place. The leaves in this tree may not necessarily end on the same vertical line, due to different rates of mutation.

**Chronogram (ultrametric tree):** Chronograms can provide information about topology and the divergence times. Branch lengths are directly related to **time**, which indicates that the longer the branches of a tree, the greater the amount of time that has passed. The leaves in this tree necessarily end on the same vertical line (i.e. they are the same distance from the root), since they are all in the present unless extinct species were included in the tree. Although there is a correlation between branch lengths and genetic distance on a chronogram, they are not necessarily exactly proportional because evolution rates / mutation rates are not constant. Some species evolve and mutate faster than others, and some historical time periods foster faster rates of evolution than others.

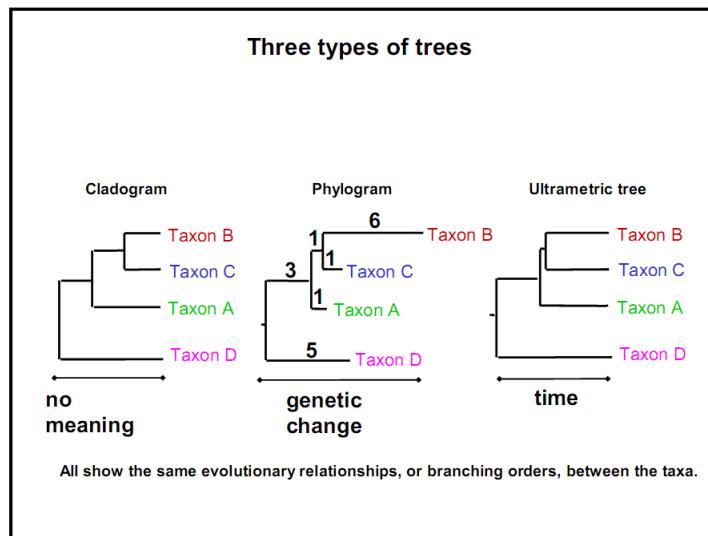


Figure 23.3: Three types of trees.

### 23.2.3 Methods for Tree Reconstruction

Once we have found genetic data for a set of species, we are interested in learning how these species relate to each other. Since we can, for the most part, only obtain DNA from living creatures, we must infer the ancestral sequence implied by each node and ultimately that of the common ancestor. The following sections will explore the modern methods for inferring ancestry from sequence data. They can be classified into two approaches, distance-based methods and character-based methods.

**Distance-based approaches** quantify the amount of mutation that separates each pair of sequences (which may or may not be proportional to the time since they have been separated) and fits the most likely tree according to the pair-wise distance matrix. The second step is usually a direct algorithm, based on some assumptions, but may be more complex. This approach uses a number to represent the relationship between two sequences.

**Character-based approaches** instead try to find the tree that best explains the observed sequences. As opposed to direct reconstruction, these methods rely on tree proposal and scoring techniques to perform a heuristic search over the space of trees. This approach uses alignments to represent the relationship between two sequences.

***Did You Know?***

**Occam's Razor**, as discussed in previous chapters, does not always provide the most accurate hypothesis. In many cases during tree reconstruction, the simplest explanation is not the most probable. For example, a set of possible ancestries may be possible, given some observed data. In this case, the simplest ancestry may not be correct if a trait arose independently in two separate lineages. This issue will be considered in a later section.

## 23.3 Distance Based Methods

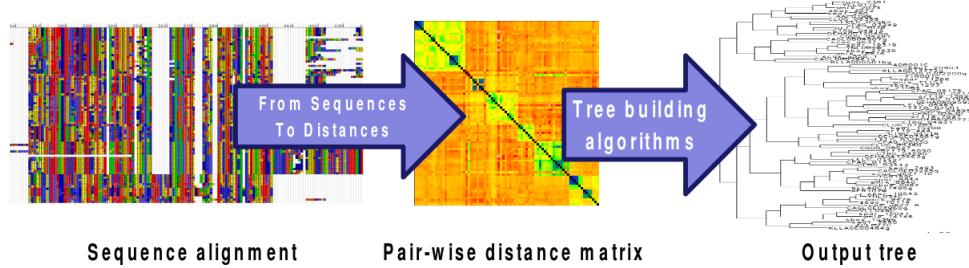


Figure 23.4: The two steps of distance based phylogenetic reconstruction.

The first approach to determining the evolutionary history of a species or sequence is to use distance-based methods. In general, the two steps to this approach are:

1. Map each pair of sequences to a distance separating those sequences.
2. Use those distances to build an optimal tree.

Though some information is lost by producing these pairwise distances, this approach sets up the platform for direct tree reconstruction.

### 23.3.1 From alignment to distances

In order to understand how a distance-based model works, it is important to think about what distance means when comparing two sequences. The most basic way is to measure **nucleotide divergence** (also called pair wise distance) by aligning two sequences and counting nucleotide sites that don't match. This gives equal weight to all changes, and assumes that evolution happens at a uniform rate across the genome. Because these assumptions only loosely hold, the method can be unreliable at times. However, this procedure is useful because nucleotide divergence can be easily assessed and requires little inference. Rather than assuming rates are uniform, we can augment our distance metric to include known information about mutation rates:

**Transitions and Transversions (see Kimura model)** Due to structural similarities, one might expect mutations between purines (A to G, G to A) and between pyrimidines (G to C, C to G), known as transitions, to be more likely than transversions, or mutations between a purine nucleotide to a pyrimidine nucleotide. Therefore, it would make sense to have two parameters, one for each type of mutation.

**Synonymous and non-synonymous substitutions** This method keeps tracks of substitutions that affect the coded amino-acid. It assumes that substitutions that do not change the coded protein will not be selected against, and will thus have a higher probability of occurring than those substitutions which do change the coded amino acid. Under this model, you can expect to see higher substitution rates in the 3rd codon position than the 1st and 2nd codon positions.

The naive way to interpret the separation between two sequences may be simply the number of mismatches, as described by nucleotide divergence above. While this does provide us a distance metric (i.e.  $d(a, b) + d(b, c) \geq d(a, c)$ ) this does not quite satisfy our requirements, because we want **additive distances**, i.e. those that satisfy  $d(a, b) + d(b, c) = d(a, c)$  for a path  $a \rightarrow b \rightarrow c$  of evolving sequence. This is because the number of mutations accumulated along a path in the tree should be the sum of that of its individual components. This may not always be the case in evolution, however. Take the example in which sequence A evolves into sequence B, which evolves into sequence C. It may be that sequence B is a version of sequence A in which 30 percent of the bases have mutated, and sequence C is a version of sequence B in which 30 percent of the bases have mutated, but C is only 50 percent different from A (as opposed to 60 percent). It is also obvious to see that any sequence can diverge by 50 percent or more, while no decedent can ever exceed a 100 percent mutation rate from its earliest ancestor.

The reason for this phenomenon is **back-mutations**. When a large number of mutations accumulate on a sequence, not all the mutations introduce new mismatches - some of them may occur on already mutated base pair, resulting in the mismatch score remaining the same or even decreasing. In other words, the number of actual mutations are always greater than the number of observed substitutions. For small mismatch-scores, however, this effect is statistically insignificant, because there are vastly more identical pairs than mismatching pairs. However, for sequences separated by longer evolutionary distance, back-mutations must be accounted for. The Jukes-Cantor model is one such simple Markov model that incorporates back-mutations.

### Jukes-Cantor distances

To illustrate this concept, consider a nucleotide in state 'A' at time zero. At each time step, it has a probability 0.7 of retaining its previous state and probability 0.1 of transitioning to each of the other three states. The probability  $P(B|t)$  of observing state (base)  $B$  at time  $t$  essentially follows the recursion

$$P(B|t+1) = 0.7P(B|t) + 0.1 \sum_{b \neq B} P(b|t) = 0.1 + 0.6P(B|t)$$

Figure 23.5: Markov chain accounting for back mutations

If we plot  $P(B|t)$  versus  $t$ , we observe that the distribution starts off as concentrated at the state 'A' and gradually spreads over to the rest of the states, eventually going towards an equilibrium of equal probabilities. This progression makes sense, intuitively. Over millions of years, species can evolve so dramatically that they no longer resemble their ancestors. At that extreme, a given base location in the ancestor is just as likely to have evolved to any of the four possible bases in that location over time.

time:-	0	1	2	3	4
A	1	0.7	0.52	0.412	0.3472
C	0	0.1	0.16	0.196	0.2196
G	0	0.1	0.16	0.196	0.2196
T	0	0.1	0.16	0.196	0.2196

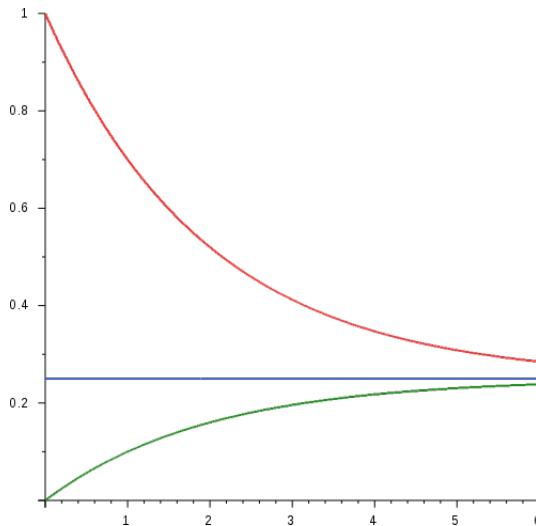


Figure 23.6: The y axis denotes probability of observing the bases - A(red), others(green). x axis denotes time.

The essence of the Jukes-Cantor model is to backtrack  $t$ , the amount of time elapsed from the fraction of altered bases. Conceptually, this is just inverting the x and y axis of the green curve. To model this quantitatively, we consider the following matrix  $S(t)$  which denotes the respective probabilities  $P(x|y, \Delta t)$  of observing base  $x$  given a starting state of base  $y$  and change in time  $\Delta t$ .

$$S(\Delta t) = \begin{pmatrix} P(A|A, \Delta t) & P(A|G, \Delta t) & \cdots & P(A|T, \Delta t) \\ P(G|A, \Delta t) & \cdots & & \cdots \\ \cdots & & & \cdots \\ P(T|A, \Delta t) & \cdots & \cdots & P(T|T\Delta t) \end{pmatrix}$$

We can assume this is a stationary markov model, implying this matrix is multiplicative, i.e.

$$S(t_1 + t') = S(t_1)S(t')$$

which implies that

$$P(x|y, t + t') = \sum_z P(x|z, t)P(z|y, t')$$

For a very short time  $\epsilon$ , we can assume that there is no second order effect, i.e. there isn't enough time for two mutations to occur at the same nucleotide. So the probabilities of cross transitions are all proportional to  $\epsilon$ . Further, in Jukes-Cantor model, we assume that all the transition rates are same from each nucleotide to another nucleotide. Hence, for a short time  $\epsilon$

$$S(\epsilon) = \begin{pmatrix} 1 - 3\alpha\epsilon & \alpha\epsilon & \alpha\epsilon & \alpha\epsilon \\ \alpha\epsilon & 1 - 3\alpha\epsilon & \alpha\epsilon & \alpha\epsilon \\ \alpha\epsilon & \alpha\epsilon & 1 - 3\alpha\epsilon & \alpha\epsilon \\ \alpha\epsilon & \alpha\epsilon & \alpha\epsilon & 1 - 3\alpha\epsilon \end{pmatrix}$$

At a longer time  $t$ , the matrix is given by

$$S(t) = \begin{pmatrix} r(t) & s(t) & s(t) & s(t) \\ s(t) & r(t) & s(t) & s(t) \\ s(t) & s(t) & r(t) & s(t) \\ s(t) & s(t) & s(t) & r(t) \end{pmatrix}$$

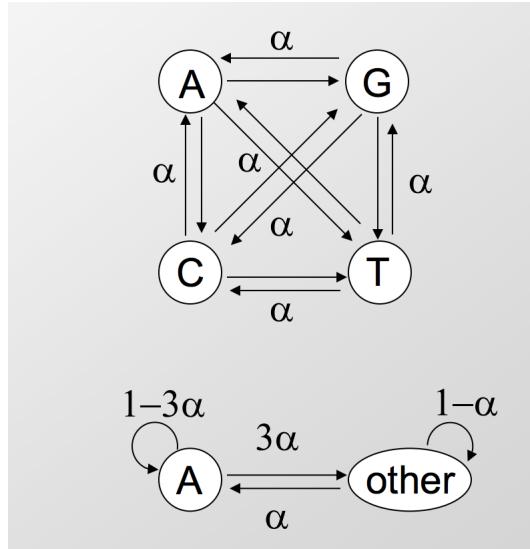


Figure 23.7: Summary of the state transitions.  $\alpha$  indicates the rate of substitution.

From the equation  $S(t + \epsilon) = S(t)S(\epsilon)$  we obtain

$$r(t + \epsilon) = r(t)(1 - 3\alpha\epsilon) + 3\alpha\epsilon s(t) \text{ and } s(t + \epsilon) = s(t)(1 - \alpha\epsilon) + \alpha\epsilon r(t))$$

Which rearrange as the coupled system of differential equations

$$r'(t) = 3\alpha(-r(t) + s(t)) \text{ and } s'(t) = \alpha(r(t) - s(t))$$

With the initial conditions  $r(0) = 1$  and  $s(0) = 0$ . The solutions can be obtained as

$$r(t) = \frac{1}{4}(1 + 3e^{-4\alpha t}) \text{ and } s(t) = \frac{1}{4}(1 - e^{-4\alpha t})$$

Now, in a given alignment, if we have the fraction  $f$  of the sites where the bases differ, we have:

$$f = 3s(t) = \frac{3}{4}(1 - e^{-4\alpha t})$$

implying

$$t \propto -\log\left(1 - \frac{4f}{3}\right)$$

To agree asymptotically with  $f$ , we set the evolutionary distance  $d$  to be

$$d = -\frac{3}{4} \log\left(1 - \frac{4f}{3}\right)$$

Note that distance is approximately proportional to  $f$  for small values of  $f$  and asymptotically approaches infinity when  $f \rightarrow 0.75$ . Intuitively this happens because after a very long period of time, we would expect the sequence to be completely random and that would imply about three-fourth of the bases mismatching with original. But the uncertainty values of the Jukes-Cantor distance also becomes very large when  $f$  approaches 0.75.

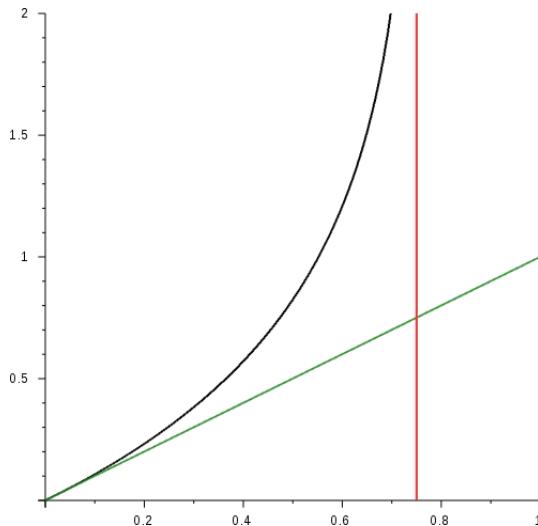


Figure 23.8: Fraction of altered bases (x axis) versus the Jukes-Cantor distance(y axis).  
Black line denotes the curve, green is the trend line for small values of  $f$  while the red line denotes the asymptotic boundary.

### Kimura's model

The Jukes-Cantor model is the simplest model that gives us a theoretically consistent additive distance model. However, it is a one-parameter model that assumes that mutations from each base to every other base occur with equal probability. In reality, **transitions**, which are substitutions between pairs of purines or between pairs of pyrimidines, are more common than **transversions**, which are substitutions between one purine and one pyrimidine.

Kimura's two parameter model is one of the simpler models that take into account the different rates of transitions (rate  $\alpha$ ) and transversion (rate  $\beta$ ).

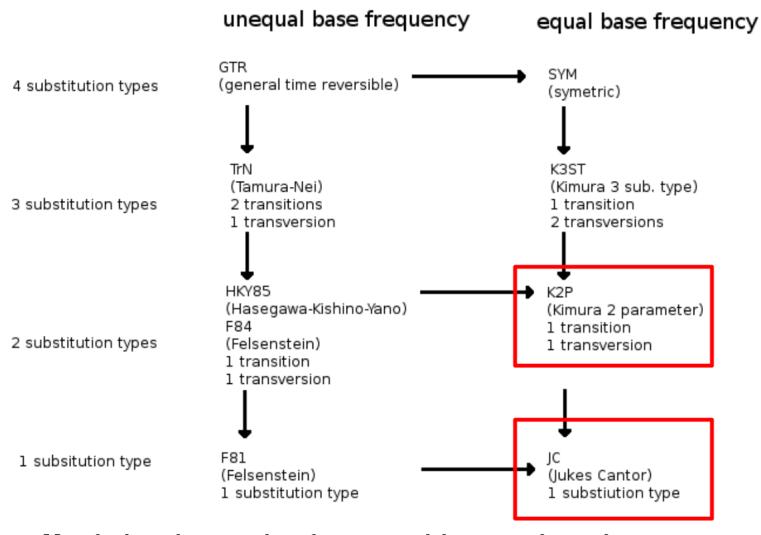
$$S(t) = \begin{pmatrix} r(t) & s(t) & u(t) & u(t) \\ s(t) & r(t) & u(t) & u(t) \\ u(t) & u(t) & r(t) & s(t) \\ u(t) & u(t) & s(t) & r(t) \end{pmatrix}$$

Following the same steps of reasoning as the Jukes-Cantor model, we get

$$\begin{aligned} s(t) &= 0.25(1 - e^{-4\beta t}) \\ u(t) &= 0.25(1 + e^{-4\beta t} - e^{-2(\alpha+\beta)t}) \\ r(t) &= 1 - 2s(t) - u(t) \end{aligned}$$

## Other Models

Other models with more complex parameters are depicted below.



**Models also exist for peptides and codons**

Figure 23.9: Distance models of varying levels of complexity(parameters).

## FAQ

**Q:** Can we use different parameters for different parts of the tree? To account for different mutation rates?

**A:** It's possible, it is a current area of research.

Other parameters are often added to these models. Where a typical model assumes that rates are evenly distributed across sites, adding the parameter **G**, referring to a gamma distribution, models heterogeneity in mutations rates. Another common parameter **I**, can be added to account for some proportion of sites that can be expected to be completely invariant.

### 23.3.2 Distances to Trees

If we have a weighted phylogenetic tree, we can find the total weight (length) of the shortest path between a pair of leaves by summing up the individual branch lengths in the path. Considering all such pairs of leaves, we have a distance matrix representing the data. In distance based methods, the problem is to reconstruct the tree given this distance matrix, while minimizing the discrepancy between the observed distances and the tree-based distances ( $\min \sum_{ij} (D_{ij} - M_{ij})^2$ ).

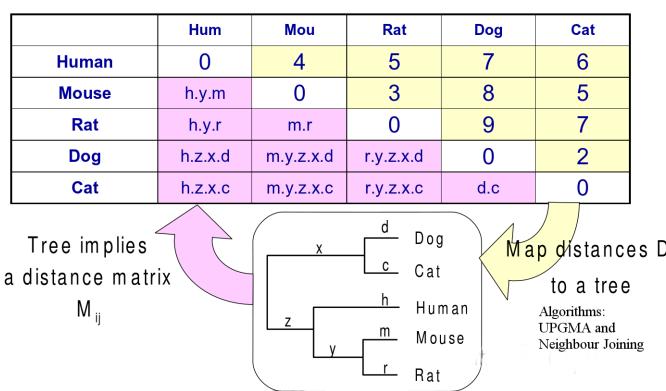


Figure 23.10: Mapping from a tree to a distance matrix and vice versa

## FAQ

**Q:** In Figure 23.10 The m and r sequence divergence metrics can have some overlap so distance between mouse and rat is not simply m+r. Wouldn't that only be the case if there was no overlap?

**A:** If you model evolution correctly, then you would get evolutionary distance. It's an inequality rather than an equality and we agree that you can't exactly infer that the given distance is the precise distance. Therefore, the sequences' distance between mouse and rat is probably less than m + r because of overlap, convergent evolution, and transversions.

However, note that there is not a one-to-one correspondence between a distance matrix and a weighted tree. Each tree does correspond to one distance matrix, but the opposite is not always true. A distance matrix has to satisfy additional properties in order to correspond to some weighted tree. In fact, there are two models that assume special constraints on the distance matrix.

### Ultrametric distances

**3-point condition:** For all triplets  $(a, b, c)$  of leaves, two pairs among them have equal distance, and the third distance is smaller; i.e. the triplet can be labelled  $i, j, k$  such that

$$d_{ij} \leq d_{ik} = d_{jk}$$

Conceptually this is because the two leaves that are more closely related (say  $i, j$ ) have diverged from the third ( $k$ ) at exactly the same time. And the time separation from the third should be equal, whereas the separation between themselves should be smaller. This condition results in

1. All paths from leaves are equidistant to the root
2. The rooted tree has uniform rates of evolution

which makes the interpretation of the tree much simpler - the distances will fit perfectly onto a tree.

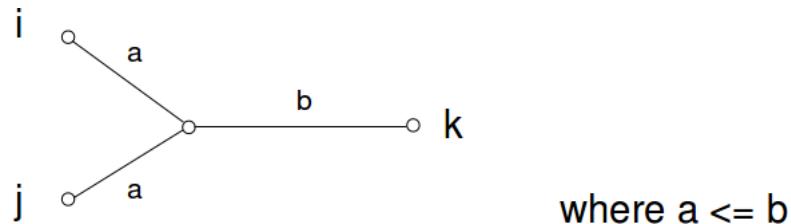


Figure 23.11: Ultrametric distances.

**‘Ultrametric trees:** Ultrametric trees are represented by symmetric 0-diagonal matrices, where the indices represent the divergence times. It makes intuitive sense when you imagine that the indices are just the number of substitutions between the pairs. Given a symmetric  $n \times n$  0-diagonal matrix  $D$ , an ultrametric tree  $T$  for that matrix is one in which:

1. There are  $n$  leaves, one for each row and column of  $D$ .
2. Each internal nodes is labeled by a time in  $D$  and has exactly two children.
3. Along any path from the root to a leaf, the divergence times at the internal nodes strictly decrease.
4. For any two leaves  $i, j$  of  $T$ , the least common ancestor of  $i, j$  is labeled with time  $D(i, j)$ .

#### ‘Ultrametrification’ of non-ultrametric trees:

If a tree does not satisfy ultrametric conditions, we can attempt to find a set of alterations to an  $n \times n$  symmetric distance matrix that will make it ultrametric. This can be accomplished by constructing a completely connected graph with weights given by the original distance matrix, finding a minimum spanning tree (MST) of this graph, and then building a new distance matrix with elements  $D(i,j)$  given by the *largest* weight on the unique path in the MST from  $i$  to  $j$ .

A spanning tree of the fully connected graph simply identifies a subset of edges that connects all nodes without creating any cycles, and a minimum spanning tree is a spanning tree that minimizes the total sum of edge weights. It should be noted that there is a *unique* path between any two vertices in a spanning tree. An MST can be found using methods such as Prim’s algorithm(graph traversal), and then be used to “correct” a non-ultrametric tree.

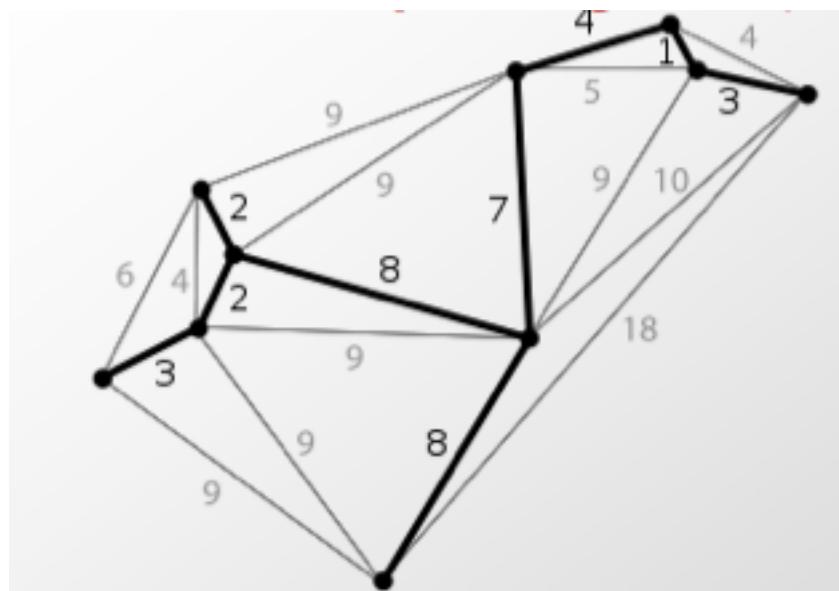


Figure 23.12: Example of a minimum spanning tree generated from a graph

### Additive distances

**4-point condition:** Additive distance matrices satisfy the property that all quartets of leaves can be labelled  $i, j, k, l$  such that

$$d_{ij} + d_{kl} \leq d_{ik} + d_{jl} = d_{il} + d_{jk}$$

This is in fact true for all positive-weight trees. For any 4 leaves in a tree, there can be exactly one topology, i.e.

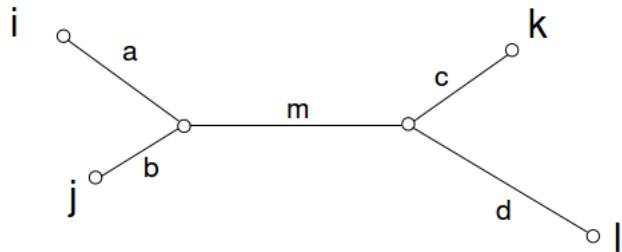


Figure 23.13: Additive distances.

Then the above condition is term by term equivalent to

$$(a + b) + (c + d) \leq (a + m + c) + (b + m + d) = (a + m + d) + (b + m + c)$$

. This equality corresponds to all pairwise distances that are possible from traversing this tree. Additive distances, unlike ultrameric distances, do not require that all leaves be equidistant from common ancestors.

These types of redundant equalities must occur while mapping a tree to a distance matrix, because a tree of  $n$  nodes has  $n - 1$  parameters, one for each branch length, while a distance matrix has  $n^2$  parameters. Hence, a tree is essentially a lower dimensional projection of the higher dimensional matrix. A corollary of this observation is that not all distance matrices have a corresponding tree, but all trees map to a unique distance matrix.

### Limitations of ultrametric and additive distances

However, real datasets do not exactly satisfy either ultrameric or additive constraints. This can be due to noise (when our parameters for our evolutionary models are not precise), stochasticity and randomness (due to small samples), fluctuations, different rates of mutations, gene conversions, and horizontal transfer. Because of this, we need tree-building algorithms that are able to handle noisy distance matrices. A couple possibilities are: enumeration and scoring of all trees, which is too expensive, UPGMA, which typically gives a poor tree, and neighbor-joining, which typically gives a good tree.

### UPGMA - Unweighted Pair Group Method with Arithmetic Mean

This is exactly same as the method of **Hierarchical clustering** discussed in Lecture 13, Gene Expression Clustering. It forms clusters step by step, from closely related nodes to ones that are further separated. A branching node is formed for each successive level. The algorithm can be described properly by the following steps:

#### Initialization:

1. Define one leaf  $i$  per sequence  $x_i$ .
2. Place each leaf  $i$  at height 0.
3. Define Clusters  $C_i$  each having one leaf  $i$ .

#### Iteration:

1. Find the pairwise distances  $d_{ij}$  between each pairs of clusters  $C_i, C_j$  by taking the arithmetic mean of the distances between their member sequences.
2. Find two clusters  $C_i, C_j$  such that  $d_{ij}$  is minimized.
3. Let  $C_k = C_i \cup C_j$ .
4. Define node  $k$  as parent of nodes  $i, j$  and place it at height  $d_{ij}/2$  above  $i, j$ .
5. Delete  $C_i, C_j$ .

**Termination:** When two clusters  $C_i, C_j$  remain, place the root at height  $d_{ij}/2$  as parent of the nodes  $i, j$

#### Weaknesses of UPGMA

Although this method is guaranteed to find the correct tree if the distance matrix obeys the ultrameric property, it turns out to be an inaccurate algorithm in practice. Apart from lack of robustness, it suffers from the molecular clock assumption that the mutation rate over time is constant for all species. This is certainly not true as species such as rat and mice evolve much more quickly than others. Such differences in mutation rate can lead to long branch attraction; nodes sharing a lower mutation rate but found in distinct lineages may be merged, leaving those nodes with higher mutation rates (long branches) to appear together in the tree. The following figure illustrates an example where UPGMA fails:

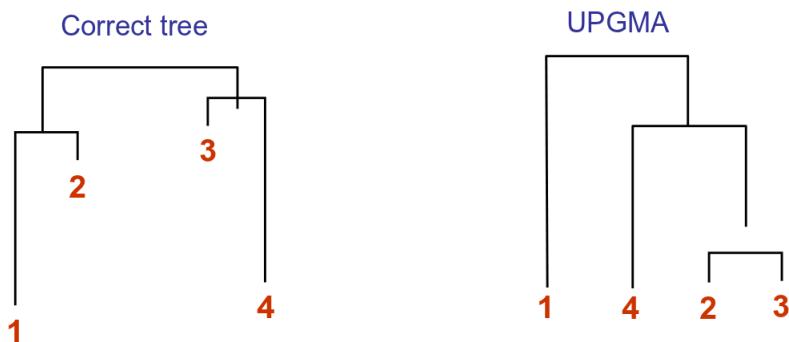


Figure 23.15: UPGMA fails to find the correct tree in this case

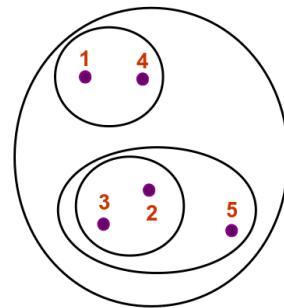


Figure 23.14: UPGMA / Hierarchical Clustering

### Neighbor Joining

The neighbor joining method is guaranteed to produce the correct tree if the distance matrix satisfies the additive property. It may also produce a good tree when there is some noise in the data, as it tries to rescale distances to correct for long branch attraction. The algorithm is described below:

**Finding the neighboring leaves:** Let

$$D_{ij} = d_{ij} - (r_i + r_j) \text{ where } r_a = \frac{1}{n-2} \sum_k d_{ak}, a \in \{i, j\}$$

Here  $n$  is the number of nodes in the tree; hence,  $r_i$  is the average distance of a node to the other nodes. It can be proven that the above modification ensures that  $D_{ij}$  is minimal only if  $i, j$  are neighbors. (A proof can be found in page 189 of Durbin's book).

**Initialization:** Define  $T$  to be the set of leaf nodes, one per sequence. Let  $L = T$

**Iteration:**

1. Pick  $i, j$  such that  $D_{ij}$  is minimized.
2. Define a new node  $k$ , and set  $d_{km} = \frac{1}{2}(d_{im} + d_{jm} - d_{ij}) \forall m \in L$
3. Add  $k$  to  $T$ , with edges of lengths  $d_{ik} = \frac{1}{2}(d_{ij} + r_i - r_j)$
4. Remove  $i, j$  from  $L$
5. Add  $k$  to  $L$

**Termination:** When  $L$  consists of two nodes  $i, j$ , and the edge between them of length  $d_{ij}$ , add the root node as parent of  $i$  and  $j$ .

### Distance-Fitting Algorithms

Besides making a tree and clustering them using algorithms such as UPGMA and Neighbor-Joining, we can also aim to directly minimize discrepancy between original distance matrix and tree-based distance matrix with standard distance-fitting algorithms such as least squares (minimize the sum of squares of differences).

		COMPUTATIONAL METHOD	
		Optimality criterion Clustering algorithm	
DATA TYPE	Characters	PARSIMONY	
		MAXIMUM LIKELIHOOD	
Distances	MINIMUM EVOLUTION		UPGMA
	LEAST SQUARES		NEIGHBOR-JOINING

Figure 23.16: Categories of different algorithms used for phylogenetics.

### 23.3.3 Summary of Distance Methods Pros and Cons

The methods described above have been shown to capture many interesting features of phylogenetic relationships, and are typically very fast in the algorithmic sense. However, there is some information loss when transferring between distance matrices and trees, and typically only one of many possible trees is proposed. Long branch attraction, in which convergent evolution of sequences makes species seem more related than they really are, is one example of a serious error that can occur when basic assumptions about mutation rate are violated. Finally, distance methods make no inference about the history of a particular site, and thus do not make suggestions about the ancestral state of a sequence.

## 23.4 Molecular Clocks and themakers

The molecular clock model, as proposed by Zuckerkandl & Pauling in 1962, assumes that different genes evolve at specific, roughly constant rates. This implies that divergence between orthologous sequences is proportional to time separating the species. Under this model, all individual gene trees are ultrametric (up to a sampling error), and identical to the species tree (up to a scaling factor/evolution rate).

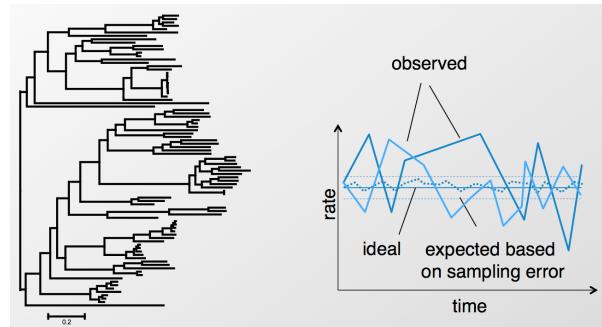


Figure 23.17: Since most of the real phylogenetic trees are far from being ultrametric, molecular clock is substantially dispersed.

However, as we have discussed earlier, most of the real phylogenetic trees are not ultrametric, and the observed molecular clock rates are very much overdispersed. This led to the rise of relaxed molecular clock models, which allow for rate variation. These rates are sampled from prior distributions with limited variance, independently or in an autocorrelated manner. And these deviations are expected to be uncorrelated. Genes are either analyzed individually, or as concatenated alignments implying modular units that evolve together.

On the other hand, the universal pacemaker model (UPM) assumes that evolutionary time runs at a different pace in each lineage. Under this model, species trees are intrinsically non-ultrametric, so all deviations are expected to be correlated. It means

that although changes in the rate of evolution might change arbitrarily, they change in the same direction altogether. Several studies have been done to test and compare these models. For example, in one study conducted by [10], phylogenetic trees of more than two thousand gene families from 41 archeal and 59 bacterial genomes are used to create 'supertrees' that best summarize the overall evolutionary direction. Then tested how well the molecular clock, relaxed molecular clock and the UPM fit, and it turned out that UPM was significantly better than the others. Furthermore, they concluded that the UPM sufficiently explains the observed correlations between the evolutionary rates among the genes of a genome.

These results attract our attention to the biologically relevant implications of the UPM. Investigating the underlying factors that determine this pace, or these correlated shifts in the rate of evolutions among various lineages, is likely to be an active research area in the coming years.

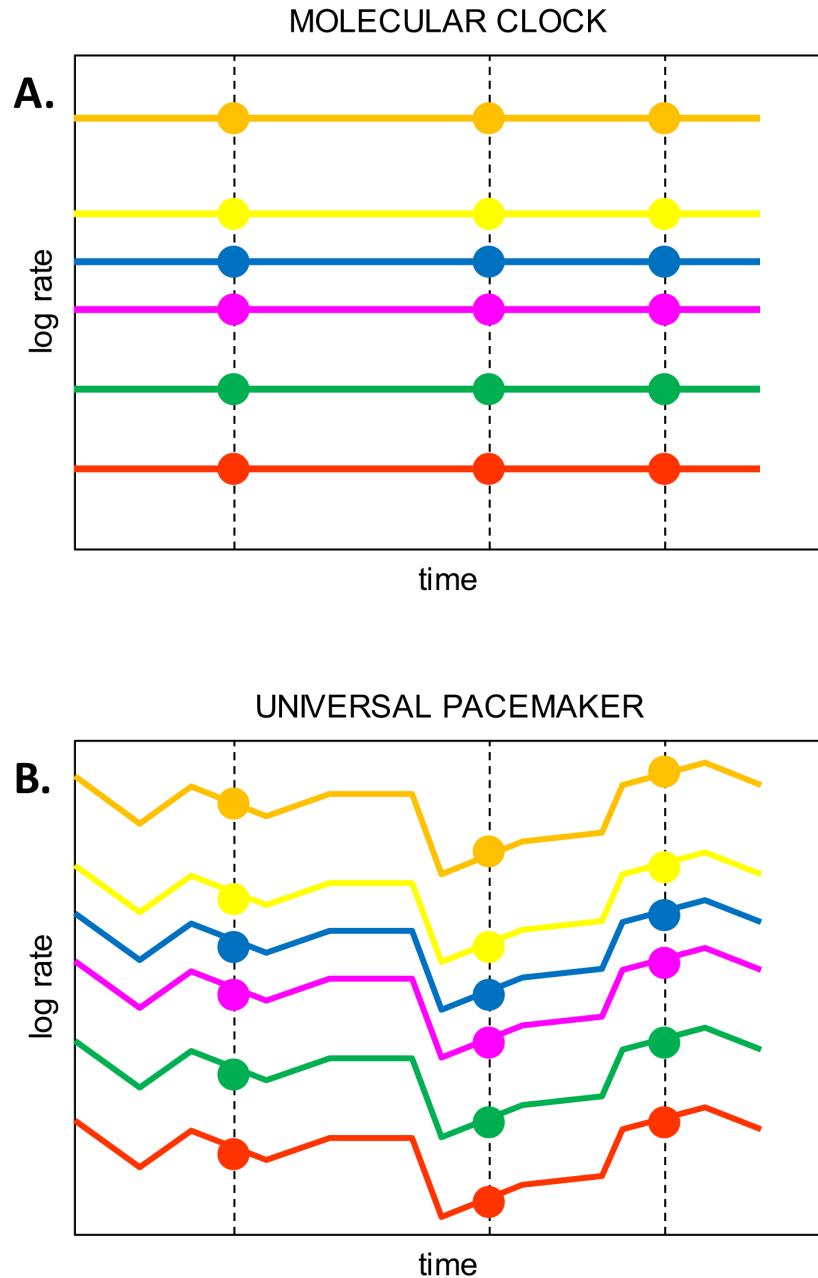


Figure 23.18: Illustrates the difference between the molecular clock and the universal pacemaker model (UPM) [10].

## 23.5 Character-Based Methods

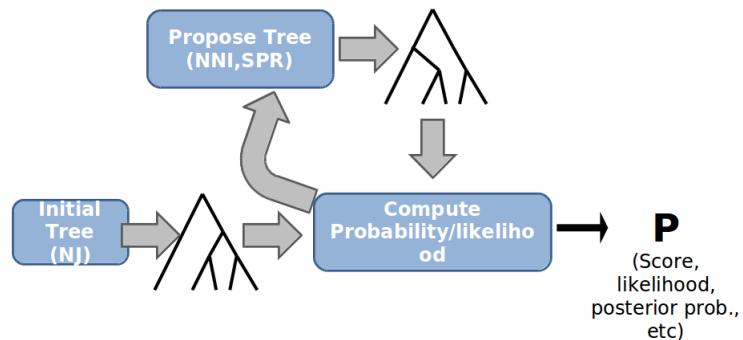


Figure 23.19: An overview of the character based methods

While distance-based methods sought to generate a tree from a series of observations, character-based methods seek to take this information and find an optimal preexisting tree. This approach has two main steps. First, a procedure is developed that will score a tree based on the probability that it would produce the observed sequence at its leaves. Then, the entire space of trees is searched for one that maximizes this probability. Though this problem is theoretically NP-Hard due to the large number of possible trees, heuristic methods are frequently able to find trees that are a good fit. In the following section we will discuss first tree scoring algorithms, and then search techniques.

### 23.5.1 Scoring

There are two main algorithms for tree scoring. The first approach, which we will call parsimony reconstruction, is based on Occam's razor, and scores a topology based on the minimum number of mutations it implies, given the (known) sequences at the leaves. This method is simple, intuitive, and fast. The second approach is a maximum likelihood method which scores trees by explicitly modeling the probability of observing the sequences at the leaves given a tree topology.

#### Parsimony

Conceptually, this method is simple. It simply assigns a state(s) at each ancestral node (in the case of DNA sequences, nucleotide(s)) such that the number of substitutions is minimized. The score is then just the sum over all state changes in the topology.

This is performed by first reconstructing the ancestral sequences at all internal nodes on the tree. The algorithm begins by scanning up from the (known) leaf sequences, assigning a set of bases at each internal node based on what bases are found in the children. In the case of Figure 18, the two taxa on the left have states A and B. As a result, their ancestor could have had either character A or B. For example, if the ancestor had A, then a mutation occurred, post divergence, in the right branch to state B. Next, the algorithm iterates down the tree, picking bases out of the allowed sets at each node, this time based on the states at the node's parents. The following illustrates this algorithm in detail (note that there are  $2N - 1$  total nodes, indexed from the root, such that the known leaf nodes have indices  $N - 1$  through  $2N - 1$ ):

Given a tree, and an alignment column

Label internal nodes to minimize the number of required substitutions

**Initialization:**

Set cost  $C = 0$ ;  $k = 2N - 1$

**Iteration:**

If  $k$  is a leaf, set  $R_k = \{x^k[u]\}$

If  $k$  is not a leaf,

Let  $i, j$  be the daughter nodes;

Set  $R_k = R_i \cap R_j$  if intersection is nonempty

Set  $R_k = R_i \cup R_j$ , and  $C += 1$ , if intersection is empty

**Termination:**

Minimal cost of tree for column  $u, = C$

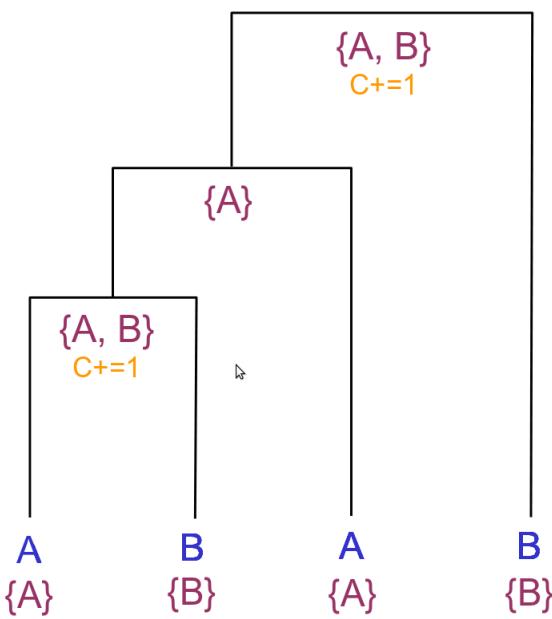


Figure 23.20: Parsimony scoring: union and intersection

**Traceback:**

1. Choose an arbitrary nucleotide from  $R_{2N-1}$  for the root
2. Having chosen nucleotide  $r$  for parent  $k$ ,  
If  $r \in R_i$  choose  $r$  for daughter  $i$   
Else, choose arbitrary nucleotide from  $R_i$

Easy to see that this traceback produces some assignment of cost C

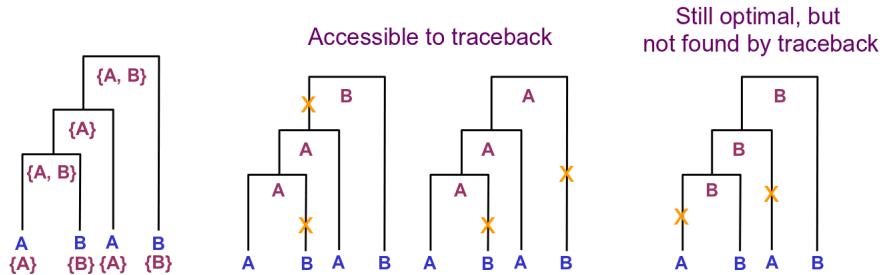


Figure 23.21: Parsimony traceback to find ancestral nucleotides

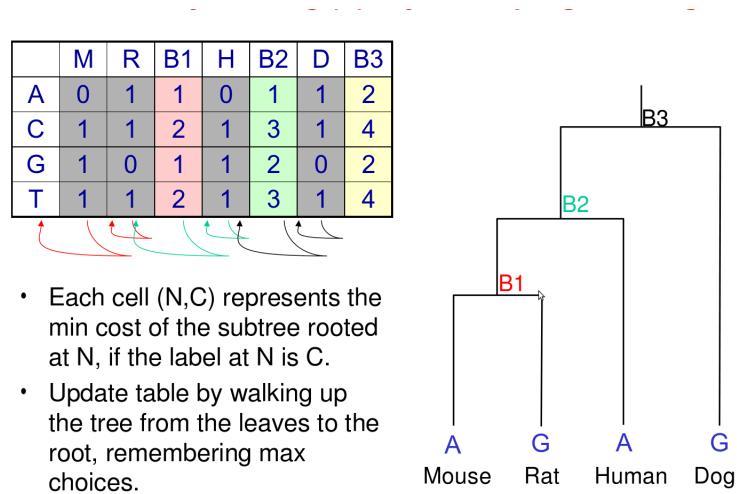


Figure 23.22: Parsimony scoring by dynamic programming

As we mentioned before, this method is simple and fast. However, this simplicity can distort the scores it assigns. For one thing, the algorithm presented here assumes that a given base pair undergoes at most one substitution along a branch from a given node, which may lead it to ignore highly probable internal sequences that violate this assumption. You can imagine that a sequence of mutations and back-mutations could have occurred in any given branch, which we are unable to observe and account for under this simple model. Furthermore, this method does not explicitly model the time represented along each edge, and thus cannot account for the increased chance of a substitution along edges that represent a long temporal duration, or the possibility of different mutation rates across the tree. Maximum likelihood methods largely resolve these shortcomings, and are thus more commonly used for tree scoring.

### Maximum Likelihood - Peeling Algorithm

As with the general maximum likelihood methods, this algorithm scores a tree according to the (log) joint probability of observing the data and the given tree, i.e.  $P(D, T)$ . The peeling algorithm again considers individual base pairs and assumes that all sites evolve independently. As in the parsimony method, this algorithm considers all base pairs independently: it calculates the probability of observing the given characters at each base pair in the leaf nodes, given the tree, a set of branch lengths, and the maximum likelihood assignment of the internal sequence, then simply multiplies this probabilities over all base pairs to get the total probability of observing the tree. Note that the explicit modeling of branch lengths is a difference from the previous approach.

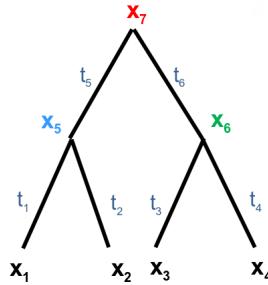


Figure 23.23: A tree to be scored using the peeling algorithm.  $n=4$

Here each node has a character  $x_i$  and  $t_i$  is the corresponding branch length from its parent. Note that we already know the values  $x_1, x_2 \dots x_n$ , so they are constants, but  $x_{n+1}, \dots x_{2n-1}$  are unknown characters at ancestral nodes which are variables to which we will assign maximum likelihood values. (Also note that we have adopted a leaves-to-root indexing scheme for the nodes, the opposite of the scheme we used before.) We want to compute  $P(x_1x_2 \dots x_n|T)$ . For this we sum over all possible combinations of values at the ancestral nodes. this is called marginalization. In this particular example

$$P(x_1x_2x_3x_4|T) = \sum_{x_5} \sum_{x_6} \sum_{x_7} P(x_1x_2 \dots x_7|T)$$

There are  $4^{n-1}$  terms in here, but we can use the following factorization trick:

$$= \sum_{x_5} \sum_{x_6} \sum_{x_7} P(x_1|x_5, t_1) P(x_2|x_5, t_2) P(x_3|x_6, t_3) P(x_4|x_6, t_4) P(x_5|x_7, t_5) P(x_6|x_7, t_6) P(x_7)$$

Here we assume that each branch evolves independently. And the probability  $P(b|c, t)$  denotes the probability of base  $c$  mutating to base  $b$  given time  $t$ , which is essentially obtained from the Jukes Cantor model or some more advanced model discussed earlier. Next we can move the factors that are independent of the summation variable outside the summation. That gives:

$$= \sum_{x_7} \left[ P(x_7) \left( \sum_{x_5} P(x_5|x_7, t_5) P(x_1|x_5, t_1) P(x_2|x_5, t_2) \right) \left( \sum_{x_6} P(x_6|x_7, t_6) P(x_3|x_6, t_3) P(x_4|x_6, t_4) \right) \right]$$

Let  $T_i$  be the subtree below  $i$ . In this case, our  $2n - 1 \times 4$  dynamic programming array computes  $L[i, b]$ , the probability  $P(T_i|x_i = b)$  of observing  $T_i$ , if node  $i$  contains base  $b$ . Then we want to compute the probability of observing  $T = T_{2n-1}$ , which is

$$\sum_b P(x_{2n-1} = b) L[2n-1, b]$$

Note that for each ancestral node  $i$  and its children  $j, k$ , we have

$$L[i, b] = \left( \sum_c P(c|b, t_j) L[j, c] \right) \left( \sum_c P(c|b, t_k) L[k, c] \right)$$

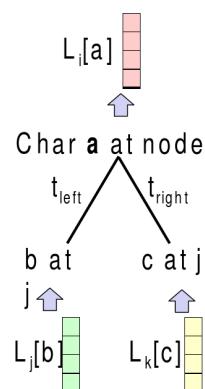


Figure 23.24: The

Subject to the initial conditions for the leaf nodes, i.e. for  $i \leq n$ :

$$L[i, b] = 1 \text{ if } x_i = b \text{ and } 0 \text{ otherwise}$$

Note that we still do not have the values  $P(x_{2n-1} = b)$ . It is usually assigned equally or from some prior distribution, but it does not affect the results greatly. The final step is of course to multiply all the probabilities for individual sites to obtain the probability of observing the set of entire sequences. In addition, once we have assigned the maximum likelihood values for each internal node given the tree structure and the set of branch lengths, we can multiply the resulting score by some prior probabilities of the tree structure and the set of branch lengths. These prior probabilities are often generated from distributions that specifically model biological processes such as the Yule-Simon distribution or the Moran process for genetic drift. The result of this final multiplication is called the a posteriori probability, using the language of Bayesian inference. The overall complexity of this algorithm is  $O(nmk^2)$  where  $n$  is the number of leaves (taxa),  $m$  is the sequence length, and  $k$  is the number of characters.

There are addvantages and disadvantages of this algorithm. Such as

#### Advantages:

1. Inherently statistical and evolutionary model-based.
2. Usually the most ‘consistent’ of the methods available.
3. Used for both character and rate analyses
4. Can be used to infer the sequences of the extinct ancestors.
5. Account for branch-length effects in unbalanced trees.
6. Nucleotide or amino acid sequences, other types of data.

#### Disadvantages:

1. Not as simple and intuitive as many other methods.
2. Computationally intense Limited by, number of taxa and sequence length).
3. Like parsimony, can be fooled by high levels of homoplasy.
4. Violations of model assumptions can lead to incorrect trees.

### 23.5.2 Search

A comprehensive search over the space of all trees would be extremely costly. The number of full rooted trees with  $n + 1$  leaves is the  $n$ -th catalan number

$$C_n = \frac{1}{n+1} \binom{2n}{n} \approx \frac{4^n}{n^{3/2}\sqrt{\pi}}$$

Moreover, we must compute the maximum likelihood set of branch lengths for each of these trees. Thus, it is an NP-Hard problem to maximize the score absolutely for all trees. Fortunately, heuristic search algorithms can generally identify good solutions in the tree space. The general framework for such search algorithms is as follows:

**Inititalization:** Take some tree as the base of iteration (randomly or according to some other prior, or from the distance based direct algorithms).

**Proposal:** Propose a new tree by randomly modifying the current tree slightly.

**Score:** Score the new proposal according to the methods described above.

**Select:** Randomly select the new tree or the old tree (corresponding probabilities according to the score(likelihood) ratio.

**Iterate:** Repeat to proposal step unless some termination criteria is met (some threshold score or number of steps reached).

The basic idea here is the heuristic assumption that the scores of closely related trees are similar, so that good solutions may be obtained by successive local optimization, which is expected to converge towards a overall good solution.

### Tree Proposal

One method for modifying trees is the Nearest Neighbor Interchange (NNI), illustrated below.

Figure 23.25: An unit step using Nearest Neighbor Interchange scheme

Another common method, not described here, is Tree Bisection and Join (TBJ). The important criteria for such proposal rules is that:

- (a) The tree space should be connected, i.e. any pair of trees should be obtainable from each other by successive proposals.
- (b) An individual new proposal should be sufficiently close to the original. So that it is more likely to be a good solution by virtue of the proximity to an already discovered good solution. If individual steps are too big, the algorithm may move away from an already discovered solution (also depends on the selection step). In particular, note that the measure of similarity by which the measure these step sizes is precisely the difference in the likelihood scores assigned to the two trees.

### Selection

Choosing whether or not to adopt a given proposal, like the process of generating the proposal itself, is inherently heuristic and varies. A general rules of thumb is:

1. If the new one has a better score, always accept it.
2. If it has a worse score, there should be some probability of selecting it, otherwise the algorithm will soon fixate in a local minima, ignoring better alternatives a little far away.
3. There should not be too much probability of selecting a worse new proposal, as this risks rejecting a known good solution.

It is the trade-off between steps 2 and 3 that determines a good selection rule. Metropolis Hastings is a Markov Chain Monte Carlo Method (MCMC) that defines specific rules for exploring the state space in a way that makes it a sample from the posterior distribution. These algorithms work somewhat well in practice, but there is no guarantee for finding the appropriate tree. To improve the outcome, a resampling method called bootstrapping is often used. Bootstrapping essentially involves running the algorithm over and over using subsets of the base pairs in the leaf sequences, and then favoring global trees that match the topologies generated by only these subsequences.

## 23.6 Possible Theoretical and Practical Issues with Discussed Approach

### 23.6.1 Major Issues

There is at least one notable source of bias in calculating sequence distances. Because aligned gene sequences are generally used to calculate distance, most current tree reconstruction methods depend heavily on the most highly conserved genes. However, it is often less conserved genes that provide great insight into the development of various morphological features. Some algorithms exist that take this bias into account, but they tend to be very computationally costly due to the NP-Hard nature of reconstructing trees.

Additionally, aligned sequences do not explicitly identify the evolutionary events that gave rise to them. Combinations of speciation, duplication, loss, and horizontal gene transfer (HGT) events are easy to misinterpret because only the DNA of recent species is available (see [11] for a commentary on such theoretical issues). For example, a duplication followed by a loss would be very hard to detect. A duplication followed by a speciation might be misinterpreted as an HGT event. Even the prior probabilities of different events, particularly HGT, is contested by modern scientists.

Another issue that arises is that frequently multiple marker sequences are concatenated and the concatenated sequence is used for distance calculations. This approach assumes that all of the concatenated genes had the same history, which may not be the case as HGT, duplications, and other events can occur differently in different genes. [8] show that different phylogenetic relationships can be inferred when individual genes rather than multiple concatenated genes are used for tree reconstruction. On the other hand, [4] claim that while HGT is prevalent, orthologs used for phylogenetic reconstruction are consistent with a single tree of life. Overall, debate exists in the field, and there remains to be found a non-arbitrary way to define species and infer phylogenetic relationships.

### 23.6.2 Forest of Life

Current research suggests that the evolution of prokaryotes has not actually followed a neat tree in which each descendent has a single ancestor. Rather than an organized evolution of species, there is instead just a network of genetic material that is transferred frequently between organisms. A new representation of the transfer of genetic material is clearly necessary. New analytical techniques rest on the construction of networks to describe this phenomenon, rather than single individual trees.

The techniques described throughout this chapter can still be applied to this new problem. The first step here is to generate trees that can explain some portion of the observed data. Those that are consistent with the most observations (generally 90 - 100 percent) are termed "Nearly Universal Trees" (NUT) and are selected for the pool of trees used to generate a network. Because NUTs are those trees consistent with the most observations, they have a number of notable characteristics. NUTs are closer to each other than what is expected by chance; in other words, the topologies of NUTs are highly consistent and non-random. The majority of NUTs are over genes encoding proteins involved in translation and the core aspects of transcription, which are likely to be highly conserved [9]. This fact coupled with the observation that NUTs form tightly connected networks when clustered by similarity proves that the genes that NUTs represent are very similar. These NUTs can be used to generate the consensus tree - the tree that on average, best represents trends in the data. The consensus tree is a bundle of the individual NUTs, and at any given node, there are a certain number of deviations from the central tree structure. Overall, the trend of the tree is clearly observable as the sum of many NUTs, but individual NUTs branch off giving the tree its signature network-like structure. The final result is a forest of life rather than a tree of life.

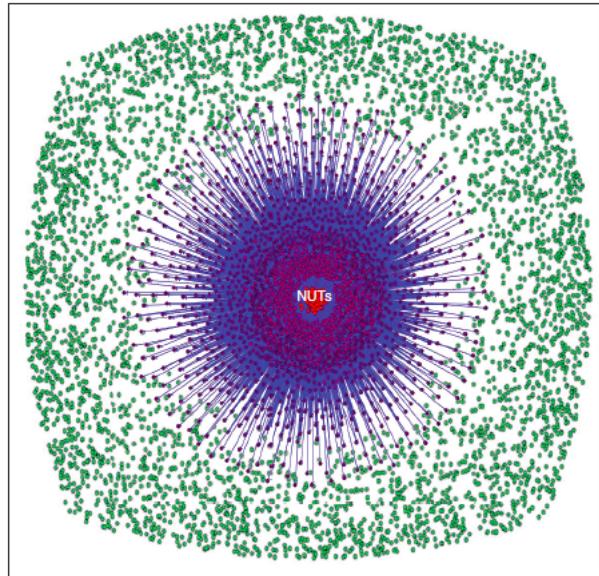


Figure 23.26: Network representation of the forest of life, with NUTs in the middle

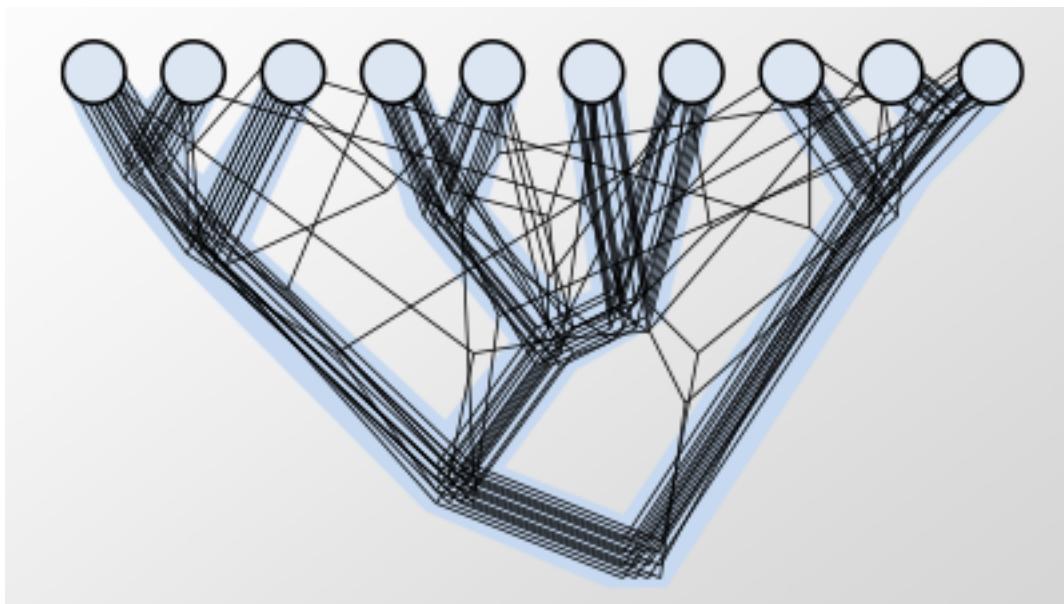


Figure 23.27: Forest generated from multiple trees

## 23.7 Towards final project

### 23.7.1 Project Ideas

1. Creating better distance models such as taking into account duplicate genes or loss of genes. It may also be possible to analyze sequences for peptide coding regions and calculate distances based on peptide chains too.
2. Creating a faster/more accurate search algorithm for turning distances into trees.

3. Analyze sequences to calculate probabilities of speciation, duplication, loss, and horizontal gene transfer events.
4. Extending an algorithm that looks for HGTs to look for extinct species. A possible use for HGTs is that if a program were to infer HGTs between different times, it could mean that there was a speciation where one branch is now extinct (or not yet discovered) and that branch had caused an HGT to the other extant branch.

### 23.7.2 Project Datasets

1. 1000 Genomes Project <http://www.1000genomes.org/>
2. Microbes Online <http://microbesonline.org/>

## 23.8 What Have We Learned?

In this chapter, we have learnt different methods and approaches for reconstructing Phylogenetic trees from sequence data. In the next chapter, its application in gene trees and species trees and the relationship between those two will be discussed, as well as modeling phylogenies among populations within a species and between closely related species.

## Bibliography

- [1] 1000 genomes project.
- [2] et al Ciccarelli, Francesca. Toward automatic reconstruction of a highly resolved tree of life. *Science*, 311, 2006.
- [3] Tal Dagan and William Martin. The tree of one percent. *Genome Biology*, Nov 2006.
- [4] Ochman Howard Daubin Vincent, Moran Nancy A. Phylogenetics and the cohesion of bacterial genomes. *Science*, 301, 2003.
- [5] A.J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584, Apr 2002.
- [6] Stephanie Guindon and Olivier Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systems Biology*, 52(5):696–704, 2003.
- [7] Sanderson MJ. r8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, 19(2):301–302, Jan 2003.
- [8] R. Thane Papke, Olga Zhaxybayeva, Edward J Fiel, Katrin Sommerfeld, Denise Muise, and W. Ford Doolittle. Searching for species in haloarchaea. *PNAS*, 104(35):14092–14097, 2007.
- [9] Pere Puigbo, Yuri I Wolf, and Eugene V Koonin. Search for a 'tree of life' in the thicket of the phylogenetic forest. *Journal of Biology*, 8(59), July 2009.
- [10] Sagi Snir, Yuri I Wolf, and Eugene V Koonin. Universal pacemaker of genome evolution. *PLoS computational biology*, 8(11), 2012.
- [11] Douglas L Theobald. A formal test of the theory of universal common ancestry. *Nature*, 465:219–222, 2010.



---

CHAPTER  
**TWENTYFOUR**

---

**PHYLOGENOMICS II**

Guest Lecture by  
Matt Rasmussen  
2012: Updated by Orit Giguzinsky and Ethan Sherbondy

**Figures**

---

24.1 Species Tree . . . . .	384
24.2 Gene Tree . . . . .	385
24.3 Gene Tree Inside a Species Tree . . . . .	385
24.4 Gene Family Evolution: Gene Trees and Species Trees . . . . .	386
24.5 Gene Duplication . . . . .	386
24.6 Maximum Parsimony Reconciliation (MPR) . . . . .	387
24.7 Labeling Duplication and Losses in MPR . . . . .	387
24.8 Reconciliation Example 1, simple mapping case . . . . .	388
24.9 Reconciliation Example 2, parsimonious reconciliation for complex case . . . . .	388
24.10 Reconciliation Example 3, non parsimonious reconciliation for complex case . . . . .	389
24.11 Reconciliation Example 4, invalid Reconciliation . . . . .	389
24.12 Species Tree Reconstruction . . . . .	390
24.13 Using species trees to improve gene tree reconstruction. . . . .	391
24.14 Developing Rates Model . . . . .	391
24.15 SPIMAP Generative Model . . . . .	391
24.16 SPIMAP equation to find the maximum a posteriori tree . . . . .	392
24.17 The Wright-Fisher model . . . . .	393
24.18 Many iterations of Wright-Fisher yielding a lineage tree . . . . .	394
24.19 $k$ lineages don't have any coalescent events in parental generation. . . . .	394
24.20 The first coalescence event occurs at $t$ . . . . .	395
24.21 The coalescent model. . . . .	395
24.22 Multispecies Coalescent Model . . . . .	397
24.23 MPR reconciliation of genes and species tree. . . . .	398
24.24 Inaccuracies in gene tree. . . . .	398
24.25 Recombination. . . . .	399
24.26 Gene tree with or without recombination. . . . .	400
24.27 Gene trees describing ancestral relationships of two sites (A and B) on a chromosome, separately and overlaid. . . . .	401
24.28 The blue, green, and yellow trees represent the regions below them. . . . .	401

---

## 24.1 Introduction

In the previous chapter, we covered techniques for reasoning about evolution in terms of trees of descent. The algorithms we covered for tree-building, UPGMA and neighbor-joining, assumed that we were comparing fully aligned sections of sequences.

In this section, we present additional models for using phylogenetic trees in different contexts. Here we clarify the differences between species and gene trees. We then cover a framework called reconciliation which lets us effectively combine the two by mapping gene trees onto species trees. This mapping gives us a means of inferring gene duplication and loss events.

We will also present a phylogenetic perspective for reasoning about population genetics. Since population genetics deals with relatively recent mutation events, we offer the Wright-Fisher model as a tool for representing changes in whole populations. Unfortunately, when dealing with real-world data, we usually are only able to sequence genes from the current living descendants of a group. As a remedy to this shortcoming, we cover the Coalescent model, which you can think of as a time-reversed Wright-Fisher analog.

By using coalescence, we gain a new means for estimating divergence times and population sizes across multiple species. At the end of the chapter, we touch briefly on the challenges of using trees to model recombination events and summarize recent work in the field along with frontiers open for exploration.

## 24.2 Inferring Orthologs/Paralogs, Gene Duplication and Loss

There are two commonly used trees, Species tree and Gene tree. This section explains how these trees can be used and how to fit a gene tree inside a species tree by a process called **reconciliation**.

### 24.2.1 Species Tree

**Species trees** show how different species evolved from one another. These trees are created using morphological characters, fossil evidence, etc. The leaves of each tree are labeled as species and the rest of the tree shows how these species are related. An example of a species tree is shown in Figure 24.1. Note: in lecture it is mentioned that a species can be thought of as a "bag of genes", that is to say the group of common genes among members of a species.

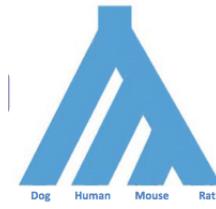


Figure 24.1: Species Tree

### 24.2.2 Gene Tree

**Gene trees** are trees that look at specific genes in different species. The leaves of gene trees are labeled with gene sequences or gene ids associated with specific sequences. Figure 24.2 shows an example of a gene tree that has 4 genes (leaves). The sequences associated with each gene are presented on the right side of Figure 24.2. Unlike Species trees, multiple nodes in a gene tree can correspond to the same species, because there can be multiple copies of that gene within that species' genome.

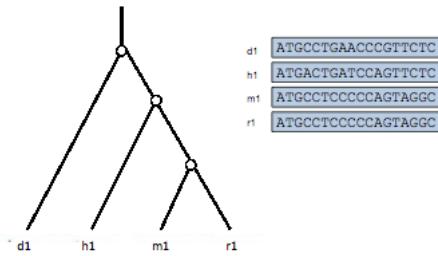


Figure 24.2: Gene Tree

### 24.2.3 Gene Family Evolution

Gene trees evolve inside a species tree. An example of a gene tree contained in a species tree is shown in Figure 24.3 below.

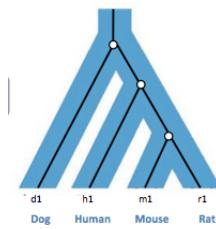


Figure 24.3: Gene Tree Inside a Species Tree

The next sub section explains how we can fit gene trees inside a species trees using Reconciliation.

### 24.2.4 Reconciliation

**Reconciliation** is an algorithm that helps compare gene trees to genome trees by fitting a gene tree inside a species tree. This is done by mapping the vertices in the gene tree to vertices in the species tree. This subsection will focus on Reconciliation, related definitions, algorithms (Maximum Parsimony Reconciliation and SPIDIR) and examples.

#### Definitions

Two genes are **orthologs** if their most recent common ancestor (MRCA) is a speciation (splitting into different species).

**Paralogs** are genes whose MRCA is a duplication.

Figure 24.5 below illustrates how these types of genes can be represented in a gene tree. The tree below has 4 speciation nodes, one duplication and one loss.

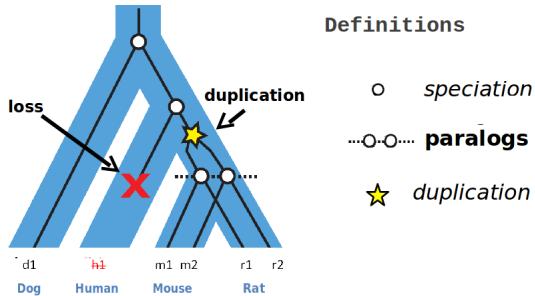


Figure 24.4: Gene Family Evolution: Gene Trees and Species Trees

Losses and duplications cause incongruences between the species and gene trees. By mapping the gene tree vertices to the species tree, reconciliation locates where the losses and duplication events occur.

**Losses** occur when a species loses a copy of a certain gene.

**Duplications** cause there to be two copies of the same parent gene within a species. Because there is always at least one functional copy within the genome, it frees up mutation of the second copy, and is a major mechanism for creating new genes and functions.

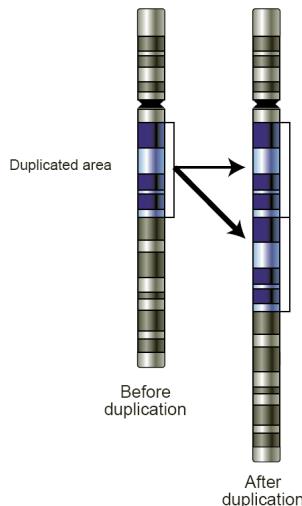


Figure 24.5: Gene Duplication

There are three ways duplication can affect the genome.

In **nonfunctionalization**, since there are two originally functional copies of the gene, one becomes mutated and non-functional.

In **neofunctionalization**, one of the copies develops an entirely new function due to mutation. One prominent example of this is the gain of tri-color vision in old world monkeys.

In **subfunctionalization**, one of the copies mutate to perform different parts of the original function of the gene.

### Maximum Parsimony Reconciliation (MPR) Algorithm

MPR is an algorithm that fits a gene tree into a species tree while minimizing the number of duplications and deletions.

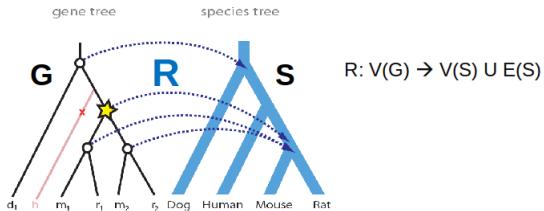


Figure 24.6: Maximum Parsimony Reconciliation (MPR)

Figure 24.6 above shows an example of a possible mapping from a gene tree to a species tree. Below is pseudocode for the MPR algorithm. In the algorithm, each node in the gene tree is mapped to a corresponding node in the species tree. Below is the pseudocode for this recursive algorithm. The mapping must preserve ancestor-descendant relations. This means that if some node  $w$  is a decendent of  $v$  in the gene tree, then the mapping  $R(w)$  must be a descendant of  $R(v)$  in the species tree, or they must map to the same spot.

The base case involves canonically matching the leaves of the gene tree to the leaves of the species tree, since each gene belongs to some species. The algorithm then progresses up the vertices of the gene tree, drawing a relationship between the MRCA of all leaves within a given vertex's sub-tree and the corresponding MRCA vertex in the species tree.

$$R[v] = \begin{cases} \text{the species of } v & \text{if } v \text{ is a leaf of the gene tree} \\ \text{MRCA}(R[\text{left child of } v], R[\text{right child of } v]) & \text{if } v \text{ is an internal node of the gene tree} \end{cases}$$

Using the MPR algorithm, we can identify duplications and losses as follows:

A node  $v$  in the gene tree corresponds to a **duplication event** if  $R(v)$  equals either  $R[\text{right}(v)]$  or  $R[\text{left}(v)]$ . In this case we can think of  $v$  as mapping to the edge above  $R[v]$ . This intuitively makes sense because it implies that the two children of  $v$  did **not** correspond to two different species groups.

Furthermore, the branch above a node  $v$  has one or more **loss** if  $R[\text{parent}(v)]$  is not equal to either  $R[v]$  or the parent *in the species tree* of the  $R[v]$ . The figure below demonstrates these two cases.

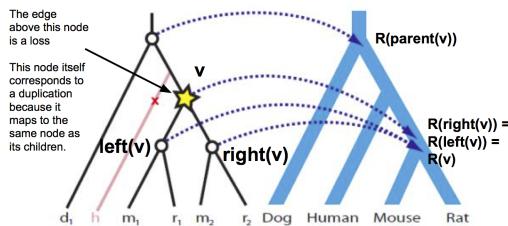


Figure 24.7: Labeling Duplication and Losses in MPR

As is evident in the algorithm, we map the arrows low as possible, to minimize duplication and losses without violating the ancestor-descendant relations.

## Reconciliation Examples

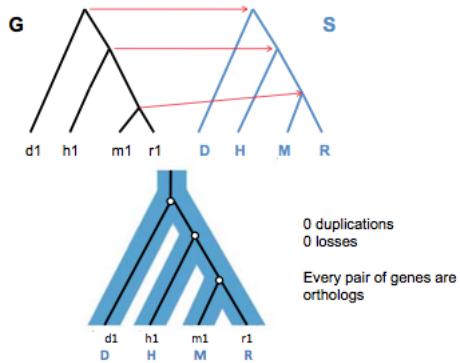


Figure 24.8: Reconciliation Example 1, simple mapping case

In Figure 24.8, the nodes can be mapped straight across, since there are no duplications or losses.

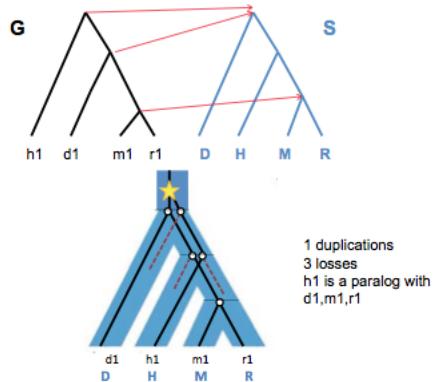


Figure 24.9: Reconciliation Example 2, parsimonious reconciliation for complex case

In Figure 24.9, we see a parsimonious (minimum number of losses and duplications) reconciliation for a case in which nodes from the gene tree cannot be mapped straight across. This is a result of the swapped locations of h1 and d1 in the gene tree; the least common ancestor for d1, m1, and r1 is now the root vertex of the species tree.

does a non-parsimonious reconciliation look like?

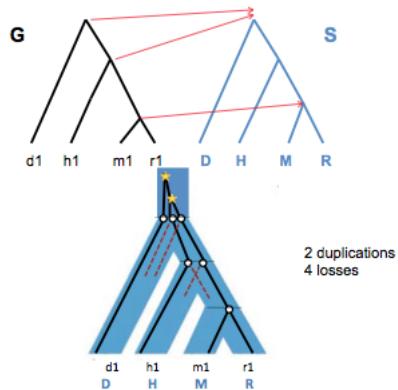
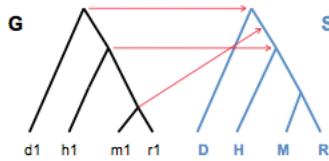


Figure 24.10: Reconciliation Example 3, non parsimonious reconciliation for complex case

Figure 24.10 shows a non-parsimonious reconciliation . The parsimonious mapping for the same trees is shown in Figure 24.8.



**Valid reconciliations must respect descendant-ancestor relationships.**

If  $a < b$  in G, then  $R[a] \leq R[b]$  in S

Figure 24.11: Reconciliation Example 4, invalid Reconciliation

Figure 24.11 shows an invalid reconciliation. This reconciliation is invalid since it does not respect descendant-ancestor relationships. In order for this reconciliation to be possible, the descendant would have to travel back in time and be created before its ancestor. Clearly, such a scenario would be impossible. A valid reconciliation must satisfy the following: **If  $a < b$  in G, then  $R[a] \leq R[b]$  in S.**

### 24.2.5 Interpreting Reconciliation Examples

Gene trees, when reconciled with species trees, offer significant insight into evolutionary events, namely duplications and losses.

In Figure 4, we see that a duplication event occurred before the divergence of mice and rats as species. This is why we see similar genes at both m1 and m2, which represent two separate loci. d2 and h2 are not included in the graph because the gene being considered is not present at those loci (since no duplication event occurred), whereas it is at both m2 and r2.

If the duplication event were to have occurred one level higher in Figure 4, without seeing a corresponding h2 in the gene tree, this would imply a loss within the h branch of the species tree.

## 24.3 Reconstruction

In the previous section we learned how to compare and combine gene trees and species trees. In this section, we will use this information to reconstruct gene trees and species trees.

### 24.3.1 Species Tree Reconstruction

In the past, it was very difficult to identify a marker gene that would give insight into the differentiation for a specific species. As sequencing improved, we started having large amounts of sequencing data on various genes. Based on different sets of loci, people built different trees, which were highly dependent on the set of loci chosen. Possible reasons why trees differ include noise (from statistical estimate errors and noise), hidden duplications and losses, and allele sorting in a population. Species Tree Reconstruction provides us with a way to reconcile many different gene trees into a single overarching species tree. Later we will see that we can also run this process in reverse and use a species tree to help construct gene trees.

#### Species Tree Reconstruction Problem

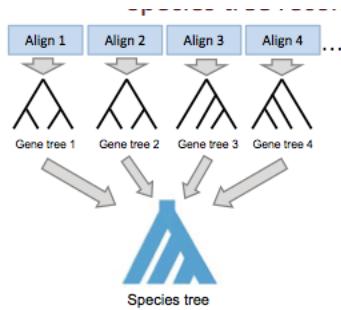


Figure 24.12: Species Tree Reconstruction

Given lots of different gene trees that disagree, our goal is to make them into one species tree (as shown in Figure 24.12). There are lots of different algorithms that reconstruct species trees. These algorithms include Supermatrix methods (Rokas 2003, Ciccarelli 2006), Supertree methods (Creevey & McInerney 2005), Minimizing Deep Coalescence (Maddison & Knowles 2006) and Modeling coalescence (Liu & Pearl 2007).

Supermatrix methods, most effective for noisy data, pull more data together in order to increase accuracy by concatenating gene alignments into a super-matrix. Supertree methods involve building a tree for each one and using a consensus method to summarize these trees, then identifying analogous branches across the lot of trees and building a species tree that has the branches that occur most frequently.

Minimizing Deep Coalescences is most effective for the case when gene trees disagree because of duplications and losses, and works by finding the species tree that applies the fewest duplications. It first proposes a species tree from information given by all the gene trees. Next, it uses reconciliation to determine the number of events each gene tree combined with the proposed species tree implies. Then, it propose other species trees and moves branches around to try to minimize contradictions with the gene trees. Wrong species trees tend to have many of these contradicting events. The correct tree should have the fewest number of contradicting events.

### 24.3.2 Improving Gene Tree Reconstruction and Learning Across Gene Trees

We can use methods similar to those described above to build better gene trees. This can be done by using information from a species tree to study a gene tree of interest. For example, species trees can be used to determine when losses and duplications occurred. The idea is that we can use the fact that species trees are often built from the entire genome, to obtain more information about related gene trees. We can use both the branch length and the number of events to do this.

If we know the species tree, we can develop a model for what kind of branch lengths we can expect. We can use conserved gene order to tell orthologs and build trees.

When a gene is fast evolving in one species, it is fast evolving in all species. We can model a branch length as two different rate components. One is gene specific (present across all species) and the other is species specific. This rates model method greatly improves reconstruction accuracy.

The SPIMAP (Species Informed Maximum A Posteriori) method uses a generative model for gene tree reconstruction. First, the algorithm takes in a given species tree. Second, it generates a gene tree topology

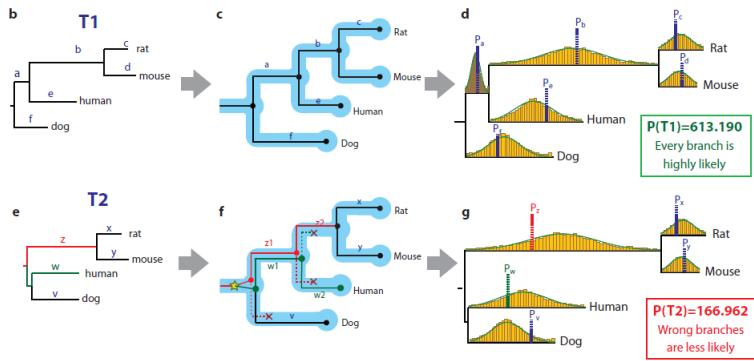


Figure 24.13: Using species trees to improve gene tree reconstruction.

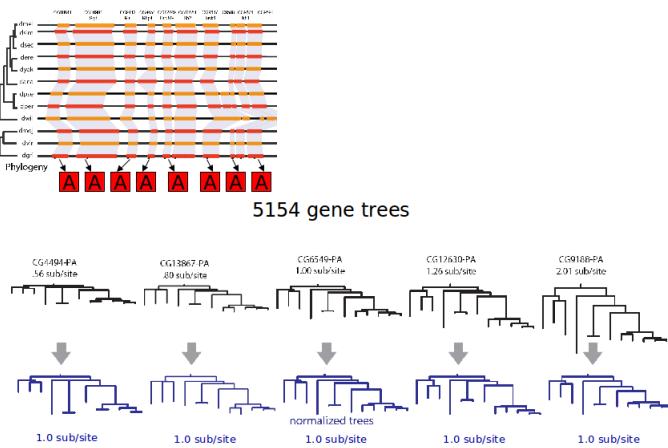


Figure 24.14: Developing Rates Model

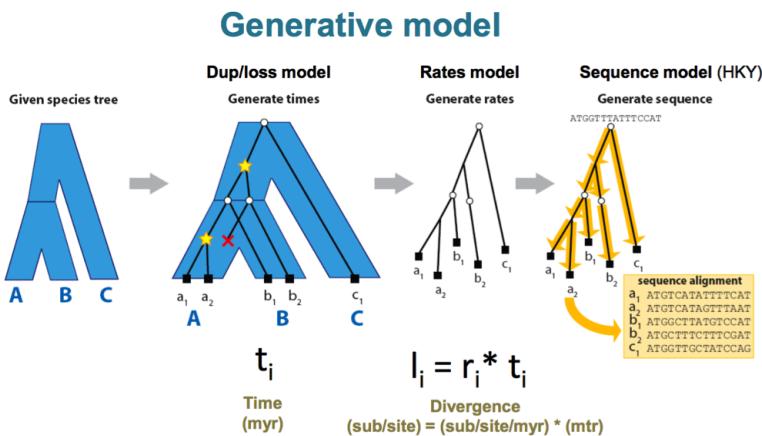


Figure 24.15: SPIMAP Generative Model

from the species tree. The gene tree branches either at duplication events, denoted by yellow stars located at branches, or at speciation events, denoted by white circles located at nodes. Gene loss events are denoted by a red "X". Third, it generates substitution rates using a rates model as discussed earlier. Finally, it uses Bayesian methods to narrow down sequences in the gene tree and output a sequence alignment.

The equation used to find the maximum a posteriori tree is included down below. It uses the topology prior inferred from the Dup/loss model and the branch prior inferred from the rates model, and the likelihood inferred from the Sequence model to come up with the most likely gene tree.

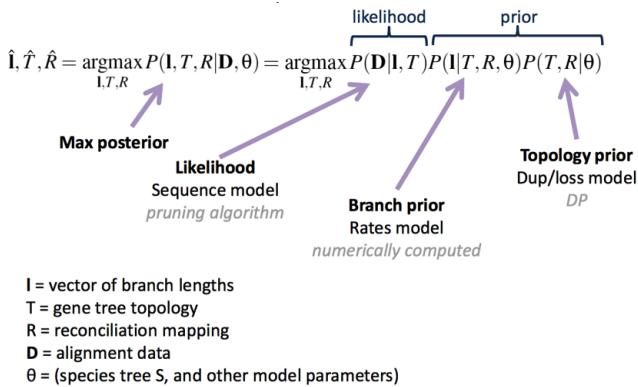


Figure 24.16: SPIMAP equation to find the maximum a posteriori tree

## 24.4 Modeling Population and Allele Frequencies

With the advent of next-gen sequencing, it is becoming economical to sequence the genomes of many individuals within a population. In order to make sense of how alleles spread through a population, it's helpful to have a model to compare data against. The **Wright-Fisher** reproduction model has filled this role for the past 70 years.

### 24.4.1 The Wright-Fisher Model

The Wright-Fisher model is an idealized, forward-time model that is used to study the effect of drift in a population. It models a case of perfectly non-overlapping generations, in contrast with the Moran model, which models perfectly overlapping generations. While the Wright-Fisher model does not represent the true biology of any particular species, it allows us to model evolution within a population in a practicable way.

Like HMMs, Wright-Fisher is a Markov process: at each step, the system randomly progresses, and the current state of the system depends only on the previous state. In this case, state transitions represent reproduction. It models the stochastic behavior of population structure (Wright 1931) by modeling the transmission of chromosomes to offspring. We can use this process to study allele frequency variations under a number of biologically relevant scenarios, including drift, migration, population bottlenecks and selection.

The model makes a number of simplifying assumptions:

1. Population size,  $N$ , is constant at each generation.
2. Only members of the same generation reproduce (no overlap).
3. Reproduction occurs at random.
4. The gene being modeled only has 2 alleles.
5. Genes undergo neutral selection.

Note that Wright-Fisher is not an appropriate choice if you're trying to model the change in frequency of a gene that is positively or negatively selected for.

The Wright-Fisher model works by modeling reproduction in the following way:

1. Establish a population of size  $N$ .

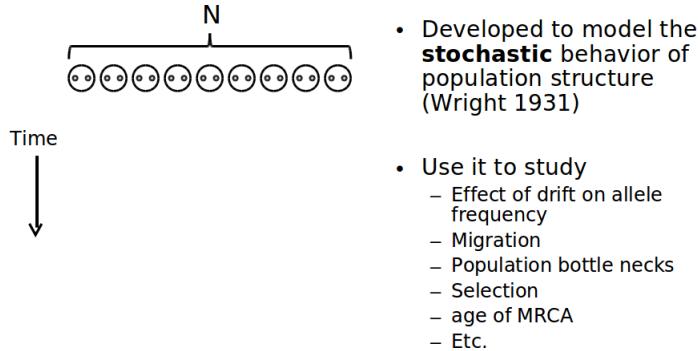


Figure 24.17: The Wright-Fisher model

2. Each individual generates an infinite number of gametes.
3. From this infinite pool of gametes, randomly choose gametes to make up N individuals in the next generation

This is equivalent to a process in which each chromosome is chosen randomly with replacement from the previous generation. At every generation, for each child, we randomly select from the parents (with replacement). The allele of the child becomes that of the randomly selected parent. We repeat this process for many generations, with the children serving as the new parents, ignoring the ordering of chromosomes.

By continuing this process for many generations, while keeping track of the parent-child relationships between generations, we can then trace individual lineages back through time, and determine when they coalesce. The Wright-Fisher model is particularly useful for this, because it is only necessary to focus on extant chromosomes in the final generation. Additionally, when considering a diploid population without sexual reproduction, we can focus on just the chromosomes, rather than the diploid organism as a whole, and consider a model with population size  $2N$ .

To determine the probability of  $k$  copies of an allele existing in the child generation when it had a frequency of  $p$  in the parent generation, we can use this formula:

$$\binom{2N}{k} p^k q^{2N-k} \quad (24.1)$$

Here,  $q = (1 - p)$ . It is the frequency of non-p alleles in the parent generation.

Now we can begin to explore such questions as: how probable is it and how many generations is it expected to take for a given allele to become **fixed**, meaning the allele is present in *every* member of the population?

The expected time (in generations) for fixation, given the assumptions made by Wright-Fisher, is proportionate to  $4N_E$ , where  $N_E$  is the effective population size.

Again, it's important to keep in mind the limitations of this model and ask if it actually makes sense for the system you're trying to represent. Consider how you could tweak the proposed model to account for a selection coefficient ranging between -1 (lethal negative selection) and 1 (strong positive selection).

### 24.4.2 The Coalescent Model

The problem with the Wright-Fisher model is that it assumes you know the allele frequencies of the ancestral generation. When dealing with the genomes of real species, these quantities are unknown. The Coalescent Model solves this conundrum by thinking retrospectively. That is to say: we start with the alleles of the *current* generation, and work our way *backwards* in time. The basic Coalescent Model makes the same assumptions as Wright-Fisher. And at each generation, we ask: what is the probability of the two identical alleles coalescing, or sharing a parent, in the previous generation.

The coalescent model focuses on reporting the distribution of coalescent times  $t_i$ . It assumes that each generation, each pair of lineages has an equal probability of coalescence. The probability that one lineage

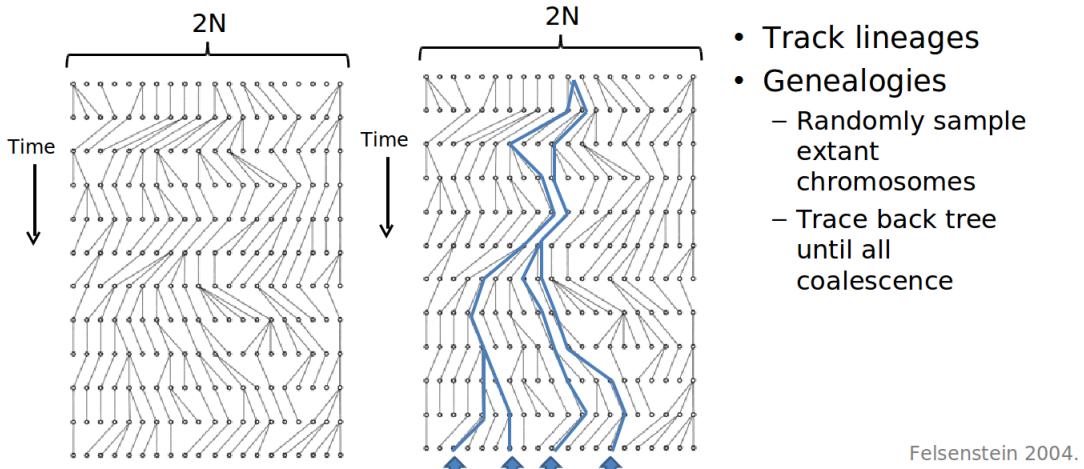


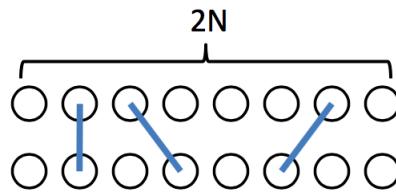
Figure 24.18: Many iterations of Wright-Fisher yielding a lineage tree

has a specific parent in the previous generation is equal to  $\frac{1}{2N}$ . Therefore the probability that two lineages have the same parent in the previous generation is also  $\frac{1}{2N}$ . When population size ( $N$ ) is large, it can be assumed that the time until coalescence ( $t_i$ ) will also be large. Therefore, we can treat coalescence times as continuous.

In order to determine the probability distribution of the coalescent times, we first consider the probability that  $k$  lineages *don't* have any coalescent events in the parental generation, or that all  $k$  lineages have different parents (Figure 24.19). This can be described by the following equation:

$$\frac{(2N-1)}{2N} \frac{(2N-2)}{2N} \dots \frac{(2N-k+1)}{2N} = \prod_{k=1}^{i=1} \left(1 - \frac{i}{2N}\right) = 1 - \sum_{k=1}^{i=1} \frac{j}{2N} + O\left(\frac{1}{N^2}\right) = 1 - \binom{k}{2} \frac{1}{2N} + O\left(\frac{1}{N^2}\right) \quad (24.2)$$

When  $k$  is significantly smaller than  $N$ ,  $O(\frac{1}{N^2})$  is very small, and can be ignored. This assumption is known as Kingman's approximation.

Figure 24.19:  $k$  lineages don't have any coalescent events in parental generation.

The probability that the first coalescence of  $k$  lineages occurs at time  $t$  generations (Figure 24.20), is therefore the probability that no coalescent events occurred before  $t-1$  (i.e. more recently) multiplied by the probability that a coalescent event occurs at  $t$ . This can be described by the following equation:

$$P(T_k = t) \approx \left\{1 - \binom{k}{2} \frac{1}{2N}\right\}^{t-1} \binom{k}{2} \frac{1}{2N} = Geom\left(\binom{k}{2} \frac{1}{2N}\right) \approx Exp\left(\binom{k}{2} \frac{1}{2N}\right) \quad (24.3)$$

This equation again assumes that when  $k$  is significantly less than  $N$ , the probability of multiple coalescence events occurring in one generation is negligible. For large  $t$ , the geometric distribution approaches an exponential.

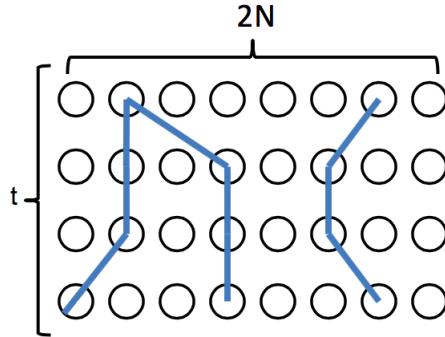


Figure 24.20: The first coalescence event occurs at  $t$

For example, when we consider the coalescence of only two lineages, the probability distribution of time until coalescence ( $P_c(t)$ ) is described by the following equation:

$$P_c(t) = \left(1 - \frac{1}{2N_e}\right)^{t-1} \left(\frac{1}{2N_e}\right) \quad (24.4)$$

Where  $N_e$  is the effective population size, or the effective number of individuals in the population that are actually reproducing. This number can be different than the actual population size due to factors such as non-random mating, population substructure, etc.

By approximating this geometric distribution as an exponential one:  $P_c(t) = \frac{1}{2N_e} e^{-(\frac{t-1}{2N_e})}$ , we can determine the expected number of generations back until coalescence, which is  $2N_e$ , with a standard deviation of  $2N_e$ . When we expand this to  $k$  lineages, the time to coalescence of  $k$  gene copies into  $k-1$  copies is exponentially distributed with mean =  $\frac{4N}{k(k-1)}$

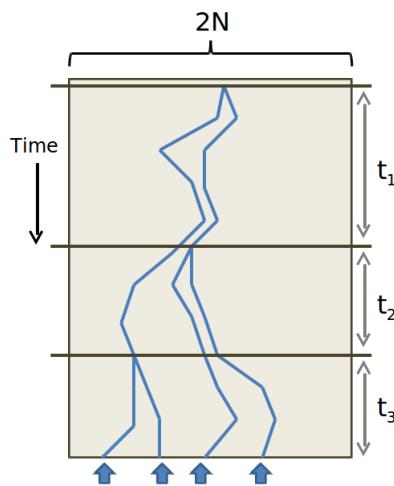


Figure 24.21: The coalescent model.

The individual at which two lineages converge is referred to as the **Most Recent Common Ancestor**. By continually moving backwards until all ancestors coalesce, we end up with a new kind of tree (Figure 24.21).

And by comparing the tree resulting from coalescence with a gene tree we've constructed, discrepancies between the two may signal that certain assumptions of the Coalescent Model have been violated, for instance non-random mating, selection, or population substructure may be present.

### 24.4.3 The Multispecies Coalescent Model

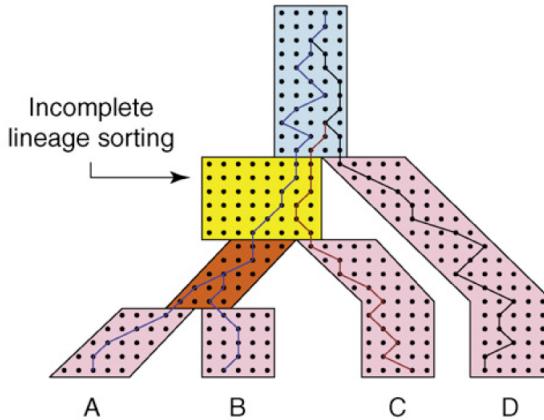


Figure 24.22: Multispecies Coalescent Model.

We can take this idea once step further and track coalescence events across multiple species. Here, each genome of an individual species is treated as a lineage.

Note that there is a lag time between the separation of two populations and the time at which two gene lineages coalesce into a common ancestor. Also note how the rate of coalescence slows down as N gets bigger and for short branches.

In the image above, deep coalescence is depicted in light blue for three lineages. The species and gene trees here are incongruent since C and D are sisters in gene tree but not the species tree.

There is a  $\frac{2}{3}$  chance that incongruence will occur because once we get to the light blue section, Wright-Fisher is memoryless and there is only  $\frac{1}{3}$  chance that it will be congruent. The effect of incongruence is called **Incomplete Lineage Sorting**. By measuring the frequency at which **ILS** occurs, we gain insight into unusually large populations or unusually short branch lengths within the species tree.

You can build a maximum parsimony species tree based on the notion of minimizing the number of ILS events rather than minimizing implied duplication/loss events as covered previously. It is even possible to combine these two methods to, ideally, create a phylogeny that is more accurate than either of them would be individually.

## 24.5 SPIDIR

### 24.5.1 Background

As presented in the supplementary information for SPIDIR, a gene family is the set of genes that are descendants of a single gene in the most recent common ancestor (MRCA) of all species under consideration. Furthermore, genetic sequences undergo evolution at multiple scales, namely at the level of base pairs, and at the level of genes. In the context of this lecture, two genes are orthologs if their MRCA is a speciation event; two genes are paralogs if their MRCA is a duplication event.

In the genomic era, the species of a modern genes is often known; ancestral genes can be inferred by reconciling gene- and species-trees. A reconciliation maps every gene-tree node to a species-tree node. A common technique is to perform Maximum Parsimony Reconciliation (MPR), which finds the reconciliation R implying the fewest number of duplications or losses using the recursion over inner nodes  $v$  of a gene tree  $G$ . MPR first maps each leaf of the gene tree to the corresponding species leaf of the species tree. Then the internal nodes of  $G$  are mapped recursively:

$$R(v) = \text{MRCA}(R(\text{right}(v)), R(\text{left}(v)))$$

If a speciation event and its ancestral node are mapped to the same node on the species tree. Then the ancestral node must be a duplication event.

Using MPR, the accuracy of the gene tree is crucial. Suboptimal gene trees may lead to an excess of loss and duplication events. For example, if just one branch is misplaced (as in ??) then reconciliation infers 3 losses and 1 duplication event. In [6], the authors show that the contemporaneous current gene tree methods perform poorly (60% accuracy) on single genes. But if we have longer concatenated genes, then accuracy may go up towards 100%. Furthermore, very quickly or slowly evolving genes carry less information as compared with moderately diverging sequences (40-50% sequence identity), and perform correspondingly worse. As corroborated by simulations, single genes lack sufficient information to reproduce the correct species tree. Average genes are too short and contains too few phylogenetically informative characters. While many early gene tree construction algorithms ignored species information, algorithms like SPIDIR capitalize on the insight that the species tree can provide additional information which can be leveraged for gene tree construction. Synteny can be used to independently test the relative accuracy of different gene tree reconstructions. This is because syntenic blocks are regions of the genome where recently diverged organisms have the same gene order, and contain much more information than single genes.

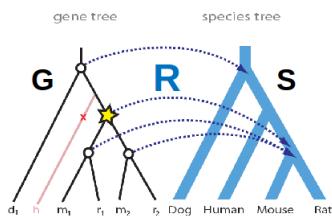


Figure 24.23: MPR reconciliation of genes and species tree.

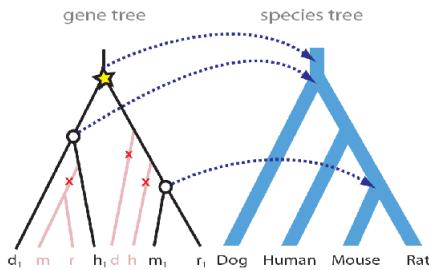


Figure 24.24: Inaccuracies in gene tree.

There have been a number of recent phylogenomic algorithms including: RIO [2], which uses neighbor joining (NJ) and bootstrapping to deal with incogruencies, Orthostrapper [8], which uses NJ and reconciles to a vague species tree, TreeFAM [3], which uses human curation of gene trees as well as many others. A number of algorithms take a more similar track to SPIDIR [6], including [4], a probabilistic reconciliation algorithm [9], a Bayesian method with a clock,[11],and parsimony method using species tree , as well as more recent developments: [1] a Bayesian method with relaxed clock and [5], a Bayesian method with gene and species specific relaxed rates (an extension to SPIDIR) .

### 24.5.2 Method and Model

SPIDIR exemplifies an iterative algorithm for gene tree construction using the species tree. In SPIDIR, the authors define a generative model for gene-tree evolution. This consists of a prior for gene-tree topology and branch lengths. SPIDIR uses a birth and death process to model duplications and losses (which informs the prior on topology) and then then learns gene-specific and species-specific substitution rates (which inform the prior on branch lengths). SPIDIR is a *Maximum a posteriori (MAP)* method, and, as such, enjoys several nice optimality criteria.

In terms of the estimation problem, the full SPIDIR model appears as follows:

$$\text{argmax}_{L,T,R} P(L,T,R|D,S,\Theta) = \text{argmax}_{L,T,R} P(D|T,L)P(L|T,R,S,\Theta)P(T,R|S,\Theta)$$

The parameters in the above equation are:  $D$  = alignment data ,  $L$  = branch length  $T$  = gene tree topology ,  $R$  = reconciliation ,  $S$  = species tree (expressed in times) ,  $\Theta$  = ( gene and species specific parameters [estimated using EM training],  $\lambda$ ,  $\mu$  dup/loss parameters)). This model can be understood through the three terms in the right hand expression, namely:

1. the sequence model–  $P(D|T, L)$ . The authors used the common HKY model for sequence substitutions, which unifies Kimura’s two parameter model for transitions and transversions with Felsenstein’s model where substitution rate depends upon nucleotide equilibrium frequency.
2. the first prior term, for the rates model–  $P(L|T, R, S, \Theta)$ , which the authors compute numerically after learning species and gene specific rates.
3. the second prior term, for the duplication/loss model–  $P(T, R|S, \Theta)$ , which the authors describe using a birth and death process.

Having a rates model is very useful, since mutation rates are quite variable across genes. In the lecture, we saw how rates were well described by a decomposition into gene and species specific rates. In lecture we saw that an inverse gamma distribution appears to parametrize the gene specific substitution rates, and we were told that a gamma distribution apparently captures species specific substitution rates. Accounting for gene and species specific rates allows SPIDIR to build gene trees more accurately than previous methods. A training set for learning rate parameters can be chosen from gene trees which are congruent to the species tree. An important algorithmic concern for gene tree reconstructions is devising a fast tree search method. In lecture, we saw how the tree search could be sped up by only computing the full  $\text{argmax}_{L,T,R} P(L,T,R|D,S,\Theta)$  for trees with high prior probabilities. This is accomplished through a computational pipeline where in each iteration 100s of trees are proposed by some heuristic. The topology prior  $P(T, R|D, S, \Theta)$  can be computed quickly. This is used as a filter where only the topologies with high prior probabilities are selected as candidates for the full likelihood computation.

The performance of SPIDIR was tested on a real dataset of 21 fungi. SPIDER recovered over 96% of the synteny orthologs while other algorithms found less than 65%. As a result, SPIDER invoked much fewer number of duplications and losses.

## 24.6 Ancestral Recombination Graphs

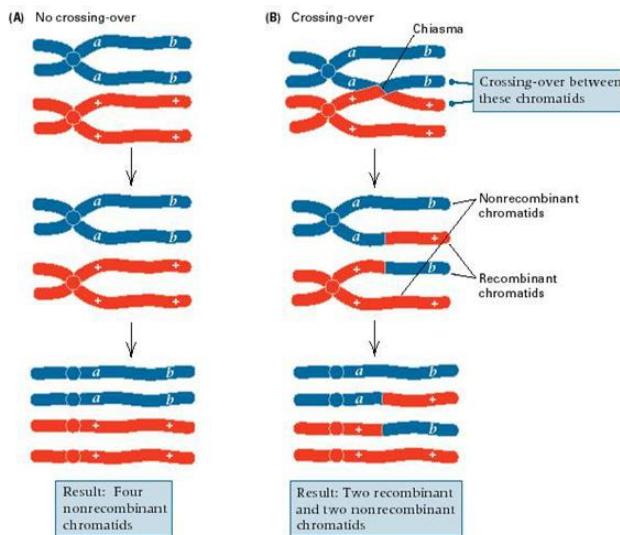


Figure 24.25: Recombination.

In Figure 24.25 a, the two chromosomes at the top represent the homologous chromosomes of a parent. The red chromosome represents the genetic information from the mother and the blue chromosome represents the genetic information from the father (of the grandparent generation). Without crossing-over (recombination), the parent will either pass on the red or the blue genetic information to the offspring. In reality, recombination happens during meiosis so that a parent will pass on some genetic information from both grandparents, effectively passing on a better representation of the parent genetic information.

At each generation, a recombination event can occur at any loci. This complicates our ability to reconstruct the ancestry of a particular chromosome, because a recombination event means that regions of the chromosome on opposite sides of the recombination point may have independent ancestries. Instead of being able to draw a simple tree, in which every node has one parent (except the root), with recombination a chromosome has two parents, and therefore requires a more complex tree (see Figure 24.26). Chromosomes may undergo recombination at every generation, creating increasingly complex gene trees.

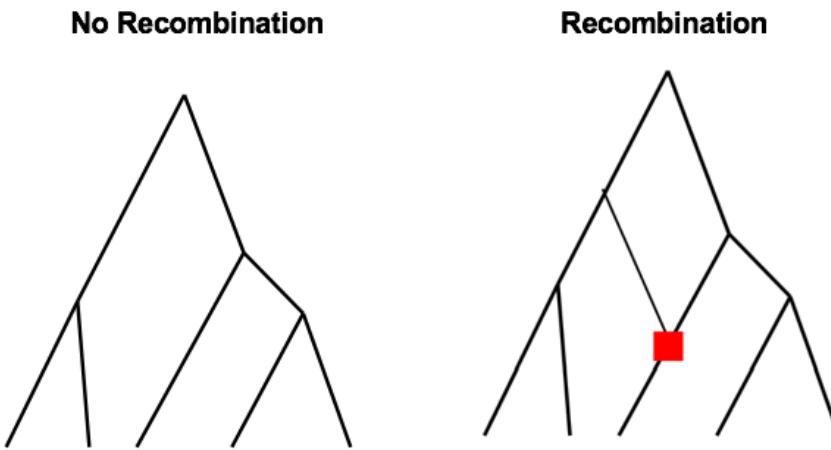


Figure 24.26: Gene tree with or without recombination.

Recombination complicates the process of reconstruction of the ancestry of a loci (or portion of the chromosome), because a single locus may have multiple lineages that are ancestral to it. In order to make meaningful inferences about the coalescent process in the presence of recombination, we must take these complex trees into account. One way to do this, is to consider each location in the chromosome separately.

An ancestral recombination graph (ARG) defines a collection of ancestry trees, and indicates which trees define the evolutionary history of each site on the chromosome. By treating each site in the genome separately and modeling their evolutionary history, we allow different sites to have different gene genealogies (see Figure 24.27). When recombination rates are high, the sites will segregate independently and have independent gene genealogies, while with negligible levels of recombination, sites will share a gene genealogy. Intermediate levels of recombination, some regions along the chromosome will have shared ancestries, while others will not. [10] It is impossible to know exactly how many recombination events have occurred in the evolutionary history of a chromosome, however it can be informative to determine the minimum number of recombination events that must have occurred in its evolutionary history. [7]

### 24.6.1 The Sequentially Markov Coalescent

Whereas the Wright-Fisher model makes the simplifying assumption that we are dealing with a region with no recombination events, the Sequentially Markov Coalescent Model directly addresses the role of recombination in tree construction. With recombination involved, a sequence may have two different parents, which complicates construction. The Sequentially Markov Coalescent Model tells us that moving sequentially from left to right is a simpler and much more efficient approach to analyzing the tree; the approach essentially breaks the tree into local trees and overlays them to describe recombination events.

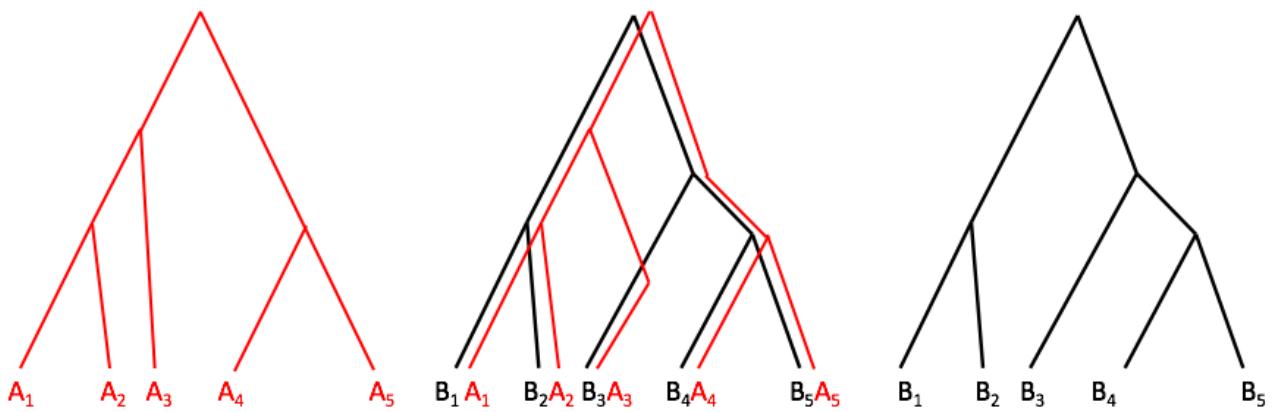


Figure 24.27: Gene trees describing ancestral relationships of two sites (A and B) on a chromosome, separately and overlaid.

A key observation to make is that every ancestor coalesces at some point, so we must find these coalescence points. First, we build a graph of all ancestral recombination events where every sequence has two parents, called an ancestral recombination graph or an ARG. We can model this graph as a process that moves from one end of the chromosome to the other end and every time there is a recombination event, there is a change in topology between the ancestral histories of two groups of alleles. This allows us split up different regions and have a single tree for each region with a single ancestor instead of trying to model a region with two ancestors. These trees are illustrated in the figure below. To transition between one topology to the next, we can perform a single subtree pruning and regrafting (SPR) operation due to the single recombination event. Using the model, we can infer the coalescence events over time.

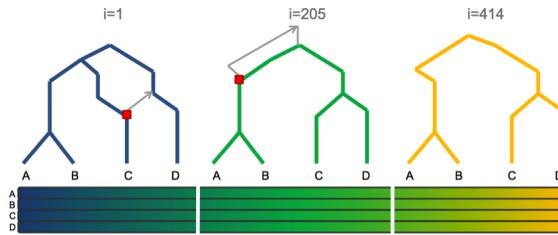


Figure 24.28: The blue, green, and yellow trees represent the regions below them.

More can be read in the following paper:

SCRIBE: Elaha  
upon intricacies  
the model itse  
<http://www.n>

## 24.7 Conclusion

Incorporating species tree information into the gene tree building process via introducing separate gene and species substitution rates allows for accurate parsimonious gene tree reconstructions. Previous gene tree reconstructions probably vastly overestimated the number of duplication and loss events. Reconstructing gene trees for large families remains a challenging problem.

## 24.8 Current Research Directions

### 24.9 Further Reading

- Paper on discovering Whole Genome Duplication event in yeast:  
<http://www.nature.com/nature/journal/v428/n6983/pdf/nature02424.pdf>

### 24.10 Tools and Techniques

### 24.11 What Have We Learned?

In this chapter, we drew conclusions regarding the relationship between gene trees and species trees. We then explored methods using gene trees to develop more accurate species trees and vice versa, involving the mutation rates of specific to both genes and species. The Wright-Fisher Model, as well as the Coalescent Model, helped us further interpret these mutation rates and understand the dynamics of allele frequencies within a population.

## Bibliography

- [1] O. Akerborg, B. Sennblad, L. Arvestad, and J. Lagergren. Bayesian gene tree reconstruction and reconciliation analysis. *Proc Natl Acad Sci*, 106(14):5714–5719, Apr 2009.
- [2] Zmasek C.M. and Eddy S.R. Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, 3(14), 2002.
- [3] Li H, Coghlan A, Ruan J, Coin LJ, Heriche JK, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, Wong GK, Zheng W, DEhal P, Wang J, and Durbin R. Treefam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res*, 34, 2006.
- [4] Arvestad L., Berglund A., Lagergren J., and Sennblad B. Bayesian gene/species tree reconciliation and orthology analysis using mcmc. *Bioinformatics*, 19 Suppl 1, 2003.
- [5] M. D. Rasmussen and M. Kellis. A bayesian approach for fast and accurate gene tree reconstruction. *Mol Biol Evol*, 28(1):273–290, Jan 2011.
- [6] Matthew D. Rasmussen and Manolis Kellis. Accurate gene-tree reconstruction by learning gene and species-specific substitution rates across multiple complete genomes. *Genome Res*, 17(12):1932–1942, Dec 2007.
- [7] Yun S. Song and Jotun Hein. Constructing minimal ancestral recombination graphs. *Journal of Computational Biology*, 12(2):147–169, March 2005.
- [8] C.E.V. Storm and E.L.L. Sonnhammer. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics*, 18(1):92–99, Jan 2002.
- [9] Hollich V., Milchert L., Arvestad L., and Sonnhammer E. Assessment of protein distance measures and tree-building methods for phylogenetic tree reconstruction. *Mol Biol Evol*, 22:2257–2264, 2005.
- [10] John Wakeley. *Coalescent theory: an introduction*. Number 575: 519.2 WAK. Roberts and Company Publishers, 1 edition, 2009.
- [11] Wapinski, I. A. Pfeffer, N. Friedman, and A. Regev. Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*, 23(13):i549–i558, 2007.

---

CHAPTER  
**TWENTYFIVE**

---

POPULATION HISTORY

Guest Lecture by David Reich  
Scribed by Deena Wang (2013) Brian Cass (2010)  
Layla Barkal and Matt Edwards (2009)

**Figures**

---

25.1 Similarity between two subpopulations can be measured by comparing allele frequencies in a scatterplot. The plots show the relative dissimilarity of European American and American Indian populations along with greater similarity of European American and Chinese populations. . . . .	405
25.2 Populations can be projected onto the principal components of other populations: South Asians projected onto Chinese and European principal components produces a linear effect (the India Cline), while Europeans projected onto South Asian and Chinese principal components does not. . . . .	407
25.3 An admixture graph that fits Indian history . . . . .	407
25.4 Projection onto two dimensions of a principle component analysis of different human populations. . . . .	409
25.5 Data and models for ancestral gene flow. . . . .	410
25.6 Ancient Europeans projected onto the two dimensional PCP of all modern European populations. The modern Western Europeans, represented primarily by the bottom left cline, cannot be described as a mixture of only EEF and WHG populations. However, with the addition of an ANE component, the variations can be explained. . . . .	411
25.7 European genetic composition over time shows two massive migrations: first, the migration of the EEF population, almost completely replacing the native WHG population; and second, the migration of the ANE Yamnaya population, replacing about 75% of the native population at that point. . . . .	412
25.8 Height selection in European populations from 8000 years ago to the present. . . . .	413

---

## 25.1 Introduction

Humans share 99.9% of the same genetic information, and are 99% similar to chimpanzees. Although humans have less genetic diversity than many other species [6], polymorphisms in populations can nonetheless lead to differences in disease risk. Learning about the 0.1% difference between humans can be used to understand population history, trace lineages, predict disease, and analyze natural selection trends.

In this lecture, Dr. David Reich of Harvard Medical School describes three historic examples of gene flow between human populations: gene flow between Africans and Europeans due to the slave trade, Indian intermixing due to migration, and interbreeding between Neanderthals, Denisovans and modern humans of Western Eurasian decent.

## 25.2 Quick Survey of Human Genetic Variation

In the human genome, there is generally a polymorphism every 1000 bases, though there are regions of the genome where this rate can quadruple. These Single Nucleotide Polymorphisms (SNPs) are one manifestation of genetic variation. When SNPs occur, they segregate according to recombination rates, advantages or disadvantages of the mutation, and the population structure that exists and continues during the lifespan of the SNP. Following a genetic mixing event, for example, one initially sees entire chromosomes, or close to entire chromosomes, coming from each constituent. As generations pass, recombination splits the SNP haplotype blocks into smaller pieces. The rate of change of the length of these blocks, then, is dependent on the rate of recombination and the stability of the recombination product. Therefore, the length of conserved haplotypes can be used to infer the age of a mutation or its selection. An important consideration, however, is that the rate of recombination is not uniform across the genome; rather, there are recombination “hot spots” that can skew the measure of haplotype age or selectivity. This makes the haplotype blocks longer than expected under a uniform model.

Every place in the genome can be thought of as a tree when compared across individuals. Depending on where you look within the genome, one tree will be different than another tree you may get from a specific set of SNPs. The trick is to use the data that we have available on SNPs to infer the underlying trees, and then the overarching phylogenetic relationships. For example, the Y chromosome undergoes little to no recombination and thus can produce a highly accurate tree as it passed down from father to son. Likewise, we can look at mitochondrial DNA passed down from mother to child. While these trees can have high accuracy, other autosomal trees are confounded with recombination, and thus show lower accuracy to predict phylogenetic relationships. Gene trees are best made by looking at areas of low recombination, as recombination mixes trees. In general, there are about 1 to 2 recombinations per generation.

Humans show about 10,000 base-pairs of linkage, as we go back about 10,000 generations. Fruit fly linkage equilibrium blocks, on the other hand, are only a few hundred bases. Fixation of an allele will occur over time, proportional to the size of the population. For a population of about 10,000, it will take about 10,000 years to reach that point. When a population grows, the effect of gene drift is reduced. Curiously enough, the variation in humans looks like what would have been formed in a population size of 10,000.

If long haplotypes are mapped to genetic trees, approximately half of the depth is on the first branch; most morphology changes are deep in the tree because there was more time to mutate. One simple model of mutation without natural selection is the Wright-Fisher neutral model which utilizes binomial sampling. In this model, a SNP will either reach fixation (frequency 1) or die out (frequency 0).

In the human genome, there are 10-20 million common SNPs. This is less diversity than chimpanzees, implying that humans are genetically closer to one another.

With this genetic similarity in mind, comparing human sub-populations can give information about common ancestors and suggest historical events. The similarity between two sub-populations can be measured by comparing allele frequencies in a scatter plot. If we plot the frequencies of SNPs across different populations on a scatterplot, we see more spread between more distant populations. The plot below, for example, shows the relative dissimilarity of European American and American Indian populations along with the greater similarity of European American and Chinese populations. The plots indicate that there was a divergence in the past between Chinese and Native Americans, evidence for the North American migration bottleneck that has been hypothesized by archaeologists. The spread among different populations within Africa is quite large. We can measure spread by the fixation index ( $F_{st}$ ) which describes the variance.

Several current studies have shown that unsupervised clustering of genetic data can recover self-selected labels of ethnic identity.<sup>[8]</sup> Rosenberg's experiment used a Bayesian clustering algorithm. They took a sample size of 1000 people (50 populations, 20 people per population), and clustered those people by their SNP genetic data, but they did not tag any of the people with their population, so they could see how the algorithm would cluster without knowledge of ethnicity. They tried many different numbers of clusters to find the optimal number. With 2 clusters, East-Asians and non-East-Asians were separated. With 3 clusters, Africans were separated from everyone else. With 4, East-Asians and Native Americans were separated. With 5, the smaller sub-populations began to emerge.

When waves of humans left Africa, genetic diversity decreased; the small numbers of people in the groups that left Africa allowed for serial founder events to occur. These serial founder events lead to the formation of sub-populations with less genetic diversity. This founder effect is demonstrated by the fact that genetic

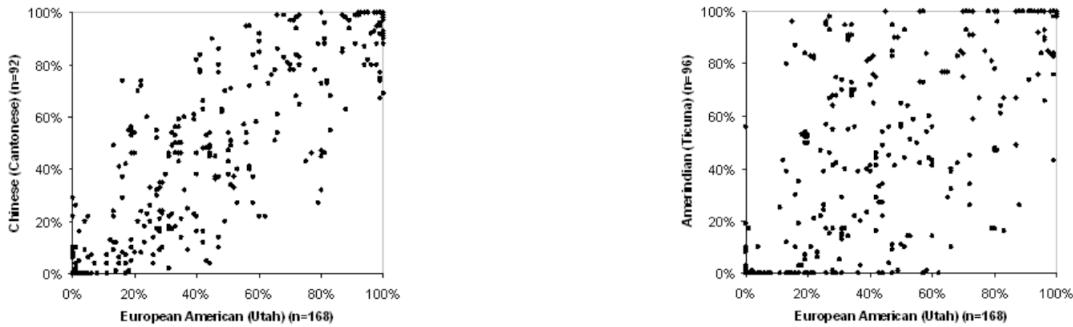


Figure 25.1: Similarity between two subpopulations can be measured by comparing allele frequencies in a scatterplot. The plots show the relative dissimilarity of European American and American Indian populations along with greater similarity of European American and Chinese populations.

diversity decreases moving out of Africa and that West Africans have the highest diversity of any human sub-population.

### 25.3 African and European Gene Flow

The Atlantic Slave Trade took place from the 16th century to the 19th century, and moved about 5 million people from Africa to the Americas. Most African-Americans today have a mixture of 80% African and 20% European heritage. When two parents of different ethnicities have children, their children will inherit one chromosome from each parent, and their grandchildren will inherit chromosomes that are a mosaic of the two ethnicities due to recombination. As time passes, the increasing number of recombination events will decrease the length of the “African” or “European” stretches of DNA.

Recombination events are not spread evenly throughout the chromosomes, but happen at hotspots. African and European DNA have different hot spots, which could be due to differences in the amino acid composition of *PRDM9*, a histone H3(K4) trimethyltransferase which is essential for meiosis.

Difference in disease susceptibility can be predicted for African and European populations. With sequencing, this knowledge can also be applied to mixed populations. For example, Africans have a higher risk of prostate cancer which is directly linked to an area in chromosome 8 that maps to a cancer proto-oncogene[4]. If a mixed individual has the African sequence in that area, he or she will have the increased risk, but if the individual has the European sequence, he or she will not have an increased risk. The same approach can be applied to breast cancer, colon cancer, multiple sclerosis, and other diseases.

### 25.4 Gene Flow on the Indian Subcontinent

Genetic evidence suggests that modern populations on the Indian subcontinent descended from two different ancestral populations that mingled 4,000 years ago. SNP array data was collected from about 500 different people from 73 Indian groups with different language families [5]. A principle component analysis plot reveals that the Dravidian/Indo-European language groups and the Austro-Asiatic language groups are in two different clusters, which suggests they have different lineages. Within the Dravidian/Indo-European language groups, there is a gradient of relatedness to West Eurasian groups.

The same mosaic technique used in the African/European intermixing study was used to estimate the date of mixture. The Indian population is a mixture of a Central Asian/European group and another group most closely related to the people of the Andaman Islands. The chunk size of the DNA belonging to each group suggests a mixture about 100 generations old, or 2,000 to 4,000 years ago. Many groups have this mixed heritage, but mixture stops after the creation of the caste system.

Knowledge of the heritage of genes can predict diseases. For example, a South Asian mutation in myosin binding protein C causes a seven-fold increase in heart failure. Many ethnic groups are endogamous and have a low genetic diversity, resulting in a higher prevalence of recessive diseases.

Past surveys in India have studied such aspects as anthropometric variation, mtDNA, and the Y chromosome. The anthropometric study looked at significant differences in physical characteristics between groups separated by geography and ethnicity. The results showed variation much higher than that of Europe. The mtDNA study was a survey of maternal lineage and the results suggested that there was a single Indian tree such that age of lineage could be inferred by the number of mutations. The data also showed that Indian populations were separated from non-Indian populations at least 40,000 years ago. Finally, the Y chromosome study looked at paternal lineage and showed a more recent similarity to Middle Eastern men and dependencies on geography and caste. This data conflicts with the mtDNA results. One possible explanation is that there was a more recent male migration. Either way, the genetic studies done in India have served to show its genetic complexity. The high genetic variation, dissimilarity with other samples, and difficulty of obtaining more samples lead to India being left out of HapMap, the 1000 Genomes Project, and the HGDP.

In David Reich and collaborators study of India, 25 Indian groups were chosen to represent various geographies, language roots, and ethnicities. The raw data included five samples for each of the twenty five groups. Even though this number seems small, the number of SNPs from each sample has a lot of information. Approximately five hundred thousand markers were genotyped per individual. Looking at the data to emerge from the study, if Principal Components Analysis is used on data from West Eurasians and Asians, and if the Indian populations are compared using the same components, the ‘India Cline’ emerges. This shows a gradient of similarity that might indicate a staggered divergence of Indian populations and European populations.

#### 25.4.1 Almost All Mainland Indian Groups are Mixed

Further analysis of the ‘India Cline’ phenomenon produces interesting results. For instance, some Pakistani sub-populations have ancestry that also falls along the ‘India Cline’. Populations can be projected onto the principal components of other populations: South Asians projected onto Chinese and European principal components produces a linear effect (the India Cline), while Europeans projected onto South Asian and Chinese principal components does not. One interpretation is that Indian ancestry shows more variability than the other groups. A similar variability assessment appears when comparing African to non-African populations. Two tree hypotheses emerge from this analysis:

1. there were serial founder events in India’s history or
2. there was gene flow between ancestral populations.

The authors developed a formal “four population test” to test ancestry hypotheses in the presence of admixture or other confounding effects. The test takes a proposed tree topology and sums over all SNPs of  $(Pp_1 - Pp_2)(Pp_3 - Pp_4)$ , where P values are frequencies for the four populations. If the proposed tree is correct, the correlation will be 0 and the populations in question form a clade. This method is resistant to several problems that limit other models. A complete model can be built to fit history. The topology information from the admixture graphs can be augmented with  $F_{st}$  values through a fitting procedure. This method makes no assumptions about population split times, expansion and contractions, and duration of gene flow, resulting in a more robust estimation procedure.

Furthermore, estimating the mixture proportions using the 4 population statistic gives error estimates for each of the groups on the tree. Complicated history does not factor into this calculation, as long as the topology as determined by the 4-population test is valid.

These tests and the cline analysis allowed the authors to determine the relative strength of Ancestral North Indian and Ancestral South Indian ancestry in each representative population sample. They found that high Ancestral North Indian ancestry is correlated with traditionally higher caste and certain language groupings. Furthermore, Ancestral North Indian (ANI) and South Indian (ASI) ancestry is as different from Chinese as European.

#### 25.4.2 Population structure in India is different from Europe

Population structure in India is much less correlated with geography than in Europe. Even correcting populations for language, geographic, and social status differences, the  $F_{st}$  value is 0.007, about 7 times that

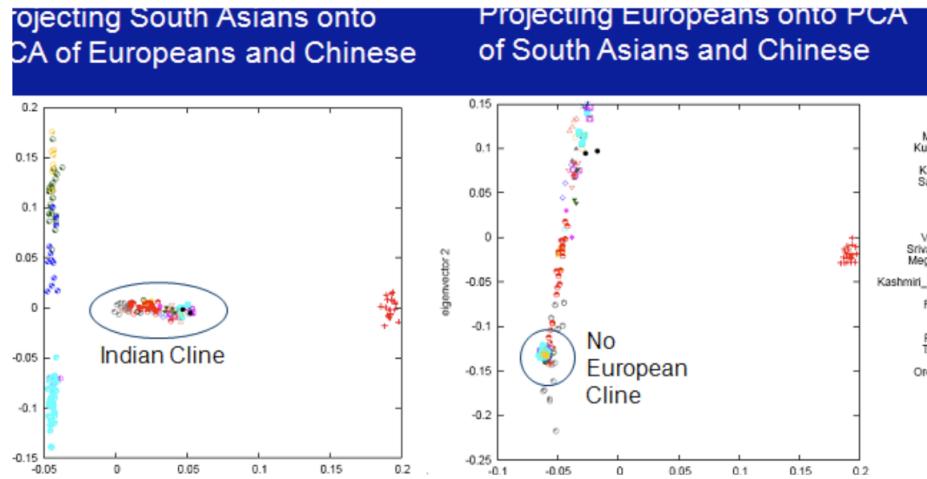


Figure 25.2: Populations can be projected onto the principal components of other populations: South Asians projected onto Chinese and European principal components produces a linear effect (the India Cline), while Europeans projected onto South Asian and Chinese principal components does not.

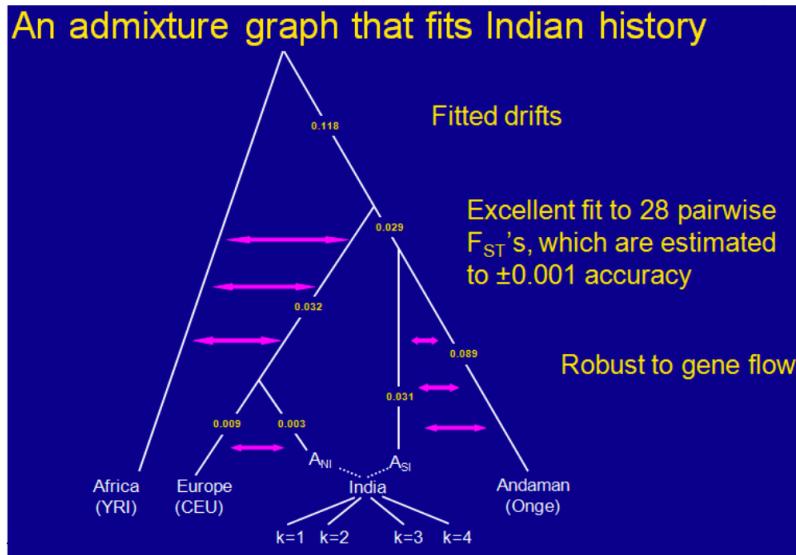


Figure 25.3: An admixture graph that fits Indian history

of the most divergent populations in Europe. An open question is whether this could be due to missing (largely India-specific) SNPs on the genotyping arrays. This is because the set of targeted SNPs were identified primarily from the HapMap project, which did not include Indian sources.

Most Indian genetic variation does not arise from events outside India. Additionally, consanguineous marriages cannot explain the signal. Many serial founder events, perhaps tied to the castes or precursor groups, could contribute. Analyzing a single group at a time, it becomes apparent that castes and subcastes have a lot of endogamy. The autocorrelation of allele sharing between pairs of samples within a group is used to determine whether a founder event occurred and its relative age. There are segments of DNA from a founder, many indicating events more than 1000 years old. In most groups there is evidence for a strong, ancient founder event and subsequent endogamy. This stands in contrast to the population structure in most of Europe or Africa, where more population mixing occurs (less endogamy).

These serial founder events and their resulting structure have important medical implications. The strong founder events followed by endogamy and some mixing have lead to groups that have strong propensities for

various recessive diseases. This structure means that Indian groups have a collection of prevalent diseases, similar to those already known in other groups, such as Ashkenazi Jews or Finns. Unique variation within India means that linkages to disease alleles prevalent in India might not be discoverable using only non-Indian data sources. A small number of samples are needed from each group, and more groups, to better map these recessive diseases. These maps can then be used to better predict disease patterns in India.

### 25.4.3 Discussion

Overall, strong founder events followed by endogamy have given India more substructure than Europe. All surveyed tribal and caste groups show a strong mixing of ANI and ASI ancestry, varying between 35% and 75% ANI identity. Estimating the time and mechanism of the ANI-ASI mixture is currently a high priority. Additionally, future studies will determine whether and how new techniques like the 4-population test and admixture graphs can be applied to other populations.

## 25.5 Gene Flow Between Archaic Human Populations

Dr. Reich worked with the Max Planck Institute as a population geneticist studying Neanderthal genetic data. This section will discuss the background of his research as part of the Neanderthal genome project, the draft sequence that they assembled, and the evidence that has been compiled for gene flow between modern humans and Neanderthals.

### 25.5.1 Background

Neanderthals are the only other hominid with a brain as large as *Homo sapiens*. Neanderthal fossils from 200,000 years ago have been found in West Eurasia (Europe and Western Asia), which is far earlier than *Homo erectus*. The earliest human fossils come from Ethiopia dating about 200,000 years ago. However, there is evidence that Neanderthals and humans overlapped in time and space between 135,000 and 35,000 years ago.

The first place of contact could have occurred in The Levant, in Israel. There are human fossils from 120,000 years ago, then a gap, Neanderthal fossils about 80,000 years ago, another gap, and then human fossils again 60,000 years ago. This is proof of an overlap in place, but not in time. In the upper paleolithic era, there was an explosion of populations leaving Africa (the migration about 60,000 to 45,000 years ago). In Europe after 45,000 years ago, there are sites where Neanderthals and humans exist side by side in the fossil record. Since there is evidence that the two species co-existed, was there interbreeding? This is a question that can be answered by examining population genomics.

See Tools and Techniques for a discussion of DNA extraction from Neanderthals.

### 25.5.2 Evidence of Gene Flow between Humans and Neanderthals

1. A comparison test between Neanderthal DNA and human DNA from African and non-African populations demonstrates that non-African populations are more related to Neanderthals than African populations. We can look at all the SNPs in the genome and see whether the human SNP from one population matches the Neanderthal SNP. When different human populations were compared to Neanderthals, it was found that French, Chinese, and New Guinea SNPs matched Neanderthal SNPs much more than Nigerian Yoruba SNPs matched Neanderthal SNPs. San Bushmen and Yoruba populations from Africa, despite being very distinct genetically, both had the same distance from Neanderthal DNA. This evidence suggests that human populations migrating from Africa interbred with Neanderthals.
2. A long-range haplotype study demonstrates that when the deepest branch of a haplotype tree was in non-African populations, the regions frequently matched Neanderthal DNA. African populations today are the most diverse populations in the world. When humans migrated out of Africa, diversity decreased due to the founder effect. From this history, one would expect that if you built a tree of relations, the deepest split would be African.

To show Neanderthal heritage, Berkley researchers picked long range sections of the genome and compared them among randomly chosen humans from various populations. The deepest branch of the tree constructed from that haplotype is almost always from the African population. However, occasionally non-Africans have the deepest branch. The study found that there were 12 regions where non-Africans have the deepest branch. When this data was used to analyze the Neanderthal genome, it was found that 10 out of 12 of these regions in non-Africans matched Neanderthals more than they matched the human reference sequence (a compilation of sequences from various populations). This is evidence of that haplotype actually being of Neanderthal origin.

3. Lastly, there is a bigger divergence than expected among humans. The average split between a Neanderthal and a human is about 800,000 years. The typical divergence between two humans is about 500,000 years. When looking at African and non-African sequences, regions of low divergence emerged in non-African sequences when compared with Neanderthal material. The regions found were highly enriched for Neanderthal material (94% Neanderthal), which would increase the average divergence between humans (as the standard Neanderthal - human divergence is about 800,000 years).

### 25.5.3 Gene Flow between Humans and Denisovans

In 2010, scientists discovered a 50,000 year old finger bone in southern Siberia. The DNA in this Denisovan sample was not like any previous human DNA. Denisovan mitochondrial DNA is an out-group to both Neanderthals and modern humans. (Mitochondrial DNA was used because it is about 1000 times more frequent than somatic DNA. The polymorphism rate is also 10 times higher.) Denisovans are more closely related to Neanderthals than humans.

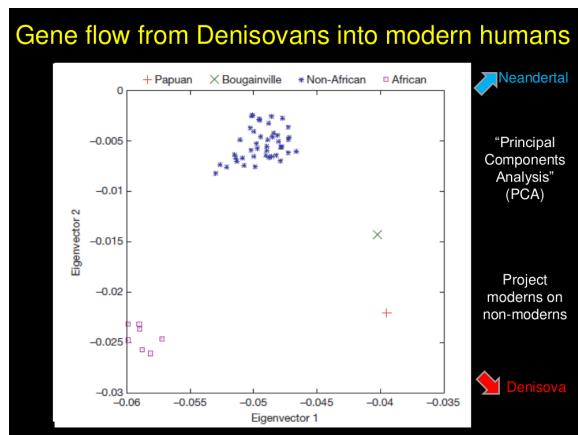


Figure 25.4: Projection onto two dimensions of a principle component analysis of different human populations.

Using the same SNP matching technique from the Neanderthal example, it was discovered that Denisovan DNA matches New Guinean DNA more than Chinese DNA or European DNA. It is estimated that Denisovans contributed about 5% of the ancestry of New Guineans today. A principle component analysis projection (see figure) between relatedness to chimpanzees, Neanderthals, and Denisovans shows that non-African populations are more related to Neanderthals, and New Guinean/Bougainvillians are more related to Denisovans.

This evidence suggests a model for human migration and interbreeding. Humans migrated out of Africa and interbred with Neanderthals, then spread across Asia and interbred with Denisovans in Southeast Asia. It is less plausible that humans interbred with Denisovans in India because not all of the populations in Southeast Asia have Denisovan ancestry.

### 25.5.4 Analysis of High Coverage Archaic Genomes

High-coverage archaic genomes can tell us a lot about the history of hominid populations. A high coverage Altai Neanderthal sequence was acquired from a toe bone found in Denisova cave. From this sequence, we

can look at the time to convergence of the two copies of chromosomes to estimate the size of the population. Neanderthal DNA contains many long stretches of homozygosity, indicating a persistent small population size and inbreeding. For the Altai Neanderthal, one eighth of the genome was homozygous, about the expected level of inbreeding of half-siblings. Applying the technique to non-African populations shows a bottleneck 50,000 years ago and a subsequent population expansion, which is consistent with the Out Of Africa theory.

Neanderthals and Denisovans also interbred, demonstrating the remarkable proclivity of humanoids towards reproduction. Although most of the Neanderthal genome has a minimum depth of hundreds of thousands of years from the Denisovan genome, at least 0.5% of the Denisovan genome has a much shorter distance from Neanderthal genome, especially for immune genes.

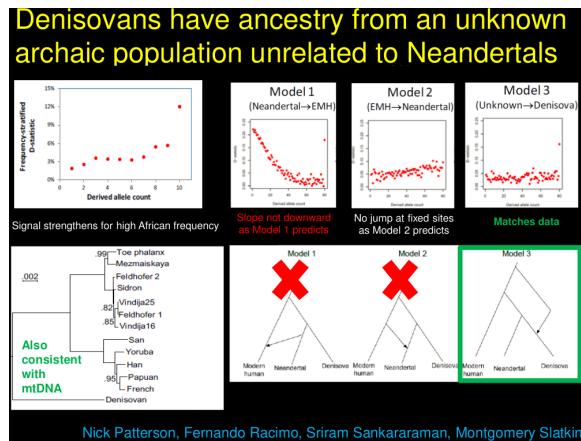


Figure 25.5: Data and models for ancestral gene flow.

Denisovans most likely have ancestry from an unknown archaic population unrelated to Neanderthals. An African sequence has a 23% match with Neanderthal DNA and 47% match with Denisovan DNA, which is statistically significant. If you stratify the D-statistic by the frequency of an allele in the population, you see an increasing slope and a sharp jump when you reach fixation which most closely matches the predictions one would obtain from an unknown population flowing into Denisovans (see figure).

### 25.5.5 Discussion

The bottleneck caused by the migration from Africa is only one example of many. Most scientists usually concentrate on the age and intensity of migration events and not necessarily the duration, but the duration is very important because long bottlenecks create a smaller range of diversity. One way to predict the length of a bottleneck is to determine if any new variations arose during it, which is more likely during longer bottlenecks. The change in the range of diversity is also what helped create the different human sub-populations that became geographically isolated. This is just another way that population genetics can be useful for helping to piece together historical migrations.

Genetic differences between species (here within primates) can be used to help understand the phylogenetic tree from which we are all derived. We looked at the case study of comparisons with Neanderthal DNA, learned about how ancient DNA samples are obtained, how sequences are found and interpreted, and how that evidence shows high likelihood of interbreeding between modern humans (of Eurasian descent) and Neanderthals. Those very small differences between one species and the next, and within species, allow us to deduce a great deal of human history through population genetics.

## 25.6 European Ancestry and Migrations

### 25.6.1 Tracing the Origins of European Genetics

Before 2014, it was believed that modern European genetics was primarily a mixture of two ancestral populations. The first population is what is known as the Western hunter-gatherer (WHG) population, and

is considered the indigenous European population. The second population is known as the Early European farmer (EEF) population, and represents the rapid migration of farming peoples into Europe, and the subsequent mixing of the new farming population with the original WHG population. However, in 2012, Patterson et al [9] used the principal component analysis in Figure 25.6 to show that European genetics does not match up with being a mixture of only these two populations. Rather, genetic mixture analysis showed that some Europeans could only be explained as a mix of EEF/WHG populations with a third population whose genetics resembled Native Americans. While this does not mean that Native Americans are ancestral to Europeans, the study concluded that the most likely hypothesis was the mixture of these two known populations with an Ancient North Eurasian (ANE) population which migrated to both Asia and Europe, and is no longer found in North Eurasia. This study called this mystery population the "Ghost of North Eurasia."

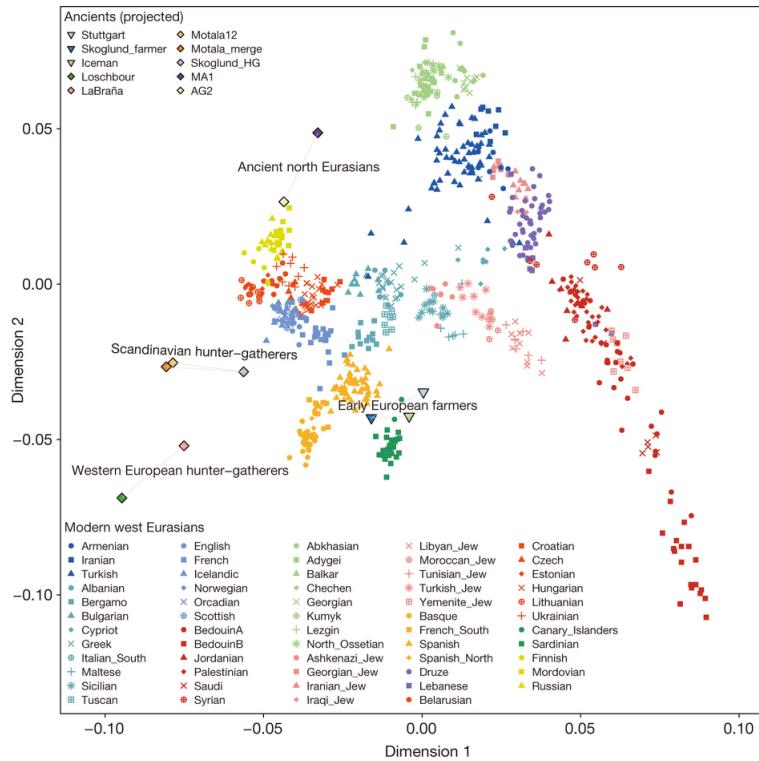


Figure 25.6: Ancient Europeans projected onto the two dimensional PCP of all modern European populations. The modern Western Europeans, represented primarily by the bottom left cline, cannot be described as a mixture of only EEF and WHG populations. However, with the addition of an ANE component, the variations can be explained.

Two years later, in 2014, however, a sample was found confirming the existence of this population. Proclaiming that "The ghost is Found", Raghavan, Skoglund et al. [10] studied the newly found "Mal'ta" sample from Lake Baikal (currently in Southern Russia) and determined that it matched the predicted ghost population from 2012, and could explain the two dimensional variation in modern European populations. In particular, modern Europeans were found to be composed of 0-50% WHG, 32-93% EEF, and 1-18% ANE populations.

### 25.6.2 Migration from the Steppe

Given this new population as a source of European ancestry, the natural questions are when and why did the members of the ANE population migrate to Europe? The answer, of course, can be teased out of further genetic data about the history of European populations. The first clue was found in mitochondrial DNA data, in a 2013 paper by Brandt, Haak et al. [3], which found that there were two discontinuities in

European mitochondrial DNA: one between the Mesolithic and the early Neolithic ages, and one between the mid Neolithic age and the Late Neolithic and Bronze ages. In 2014, studies of 9, and then 94 samples of ancient European individuals showed clearly the two migration events, visualized in Figure 25.7. The first migration, at roughly 6500 BCE, was a migration of the EEF population, which replaced the existing WHG population at a rate of between 60 and 100%. The second migration was a migration of steppe pastoralists, known as the Yamnaya, which replaced the existing population with a rate between 60 and 80%. In both cases, the migrating population takes over a chunk of the genetic composition almost immediately, and then the previous population gradually resurges over several thousand years.

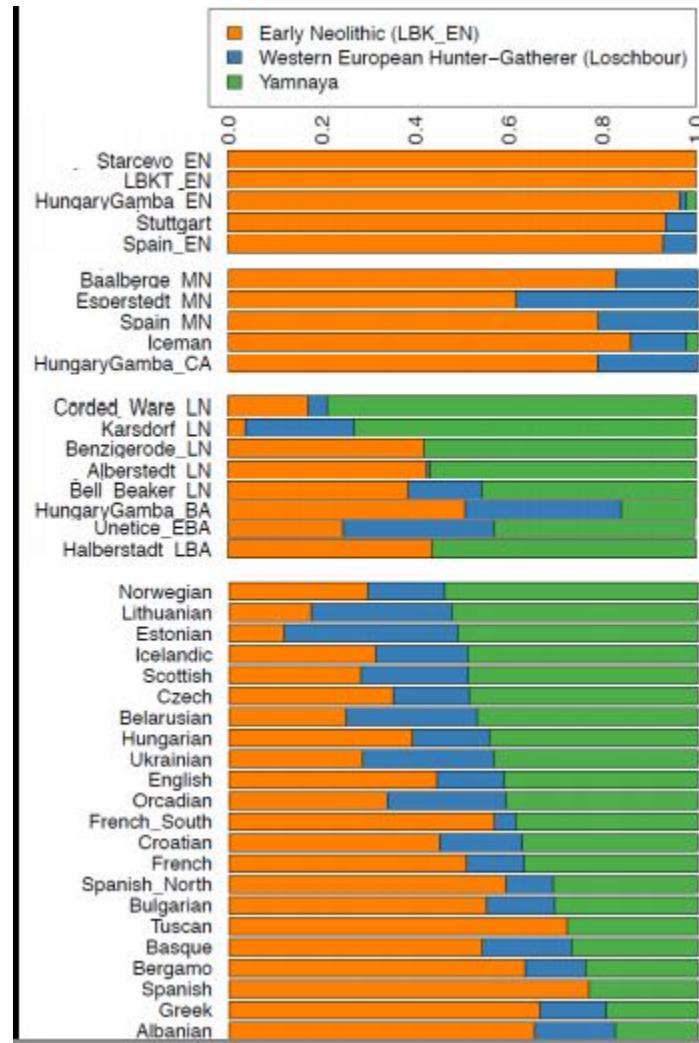


Figure 25.7: European genetic composition over time shows two massive migrations: first, the migration of the EEF population, almost completely replacing the native WHG population; and second, the migration of the ANE Yamnaya population, replacing about 75% of the native population at that point.

### 25.6.3 Screening for Natural Selection

Another application of DNA data to history is in tracing natural selection events. Essentially, one can look at the frequencies of various alleles in modern European DNA data, and find cases where it does not match the ancestral mixing model of the population. Such cases will tend to signify alleles that have been selected for or against since the ancestral mixing events occurred. The easiest to identify, and most well known example of such a trait is lactase persistence. The current level of prevalence of this trait is well above any

of the levels represented by ancestral populations, suggesting that it underwent positive selection (due to the domestication and milking of animals) since the ancestral mixing events.

Several other traits can also be detected as candidates for selection. Another straightforward example is skin pigmentation. More interesting is the tale of height selection shown by the genetics of Northern and Southern Europeans. In particular, the data shows that two distinct selection effects occurred. First, the early farmers of Southern Europe underwent selection for decreasing height between 8000 and 4000 years ago. Second, the peoples of Northern Europe (modern Scandinavians, etc.) underwent positive selection around the same time period and through the present. While the anthropological explanations of these effects are disputed, the effects themselves are shown clearly in the genetic data.

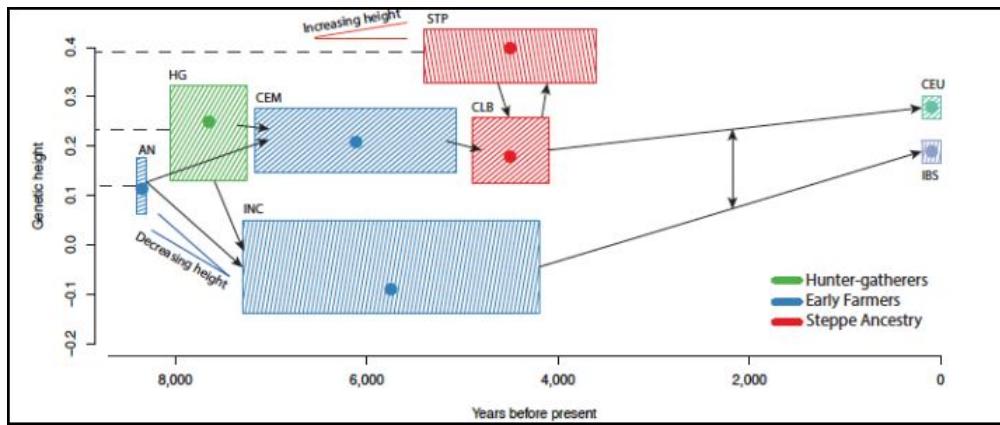


Figure 25.8: Height selection in European populations from 8000 years ago to the present.

## 25.7 Tools and Techniques

### 25.7.1 Techniques for Studying Population Relationships

There are several different methods for studying population relationships with genetic data. The first general type of study utilizes both phylogeny and migration data. It fits the phylogenies to  $F_{st}$  values, values of sub-population heterozygosity (pioneered by Cavalli-Sforza and Edwards in 1967 [2]). This method also makes use of ‘synthetic maps’ and Principal Components Analysis. [7] The primary downside to analyzing population data this way is uncertainty about results. There are mathematical and edge effects in the data processing that cannot be predicted. Also, certain groups have shown that separate, bounded mixing populations can produce significant-seeming principal components by chance. Even if the results of the study are correct, then, they are also uncertain.

The second method of analyzing sub-population relationships is genetic clustering. Clusters can be formed using self-defined ancestry [1] or the STRUCTURE database. [8] This method is overused and can over-fit the data; the composition of the database can bias the clustering results.

Technological advances and increased data collection, though, have produced data sets that are 10,000 times larger than before, meaning that most specific claims can be disproved by some subset of data. So in effect, many models that are predicted either by phylogeny and migration or genetic clustering will be disproved at some point, leading to large-scale confusion of results. One solution to this problem is to use a simple model that makes a statement that is both useful and has less probability of being falsified.

### 25.7.2 Extracting DNA from Neanderthal Bones

Let’s take a look at how you go about finding and sequencing DNA from ancient remains. First, you have to obtain a bone sample with DNA from a Neanderthal. Human DNA and Neanderthal DNA is very similar (we are more similar to them than we are to chimps), so when sequencing short reads with very old DNA, it is impossible to tell if the DNA is Neanderthal or human. The cave where the bones were found is first

classified as human or non-human using trash or tools as an identifier, which helps predict the origin of the bones. Even if you have a bone, it is still very unlikely that you have any salvageable DNA. In fact, 99% of the sequence of Neanderthals comes from only three long bones found in one site: the Vindija cave in Croatia (5.3 Gb, 1.3x full coverage).

Next, the DNA is sent to an ancient-DNA lab. Since they are 40,000 year old bones, there is very little DNA left in them. So, they are first screened for DNA. If they find DNA, the next question is whether it is primate DNA? Usually it is DNA from microbes and fungi that live in soil and digest dead organisms. Only about 1-10% of the DNA on old bones is the primate's DNA. If it is primate DNA, is it contamination from the human (archeologist or lab tech) handling it? Only one out of 600 bp are different between humans and Neanderthals DNA. The size of reads from a 40,000 year old bone sample is 30-40 bp. The reads are almost always identical for a human and Neanderthal, so it is difficult to distinguish them.

In one instance, 89 DNA extracts were screened for Neanderthals DNA, but only 6 bones were actually sequenced (requires lack of contamination and high enough amount of DNA). The process of retrieving the DNA requires drilling beneath the bone surface (to minimize contamination) and taking samples from within. For the three long bones, less than 1 gram of bone powder was able to be obtained. Then the DNA is sequenced and aligned to a reference chimp genome. It is mapped to a chimp instead of a particular human because mapping to a human might cause bias if you are looking to see how the sequence relates to specific human sub-populations.

Most successful finds have been in cool limestone caves, where it is dry and cold and perhaps a bit basic. The best chance of preservation occurs in permafrost areas. Very little DNA is recoverable from the tropics. The tropics have a great fossil record, but DNA is much harder to obtain. Since most bones don't yield enough or good DNA, scientists have the screen samples over and over again until they eventually find a good one.

### 25.7.3 Reassembling Ancient DNA

DNA extracted from Neanderthal bones have short reads, about 37 bp on average. There are lots of holes due to mutations caused by time eroding the DNA. It is difficult to tell whether a sequence is the result of contamination because humans and Neanderthals only differ in one out of one thousand bases. However, we can use DNA damage characteristic of ancient DNA to distinguish old and new DNA. Old DNA has a tendency towards C to T and G to A errors. The C to T error is by far the most common, and is seen about 2% of the time. Over time, a methyl group gets knocked off of a C, which causes it to resemble to U. When PCR is used to amplify the DNA for sequencing, the polymerase sees a U and repairs it to a T. In order to combat this error, scientists use a special enzyme that recognizes the U, and cuts the strand instead of replacing it with a T. This helps to identify those sites. The G to A mutations are the result of seeing that on the opposite strand.

The average fragment size is quite small, and the error rate is still 0.1% - 0.3%. One way to combat the mutations is to note that on a double stranded fragment, the DNA is frayed towards the ends, where it becomes single stranded for about 10 bp. There tend to be high rates of mutations in the first and last 10 bases, but high quality DNA elsewhere, i.e. more C to T mutations in the beginning and G to A in the end. In chimps, the most common mutations are transitions (purine to purine, pyrimidine to pyrimidine), and transversions are much rarer. The same goes for humans. Since the G to A and C to T mutations are transitions, it can be determined that there are about 4x more mutations in the old Neanderthal DNA than if it were fresh by noting the number of transitions seen compared to the number of transversions seen (by comparing Neanderthal to human DNA). Transversions have a fairly stable rate of occurrence, so that ratio helps determine how much error has occurred through C to T mutations.

We are now able to get human contamination of artifact DNA down to around  $\pm 1\%$ . When the DNA is brought in, as soon as it is removed from the bone it is bar coded with a 7 bp tag. That tag allows you to avoid contamination at any later point in the experiment, but not earlier. Extraction is also done in a clean room with UV light, after having washed the bone. Mitochondrial DNA is helpful for distinguishing what percent of the sample is contaminated with human DNA. Mitochondrial DNA is filled with characteristic event sites because humans and Neanderthals are reciprocally monophylogenetic. The contamination can be measured by counting the ratio of those sites. In the Neanderthal DNA, contamination was present, but it was  $\pm 0.5\%$ .

In sequencing, the error rate is almost always higher than the polymorphism rate. Therefore, most sites in the sequence that are different from humans are caused by sequencing errors. So we can't exactly learn about Neanderthal biology through the sequence generated, but we can analyze particular SNPs as long as we know where to look. The probability of a particular SNP being changed due to an error in sequencing is only  $\frac{1}{300}$  to 11000, so usable data can still be obtained.

After aligning the chimp, Neanderthal, and modern human sequences, we can measure the distance from Neanderthals to humans and chimps. This distance is only about 12.7% from the human reference sequence. A French sample measures about 8% distance from the reference sequence, and a Bushman about 10.3%. What this says is that the Neanderthal DNA is within our range of variation as a species.

## 25.8 Research Directions

Currently, the most exciting trend in the field is the existence of more and more data on both ancient and modern population genetics. With more samples, we can devise more fine statistical tests, and tease more and more information about population composition and history.

## 25.9 Further Reading

### Bibliography

- [1] Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, and Cavalli-Sforza LL. High resolution of human evolutionary history trees with polymorphic microsatellites. *Nature*, 368:455–457, 1994.
- [2] Luigi L. Cavalli-Sforza and Anthony WF Edwards. Phylogenetic analysis. models and estimation procedures. *American Journal of Human Genetics*, 19.3:233, 1967.
- [3] Clio Der Sarkissian, Oleg Balanovsky, Guido Brandt, Valery Khartanovich, Alexandra Buzhilova, Sergey Koshel, Valery Zaporozhchenko, Detlef Gronenborn, Vyacheslav Moiseyev, Eugen Kolpakov, et al. Ancient dna reveals prehistoric gene-flow from siberia in the complex human population history of north east europe. *PLoS Genet*, 9(2):e1003296, 2013.
- [4] Christopher A Haiman et al. Multiple regions within 8q24 independently affect risk for prostate cancer. *Nature Genetics*, 39:638–644, 2007.
- [5] Priyasamy et al. Genetic evidence for recent population mixture in india. *The American Journal of Human Genetics*, 93:1–17, 2013.
- [6] Lynn B. Jorde and Stephen P. Wooding. Genetic variation, classification, and race. *Nature Genetics*, 36, 2004.
- [7] Menozzi. Synthetic maps of human gene frequencies in europeans. *Science*, 201(4358):768–792, Sep 1978.
- [8] Rosenberg N. Genetic structure of human populations. *Science*, 298(5602):2381–2385, 2002.
- [9] Nick Patterson, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. Ancient admixture in human history. *Genetics*, 192(3):1065–1093, 2012.
- [10] Pontus Skoglund, Helena Malmström, Maanasa Raghavan, Jan Storå, Per Hall, Eske Willerslev, M Thomas P Gilbert, Anders Götherström, and Mattias Jakobsson. Origins and genetic legacy of neolithic farmers and hunter-gatherers in europe. *Science*, 336(6080):466–469, 2012.



---

CHAPTER  
**TWENTYSIX**

---

## POPULATION GENETIC VARIATION

Guest Lecture by Pardis Sabeti

Scribed by Aasavari Phanse, Katy Johnson, Yilun Gu (2016),  
Mohammad Ghassemi, Jonas Helfer, Ben Mayne (2012),  
Alex McCauley (2010), Matthew Lee (2009), Arjun K. Manrai and Clara Chan (2008)

### Figures

---

26.1 Plot of genotype frequencies for different allele frequencies . . . . .	420
26.2 Changes in allele frequency over time . . . . .	420
26.3 A comparison of the hetrozygous and homozygous derived and damaging genotypes per individual in an African American (AA) and European American (EA) population study. . . . .	422
26.4 Two isolated populations . . . . .	423
26.5 Trio phasing . . . . .	425
26.6 Approximate Time Table of Effects Sabeti et al. <i>Science</i> 2006 . . . . .	427
26.7 Localized positive selection for Malaria resistance within species Sabeti et al. <i>Science</i> 2006 . . . . .	428
26.8 Localized positive selection for lactase persistence allele Sabeti et al. <i>Science</i> 2006 . . . . .	428
26.9 Mean allele frequency difference of height SNPs, matched SNPS, and genome-wide SNPS between Northern- and Southern-European populations Turchin et al., <i>Nature Genetics</i> (2012) . . . . .	429
26.10 Broken haplotype as a signal of natural selection . . . . .	429
26.11A depiction of two major bottleneck events, one in the founding population from Africa, and other, smaller subsequent bottleneck events in the East Asian and Western European populations. . . . .	431
26.12 An illustration of two bottleneck events . . . . .	432
26.13 The figure illustrate the effects of a bottleneck events on the number of rare Alleles in a population. . . . .	433
26.14A depiction of European admixture levels in the Mexican, and African American populations. . . . .	433
26.15 As illustration of the magnitude and origin of migrants based on the tract length and number of tracts in the admixed population. . . . .	434

---

### 26.1 Introduction

For centuries, biologists had to rely on morphological and phenotypical properties of organisms in order to infer the tree of life and make educated guesses about the evolutionary history of species. Only recently, the ability to cheaply sequence entire genomes and find patterns in them has transformed evolutionary biology. Sequencing and comparing genomes on a molecular level has become a fundamental tool that allows us to gain insight into much older evolutionary history than before, but also to understand evolution at a much

smaller resolution of time. With these new tools, we can not only learn the relationship between distant clades that separated billions of years ago, but also understand the present and recent past of species and even different populations inside a species.

In this chapter we will discuss the study of Human genetic history and recent selection. The methodological framework of this section builds largely on the concepts from previous chapters. Most specifically, the methods for association mapping of disease and phylogenetic constructs such as tree building among species and genes, and the history of mutations using coalescence. Having learned about these methods in the last chapter, we now will study how their application can inform us about the relationships, and differences between human populations. Additionally, we will look for how these differences can be exploited to look for signals of recent natural selection and the identification of disease loci. We will also discuss in this chapter what we currently know about the differences between human populations and describe some parameters we can infer that quantify population differences, using only the extent genetic variation we observe. In the study of Human Genetic history and recent selection, there are two principal topics of investigation which are often studied. The first is the history of population sizes. The second is the history of interactions between populations. Questions are often asked about these areas because the answers can often provide knowledge to improve the disease mapping process. Thus far, all present research based knowledge of human history was found by investigating functionally neutral regions of the genome, and assuming genetic drift. The reason that neural regions are employed is because mutations are subject to positive, negative and balancing selection pressure, when they take place on a functional region. Hence investigating a neural regions provides a selection unbiased proxy for the drift between species. In this chapter we will delve into some of the characteristics of selection process in humans and look for patterns of human variation in terms of cross species comparisons, comparison synonymous and non-synonymous mutations, and haplotype structure.

## 26.2 Population Selection Basics

### 26.2.1 Polymorphisms

**Polymorphisms** are differences in appearance amongst members of the same species. Many of them arise from mutations in the genome. These mutations constitute the 1% difference between individuals representing the predisposition differences. The remaining variation is due to environmental/stochastic differences. These mutations, or genetic polymorphisms, can be characterized into different types.

#### Single Nucleotide Polymorphisms (SNPs)

- The mutation of only a single nucleotide base within a sequence. In most cases, these changes are without consequence. However, there are some cases where the mutation of a single nucleotide has a major effect.
- For example, **Sickle Cell Anemia** is caused by a **single nucleotide polymorphism** from A to T, that causes a change from glutamic acid (GAG) to valine (GTG) in hemoglobin.
- These occur 1 per 1,000 base pairs

#### Insertion/Deletion

- Through faulty copying or DNA-repair, insertions **insertions** or deletions **deletions** of one or multiple nucleotides can occur.
- If the insertion or deletion is inside an exon (the protein-coding region of a gene) and does not consist of a multiple of three nucleotides, a frameshift **frameshift** will occur.
- Prime example is deletions in the CFTTR gene, which codes for chloride channels in the lungs and may cause **Cystic Fibrosis** where the patient cannot clear mucous in the lungs and causes infection
- These occur 1 per 10,000 base pairs

## Variable Number Tandem Repeats

- When a short sequence is repeated multiple times, DNA Polymerase can sometimes "slip", causing it to make either too many or too few copies of the repeat. This is called a variable number tandem repeat**variable number tandem repeat**.
- These vary greatly in size, expanding and contracting every generation. These occur 1 per 10,000 base pairs
- For example, **Huntington's disease** that is caused by too many repeats of the trinucleotide CAG repeat in the HTT gene. Having more than 36 repeats can lead to gradual muscle control loss and severe neurological degradation. Generally, the more repeats there are, the stronger the symptoms.

## Structural Variants/Copy Number Variants

- Correspond to large (median of 5,000 base pair) deletions, duplications, and insertions. **structural variants/copy number**.
- These occur 1 per 1,000,000 base pairs
- For example, **Huntington's disease** that is caused by too many repeats of the trinucleotide CAG repeat in the HTT gene. Having more than 36 repeats can lead to gradual muscle control loss and severe neurological degradation. Generally, the more repeats there are, the stronger the symptoms.

### **Did You Know?**

DNA profiling is based on short variable number tandem repeats (STR). DNA is cut with certain restriction enzymes, resulting in fragments of variable length that can be used to identify an individual. Different countries use different (but often overlapping) loci for these profiles. In North America, a system based on 13 loci is used.

### 26.2.2 Allele and Genotype Frequencies

In order to understand the evolution of a species through analysis of alleles or genotypes, we must have a model of how the alleles are passed on from one generation to another. It is of immense importance that the reader has a firm intuition for the Hardy-Weinberg Principle and Wright Fisher model before continuing. Hence, we will provide here a short reminder of modelling the history of mutations via the these methods. First introduced over a hundred years ago, the Wright-Fisher Model is a mathematical model of genetic drift in a population. Specifically, it describes the probability of obtaining  $k$  copies of a new allele  $p$  within a population of size  $N$ , with a non-mutant frequency of  $q$ , and what its expected frequency will be in successive generations.

#### Hardy-Weinberg Principle

The **Hardy-Weinberg Principle** states that allele and genotype frequencies within a population will remain constant unless there is an outside influence that pushes them away from that equilibrium.

The Hardy-Weinberg principle is based on the following assumptions:

- The population observed is very large
- The population is isolated, i.e. there is no introduction of another subpopulation into the general population
- All individuals have equal probability of producing offspring
- All mating in the population is at random

- No random mutations occur in the population from one generation to the next
- Allele frequency drives future genotype frequency (Prevalent allele drives Prevalent genotype)

In a Hardy-Weinberg Equilibrium, for two alleles A and a, occurring with probability  $p$  and  $q = 1-p$ , respectively, the probabilities of a randomly chosen individual having the homozygous AA or aa (pp or qq, respectively) or heterozygous Aa or aA (2pq) genotypes can be described by the equation:

$$p^2 + 2pq + q^2 = 1$$

This equation gives a table of probabilities for each genotype, which can be compared with the observed genotype frequencies using statistical error tests such as the chi-squared test to determine if the Hardy-Weinberg model is applicable. Figure 26.1 shows the distribution of genotype frequencies at different allele frequencies.

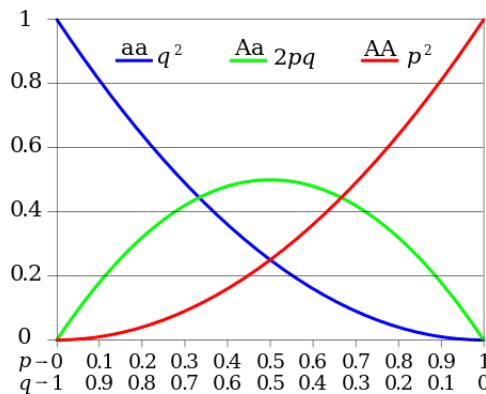


Figure 26.1: Plot of genotype frequencies for different allele frequencies

In natural populations, the assumptions made by the Hardy-Weinberg principle will rarely hold. Natural selection occurs, small populations undergo genetic drift, populations are split or merged, etc. In Nature a mutation will always either disappear (frequency = 0) from the population or become prevalent in a species - this is called “fixation”; in general, 99% of mutations disappear. Figure 26.2 shows a simulation of a mutation’s prevalence in a finite-sized population over time: both perform random walks, with one mutation disappearing and the other becoming prevalent:

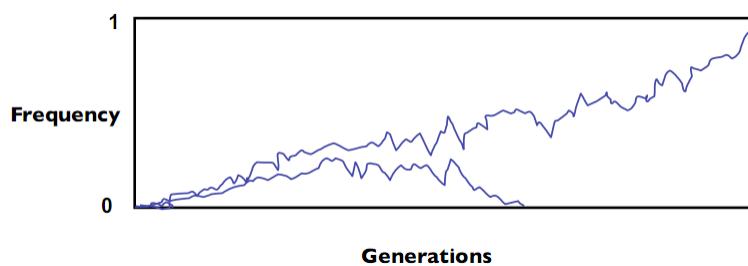


Figure 26.2: Changes in allele frequency over time

Once a mutation has disappeared, the only way for it to reappear is the introduction of a new mutation into the population. For humans, it is believed that a given mutation under no selective pressure should fixate to 0 or 1 (within, e.g., 5%) within a few million years. However, under selection this will happen much faster.

### Wright-Fisher Model

Under this model the time to fixation is  $4N$  and the probability of fixation is  $1/2N$ . In general Wright-Fisher is used to answer questions related to fixation in one way or another. To make sure your intuitions about

the method are absolutely clear considering the following questions:

### **FAQ**

**Q:** Say you have a total of 5 mutations on a chromosome among a population of size 30, on average, how many mutations will be present in the next generation if each entity produces only one child?

**A:** If each parent has only one offspring, then there will be, on average, 5 mutations in the next generation because the expectation of allele frequencies is to remain constant according to the Hardy-Weinberg equilibrium principle in basic biology.

### **FAQ**

**Q:** Is the Hardy-Weinberg Equilibrium principle's assumption about constant allele frequency reasonable?

**A:** No, the reality is far more complex as there is stochasticity in population size and selection at each generation. A more appropriate way to envision this is to imagine drawing alleles from a set of parents, with the amount of alleles in the next generation varying with the size of the population. Hence the frequency in the next generation could very well go up or down. Note here that if the allele frequency goes to zero it will always be at zero. The probability at each successive generation is lower if it's under negative selection and higher if it's under positive selection. Hence if it's a beneficial mutation the fixation time will be smaller, if the mutation is deleterious the fixation will be larger. If there are no offspring with a given mutation, then there won't be any descendants with that mutation either. If one produces multiple offspring however, who in turn produce multiple offspring of their own, then there is a greater chance that this allele frequency will rise.

### **FAQ**

**Q:** Consider that the average human individual carries roughly 100 entirely unique mutations. So, when an individual produces offspring we could expect that half (or 50) of those mutations may appear in the child because in each sperm or egg cell, 50 of those mutations will be present, on average. Hence the offspring of an individual are likely to inherit approximately 100 mutations, 50 from one parent, and 50 from another in addition to their own unique mutations which come from neither parent. With this in mind, one might be interested in understanding what the chances are of some mutations appearing in the next generation if an individual produces, say, n children. How can one do this?

The

**A:** Hint: To compute this value, we assume that some allele originates in the founder, at some arbitrary chromosome (1 for example). Then we ask the question, how many chromosome 1s exist in the entire population? At the moment, the size of the human population is 7 Billion, each carrying two copies of chromosome 1.

above questions and answers should make it painfully clear that the standard Hardy-Weinberg assumption of allele frequencies remaining constant from one generation to the next is violated in many natural cases including migration, genetic mutation, and selection. In the case of selection, this issue is addressed by

modifying the formal definition to include a  $S$ , term which measures the skew in genotypes due to selection. See table 26.1 for a comparison of the original and selection compensated versions:

Behavior	With only drift	With drift and selection
n in next generation	Mean: $n(= 2Np)$ , Dist: $\text{Binomial}(2N, p)$	Mean: $n(1 + \frac{s}{1+ps})$ , Dist: $\text{Binomial}(2N, p \frac{1+s}{1+ps})$
Time to fixation	$4N$	$\frac{4N}{1+\frac{3}{8}N s } \left( \frac{1+\frac{1}{2}(\ln N) s }{1+ s } \right)$
Probability of fixation	$\frac{1}{2N}$	$\frac{1-e^{-2s}}{1-e^{-4Ns}}$

Table 26.1: Comparison of Wright-Fisher Model With Drift, Versus Drift and Selection

The main point to take away from Table 26.1, and this section of the chapter is that whether you have selection or not, it is highly unlikely that a single allele will fixate in a population. If you have a very small population, however, then the chances of an allele fixating are much better. This is often the case in human populations, where there are often small, interbred populations which allow for mutations to fix in a population after only a few generations, even if the mutation is deleterious in nature. This is precisely why we tend to see recessive deleterious mandolin disorders in isolated populations.

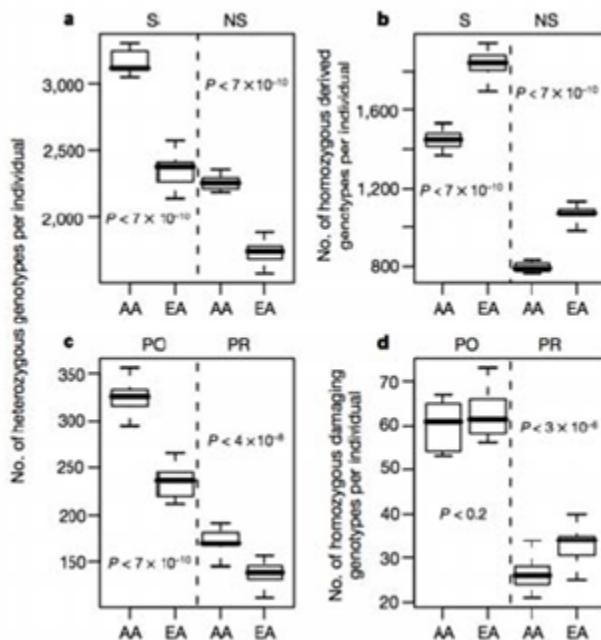


Figure 26.3: A comparison of the heterozygous and homozygous derived and damaging genotypes per individual in an African American (AA) and European American (EA) population study.

### 26.2.3 Ancestral State of Polymorphisms

How can we determine for a given polymorphism which version was the **ancestral state** and which one is the mutant? The ancestral state can be inferred by comparing the genome to that of a closely related species (e.g. humans and chimpanzees) with a known phylogenetic tree. Mutations can occur anywhere along the phylogenetic tree – sometimes mutations at the split fix differently in different populations (“fixed difference”), in which case the entire populations differ in genotype. However, recent mutations will not have had enough time to become fixed, and a polymorphism will be present in one species but fully absent in the other as simultaneous mutations in both species are very rare. In this case, the “derived variant” is the version of the polymorphism appearing after the split, while the “ancestral variant” is the version occurring in both species.

### 26.2.4 Measuring Derived Allele Frequencies

The frequency of the derived allele in the population can be easily calculated, if we assume that the population is homogeneous. However, this assumption may not hold when there is an unseen divide between two groups that causes them to evolve separately as shown in figure 26.4.

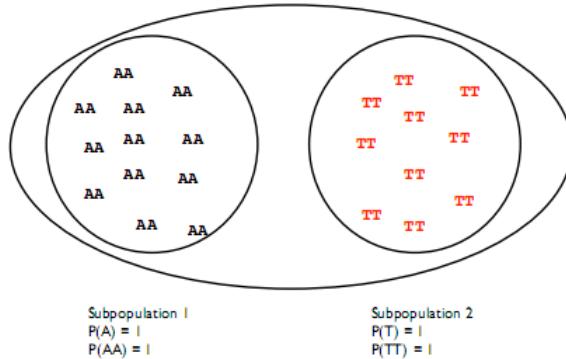


Figure 26.4: Two isolated populations

In this case the prevalence of the variants among subpopulations is different and the Hardy-Weinberg principle is violated.

One way to quantify this difference is to use the **Fixation Index ( $F_{st}$ )** to compare subpopulations within a species. In reality only a portion of the total heterozygosity in a species is found in a given subpopulation.  $F_{st}$  estimates the reduction in heterozygosity ( $2pq$  with alleles  $p$  and  $q$ ) expected when 2 different populations are erroneously grouped together. Given a population having  $n$  alleles with frequencies  $p_i$  where  $(1 \leq i \leq n)$ , the homozygosity  $G$  of the population is calculated as:

$$\sum_{i=1}^n p_i^2$$

The total heterozygosity in the population is given by  $1-G$ .

$$F_{st} = \frac{\text{Heterozygosity(total)} - \text{Heterozygosity(subpopulation)}}{\text{Heterozygosity(total)}}$$

In the case shown in figure 26.4 there is no heterozygosity between the populations, so  $F_{st} = 1$ . In reality the  $F_{st}$  will be small within one species. In humans, for example, it is only 0.0625. For in practise, the  $F_{st}$  is computed either by clustering sub-populations randomly or using an obvious characteristic such as ethnicity or origin.

## 26.3 Genetic Linkage

Before discussing genetic linkage, it is important to understand how genes are passed from generation to generation. Meiosis is the biological process that gives rise to gametes i.e. egg and sperm. There are haploid cells meaning they contain only one copy of each chromosome. The haploid gametes combine to form a diploid organism. During meiosis, each of the parent's two chromosomes can recombine to form new haplotypes. Recombination hotspots occur every 100kb and recombinations are more likely to occur there.

In the simple models we've seen so far, alleles are assumed to be passed on independently of each other. While this assumption generally holds in the long term, in the short term we will generally observe a that certain alleles are passed on together more frequently than expected. This is termed genetic linkage.

The **Law of Independent Assortment**, also known as Mendel's second law states:

*Alleles of different genes are passed on independently from parent to offspring.*

When this “law” holds, there is no correlation between different polymorphisms and the probability of a haplotype (a given set of polymorphisms) is simply the product of the probabilities of each individual polymorphism.

In the case where the two genes lie on different chromosomes this assumption of independence generally holds, but if the two genes lie on the same chromosome, they are more often than not passed on together. Without genetic recombination events, in which segments of DNA on homologous chromosomes are swapped (crossing-over), the alleles of the two genes would remain perfectly correlated. With **crossing-overs** however, the correlation between the genes will be reduced over several generations. Over a suitably long time interval, recombination will completely remove the linkage between two polymorphisms; at which point they are said to be in equilibrium. When, on the other hand, the polymorphisms are correlated, we have **Linkage Disequilibrium (LD)**. The amount of disequilibrium is the difference between the observed haplotype frequencies and those predicted in equilibrium.

The linkage disequilibrium can be used to measure the difference between observed and expected assortments. If there are two alleles (1 and 2) and two loci (A and B) we can calculate haplotype probabilities and find the expected allele frequencies.

- Haplotype frequencies

- $P(A_1) = x_{11}$
- $P(B_1) = x_{12}$
- $P(A_2) = x_{21}$
- $P(B_2) = x_{22}$

- Allele frequencies

- $P_{11} = x_{11} + x_{12}$
- $P_{21} = x_{21} + x_{22}$
- $P_{12} = x_{11} + x_{21}$
- $P_{22} = x_{12} + x_{22}$

- $D = P_{11} * P_{22} - P_{12} * P_{21}$

$D_{max}$ , the maximum value of D with given allele frequencies, is related to D in the following equation:

$$D' = \frac{D}{D_{max}}$$

$D'$  is the maximum linkage disequilibrium or “complete skew” for the given alleles and allele frequencies.

$D_{max}$  can be found by taking the smaller of the expected haplotype frequencies  $P(A_1, B_2)$  or  $P(A_2, B_1)$ . If the two loci are in complete equilibrium, then  $D' = 0$ . If  $D' = 1$ , there is full linkage.

The key point is that relatively recent mutations have not had time to be broken down by crossing-overs. Normally, such a mutation will not be very common. However, if it is under positive selection, the mutation will be much more prevalent in the population than expected. Therefore, by carefully combining a measure of LD and derived allele frequency, we can determine if a region is under positive selection.

Decay of **linkage disequilibrium** is driven by recombination rate and time (in generations) and has an exponential decay. For a higher recombination rate, linkage disequilibrium will decay faster in a shorter amount of time. However, the background recombination rate is difficult to estimate and varies depending on the location in the genome. Comparison of genomic data across multiple species can help in determining these background rates.

### 26.3.1 Correlation Coefficient $r^2$

Answers how predictive an allele at locus A is of an allele at locus B

$$r^2 = \frac{D^2}{P(A_1)P(A_2)P(B_1)P(B_2)}$$

As the value of  $r^2$  approaches 1, the more two alleles at two loci are correlated. There may be linkage disequilibrium between two haplotypes, even if the haplotypes are not correlated at all. The correlation coefficient is particularly interesting when studying associations of diseases with genes, where knowing the genotype at locus A may not predict a disease whereas locus B does. There is also the possibility where neither locus A nor locus B are predictive of the disease alone but loci A and B together are predictive.  $r^2$  scores are the same as the GWAS association summary statistics of the SNPs.

## 26.4 Haplotype Phasing

There are a limited number of distinct haplotypes. Given a region with 36 SNPs, there are potentially  $2^{36}$  combinations. However, looking at 120 European chromosomes, only 5 recurrent haplotypes and 2 singleton haplotypes were found. This implies a high level of genotype sharing even for unrelated individuals.

### 26.4.1 Phasing of related individuals

The goal of phasing is to resolve genotypes into their underlying haplotypes which requires auxiliary information such as parent genotypes. As seen in figure 26.5, homozygous sites can be trivially phased. If at least one parent is homozygous, there is no ambiguity when the child is heterozygous. However, if both parents are heterozygous, LD is needed to resolve the remaining ambiguous sites.

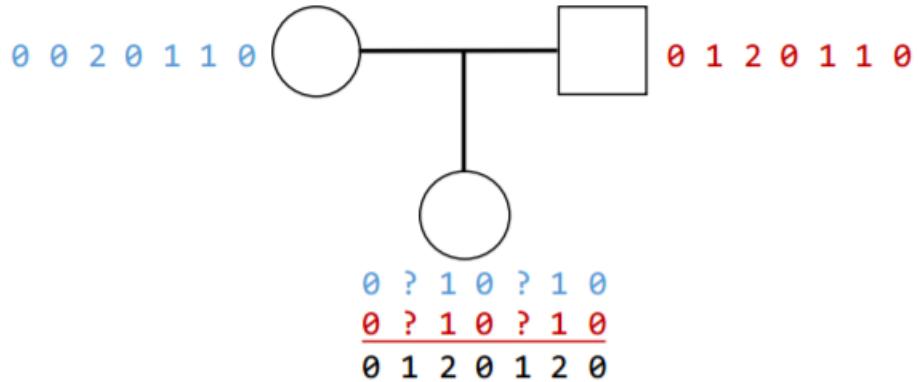


Figure 26.5: Trio phasing

### 26.4.2 Phasing of unrelated individuals

Phasing of unrelated individuals is also possible. Modern analyses often consider collections of unrelated individuals and need to rely on patterns of LD. The unobserved haplotypes underlying observed genotypes can be traced back to a common ancestor with a reference panel. Directly fitting the ARG is not possible so it is approximated. The unobserved haplotypes are generated by copying segments from reference haplotypes so that the resulting genotypes match the observations. Recombination corresponds to switching which reference haplotype that gets copied from. Therefore, the model needs to infer where recombination events occur. Trying all possible sets of recombination events is too difficult. Instead one slides a window over the genome and considers whether recombination occurs at the right edge or if no recombination occurs. Redundancy is compressed in the reference haplotype by representing them as paths through a graph. Only some of the paths are consistent with the observed genotypes as the sum of the haplotypes must equal that of the observed genotype. A HMM can be used as a to compute the posterior probability of all feasible paths through the haplotype graph using a forward-backward algorithm. The hidden state is the two reference haplotypes to copy at each position for each observed individual. The emissions are the unobserved haplotypes for individual allowing for substitutions not seen in the reference individuals. The transitions are the probability of a recombination event. This means one can phase known parts and impute the rest.

### 26.4.3 Fine-mapping Disease Associations

Association mapping refers to identifying variants/gene associated with disease. This is confounded by LD. Many variants are strongly correlated to the true causal variant and will show nearly as strong associations. One can use estimated correlations to explain correlated associations and recover the true underlying effects. SNPs that are causal are will have high  $r^2$  values with each other. Allele frequencies and LD patterns can differ between populations. The disease associations are currently biased towards European cohorts however as association studies are conducted in Asia and Africa there is a need to develop statistical methods which can account for the population genetic differences.

## 26.5 Natural Selection

In the mid 1800's the concept of evolution was not an uncommon idea, but it wasn't before Darwin and Wallace proposed natural selection as the mechanism that drives evolution in nature that the theory of evolution got widespread recognition. It took 70 years (1948) until J.B.S Haldane's "Malaria Hypothesis" found the first example for natural selection in humans. He showed a correlation between genetic mutations in red blood cells and the distribution of malaria prevalence and discovered that individuals who had a specific mutation that made them suffer from sickle cell anaemia also gave made them resistant to malaria.

Lactose tolerance (lasting into adulthood) is another example of natural selection. Such explicit examples were hard to prove without genome sequences. With whole genome sequencing readily available, we can now search the genome for regions with the same patterns as these known examples to identify further regions undergoing natural selection.

### 26.5.1 Genomics Signals of Natural Selection

- $K_a/K_s$  ratio of non-synonymous to synonymous changes per gene
- Low diversity and many rare alleles over a region (ex Tajima's D with regard to sickle-cell anemia)
- High derived allele frequency (or low) over a region (ex Fay and Wu's H)
- Differentiation between populations faster than expected from drift (Measured with  $F_{st}$ )
- Long haplotypes: evidence of selective sweep.
- Exponential prevalence of a feature in sequential generations
- Mutations that help a species prosper

#### Examples of Negative (Purifying) Selection

- **Across species** we see negative selection of new mutations in conserved functional elements (exons, etc.).
- **New alleles** within one species tend to have lower allele frequencies if the allele is non-synonymous than synonymous. Lethal alleles have very low frequencies.

#### Examples of Positive (Adaptive) Selection

- **Similar to negative selection** in that positive selection more likely in functional elements or non-synonymous alleles.
- **Across species** in a conserved element, a positively selected mutation might be the same over most mammals, but change in a specific species because a positively selected mutation appeared after speciation or caused speciation.

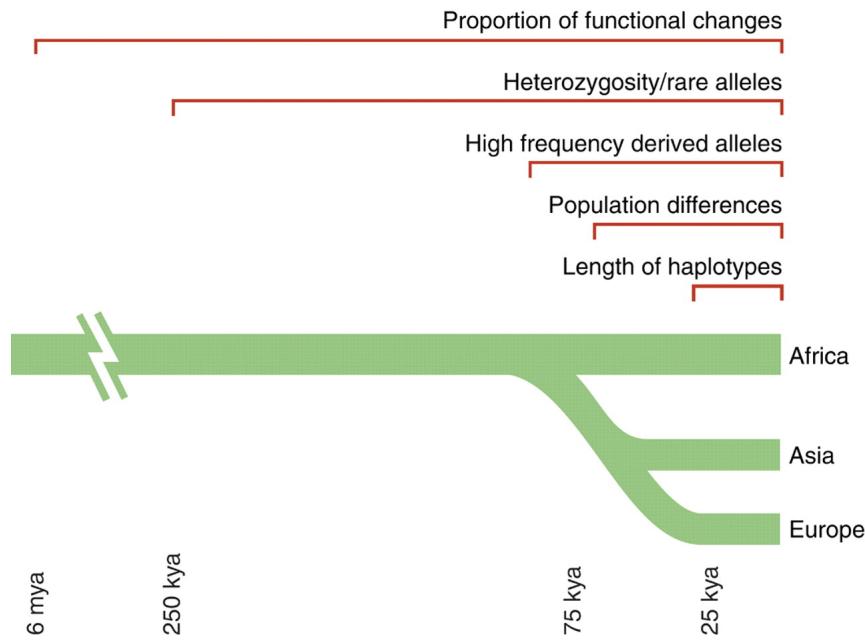


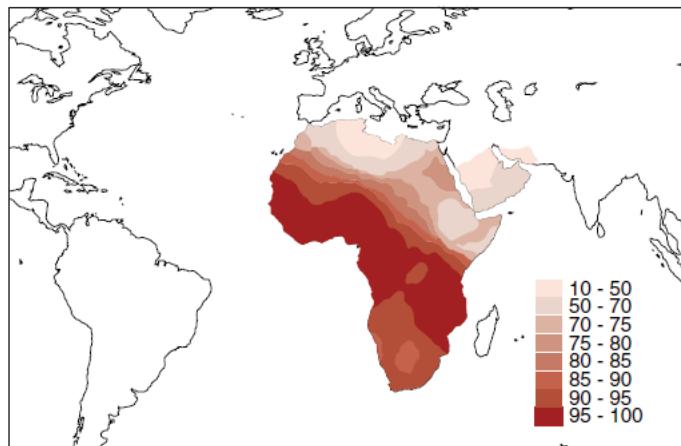
Figure 26.6: Approximate Time Table of Effects  
Sabeti et al. *Science* 2006

- **Within a species** positively selected alleles likely differ in allele frequency ( $F_{st}$ ) across populations. Examples include malaria resistance in African populations (26.7) and lactose persistence in European populations (26.8).
- **Polygenic selection** within species can arise when a trait is selected for that depends on many genes. An example is human height where 139 SNPs are known to be related to height. Most are not population specific mutations but alleles across all humans that are selected for in some populations more than others. (26.9)

### Statistical Tests

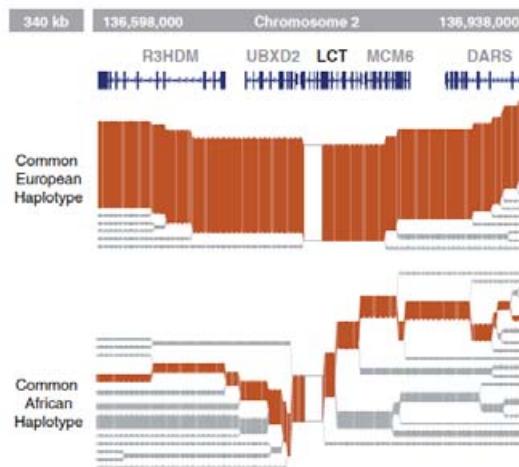
- **Long range correlations (iHs, Xp, EHH):** If we tag genetic sequences in an individual based on their ancestry, we end up with a “broken haplotype”, where the number of breaks (color changes) is correlated with the number of recombinations and can tell us how long ago a particular ancestry was introduced.
- **SWEEP** – A program developed by Pardis Sabeti, Ben Fry and Patrick Varilly. SWEEP detects evidence of natural selection by analyzing haplotype structures in the genome using the long range haplotype test (LRH). It looks for high frequency alleles with long range linkage disequilibrium that hints to large scale proliferation of a haplotype that occurred at a rate greater than recombination could break it from its markers .
- **High Frequency Derived Alleles** Look for large spikes in the frequency of derived alleles in set positions.
- **High Differentiation ( $F_{st}$ )** Large spikes in differentiation at certain positions.

Using these tests, we can find genomic regions under selective pressure. One problem is that a single SNP under positive selection will allow nearby SNPs to piggy-back and ride along. It is difficult to distinguish the SNP under selection from its neighbours with only one test. Under selection, all the tests are strongly



Extreme population differences in *FY\*O* allele frequency. The *FY\*O* allele, which confers resistance to *P. vivax* malaria, is prevalent and even fixed in many African populations, but virtually absent outside Africa (38).

Figure 26.7: Localized positive selection for Malaria resistance within species  
Sabeti et al. *Science* 2006



Long haplotype surrounding the lactase persistence allele. The lactase persistence allele is prevalent (~77%) in European populations but lies on a long haplotype, suggesting that it is of recent origin (6).

Figure 26.8: Localized positive selection for lactase persistence allele  
Sabeti et al. *Science* 2006

correlated; however, in the absence of selection they are generally independent. Therefore, by employing a composite statistic built from all of these tests, it is possible to isolate the individual SNP under selection.

Examples where a single SNP has been implicated in a trait:

- Chr15 – Skin pigmentation in Northern Europe
- Chr2 – Hair traits in Asia
- Chr10 – Unknown trait in Asia
- Chr12 – Unknown Trait in Africa

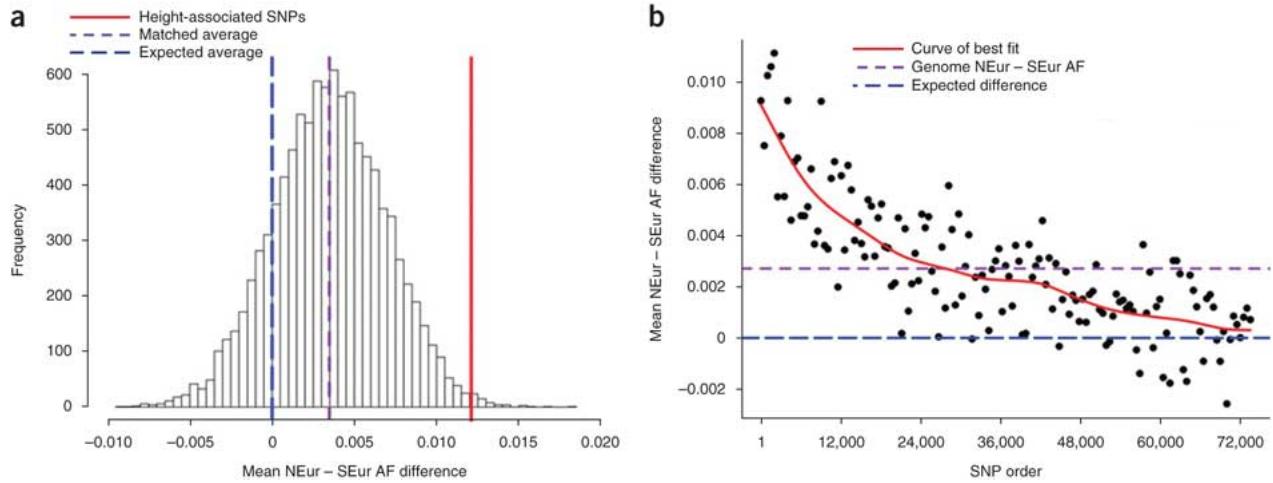


Figure 26.9: Mean allele frequency difference of height SNPs, matched SNPs, and genome-wide SNPs between Northern- and Southern-European populations

Turchin et al., *Nature Genetics* (2012)

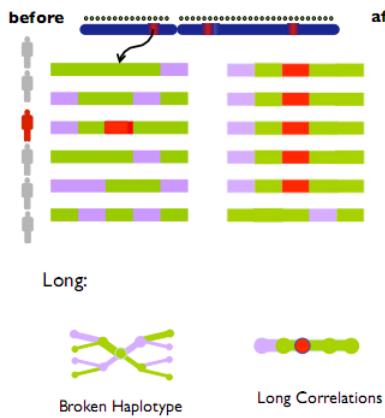


Figure 26.10: Broken haplotype as a signal of natural selection

## 26.6 Human Evolution

### 26.6.1 A History of the Study of Population Dynamics

Not surprisingly, the scientific community has a long, and somewhat controversial history of interest in recent population dynamics. While indeed some of this interest was applied toward more nefarious aims, such as the scientific justifications for racism for eugenics but these are increasingly the exception and not the rule. Early studies of population dynamic were primitive in many ways. Quantifying the differences between human populations was originally performed using blood types, as they seemed to be phenotypically neutral, could be tested for outside of the body, and seemed to be polymorphic in many different human populations. Fast forward to the present, and the scientific community has realized that there are other glycoproteins beyond the A,B and O blood groups that are far more polymorphic in the population. As science continued to advance and sequencing became a reality, they began whole genome sequencing of the Y-chromosome, mitochondrial and microsatellite markers around them. What's special about those two types of genetic data? First and foremost, they are quite short so they can be sequenced more easily than other chromosomes. Beyond just the size, the reason that the Y and mitochondrial chromosomes were of such interest is because they do not recombine, and can be used to easily reconstruct inheritance trees.

This is precisely what makes these chromosomes special relative to a short chunk on an autosome; we know exactly where it comes from because we can trace paternal or maternal lineage backward in time.

This type of reconstruction does not work with other chromosomes. If one were to generate a tree using a certain chunk of all of chromosome 1 in a certain population, for instance, they would indeed form a phylogeny but that phylogeny would be picked from random ancestors in each of the family trees.

As sequencing continued to develop and grow more effective, the human genome project was being proposed, and along with it there was a strong push to include some sort of diversity measure in genomic data. Technically speaking, it was easiest to simply look at microsatellites for this diversity measure because they can be studied on gel to see size polymorphisms instead of inspecting a sequence polymorphism. As a reminder, a microsatellite is a region of variable length in the human genome often characterised by short tandem repeats. One reason for microsatellites is retroviruses inserting themselves into the genome, such as the ALU elements in the human genome. These elements sometimes become active and will retro-transpose as insertion events and one can trace when those insertion events have happened in human lineage. Hence, there was a push, early on to assay these parts of the genome in a variety of different populations. The really attractive thing about microsatellites is that they are highly polymorphic and one can actually infer their rate of mutation. Hence, we can not only say that there is a certain relationship between populations based on these rates, but we can also say how long they have been evolving and even when certain mutations occurred, and how long it's been on certain branches of the phylogenetic tree.

## FAQ

**Q:** Can't this simply be done with SNPs

**A:** You can't do it very easily with SNPs.

You can get an idea of how old they are based on their allele frequency, but they're also going to be influenced by selection.

After the human genome project, came the Haplotype inheritance Hapmap project which looked at SNPs genome wide. We have discussed Haplotype inheritance in detail in prior chapters where we learned the importance of Hapmap in designing genotyping arrays which look at SNPs that mark common haplotypes in the population.

The effects of Bottlenecks on Human diversity Using this wealth of data across studies and a plethora of mathematical techniques has led to the realization that humans, in fact, have a very low diversity given our census population; which implies a small effective population size. Utilizing the Wright-Fisher model it is possible to work back from the level of diversity and the number of mutations we see in the population today to generate a founding population size. When this computation is performed it works out to being around 10,000.

## FAQ

**Q:** Why is this so much smaller than our census population size?

**A:** There was A population bottleneck somewhere.

Most of the total variation between humans is happening within-continent. One can measure how much diversity is explained by geography and how much is not. It turns out that most of it is not explained by geography. In fact, most common variants are polymorphic in every population and if a common variant is unique to a given population, there probably hasn't been enough time for that to happen by drift itself.

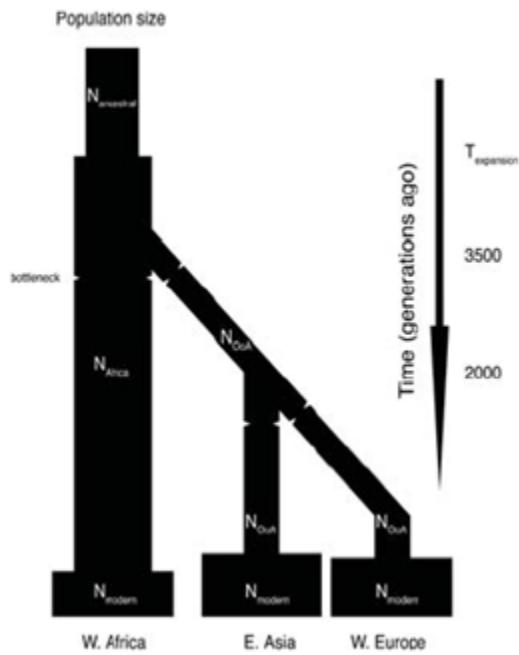


Figure 26.11: A depiction of two major bottleneck events, one in the founding population from Africa, and other, smaller subsequent bottleneck events in the East Asian and Western European populations.

Recall what an unlikely process it is to get to a high allele frequency over the course of several generations by mere chance alone. Hence, we may interpret this as a signal of selection when it occurs. All of the evidence in terms of comparing diversity patterns and trees back to ancestral haplotypes converges to an Out-of-Africa hypothesis which is the overwhelming consensus in the field and is the lens through which we review all the genetic population data. Starting from the African founder population, there have been works which have demonstrated that it's possible to model population growth using the Wright Fisher model. The studies have shown that the growth rate we see in Asian and European populations are only consistent with large exponential growth after the out-of-Africa event.

Study	Sample size (n)*	Time growth started (years ago)†	Initial $N_e$ ‡	Growth per generation (%)
Gravel <i>et al.</i> (5)	60	23,000§ (21,000–27,000)	1032 (677–1290)	0.48 (0.30–0.75)
Gutenkunst <i>et al.</i> (6) (including New World modeling)	22	26,400§ (21,700–30,700)	1500 (900–2200)	0.23 (0.16–0.34)
Gutenkunst <i>et al.</i> (6) (excluding New World modeling)	22	21,200§ (17,600–23,900)	1000 (500–1500)	0.4 (0.26–0.57)
Schaffner <i>et al.</i> (29)	62	8750	7700	0.73
Coventry <i>et al.</i> (18)	10,422	1400 (900–2800)	7700#	9.4 (4.5–14.5)

Table 26.2: Genetic Estimates of Recent Population Growth in Europe

This helps us understand the reasons for phenotypical differences between the races as Bottlenecks which are followed by exponential growth can lead to an excess of rare alleles. The present theory on human diversity states that there were secondary bottleneck events after the founding population migrated out of Africa. These founders were, at some earlier point subject to an even smaller bottleneck event which is

now reflected in every human genome on the planet, regardless of their immediate ancestry. It is possible to estimate how small the original bottle neck was by looking at differences between African and European origin individuals, inferring the effects of the secondary bottleneck, and the term of exponential growth of the European population. The other way of approaching bottleneck event estimation is to simply inspect the allele frequency spectrum needed to build coalescent trees. In this way, one can take haplotypes across the genome and ask what the most recent common ancestor was by observing how the coalescence varies across the genome. For instance, one may guess that some haplotype was positively selected for only recently given the length of the haplotype. An example of one such recent mutation in the European population is the lactase gene. Another example for the Asian population is the ER locus.

There is a wealth of literature showing that when one draws a coalescence tree for most haplotypes it ends up going way back before when we think speciation happened. This indicates that certain features have been kept polymorphic for a very long time. One can, however, look at this distribution of features across the whole genome and infer something about population history from it. If there was a recent bottle neck in a population, it will be reflected by the ancestors being very recent whereas more ancient things will have survived the bottleneck. One can take the distribution of coalescent times and run simulations for how the effect of population size would have varied with time. The model for doing this type of study was outlined by Li and Durbin. The Figure 26.12 from their study illustrates two such bottleneck events. The first is the bottleneck which occurred in Africa long before migrations out of the continent. This was then followed by a population specific bottleneck that resulted from migration groups out of Africa. This is reflected in the diversity of the populations today based on their ancestry and it can be derived from looking at a pair of chromosome from any two people in these populations.

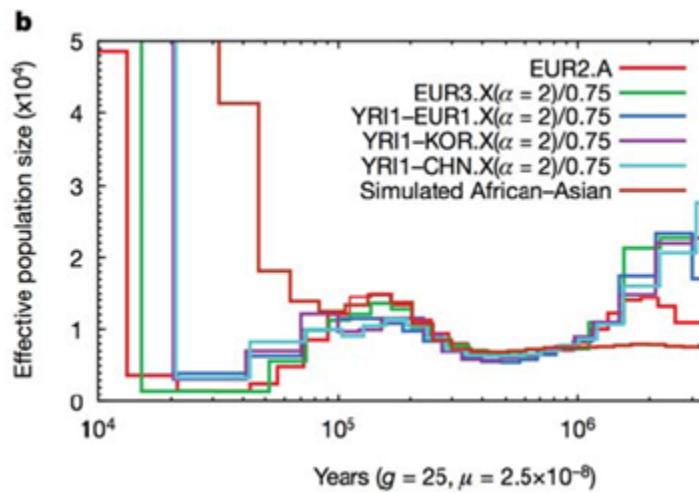


Figure 26.12: An illustration of two bottleneck events

## 26.6.2 Understanding Disease

Understanding that human populations went through bottlenecks has important implications for understanding population specific disease. A study published by Tennessen et al. this year was looking at exome sequences in many classes of individuals. The study intended to look at how rare variants might be contributing to disease and as a consequence they were able to fit population genetics models to the data, and ask what sort of deleterious variants were seen when sequencing exomes from a broad population panel. Using this approach, they were then able to generate parameters which describe how long ago exponential growth between the founder, and branching populations occurred. See figure 26.13 below for an illustration of this:

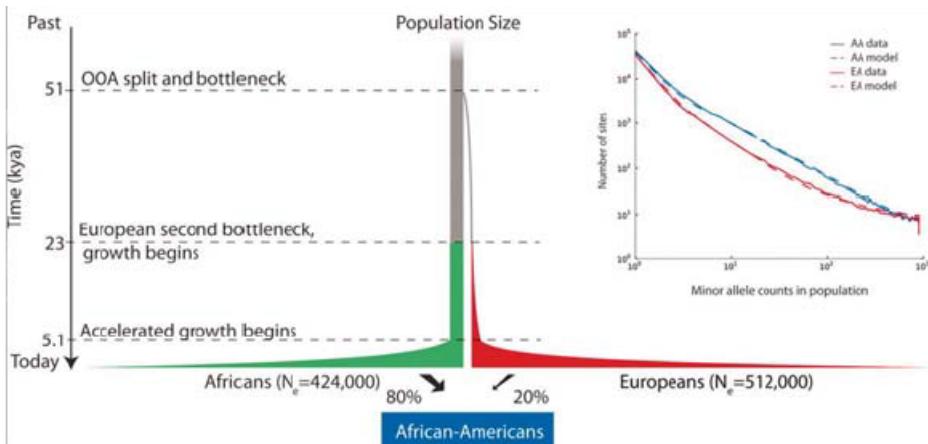


Figure 26.13: The figure illustrate the effects of a bottleneck events on the number of rare Alleles in a population.

### 26.6.3 Understanding Recent Population Admixture

In addition to viewing coalescent times, one can also perform Principal Component Analysis on SNPs to gain an understanding of more recent population admixtures. Running this on most populations shows clustering with respect to geographical location. There are some populations, however, that experienced a recent admixture for historical reason. The two most commonly referred to in the scientific literature are: African Americans, who on average are 20% European and 80% West African. And Mexican Americans, who are on average, half Native American and European. To infer this mixture level, one can simply perform PCA and look at the loading of the components across the genome. Figure 26.14 illustrates this nicely; it shows the results for two genomes with clearly defined chunks of different ancestry.

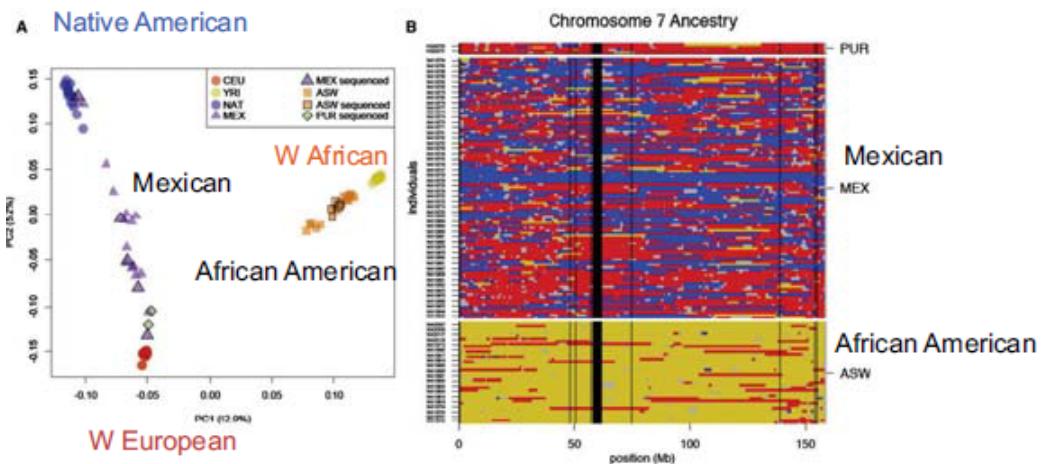


Figure 26.14: A depiction of European admixture levels in the Mexican, and African American populations.

There are two major things one can say about the admixture event of African Americans and Mexican Americans. The first and more obvious is inferring the admixture level. The second, and more interesting, is inferring when the admixture event happened based on the actual mixture level. As we have discussed in previous chapters, the racial signifiers of the genome break down with admixture because of recombination in each generation. If the population is contained, the percentage of those with European and West African origin should stay the same in each generation, but the segments will get shorter, due to the mixing. Hence, the length of the haplotype blocks can be used to date back to when the mixing originally happened. (When

it originally happened we would expect large chunks, with some gambits being entirely of African origin, for instance.) Using this approach, one can look at the distribution of recent ancestry traps and then fit a model to when these migrants entered an ancestral population as shown below:

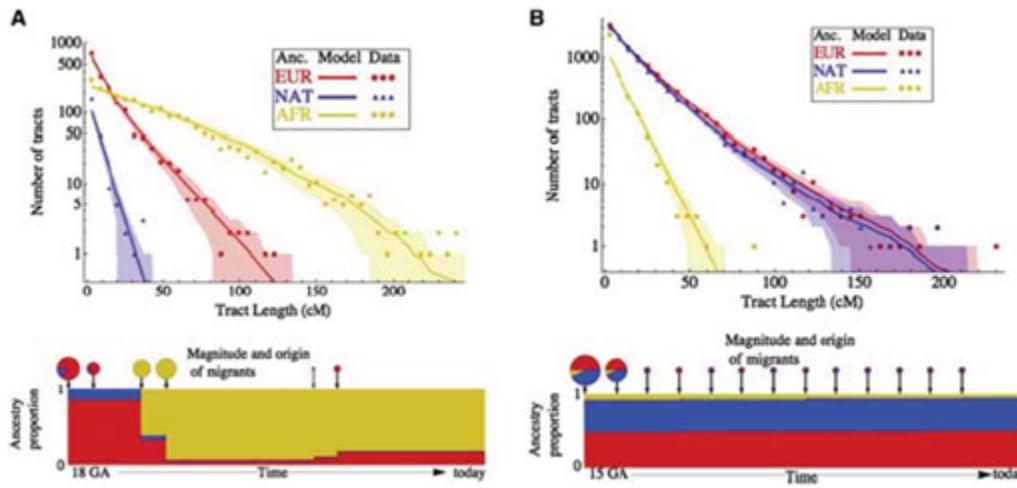


Figure 26.15: An illustration of the magnitude and origin of migrants based on the tract length and number of tracts in the admixed population.

## 26.7 Current Research

### 26.7.1 HapMap project

The International **HapMap** Project aims to catalog the genomes of humans from various countries and regions and find similarities and differences to help researchers find genes that will benefit the advance in disease treatment and administration of health related technologies.

### 26.7.2 1000 genomes project

The 1000 Genomes Project is an international consortium of researchers aiming to establish a detailed catalogue of human genetic variation. Its aim was to sequence the genomes of more than a thousand anonymous participants from a number of different ethnic groups. In October 2012, the sequencing of 1092 genomes was announced in a Nature paper. It is hoped that the data collected by this project will help scientists gain more insight into human evolution, natural selection and rare disease-causing variants. The 1000 genomes project suggested that the Non-African populations sustained a bottleneck around 15-20 kyears ago while migrating from Africa, after which there was a rapid population expansion. It also suggested that African individuals have more variation in their genomes due to containing more ancestral diversity.

## 26.8 Further Reading

- Campbell Biology, 9th edition; *Pearson*; Chapter 23: The Evolution of Populations
- The Cell, 5th edition, *Garland publishing*; Chapters 5: DNA replication, repair and recombination, Chapter 20: Germ cells and fertilization

## Bibliography

# **Part V**

## **Medical Genomics**



---

CHAPTER  
**TWENTYSEVEN**

---

## VARIATION 2: : QUANTITATIVE TRAIT MAPPING, eQTLs, MOLECULAR TRAIT VARIATION

Tim Wall, Brendan Liu, Lei Ding (2014), Tejas Sundaresan, Giri Anand (2015), Maria Fabre, Zi-Ning Choo (2016)

### Figures

---

27.1 Non-coding vs. Coding Variation and Explanation of Traits . . . . .	438
27.2 A comparison of eQTL and GWAS approaches . . . . .	439
27.3 cis-eQTL regulating gene transcription directly. [3] . . . . .	440
27.4 trans-eQTL regulating gene transcription indirectly. [3] . . . . .	440
27.5 eQTL Study Approach . . . . .	442
27.6 Effects of the search radius and MAF on the number of eQTLs detected . . . . .	443
27.7 The effects of technical and population variance on expression level assays . . . . .	444
27.8 The decision parameters researchers must make when conducting an eQTL study . . . . .	445
27.9 An example eQTL Study on asthma . . . . .	446

---

## 27.1 Introduction

Differences in gene coding regions across different organisms do not completely explain the phenotypic variation we see. For example, although the phenotypic difference is high between humans and chimpanzees and low between different squirrel species, there is more genetic variation among the squirrel species [1]. These observations lead us to conclude that there must be more than just gene-coding variation that accounts for phenotypic variation; specifically, non-coding variation also influences how genes are expressed, and consequently influences the phenotype of an organism. In fact, previous research has shown that most genetic variation occurs in non-coding regions [2]. Furthermore, most expression patterns have been found to be heritable traits.

## 27.2 Motivations for studying eQTLs

### 27.2.1 Effects of noncoding regions on phenotype

Understanding how variation in non-coding regions affects co-regulating genes would allow us not only to understand but also control the expression of these and other related genes. This is especially relevant to the control of undesirable trait expressions like complex, polygenic diseases (Figure 27.1). In Mendelian disease,

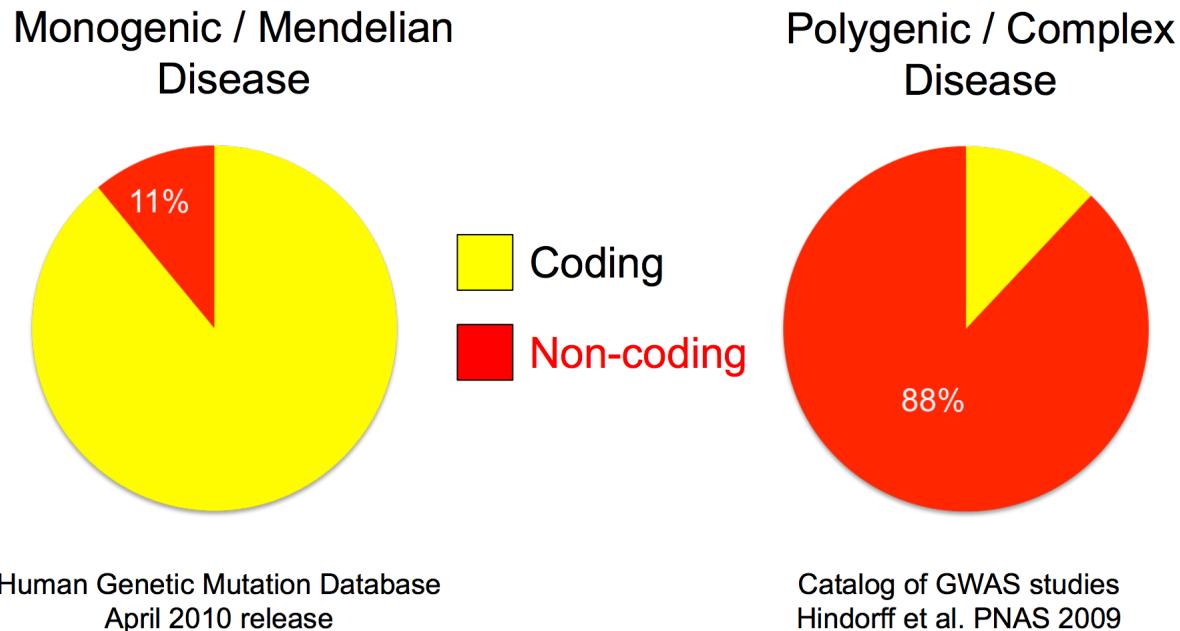


Figure 27.1: Non-coding vs. Coding Variation and Explanation of Traits

the majority of disease risk is predicted by coding variation, whereas in polygenic diseases the vast majority of causal variation is found outside of coding regions. This suggests that variation in the regulation of gene expression may play a greater role than genotypic variation in these polygenic diseases. Thus, the study of these trait associated variants is a step in the direction of understanding how genetic sequences both code for and control the expression of such diseases and their associated phenotypes.

### 27.2.2 Understanding and treating diseases at multiple phenotypic layers

As shown in Figure X, there are many phenotypic layers between genotype and organism phenotype, including tissue/cell type-specific expression, epigenetic changes, expression level changes, and endo phenotypes (phenotypic differences at the molecular level). It is unlikely for a genetic variant in a to directly cause organism phenotypic changes because changes that cause strongly adverse effects are under strong purifying selection. The intermediate phenotypic layer of expression level is under weaker purifying selection, and thus may be more easily connected to genetic variation associated with organism phenotypes.

In addition, by understanding changes at intermediate phenotypes implicated in disease, we may be able to treat diseases by targeting those changes. The caveat here is that unlike genomic variation, which is always causative, it is difficult to determine whether these intermediate phenotypic changes cause or result from disease.

## 27.3 eQTL Basics

eQTLs (*expression quantitative trait loci*) encapsulate the idea of non-coding regions influencing mRNA expression introduced above: we can define an eQTL as a region of variants in a genome that are quantitatively correlated with the expression of another gene encoded by the organism. Usually, we will see that certain SNPs in certain non-coding regions will either enhance or disrupt the expression of a certain gene. The field of identifying, analyzing, and interpreting eQTLs in the genome has grown immensely over the last couple

of years with hundreds of research papers being published.

There are four main mechanisms for how eQTLs influence the expression of their associated genes:

1. Altered transcription factor binding
2. Histone modifications
3. Alternative splicing of mRNA
4. miRNA silencing

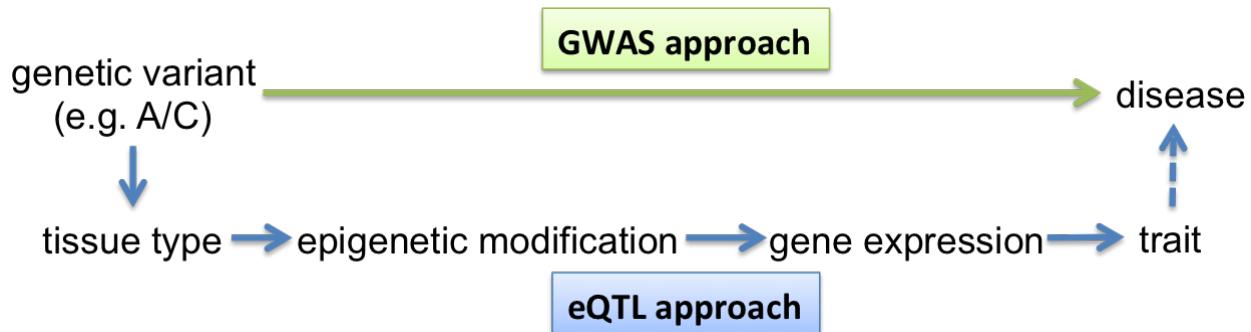


Figure 27.2: A comparison of eQTL and GWAS approaches.

## FAQ

**Q:** What is the difference between an eQTL study and a GWAS?

**A:** There are two fundamental differences. The first is in the nature of the phenotype being examined. In an eQTL, the phenotype checked is usually on a lower level of biological abstraction (normalized gene expression levels) instead of a more higher-level, sometimes visible phenotype used in GWAS, such as "black hair"). Secondly, in GWAS, usually because the phenotype being correlated with various SNPs is a higher-level phenotype, we very rarely see tissue-specific GWAS. However, in eQTLs, the expression patterns of mRNA could vary greatly between tissue-types within the same individual, and eQTL studies for a specific tissue-type, such as neuron and glial cells, can be performed (Figure 27.2)

### 27.3.1 Cis-eQTLs

The use of whole genome eQTL analysis has separated eQTLs into two distinct types of manifestation. The first is a **cis-eQTL** (Figure 27.3) in which the position of the eQTL maps near the physical position of the gene. Because of proximity, cis-eQTL effects tend to be much stronger, and thus can be more easily detected by GWAS and eQTL studies. Often, these function as promoters of certain polymorphisms, affect methylation and chromatin conformation (thus increasing or decreasing access to transcription), and can manifest as insertions and deletions to the genome. Cis-eQTLs are generally classified as variants that lie within 1 million base pairs of the gene of interest. However, this is indeed an arbitrary cutoff and can be altered by an order of magnitude, for instance.

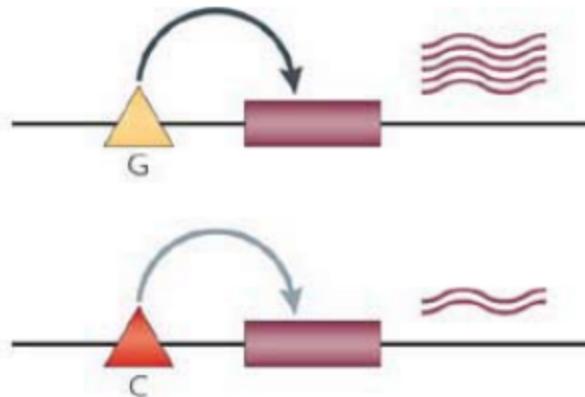


Figure 27.3: cis-eQTL regulating gene transcription directly. [3]

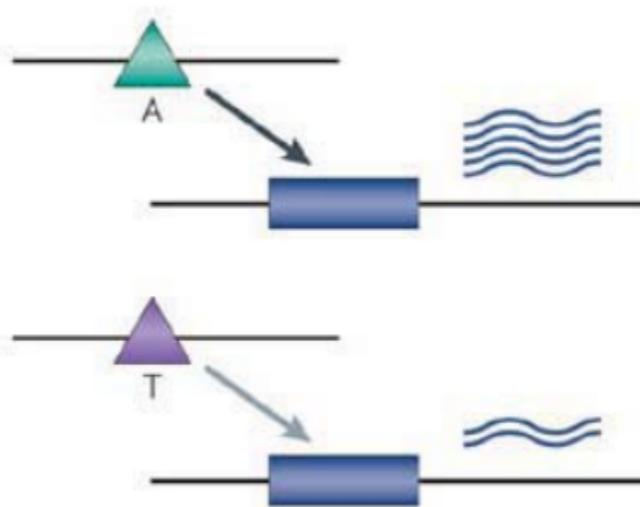


Figure 27.4: trans-eQTL regulating gene transcription indirectly. [3]

### 27.3.2 Trans-eQTLs

The second distinct type of eQTL is a **trans-eQTL** (Figure 27.4). A trans-eQTL does not map near the physical position of the gene it regulates. Its functions are generally more indirect in their effect on the gene expression (not directly boosting or inhibiting transcription but rather, affecting kinetics, signaling pathways, etc.). Since such effects are harder to determine explicitly, they are harder to find in eQTL analysis; in addition, such networks can be extremely complex, further limiting trans-eQTL analysis. However, eQTL analysis has led to the discovery of **trans hotspots** which refer to loci that have widespread transcriptional effects [11].

Perhaps the biggest surprise of eQTL research is that, despite the location of trans hotspots and cis-eQTLs, no major trans loci for specific genes have been found in humans [12]. This is probably attributed to the current process of whole genome eQTL analysis itself. As useful and widespread whole genome eQTL analysis is, we find that genome-wide significance occurs at  $p = 5 \times 10^{-8}$  with multiple testing on about 20,000 genes. Thus, studies generally use an inadequate sample size to determine the significance of many trans-eQTL associations, which start with priors of very low probability to begin with as compared to cis-eQTLs [4].

Further, the bias reduction methods described in earlier sections deflate variance, which is integral to capture the microtrait associations inherent in trans loci. Finally, non-normal distributions limit the statistical significance of associations between trans-eQTLs and gene expression[4]. This has been slightly remedied by the use of cross-phenotype meta-analysis (CPMA)[5] which relies on the summary statistics from GWAS rather than individual data. This cross-trait analysis is effective because trans-eQTLs affect many genes and thus have multiple associations originating from a single marker. Sample CPMA code can be found in *Tools and Resources*.

However, while trans loci have not been found, trans-acting *variants* have been found. Since it can be inferred trans-eQTLs affect many genes, CPMA and ChIP-Seq can be used to detect such cross-trait variants. Indeed, 24 different significant trans-acting transcription factors were determined from a group of 1311 trans-acting SNP variants by observing allelic effects on populations and target gene interactions/connections.

## 27.4 Structure of an eQTL Study

The basic approach behind an eQTL study is to consider each gene's expression as a quantitative multi-factor trait and regress on principal components that explain the variance in expression. First, cells of the tissue of interest are extracted and their RNA extracted. Expression of proteins of interest is measured either by microarray or by RNA-seq analysis. Expression levels of each gene are regressed on genotypes, controlling for biological and technical noise, such that

$$Y_i = \alpha + X_i\beta + \epsilon_i$$

Where  $Y_i$  is the gene expression of gene  $i$ ,  $X_i$  is a vector containing the allelic composition of each SNP associated with the gene (and can take on values 0, 1, or 2 given a reference allele),  $\alpha$  and  $\beta$  are column vectors containing the regression coefficients, and  $\epsilon_i$  is the residual error (See Figure 27.5) [9]. In concept, such a study is extremely simple. In practice, there are hundreds of potential confounders and statistical uncertainties which must be accounted for at every step of the process. However, the same regression model can be used to account for these covariates.

Figure 27.9 contains an example eQTL study conducted on asthma. The key result from the study is the linear model in the upper right: we can see as the genotype tends more towards the "A" variant, the target gene expression decreases.

### 27.4.1 Considerations for Expression Data

Quantifying expression of genes is fraught with experimental challenges. For a more detailed discussion of these issues, see Chapter 14. One important consideration for this type of expression analysis is the **SNP-under-probe effect**: probe sequences that map to regions with common variants provide inconsistent results due to the effect of variation within the probe itself on binding dynamics. Thus, experiments repeated with multiple sets of probes will produce a more reliable result. Expression analysis should also generally exclude **housekeeping genes**, which are not differentially regulated across members of a population and/or cell types, since these would only dilute the statistical power of the study.

Determining the **minimum expression level** required for a potential target gene is another consideration. In general, most cells express fewer than 50% of all known genes, and testing a target that is not expressed is inefficient. It is necessary to select some expression threshold at which a gene is considered a potential regulatory target in eQTL analysis.

Another consideration is **population variance**, or the variation in intensity of expression level throughout samples in the analysis. If mean gene expression level changes but the population variance is high, then that change is less likely to be significant.

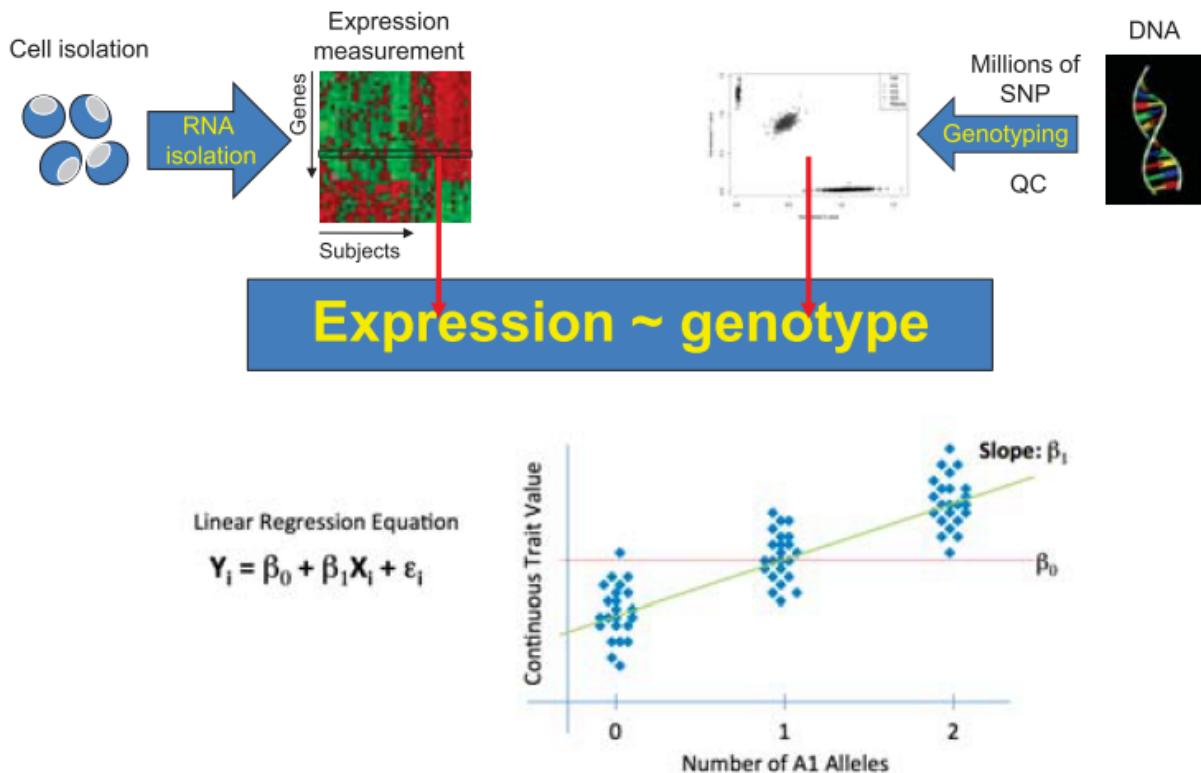


Figure 27.5: eQTL Study Approach

### 27.4.2 Considerations for Genomic Data

There are two main considerations for the analysis of genomic data: the minor allele frequency and the search radius. The **search radius** determines the generality of the effect being considered: an infinite search radius corresponds to a full-genome cis and trans-eQTL scan, while smaller radii restrict the analysis to cis-eQTLs. The **minor allele frequency** (MAF) determines the cutoff under which a SNP site is not considered: it is a major determinant of the statistical power of the study. A higher MAF cutoff generally leads to higher statistical power, but MAF and search radius interact in nonlinear ways to determine the number of significant alleles detected (see Figure 27.6).

### 27.4.3 Covariate Adjustment

There are many possible statistical confounders in an eQTL study, both biological and technical. Many biological factors can affect the observed expression of any given mRNA in an individual; this is exacerbated by the impossibility of controlling the testing circumstances of the large population samples needed to achieve significance. Population stratification and genomic differences between racial groups, age groups, and sex are additional contributing factors. Statistical variability also exists on the technical side. Even samples run on the same machine at different times show markedly different clustering of expression results. (Figure 27.7).

Researchers have successfully used the technique of **Principal Component Analysis** (PCA) to separate the effects of these confounders. PCA can produce new coordinate axes along which SNP-associated gene expression data has the highest variance, thereby isolating unwanted sources of consistent variation (see Chapter 20.4 for a detailed description of Principal Component Analysis). After extracting the principal

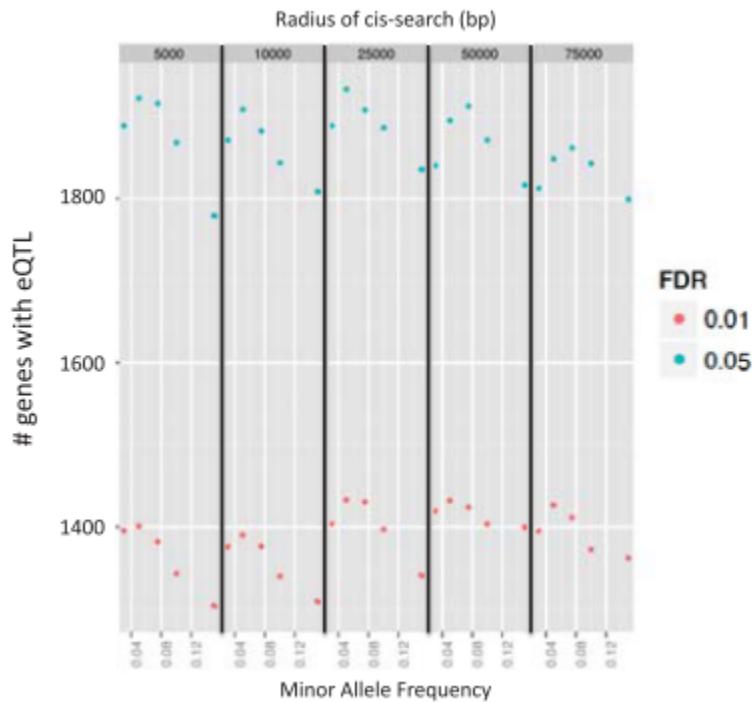


Figure 27.6: Effects of the search radius and MAF on the number of eQTL's detected

components of the gene expression data, we can extend the linear regression model to account for these confounders and produce a more accurate regression.

Figure X shows the effect of the number of principle components on the number of eQTL's discovered. Principle components and population covariation may be taken into account in an expanded eQTL model, such as the one shown below, where  $gPC_i$  refers to genotypic principal components, and  $ePC_i$  refers to expression phenotypic components:

$$Y_{ij} = \alpha + \beta_1 genotype + \beta_2 gender + \beta_3 age + \beta_4 gPC_1 + \beta_5 gPC_2 + \beta_6 ePC_1 + \beta_7 ePC_2 + \beta_8 ePC_3 \quad (27.1)$$

## FAQ

**Q:** Why is PCA an appropriate statistical tool to use in this setting and why do we need it?

**A:** Unfortunately, our raw data has several biases and external factors that will make it difficult to infer good eQTLs. However, we can think of these biases as being independent influences on the datasets that create artificial variance in the expression levels we see, confounding the factors that give rise to actual variance. Using PCA, we can decompose and identify these variances into their principal components, and filter them out appropriately. Also, due to the complex nature of the traits being analyzed, PCA can help reduce the dimensionality of the data and thereby facilitate computational analysis.

## FAQ

**Q:** How do we decide how many principal components to use?

**A:** This is a tough problem; one possible solution would be to try a different number of principal components and examine the eQTLs found afterwards - very this number for future tests by seeing whether the outputted eQTLs are viable. Note that it would be difficult to "optimize" different parameters for the eQTL study because each dataset will have an optimal number of principal components, a best value for MAF, etc...

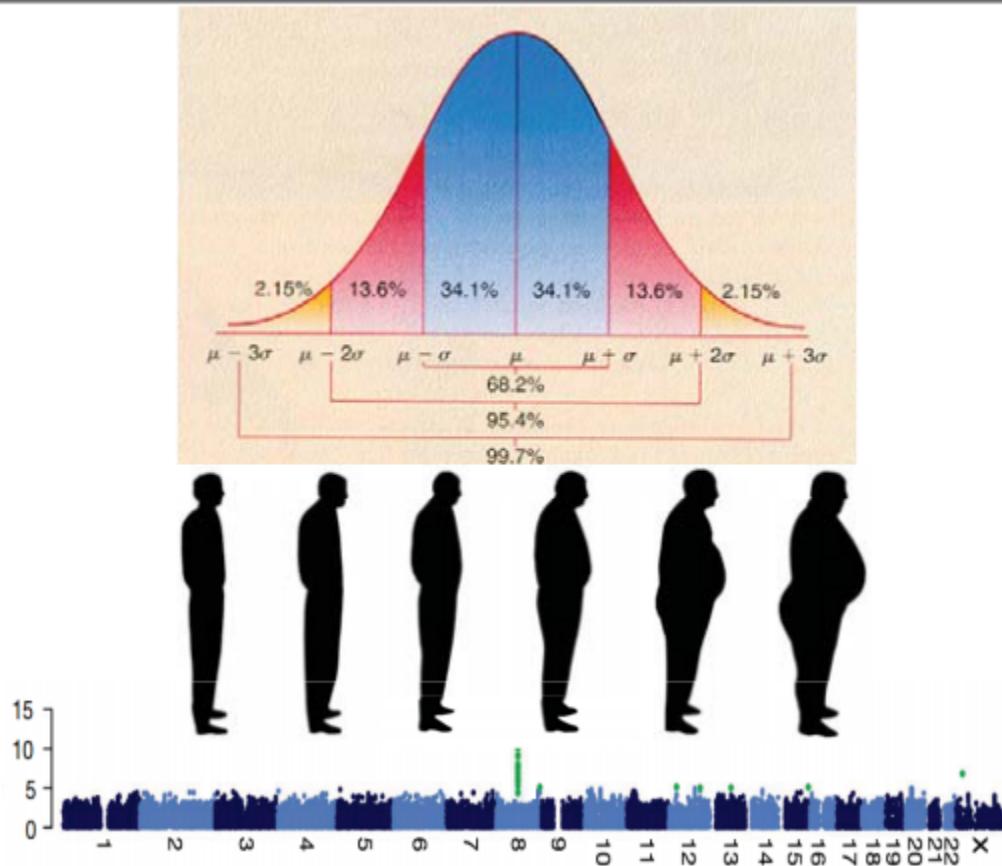


Figure 27.7: The effects of technical and population variance on expression level assays

### 27.4.4 Insights on eQTL's and biology

The following are some points to consider when conducting an eQTL study.

- The optimal strategy for eQTL discovery in a specific dataset out of all different ways to conduct normalization procedures, non-specific gene filtering, search radius selection, and minor allele frequency cutoffs may not be transferable to another eQTL study. Many scientists overcome this using *greedy tuning* of these parameters, running the eQTL study iteratively until a maximum number of significant eQTLs are found.
- It is important to note that eQTL studies only find *correlation* between genetic markers and gene expression patterns, and do not imply causation.

- When conducting an eQTL study, note that most significant eQTLs are found within a few kb of the regulated gene. In fact, there is a 34-fold enrichment of eQTL SNP's within 5 kb of a target gene. Because the pre-test probability of finding a distal eQTL is lower, these eQTL's have a higher significance threshold. This compounds the fact that the effect of distal eQTL's on differential gene expression is generally lower, making these eQTL's harder to find[4].
- Historically, it has been found that most eQTL studies are about 30-40% reproducible, and this is a relic of how the dataset is structured and the different normalization and filtering strategies the respective researchers use. However, eQTLs that are found in two or more cohorts consistently follows similar expression influence within each of the cohorts.
- Many eQTLs are tissue-specific; that is, their influence in gene expression could occur in one tissue but not in another, and a possible explanation of this is the co-regulation of a single gene by multiple eQTLs that is dependent on one gene having multiple alleles. It is estimated that approximately 30-35% of eQTL's are shared across tissues.
- co-regulation* is sometimes observed in eQTL's in which an eQTL regulates more than one gene. One example is the 17q locus, which contains SNPs correlated with differential regulation of the genes ZPB2, GSDMB, and ORMDL3 [4]
- eQTL's exhibit *disease enrichment*, as diseases are generally complex and polygenic rather than monogenic [4]. Hence, eQTL potential is a useful predictor of GWAS-tagging variants in GWAS analyses of disease phenotypes.
- Allele-specific eQTL's (*aseQTL*) analysis has been developed to analyze the effects of heterozygosity in eQTL analysis. Differential expression regulation due to a local eQTL may actually be due to a SNP in an eQTL on a homologous chromosome. Hence, *aseQTL* allows us to identify true cis and trans eQTL effects.

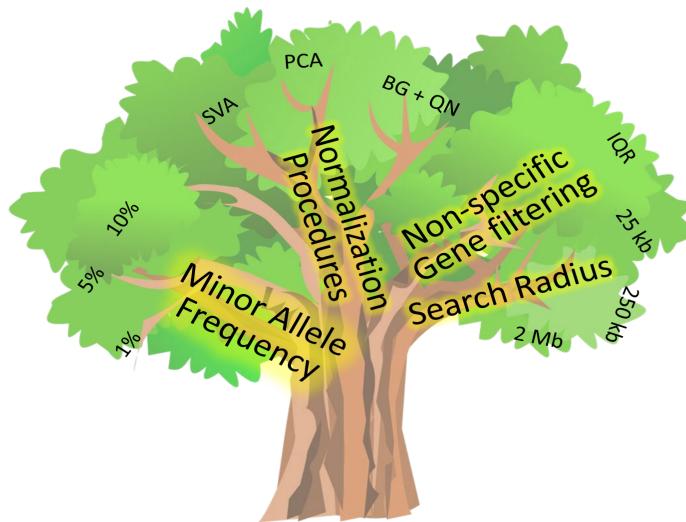


Figure 27.8: The decision parameters researchers must make when conducting an eQTL study

## 27.5 Current Research Directions

### 27.5.1 Quantifying Trait Variation

Because the study of eQTLs is a study in the level of expression of a gene, the primary step towards conducting an informative study is picking traits that have varying levels of expression rather than binary

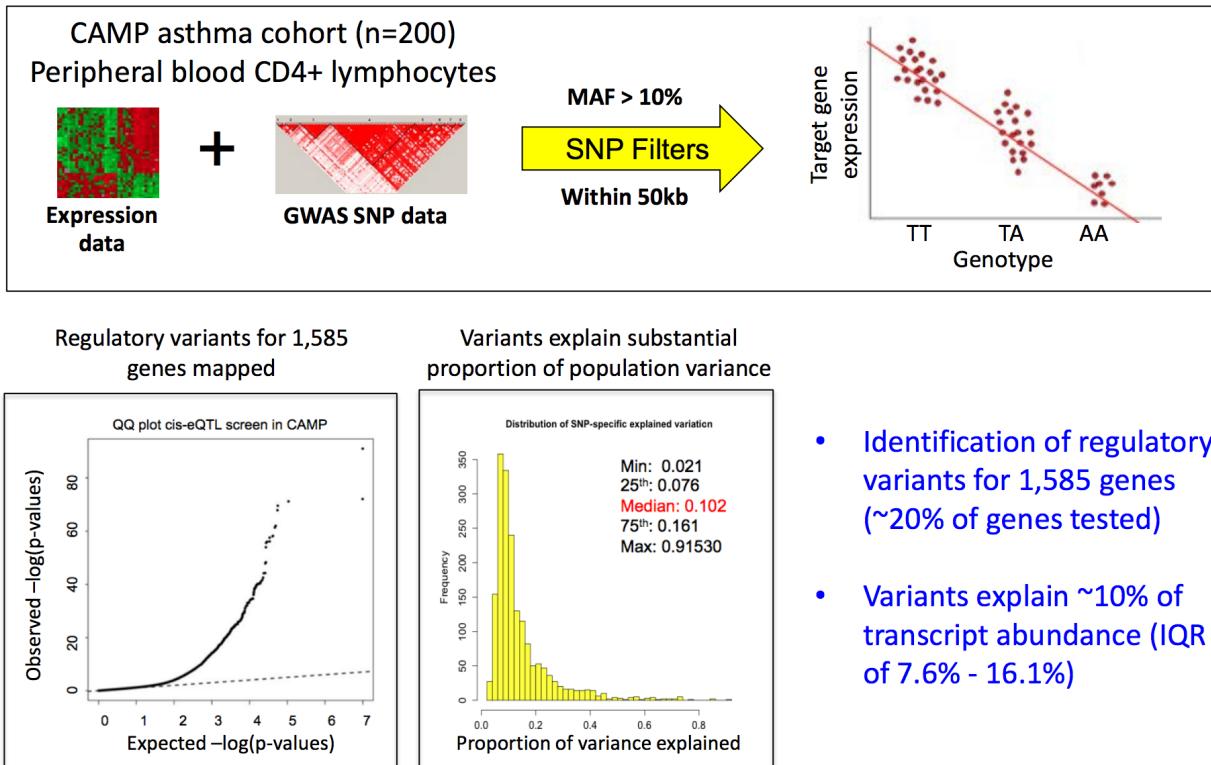


Figure 27.9: An example eQTL Study on asthma

expression. Examples of such quantitatively viable traits are body mass index (BMI) and height. In the late 1980's and early 1990's, the first studies of gene expression through genome-wide mapping studies were initiated by Damerval and de Vienne [8] [6]. However, their use of 2-D electrophoresis for protein separation was inefficient and not thoroughly reliable as it introduced a lot of noise and could not be systematically and quantitatively summarized. It was only in the early 2000s when the introduction of high-throughput array-based methods to measure mRNA incidence accelerated the successful use of this method, first highlighted in a study by Brem [10].

### 27.5.2 New Applications

There are two directions that eQTL studies are headed. First, there is a rush to use whole genome eQTL analysis to validate associations among variances in the human population such as differences in gene expression among ethnic groups, as the statistical power for being able to do so is beginning to reach the threshold of significance. A second direction of research seeks to dislocate genetic associations with varying phenotypes and and population differences based on a non-genetic basis. These non-genetic factors include environment, cell line preparation, and batch effects.

## 27.6 What Have We Learned?

In summary, most causal variation for complex polygenic diseases that we have discovered so far is noncoding. Moreover, phenotypic differences between species are not well explained by coding variation, while gene expression is highly heritable between generations. Thus, it is proposed that genetic control of expression levels are a crucial factor in determining phenotypic variance.

eQTLs are SNP variant loci that are correlated with gene expression levels. They come in one of two

forms. Cis-eQTLs are sites whose loci map to near the affected genes, are relatively easy to detect due to their proximity, and generally have clear mechanisms of action. Trans-eQTLs map to distance areas of the genome, are more difficult to detect, and their mechanisms are not as direct.

eQTL studies combine a whole-genome approach similar to GWAS with a expression assay, either microarray or RNA-seq. Expression levels of each gene are correlated by linear regression with genotypes after using PCA to extract confounding factors. Determining the optimal parameters for MAF, search radius, and confounder normalization is an open research question. Applications of eQTLs include the identification of disease-associated variants as well as variants associated with population subspecies and the genetic and environmental variance that gives rise to complex traits,

## 27.7 Further Reading

The following is a very good introductory literature review on eQTLs, including their history and current applications:

### **The role of regulatory variation in complex traits and disease**

Frank W. Albert and Leonid Kruglyak

*Nature Reviews Genetics* 16 2015

There are also some research papers that are trailblazers in what is current in eQTL studies. One such paper is informative on the occurrence of DNA methylation affecting gene expression in the human brain. Another is a study on changes in expression during development in the nematode *C. elegans*, using age as a covariate during eQTL mapping:

1. **Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain**  
Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, et al.  
*PLOS Genet* 6 2010
2. **The effects of genetic variation on gene expression dynamics during development**  
Francesconi, M. and Lehner, B.  
*Nature* 505 2013

In addition, eQTL variants have recently been found to be implicated in diseases such as Crohn's disease and multiple sclerosis [4].

As mentioned in Section 4.2, there have also been a recent surge in studies applying eQTL studies to delineating differences among human subpopulations and characterizing the contributions of the environment toward trait variation:

1. **Common genetic variants account for differences in gene expression among ethnic groups**  
Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, Cheung VG.  
*Nature Genetics* 2007
2. **Gene-expression Variation Within and Among Human populations**  
Storey JD, Madeoy J, Strout JL, Wurfel M, Ronald J, Akey JM  
*The American Journal of Human Genetics* 2007
3. **Population genomics of human gene expression [12]**  
Stranger BE, Nica AC, Forrest MS, et. al.  
*The American Journal of Human Genetics* 2007
4. **Evaluation of genetic Variation Contributing to Differences in Gene expression between populations**  
Zhang W, Duan S, Kistner EO, Bleibel WK, Huang RS, Clark TA, Chen TX, Schweitzer AC, Blume JE, Cox NJ, Dolan ME  
*The American Journal of Human genetics* 2008

**5. A Genome-Wide Gene Expression Signature of Environmental Geography in Leukocytes of Moroccan Amazighs**

Idaghdour Y, Storey JD, Jadallah SJ, Gibson G

PLoS 2008

**6. On the design and analysis of gene expression studies in human populations**

Joshua M Akey, Shameek Biswas, Jeffrey T Leek, John D Storey

Nature Genetics 2007

## 27.8 Tools and Resources

- The Costapas Lab distributes code to calculate CPMA from GWAS association p-values which can be found here: <http://www.cotsapaslab.info/index.php/software/cpma/>
- The Pritchard lab has several resources (found here: <http://eqtl.uchicago.edu/Home.html>) for eQTL research and gene regulation including:
  - DNase-seq data from 70 YRI lymphoblastoid cell lines
  - Downloading positions of transcription factor binding sites inferred in the HapMap lymphoblastoid cell lines by CENTIPEDE
  - Raw and mapped RNA-Seq data from Pickrell et al.
  - Assorted scripts for identifying sequencing reads covering genes, polyadenylation sites and exon-exon junctions
  - Data and meQTL results for Illumina27K methylation data in HapMap lymphoblastoid cell lines.
  - Files to ignore areas of the genome that are prone to causing false positives in ChIP-seq and other sequencing based functional assays
  - Browser for eQTLs identified in recent studies in multiple tissues
- The Wellcome Trust Sanger Institute has developed packaged database and web services (*Genevar*) that are designed to help integrative analysis and visualization of SNP-gene associations in eQTL studies. This information can be found here: <http://www.sanger.ac.uk/resources/software/genevar/>
- The Wellcome Trust Sanger Institute has also developed databases that contain information relevant to eQTL studies such as finding and identifying all functional elements in the human genome sequence and maintaining automatic annotation on selected eukaryotic genomes. This information can be found here: <http://www.sanger.ac.uk/resources/databases/>.
- Finally, the NIH is progressing on the Genotype-Tissue Expansion Project (GTEx). Currently, the project stands at 35 tissues from 50 donors; the aim is to acquire and analyze 20,000 tissues from 900 donors, with the hope of gathering even more data for further genetic analyses, especially for eQTL and trans-eQTL analyses that require larger sample sizes.

## Bibliography

- [1] King, Mary-Claire and Wilson, A.C. (April 1975) *Evolution at Two Levels in Humans and Chimpanzees* Science Vol.188 No. 4184
- [2] 1000 Genomes Project Consortium. Nature. 2010; 467:1061-73.
- [3] Cheung Vivien G. and Spielman Richard S. (2009) *Genetics of Human Gene Expression: Mapping DNA Variants that Influence Gene Expression* Nature Reviews Genetics
- [4] C. Cotsapas, *Regulatory variation and eQTLs*. 2012 Nov 1.
- [5] C. Cotsapas, BF Voight, E Rossin, K Lage, BM Neale, et al. (2011) *Pervasive Sharing of Genetic Effects in Autoimmune Disease*. PLoS Genet 7(8):e1002254. doi:10.1371/journal.pgen.1002254

- [6] Damerval C, Maurice A, Josse JM, de Vienne D (May 1994). *Quantitative Trait Loci Underlying Gene Product Variation: A Novel Perspective for Analyzing Regulation of Genome Expression* Genetics 137 (1): 289–301.PMC 1205945. PMID 7914503.
- [7] Dimas AS , et. al. (Sept. 2009) *Common regulatory variation impacts gene expression in a cell type-dependent manner*. Science 325(5945):1246-50. 2 Epub 2009 Jul 30.
- [8] D. de Vienne, A. Leonardi, C. Damerval (Nov 1988). *Genetic aspects of variation of protein amounts in maize and pea*. Electrophoresis 9 (11): 742–750. doi:10.1002/elps.1150091110. PMID 3250877.
- [9] Shengjie Yang, Yiyuan Liu, Ning Jiang, Jing Chen, Lindsey Leach, Zewei Luo, Minghui Wang. *Genome-wide eQTLs and heritability for gene expression traits in unrelated individuals*. BMC Genomics 15(1): 13. 2014 Jan 9.
- [10] Rachel B. Brem and Leonid Kruglyak. *The landscape of genetic complexity across 5,700 gene expression traits in yeast*. PNAS 102(5): 1572–1577. 23 Nov 2004.
- [11] Michael Morley, Cliona M. Molony, Teresa M. Weber, James L. Devlin, Kathryn G. Ewens, Richard S. Spielman, Vivian G. Cheung. *Genetic analysis of genome-wide variation in human gene expression*. Nature 430: 743-747. 12 Aug 2004.
- [12] Barbara E Stranger, Alexandra C Nica, Matthew S Forrest, Antigone Dimas, Christine P Bird, Claude Beazley, Catherine E Ingle, Mark Dunning, Paul Flicek, Daphne Koller, Stephen Montgomery, Simon Tavaré, Panos Deloukas, Emmanouil T Dermitzakis. *Population genomics of human gene expression*. Nature Genetics 39: 1217 - 1224. 16 Sep 2007.

## Bibliography



---

CHAPTER  
**TWENTYEIGHT**

---

PERSONAL GENOMES, SYNTHETIC GENOMES, COMPUTNG IN C  
VS. SI

Guest Lecture by George Church  
Scribed by Lawson Wong (2011)

## 28.1 Introduction

George Church discussed a variety of topics that have motivated his past and present research. He first discussed about reading and writing genomes, including his own involvement in the development of sequencing and the Human Genome Project. In that latter half, he discussed about his more recent endeavor, the Personal Genome Project, which he initiated in 2005.

## 28.2 Reading and Writing Genomes

As a motivation, consider the following question: Is there any technology that is not biologically motivated or inspired? Biology and our observations of it influence our lives pervasively. For example, within the energy sector, biomass and bioenergy has always existed and is increasingly becoming the focus of attention. Even in telecommunications, the potential of quantum-level molecular computing is promising, and is expected to be a major player in the future.

Church has been involved in molecular computing in his own research, and claims that once harnessed, it has great advantages over their current silicon counterparts. For example, molecular computing can provide at least 10% greater efficiency per Joule in computation. More profound perhaps is its potential effect on data storage. Current data storage media (magnetic disk, solid-state drives, etc.) is much less (billions times) dense than DNA. The limitation of DNA as data storage is that it has a high error rate. Church is currently involved in a project exploring reliable storage through the use of error correction and other techniques.

In a 2009 Nature Biotechnology review article [1], Church explores the potential for efficient methods to read and write to DNA. He observes that in the past decade there has been a  $10\times$  exponential curve in both sequencing and oligo synthesis, with double-stranded synthesis lagging behind but steadily increasing. Compared to the  $1.5\times$  exponential curve for VLSI (Moore's Law), the increase on the biological side is more dramatic, and there is no theoretical argument yet for why the trend should taper off. In summary, there is great potential for genome synthesis and engineering.

### **Did You Know?**

George Church was an early pioneer of genome sequencing. In 1978, Church was able to sequence plasmids at \$10 per base. By 1984, together with Walter Gilbert, he developed the first direct genomic sequencing method [3]. With this breakthrough, he helped initiate the Human Genome Project in 1984. This proposal aimed to sequence an entire human haploid genome at \$1 per base, requiring a total budget of \$3 billion. This quickly played out into the well-known race between Celera and UCSC-Broad-Sanger. Although the latter barely won in the end, their sequence had many errors and gaps, whereas Celera's version was much higher quality. Celera initially planned on releasing the genome in 50 kb fragments, which researchers could perform alignments on, much like BLAST. Church once approached Celera's founder, Craig Venter, and received a promise to obtain the entire genome on DVD after release. However, questioning the promise, Church decided instead to download the genome directly from Celera by taking advantage of the short fragment releases. Using automated crawl and download scripts, Church managed to download the entire genome in 50 kb fragments within three days!

## **28.3 Personal Genomes**

In 2005, George Church initiated the Personal Genome Project [2]. Now that sequencing costs have rapidly decreased to the point that we can currently get the entire diploid human genome for \$4000 (compare to \$3 billion for a haploid human genome in the Human Genome Project), personal genome and sequence information is becoming increasingly affordable.

One important application for this information is in personalized medicine. Although many diseases are still complicated to predict, diagnose, and study, we currently already have a small list of diseases that are highly predictable from genome data. Examples include phenylketonuria (PKU), BRCA-mutation-related breast cancer, and hypertrophic cardiomyopathy (HCM). Many of these and similar diseases are uncertain (sudden onset without warning symptoms) and not normally checked for (due to their relative rareness). As such, they are particularly suitable as targets for personalized medicine by personal genomes, because genomic data provide accurate information that otherwise cannot be obtained. Already, there are over 2500 diseases (due to  $\sim 6000$  genes) that are highly predictable and medically actionable, and companies such as 23andMe are exploring these opportunities.

As a final remark on the subject, Church remarked on some of his personal philosophy regarding personalized medicine. He finds many people reluctant to obtain their genomic information, and attributes this to a negative view among the general public toward GWAS and personalized medicine. He thinks that the media focuses too much on the failure of GWAS. The long-running argument against personalized medicine is that we should focus first on common diseases and variants before studying rare events. Church counterargues that in fact there is no such thing as a common disease. Phenomena such as high blood pressure or high cholesterol only count as symptoms; many 'common diseases' such as heart disease and cancer have many subtypes and finer categories. All along, lumping these diseases into one large category only has the benefit of teaching medical students and to sell pharmaceuticals (e.g., statins, which have fared well commercially but only benefit very few). Church argues that lumping implies a loss of statistical power, and is only useful if it is actually meaningful. Ultimately, everyone dies due to their own constellation of genes and diseases, so Church sees that splitting (personalized genomics) is the way to proceed.

Personal genomics provide information for planning and research. As a business model, it is analogous to an insurance policy, which provides risk management. As an additional benefit however, the information received allows for early detection, and consequences may even be avoidable. Access to genomic information allows one to make more informed decisions.

## **28.4 Current Research Directions**

## **28.5 Further Reading**

Personal Genome Project: <http://www.personalgenomes.org/>

## 28.6 Tools and Techniques

## 28.7 What Have We Learned?

## Bibliography

- [1] Peter A. Carr and George M. Church. Genome engineering. *Nature biotechnology*, 27(12):1151–1162, December 2009.
- [2] G. M. Church. The Personal Genome Project. *Molecular Systems Biology*, 1(1):msb4100040–E1–msb4100040–E3, December 2005.
- [3] G. M. Church and W. Gilbert. Genomic sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 81(7):1991–1995, April 1984.



---

CHAPTER  
TWENTYNINE

---

PERSONAL GENOMICS

Deniz Aksel, Molly Schmidt, Jonathan Uesato

**Figures**

---

29.1 Factors that contribute to the probability of getting a disease. Each relationship shown represents correlation except for the link between genome and disease. Correlation does not mean causality, but we can use the genome to resolve causality. . . . .	456
29.2 SNPs associated with Mendelian diseases often lie in coding regions whereas those associated with polygenic diseases are usually found in non-coding regions. This is because large effect variants, protein coding variants, associated with Mendelian diseases are at low frequencies due to selection. Common variants associated with polygenic diseases, tend to have a lower effect, so selection does not play as big of a role. . . . .	457
29.3 The multiple factors and datasets in determining the role of methylation on disease states, and methods for linking these datasets. . . . .	459
29.4 Modeling Human Disease . . . . .	460
29.5 Polygenic Risk Prediction . . . . .	461
29.6 A diagram of how the true patient state is translated into the raw EHR data that informs our health care process and knowledge of disease . . . . .	462
29.7 Assocations found through PheWAS, GWAS, and a combination of both [? ] . . . . .	462
29.8 Correlations between varying phylogenetic traits found from cross-trait LD score regression [? ] . . . . .	463
29.9 Example of an EHR matrix with and without missing indicators . . . . .	463
29.10Example of an RBM . . . . .	464
29.11Comparison of site-wide and specific polls for music rating [? ] . . . . .	464
29.12Joint model of data and NMAR with ratings [? ] . . . . .	465
29.13Finding patterns in longitudinal EHR data [? ] . . . . .	465

---

## 29.1 Introduction

Personalized genomics focuses on the analysis of individuals' genomes and their predispositions for diseases rather than looking at the population level. Personalized medicine is only possible with information about genetics along with information about many other factors such as age, nutrition, lifestyle, or epigenetic markers (such as methylation). To make personalized medicine more of a reality, we need to learn more about the causes and patterns of diseases in populations and individuals.

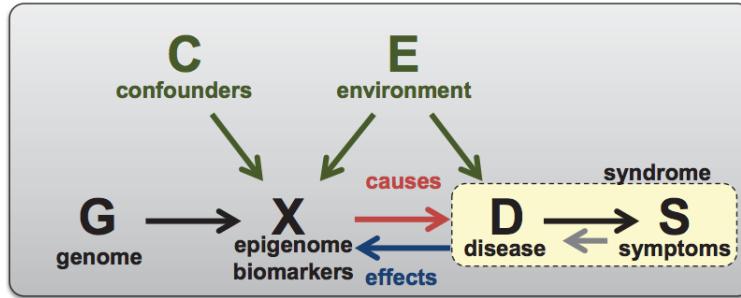


Figure 29.1: Factors that contribute to the probability of getting a disease. Each relationship shown represents correlation except for the link between genome and disease. Correlation does not mean causality, but we can use the genome to resolve causality.

## 29.2 Epidemiology: An Overview

Epidemiology is the study of patterns, causes, and effects of health and disease conditions in defined populations. In order to talk about epidemiology, we need to first understand some basic definitions and terms: **Morbidity level** is how sick an individual is whereas mortality is whether an individual is dead or not. The **incidence** is a rate which describes the number of new cases/people with a disease that appears during a period of time. The **prevalence** is the total steady state number of cases in the population. The **attributable risk** is the difference in rate of a disease between those exposed to the disease and those not exposed to the disease. **Population burden** refers to the years of potential life lost (YPLL), quality-adjusted or disability-adjusted life year (QALY/DALY). **Syndrome** refers to co-occurring signs or symptoms of a disease that are observed. The **prevention challenge** is to determine a disease and its cause and understand whether, when, and how to intervene.

In order to determine disease causes, studies must be designed according to certain principles of experimental design. These principles include control, randomization, replication, grouping, orthogonality, and combinatorics. Control groups are needed so that comparison to a baseline can be done. The placebo effect is real, so having a control group is necessary. The people who get the putative treatment being tested must also be random so that there is no bias. The study needs to be replicated as well in order to control for variability in the initial sample. (This is similar to the “winners curse.” Someone may win a race because they did outstanding in that particular round and surpassed their personal average, but in the next round they probably will regress back to performing close to their average.) Understanding variation between different subgroups may also play a large role in the outcomes of experiments. These may include subgroups based on age, gender, or demographics. One subgroup of the population may be contributing in a more profound way than the rest, so looking at each subgroup specifically is important. Orthogonality, or the combination of all factors and treatments, and combinatorics, factorial design, must also be taken into account when designing an experiment. With disease studies in particular, ethics when dealing with human subjects must be taken into account. There are legal and ethical constraints which are overseen by review boards. Clinical trials must be performed either blind (patient does not know if they are getting treatment or not) or double-blind (doctor also doesn't know). A patient who knows if they have gotten a treatment may change their habits causing bias, or a doctor who knows a patient got the treatment may treat them differently or analyze their results differently. Both considerations need to be taken into account to lower the bias that may cause different results of a clinical trial.

**Example** An example of the need for a randomized control trial is the treatment of ebola. A treatment must be distributed randomly to individuals being treated in different hospitals and it must be blind. If someone believes they are getting the vaccine, they may alter their habits to protect themselves which may affect the outcome. If only patients of one hospital get the vaccine, there is a possibility that the effects seen are just from that hospital being more careful.

## FAQ

**Q:** In poorly designed experiments, is there one aspect that is most commonly overlooked?

**A:** The most commonly missed is subgroup structure. It is sometimes not obvious what the different subgroups could be. To help with this, researchers can look at general properties of a predictor by trying to cluster cases and controls independently and visualize the clustering. If there is substructure other than case/control in the clustering, researchers can look for variables within each cluster to see what is driving substructure.

### 29.3 Genetic Epidemiology

Genetic epidemiology focuses on the genetic factors contributing to disease. Genome-Wide association studies (GWAS), previously described in depth, identify genetic variants that are associated with a particular disease while ignoring everything else that may be a factor. With the decrease of whole genome sequencing, these types of studies are becoming much more frequent.

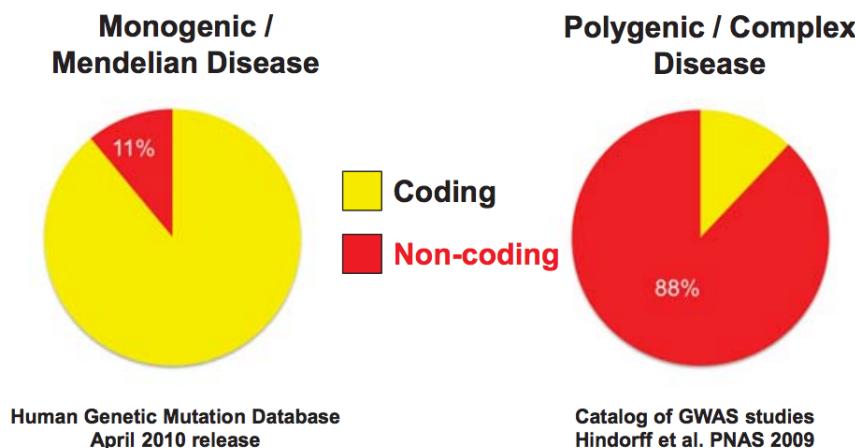


Figure 29.2: SNPs associated with Mendelian diseases often lie in coding regions whereas those associated with polygenic diseases are usually found in non-coding regions. This is because large effect variants, protein coding variants, associated with Mendelian diseases are at low frequencies due to selection. Common variants associated with polygenic diseases, tend to have a lower effect, so selection does not play as big of a role.

In genetic epidemiology there are many genetic factors that you can test to identify diseases in a particular individual. You can look at family risk alleles which are inherited with a common trait in specific genes or variants. You can study monogenic, actionable protein-coding mutations which are the most understood, would have the highest impact, and would be the easiest to interpret. There is the possibility of testing all coding SNPs (single nucleotide polymorphisms) with a known disease association. There are debates over whether a patient needs to or would want to know this information sometimes especially if the disease is not treatable. A person's quality of life may decrease just from knowing they may have the untreatable disease even if no symptoms are exhibited. You can also test all coding and non-coding associations from GWAS, all common SNPs regardless of association to any disease, or the whole genome.

### Did You Know?

23andMe is a personal genomics company that offers saliva-based direct-to-consumer genome tests. 23andMe gives consumers raw genetic data, ancestry-related results, and estimates of predisposition for more than 90 traits and conditions. In 2010, the FDA notified several genetic testing companies, including 23andMe, that their genetic tests are considered medical devices and federal approval is required to market them. In 2013, the FDA ordered 23andMe to stop marketing its Saliva Collection Kit and Personal Genome Service (PGS) as 23andMe had not demonstrated that they have “analytically or clinically validated the PGS for its intended uses” and the “FDA is concerned about the public health consequences of inaccurate results from the PGS device” [1]. The FDA expressed concerns over both false negative and false positive genetic risk results, saying that a false positive may cause consumers to undergo surgery, intensive screening, or chemoprevention in the case of BRCA-related risk, for example, while a false negative may prevent consumers from getting the care they need. In class, we discussed whether people should be informed about potential risk alleles they may carry. Often, people may misunderstand the probabilities provided to them and either underestimate or overestimate how concerned they should be. The argument was also raised that people should not be told they are at risk if there is nothing current medicine and technology can do to mitigate the risk. If people are going to be informed about a risk, the risk should be actionable; i.e. they should be able to do something about it, instead of just live in worry, as that added stress may cause other health problems for them.

Not only is there the choice of what to test, there is the question of when to test someone for a particular condition. Diagnostic testing occurs after symptoms are displayed in order to confirm a hypothesis or distinguish between different possibilities of having a condition. You can also test predictive risk which occurs before symptoms are even shown by a patient. You may test newborns in order to intervene early or even do pre-natal testing via an ultrasound, maternal serum, probes or chorionic villus sampling. In order to test which disorders you may pass on to your child, you can do pre-conception testing. You can also do carrier testing to determine if you are a carrier of a particular mutant allele that may run in your family history. Testing genetics and biomarkers can be tricky because it can be unknown if the genetics or biomarker seen is causing the disease or is a consequence of having the disease.

To interpret disease associations, we need to use epigenomics and functional genomics. The genetic associations are still only probabilistic: if you have a genetic variant, there is still a possibility that you will not get the disease. Based on Bayesian statistics however, the posterior probability increases if the prior increases. As we find more and more associations and variants, the predictive value will increase.

## 29.4 Molecular Epidemiology

Molecular Epidemiology involves looking at the molecular biomarkers of a disease state. This includes looking at gene expression profiles, DNA methylation patterns i.e. epigenomics, and chromatin structure and organization in specific cell types. In earlier chapters, we discussed the link between gene expression (as RNA or proteins) and SNPs in the context of eQTL studies. As a reminder, eQTLs (expression quantitative trait loci) seek linear correlations between gene expression levels and different variants of a genetic locus.

This section will focus on understanding the role of **epigenomic markers** as molecular indicators of a disease. It is important to understand that multiple factors, and thus multiple datasets come into play in understanding the epigenomic basis of disease: methylation patterns of sample patients (M), genomic information (G) for the same individuals, environmental data (E, covering covariates like age, gender, smoking habits etc.), and phenotype quantifications (P, can capture multiple phenotypic markers, for example in Alzheimer's Disease, the number of neuronal plaques per patient). Furthermore, we need to understand the various interconnections and dependencies between these data sets to make meaningful conclusions about the influence of methylation for a certain disease.

To remove experimental, technical or environmental covariants, we rely on either known, or ICA (Independent component analysis)-inferred corrections. To link genetic data to methylation patterns, we look for meQTLs (methylation quantitative trait loci), which is equivalent to eQTLs. Molecular phenotypes such as expression level or methylation level are also quantitative traits. Finally, to link methylation patterns with diseases, we implement EWAS (Epigenome-wide association studies).

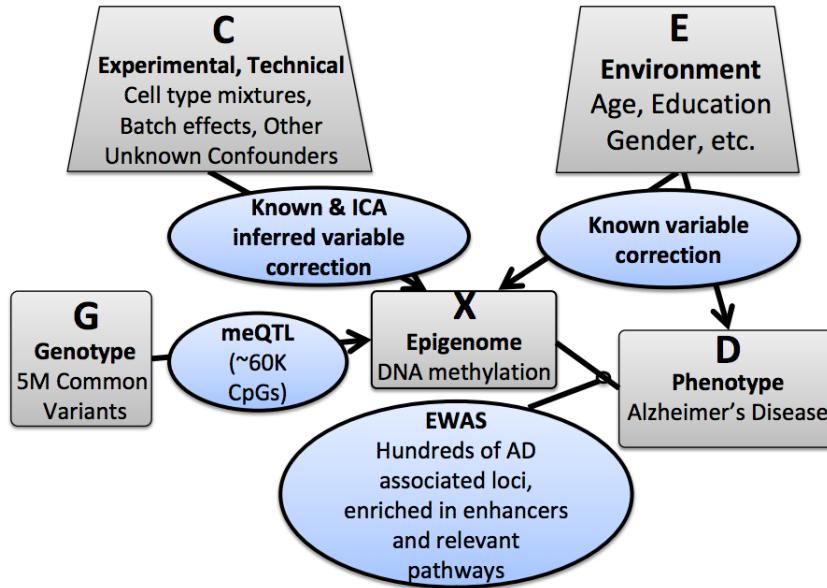


Figure 29.3: The multiple factors and datasets in determining the role of methylation on disease states, and methods for linking these datasets.

#### 29.4.1 meQTLs

The discovery of meQTLs follows a process that is highly similar to the methodology used for discovering eQTLs. To discover cis-meQTLs (i.e. meQTLs where the effect on methylation is proximal to the tested locus) we select a genomic window, and use a linear model to test whether or not we see a correlation between methylation and SNP variants in that region. We test to see if the correlation is significant via an F-test, where our null hypothesis is that the additional model complexity introduced via the genomic information does not explain a significant portion of variation in methylation patterns. Other methods of discovering meQTLs include permutation and Linear Mixed Models (LMM).

**Example** An example of using meQTLs in discovering the connection between methylation, genotype, and disease is the Memory and Aging Project. 750 elderly people enrolled in the project many years ago and today, they have mostly died and given their brain to science. The genotype and methylation of the dorsal lateral prefrontal cortex were determined in order to study the connection between methylation and the phenotype of Alzheimer's and how the genotype may affect the methylation profile. SNP data, methylation, environmental factors (such as age, gender, sample batch, smoking status, etc..), and phenotype were taken into account. First covariants needed to be discovered and excluded to make sure the results obtained are not due to confounding factors. This is done by decomposing the matrix of methylation data by doing ICA. This enables the discovery of variables that are driving the most variability in the trait. The batch sample and cell mixture can have the biggest effect in the variation between individuals. After this is corrected for, linear models, permutation tests, and linear mixed models are used to determine cis-meQTLs—how much the genotype explains the methylation level.

### 29.4.2 EWAS

Epigenome-Wide Genome Studies (EWAS) aim to find connections between the methylation pattern of a patient and their phenotype. Much like GWAS, EWAS relies on linear models and p-value testing for finding linkages between epigenomic profiles and disease states. Together with meQTLs, EWAS can also potentially shine light on whether a given methylation pattern is the cause or result of a disease. Ideally, the idea is to be able to generate models that allow us to predict disease states (phenotypes) based on methylation.

There are some drawbacks to EWAS. First, the variance in methylation patterns due to phenotype is typically very small, making it difficult to link epigenomic states to disease states, similar to seeking a needle in a haystack. To improve this situation, we need to control for other sources of variance in our methylation data, such as gender, age etc. Gender, for example, incorporates a large variance for the case of Alzheimer's Disease. We additionally need to account for variance due to genotype (in the form of meQTLs). Additionally, variability across samples is a major issue in collecting methylation data for EWAS[2]. As different cell types in the same individual will have different epigenomic signatures, it is important that relevant tissue samples are collected, and the data is corrected for the different cell/tissue types involved in a study.

## 29.5 Causality Modeling and Testing

A central question for personal genomics is the question of which markers are causal of disease. For example, one might ask whether methylation at a certain loci, or a certain histone modification, increases a person's risk for a certain disease. This question is difficult because we need to separate spurious correlations from causal effects - for example, it is possible that a mutation elsewhere in the genome causes the disease, and also increases the chance of observing a particular marker, but that the marker has no causal effect on the disease. In this case, we would find a correlation between the disease phenotype and the presence of the marker despite the lack of any causal effect.

The key insight that allows us to determine causal effects, as opposed to mere correlations, is the observation that while the genotype may influence a person's risk for a particular disease, the disease will not modify a person's genotype. This allows us to use genotype as an instrumental variable for methylation. This limits the number of possible models so that we can statistically test which model is most consistent with the observed data.

There are three possibilities for modeling complex human diseases: the independent associations model, the interaction model, and the causal pathway model, depicted in Figure 29.4. We will use the example of studying the causal relationship between methylation at a certain loci and disease to demonstrate how to test for a causal effect.

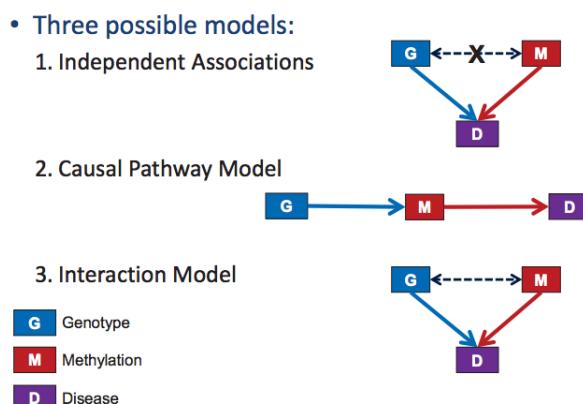


Figure 29.4: Modeling Human Disease

Under the independent associations model, the data should contain no correlation between the genotype and the disease, which distinguishes this model from the interaction and causal pathway models. However, there will be correlations between each of the factors and the disease separately. Thus, this model is straightforward to test for. An example of this would be two independent risk genes.

Under the interaction model, factor B's effect on a disease may vary depending on the value for A. For example, a drug's effect on someone can be different based on their genotype. To test for this, we determine the statistical significance of the effect of the interaction term,  $\beta_2$ , in the regression  $D = \beta_0A + \beta_1B + \beta_2A*B + c$ . If there is a significant interaction effect, we can isolate the separate effects by stratifying across different levels of A.

The causal pathway model is a little more complex. If we notice a correlation between a risk factor and a disease, we may wonder whether there is a direct link between risk factor A and a disease, or does the risk factor A affect risk factor B which then affects the disease. In the case that risk factor A only has an effect on the disease through B, we will observe that after conditioning upon B, the correlation between A and D disappears, that is, B “mediates” this interaction. In reality, the effect of A on a disease is usually only partially mediated through B, so we can instead look for if the effect size of A on the disease is decreased when B is observed.

### 29.5.1 Polygenic Risk Prediction

One of the most central questions of personal genomics is prediction of genetic predispositions to various genetic traits, using multiple genes to inform our predictions. The basic approach is explained in Figure 29.5. First, the data set is divided into a training and test set, and in the training cohort, we select which SNPs are most important and their appropriate weightings. Then we use the test set to evaluate the accuracy of our predictions. Finally, we use this model to predict genetic predispositions for the target cohort by using the confidences we determined from the test set.

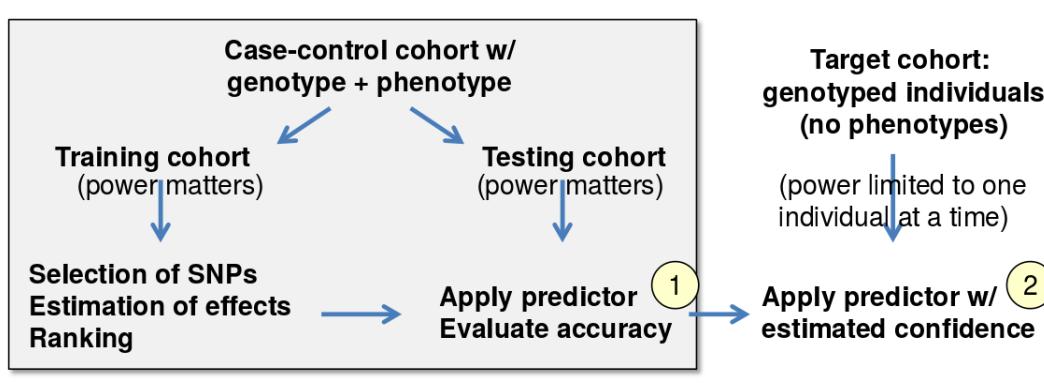


Figure 29.5: Polygenic Risk Prediction

## 29.6 Current Research Directions

An exciting new direction for personal genomics is PheWAS: phenome-wide association studies. These studies examine relationships between various phenotypic attributes, in contrast to GWAS, which are more focused on the genetic level. In this section we discuss the advantages of these PheWAS and how these studies are conducted.

### 29.6.1 PheWAS motivation

So what do PheWAS give us that other types of studies don't? Because of their focus on phenotypes, they allow us to draw associations of otherwise unknown or underappreciated phenotypes with well-studied

ones, exposing new relationships. This also gives additional insight into biological mechanisms that create phenotypes, as we can examine easy-to-measure 'biomarker' phenotypes. Finally, they can boost the power of GWAS, allowing us to better infer causal variants and consistent pathways.

Another major advantage is the ability to use the Electronic Health Record, or EHR, as a source of data to conduct these studies. The EHR contains extremely **rich** data on each individual patient, with such information as prescription data, laboratory data, and digital imaging files. We are then able to leverage the wealth of data to perform better phenotypic studies. However, we must be careful with such data, as it is only an indirect reflection of the true state of a patient due to the recording process.

to replace  
e



Figure 1: Dummy

Lorum ipsum dolor sit amet, consecetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Cumshir dictum gravidae manus non accumsan. Nam arcu libero, nonummy eget, consetetur id, vulputate a, magna. Donec vehicula augue eu neque. Nullentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus arcu. Nulla et lectus vestibulum orci fringilla ultrices. Phasellus

Figure 29.6: A diagram of how the true patient state is translated into the raw EHR data that informs our health care process and knowledge of disease

### 29.6.2 Modeling PheWAS with genetic information

PheWAS can be used to both confirm GWAS associations and also discover new associations, which can be seen in figure 7. In the figure, red markers are associations found with only GWAS, blue markers from only PheWAS, and green markers from both. It is clear the combination of the two techniques allows us to find more associations and develop a better understanding of these relationships.

to replace  
e



Figure 1: Dummy

Lorum ipsum dolor sit amet, consecetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Cumshir dictum gravidae manus non accumsan. Nam arcu libero, nonummy eget, consetetur id, vulputate a, magna. Donec vehicula augue eu neque. Nullentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus arcu. Nulla et lectus vestibulum orci fringilla ultrices. Phasellus

Figure 29.7: Associations found through PheWAS, GWAS, and a combination of both [? ]

Insights into phenotypic associations can be accomplished through cross-trait LD score regression. In one such study, 276 genetic correlations of 24 traits were examined, finding correlations between anorexia, depression, and schizophrenia (see figure 8). This study demonstrated the power of PheWAS, as it was able to leverage information from genome-wide signals, rather than a single SNP, to find these correlations.



Figure 1: Dummy

...ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravidae  
mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna.  
Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus  
et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra  
metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultricies. Phasellus...

Figure 29.8: Correlations between varying phylogenetic traits found from cross-trait LD score regression [?]

### 29.6.3 Modeling PheWAS with EHR data

Oftentimes, genetic information is not available to use for PheWAS, especially over a large patient cohort. However, given the causal mediating phenotypes, the presence of some end disease is independent from the underlying genotype, allowing us to use the mediating phenotypes as a replacement when performing studies. Getting the data for these phenotypes is done through the EHR, which holds such data as diagnoses, prescriptions, and procedures. This EHR data is highly sparse, with many entries non-marked. In fact, we describe the EHR data as **non-missing at random**, which is a specific description of the underlying distribution of the missing data.

Find a figure  
missing figure



Figure 1: Dummy

...ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravidae  
mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna.  
Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus  
et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra  
metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultricies. Phasellus...

Figure 29.9: Example of an EHR matrix with and without missing indicators

### 29.6.4 Missing Mechanism

There are 3 possible ways to describe missing data. First is missing completely at random, where there is no underlying mechanism that dictates why certain data is missing. The second is missing at random (MAR), where the missing status depends only on the values we have already observed. Models that assume MAR are simpler, but sacrifice accuracy if the assumption is incorrect. Finally, there is non-missing at random (NMAR), where the missing status depends on the non-observed values as well. EHR data is NMAR, and requires more sophisticated models to perform accurate inference. Our goal is then to create models that can **impute**, or fill out, these missing data fields so we can then infer relationships between phenotypes that have been measured in patients and underlying phenotypes that may not have been observed yet.

Find a figure  
missing figure

### 29.6.5 EHR with MAR

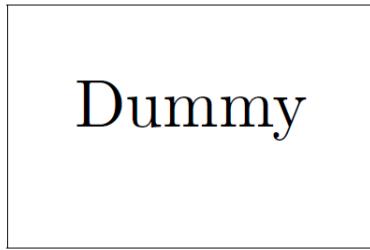


Figure 1: Dummy

...ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. ...

Figure 29.10: Example of an RBM

### 29.6.6 EHR with NMAR

In our real datasets, we believe our data is non-missing at random. Making the correct assumption about the distribution of missing data is very significant in obtaining accurate results. For example, in figure 10 the distribution of the ratings of music from a random site-wide poll and a specific user poll are compared, where the site-wide poll is NMAR while the specific user poll is MAR. It is clear the distributions differ heavily, and making the wrong assumption can lead to biased, incorrect results.

to replace  
e

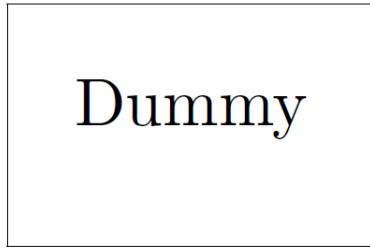


Figure 1: Dummy

...ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. ...

Figure 29.11: Comparison of site-wide and specific polls for music rating [? ]

To account for this, one approach that has been taken by Marlin & Zemel was to jointly model the data and the NMAR status. A successful model they found assumed the data was MAR, and used a multinomial mixture model to allow for the distribution of missing data to depend on their underlying rating and identity. Using such a model, they were able to achieve higher accuracy than any model that only had the MAR assumption.

to replace  
e

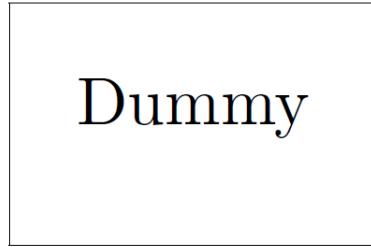


Figure 1: Dummy

LOREM IPSUM DOLOR SIT AMET, CONSECTETUR ADIPISCING ELIT. UT PURUS ELIT, VESTIBULUM UT, PLACERAT AC, ADIPISCING VITAE, FELIS. CURABITUR DICTUM GRAVIDA MAURIS. NAM AREU LIBERO, NOMNMMY EGIT, CONSECTETUR ID, VULPITATE A, MAGNA. DONEC VEHICULA AUGUE EU NEQUE. PELLentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus

Figure 29.12: Joint model of data and NMAR with ratings [? ]

Similar models could thus be adapted to EHR data to capture the NMAR property and allow for better imputation of missing data.

### 29.6.7 Modeling longitudinal EHR data

Another significant use case for EHR data is finding patterns over time for patient phenotypes. Detecting these patterns would allow for new insights on the progression of disease or the effectiveness of treatments. In Merline et al, ACM 2012, unsupervised learning techniques were used to find such patterns. The challenge in this application is that, like before, the longitudinal EHR data is highly sparse, incomplete, and high-dimensional. Their paper aimed to overcome such sparsity with probabilistic clustering models, and was shown to be able to find such patterns in the data.

Find a figure  
missing figure



Figure 1: Dummy

LOREM IPSUM DOLOR SIT AMET, CONSECTETUR ADIPISCING ELIT. UT PURUS ELIT, VESTIBULUM UT, PLACERAT AC, ADIPISCING VITAE, FELIS. CURABITUR DICTUM GRAVIDA MAURIS. NAM AREU LIBERO, NOMNMMY EGIT, CONSECTETUR ID, VULPITATE A, MAGNA. DONEC VEHICULA AUGUE EU NEQUE. PELLentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus

Figure 29.13: Finding patterns in longitudinal EHR data [? ]

## 29.7 Further Reading

## 29.8 Tools and Techniques

## 29.9 What Have We Learned?

In this section we have learned about the basics of epidemiology, both genetic and molecular. We have learned techniques of designing an epidemiological experiment and how and when to use genetic screens for identifying diseases. We also focused on resolving causality vs. correlation between epigenetic markers and diseases using genetics as an instrument variable. Finally, we learned about PheWAS and how it can improve

the effectiveness of GWAS in learning phenotypic relationships, and how it can harness the data from the EHR to find otherwise missing patterns and relationships.

## Bibliography

- [1] Alberto Gutierrez. Inspections, Compliance, Enforcement, and Criminal Investigations: 23andMe, Inc. 11/22/13, November 2013. FDA.
- [2] Mukesh Verma. Epigenome-Wide Association Studies (EWAS) in Cancer. *Curr Genomics*, 13(4):308–313, 2012.

---

CHAPTER  
**THIRTY**

---

## CANCER GENOMICS

Pasha Muravyev (2014), Casper Enghuus (2015), Owen Gray (2015), Nathan Hunt (2016), Maha Shady (2017)

### 30.1 Introduction

What is cancer? Cancer represents a group of diseases or tumors that trigger abnormal cell growth and have the potential to spread to many parts of the body. A cancer usually starts with mutations in one or more "driver genes" which are genes that can drive tumorigenesis. These mutations are called driver events, meaning that they provide a selective fitness advantage for the individual; other mutations that do not provide a fitness advantages are called passenger mutations.

In order for a cell to become cancerous, it must find ways to avoid the numerous checkpoints that restrains any single cell from excessive proliferation. Weinberg outlines six hallmarks of cancer that differentiates a tumor cell from a normal cell, and added four more to his list in 2011. Mutations leading to each of these characteristics occur in every cancer, though in different orders and through different mechanisms. These hallmarks are:

- **Weinberg's 6 Hallmarks**

1. Evading apoptosis
2. Growth signal self-sufficiency
3. Insensitivity to anti-growth signals
4. Sustained angiogenesis
5. Tissue invasion + metastasis
6. Limitless replicative potential

- **Emerging Hallmarks:**

1. Deregulate cellular energetics
2. Avoid immune destruction

- **Enabling Hallmarks:**

1. Genome instability and mutation
2. Tumour-promoting inflammation

Cancers rarely occur due to single mutation events, but are rather a product of a large number of accumulated changes. Certain risk factors may increase the rate of this process and increase the likelihood of developing cancer. These risk factors include:

- Viruses (i.e. HPV)
- Carcinogens
- Radiation (like UV light)
- Oxidative damage (aging)
- Genetic risk factors (i.e. mutations in BRCA1)

Detecting which mutations contribute to a cancer is a challenging task, since most (> 99.9%) of mutations are passenger mutations and do not directly contribute to tumorigenesis. There are generally 4 types of genetic aberrations that contribute to cancers: germline mutations, which may lead to a heritable predisposition towards developing cancer; somatic mutations; epigenetic changes, which may lead to repression or activation of unintended genes; and gene-regulatory changes (i.e. mutation of transcription factors or their promoters) that may give rise to complex and detrimental phenotypes.

The main objective of cancer genomics is thus to generate a comprehensive catalog of cancer genes, pathways, and their driver mutations. Many cancer genome projects have been started within the last ten years (mainly due to the drop in genome sequencing costs); for example, The Cancer Genome Atlas was started in 2006 with the aim of analyzing 20-25 tumor types with 500 tumor-normal pairs each via a large number of experiments (SNP arrays, whole-exome sequencing, RNA seq, and others). The ICGC (International Cancer Genome Consortium) is a larger, umbrella organization that organizes similar projects around the world with the end goal of studying 50 tumor types with 500 tumor-normal types each.

## 30.2 Characterization

There are two sides of cancer: recurrence and heterogeneity. Recurrence refers to the small set of pathway alterations necessary for cancer (hallmarks), whereas heterogeneity refers to the generally large number of mutations that are found in cancer. Cancer develops by an evolutionary process that positively selects for mutations that increase fitness of the cancer cells. As cancer cells proliferate, there is an increased chance that they will accumulate mutations. Moreover, many cancers have mutations in genes ensuring proper chromosome segregation, mismatch repair, etc. during DNA replication leading to decreased genomic stability. This may further increase the accumulation of mutations and cell heterogeneity. For each tumor, our aim is thus to obtain a complete, base-level characterization of that tumor, its evolutionary history and the mechanisms that shaped it. We can use *massively parallel sequencing* to get the base level genome characterization, but this approach brings with it some associated challenges.

1. **Massive amounts of data:** The main challenge with increased amounts of data is an increase in the computational power required to analyze this data, as well as storage costs associated with keeping track of all of the sequenced genomes. There also needs to be an analysis pipeline (automated, standardized, reproducible) to have consistent findings across the different characterization efforts. Finally, we need to come up with new ways of visualizing and reporting on large scale data.
2. **Sensitivity versus Specificity:** Cancer characterization starts with the proper identification of SNP mutations present in cancer cells, and maximal removal of false positive reads. When selecting tumor samples, the extracted DNA is a mix of normal genomes and complex tumor genomes. The mutational allelic fraction (the fraction of DNA molecules from a locus that carry a mutation), is used to study significance of a mutation and its prevalence in the cancer subtype. This fraction depends on the purity, local copy number, multiplicity of the tumor sample, and the cancer cell fraction (CCF, amount of cancer cells that carry the mutation). Clonal mutations are carried by all cancer cells, and sub-clonal mutations are carried by a subset of the tumor cells.

As well as detecting the presence of clonal and subclonal mutations, proper analysis requires removal of false positive mutagenic events. Two types of false positives include sequencing errors and germline

mutations. Sequencing errors can come from misread bases, sequencing artifacts, and misaligned reads, while germline mutations usually occur in predictable places in the genome (1000/MB known, 10-20/MB novel). By having multiple reads of the same sequence the likelihood of repeated errors in sequencing drops rapidly, and by knowing where in the genome a germline mutation is likely, a filter can correct for the additional false positive probability. The overall sensitivity of detecting single nucleotide variations depends on the frequency of background mutations and the number of alternative reads.

A third type of false positive can come from cross patient contamination if the “tumor” sample contains DNA from another person. ContEst is a method to accurately detect contamination by comparison to a SNP array.

A “mutation caller” is a classifier asking at every genomic locus, “Is there a mutation here?”. Their classifications depend on the allele fraction, coverage of tumor and normal sample, and sequencing and alignment noise. They are evaluated using Receiver Operators Characteristic (ROC) curves, which plot the true positive classifications versus the false positives. MuTect is a highly sensitive Somatic Mutation Caller. The MuTect pipeline is as follows: tumor and normal samples are passed into a variant detection statistic (which compares the variant model to the null hypothesis), which is passed through site-based filters (proximal gap, strand bias, poor mapping, triallelic site, clustered position, observed in control), then compared to a panel of normal samples, and finally classified as candidate variants. MuTect can detect low allele fraction mutations and is thus suited for studying impure and heterogenous tumors.

3. **Discovering mutational processes:** Instead of detecting the presence of mutations in cancer genes, a different approach could be to discover if there were specific patterns among mutations in the cancer samples. A ”Lego plot” is a way to visualize patterns of mutations, in which the heights of each of the colors represents frequencies of the 6 types of base pair substitutions, and the frequency of each is plotted relative to the 16 different contexts this mutation could occur in (neighboring nucleotides). The specific types of mutagenic events in each type of cancer can be plotted and analyzed. As an example, a novel mutation pattern (AA  $\rightarrow$  AC) is found in esophageal cancer. Cancers can be grouped by these specific mutational spectra. Dimensionality reductions using non-negative Matrix Factorization (NMF) of Lego plot data can be used to identify fundamental spectral signatures.
4. **Estimating purity, ploidy and cancer cell functions:** As well as detecting mutations in cancer cells, removing false positives, and detecting patterns of mutations, a proper characterization of each tumor sample is required. Because of heterogeneity and sample impurities, estimating the purity, absolute copy number and cancer cell fraction (CCF) of the “tumor” sample being sequenced is needed to get the correct total number and prevalence of the mutated alleles.
5. **Tumor heterogeneity and evolution:** Samples can have large distributions of point mutations and copy number alterations, but a Bayesian clustering algorithm can help identify the mutations and copy number alterations in distinct subpopulations.

### 30.3 Background Mutation Models

The fundamental challenge in interpreting the sequencing results lies in differentiating driver mutations from passenger mutations. In order to accomplish this, we need to model the background mutational processes of the analyzed sequences and identify pathways/regions with more mutations than would have been predicted solely by the background model. Those regions then become our candidate cancer genes.

However, we run into the potential issue of selecting an incorrect background model or we can encounter systematic artifacts in mutation calling. In this case, we have to go back to the drawing board and attempt to come up with a better background model before we can proceed with candidate gene identification.

Many tools have been developed in an effort to accurately detect candidate cancer genes and pathways (sub-networks) including NetSig, GISTIC, and MutSig. NetSig is used to identify clusters of mutated genes in protein-protein interaction networks. GISTIC can be used to score regions according to frequency and

amplitude of copy-number events. MutSig is used to score genes according to number and types of mutations. The main analysis steps in finding candidate cancer genes are 1) estimation of the background mutation rate (which varies across samples, 2) calculate p-values based on statistical models, and 3) correct for multiple testing hypothesis (N genes).

As sample size and or mutation rate increases, the significant gene list for cancer genes increases and contains many ‘fishy’ genes. One major breakthrough to reduce fishy genes has been the proper modeling of background mutations. Standard tools use consistent background rate (rates for CpG, C/G, A/T, indel) while ignoring heterogeneity across samples, additional sequence contexts, and the genome. However, better background models can be developed if considering the multiple factors that increase variability in mutation rates. Importantly, different cancer types have different mutation rates. For instance melanomas mutate at an  $\approx$ 100-fold higher rate than sarcomas. Moreover, different cancers also have distinct mutational profiles that may target specific genes unique to the cancer/tissue. Mutation rate also vary with gene expression level where highly expressed genes show less mutations, and the frequency of somatic mutations correlates with DNA replication time, i.e. regions that are replicated early show less mutations than regions replicated late. DNA accessibility and chromatin state also affects mutation rates with accessible regions being more prone to mutations. Finally, the mutation rate of a cancer also vary between each individual patient. All of these effects need to be considered when attempting to call driver mutations. MutSigCV is a tool which incorporates these factors into the background model and uses them to more accurately predict significant somatic mutations. It does so by calculating a score for each mutation, aggregating the scores, and then calling significance after controlling for FDR.

## 30.4 Immunotherapy

A really important class of cancer therapy is immunotherapy. The goal of immunotherapy is to train the body’s immune system to attack the tumor. To understand cancer immunotherapy, you should first get an overview of how the immune system attacks foreign cells.

Our cells have a set of cell surface proteins called the major histocompatibility complex or MHC (the human MHC is also called the human leukocyte antigen - HLA). These proteins bind to antigens derived from proteins inside the cell, and presents those antigens on the cell surface allowing the body’s immune cells to inspect them. The antigens may be self or non-self, if there is a foreign body or protein inside the cell. If a cytotoxic T cell binds to a non-self antigen, it will begin to replicate into memory T cells and effector T cells. The memory T cells are kept in the body, to be ready to act if the same foreign antigen is presented again in the future, and the effector T cell will start to act to kill the foreign cells. The effector T cells will bind to other cells that present the same antigen, it starts to release perforins, which will make holes in the foreign cells, then the T cell will also release enzymes into the foreign cell which will cause it to kill itself.

Now, what factors drive tumor immunity?

1. Tumor antigens
2. Tumor resistance (B2M/HLA, Ag loss, IFNGR pathways)
3. Immune cell states and pathways (T cells, fibroblasts, etc.)
4. Local tumor-related triggers (viruses, damaged DNA, tissue injury, hypoxia, metabolism)
5. Environmental factors (microbiome, smoking, pollutants, food)
6. Genetics (susceptibility to autoimmunity)

Somatic mutations in the tumor have the potential to generate neo-antigens (non-self antigens), making tumors with high mutational burden more susceptible to immunotherapy. Some tumors are also associated with viral infections, and thus, the infected cancer cells can also generate neoantigens. Presentation of

these neo-antigens on the cell surface would trigger an immune response. However, there can be other characteristics in the cancer cells that evade the immune response. For example, HLA mutations may prevent the presentation of the neo-antigens. Furthermore, even if the immune system attacks the cell, suppressive subclones may arise and continue to grow.

## 30.5 Current Research Directions

Much work has been done in characterizing coding driver mutations. While there certainly remain driver genes for various cancers to be discovered, there is also an increasing amount of research directed at identifying non-coding driver mutations (NDMs). The vast majority of mutations are non-coding, so there may be much more to be understood about cancer when we can understand the effects of these mutations, but because they aren't translated into amino acid differences, predicting the functional impact of such mutations is more difficult.

Two recent approaches to detecting NDMs were OncodriveFML, which tries to identify mutations undergoing selection [Mularoni *et al.*, *Genome Biology* 2016], and ParsSNP, a 'parsimony' guided approach that assumes drivers are more equitably distributed among samples than passengers [Kumar *et al.*, *Nature Genetics* 2016]. In addition, newly developed tools like DeepSEA, which can predict the functional impact of non-coding mutations (e.g. the effect on transcription factor binding or chromatin state) may be useful in mapping non-coding mutations to their putative functions [Zhou and Troyanskaya, *Nature Methods* 2015].

## 30.6 Further Reading

1. <http://www.broadinstitute.org/cancer/cga/mutect>
2. <http://www.broadinstitute.org/cancer/cga/ABSOLUTE>

## 30.7 Tools and Techniques

- **MutSigCV**, mentioned above, is a widely-used tool for detecting mutations based on their frequency which has a sophisticated background model built from multiple data sources.
- **OmicTools** has a list of several different tools for detecting driver mutations. These are generally less well-known and thus may not be as well tested, but they include many different approaches and the site appears to be updated at times with newly published tools.

## 30.8 What Have We Learned?

The drop in sequencing costs over the last ten years has led to a need for automated analysis pipelines and more computational / storage power to handle the vast flood of data being generated by a multitude of parallel sequencing efforts. Two major tasks of cancer genome projects going forward can be roughly grouped into two areas: characterization and interpretation.

For characterization, there seems to still a need for a systematic benchmark of analysis methods (one example is ROC curves - curves that illustrate the performance of a classifier with a varying discrimination threshold). We saw that cancer mutation rates tend to vary more than 1,000-fold across different tumor types. We also learned that clonal and subclonal mutations could be used for studying tumor evolution and heterogeneity.

Running a significance analysis on the sequencing results identified a long-tailed distribution of significantly mutated genes. Since we're dealing with a long tail distribution, we can increase the predictive power of our models and detect more cancer genes by integrating multiple sources of evidence. However we have to take into account that mutation rates differ according to the original sample, gene, and category from each study.

## Bibliography

- [1] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, 2015.
- [2] Runjun D Kumar, S Joshua Swamidass, and Ron Bose. Unsupervised detection of cancer driver mutations with parsimony-guided learning. *Nature Genetics*, 48(10), 2016.
- [3] Loris Mularoni, Radhakrishnan Sabarinathan, Jordi Deu-Pons, Abel Gonzalez-Perez, and Núria López-Bigas. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome biology*, 17(1):128, 2016.
- [4] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods*, 12(10):931–4, 2015.

---

CHAPTER  
**THIRTYONE**

---

## GENOME EDITING

Guest Lecture by Feng Zhang

Scribed by: Jonathan King Him Ching (Fall 2017)

Pranam Chatterjee (Fall 2017)

Roger Jin (Fall 2017)

Previously contributed by: Eunice Wu

Ye Tao

### 31.1 Introduction

#### 31.1.1 What is CRISPR/Cas?

The CRISPR/Cas system is the prokaryotic immune system. When a virus or other foreign attacker attempts to infect a prokaryotic cell and inject its own DNA into a prokaryote's genome, the prokaryote's CRISPR/Cas system is responsible for removing the foreign DNA. How does it do this? The CRISPR/Cas system has two parts, CRISPR and Cas. The CRISPR part (a CRISPR array), is responsible for "remembering" the foreign DNA, while the Cas part (Cas proteins), is responsible for cutting out the recognized foreign DNA. A CRISPR array is made up of segments of short spacer DNA, which are the results of previous exposure to foreign DNA. These spacer DNA are transcribed to RNAs, which can be used to match the foreign DNA that the spacer DNA was built from. These RNA are then picked up by Cas proteins. When a Cas protein picks up a particular RNA, it becomes sensitive to the matching DNA sequences. The next time the same foreign DNA is inserted into the prokaryote, the Cas proteins sensitive to it will match the foreign DNA and cut it out of the genome, causing it to become inactive.

#### 31.1.2 Why is CRISPR/Cas important to us?

Because nature is giving us an effective way of editing a genome! In order to accurately edit a genome, it is important to be able to cut a sequence at precisely the targeted location. Once a cut is made, repair mechanism can go in and make a modification at the target site. The CRISPR/Cas system is a naturally occurring time tested method of doing making alterations to DNA sequences.

Currently, the ability of researchers to perturb and interrogate the genome is lagging behind the current level of techniques for reading. CRISPR provides an effective way to write to the genome that we are capable of reading, allowing us to determine what variations in the genetic code give rise to diseases of interest.

#### 31.1.3 Cas9

The CRISPR/Cas9 system in particular has emerged as a promising new genetic perturbation technology capable of precisely recognizing and cleaving target DNA. Cas9 is a programmable endonuclease that can be guided to specific genomic targets by single guide RNAs (sgRNAs) that form Watson-Crick base pairs with

the DNA. Thus Cas9 is easily retargetable to 20 bp sites flanked by a protospacer adjacent motif (PAM), or NGG.

Cas9 generates precise double-strand breaks (DSBs) at target loci that are repaired through either homology-directed repair (HDR) or, more often, non-homologous end-joining (NHEJ). HDR precisely repairs the DSB using a homologous DNA template, whereas NHEJ is error-prone and introduces indels. When Cas9 is targeted to a coding region, loss-of-function (LOF) mutations can occur as a result of frameshifting indels that produce a premature stop codon and subsequent nonsense-mediated decay of the transcript or generate a non-functional protein. These features make Cas9 ideal for genome editing applications.

## 31.2 Engineering Cas9 for Mammalian Genome Editing

### 31.2.1 Cas9 Specificity

Although Cas9 can precisely edit genomic targets that are complementary to its sgRNA, it can also sometimes edit genomic sites with 1-4 nucleotide mismatches to its sgRNA target sequence. This off-target activity of Cas9 is undesirable, and thus several methods have been developed to improve the specificity of Cas9. Truncating the sgRNA target sequence can reduce off-target activity without affecting on-target activity. A different approach focused on mutating the positive DNA-interacting residues of Cas9 to alanine to improve specificity. Additionally, mutating the residues that form the hydrogen bonds between the target DNA and Cas9 also improves specificity.

### 31.2.2 Cas9 for transcriptional activation and repression

In addition to generating LOF mutations, Cas9 can modulate transcription without modifying the genomic sequence through fusing catalytically inactive Cas9 (dCas9) to transcriptional activation and repression domains. CRISPR activation (CRISPRa) and CRISPR inhibition (CRISPRi) can be achieved by direct fusion or recruitment of activation and repression domains, such as VP64 and KRAB, respectively. CRISPRa in particular offers a significant improvement over other activation approaches. Previously gain-of-function tools were primarily limited to cDNA overexpression, which suffered from overexpression beyond physiological levels and endogenous regulation, lack of isoform diversity, and high cost of construction. CRISPRa overcomes these limitations because it activates gene transcription at the endogenous locus and simply requires the synthesis and cloning of RNA guides, making it much more affordable.

The first generation of CRISPRa fused dCas9 to a VP64 or p65 activation domain to produce modest transcriptional upregulation, the range of which was not suitable for genome-scale screening. Second generation CRISPRa designs produced more robust upregulation by recruiting multiple activation domains to the dCas9 complex. For instance, SunTag recruits multiple VP64 activation domains via a repeating peptide array of epitopes paired with single-chain variable fragment antibodies. Another activation method, VPR, uses a tandem fusion of three activation domains, VP64, p65, and Rta to dCas9 to enhance transcriptional activation. We devised an alternative approach to CRISPRa that involved incorporating MS2 binding loops in the sgRNA backbone to recruit two different activation domains, p65 and HSF1, to a dCas9-VP64 fusion. By recruiting three distinct transcriptional effectors, this synergistic activation mediator (SAM) complex could robustly and reliably drive transcriptional upregulation. A comparison of SunTag, VPR, and SAM across various cell types and species suggested that SAM induced more potent activation in some contexts, but further analysis is needed to determine which approach is most effective for GOF screening<sup>36</sup>.

### 31.2.3 Cas9 as a tool for screening

Together with large pooled single guide RNA (sgRNA) libraries, Cas9 can mediate high-throughput LOF and gain-of-function (GOF) dissection of many selectable phenotypes and elucidate complex biological questions. As a proof of principle to demonstrate the CRISPR-Cas9 system's utility for screening, genome-scale CRISPR-Cas9 knockout (GeCKO) and SAM libraries have been constructed to identify genes that, upon knockout or activation, confer resistance to the BRAF-inhibitor vemurafenib in a melanoma cell line.

In addition to vemurafenib resistance, CRISPR-Cas9 screens have provided insight into the molecular basis of gene essentiality, drug and toxin resistance, the hypoxia response, and the role of flavivirus host factors in infection. Although most screens have been performed in *in vitro* systems, the Cas9 system has also been applied *ex vivo* in dendritic cells to study the bacterial lipopolysaccharides response and *in vivo* to identify key factors that allow a non-metastatic lung cancer cell line to metastasize. CRISPR-Cas9 screens have also been expanded to the noncoding genome through saturated mutagenesis by tiling sgRNAs across a noncoding locus to uncover functional elements in the BCL11A enhancer, POU5F1 locus, and CUL3 locus, as well as p53 and ESR1 transcription factor binding sites.

### 31.3 Novel CRISPR enzymes

While CRISPR-Cas9 has been a major focus of genome editing research, CRISPR systems are incredibly diverse and are found in most bacterial and archaeal species and include dozens of families of proteins. While CRISPR systems were first uncovered in 1989, up until only a few years ago in 2014, only three types of CRISPR systems were known of, with Cas9 being among one of these. Recent advances in computational data mining amongst the thousands of bacterial genomes that have been sequenced, have allowed our understanding of CRISPR systems to deepen in the past few years. Certain features of CRISPR systems, such as the highly conserved adaptation protein Cas1, or the distinct signature of CRISPR repeats, allows algorithms to quickly identify probably CRISPR sites in genomes. Clustering of proteins and open reading frames around these sites, allows new families of CRISPR enzymes to be identified. Two of these new CRISPR effector proteins, Cpf1 and C2c2, have unique features from Cas9 and have been useful in genome editing applications.

#### 31.3.1 Cpf1

Cpf1 was the first novel type II CRISPR system to be discovered after Cas9. Like Cas9, it is an RNA guided DNA nuclease, capable of robust interference against DNA phage infections and displaying adaptation in response to previous infections. It is considered to be a unique family of CRISPR enzymes within the type II subclass. It shares a RuvC endonuclease domain with Cas9, but initially showed no identifiable additional nuclease domain. Initial experiments showed that it was capable of dsDNA cleavage and later structural information from crystallography experiments revealed the second novel DNase domain. Cpf1 has different sequence requirements with a PAM that is 5' to the target site and AT rich, rather than the GC rich NGG PAM of Cas9. Additionally, Cpf1 generates a dsDNA cut with 5' staggered ends as opposed to the blunt ends of a Cas9 cut. These findings were exciting for the genome editing field because a different PAM expands the number of possible target sites for making edits and staggered cuts would theoretically make insertion of DNA fragments more efficient.

It was also identified that Cpf1 is capable of processing its own CRISPR array. Typically, type II CRISPR systems require an additional RNase present in the cell to process the CRISPR array into mature crRNAs. Cpf1 however has an additional RNase domain that is capable of performing this function. This unique functionality has interesting biotechnological applications because it allows for many guides to be delivered under a single CRISPR array. This enables multiplexed genome editing for changing many genes at once *in vivo* or for screens involving multiple edits.

Recently, Cpf1 was shown to possess unrestricted ssDNA cleavage activity after target recognition and activation. This activity, catalyzed by the same active site responsible for site-specific dsDNA cutting, indiscriminately shreds ssDNA with rapid multiple-turnover cleavage kinetics. Activation of ssDNA cutting requires on-target recognition of a DNA target sequence matching the guide RNA sequence with enough specificity to distinguish between closely related viral serotypes. Such activation most likely provides simultaneous protection from both dsDNA and ssDNA phages, and may also target ssDNA sequences that arise temporarily during phage replication or transcription.

#### 31.3.2 C2c2

C2c2 was initially discovered as part of a computational search to uncover novel single effector CRISPR proteins for expanding the genome editing toolbox. While many novel RNA guided DNases were discovered,

C2c2 was different because it showed no identifiable DNase domains. Rather, multiple alignment of the family members revealed two higher eukaryotic and prokaryotic nuclease (HEPN) domains, which are canonically RNase domains. Because of the highly conserved nature of these domains and the localization of C2c2 by CRISPR arrays, it was hypothesized that C2c2 might be a novel RNA guided RNA targeting CRISPR effector.

Initial experiments showed that C2c2 from *Leptotrichia shahii* was capable of robust immunity against the ssRNA phage MS2. Further biochemical characterization showed that the enzyme was generating many cuts in surrounding ssRNA regions within the secondary structure and had a very different mechanism of cleavage than CRISPR DNA targeting enzymes. Additionally, in biochemical experiments, once C2c2 was activated, it was capable of cleaving any RNAs present in the reaction. This promiscuous collateral activity is hypothesized to be part of a programmed cell death pathway that instructs the cell to commit suicide by shredding host RNAs if the cell is not capable of surviving an infection event.

C2c2 was also capable of programmed mRNA knockdown, showing some utility as a tool. Because of its high specificity to even single mismatches, it might be useful for RNA targeting applications, such as RNA knockdown or live imaging of transcripts using a catalytically dead version.

## Cas13

Further orthologs of C2c2 were tested to eliminate the collateral damage during Cas13-mediated RNA editing. After initial screening of 15 orthologues, Cas13a from *Leptotrichia wadei* (LwaCas13a) was identified as the most effective in an interference assay in *Escherichia coli*. LwaCas13a can be expressed in cells of various genomes to knockdown reporter or endogenous transcripts with comparable levels of knockdown as RNA interference and improved specificity. Furthermore, a fusion protein was generated between the catalytic domain of ADAR and a dCas13b, with the ability to conduct A-to-I editing at any site.

## Cas9 Orthologs

To expand the space of Cas9 enzymes that are capable of cleaving DNA and inducing gene editing, much work has been put into discovering novel Cas9 orthologs from a variety of bacterial species. For example, to package Cas9 and its guide RNA into a single adeno-associated virus (AAV) vector, so that delivery of the system has low immunogenic potential, reduced oncogenic risk from host-genome integration, and broad-range of serotype specificity, orthologs from organisms such as *Staphylococcus aureus* and *Campylobacter jejuni* have been characterized. Furthermore, Cas9 proteins with other useful properties, such as thermostability and divergent PAM sequences for increased targeting range, have been recently described.

## 31.4 Further Reading

## 31.5 Tools and Techniques

## 31.6 What Have We Learned?

CRISPR/Cas9 produces double stranded breaks in DNA, and has two main components:

20bp DNA

PAM

### 31.6.1 Why is CRISPR/Cas important to us?

Because nature is giving us an effective way of editing a genome! In order to accurately edit a genome, it is important to be able to cut a sequence at precisely the targeted location. Once a cut is made, repair mechanism can go in and make a modification at the target site. The CRISPR/Cas system is a naturally occurring time tested method of doing making alterations to DNA sequences.

Currently, the ability of researchers to perturb and interrogate the genome is lagging behind the current level

of techniques for reading. CRISPR provides an effective way to write to the genome that we are capable of reading, allowing us to determine what variations in the genetic code give rise to diseases of interest.

Furthermore, older methods such as fusing nuclease domains to Zinc Fingers and Transcription-Activator Like Effectors, DNA binding proteins, required tedious construction of individual proteins for every new intended DNA target. CRISPR improves on these techniques by providing a programmable strategy of utilizing a single effector protein that can be guided to sites simply by altering the (much easier to manipulate) guide RNA molecule to hybridize to different target sites.

## Bibliography

