

Bioinformatics III

Structural Bioinformatics and Genome Analysis

Summer Semester 2007

**by Sepp Hochreiter
(Chapters 2 and 3 by Noura Chelbat)**

© 2007 Sepp Hochreiter & Noura Chelbat

This material, no matter whether in printed or electronic form, may be used for personal and educational use only. Any reproduction of this manuscript, no matter whether as a whole or in parts, no matter whether in printed or in electronic form, requires explicit prior acceptance of the author.

Preface

This course is part of the curriculum of the master of science in bioinformatics at the Johannes Kepler University Linz. Focus of the course is structural bioinformatics (part 1) and genome analysis (part 2). These two topics are merged in this course because of the master's program schedule.

The spacial restriction did neither allow to introduce all methods nor allow to explain the introduced ones in more detail.

The students should gain insights into the topics and methods of structural bioinformatics and genome analysis. The students should learn how to choose appropriate methods from a given pool of approaches to structural bioinformatics (e.g. structural alignment or 3D prediction) and to genome analysis (e.g. microarray technique). The students should learn to understand and to evaluate the different approaches, know their advantages and disadvantages as well as where to obtain and how to use them. In a step further, the students should be able to adapt standard algorithms for their own purposes or to modify those algorithms for specific applications with certain prior knowledge or special constraints.

Structural Bioinformatics

A main topic in structural bioinformatics is to give computational approaches to predict and analyze the spatial structure of macromolecules like proteins, DNA, and RNA. Their 3D structure is predicted based on the 1D structure, the nucleotide or amino acid sequence, which is obtained from genome sequencing. Knowing and understanding their 3D structure is crucial for inferring and modifying their function. Direct applications could be in medical and pharmacological fields – especially for drug design, where it is important to determine which groups of ligands bind and regulate a protein, which proteins are potential targets for drugs, etc.

For detecting the 3D structure the methods from Bioinformatics I allow for homology and comparative modeling by sequence-sequence comparisons, where it is assumed that similar sequences have the same 3D structure. Another approach which includes structural information is sequence-structure comparison by computing the sequences-to-structure-fitness through “threading”, which determines how well a sequence fits to a given 3D structure.

By modeling the physical laws details about the protein function and ligand docking behavior is obtained. Modeling is often based on molecular dynamics using force fields which approximate the physical laws.

Genome Analysis

Main focus of the genome analysis will be the microarray technique and the preprocessing and analysis methods associated with it.

The microarray technique generates a gene expression profile which gives the expression states of genes in a cell by reporting the mRNA concentration. The mRNA concentration in turn reports the cell status determined by what and how many proteins are currently produced. The DNA microarray technologies such as cDNA and oligonucleotide arrays provide means of measuring tens of thousands of genes simultaneously (a snapshot of the cell). The microarrays are a large scale high-throughput method for molecular biological experimentation.

The information obtained by recognizing genes that share expression patterns and hence might be regulated together are assumed to be in the same genetic pathway. Therefore the microarray technique helps to understand the dynamics and regulation behavior in a cell.

One of the goals of microarray technology is the detection of genes that are differentially expressed in tissue samples like healthy and cancerous tissues to see which genes are relevant for cancer. It has important applications in pharmaceutical and clinical research and helps in understanding gene regulation and interactions.

Genome analysis includes also genome anatomy and genome individuality (e.g. repetitions or single nucleotide polymorphism).

We will address also actual genomic research questions about alternative splicing and nucleosome position.

Literature

- David W. Mount. Bioinformatics – Sequence and Genome Analysis. Cold Spring Harbor Labaratory Press, Cold Spring Harbor, New York, USA, 2004.
- Philip E. Bourne and Helge Weissig. Structural Bioinformatics. Wiley-Liss, Hoboken, New Jersey, USA, 2003.
- Michael J. E. Sternberg. Protein Structure Prediction. Oxford University Press, 1996.
- Steen Knudsen. Guide to Analysis of DNA Microarray Data. John Wiley & Sohns, Hoboken, New Jersey, USA, 2004.
- Ernst Wit and John McClure. Statistics for Microarrays. John Wiley & Sohns Ltd., England, 2004.
- Pierre Baldi and G. Wesley Hatfield. DNA Microarrays and Gene Expression – From Experiments to Data Analysis and Modeling. Cambridge University Press, United Kingdom, 2002.
- Geoffry J. McLachlan, Kim-Anh Do, and Christophe Ambroise. Analyzing Microarray Gene Expression Data. John Wiley & Sohns Inc., Hoboken, New Jersey, USA, 2004.
- Jerome K. Percus. Mathematics of Genome Analysis. Cambridge University Press, United Kingdom, 2002.

Contents

I Structural Bioinformatics	1
1 Introduction	3
2 Chemical and Physical Background	13
2.1 Atomic Bounds: A Basic Introduction	13
2.1.1 Non-Covalent Interactions	15
2.1.1.1 Charge-Charge Interactions or Ionic Bounds	15
2.1.1.2 Dipole Interactions	17
2.1.1.3 Van der Waals Forces or Dispersion	20
2.1.1.4 Hydrogen Bond	20
2.1.1.5 Hydrophobic-Hydrophilic Interactions	25
2.1.2 Conclusion	28
2.1.3 Glossary	29
2.2 From chain polypeptide 1D configuration to folded 2D	30
2.2.1 Amino acids: classification and chemical-physical properties	30
2.2.1.1 Peptide bond	30
2.2.1.2 Torsion angles Phi (Φ) and Psi (Ψ)	39
2.2.1.3 Ramachandran plot	41
2.2.2 Interactions and folding	45
2.2.2.1 Bonds	46
2.2.2.2 Thermodynamics	48
2.2.3 Secondary Structure Elements	50
2.2.3.1 Types	52
2.2.3.1.1 α -helix	52
2.2.3.1.2 β -sheet	52
2.2.3.1.3 Turn and Loops	56
2.2.3.1.4 Coiled coil	56
2.2.3.1.5 TIM barrels	58
2.2.3.2 Motifs and Domains	58
2.2.3.2.1 Homeodomains	60
2.2.3.2.2 Leucine zipper	60
2.2.3.2.3 Zinc finger	61
2.2.3.2.4 Transmembrane elements	61
2.2.4 3D Structure	61
2.2.5 Major methods of structure determination	64
2.2.5.1 X-ray Crystallography	64

2.2.5.2	NMR Spectroscopy	66
2.2.6	Viewers	67
2.2.6.1	Rasmol	67
2.2.6.2	Chime	69
2.2.6.3	Pymol	69
2.2.7	First approximation	69
2.2.7.1	PDB- function	69
2.2.7.2	SCOP-Classes	70
2.2.8	Concepts	71
2.2.9	Annexes	72
3	Structural Comparison and Alignment	75
3.1	Introduction	75
3.2	Methods for Structure Comparison and Alignment	77
3.2.1	Basic remind	78
3.2.1.1	Dynamic programming	78
3.2.1.2	Distance Matrix	79
3.2.2	SARF2, VAST, COMPARER	82
3.2.3	SARF2: Spatial Arrangement of Backbone Fragments	82
3.2.3.1	VAST: Vector Alignment Search Tool	83
3.2.3.2	COMPARER	84
3.2.4	CE, DALI, SSAP	85
3.2.4.1	CE: Combinatorial Extension of the Optimum Path	85
3.2.4.2	DALI: Distance Matrix Alignment	91
3.2.5	SSAP: Secondary Structure Alignment Program	95
3.3	Conclusion	101
3.4	Exercises	101
3.5	Concepts	102
4	Protein Secondary Structure Prediction	105
4.1	Introduction	105
4.2	Assigning Secondary Structure to Measured Structures	105
4.2.1	DSSP	105
4.2.2	STRIDE	110
4.2.3	DEFINE and P-Curve	111
4.3	Prediction of Secondary Structure	113
4.3.1	Chou-Fasman Method	114
4.3.2	GOR Methods	114
4.3.3	Lim's Method	116
4.3.4	Neural Networks	116
4.3.5	PHD, PSIPRED, PREDATOR, JNet, JPred2, NSSP, SSPro	118
4.4	Evaluating Secondary Structure Prediction	122
4.4.1	Non-Homologous Test Sequences	122
4.4.2	Secondary Structure Classes	123
4.4.3	Quality Measures	124
4.4.4	Problems in Quality Comparisons	126

5 Homology 3D Structure Prediction	127
5.1 Introduction	127
5.2 Comparative Modeling: Sequence-Sequence Comparison	128
5.3 Threading: Sequence-Structure Alignment	132
6 Ab Initio Prediction and Molecular Dynamics	139
6.1 Introduction	139
6.2 Ab Initio Methods	139
6.3 Molecular Dynamics	140
II Genome Analysis	143
7 Introduction	145
8 DNA Microarrays	147
8.1 Motivation	147
8.2 DNA Microarray History and Current Status	148
8.3 DNA Microarray Techniques	149
8.3.1 Oligonucleotide Arrays	151
8.3.2 cDNA / Spotted Arrays	157
8.3.3 Other Techniques	163
8.3.3.1 SAGE	163
8.3.3.2 Digital Micromirror Arrays	163
8.3.3.3 Inkjet Arrays	164
8.3.3.4 Bead Arrays	164
8.3.3.5 Nanomechanical Cantilevers	164
8.4 Microarray Noise	164
8.5 Image Analysis	164
8.6 Background Correction	165
8.7 Normalization	171
8.8 PM Correction	174
8.9 Summarization	174
8.10 Different Combinations of the Processing Steps	176
8.11 Microarray Gene Selection Protocol	178
8.11.1 Description of the Protocol	178
8.11.2 Comments on the Protocol and on Gene Selection	180
8.11.3 Classification of Samples	181
9 DNA Analysis	183
9.1 Genome Anatomy	184
9.2 Gene Finding	186
9.2.1 Hidden Markov models	186
9.2.2 Neural networks	189
9.2.3 Homology Search	190
9.2.4 Promoter Prediction	190
9.2.4.1 Prokaryotes: <i>E. coli</i>	190

9.2.4.2	Eukaryotes	191
9.2.5	EST Clusters	194
9.2.6	Performance of Gene Prediction Methods	194
9.3	Alternative Splicing and Nucleosomes	195
9.3.1	Nucleosomes	195
9.4	Comparative Genomics	196
9.5	Genomic Individuality	201
9.5.1	Sequence Repeats	201
9.5.2	SNPs	205
10	DNA Sequence Statistics	209
10.1	Local Characteristics	209
10.2	Long-Range Characteristics	209
10.2.1	Matching Probability of Subsequences	209
10.2.2	Spectral Analysis	217
10.2.3	Entropy Analysis	219
A	Probability Generating Function	221
A.1	Definition	221
A.2	Properties	221
A.2.1	Power series	221
A.2.2	Probabilities and expectations	221
A.2.3	Functions of independent random variables	222
A.3	Examples	223
A.4	Example calculation: use of bivariate generating functions	224
B	Contact Potential for Threading	227
C	3D Prediction Challenge Results (CASP7)	233

List of Figures

1.1	Protein lysozyme in cartoon representation, where the secondary structure elements are shown.	4
1.2	Balls representation of a protein consisting of an α -helix.	5
1.3	Cartoon representation of a protein consisting of an α -helix.	5
1.4	Balls and stick representation of the α -helix of the protein lysozyme.	6
1.5	Balls and stick representation with hydrogen non-covalent bonds of the α -helix of the protein lysozyme.	6
1.6	Ribbon representation of a protein.	7
1.7	The protein “mannose” represented by the SARF2 software.	7
1.8	Hydrophobicity plot for subunit M of the photosynthetic center of Rhodopseudomonas Viridis.	9
1.9	Hydrophobicity plot for the human actin.	9
2.1	Covalent and non-covalent bond energies.	14
2.2	Coulomb’s Law.	16
2.3	Non-covalent interaction energy of two close particles.	17
2.4	Dipole moments.	18
2.5	Types of non-covalent interactions.	22
2.6	Hydrogen bonds in water.	24
2.7	Triple hydrogen bond in a DNA base pair.	25
2.8	Clathrate structure.	26
2.9	Amphiphatic molecules.	27
2.10	Immersion of amphiphatic molecules in water.	28
2.11	Aliphatic/hydrophobic R.	32
2.12	Proline residue.	32
2.13	Aromatic R.	33
2.14	Sulfur-containing amino acids.	33
2.15	Hydroxyl R.	34
2.16	Basic amino acids.	36
2.17	Acids R.	36
2.18	The four constituents around the C_α are shown.	37
2.19	Formation of a dipeptide linking alanine and glycine.	38
2.20	Zwitterion form.	38
2.21	Planar nature of the double-bond character.	39
2.22	Cis- and trans-conformation of the amino acid proline.	40
2.23	Polypeptide chain.	40

2.24	Torsion angles.	41
2.25	A non-allowed conformation is shown.	42
2.26	Ramachandran plot.	43
2.27	Ramachandran representation for Alanine and Glycine.	44
2.28	Central dogma represented.	47
2.29	Folding pathway.	47
2.30	Energy folding profile.	50
2.31	Folding factors.	51
2.32	Structure hierarchy.	51
2.33	The alpha helix.	53
2.34	3_{10} helix and α -helix.	54
2.35	β -sheet bonds.	55
2.36	Parallel and antiparallel b sheet.	56
2.37	B turn.	57
2.38	Loops	57
2.39	Coiled coil SSEs	58
2.40	Peptide velcro hypothesis.	59
2.41	TIM barrel SSEs.	59
2.42	Leucine zipper.	60
2.43	Zinc finger.	61
2.44	Glycophorin C protein.	62
2.45	Lysozyme.	63
2.46	Hemoglobin.	64
2.47	P53.	65
3.1	Structure superimposition.	76
3.2	Double Dynamic Programming (DDP).	80
3.3	Distance Matrix.	81
3.4	SARF superimposed result.	83
3.5	Structural vector alignment software SARF2 and VAST.	84
3.6	Calculation of distance D_{ij} for two AFPs.	88
3.7	Calculation of distance D_{ij} for a single AFP.	88
3.8	Hypothetical structural alignment by DALI.	93
3.9	SSAP dendrogram.	97
3.10	Multiple sequence alignment derived from pairwise simple alignment concatenation generated by SSAP.	98
4.1	Distances used to compute the DSSP Coulomb hydrogen bond according to eq. (4.1).	106
4.2	Output example for DSSP.	108
4.3	Distances used to compute the STRIDE hydrogen bond according to eq. (4.2).	111
4.4	Output example for STRIDE.	112
4.5	Comparison of different methods from [Rost, 2003b].	113
4.6	Helical wheel depicting the positions of amino acids in an α -helix.	116
4.7	Helical wheel depicting for leucine zipper, a coiled-coil structure.	117
4.8	The NETTalk neural network architecture is depicted.	117
4.9	Neural network approach to secondary structure prediction.	118

4.10	Neural network approach in the second level to secondary structure prediction.	119
4.11	The PHD levels of prediction.	120
4.12	Comparison of different methods from [Rost, 2003b].	121
4.13	Comparison of different methods from [Rost, 2003b] (2).	121
4.14	Comparison of different methods from [Rost, 2003b] (3).	122
5.1	Threading alignment.	134
6.1	Schematic view of the angles and bond length which are used to compute the force field.	141
8.1	The microarray technique.	150
8.2	The Affymetrix microarray technology for oligonucleotide arrays.	152
8.3	The Affymetrix technology in a simplified version.	153
8.4	The Affymetrix GeneChip® and the scanner images.	154
8.5	The Affymetrix GeneChip® devices.	155
8.6	The Affymetrix technique to detect the expression of a gene.	156
8.7	A probe set of an Affymetrix chip.	156
8.8	The different Affymetrix array techniques.	157
8.9	The steps of spotted arrays (cDNA arrays).	158
8.10	Spotted arrays (cDNA arrays).	159
8.11	Spotted arrays (cDNA arrays).	160
8.12	Spotted arrays (cDNA arrays).	160
8.13	Examples of scanner images of red-green spotted arrays.	161
8.14	More examples of scanner images of red-green spotted arrays.	162
8.15	Different view on examples of scanner images of red-green spotted arrays.	162
8.16	The steps of image analysis.	165
8.17	An image of a microarray from Terry Speed et al.	166
8.18	Examples of segmentation at the microarray image analysis.	166
8.19	Different methods to access the shape of the spot.	167
8.20	The background correction is shown.	168
8.21	MvA plots.	171
8.22	Per line MvA plots for original data and data mapped to a linear scale are shown.	173
8.23	Different methods at different processing levels. RMA and MBEI are shown.	177
8.24	Different methods at different processing levels. RMA and FARDS are shown.	177
9.1	The GLIMMER hidden Markov model for gene finding.	187
9.2	The GENEZILLA hidden Markov model for gene finding.	188
9.3	A pattern for the exon-intron boundary.	189
9.4	A pattern for the intron-exon boundary.	190
9.5	The pattern for <i>E. coli</i> RNA polymerase promoter at position -10.	191
9.6	The pattern for <i>E. coli</i> RNA polymerase promoter at position -35.	191
9.7	The input weights to a neural network which codes the nucleotides locally can be viewed as scoring matrix.	192
9.8	The nucleosome model found by Segal et al.	196
9.9	Genome comparison between species.	197
9.10	Genome comparison of campylobacteriales.	198

9.11 Segmental duplication between the chromosomes of <i>Arabidopsis</i>	199
9.12 The mouse chromosomes mapped to the human chromosomes.	200
9.13 The human chromosomes mapped to the mouse chromosomes.	201
9.14 The human chromosomes mapped to the mouse chromosomes.	202
9.15 The chimpanzee chromosomes mapped to the human chromosomes.	202
9.16 Worm chromosomes mapped to the fruit fly (<i>drosophila</i>) chromosomes.	203
9.17 Mitochondrial genomes of the <i>Zea</i> family.	204
9.18 The COMT SNPs associated with schizophrenia.	207

List of Tables

1.1	Two approaches of comparisons of proteins with known sequence or additional known structures.	8
2.1	Some Dipole Moments.	19
2.2	Some values of Van der Waals radii. These values represent the distances of closest approach for another atom or group.	21
2.3	Types of hydrogen bonds in biological compounds.	23
2.4	Properties of water.	25
2.5	The relative strength of different bonds.	28
2.6	Genetic code.	31
2.7	Hydrophobicity indices.	34
2.8	Amino acids can be translated for more than one triplet codon.	35
2.9	Amino acids substitutions.	37
2.10	Phi and Psi ideal angles values for some secondary structures.	44
2.11	Some of the chemical interactions that stabilize polypeptide chains.	48
2.12	Parameters of the most commonly found helical SSEs.	52
2.13	Principal locations of the principal Secondary Structure Elements.	56
2.14	Preferences normalized values of individual amino acid to be found within specific SSEs.	68
3.1	Pairwise SSAP scores matrix for immunoglobulins fold.	96
3.2	Pairwise SSAP scores obtained using a representative TIM structure consensus as template.	99
3.3	Methods for structural comparison and alignment.	100
4.1	The secondary structure symbols assigned by DSSP.	107
4.2	The 8 secondary classes mapped to 3 classes.	107
4.3	The secondary structure symbols assigned by STRIDE.	111
4.4	The 8 secondary classes of DSSP mapped to 3 or 4 classes.	123
4.5	Confusion matrix.	124
5.1	Results on the SCOP benchmark data set.	131
9.1	Genomes of different species.	185
9.2	Confusion matrix.	194
9.3	Test of finding the nucleotide ends of exons.	195
9.4	Test of Tab. 9.3 with other methods according to Zhang 1997.	195
9.5	Percentage of the transposable DNA in the genome for different species.	205

9.6 SNP associated with schizophrenia.	207
10.1 Nucleotide frequencies.	209
10.2 Nucleotide frequencies for human fetal globin gene.	210
10.3 Nucleotide ratio $\frac{p_{ij}}{p_i p_j}$ of observed pairs p_{ij} and random pairs $p_i p_j$	210
B.1 The contact potential from [Williams and Doherty, 2004].	227
B.2 The contact potential from [Dombkowski and Crippen, 2000], where the potential value is computed by eq. (B.1).	228
B.3 Overall threading potential.	229
B.4 Potential for 5 Å.	229
B.5 Potential for 7 Å.	230
B.6 Potential for 9 Å.	230
B.7 Potential for 11 Å.	231
B.8 Potential for 13 Å.	231

Part I

Structural Bioinformatics

Chapter 1

Introduction

The holy grail of bioinformatics is to predict 3D protein structures from 1D amino acid sequences in order to understand the function and the folding process of the proteins.

In the Brookhaven database PDB, the main protein 3D structure database, there are 22,000 protein structures solved. The NR database, the database of all non-redundant sequences (most of them obtained by genome sequencing), contains currently over 3 million sequences giving a ratio of structures to sequences of 1:136.

The structural levels of proteins are:

- **1D:** *primary structure*, the amino acid sequence as assembled on the ribosome using the genetic code to translate mRNA (three mRNA nucleotides = one amino acid).
- **2D:** *secondary structure*, elements like loops, α -helices, and β -sheets which arise through local hydrogen bonds between amino acids and form a local minimal energy state.
- **3D:** *tertiary structure*, a global minimal energy state of the amino acid sequence through global interactions among amino acids.

Chemical-physical Properties

In order to understand how an amino acid sequence folds, to understand how these amino acid sequences choose one out of thousands folding possibilities, to understand which relationships between amino acids arise depending on fold, an overview of chemical-physical properties of atomic bonds, playing special attention to the non-covalent bonds will be given at the beginning.

Molecular Viewers

Molecular viewers are important tools in structural bioinformatics or biochemistry to visualize the 3D structure of proteins. Some of these viewers will be discussed in this course.

Molecular viewer overcome difficulties

- in converting all of the important 3D structural information about a molecule into an understandable two-dimensional representation,

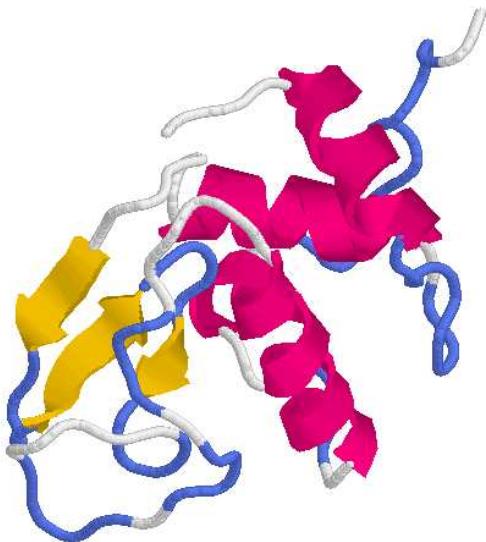


Figure 1.1: Protein lysozyme in cartoon representation, where the secondary structure elements are shown: α -helix (red) and β -sheet (yellow).

- in combining the variety of molecular representation formats which have been developed – each one designed to show a particular aspect of a molecule’s structure,
- in understanding the relationship between the structural features and its function.

Macromolecules (even when we focus on proteins) could be represented in many ways – see for examples Fig. 1.1 to Fig. 1.7.

Structural Alignment and Comparison

The most common method to compare two proteins is by alignment through Needleman-Wunsch or Smith-Waterman algorithms based on dynamic programming or through fast heuristics like BLAST. However alignments do not take the structure into account if the structure is known. The structure is an important source of information because the evolutionary relation by structure is stronger than by sequences. That means even if the evolutionary relations between proteins is no longer recognizable by sequence comparison it is still recognizable by structural similarities.

If using structure information two basic approaches are possible: (1) aligning the sequences by superimposing the structures, and (2) “structural alignment” which compares the structures to one another and ignores the type of amino acid at a certain position.

With (1) “structure comparison” we mean

similarities between two or more proteins based on their atomic 3D coordinates. For example, similarity can be measured by the distance of the backbone atoms (C_α -atoms).

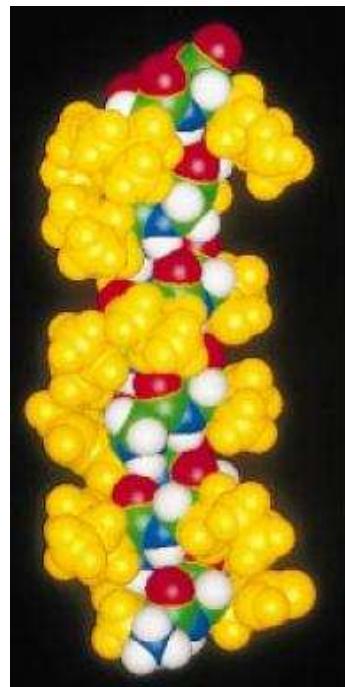


Figure 1.2: Balls representation of a protein consisting of an α -helix. Each ball represents an atom where the size of the ball shows the van der Waals radius and the color the type of the atom.

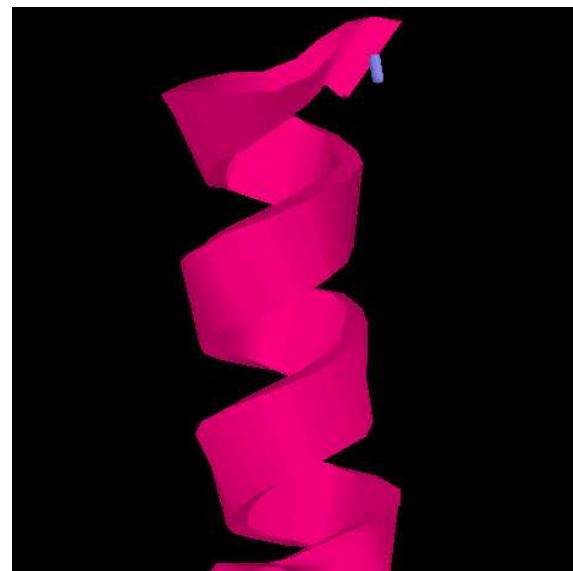


Figure 1.3: Cartoon representation of a protein consisting of an α -helix.

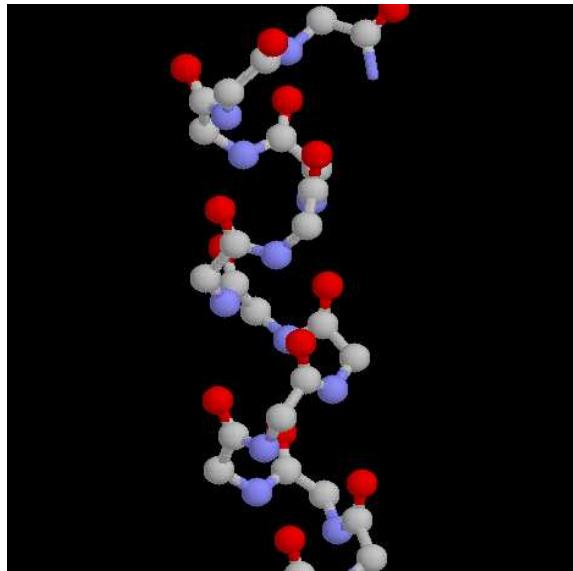


Figure 1.4: Balls and stick representation of the α -helix of the protein lysozyme. Atoms are represented by balls and covalent bonds by sticks between the balls.

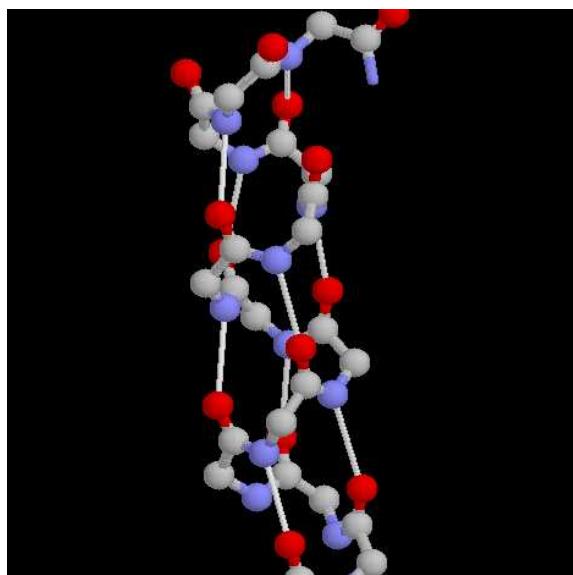


Figure 1.5: Balls and stick representation with hydrogen non-covalent bonds of the α -helix of the protein lysozyme.

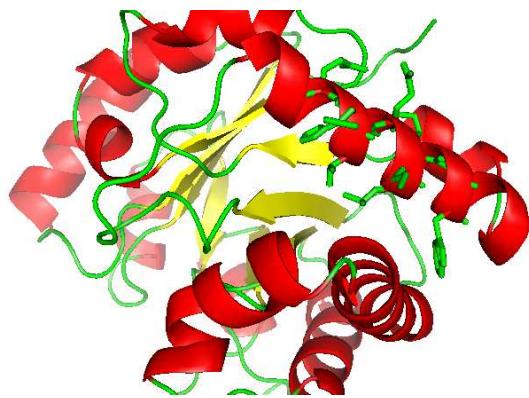


Figure 1.6: Ribbon representation of a protein. The backbone is traced in a cartoon-like representation.

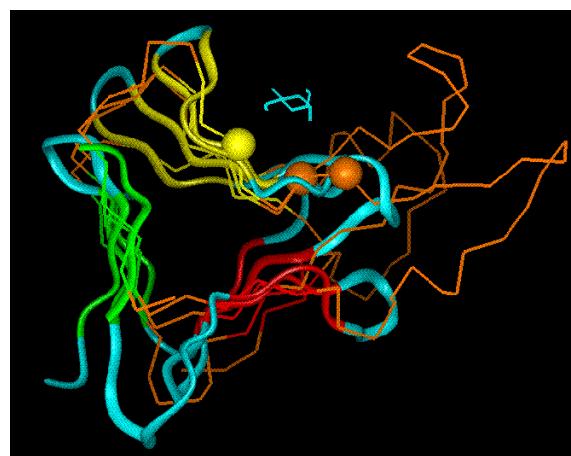


Figure 1.7: The protein “manose” represented by the SARF2 software.

	SEQUENCE ALIGNMENT	STRUCTURAL COMPARISONS
HOW TO	Sequences of proteins written one above the other, and then gaps are inserted so that similar amino acids are placed in the same columns	Protein structures are superimposed by fitting the atoms (backbone) so that the average deviation in Euclidian distance between them is minimal
EVOLUTIONARY SIGNIFICANCE	Sequence similarity = evolutionary relationship	structural similarity = evolutionary relationship (convergence to same structure by different evolutionary origins has not yet been found)

Table 1.1: Two approaches of comparisons of proteins with known sequence or additional known structures.

With (2) “structural alignment” we mean

based on known two or more 3D structures to find equivalent residues (amino acids) in both amino acid sequences, where equivalent means that they are at corresponding positions in both structures.

Structural alignment and comparison methods like CE, DALI, SSAP, VAST, SARF2 and COMPARE will be introduced later. Some of them rely on dynamic programming or on distance matrices. Structural alignment methods do not use PAM or BLOSUM matrices which have been designed for pointwise similarities but not for 3D positions:

1. Inside the protein core region: the substitutions are more restricted by space and interaction with other amino acids,
2. Outside the protein core region: the substitutions are less restricted.

Measuring the 3D Structure

Through nuclear magnetic resonance spectroscopy (NMR) and X-ray crystallography the 3D coordinates of the atoms of a protein can be determined. These coordinates beside other relevant information like which organism, how the probes are obtained, the resolution, the amino acid sequence, special bonds, etc. are stored in the Brookhaven database (PDB) in a PDB file.

Note, there are two principal types of data bases for measurements: (1) sequences (DNA) from genome sequencing and (2) structures (proteins) from NMR or x-ray. The sequence data bases are in general much larger than the structural data bases.

Hydrophobic Profiles

The hydrophobic profiles of proteins given as amino acid sequences are helpful to detect surface vs. buried regions or transmembrane helices. Fig. 1.8 shows the hydrophobic profile of subunit M of the photosynthetic center of *Rhodopseudomonas Viridis*. The plot reliably predicts the five hydrophobic membrane-spanning helices. Fig. 1.9 shows another example where membrane helices are located.

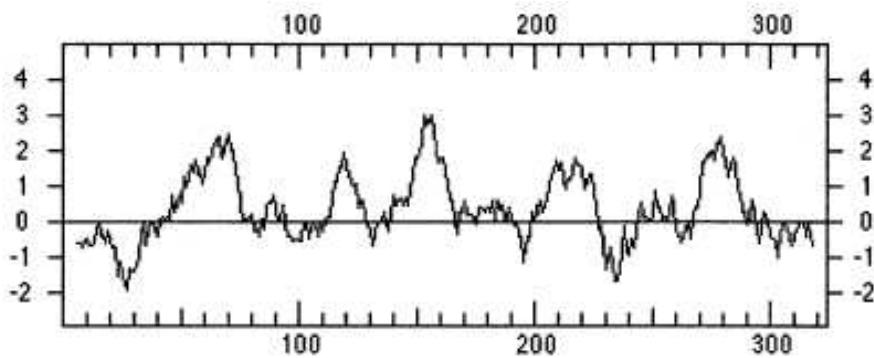


Figure 1.8: Hydrophobicity plot for subunit M of the photosynthetic center of *Rhodopseudomonas Viridis*. Membrane helices are found at positions 52-78, 110-139, 142-167, 197-225, and 259-285.

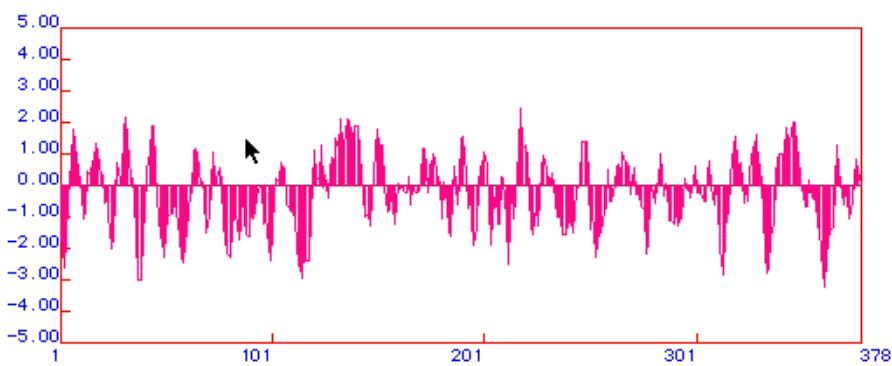


Figure 1.9: Hydrophobicity plot for the human actin. With at least 3 peaks above 2.00, actin is most likely an integral membrane protein.

Secondary Structure Prediction

The secondary structure of a measured protein can be extracted from the coordinate files by DSSP (Dictionary of Secondary Structure of Proteins) or by STRIDE (STRuctural IDEntification).

Protein secondary structure prediction methods which we will discuss include GOR, Chou-Fasman, Lim's methods, neural networks, PHD (Profile Network from Heidelberg), PSIPRED, and others.

The secondary structure is a main element of the tertiary, 3D structure of a protein as can be seen by the fact that structural databases like SCOP and CATH classify structures according to their main secondary structure type:

1. α -class: proteins only consisting of α -helices connected by loops,
2. β -class: proteins only consisting of β -sheets,
3. α/β -class: β -sheets and helices combined, e.g. parallel β -sheets connected by α -helices,
4. $(\alpha + \beta)$ -class: α -helices are separated from β -sheets,
5. $(\alpha \text{ and } \beta)$ -class: multi-domain, that means the α -helices are not in contact with the β -sheets,
6. membrane and cell surface proteins.

Tertiary Structure Prediction

To predict the 3D structure based on the 1D sequences obtained by genome sequencing has been a major goal of bioinformatics since decades. Most proteins are only given as 1D sequences and neither their function nor their folding and stability characteristics are known. If the protein 3D structure can reliably be predicted then its function can be inferred and its stability properties analyzed. Finally, the human genome can be understood.

For 3D structure prediction there exist two basic approaches: (1) compare the structure with proteins with known structure, or (2) to predict the structure just from the sequence including physical laws and empirical knowledge.

Item (1) can be subdivided into comparative modeling by (1.1) sequence-sequence comparison (alignment) and comparative modeling by (1.2) sequence-structure comparison (threading).

If after a global sequence alignment the identity between the proteins is 25-45%, then the two structures are similar. When the similarity is about 45%, then structures are equal, i.e. their structure match exactly.

For (1.1) alignment methods like BLAST are used and for (1.2) different threading methods are introduced. Threading methods put a new sequence on a known structure and compute how well the new sequence fits the known structure, e.g. how many hydrophobic amino acids are buried.

Item (2) includes “ab initio prediction” and molecular modeling or quantum mechanical modeling. Methods from category (2) can be applied if the protein possesses a novel structure which has not yet been solved by NMR or x-ray methods. More importantly, these methods can be used to design new structures and, therefore, new proteins.

Rosetta is the best known ab initio method which puts peptides of 3 to 9 amino acids together by an optimization method.

With molecular dynamics the folding process of a protein can be simulated and with quantum mechanic simulation the docking of a ligand can be analyzed. These methods use first principles and do not rely on empirical data. Therefore they are more exact than the empirical methods like threading. However with these days computers folding simulations would take hundreds of years to fold large proteins.

Chapter 2

Chemical and Physical Background

2.1 Atomic Bounds: A Basic Introduction

Macromolecules are made up of smaller units linked one to the next by specific bonds. The basic repetitive units of nucleic acids are nucleotides, while those of proteins are amino acids. In both types of macromolecule the individual constituent atoms form specific bonds according to their chemical and physical properties. The atoms are in turn composed of varying numbers of three main subatomic particles:

- **Electrons**, negatively charged, are located in shells. The main cause of chemical bonding is the interaction of electrons in the outermost shell with adjacent nuclei.
- **Protons**, positively charged, protons and neutrons are the constituents of the atomic nucleus
protons weigh about 1836 times the mass of an electron. The number of protons in the nucleus defines the chemical element and thus the properties of the atom.
- **Neutrons**, uncharged, are located in the nucleus and weigh about 1838 times the mass of an electron. The number of neutrons determines the isotope of an element.

The positive charge of a proton is equal in strength to the negative charge of an electron. If the number of protons in an atom equals the number of electrons, then the atom itself has no overall charge, it is neutral. The number of electrons in the outermost shell – the so-called valence shell – of an atom governs its bonding behavior. Atoms with a full valence shell (8 electrons for most atoms) are most stable. Hence the noble gases (the rightmost column of the periodic table of elements) are inert and, conversely, atoms with few electrons in the valence shell and atoms that need only few electrons to fill it are reactive. In order to reach the ideal, stable electron configuration of a full valence shell, atoms form chemical bonds by sharing electrons with other atoms (covalent bonds), or by electron transfer between atoms (non-covalent bonds). The linear or primary sequence of macromolecules like proteins , DNAs, RNAs is maintained by covalent bonds, whereas the 3D structure is stabilized by non-covalent intra- or intermolecular interactions. The spatial structure depends on the surrounding solvent context in which the macromolecule is placed. The most important covalent bonds in biology (C-C, C-H), have bond energies in the range of 300-400 kJ/mol while non-covalent bonds are usually 10 to 100 times weaker. In order for molecular processes of the cell to function macromolecules must exhibit a degree of conformational flexibility and be able to break and make certain bonds. This is mainly possible where weak interactions/bonds are involved indicating their essential importance.

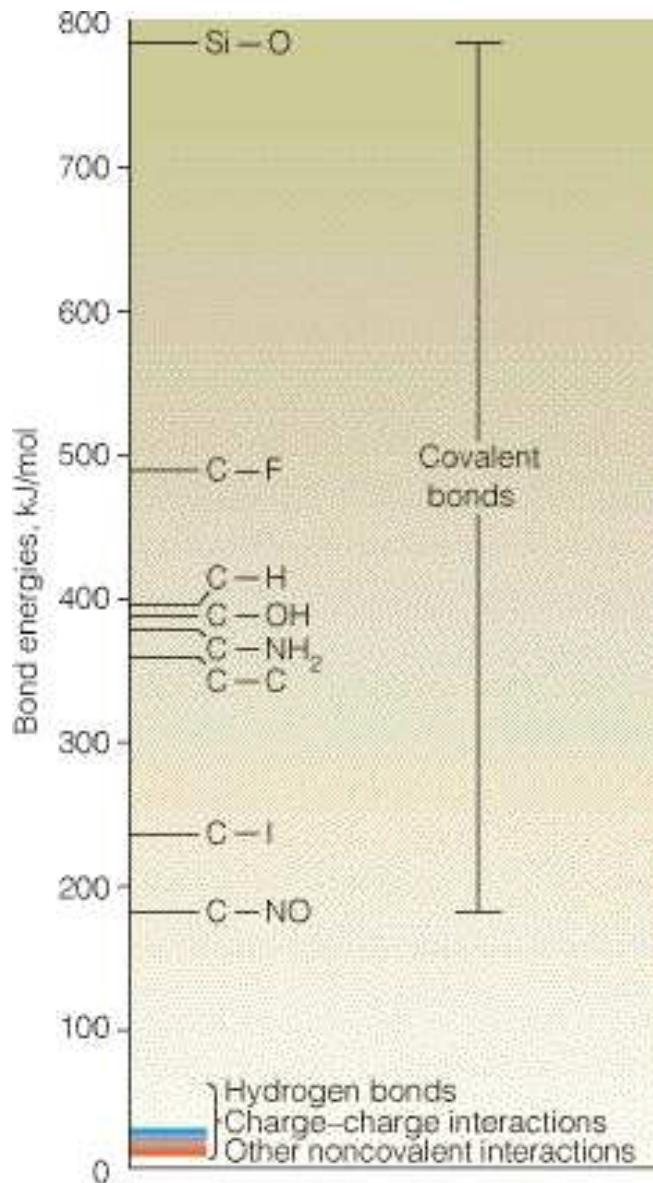


Figure 2.1: Covalent and non-covalent bond energies. Energies of non-covalent bonds are one to two orders of magnitude weaker than energies of the covalent bonds found in biochemical compounds.

As previously indicated, non-covalent interactions are responsible for the secondary and tertiary structure of proteins and hence for their functions. They also play a role in the formation of protein complexes (quaternary structure) where two or more polypeptide chains are assembled to a bigger unit.

2.1.1 Non-Covalent Interactions

Non-covalent interactions depend on the electrostatic state of the participating molecules and thus the electrostatic forces exerted upon one another. All interactions but those of hydrogen bonds become weaker with growing distance. Non-covalent bonds are weak by nature and must therefore work together to have a significant effect. In addition, the combined bond strength is greater than the sum of the individual bonds. This can be explained thermodynamically as the free Gibbs energy G of multiple bonds is greater than the sum of the enthalpies H of each bond due to entropic effects S .

$$\Delta G = \Delta H - T\Delta S$$

Non-covalent interactions always involve *electrical charges*.

2.1.1.1 Charge-Charge Interactions or Ionic Bounds

Based on electrostatic forces between two oppositely charged ions. Many cellular molecules carry a net electrical charge that makes them susceptible to reacting with other charged molecules. Coulomb's law describes the forces between a pair of charges q_1 and q_2 separated by a distance r by the formula

$$F = k \cdot \frac{q_1 \cdot q_2}{r^2}$$

Where

- F is the magnitude of the force exerted,
- q_1 is the charge on one body,
- q_2 is the charge on the other body,
- r is the distance between them,
- K is the electrostatic constant or Coulomb force constant defined by

$$k_C = \frac{1}{4\pi\epsilon_0} \approx 8.988 \times 10^9 Nm^2C^{-2} (\text{also } mF^{-1})$$

- $\epsilon_0 \approx 8.854 \times 10^{12} C^{-2}Nm^2$ (*also mF⁻¹*) is a physical constant that defines the permittivity of free space, also called electric constant

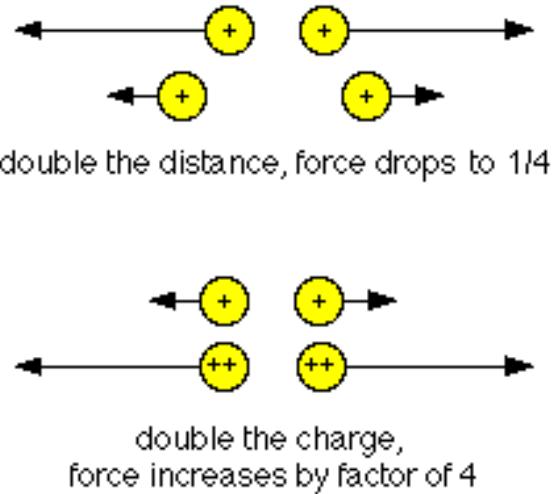


Figure 2.2: Coulomb's Law. Charged objects create an invisible surrounding electric force field. The greater the charges are, the stronger the force. The greater the distance, the weaker the force becomes.

Charged objects create an invisible surrounding electric force field. The greater the charges are, the stronger the force. The greater the distance, the weaker the force becomes.

The force is inversely proportional to r^2 . When q_1 and q_2 have the same charge, the force is positive. This corresponds to a *bf* repulsive force. When one charge is positive and the other is negative, the sign of the force is negative. This corresponds to an *bf* attractive force. Crystals of salts like NaCl are stabilized by such charged-charged interactions. Inside and outside a cell, charges are always separated by water or by other molecules, so an additional dimensionless constant, depending on the composition of the surrounding medium, is introduced to formula 2. This constant is called the dielectric constant and represents the effect of the biological environment in which the actual force is always less than that given by the equation 2.

$$F = k \cdot \frac{q_1 \cdot q_2}{\varepsilon \cdot r^2}$$

Thus, the larger the value of ε the weaker the force between the interacting charges. For organic substances this value is in the range of 1 to 10 while for water it is higher, approximately 80, the reason being that the charged particles within an aqueous environment interact weakly unless they are very close together. The interaction energy U of a pair of reacting molecules can be measured by transforming Coulomb's formula as follows:

$$U = k \cdot \frac{q_1 \cdot q_2}{\varepsilon \cdot r}$$

If the charges have opposite signs, the interaction energy U is negative signifying *attraction*. If the charges have the same sign, the interaction energy U is positive signifying *repulsion*. U approaches zero when r becomes very large, so there is *no interaction*. Conclusion:

- The force F between the charges depends only on the distance.

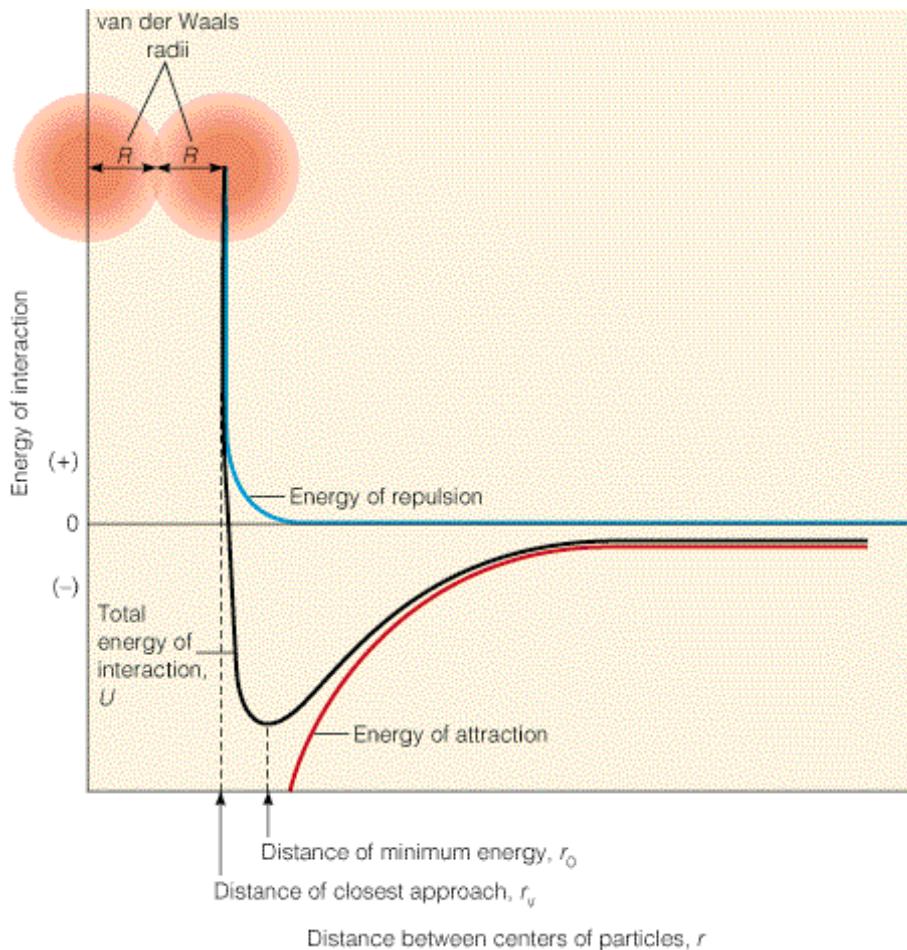


Figure 2.3: Non-covalent interaction energy of two close particles.

- The interaction energy U varies with the distance and it is inversely proportional to the first power of r .

The interaction energy U of two atoms, molecules, or ions is plotted on the Y axis versus the distance of their centers r plotted on the x -axis. The total U at any distance is the sum of the attractive (+) and repulsive (-) energies. As the distance between the particles decreases (reading right to left along the x -axis), both the attractive and repulsive energies increase but at different rates. As the repulsive energy increases the distance reaches the barrier of closest approach (r_v). The van der Waals radii are also defined. Normally the position of minimum energy (r_o) is very close to r_v .

2.1.1.2 Dipole Interactions

The protons and neutrons in the nucleus are held together very tightly. The nucleus does not change. However, some of the electrons of the outermost shell, even when this shell is complete, may be distributed asymmetrically in a way that the uncharged atom becomes partially charged. Such an atom is called a *dipole* and depending on the surrounding medium, particles and their

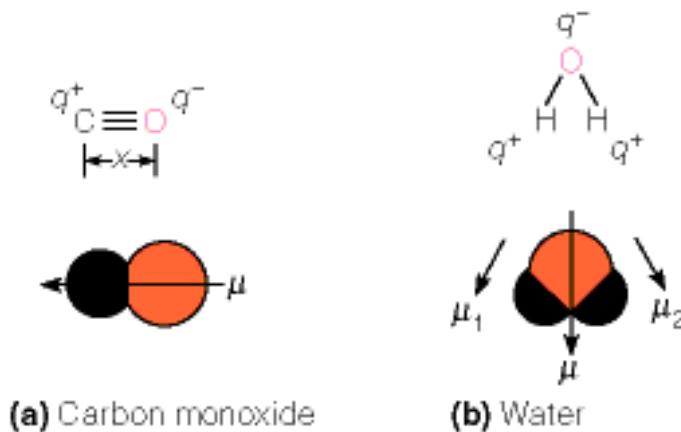


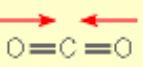
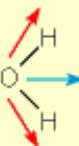
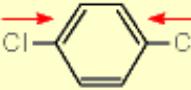
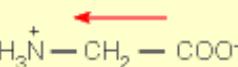
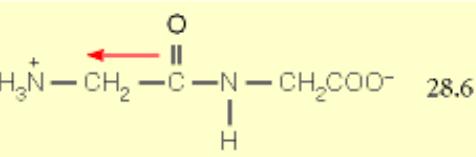
Figure 2.4: Dipole moments.

charges, it can behave in different ways. We distinguish between induced, instantaneous and permanent dipoles. The transient dipole charge allows atoms with no net charge to interact with other charged particles or dipoles in the medium. The *dipole moment* μ describes the asymmetry of a molecule by measuring the polarity.

Carbon monoxide: the dipole moment spans the O-C axis due to the slight negativity of the oxygen end compared to the carbon end. (b) Water: the vector sum (μ) is the result of the two moments along the O-H bonds. The dipole arises due to excess negative charge of oxygen and excess positive charge of hydrogen. The dipole moment can be calculated as $\mu = qx$, where q represents the vector pointing towards $q+$ and x represents the distance that separates the charges.

The larger the distance between the ionized groups the greater the dipole moment. Molecules that consist of two or more sub-molecules with their own dipole moments have a global dipole moment that results from the sum of the single dipole moments of the constitutive molecules. If the dipole vectors of a molecule have the same magnitude but opposite directions, their effects cancel each other out. We conclude that the molecules must be asymmetric to have a dipole moment.

Within a cell or in aqueous medium, a permanent dipole can be attracted by a close-by ion, establishing so-called *charge-dipole* interactions. The strength of the interaction depends on the orientation of the involved molecules. A permanent dipole can also interact with another permanent dipole leading to a *dipole-dipole* interaction that depends on the respective orientation of the dipoles. This dipole-dipole interaction works like an ionic interaction but in a weaker manner because only partial charges are involved. As mentioned above molecules with no net charge can adopt transient partial charges once they are in the presence of an electric field; they become *induced dipoles*. The field can be generated by a neighboring charged or dipolar molecule, accordingly, we distinguish between three types of interactions depending on the kind of dipole-inducing agent: *induced-dipole interaction* (inducing agent = polarizable molecule), *charge-induced dipole interaction* (cations, anions), *dipole-dipole interactions* (permanent dipoles). The interaction energy decreases from the charge-dipole interactions to the dipole-induced dipole interactions within

Molecule	Formula	Dipole Moment (D) ^a
Carbon monoxide		0.12
Carbon dioxide		0
Water		1.83
<i>ortho</i> -Dichlorobenzene		2.59
<i>para</i> -Dichlorobenzene		0
Glycine		16.7
Glycylglycine		28.6

^aThe common units of dipole moment are *debyes*; 1 debye (D) equals $3.34 \times 10^{-30} \text{ C m}$.

Table 2.1: Some Dipole Moments. Note that there is no dipole moment when the molecule is symmetric.

a range of $1/r - 1/r^5$ explaining the weakness of the latter and the small effective scope. As the electrons in the outermost shell of an atom are not static but fluctuate, it can happen that if two uncharged molecules are close enough they can interact by synchronizing the fluctuation of their electrons resulting in a net attractive force, the molecules become *instantaneous dipoles*. This kind of interaction between two induced dipoles is called London or *dispersion force*. In spite of the short range of scope, London forces become relevant when planar molecules stack on top of each other as it is the case with the internal packing of proteins and nucleic acids.

2.1.1.3 Van der Waals Forces or Dispersion

These interactions involve the attraction between temporarily induced, short-living dipoles in non-polar molecules. This polarization can be induced either by a polar molecule or by the repulsion of negatively charged electron clouds in non-polar molecules. If molecules or atoms come very close together, their outer electron orbitals can overlap and mutual repulsion occurs. The repulsion increases as the radii of the atoms approach $1/r^{12}$. An example of this effect is chlorine dissolving in water:



Water is permanently polarized as explained above. The dipole of the chlorine molecule is induced by the electric field of the permanent water dipoles.

Van der Waals interactions define the minimal distance between interacting molecules and hence determine the shape of molecular surfaces. As depicted in 2.3, the van der Waals radius R is the effective radius for closest molecular packing. For two identical spherical molecules with radius R_1 and R_2 , the boundary or distance $r_v = 2R$ and for two molecules with van der Waals radii, $r_v = R_1 + R_2$. Biological molecules are not spherical, so the concept of van der Waals radii is extended to atoms or groups of atoms within a molecule.

2.1.1.4 Hydrogen Bond

When a hydrogen atom is bound covalently to an electronegative atom, an interaction between this hydrogen and another electronegative atom can occur. Hydrogen bonds can be formed between molecules (inter-molecularly) as well as within molecules (intra-molecularly). This interaction is stronger than van der Waals forces but weaker than ionic or covalent bonds. Among other reactions, hydrogen bonds are responsible for the secondary, tertiary and quaternary structure of nucleic acids and proteins as well as for the high boiling point of water (100°C). The atom to which the hydrogen is covalently bound is called the *hydrogen bond donor* and the atom with the free electron pair is called the *hydrogen bond acceptor*. Hydrogen donors tend to be electronegative atoms such as oxygen, nitrogen and fluorine because their high electronegative level makes the hydrogen atom bonded to them partially positive so susceptible to the attraction of the pair of electrons of the hydrogen bond acceptor groups. In biological compounds only nitrogen and oxygen are sufficiently electronegative atoms to act like strong hydrogen bond donors. The bond

	R(nm)
Atoms	
H	0.12
O	0.14
N	0.15
C	0.17
S	0.18
P	0.19
Groups	
—OH	0.14
—NH ₂	0.15
—CH ₂ —	0.20
—CH ₃	0.20
Half-thickness of aromatic ring	0.17

Table 2.2: Some values of Van der Waals radii. These values represent the distances of closest approach for another atom or group.

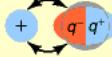
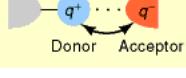
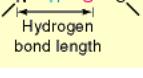
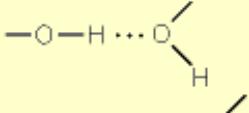
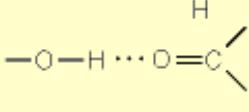
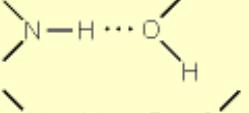
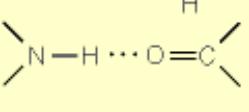
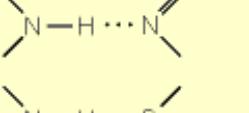
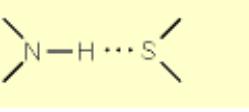
Type of Interaction	Model	Example	Dependence of Energy on Distance
(a) Charge-charge Longest-range force, nondirectional		--NH_3^+ $\text{--C}=\text{O}$	$1/r$
(b) Charge-dipole Depends on orientation of dipole		--NH_3^+ H_3O^+	$1/r^2$
(c) Dipole-dipole Depends on mutual orientation of dipoles		H_3O^+ H_3O^+	$1/r^3$
(d) Charge-induced dipole Depends on polarizability of molecule in which dipole is induced		--NH_3^+ C_6H_5^-	$1/r^4$
(e) Dipole-induced dipole Depends on polarizability of molecule in which dipole is induced		H_3O^+ C_6H_5^-	$1/r^5$
(f) Dispersion Involves mutual synchronization of fluctuating charges			$1/r^6$
(g) van der Waals repulsion Occurs when outer electron orbitals overlap			$1/r^{12}$
(h) Hydrogen bond Charge attraction + partial covalent bond			Length of bond fixed

Figure 2.5: Types of non-covalent interactions. Dipole interactions (b-f) of molecules with no net charge.

Donor···Acceptor	Bond Length ^a (nm)	Comment
	0.28 ± 0.01	H bond formed in water
	0.28 ± 0.01	
	0.29 ± 0.01	
	0.29 ± 0.01	
	0.31 ± 0.02	
	0.37	Relatively rare; weaker than above

^aDefined as distance from center of donor atom to center of acceptor atom. For example, in the $\text{N}-\text{H}\cdots\text{O}=\text{C}$ bond it is the $\text{N}-\text{O}$ distance.

Table 2.3: Types of hydrogen bonds in biological compounds.

length for the most important hydrogen interactions found in biological molecules are in the range of 0.28 to 0.31 nm and are the ones corresponding hydroxyl groups (OH-OH), carbonyl groups (C=O) and amine groups (N-H).

Hydrogen bonds have characteristics of both covalent and non-covalent interactions. On one hand they are like charged-charged interactions (non-covalent) due to the partial negative charge of the hydrogen bond acceptor and the positive charge of the hydrogen bond donor. On the other hand the fact that electrons are shared is reminiscent of covalent bonds. Linus Pauling (1901-1994) proposed for the first time the partially covalent nature of the hydrogen bond, but it was not until the late 1990's that F. Cordier employed NMR techniques to prove this proposition. He transferred information between the nuclei involved in hydrogen bonds, which is only possible if the hydrogen bond possesses some covalent character. This ambivalent character is reflected in the bond length of the hydrogen bond as the expected van der Waals radius of the bond =N-H-O=C= does not correspond to the sum of the single radii defined in 2.2

$$R_1\text{H} + R_2\text{O} = 0.12\text{nm} + 0.14\text{nm} = 0.26\text{nm}$$

The length of the covalent O-H bond is 0.10nm, the actual length of the hydrogen bond in this example is 0.19nm, so it lies between that of a covalent and a non-covalent bond. Hydrogen bonds can vary in strength from very weak as in $\text{N}-\text{H}\cdots\text{O}$ (8 kJ/mol) to extremely strong as in

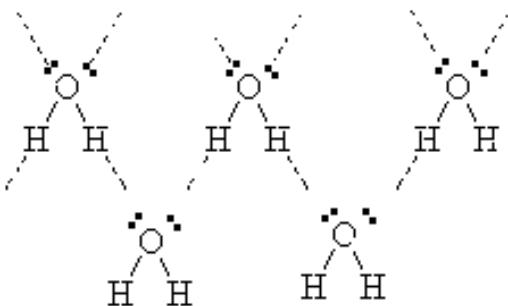


Figure 2.6: Hydrogen bonds in water.

$\text{O}-\text{H}\cdots:\text{N}$ (29 kJ/mol). Between these values others can be found like $\text{O}-\text{H}\cdots:\text{O}$ (21 kJ/mol) and $\text{N}-\text{H}\cdots:\text{N}$ (13 kJ/mol). The length of the hydrogen bond depends on the temperature, pressure and the bond strength. In addition to these three conditions also the bond angle and the value of the local dielectric constant of the environment influence the length. In the cell macromolecules are surrounded by water, so water is also the medium in which biological reactions take place. It clearly follows that hydrogen bonds are of fundamental importance in biological processes.

Water has two hydrogen atoms and one oxygen, each water molecule can bond with up to four other molecules as follows: the oxygen of one water molecule has two lone pairs of electrons each of which can form a hydrogen bond with hydrogens of two other water molecules; the two hydrogen atoms of the same water molecule can form two hydrogen bonds with two oxygens from two other water molecules. The number of hydrogen bonds a molecule participates in fluctuates with time and depends on the temperature (the higher the temperature, the less hydrogen bonds).

As mentioned above, hydrogen bonds are of crucial importance in biology since they can determine the conformation and folding ways of macromolecules like proteins and nucleic acids. This kind of interaction facilitates intermolecular and intramolecular effects causing the macromolecules to fold into a specific shape which determines their biological functions. Examples are the double helical structure of the DNA molecule in which the base pairs are linked together by hydrogen bonds, and the hydrogen bonds formed between the backbone oxygens and amide hydrogens in proteins. Depending on the number of amino acids in a protein that lie between those participating in the hydrogen interaction, the secondary structure being formed can be an α -helix ($n + 4$), a β -sheet when two strands are involved and so on. Hydrogen bonds also play a role in forming the tertiary structure of proteins through the interaction of residues.

We can conclude with three main facts about hydrogen bonds:

- Hydrogen bonds have an average length of about 0.33 nm.
- Hydrogen bonds are highly directional - the donor H tends to point directly at the acceptor electron pair.
- The energy of hydrogen bonds is greater than most other non covalent interactions.

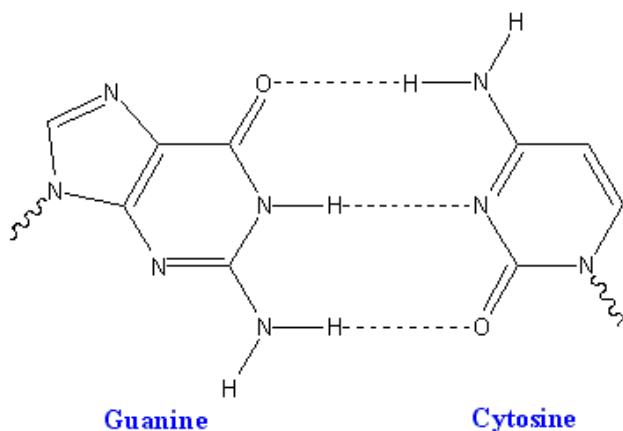


Figure 2.7: Triple hydrogen bond in a DNA base pair.

compound	molecular weight	melting point ($^{\circ}C$)	boiling point ($^{\circ}C$)	heat of vaporization (kj/mol)
CH ₄	16.04	-182	-164	8.16
NH ₃	17.03	-78	-33	23.26
H ₂ O	18.02	0	+100	40.71
H ₂ S	34.08	-86	-61	18.66

Table 2.4: Properties of water. The main explanation of these properties is the tendency of water to form hydrogen bonds.

2.1.1.5 Hydrophobic-Hydrophilic Interactions

The fact that water is in the liquid state at room temperature while other compounds of low molecular weight are gases can be explained by its dipolar character that makes it possible to form hydrogen bonds with other water molecules. It explains why water is the universal environment life has selected. Due to its composition, water can act as a permanent dipole with all its implicated consequences.

Properties of water like high viscosity and surface tension, a relatively high boiling point and the decrease of density when changing into the solid state are all due to the two lone electron pairs of the outer orbital of the oxygen atom that act as perfect hydrogen bond acceptors and to the HOH group that acts as hydrogen donor. An unusual amount of energy is required in order to break all the hydrogen bonds. When water freezes to ice a rigid tetrahedral molecular lattice is created in which each molecule is H-bonded to four others. Water is denser in its liquid than in its solid state because when the ice lattice breaks down molecules can move closer together. This property is crucial for being the medium in which evolutionary changes occur.

Water is an excellent solvent because of its hydrogen bonding potential and its polar nature. Chemical groups like hydroxyl compounds (-OH), amines (-NH₂), sulphydryl compounds (-SH),

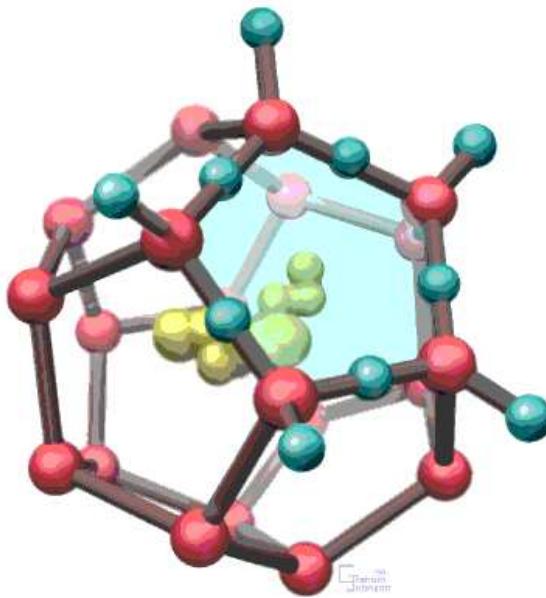


Figure 2.8: Clathrate structure. Red balls oxygen, blue balls hydrogen. The ordered structure may extend into the surrounding water.

esters (-CHO) and ketones (C=O) can be dissolved in water by *hydrophilic interactions*. All substances susceptible to solvation in water are called *hydrophilic* and are composed by ions or polar molecules that attract water molecules through electrical charge effects forming *hydration shells* in which water molecules surround the compounds covering the acceptors groups. Polar biological substances like urea and some amino acid residues can form hydrogen bonds with water and thus be dissolved. Other chemical groups, mainly hydrocarbons with a backbone composed of C-H that are hence non-polar and non-ionic, are classified as hydrophobic groups because they are insoluble in water. These molecules do not form *hydration shells* but clathrates or cages around non-polar molecules. The entropy of such molecules is decreased contributing to the low solubility in water. Water molecules of the surrounding medium are not as strongly attracted to such molecules as they are to other water molecules so the tendency to dissolve them is very low to non-existent.

There is a third type of molecules, mainly hydrocarbons, that simultaneously show both hydrophobic and hydrophilic properties. Usually one end or side of the molecule is hydrophilic and the other end or side is hydrophobic. They are called *amphipathic* and their capability of repulsing water on one hand and being attached to it on the other hand is responsible for one of the main important characteristic of life: *the isolation and partition between cellular compartments* (plasmatic membrane, nuclear membrane, organelles, etc). When dissolved in water the amphipathic molecules tend to form monolayers, micelles (spherical structures formed by a single layer of molecules) or bilayer vesicles on the water surface with only the hydrophilic groups immersed in water and the hydrocarbon tails in the interior arranged in parallel arrays which allows them to interact via van der Waals interaction. Biological membranes are composed of such amphipathic molecules; phospholipids adopt this special configuration.

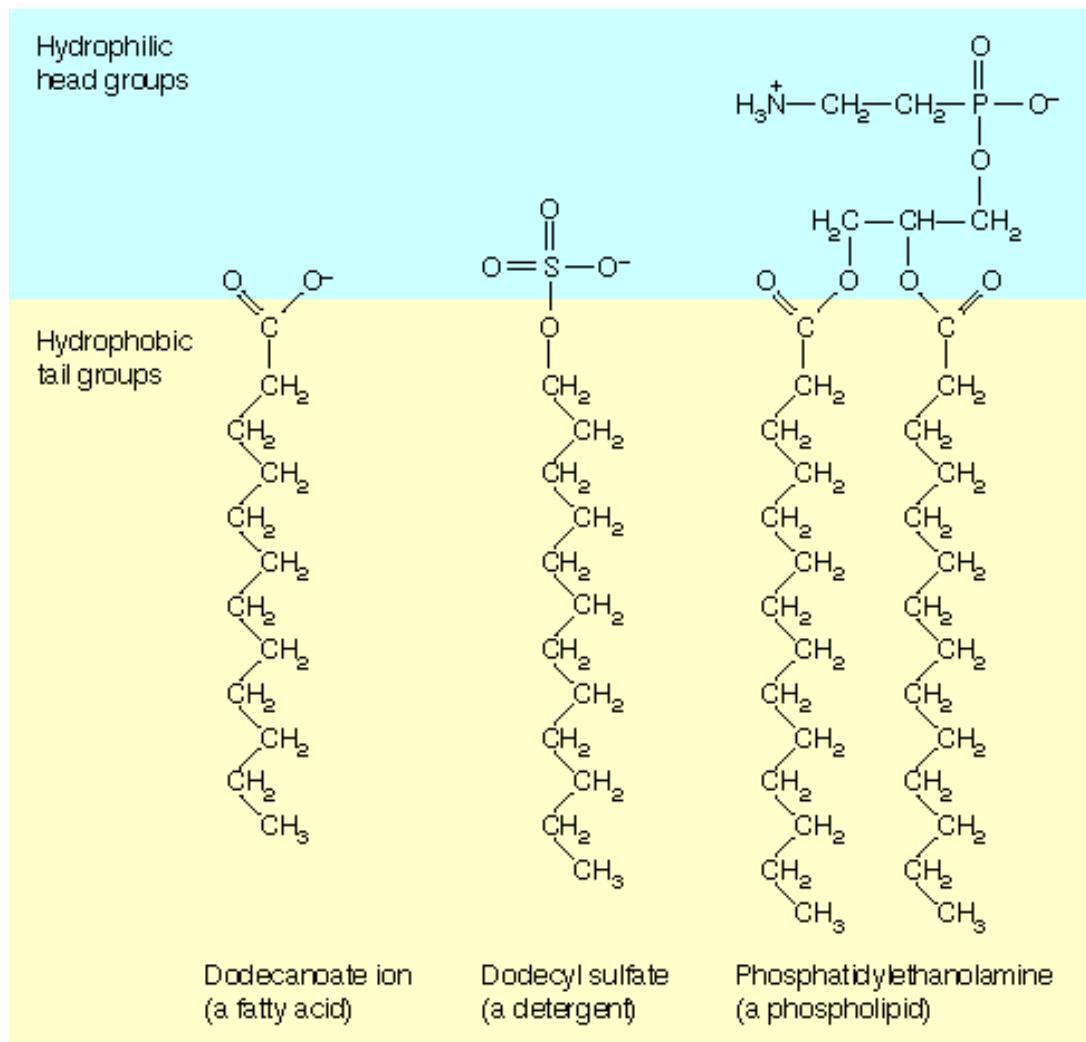


Figure 2.9: Amphipathic molecules. Three examples of the ambiguous behavior of amphipathic molecules with hydrophilic head and hydrophobic tail.

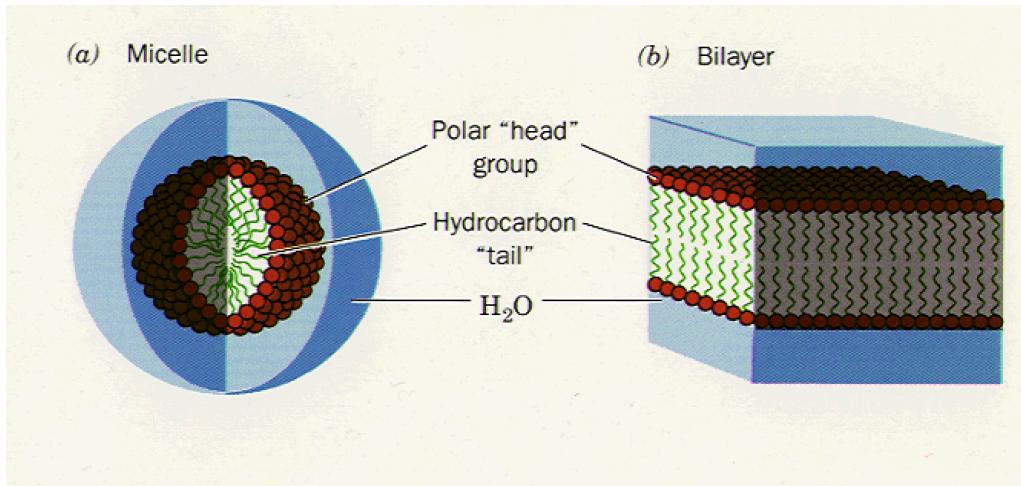


Figure 2.10: Immersion of amphipathic molecules in water. When amphipathic substances are mixed with water, micelles, bilayer vesicles and a monolayer can form. In all cases hydrophilic heads are in touch with water whereas hydrophobic tails are hidden.

Bond Type	Relative Strength
Ionic Bonds	1000
Hydrogen Bonds	100
Dipole-Dipole	10
Van der Waals	1

Table 2.5: The relative strength of different bonds.

2.1.2 Conclusion

We can summarize the different interactions by classifying them into

1. **Electrostatic interactions:** ionic interactions, hydrogen bonds, dipole-dipole interactions, hydrophobic-hydrophilic interactions.
2. **Electrodynami c interactions:** Van der Waals force/London dispersion forces.

Electrostatic interactions are described by Coulomb's law. The basic difference between them is the strength of their charge. *Ionic interactions* are the strongest with integer level charges, *hydrogen bonds* have partial charges that are about an order of magnitude weaker, and *dipole-dipole interactions* also come from partial charge further order of magnitude weaker.

All bonds can be explained by quantum theory. Covalent bonds, as the strongest bonds, alone can not explain the complexity of molecular structure in biology and the inclusion of weaker, non-covalent bonds is necessary. Information related to such weak interactions allows us to understand molecular properties and processes of the cell.

2.1.3 Glossary

Amphipathic - The molecular property of having both hydrophobic and hydrophilic portions.

Usually one end or side of the molecule is hydrophilic and the other end or side is hydrophobic.

Coulomb's Law - Describes the force between two charges in a vacuum. The force, F , is defined as $F = k \cdot \frac{q_1 q_2}{r^2}$, where q_1 and q_2 are the charges and r is the distance between the charges.

Covalent bonds - The strong chemical bonds between atoms in an organic molecule.

Dielectric constant - A constant, called ϵ , that modifies Coulomb's law to account for shielding by molecules placed between the charges in a medium. With this modification, the force F between charges in a dielectric medium (non-vacuum) is $F = k \cdot \frac{q_1 q_2}{\epsilon r^2}$, where q_1 and q_2 are the charges and r is the distance between the charges.

Dipole Moment - Molecules which have an asymmetric distribution of charge are dipoles. The magnitude of the asymmetry is defined by the dipole moment of the molecule.

Interaction Energy - The interaction energy, U , of two charged particles is the energy required to separate them from a distance, r , to an infinite distance. It is a measure of the energy required to overcome the electrostatic forces between them. $U = k \cdot \frac{q_1 q_2}{\epsilon r}$, where k is a constant, and q_1 and q_2 are the charges.

Hydration shell - The interactions of dipoles with cations and anions in aqueous solution cause the ions to become hydrated - surrounded by layers of water molecules called a hydration shell.

Hydrogen bond - An attractive interaction between the hydrogen atom of a donor group, such as OH or =NH, and a pair of non-bonding electrons on an acceptor group, such as O=C. The donor group atom that carries the hydrogen must be fairly electronegative for the attraction to be significant.

Hydrophilic - Refers to the ability of an atom or a molecule to engage in attractive interactions with water molecules. Substances that are ionic or can engage in hydrogen bonding are hydrophilic. Hydrophilic substances are either soluble in water or at least wettable.

Hydrophobic - The molecular property of being unable to engage in attractive interactions with water molecules. Hydrophobic substances are non-ionic and non-polar; they are non-wettable and do not readily dissolve in water.

Induced Dipole - A molecule such as benzene, which has a symmetric shape and no dipole moment in the absence of external interactions, can exhibit a slight redistribution of electronic charge due to interactions with an electric field. This induces a dipole in the symmetric molecule.

Non-covalent interactions - Attractive or repulsive forces, such as hydrogen bonds or charge-charge interactions, which are non-covalent in nature, are called non-covalent interactions.

Permanent Dipole - Molecules, such as water, which have an asymmetric distribution of electronic charge due to their molecular geometry and differences in electronegativity between atoms, are permanent dipoles.

van der Waals Radius - The effective radius of an atom or a molecule that defines closest molecular packing.

2.2 From chain polypeptide 1D configuration to folded 2D

2.2.1 Amino acids: classification and chemical-physical properties

Proteins are polymers of 20 different amino acids linked by specific type of bond, the peptide bond. The direct chain translated from the genetic code within ribosomes using mRNA as template is called the primary structure of the protein. When the non covalent hydrogen bonds are being formed between the N-H and -C=O groups of the invariant parts of the amino acids, the backbone chain that contains them can adopt either α -helices or β -strands given rise to the secondary structure. In a further step, the secondary structure elements fold linked also by loops, turns as well as parts without a defined structure into the globular tertiary structure. The last state of configuration, the quaternary structure, is achieved when the protein is formed by the association of more than one polypeptide folded chain. The organization of the genetic code reflects the chemical-physical grouping of the amino acids. The general formula is $\text{NH}_2\text{CHRCOOH}$. Both the amino and carboxylate groups are attached to the same carbon, which is called the α -carbon. The various alpha amino acids differ in which side chain (R group) is attached to their alpha carbon.

In solution at $pH = 7$ the amino and carboxylic acid groups ionize to NH_3^+ and COO^- . Except for glycine where R=H, amino acids are chiral and therefore enantiomers or optical isomers. Thus, the C_α is linked to four different substituents that can not be super-imposable on its mirror image. The 20 proteinogenics amino acids have a left-right asymmetry being the L-amino acid the most common and represent the vast majority of amino acids found in proteins.

Amino acids can be substituted between them within some restrictions. The values in 2.9 shows the larger the number, more common a particular replacement is. Amino acids with the smallest side chains like glycine and alanine are commonly replaced for one another. The negatively charged amino acids aspartic and glutamic have the highest frequency, which make sense since their R chains differ in just one more CH₂. Some surprises appear like replacement between serine and proline or glutamic acid and alanine. In some occasions serine substitutes praline because its OH side chain can receive an hydrogen bond from its own main-chain NH mimicking the fused ring of the proline.

2.2.1.1 Peptide bond

When the carboxylic acid (-COOH) linked to the C_α of one amino acid condenses with the amino group (-NH₂) bounded to the C_α of the next amino acid and a water molecule is expulsed, it says the peptide bond has been formed. The peptide bond is also called amide bond and its nature is covalent, thus the strongest interactions implying a pair of electrons to be shared. The reverse process is called hydrolysis and it requires the addition of water. Both synthesis and hydrolysis of peptide bond involve the action of enzymes that in case of synthesis almost always occurs in the ribosome and is directed by an mRNA template. The end of the polypeptide with the free amino group is the amino terminus (-N terminus) and the end with the carboxyl free group is the carboxyl terminus (-C terminus).

		1 st Position (5' end)			2 nd Position		3 rd Position (3' end)	
		U	C	A	G			
U	Phe	Ser	Tyr	Cys		U		
	Phe	Ser	Tyr	Cys		C		
	Leu	Ser	STOP	STOP		A		
	Leu	Ser	STOP	Tyr		G		
C	Leu	Pro	His	Arg		U		
	Leu	Pro	His	Arg		C		
	Leu	Pro	Gln	Arg		A		
	Leu	Pro	Gln	Arg		G		
A	Ile	Thr	Asn	Ser		U		
	Ile	Thr	Asn	Ser		C		
	Ile	Thr	Lys	Arg		A		
	Met	Thr	Lys	Arg		G		
G	Val	Ala	Asp	Gly		U		
	Val	Ala	Asp	Gly		C		
	Val	Ala	Glu	Gly		A		
	Val	Ala	Glu	Gly		G		

Table 2.6: Genetic code. All the 64 possibilities code for either an amino acid or a STOP signal for the end of the coding portion. Almost all of the amino acids can be specified by two or more codons differing only in the third position. Single-base changes elsewhere in the codon produces normally a different amino acid but with similar physical-chemical properties.

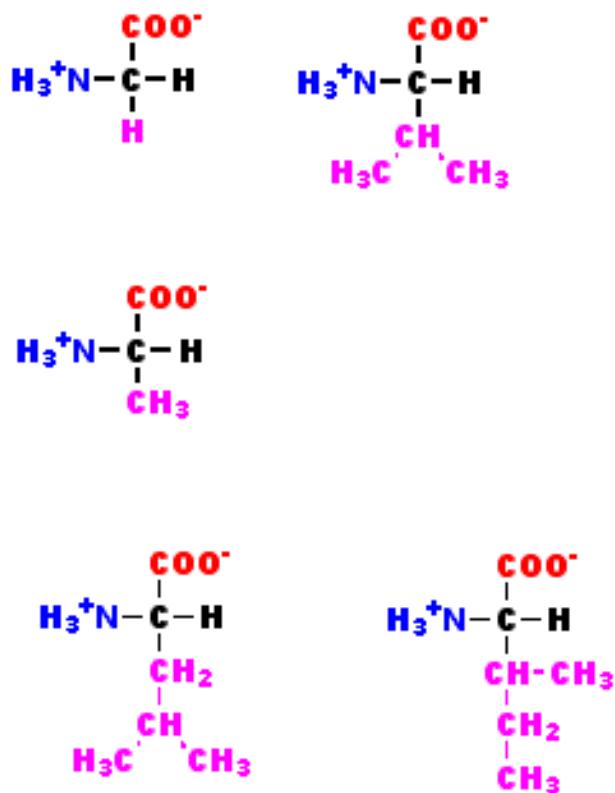


Figure 2.11: Aliphatic/hydrophobic R: Hydrocarbon side chain. Alanine (metal group), Glycine (hydrogen atom), Valine, Leucine and Isoleucine.

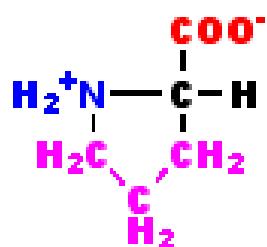


Figure 2.12: Proline residue. Proline has a hydrocarbon side and hence is also hydrophobic, but bounded either C as NH of the amine group.

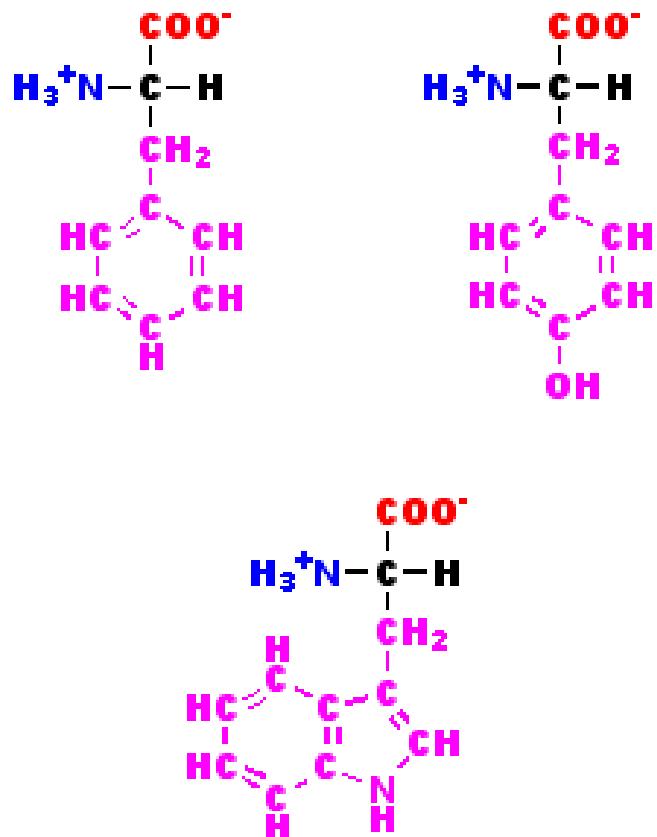


Figure 2.13: Aromatic R: Ring side chain or heterocyclic group. Phenylalanine is an Alanine with a phenyl group linked. Tyrosine is as phenylalanine but with an extra hydroxyl group which makes the amino acid less hydrophobic and more reactive. Tryptophan has an indol group.

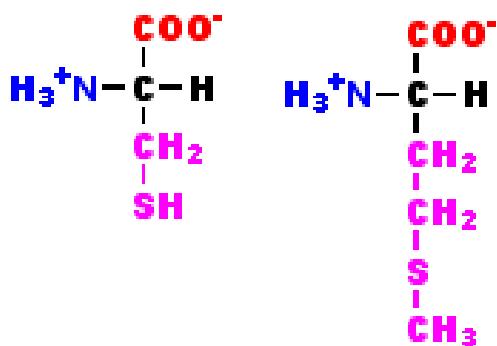


Figure 2.14: Sulfur-containing amino acids: Methionine and cysteine are the only sulfur-containing proteinogenic amino acids. Cysteine has a thiol group and methionine a thioether group. Disulfide bonds can form between two cysteine side chains.

POLAR AMINO ACIDS		NON-POLAR AMINO ACIDS	
Negative		Alanine	Ala A (1.8)
Aspartic acid	Asp D (-3.5)	Glycine	Gly G (-0.4)
Glutamic acid	Glu G (-3.5)	Valine	Val V (4.2)
Positive		Leucine	Leu L (3.8)
Arginine	Arg R (-4.5)	Isoleucine	Ile I (4.5)
Lysine	Lys K (-3.9)	Phenylalanine	Phe F (2.8)
Histidine	His H (-3.2)	Phenylalanine	Phe F (2.8)
Uncharged			
Asparagine	Asn N (-3.5)	Tryptophan	Trp W (-0.9)
Glutamine	Gln Q (-3.5)	Methionine	Met M (1.9)
Serine	Ser S (-0.8)	Proline	Pro P
Threonine	Thr T (-0.7)	Cysteine	Cys C (2.5)
Tyrosine	Tyr Y (-1.3)		

Table 2.7: Hydrophobicity indices (in brackets): The larger the number is the more hydrophobic is the amino acid. The most hydrophobic amino acids are isoleucine (4.5) and valine (4.2). The most hydrophilic ones are arginine (-4.5) and lysine (-3.9).

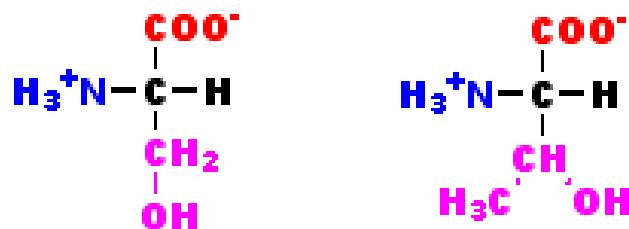


Figure 2.15: Hydroxylic R: Serine and threonine are the two amino acids with hydroxyl group which makes them more reactive and hydrophilic.

Inverse Table			
Ala	GCU, GCC, GCA, GCG	Leu	UUA, UUG, CUU, CUC, CUA, CUG
Arg	CGU, CGC, CGA, CGG, AGA, AGG	Lys	AAA, AAG
Asn	AAU, AAC	Met	AUG
Asp	GAU, GAC	Phe	UUU, UUC
Cys	UGU, UGC	Pro	CCU, CCC, CCA, CCG
Gln	CAA, CAG	Ser	UCU, UCC, UCA, UCG, AGU, AGC
Glu	GAA, GAG	Thr	ACU, ACC, ACA, ACG
Gly	GGU, GGC, GGA, GGG	Trp	UGG
His	CAU, CAC	Tyr	UAU, UAC
Ile	AUU, AUC, AUU	Val	GUU, GUC, GUA, GUG
START	AUG	STOP	UAG, UGA, UAA

Table 2.8: Amino acids can be translated for more than one triplet codon.

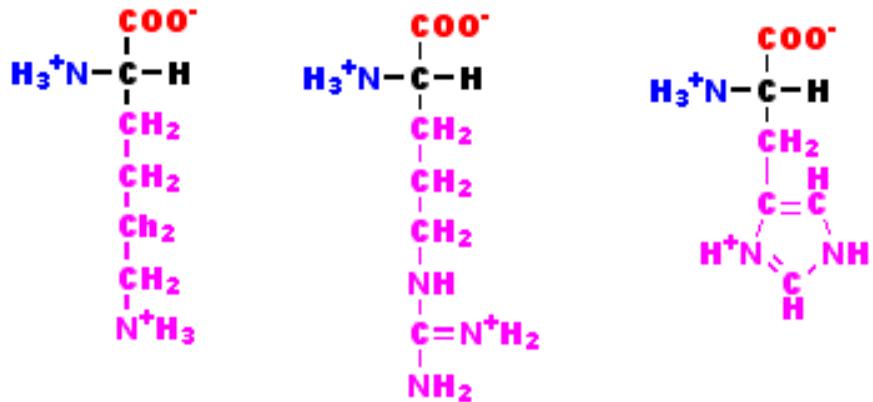


Figure 2.16: Basic amino acids.

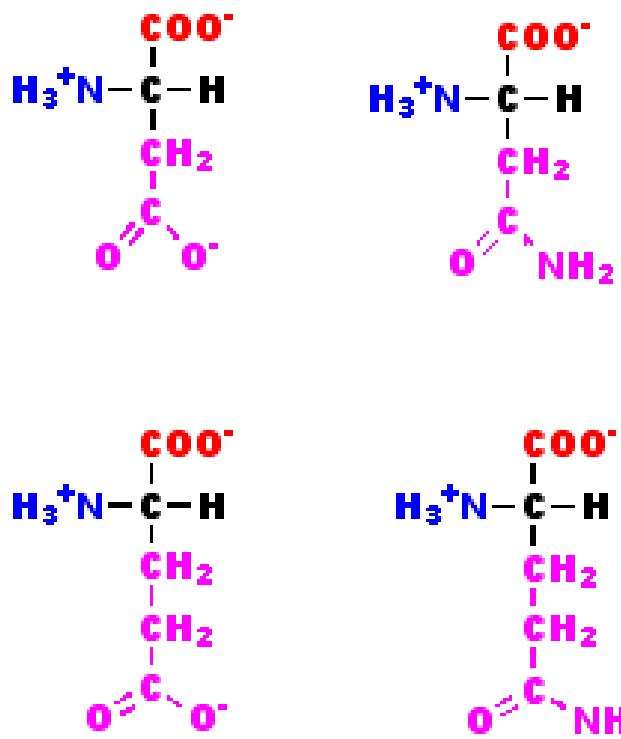


Figure 2.17: Acids R: aspartic acid and glutamic acid with their non charged derivates, the glutamine and asparagines which contain an amide terminal group instead of a free carbox.

Gly	Ala	Val	Leu	Ile	Met	Cys	Ser	Thr	Asn	Gln	Asp	Glu	Lys	Arg	His	Phe	Tyr	Trp	Pro
Gly																			
Ala 58																			
Val 10		37																	
Leu 2	10	30																	
Ile	7	66	25																
Met 1	3	8	21	6															
Cys 1	3	3			2														
Ser 45	77	4	3	2	2		12												
Thr 5	59	19	5	13	3		1	70											
Asn 16	11	1	4	4					43	17									
Gln 3	9	3	8	1	2				5	4	5								
Asp 16	15	2		1					10	6	53		8						
Glu 11	27	4	2	4	1				9	3	9	42	83						
Lys 6	6	2	4	4	9				17	20	32	15		10					
Arg 1	3	2	2	3	2		1		14	2	2	12	9		48				
His 1	2	3	4				1		3	1	23	24	4	2	2	10			
Phe 2	2	1	17	9	2				4	1	1					1	2		
Tyr	2	2	2	1			3		2	2	4			1	1		4	26	
Trp			1						2					3		1	1		
Pro 5	35	5	4	1			1		27	7	3	9	1	4	4	7	5	1	

Table 2.9: Amino acids substitutions. Frequency with which an amino acid can be substituted by others in sequences of the same protein from different organisms.

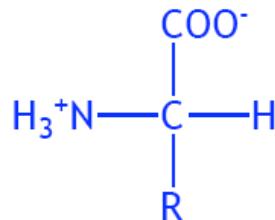


Figure 2.18: The four constituents around the C_α are shown.

There are four atoms linked to the C_α :

- The carboxyl group
- The amide group
- The hydrogen atom
- The R side chain

Within the cell the amino acids are mainly in their dipolar form or **zwitterion**. In this dipolar form, even if the carboxyl group is dissociated ($-\text{COO}^-$) and the amide group is protonated ($-\text{N}^+\text{H}_3$), the whole charge of the molecule is neutral, thus the zwitterion can act as an acid(H^+ donor) or a base(H^+ acceptor).

The properties of the peptide bond have important consequences on the stability and flexibility of polypeptide chains in water. The pairs of electrons being shared between the oxygen of the

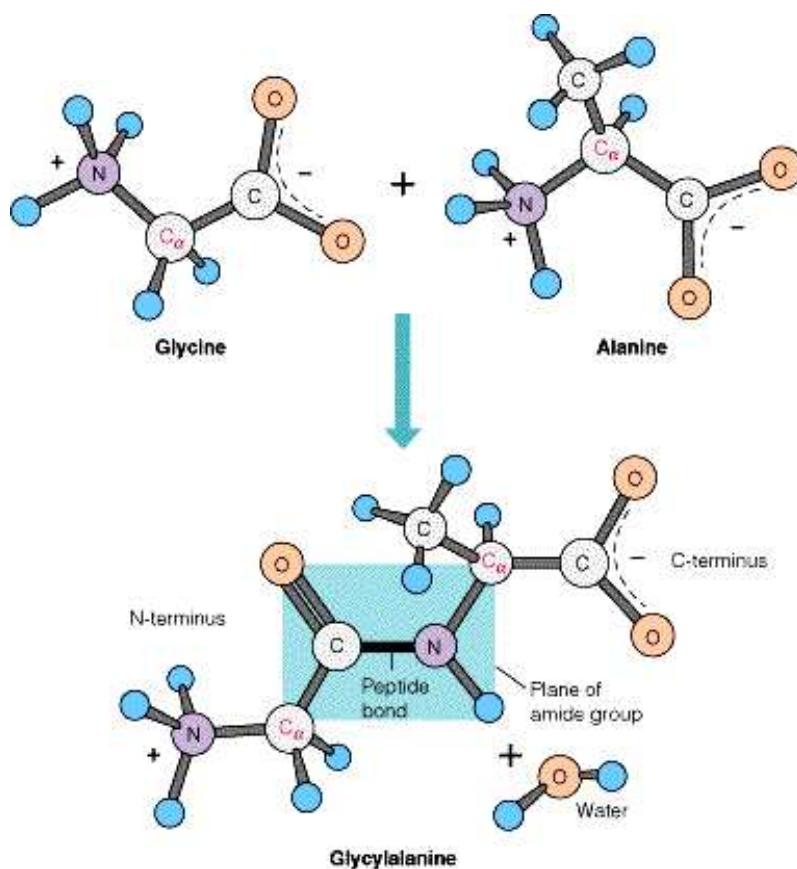


Figure 2.19: Formation of a dipeptide linking alanine and glycine. Summarizing, the backbone of every protein is constituted by blocks of $N-C_{\alpha}-C$ repetitions that are linked one to another through peptide bonds.

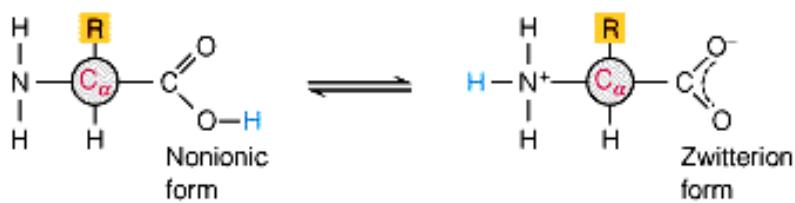


Figure 2.20: Zwitterion form. Amino acid forming zwitterions at neutral pH. Groups NH_3^+ and COO^- are ionized.

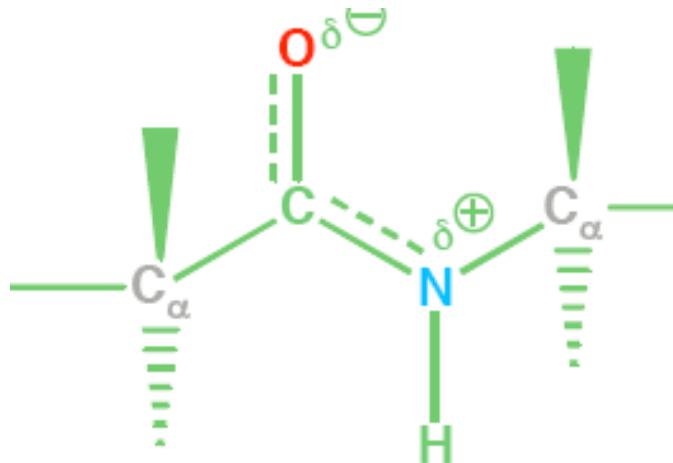


Figure 2.21: Planar nature of the double-bond character.

carboxyl group and the nitrogen of the amide are de-localized leading to the so called resonance. This is the responsible of the rigid character of the peptide bond and therefore of the movement and structure restrictions the polypeptide backbone can adopt. The six atoms of the peptide bond are placed in the same plane because of its partial double-bond character. The resonance is the main responsible of two transcendental effects:

1. Increasing the polarity of the peptide bond by maintaining longer the dipole moment ($\mu = qx$)
2. The three non hydrogen atoms that perform the bond (the carbonyl oxygen O, the carbonyl carbon C and the amide nitrogen N) are coplanar and there is : **no free rotation** about the bonds. The other two bonds the N-C_α and C_α-C, are single bonds and free rotation is permitted.

We can conclude that proteins are polymers with altered rotatable covalent bonds and planar-rigid ones and therefore the possible folds the polypeptide chain can adopt are restricted making likely to determine or infer which kind of fold a chain will achieve if the constituent amino acids are known.

Due to this partial double-bond, there are two possible conformations the substituents around C_α can adopt named cis- and trans-. In the trans- conformation both C_α are placed in opposite corners of the planar square formed by the peptide bond while in the cis- conformation are in the same side of the peptide bond and hence are located closer one to another.

2.2.1.2 Torsion angles Phi (Φ) and Psi (Ψ)

The angle of the N-C_α to the next adjacent bond is called phi torsion angle and the angle of the C_α-C to the adjacent peptide bond is called psi torsion bond. The phi angle is normally close to values of 180 (trans-conformation) but can have 0 (cis-conformation) mainly due to the planarity of the

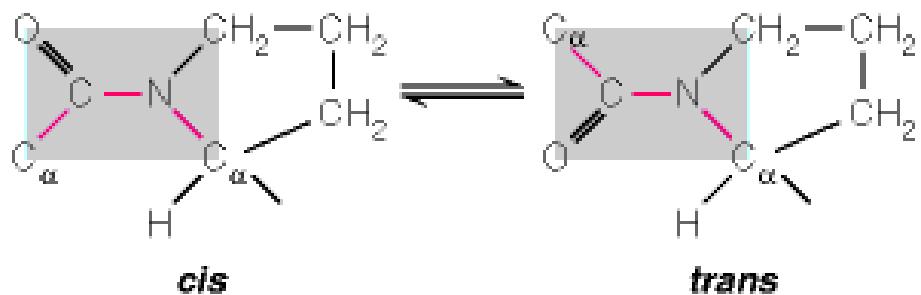


Figure 2.22: Cis- and trans-conformation of the amino acid proline. Unlike the rest of amino acids, in which the formation of the trans-configuration is the most likely one, in praline, the possibility to form the functional wrong isomer is higher.

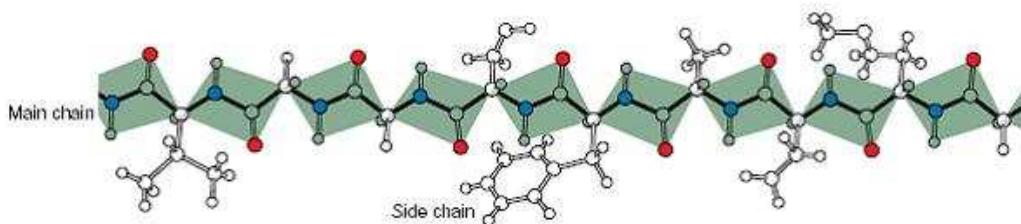


Figure 2.23: Polypeptide chain in which the coplanarity of the atoms around the peptide bond is shown as the shadowed light-blue squares.

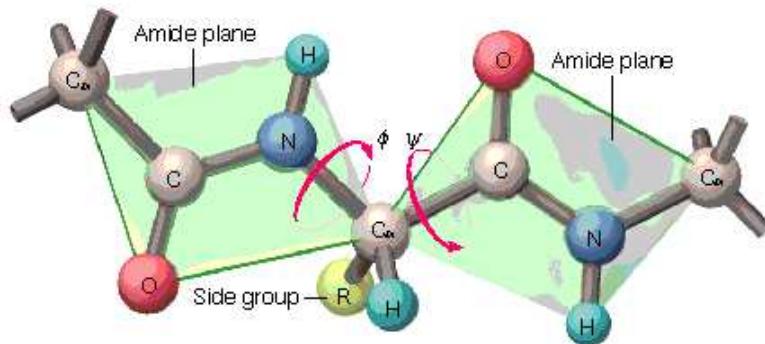


Figure 2.24: Torsion angles. Rotation is allowed only for the torsion angles phi and psi. The positive rotation is clockwise. Here the extended conformation corresponds to +180 for both angles.

peptide bond. The distance between the C_α atoms in the trans- and cis isomers is approximately 3.8 and 2.8 . The cis isomer is mainly observed in X-Pro peptide bonds due to its limited flexible geometry. These two torsion angles are included within the backbone dihedral angles of proteins (the angle between two planes is called their dihedral angle). Besides phi ϕ and psi ψ , proteins also include the omega ω that involves the backbone atoms C_α -C-N-C $_\alpha$. Thus, Φ controls the C_α -C distance, Ψ controls the N-N distance and Ω controls the C_α -C $_\alpha$ distance.

The side chain dihedral angles tend to cluster near 180, 60 and -60 (trans-, gauche + and gauche - conformations). The choice of side chain dihedral angles is affected by the neighboring backbone and side chain dihedrals.

2.2.1.3 Ramachandran plot

To know how secondary structure elements are arranged could provide us with a suitable way to classify types of fold and hence to predict in a forward step the potential biological function of a protein. As explained above, for each amino acid there are only two dihedral angles that could rotate: the phi and psi. The allowed values these torsion angles are able to cover can be plotted on the so-called Ramachandran diagram. On this graph the conformation of every individual residue in the protein with boundaries limiting regions of favorable conformation can be observed. The diagram was developed by Gopalasamudram Narayana Ramachandran and represents a way to visualize dihedral angles F against of amino acid residues in protein structure. It shows the possible conformations of F and angles for a polypeptide. It used as diagnosis method to accurate the prediction of protein structures conformations such a way that when experimentally determined protein torsion angles are determined and the values are plotted, these observed values should fall principally in this allowed region. Two large regions of phi and psi space are permitted by steric constrains:

- Regions including torsion angles values for the right-handed α -helix
- Regions including torsion angles values for the or platted sheet

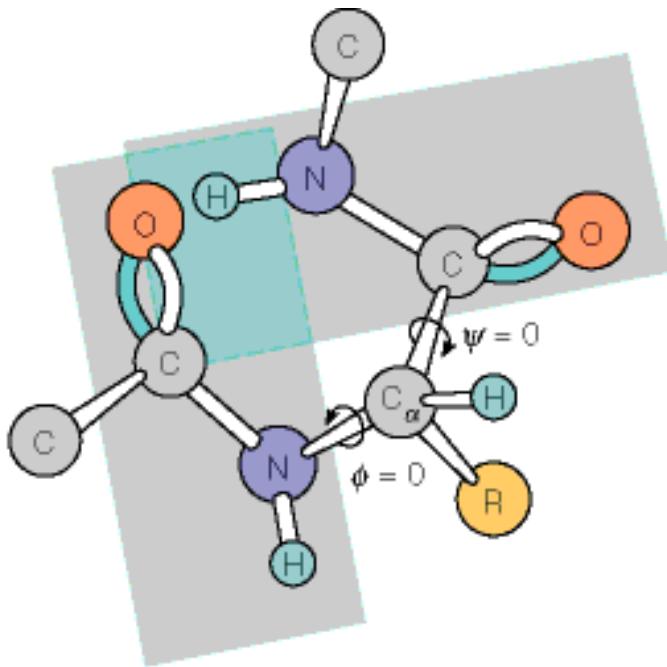


Figure 2.25: A non-allowed conformation is shown. Angles phi and psi can not cover the values both 0 due to the clash restrictions between the carbonyl oxygen and the amino proton.

In order to understand what meant the values adopted by both angles we define positive rotation as the one follows clockwise when looking either direction from the C_α (from left to right), negative rotation (opposite to clockwise) and the zero-angle conformation for each one. We must keep in mind that the fully extended form of the polypeptide chain corresponds to a value of 180 for each torsion angle. Taking the mentioned information into account, different combinations of the backbone are detected: polyproline helix and antiparallel β -sheet (left up side of the diagram with Φ values between $-180 - 90$, and $\Psi > 90$ values) as allowed values because they do not result as sterical interference Allowed folds if some relaxation of steric hindrance or obstruction is permitted (Φ value from -180 to -90 and 4Ψ values from 0 to 90) Isolated left-handed a helix which is rarely observed I short segments of the protein ($0 < \Phi < 90$ and $0 < \Psi < 90$) represented on the right up-center side of the diagram.

When proline is represented, the plot shows only a very limited number of possible combinations of ϕ and ψ due to its ring R that includes the NH group of the peptide bond. The praline amino acid can be considered as an indicator to find turns and loops.

Glycine has a hydrogen atom, with a smaller van der Waals radius, instead of a methyl group at the R position, therefore is less restricted and this is apparent in the Ramachandran plot for Glycine for which the allowable area is considerably superior and the possible phi and psi combinations are larger. Alanine has a metal group linked to the C at the R position so it has more restriction

We can summarize: That every backbone conformation of any particular residue in any protein could be described by specifying those two angles In similar secondary structure types all residues would be drawn as super-imposable points because are in equivalent conformation and hence have corresponding Phi and Psi angles The allowed conformations of a polypeptide chain depend on

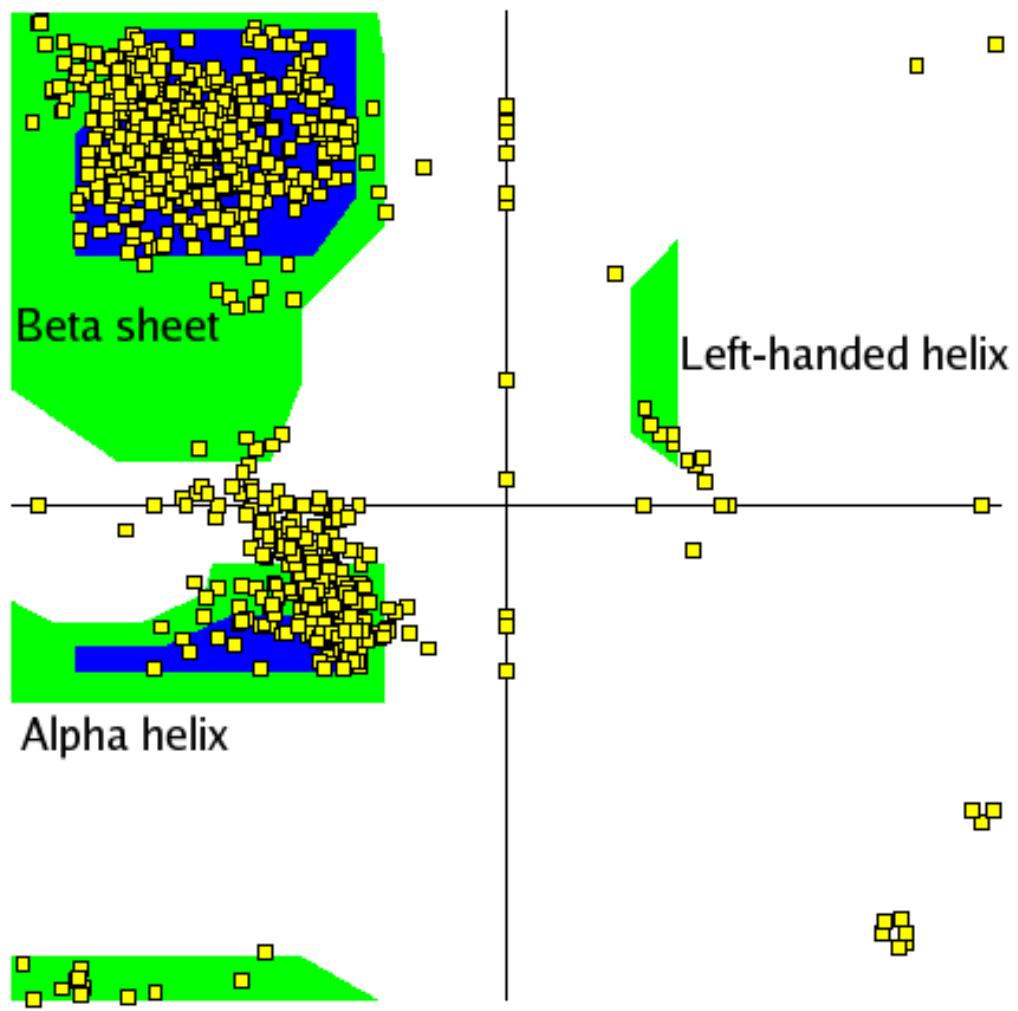


Figure 2.26: Ramachandran plot. The plot is divided into four identical squares by an axis with values equal to 0 for both Phi and Psi angles.

Conformation	Phi (N-C _{α})(Φ)	Psi (C _{α} -C)(Ψ)
Right-handed α -helix	-57	-47
Left-handed α -helix	+57	+47
3 ₁₀ helix	-49	-26
Antiparallel β -sheet	-139	+135
Parallel β -sheet	-119	+113
Turn II (second residue)	-60	+120
Turn II (third residue)	+90 0	0
Extended chain	-180	-180

Table 2.10: Phi and Psi ideal angles values for some secondary structures.

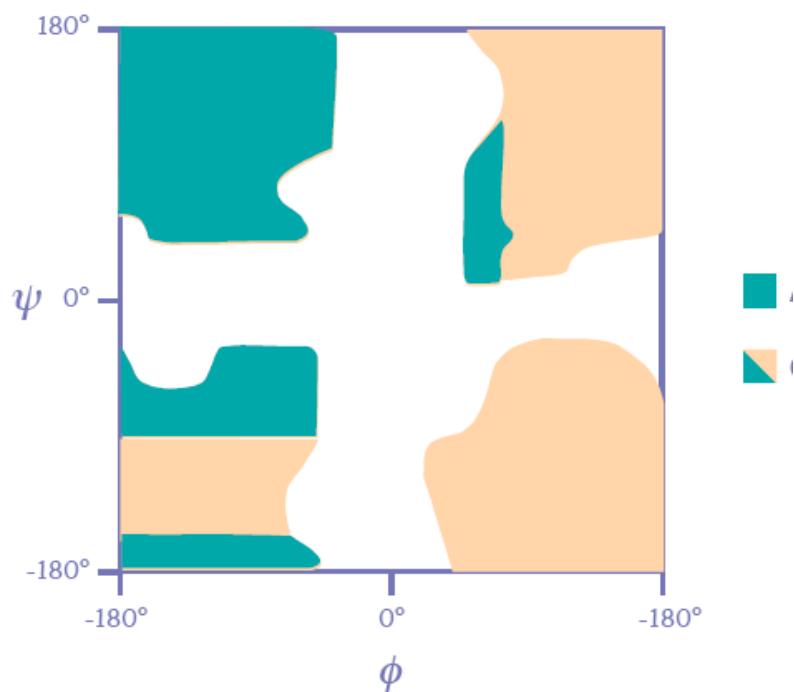


Figure 2.27: Ramachandran representation for Alanine and Glycine.

the bulkiness of the side chains and consequently on the amino acids residue constitution.

2.2.2 Interactions and folding

Amino acids side chains have different tendency to participate in interactions between each other and with water, this difference influence the stability, function and state folding of the protein they belong to.

- **Hydrophobic** amino acids are the ones exploiting Van der Waals interactions as they avoid contact with water packing against each other and being the basis of the Hydrophobic effect (explained in detail above). Alanine and leucine are the ones found in helix as their nitrogen backbone is available too hydrogen bonding as required for helix formation. By contrast praline has no nitrogen to hydrogen bond so rarely forms part of a helix. Weakly polar interactions can be performed by aromatic ring of phenylalanine.
- **Hydrophilic** amino acids are the ones in which hydrogen bonds interactions are maximal working, thus these residues hydrogen bonding to water, to one another, to the peptide backbone and to polar organic molecules. For aspartic and glutamic acid as their charges state changes depending on the pH of the micro-environment, they can function as proton donors when placed in the hydrophobic interior of a protein at physiological pH or when a negative charge is placed nearby (the pKa shifts from 5 in aqueous solution to 7) or as acceptors in aqueous solution (unprotonated and negatively charged). For positive charged amino acids as lysine, the behavior is almost the same one, being proton donor in aqueous solution and proton acceptor when functions as neutral specie in a non polar environment or in presence of a neighboring positive charge (pKa shifts from 10 in aqueous solution to 6). The NH₂₊ group of the arginine is always positively charged at neutral pH being stabilized by resonance. Histidine can function as double proton donor when both NH groups (titratable N-H groups with a pKa= 6 for each one) are protonated and hence the whole charge is positive but it can also function as donor-acceptor at the same time when one of the NH group loses one proton increasing the pKa of the other one until a value of 10. The fully deprotonate state is negatively charged but occurs rarely. This versatility makes the histidine to be the most common amino acid found in the active site of the enzymes. Serine, threonine, asparagines and glutamine do not ionize but are able both to accept and to donate hydrogen bonds simultaneously, although the amide N for asn and gln is not charged at neutral pH but is polar. Cysteine is also found mainly in the active site of enzymes because its thiolate anion is the most powerful nucleophile available
- **Amphiphatic** amino acids in which both polar and non polar character take place making them to form interfaces. Hydrophobic regions of amino acids positively charged as lysine can interact through van der Waals interactions with other hydrophobic side chains. The OH group of tyrosine, serine and threonine is able both to donate and to accept hydrogen bonds and the aromatic ring of the tryptophan can form weakly polar interactions. The least polar of the amphiphatic is the methionine even thought the thioether sulfur is an exceptional ligand for many metal ions.

Weak acid, the pKa is a measure of the tendency of the acid to dissociate (give of an H⁺ ion)
Key rule:

- $\text{pH} = \text{pK}_a$: protonated and unprotonated forms are at equilibrium
- $\text{pH} < \text{pK}_a$: positively charged, more protonated and proton donor
- $\text{pH} > \text{pK}_a$: negatively charged, less protonated and proton acceptor

2.2.2.1 Bonds

It is a hard problem to figure out the structure of a protein from its 1D sequence. One reason is because of space: If we assume that each amino acid can adopt one of the three conformations (alpha, beta, coil) then the chain of 100 amino acids has $3^{100} = 5 \times 10^{47}$ possible folds. Another reason is because of time: A fold takes 10 -13 seconds, so it would take 10 10 years (universe is 1010 years old) The last reason could be because of correct fold: Interactions between thousands of atoms one with each other, surrounding water, and surrounding molecules are a direct condition to the way the proteins finally folds. The proteins fold in order time of seconds, thus the nature knows the correct criterion in order not to lose time trying different conformations.

Due to the chemical-physical properties of amino acids in aqueous solution, the chain being translated folds up to reach the native state of an active protein. This state is characterized for its closely packed interior core, in which secondary structure elements are made up by non polar and non charged amino acids while the polar and charged are located in the outer core of the protein where hydrogen bonds with water can occur. The amino acid sequence chain can fold spontaneously in a quite manner. However, during the protein synthesis, special proteins called chaperones (or housekeepers) assist to fold the polypeptides chains in the proper way avoiding mistakes to miss folding. In both cases, before to reach the final folded configuration, partially folded intermediate states exist and secondary structures elements can be recognized but are not stable enough to allow their isolation and posterior study.

The particular chemical-physical properties of the peptide bond and the nature of each one of the 20 proteinogenics amino acids are crucial in determining the native state defined above. The chemical interactions that stabilize the polypeptides are summarized in the table above depicted. Disulfide bonds and the amide bonds in the backbone are the only covalent interactions that embrace the polypeptide chain. Even though their high contribution to the general enthalpy, the non covalent polar weak interactions as hydrogen- Van der Waals bonds are the ones that contribute in larger manner to the stabilization of the final fold state of the proteins because they can sum up to substantial energetic contribution. Van der Waals interactions depend directly of the change in charge induction the electron clouds fluctuations of one atom or groups of atoms generate on a neighbor atom. This effect is greater with those most polarizable groups and decreases with the distance thus, carbohydrate or methyl groups of hydrophobic sides chain as leucine and valine and 5A or less distance, are favorable settings to their arrangement.

The hydrogen bonds are included within the polar weak interactions though their strength increases with the number of interactions stabilized and depends mightily on the environment. When both donor and acceptor are fully charged the hydrogen bond is called salt bridge and its bonding energy is higher since both atoms contribute (in hydrogen bonds just one fully charged participant contribute)

Hydrogen bonds in water are the most important interactions since the same water molecule can act both as donor and acceptor being also the reason the water to be liquid at ordinary temperatures. Within the polypeptide nearly all potential donors or acceptors are participating in

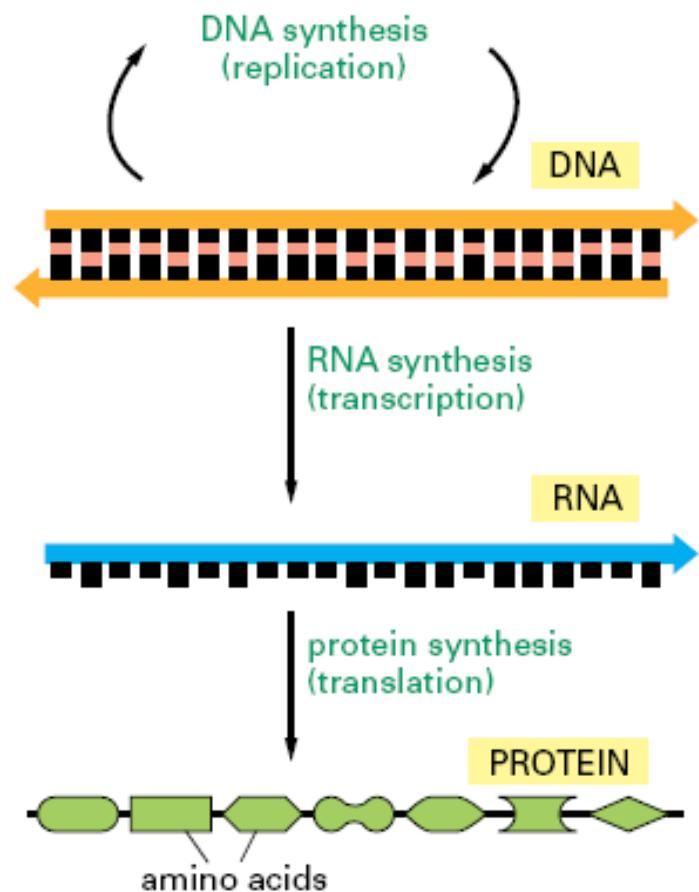


Figure 2.28: Central dogma represented. From a DNA sequence to a polypeptide chain using mRNA as a template.

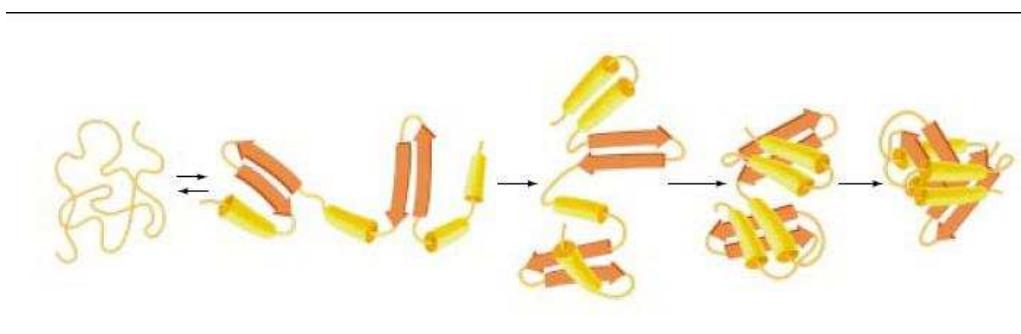


Figure 2.29: Folding pathway.

INTERACTION	EXAMPLE	DISTANCE DEPENDENCE	TYPICAL DISTANCE(Å)	FREE ENERGY (kJ/mol) (bond dissociation enthalpies for the covalent bonds)
Covalent bond	Cα-Cα	-	1.5	356
Disulfide Bond	-Cys-S-S-cys	-	2.2	167
Hydrogen bond	-NH-O=C-	Donor(N) and acceptor(O)	3.0	2-6 in water and 12.5-21 if either donor and acceptor is charged
Van der Waals	-CH ₃ -CH ₃	Short range and falls rapidly beyond 4 Å separation	3.5	4 (4-7 in protein interior) depending on the size of the group

Table 2.11: Some of the chemical interactions that stabilize polypeptide chains. Adapted from Disulfide bonds are not favored in the interior cells environment since their characteristic is reduced state which makes free SH groups favorable over the S-S.

such reactions because no to do it is energetically unfavorable. As demonstrate in the thermodynamic section not to make hydrogen bonds would leave one or more uncompensated partial or full charges.

Hydrogen bonds and hydrophobic interactions plus the aforementioned van der Waals interactions are the three mainly non covalent interactions responsible of the final configuration and active function of the proteins. Amino acids with non polar R sides are restricted to the inner core of the proteins clustering together due to the so called **hydrophobic effect**. This reaction is the one causing the polypeptide to become compact. This repulsion (or attraction in case of hydrophilic amino acids) to water and the possible hydrogen bonds make possible the protein secondary structure elements to compact and hence to give the globular shape of the final active state

We have to take into account the reduce environment of the inner cells when relating to disulfide bonds. The inner cell space, the cytosol is reduced, proteins with disulfide bonds are not found since these bonds can not occur and cysteine SH groups can not link. Outside the cell or within the plasma membrane, the redox state is ideal for the oxidation of the cysteine residues leading to the formation of afore mentioned bonds. During protein synthesis special proteins as the Protein disulfide isomerase (PDI) catalyzes the correct formation for such bonds.

The steps followed during the folding process are:

SSEs + Burial of hydrophobic aromatic groups + hydrogen bonds

2.2.2.2 Thermodynamics

Once we want to understand the full and fold pathway, all states of the process must be explained both energetically and structurally. Free Gibbs energy, enthalpy and entropy are critical terms to take into account in folding pathways since the contributions of the forces to protein stability are

quantified in terms of the energy associated to one of them. Enthalpy is taken as the heat released once the bond is formed through one of the interaction within an isolated system. Entropy is a way to measure the randomness or disorder of a system and, when regarding to folding process is mainly due to water and hydrophobic effect. Free Gibbs energy is a combined effect of entropy and enthalpy. Therefore its value, but more important its sign depends directly on the values and sign both system disorder and heat released.

When trying to find explanation to the change state from unfolded to folded one, terms of thermodynamics, specially **free energy net loss**, is thought to be the driven force to explain the higher stability when the hydrogen bonds have been formed in α -helix and β -sheets. However these bonds suppose a small change in free energy besides the fact that the unfolded states can also hydrogen bond with water turning it as the no driven force for folding processes. The hydrophobic effect by contrast can explain the spontaneous fold of the polypeptide chains in terms of increasing the system entropy (protein) and hence varying the free energy in terms of spontaneously. Spontaneous process following the second law of thermodynamics, are based in an increasing of total entropy of a system plus its surroundings. Such an increment of the system disorder happens once the buried amino acids clump together expelling water to the exterior even if their aromatic side chains are initially surrounded by ordered water molecules (decreasing in entropy). The hydrophobic amino acids hidden from water early in the folding process have the consequence of reducing the number of possible conformations to look for and of avoiding the possible hydrogen bonds their -NH and CO groups in the backbone could achieve. As this state is energetically unfavorable and the only way, once their chain side is completely buried in the internal hydrophobic core, is hydrogen bonding one to another, it results in the formation of a helix and b sheets secondary structure elements. Therefore we can conclude that the SSEs formation as well as the gain in solvent entropy is both consequences of the burying hydrophobic side chains. Stability (ΔG): net loss of free energy such the difference in free energy between the folded and the unfolded state is < 0 Decreasing in $\Delta H < 0$ due to bonds formation $\Delta G = \Delta H - T\Delta S$ Increasing Gain of $\Delta S > 0$ due to hydrophobic effect The folded state of a protein is a thermodynamic compromise since the free net energy of stabilization is rather small, even when hundreds of interactions as hydrogen bonds and van der Waals occur. The energy released once the protein is folded and such interactions are formed, is just equilibrated with the loss of entropy (and hence of conformational flexibility). As not all hydrophobic amino acids within a polypeptide chain are buried and clustered inside the internal core but scattered alongside the sequence, other reasons to explain the early and characteristic fold must be addressed. During the folding process the protein proceeds from high unfolded energy level to a low native state through meta-stable intermediate states with local low energy minima separated by unstable transition high energy state. There are many different experimental techniques that do try to characterize such states (NMR, Hydrogen exchange, etc). A recent technique implies molecular engineering and genetic punctual mutations in order to understand which energetic changes take place during unfold-fold course. For instance by mutation of Ala to Gly amino acid residue in the solvent-exposed side of an α helix would destabilize both intermediate and native state: When both are destabilized this single-site mutation is being formed in the early state of the folded protein, thus is the helix is already fully formed in the intermediate state. Instead if just the native form is affected the helix is not formed until the transition state is accomplished.

Proteins are temperature-sensitive as they are made up by hundreds of weak interactions. The native conformation can be disrupted leading to the denatured state which is characterized by the unfolded state. The denaturalization nature of the unfolded state can be attained by chemical

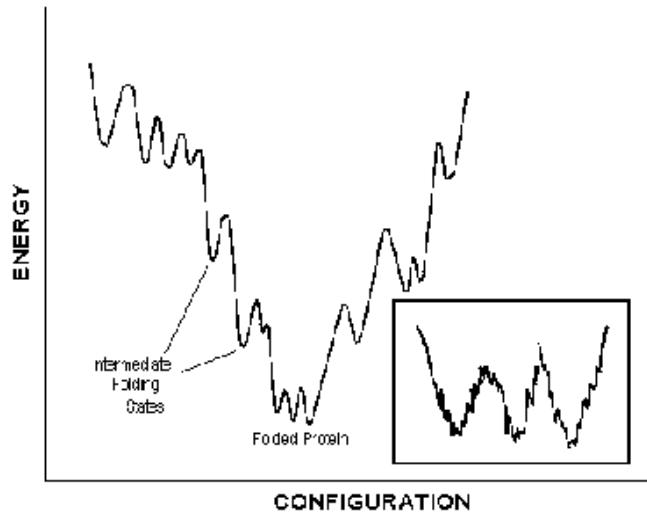


Figure 2.30: Energy folding profile. The types of folds are made by searching the global minimum so the potential energy function is minimized.

substances as SDS detergent or guanidine hydrochloride. The difference between the temperature factor as a denaturalization factor is the competition to hydrogen bonding with the polar groups of the backbone and side chains.

2.2.3 Secondary Structure Elements

Polypeptide segments due to the physical-chemical properties, bond lengths and types as well as torsion restrictions, fold into a higher level structure from the linear sequence of amino acids to the native fold configuration which is characterized for phi and psi values and non covalent interactions, mainly hydrogen bonds between the peptide NH and CO groups of different residues. One consequence of the hydrophobic effect described above is that the formation of hydrogen bonds from the amide and carbonyl groups of the peptide backbone to the water is not possible anymore. They are hidden in the inner core so to satisfy their hydrogen-bonding potential, most of them interact with their selves leaving the secondary structure elements to form as a way to gain in free available energy against the increase of environment entropy.

Unfolded chain: side chains interact with water and hydrophobic groups faced the interior side
 Compact structure: buried hydrophobic chains interact with each other, polar backbone hydrogen bonding with each other and hydrophilic polar side chains on the surface interacting with water

The formation of secondary structure is driven by the burial of hydrophobic side chain residue when they associate to each other and exclude water. Small sequences which show semi-stable helices in water could work as nucleation point over which the reminder amino acids forming the whole protein are going to be placed. Additional levels of classification can be added when including Super-secondary structure (recurrent patterns of interaction between helices and sheets close together in the sequence), domains and modular proteins, when the proteins show compact units within the folding pattern of a single chain or when many copies of close related domains

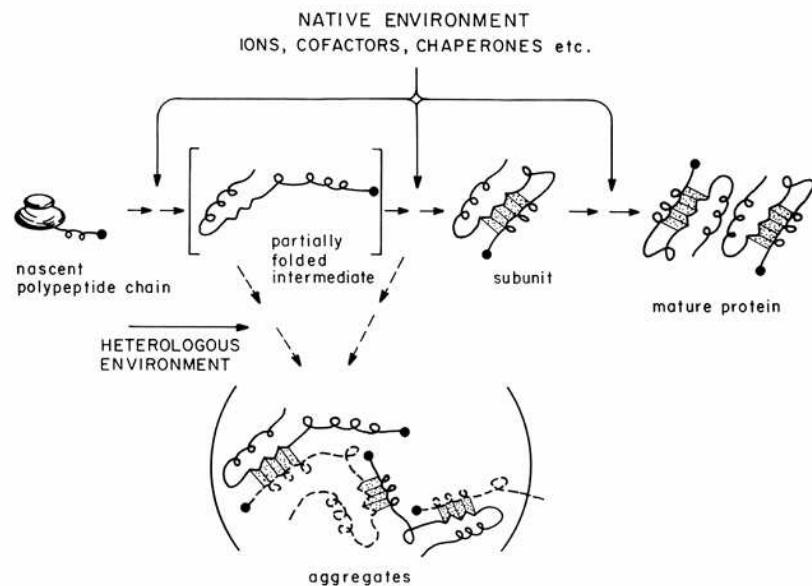


Figure 2.31: Folding factors.

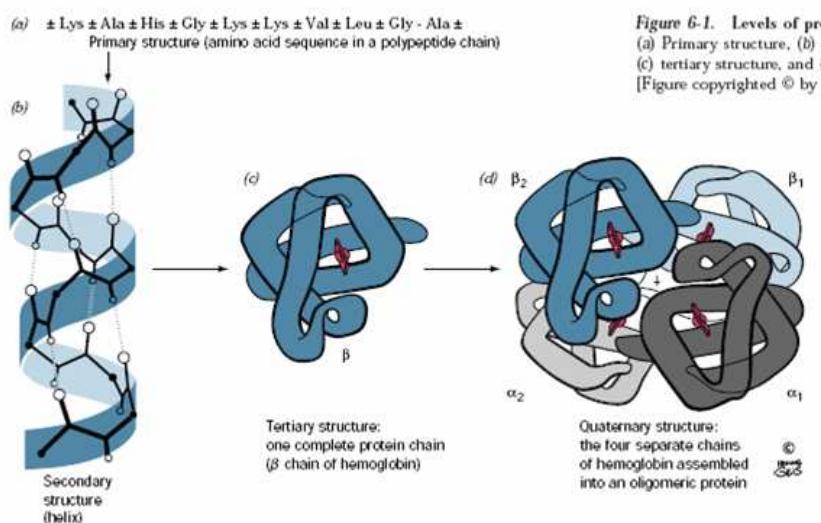


Figure 2.32: Structure hierarchy. Secondary structure as the assignment of helices and sheets through the hydrogen-bonding pattern of the main chain ; Tertiary structure as the assembly and interactions of the helices and sheets; Quaternary structure as the assembly of monomers when the protein is composed by more than one subunit.

CONFORMATION	PHI (°)	PSI (°)	RESIDUES PER TURN	TRANSLATION PER RESIDUE (distance from two consecutive residues) (Å)
Alpha helix	-57	-47	3.6	1.5
3-10 helix	-49	-26	3.0	2.0
Pi-helix	57	-70	4.4	1.15
Polyproline I	-83	+158	3.33	1.9
Polyproline II	-78	+149	3.0	3.12
Polyproline III	-80	+150	3.0	3.1

Table 2.12: Parameters of the most commonly found helical SSEs. The different values define different helical geometries.

are setting up a multi-domain within the same protein respectively.

2.2.3.1 Types

Steric limitations due to physical size of atoms and possible bonds allowed in the backbone of a polypeptide chain, limit the possible types of secondary structures. Individual secondary structure elements are rarely associated with a specific function.

2.2.3.1.1 α -helix α -helices are regular cylindrical structures and one of the most common secondary structures in proteins. They are generated by hydrogen bonding between the CO group of one residue n and the NH group of the n+4 residue being all close together. All the carboxyl and amide groups are hydrogen bonding except the ones corresponding to the carboxy-terminal end and amide-terminal end. It can be also defined as a versatile cylindrical structures stabilized by a network of backbone hydrogen bonds. This backbone forms the wall of the cylinder being the outside studded with side chains.

As defined in the 2.12, there are 3.6 residues per turning a common α -helix which corresponds with a rotation of 100 so that side chains projects out from the helical axis at 100 intervals. The structural meaning of this periodicity is that residues 3-4 amino acid apart in the linear sequence axis will project from the same face making possible the alpha helices to be amphipathic with one polar hydrophilic side and one non polar hydrophobic side where similar chemical-physical characterized amino acids are placed. This feature stabilize the helix-helix packing.

It is important when predicting structure to take into account the position of such amphipathic helices normally occurred on the surfaces of proteins where polar residues are in touch with water or also placed on interfaces where polar residues interact with one another. For short helices the main profile is the one explained above however for longer length helices it would coil about the helix axis such a way that if two long helices have a pattern of hydrophobic groups four residues apart , they would interact by forming a coiled coil.

2.2.3.1.2 β -sheet Another common secondary structure. In contrast to the α -helix, it is formed by hydrogen bonds between backbone atoms on adjacent regions of the peptide backbone, called

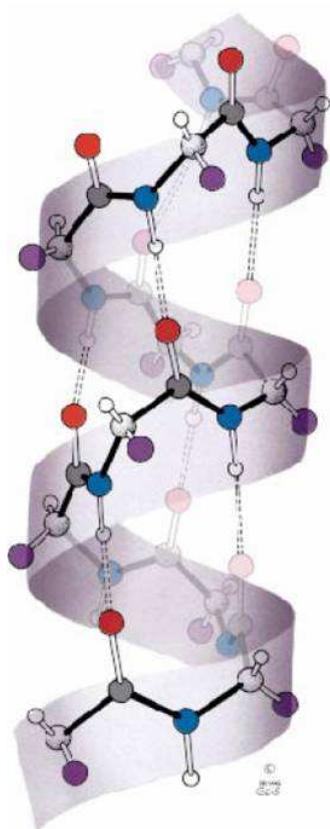


Figure 2.33: The alpha helix. Backbone of three turns is shown. Note the non hydrogen bonding ends amide and carboxy terminals up down the helix. Those groups form a permanent dipole with its positive charge at the amino terminal end and the negative charge at the carboxy terminal end. Usually a polar side chain is found at the end of the helix bridging the hydrogen bonds to these lacks donors and acceptors. Red: oxygen; White: Hydrogen; Blue: Nitrogen; Black: carbon.

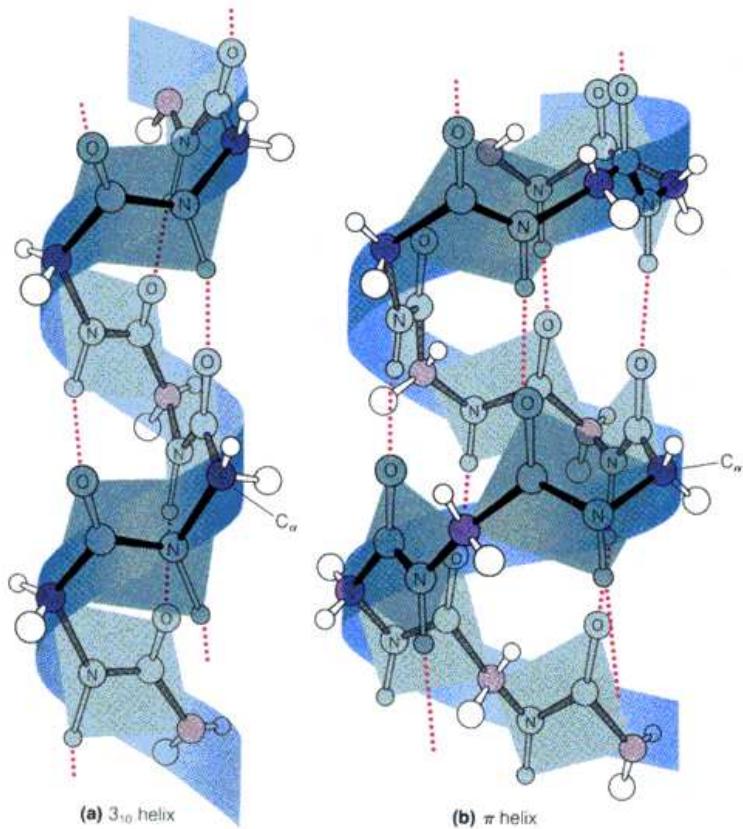


Figure 2.34: 3_{10} helix and α -helix. Other types of secondary structure involving helices. The 3_{10} has 3 residues per turn and a 10 member hydrogen bonded loop. The helix is a theoretical protein secondary structure with 4.4 residues per turn and a 16 atom hydrogen bonded ring that is sterically possible but has not yet observed.

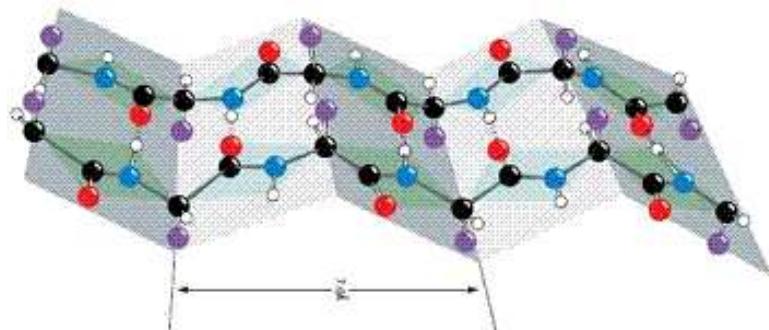


Figure 2.35: β -sheet bonds. The distance between two consecutive residues is 3.3 and the phi and psi torsion angles are -130 and +125 respectively. As in the case of a helix the right-handed twist direction is favored against the left-handed due to steric constraints of L-amino acids introduced by chirality in the C_α .

β -strands. Thus involving hydrogen bonds between backbone groups from residues distant from each other in the linear sequence making possible that two or more strands that might be widely separated in the protein sequence are arranged side by side leaving hydrogen bonds between the strands. These interactions do not involve side chains. Thus, many different sequences can form a β -sheet.

A β -sheet is a regular and rigid structure often represented as a series of flattened arrows. Each arrow points towards the protein's C-terminus side to have distinct properties from the other. β -sheets are usually twisted and not completely flat. Almost all polar amide groups are hydrogen bonded to one another except for the NH and CO groups on the outer side of the edge strands. Possible hydrogen bonds could be done with water when exposed to the solvent, with packing against polar side chains (a neighbor a helix), by interacting to an edge strand in another protein chain, etc, increasing the beta structure. One important way to hydrogen bonds is via the formation of beta barrels in which the last strand of the edge interact with the first one closing a cylinder, thus curving around itself. Such structures are the ones stabilizing quaternary structure. The polypeptide chain is almost fully extended and amino acids side chains as valine and isoleucine (aliphatic amino acids) are more easily accommodated within a β sheet than in a α -helix because the β structure are not that tightly close together. This feature performs better the search for possible structures prediction since the couple of amino acids are most frequently found in sheet than other residues. There are two types depending on the orientation the strands run: parallel β sheets when the strands run in the same direction and antiparallel β sheets when they run to one another in opposite directions. Also mixed β sheets have been observed. Parallel β sheet are placed internally buried while antiparallel are on one face in direct contact with aqueous solution which makes them to be more stable (their hydrogen bonds are more linear). Antiparallel β strands are connected mostly via turns reversing the direction of the strands while parallel strands connect via more complex unions that might include segments of a helix. In such cases the molecule built up is stronger i.e. silk. The ways in which the parallel β strands are linked force them to be discontinuous.

Because nearly all peptide are trans-, with C=O and N-H groups point in opposite directions, as we walk along the side, chains also point in opposite directions making possible the hydrophilic

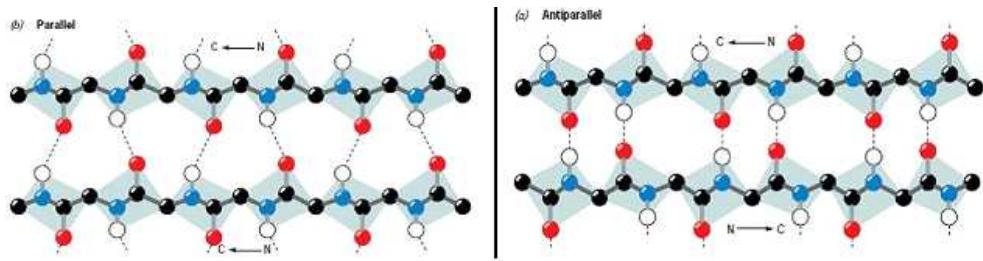


Figure 2.36: Parallel and antiparallel β sheet. Hydrogen bonds are more linear in the antiparallel sheet. Due to the corrugated appearance, these SSEs are also called pleated sheet. Note the correspondence of carbons and nitrogens in both strands in the parallel sheet and the inverse way in which the pairs C-N correspond one to another in the antiparallel sheet. Parallel strands are less twisted than antiparallel ones Red: oxygen; White: Hydrogen; Blue: Nitrogen; Black: carbon.

	Turns	A helix	B sheet parallel	B sheet antiparallel
Located	Surface of proteins Direct contact with aqueous environment	Intermembrane space Surface	Buried into internal core of proteins	Surface of proteins Direct contact with aqueous solution (on one face)

Table 2.13: Principal locations of the principal Secondary Structure Elements.

amino residues to be placed grouped in the same face and hydrophobic grouped in the other face and hence creating as in case of a helix an amphipathic beta strands structure. As a direct consequence these strands and sheets are found on the surface of proteins.

2.2.3.1.3 Turn and Loops Turns are also known as hairpin reverse turn or beta turn. It is considered as the simplest secondary structure element and the simple way to satisfy the hydrogen bonding capability of the peptide bond. It makes up as hydrogen bond between the carbonyl oxygen (-CO) of the residue n and the amide hydrogen (-NH) of the residue n+3 causing a reversion in the direction. Due to this capability those elements can limit the size of the molecule and maintain the compact state. This interaction can be made between residues n and n+2 but is not the common one. Because this kind of pattern is too tight, it can not continue alongside the chain. When the turn is not buried water molecules can donate and accept hydrogen avoiding the four residues that build up the turn to not interact with each other. This is the main reason the beta turns are placed in the surface of folding proteins in direct contact with the aqueous solution

Loops are tails of the polypeptide chains that connect regions of secondary structure involving hydrogen bonding and packing interactions with the rest of the structure.

2.2.3.1.4 Coiled coil In a typical coiled-coil two α -helices wrap around each other to form a stable structure. One side of each helix contains mostly aliphatic amino acids, such as leucines and valines, while the other side contains mostly polar residues. Helices containing distinct hydropho-

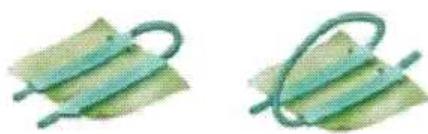


Figure 2.37: B turn. Amino acids reversing completely the direction of the polypeptide chain via hydrogen bonding between residues 1 and 4

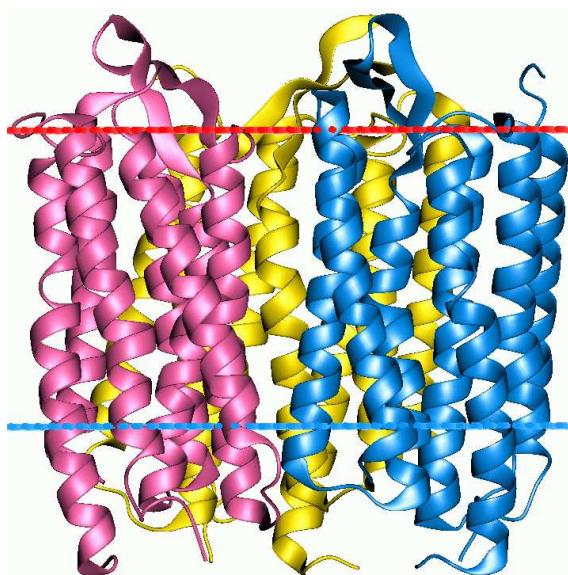


Figure 2.38: Loops. A typical protein contains approximately one third of its residues as loops that sometimes connect consecutive helices and strands that interact with each other to form supersecondary structure.

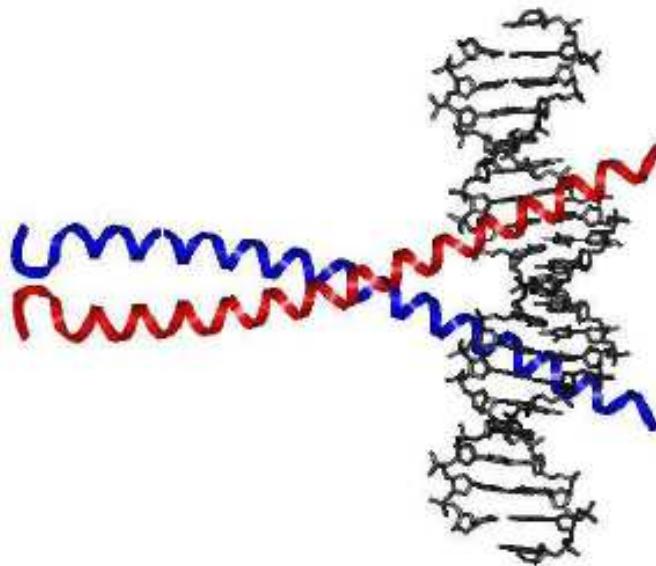


Figure 2.39: Coiled coil SSEs. Associated in parallel or antiparallel orientation and might be the same (homo-oligomer) or different (hetero-oligomer). Heptad repeat (abcdefg)_n spread out along two turns of the helix with positions a and d hydrophobic , e and g charged and b, c, f hydrophilic.

bic and polar sides are called amphipathic. In a coiled-coil, two amphipathic helices are aligned so their hydrophobic sides snuggle tightly together in the center, with their polar faces exposed to the solvent. A triple coiled-coil is another stable structure formed by α -helices. In this case, three amphipathic helices twist around a central axis. The hydrophobic sides of all three helices face the center of the coil, creating a stable hydrophobic core. Coiled-coils are often found in elongated, fibrous proteins. A triple coiled-coil is the major structural theme in fibrinogen, a protein involved in blood clotting. The fibrous nature of this protein is intimately related to its ability to form clots.

2.2.3.1.5 TIM barrels Fold characterized for a β -sheet strand followed by an α helix repeated eight times. When this kind of fold are found in a sequence it can suggest a catalytically function of the protein because all known TIM barrels to date are enzymes

In the TIM barrel structure, the α -helices and -strands form a solenoid that curves around to close on itself in a ring shape, topologically known as a toroid, thus a close curve that turns around an axis not contained within the ring. The parallel -strands form the inner wall of the ring thus, a -barrel, whereas the α -helices form the outer wall of the ring.

2.2.3.2 Motifs and Domains

A motif can be referring both to a particular amino-acid sequence characteristic of a specific biochemical function like the Zinc finger, and to a contiguous set of secondary structure elements having either a specific functional significance or defining an independently folded domain. There are five classes of domains differing from each other in the main secondary structure contained:

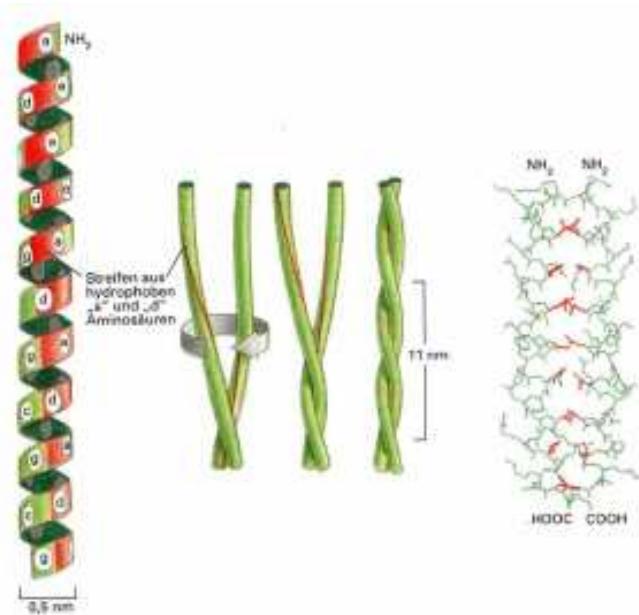


Figure 2.40: Peptide velcro hypothesis. The most favorable way for helices to arrange in an aqueous environment by wrapping around each other so hydrophobic surface is buried.

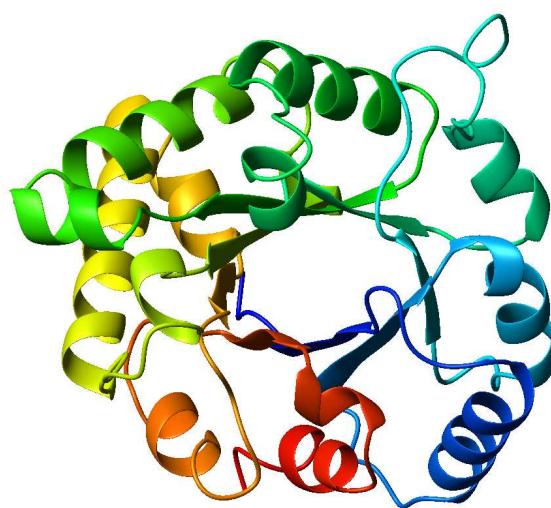


Figure 2.41: TIM barrel SSEs. The figure illustrates the enzyme triosephosphateisomerase which was the first protein discovered with this topology.

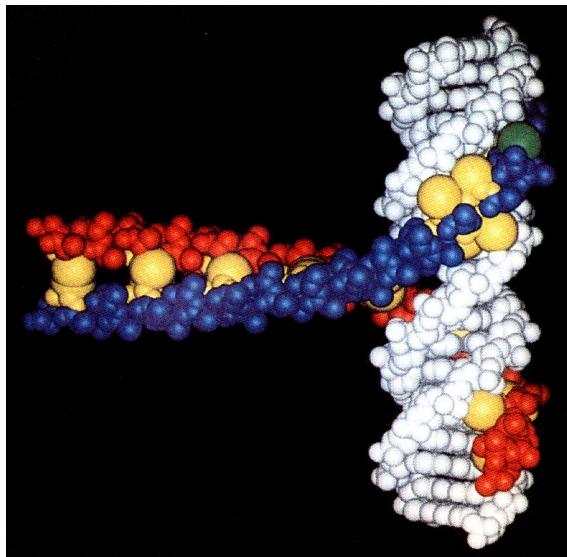


Figure 2.42: Leucine zipper. Peptide-peptide interactions in the coiled coil of the leucine zipper family of DNA- binding protein are shown. Extensions of the two leucine zipper helices straddle the DNA major groove. Side chains from both helices extend into the groove to contact DNA bases. The specific interactions between side chains and bases are hydrogen bonds.

alpha domains, comprise entirely of α -helices; beta domains contain only β -sheet; alpha/beta domains containing beta strands connecting α -helices; alpha+beta domains with separates α -helices and β -sheets regions; cross-linked domains almost no secondary structure but disulfide bonds or metal ions. Within every class many different arrangements of SSEs are possible and each one defines a structural motif.

2.2.3.2.1 Homeodomains Homeodomains are found in many transcription regulatory proteins and mediate their binding to DNA. A single homeodomain consists of three overlapping α -helices packed together by hydrophobic forces. Helix 2 and helix 3 comprise the DNA-binding element, a helix-turn-helix motif. Three side chains from the recognition helix form hydrogen bonds with bases in the DNA. In addition to the contacts between the recognition helix and the bases in the DNA major groove, an arginine residue from a flexible loop of the protein contacts bases in the minor groove

2.2.3.2.2 Leucine zipper A leucine zipper domain is composed of two long, intertwined helices. Hydrophobic side chains extend out from each helix into the space shared between them. Many of these hydrophobic side chains are leucines, giving this domain its name. A space-filling view reveals the tight packing of side chains between the leucine zipper helices; this makes the domain especially stable. Monomers are disorders in solution but fold on dimerization through hydrophobic coiled-coil interactions in their carboxy- terminal regions and on contact with DNA through their basic amino-terminal regions

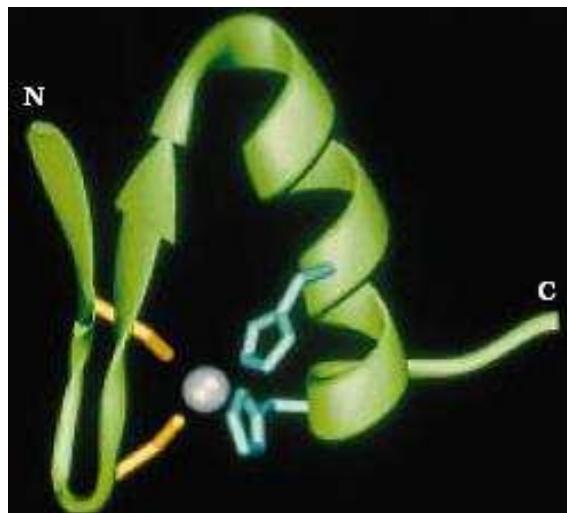


Figure 2.43: Zinc finger. Zinc Fingers are structural motifs found in many DNA-binding proteins.

2.2.3.2.3 Zinc finger Zinc finger domains are structural motifs used by a large class of DNA-binding proteins. The general formula is



They use centrally coordinated zinc atoms as crucial structural elements. A single zinc finger domain is only large enough to bind a few bases of DNA. As a result, zinc fingers are often found in tandem repeats as part of a larger DNA-binding region. The helical region of each zinc finger rests in the major groove of the DNA helix. Basic side chains project out from the helix and contact bases in the DNA. The identities of these side chains determine the precise DNA sequence recognized by each zinc finger.

Assembling different zinc finger motifs allows precise control over the sequence specificity of the protein. The specific contacts made between protein and DNA is hydrogen bonds

2.2.3.2.4 Transmembrane elements Found in proteins cross the entire membrane. Transmembrane proteins aggregate and precipitate in water. Those elements are thought to form very early in the folding process. The whole helical folding pathway of such structures is though to occur by condensation of preformed secondary structure elements. These elements make the integral membranes proteins to be unusually stable breaking down the high number of hydrogen bonds that maintain the structure only investing high levels of energy.

2.2.4 3D Structure

Tertiary structure results as the arrangement of SSEs into a stable and compact fold through weak interactions involving both polar and non polar groups. It is a hard task to determine the final shape a protein will have based just on the secondary structure elements since the same elements can come together in different ways depending on the sequence; i.e eight strands connected by

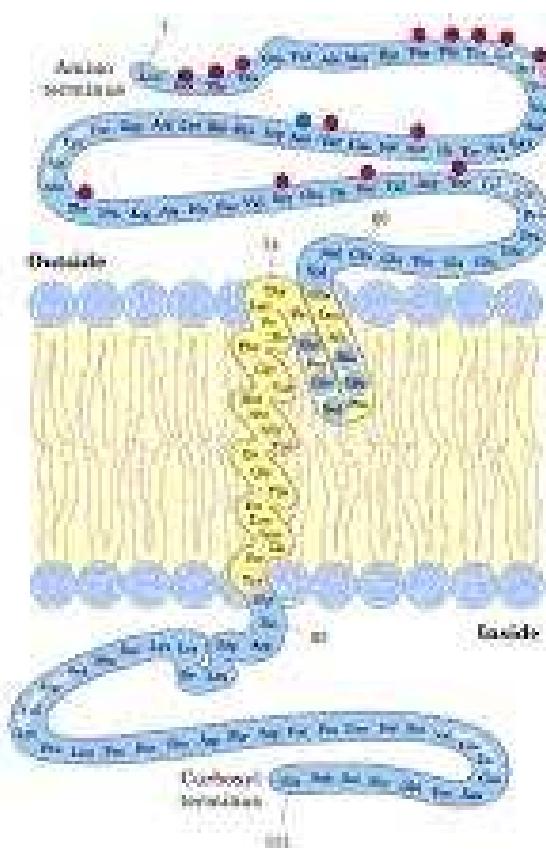


Figure 2.44: Glycophorin C protein. integral membrane protein of the erythrocyte. Possesses a single transmembrane domain (residues 49-88) and a cytoplasmic domain and in the erythrocyte interacts with band 4.1 (an 80-kDa protein) and p55 (a palmitoylated peripheral membrane phosphoprotein) to form a ternary complex that is critical for the shape and stability of erythrocyte.

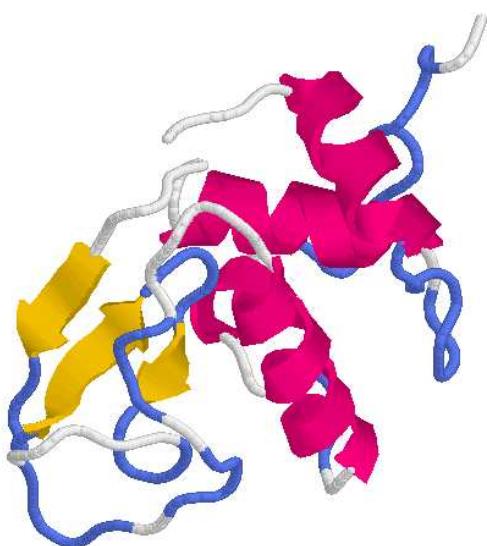


Figure 2.45: Lysozyme. A small enzyme that binds to polysaccharide chains and breaks them apart by hydrolysis. It has two structural domains. One domain is composed mostly of α helices, while the other domain is composed mostly of β strands. The interface between the two domains forms a cleft in which the substrate binds. Acts as a catalyst by adding a molecule of water to the bond between two sugars, breaking the bond. This reaction is catalyzed by two strategically positioned amino acid side chains in the enzyme's active site: glutamate 35 and aspartate 52.

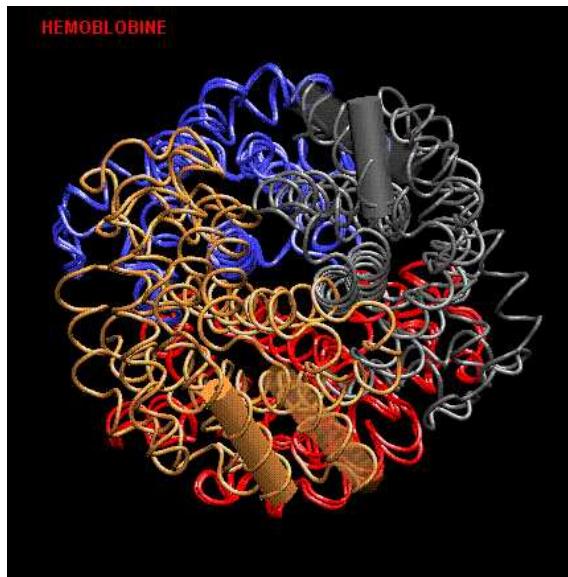


Figure 2.46: Hemoglobin. Tetrameric protein that transports oxygen. It is composed of two α -subunits and two closely related β -subunits. Oxygen binds to heme groups in the protein, which are shown in red. Each subunit can sense whether neighboring subunits contain bound oxygen. The protein subunits therefore communicate with one another through the interfaces that hold them together.

helices can lead to two different proteins as in the case of triosephosphate isomerase (TIM, PDB 1tim) and dihydrofolate reductase (DHFR, PDB 1ai9). One goal of the globular tertiary structures is to create topologies directly related with the function so the protein is able to interact either with small molecules that may bind in gaps and macromolecules with which interact via surface or region complementarity. Folded globular proteins are stabilized by packing of atoms in the internal core and by water binding to the polar side chains and potential-binding groups of the backbone. Atomic-resolution structures show a layer of bound water on the surfaces of all folded soluble proteins as a hydration shell surrounding the macromolecule. Hence, water molecules in fixed positions should also be considered as part of the tertiary structure. The efficient packing of the amino acids is achieved due to the aforementioned weak interactions (van der Waals between non polar groups and polar interactions between hydrophilic groups) which maximize the strength and the probability of such interactions to occur. The packing maintenance is accomplished through many different ways i.e. the helices and strands are finally linked together

2.2.5 Major methods of structure determination

2.2.5.1 X-ray Crystallography

After isolating and crystallizing the macromolecule, the becoming crystal is placed in an X-ray electron beam which will give the particular diffraction as a reading of the arrangement of the molecules in the crystal by bombarding it with X-ray (a form of ionizing radiation) . This diffraction pattern displays the electron density. However the displayed pattern is just the density of the

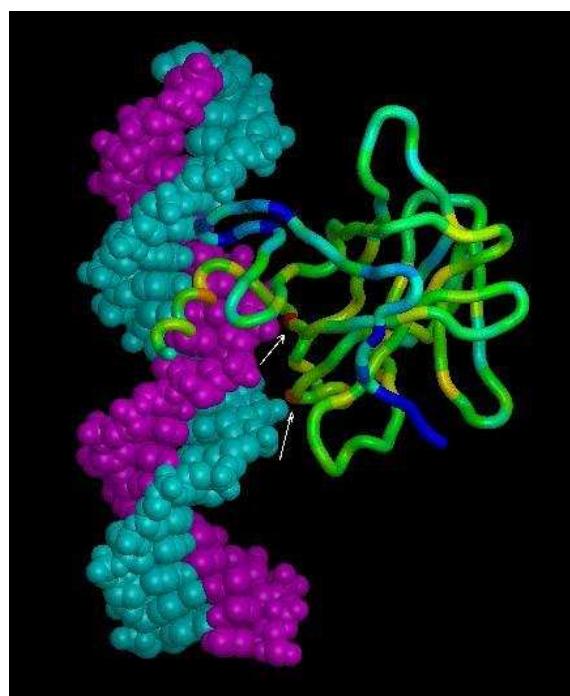


Figure 2.47: P53. The tumor suppressor protein p53 is a tetramer of four identical subunits. Each p53 subunit contains a simple tetramerization domain composed of a single β -strand connected to a α -helix. The tetrameric form of p53 assembles as a dimer of dimmers. Two copies of p53 interact via β -strands, forming a two-stranded β sheet. Two such dimmers interact via their α -helices to form the tetrameric assembled.

diffracted rays and no the electron density, hence other measurement steps should be included to solve this problem and to get more information about the phases of the diffracted rays. Once the crystallization-phases problems are resolved, a model is built up to fit into the macromolecule and the computational process proceeds in order to assure the proper stereochemistry and to maximize the agreement between the model and the measurements.

2.2.5.2 NMR Spectroscopy

Nuclei Magnetic Resonance spectra measures the levels of the magnetic nuclei in atoms with an accuracy capable of determining the values of conformational angles. It is a physical phenomenon based upon the quantum mechanical magnetic properties of the nucleus of an atom. All nuclei that contain odd numbers of protons or neutrons have an intrinsic magnetic moment and angular momentum. The most commonly measured nuclei are hydrogen-1 and carbon-13, although nuclei from isotopes of many other elements can also be observed. NMR aligns the nuclei with a very powerful external magnetic field and perturbs this alignment using an electromagnetic field. Therefore the chemical shift or signal from an atom is defined. NMR can identify pairs of atoms close together in the sequence even though they are not bonded <5 apart. The combination of both shift and bonded-non bonded pair of electrons lead NMR to define secondary structures. When protein structure determination, the common work is carried out by protons measurements even other nuclei such as N15 and C13 give also signals and can reveal additional pairs of neighboring atoms and even the tautomeric-protonated state of the Histidine ring can be determined. NMR gives ways to specify secondary structure and to infer tertiary structure.

Both X-ray and NMR have advantages and drawbacks: no crystal is needed in NMR which is a restriction sometime for X-ray and gives a time scale of approximately 10-9-10-6 seconds which permit to come into the protein dynamics. But NMR has the restriction of the size, thus a protein with more than 50-300 residues long. X-ray provides more precise values of atomic coordinates than does NMR and let us to observe the native state of the enzymatic activity. Comparisons of X-ray structure determinations of the same protein in different crystals with different intermolecular interactions, suggest that the perturbation of protein structures by the crystal environment is usually small, and localized to the regions in which the molecules are in contact the crystal

Thus, whether NMR spectroscopy or X-ray crystallography is more accurate depends on how it is considered to be artificially constrained.

Once NMR spectra and/ or crystal is obtained, mathematical approaches should be addressed: When predicting secondary structure (i.e. by Chou and Fasman or statistical methods) important inputs to take into account are the ones of the empirical rules to follow; by comparing the values of individual amino acids in a known 3D structure with results obtained randomly, thus trying to determine whether a segment of sequence will be helical, form a turn, a coiled coil, a b sheet or adopt irregular conformation. In the table bellow, normalized preferences values of individual amino acids are depicted. The values are obtained dividing the fraction of residues of each amino acid that occurred in that conformation divided by this fraction for all residues. Random occurrence would give a value of unity while values grater than 1 indicate this amino acid has a preference for this type of conformation. The gold rule for such prediction is that any amino acid can be found in any type of secondary structure. Of the twenty naturally occurring amino acids, proline is the only one that has a cyclic side chain, forming a five-membered ring that includes the backbone nitrogen. This geometry severely limits the flexibility of the backbone being disfavored

in both a helix and β sheet. Due to this restricted conformational flexibility, no hydrogen bonds can be formed. Glycine as it has a lack in one side; it can adopt a much wider range of phi and psi angles values. The two residues are found one after the other (Pro-Gly and Gly-Pro) especially in turns being therefore considered as beta turns predictors. Proline is also found a helix interrupting the helical hydrogen bonding-network producing a curve which arises to loops formation at the ends of α -helices.

2.2.6 Viewers

RasMol, Chime, Pymol and other software are free molecular visualization resources and the main is of special interest to bioinformaticists, pharmacogenetists, organic chemists engaged in pharmaceutical, agrochemical, and biotechnology R&D, in industry and academia. It is a particularly valuable tool for scientists specializing in cheminformatics. The resources employ Netscape plug-in Chime, freeware from MDL, and derived from RasMol

Secondary structures are often represented in cartoon form to clarify the underlying structure of a Protein.

Ribbon shows molecules with a "backbone" (e.g., polymers, proteins) depicting alpha helices as curled ribbons as well as the secondary structure (such as locations of any alpha helices) of a protein. It is mainly used for proteins and other polymers. As drawback it does not show individual atoms and other important structural features. This view accents α -helices and β sheets. These secondary structure elements determine the fold of most polypeptide chains. β -strands are shown as arrows pointing from the N- to the C-terminus and α -helices are shown as twisted cylinders.

Stick shows the bonds between atoms as three-dimensional "sticks", often color-coded to show atom type. Connectivities between atoms give some idea of the molecule's three-dimensional shape. Do not depict the size (volume) of the molecule or its constituent atoms, and hence gives a limited view of the molecule's three-dimensional shape.

CPK shows atoms as three-dimensional spheres whose radii are scaled to the atoms' van der Waals radii. Relative volumes of the molecule's components can be noticed. It is usually a good indicator of the molecule's three-dimensional shape and size but as a drawback this kind of representation has difficult to view all atoms in the molecule, and to determine how atoms are connected to one another.

2.2.6.1 Rasmol

Protein Explorer is a RasMol-derivate software directed to look at macromolecular structure and its relation with functions. Is much easier to use, and much more powerful because the first image of a molecule is maximally informative and explained while RasMol's is an uninformative wire-frame display without explanation. RasMol was developed by Roger Sayle. RasMol stopped its actualization in 1999 when Protein Explorer coming on

RasMol can display any molecule for which a 3-dimensional structure is available. 3D structures have not been determined for many molecules of great interest; these, RasMol cannot display. In order to display a molecule, RasMol needs a data file called an atomic coordinate file. This data file specifies the position of every atom in the molecule, as Cartesian coordinates X, Y, and Z. Three-dimensional structures can be predicted for many small molecules, but must be

AMINO ACID	ALPHA HELIX	B STRAND	REVERSE TURN
ALA	1.41	0.72	0.82
LEU	1.34	1.22	0.57
MET	1.30	1.14	0.52
GLN	1.27	0.98	0.84
GLU	1.59	0.52	1.01
LYS	1.23	0.69	1.07
ARG	1.21	0.84	0.90
HIS	1.05	0.80	0.81
VAL	0.90	1.87	0.41
ILE	1.09	1.67	0.47
PHE	1.16	1.33	0.59
TYR	0.74	1.45	0.76
CYS	0.66	1.40	0.54
TRP	1.02	1.35	0.65
THR	0.76	1.17	0.90
GLY	0.43	0.58	1.77
ASN	0.76	0.48	1.34
PRO	0.34	0.31	1.32
SER	0.57	0.96	1.22
ASP	0.99	0.39	1.24

Table 2.14: Preferences normalized values of individual amino acid to be found within specific SSEs. Leucine, methionine, glutamine, glutamic acid: long side chain amino acids found in helices principally because the side chain can project out of the central core of the cylinder being formed by the α helices. Valine, isoleucine, phenylalanine: side chains branched at the beta carbon are found in β sheets mainly because the other residues are pointing out leaving space in which the beta-branched amino acids can pack. Glycine, asparagine and proline are found predominantly in turns.

determined empirically for macromolecules. The most common method for determining structure is X-ray diffraction analysis of a crystal. Nuclear magnetic resonance (NMR) can also be used. Some structures are available only as theoretical models, often based on related molecules for which empirical structures have been determined. There are several "standard" formats for atomic coordinate files. One of the most common is the Protein Data Bank or PDB format <http://www.umass.edu/microbio/rasmol/index2.htm>

2.2.6.2 Chime

Developed by MDLI Chime is software that enables to view chemical structures from within popular Web browsers, Java Applets, and Java applications.

Chime is a browser plug-in that renders 2D and 3D molecules directly within a Web page. The molecules are "live", meaning they are not just static pictures, but chemical structures that scientists can rotate, reformat, and save in various file formats for use in modeling or database applications <http://www.mdl.com/products/framework/chime/index.jsp>

2.2.6.3 Pymol

Is a user-sponsored molecular visualization system on an OPEN-SOURCE foundation. PyMOL is a molecular graphics system with an embedded Python interpreter designed for real-time visualization and rapid generation of high-quality molecular graphics images and animations. It can also perform many other valuable tasks (such as editing PDB files) to assist you in your research. The extensible core PyMOL module (hosted here at SourceForge) is available free to everyone via the "Python" license (a simple BSD-like permission statement), but we ask all users to purchase a license and maintenance agreement in order to cover our development and support costs. In order to motivate such sponsorship, we offer support and other incentives to PyMOL licensees with current maintenance subscriptions. In this way, we seek to insure the viability of the Open-Source project by providing a specific incentive (or reward) for outside support. However, our hope is that only a small subset of PyMOL's total value will need to be restricted to Incentive packages – just enough to justify regular contributions and keep the project self-sustaining. <http://pymol.sourceforge.net/>

2.2.7 First approximation

2.2.7.1 PDB- function

The high level of atomic fluctuations allow among other things, protein adjustment to another molecule, in case of ligand binding, structure changes, in case of allosteric reactions, penetration of small molecules as water in case of reactions as hydrolysis. This flexibility is only possible due to the non covalent and hence non rigid bridges as bonds linking one amino acid to the next.

Enter the PDB code in the PDB web site and look for the structure and characteristics of the next four proteins relating them to the main function.

Binding: Specific recognition of other molecules is a crucial step in protein active function and is governed by the shape complementarity of both the ligand and the protein and polar interactions as the non covalent hydrogen bonds TATA binding protein (1tgh): Bind a specific DNA

sequence serving as platform for the complex that initiates the transcription Myoglobin (1a6k): Bind reversibly one oxygen molecule to the iron atom in its heme group and it stored oxygen for use in muscles tissues.

Catalysis: the structural features contribute in a power manner to the reactions the proteins can perform making the orientation, proximity, redox state among others, constrained conditions HV protease (1a8k): responsible of the protein-cleaving necessary in the replication of the AIDS virus HIV. In pharmacology is used as target for drugs that act as inhibitors.

Switching: The shape and flexibility in terms of conformational changes are used to switch from diverse molecular states. Ras protein (121p on and 1pll off): Bind GDP or GTP (Guanidyl di/triphosphate) groups depending on its state. This bind makes possible different proteins to recognize it in the pathway metabolism reaching different results within the same cell. Ras is an important protein in cancer pathways. It is active on state when is linked to GTP making the cell growth signals to be achieved. When the GTP is hydrolyzed to GDP the state is the non active off one and no growth is reached.

Structural proteins: directly dependent on the association of proteins subunits to with themselves as well as with other proteins, carbohydrates among other molecules. Silk (1slk): its structure defines directly the strength and flexibility. It is composed of antiparallel sheet. The strengths come from the hydrogen bonds every sheet makes one to another being the van der Waals weak interactions the ones responsible of the flexibility.

2.2.7.2 SCOP-Classes

Based on the structure, proteins are classified into four primary groups: class α class β , class α/β and class $\alpha + \beta$. This classification is available in the SCOP SCOP ("Structural Classification Of Proteins") data base. Almost all proteins present structural similarities with other proteins and in some cases this similarity could be a signal of a common phylogenetic origin. Relations between structure and evolution are provided in SCOP database based principally in a hierarchical classification of proteins in families, when besides the sequence similarity also structure and function are shared features; Superfamilies when two or more families of proteins have small similarities in their primary sequence but they share structure and function.

SCOP and enter the next examples, found the number of related families, folds and superfamilies

Class α : Myoglobin

Class β : α -amylase inhibitor

Class α/β : Mainly parallel β -strands (beta-alpha-beta patterns). Tryose phosphate isomerase

Class $\alpha + \beta$: Mainly antiparallel β -strands (separated alpha and beta section). Trans-glycosilase linked to membrane.

Multi-domain proteins: Two or more domains each one from different classes. Utiryl-CoA dehydrogenase

Surface and membrane proteins (excluding those from immune system). α -hemolysine

Proteins-Ligands : (inorganic ions, small inorganic or organic molecules, water, disulfide bond-containing) . BPTI Bovine Pancreatic trypsin Inhibitor

2.2.8 Concepts

Alpha helix: Secondary structure coiled conformation in which the backbone NH group of every residue n can hydrogen bonding with the CO group of every residue n+4

Anphiphatic: When having both polar and non polar character and hence to form interactions between hydrophobic and hydrophilic molecules. For a molecule, it is the property of having both hydrophobic and hydrophilic portions. Usually one end or side of the molecule is hydrophilic and the other end or side is hydrophobic. For an α helix is a helix with both sides hydrophilic and hydrophobic

Beta sheet: (also called pleated sheet) Secondary structure element formed due to the hydrogen bonding between the segments of the extended polypeptide chain.

Beta turn: (also called reverse turn or hairpin turn) A stretched turn that reverses the direction of the polypeptide chain and stabilized by hydrogen bonds in the backbone.

Chiral: When an object cannot superimpose their mirror images. A chiral object and its mirror image are called enantiomorphs (Greek opposite forms) or, when referring to molecules, enantiomers.

Common core: Secondary structure elements that retain their folding pattern during the evolution of a protein family. Also including the active sites if enzymes.

Entropy: Measure of randomness or disorder of a system or molecule

Enthalpy: Energy measured in terms of work that can be released or adsorbed as heat at constant pressure.

Folding pattern: The way or spatial course of the main chain of a native structure of a protein

Helical parameters: Collection of numerical values that define the geometry of a helix. The set include among others the values of phi and psi angles, the number of residues per turn, the translational rise per residue

Helix dipole: The macro-dipole formed by accumulated effect of the individual peptide dipoles in a helix. The positive end is the amino terminal (-NH) and the negative end is the carboxy terminal (-CO) bonds

Loop: peptide chains segments with no regular conformation that could lead to changes in the direction of the polypeptide.

Native state: The unique stable, active structure of a protein built up spontaneously in favorable conditions of solvent and temperature.

Nucleophile: A reagent that has two electrons to donate by forming a chemical bond to its reaction partner, the electrophile. Oxygen nucleophiles are the water molecules, sulfur nucleophiles are thiol groups (-SH) and nitrogen nucleophiles are ammonia and amines

Ramachandran plot: a two dimensional plot of the values of the dihedral torsion angles of the backbone phi and psi, in which the allowed conformations regions are indicated showing the non sterical interference.

Residue: the amino acid side chain that differentiate one amino acid from the others.

Resonance: Tendency of a system to oscillate at maximum amplitude at a certain frequency. In case of peptide, the term refers to the electron resonance stabilization in the peptide bond that turns out as rigidity not unlike the typical -C=C- double bond but a partial double-bond character

Rise: The distance a helix rises between adjacent polymer units

Secondary structure: Formation of standard conformation (helices and strands) by hydrogen covalent bonds between main chain atoms.

Super-secondary structure: two or more successive regions of secondary structures interacting in a standard conformation.

Optical isomer: molecules with the same chemical formula and often with the same kinds of bonds between atoms, but in which the atoms are arranged differently and differing in the way they rotate polarized light.

Torsion angle: angle between two groups on either side of a rotatable chemical bond. In case these two groups are the N-C and C-C in the peptide bond, the torsion bonds are called phi and psi respectively

Zwitterion: a molecule that is electrically neutral but carries both positive and negative charges. In case of an amino acid, at physiological pH, the positive charge is generated by the amine group (-NH₃⁺) and the negative charge by the carboxy group (-COO-)

2.2.9 Annexes

RasMol: Technical Introduction This description is taken from the documentation which accompanies RasMol (file ANNOUNCE). It is written by Roger A. Sayle, Ph.D., the author of RasMol. RasMol is a molecular graphics program intended for the visualization of proteins, nucleic acids and small molecules. The program is aimed at display, teaching and generation of publication quality images. The program has been developed at the University of Edinburgh's Biocomputing Research Unit and the Biomolecular Structure Department at Glaxo Research and Development, Greenford, UK. RasMol reads in molecular co-ordinate files in a number of formats and interactively displays the molecule on the screen in a variety of color schemes and representations. Currently supported input file formats include Brookhaven Protein Databank (PDB), Tripos' Alchemy and Sybyl Mol2 formats, Molecular Design Limited's (MDL) Mol file format, Minnesota Supercomputer Center's (MSC) XMol XYZ format and CHARMM format files. If connectivity information and/or secondary structure information is not contained in the file this is calculated automatically. The loaded molecule may be shown as wire-frame, cylinder (Dreiding) stick bonds, alpha-carbon trace, space-filling (CPK) spheres, macromolecular ribbons (either smooth shaded solid ribbons or parallel strands), hydrogen bonding and dot surface. Atoms may also be labeled with arbitrary text strings. Different parts of the molecule may be displayed and

colored independently of the rest of the molecule or shown in different representations simultaneously. The space filling spheres can even be shadowed. The displayed molecule may be rotated, translated, zoomed, z-clipped (slabbed) interactively using either the mouse, the scroll bars, the command line or an attached dials box. RasMol can read a prepared list of commands from a ‘script’ file (or via interprocess communication) to allow a given image or viewpoint to be restored quickly. RasMol can also create a script file containing the commands required to regenerate the current image. Finally the rendered image may be written out in a variety of formats including both raster and vector PostScript, GIF, PPM, BMP, PICT, Sun raster file or as a MolScript input script or Kinemage. RasMol will run on a wide range of architectures and systems including SGI, sun4, sun3, sun386i, DEC, HP and E&S workstations, IBM RS/6000, Cray, Sequent, DEC Alpha (OSF/1, OpenVMS and Windows NT), IBM PC (under Microsoft Windows, Windows NT, OS/2, Linux, BSD386 and *BSD), Apple Macintosh (System 7.0 or later), PowerMac and VAX VMS (under DEC Windows). UNIX and VMS versions require an 8bit, 24bit or 32bit X Windows frame buffer (X11R4 or later). The X Windows version of RasMol provides optional support for a hardware dials box and accelerated shared memory rendering (via the XInput and MIT-SHM extensions) if available.

Chapter 3

Structural Comparison and Alignment

3.1 Introduction

“An important consideration when using any structural alignment method is to consider the nature of the problem you are trying to solve and to experiment with a variety of methods”

Philip E. Bourne and Ilya N. Shindyalov, Structural Bioinformatics

When trying to determine equivalences between amino acid residues by taking into account 3D structures of known proteins, structural alignment concepts have to be introduced. Four steps should be taken when attempting to gather information about an unknown protein structure:

- 1st **Structure alignment:** find equivalences of amino acid residues based on known 3D structures.
- 2nd **Structure comparison:** based on shared similarities of two or more proteins their known 3D structures are compared.
- 3rd **Structure superposition:** based on preliminary knowledge of positive matches of some residues in proteins 1 and 2 and taking into account the alignment, the main goal is to find the optimal overlap of both proteins.
- 4th **Structure classification:** based on the structural alignment, and considering results of other methods, assign the protein to a certain class.

Even when following the previous steps, relationships between primary protein sequence, 3D structure and biological function can not be extracted so easily and other sources of information must be accessed. Structural alignments provide information that is unavailable through current sequence alignment methods. It is a direct consequence of natures specific protein selection that the approximately 30 000 proteins needed for the functioning of a complex organism adopt just 1000 to 5000 (Chotia, 200) possible folds out of the 20^x possible ones (x being the number of amino acids in a polypeptide chain forming a protein). More than 3000 structures are stored in

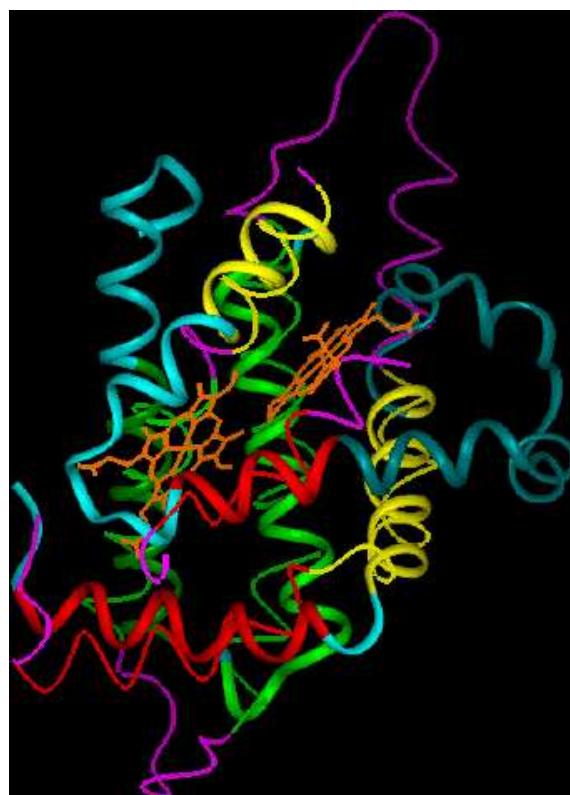


Figure 3.1: Structure superimposition. Structure superimposition of a globin fold running the software Spatial Arrangement for Backbone Fragments (Alexandrov NN) SARF2 software result when comparing myoglobin (1mgb) and cytochrome P450(1amo). The match consists of 109 Ca-atoms, superimposed with $\text{rmsd}=3.0 \text{ \AA}$. The topology of the helices is different; molecules from both structures are shown in orange.

the structural protein data bank, but due to a high level of redundancies, many of these proteins are very similar. When plotting the percentage of sequence similarity versus length of polypeptide chains in an X-Y graph to obtain relational information, we obtain clues as to which methods are suitable: sequence identities between 20-30 % are only detectable by sequence methods, whereas those lower than 20 % , in the so called midnight zone, are detectable by structure comparisons methods. Protein structures are more highly conserved than sequences. Consequently, family folds with related structures but varied sequence identities have emerged. Evolutionary changes like insertions and deletions take place mainly in loop regions, thus causing no alterations to the final fold and limiting the number of possible folds. When comparing structures two important points have to be considered:

- Similar structures may be formed by alternative folding of the amino acids C_α backbone (matched regions on the 2 proteins can be separated by unmatched segments).
- Partial local similarities do not automatically transfer to similarities in structure because there may also be local differences (proteins with similar nucleus but different ends)

Since it is computationally less expensive to align linear sequences than to compare 3D structure of proteins that have approximately 30% of common features, some algorithms have been developed so the structures can be compared, assuming they adopt the same fold. Some studies show that proteins with a sequence identity even as low as 30% adopt the same folds (homologous folds, hence they share a common ancestor); moreover similarities of just 5% can result in the same fold (analogous folds and hence no common ancestor).

3.2 Methods for Structure Comparison and Alignment

Both structure comparison and alignment methods are defined as **NP-hard problems**, non-deterministic polynomial time problems, that can be solved by an all-heuristic approach (replicable method or approach). Such kinds of problems define a set of decisions solvable in polynomial time on a non deterministic touring machine (machines that can be adapted to simulate the logic of any machine). NP-hard problems are for example optimization problems, search problems and decision problems. Structural comparison and alignment through any heuristic method can lead to the best analytical answer but that does not mean its the best biological one. Numerous methods have been developed to deal with this problem. Some also try to solve the optimization of the alignment between any given pair of proteins or to find the most suitable target in PDB when comparing a new structure. Although computers are very fast, it is still a time consuming problem to carry out the aforementioned job (for comparing 323 structures all against all 5 CPU days work on a SVN 4 are required) Methods developed are the result of two possible ways to find a solution: on one hand proteins that are released every week in PDB are added to an all-against-all comparison database and on the other hand all relationships known between sequence and structure are used so the number of computations to be performed is reduced. In addition PDB structures can be grouped so the target protein is compared against only a subset of the complete PDB database. These filtering steps give an estimated ratio of 1: 10 for new folds against the number of new protein structures. Consequently, 5 out of 10 similarities can be inferred just from the sequence without any kind of algorithm for structure comparison. The guidelines for many

methods related to the jobs described above can be summarized by following the steps either for structure comparison and alignment or for multiple structure alignment:

A. Structure comparison and alignment

1. Representation of the pair of proteins A and B, domains or fragments to be compared and aligned.
2. Compare A and B
3. Optimize the alignment between A and B
4. Statistically significant measurement of the alignment against a random set of structures (from any protein database, normally PDB)

B. Multiple structure alignment (besides A1, A2, A3, A4)

1. Starting from the initial alignment found in A3, the next step is running a search within a sequence constrain window to find the optimal alignment against all structures using profiles; HMMs or Monte Carlo approaches.

Multiple structure alignment (msa) methods compute possible alignments between all structures simultaneously in an attempt to find the best consensus alignment between all structures. The results normally display weak sequence relationships. Analysis of profiles and HMMs as well as Monte Carlo optimization can be used in such way that matches found previously when comparing a pair of structures are moved across the remaining structures in the search; a suitable pairwise alignment found by any methods described below (CE, DALI, etc) moves through a limited search space against all the structures to find the optimal alignment performed by Monte Carlo, HMMs or profile approaches. See Leibowitz, Nussinov and Wolfson, (2001) for an msa approach.

CE, DALI, SSAP, VAST, SARF2, COMPARER are available programs for structure comparisons and alignment. In order to go through the aforementioned methods an overview of dynamic programming and distance matrices must be given.

3.2.1 Basic remind

3.2.1.1 Dynamic programming

We use dynamic programming algorithms to find solutions to NP-hard problems in a computationally cheaper way. They can only be used if you can break down an optimization problem into sub-problem so that at any given stage the optimal solutions are known. There are two very important applications of dynamic programming in structural bioinformatics:

1. **Aligning sequences:** the goal is to bring as many identical or similar sequence characters into vertical register in the alignment as possible at the minimum cost of insertions and deletions. A row of amino acids in one sequence matches a row of identical or substituted positions in the second sequence; insertions or deletions show up as gaps in the respective sequence.

2. **Aligning structures:** a scoring matrix is built in order to compare the positions of the atoms in both 3D structures. Each column in the scoring matrix gives a score of how well any of the 20 amino acids fits to a single position in the structure; an optimal alignment is then calculated. The method first examines the positions of secondary structural elements (α -helix and β -sheets) within a domain and searches which types, positions and numbers of these elements are similar. Subsequently, the distances between the C_α ($NH-C_\alpha-C_\beta$) and C_β ($C_\alpha-C_\beta O=NH$) atoms within these domains, and later within the whole structure, are checked and the degree of superimposition is determined. It follows that the better the arrangement, joining and 2D alignments are, the more significant and convincing is the result.

The dynamic programming encloses two steps:

1. The atoms or molecules are treated as vectors and are given a coded value that describes the local environment of each amino acid, that is the sum of the interatomic distances plus bond angles and R groups. Subsequently, Cartesian coordinates are assigned to each (X, Y, Z) and the direction of the bond angles is included.
2. The alignment of 2D structures is carried out to determine the interatomic distances between each amino acid in the polypeptide chain. The coordinates used to draw the vectors for comparison are the ones corresponding to the beginning and the end positions of the secondary structures, thus a helices and b strands.

3.2.1.2 Distance Matrix

Distances between C_α atoms along the polypeptide chain and between C_α atoms within the protein structure are compared by a 2D matrix. The matrix compares relationships between structures without the help of alignments. Each position in the matrix represents the distance between corresponding C_α atoms in the 3D structure. The smallest distances represent the more closely packed atoms within secondary structures and regions of the 3D structures. Similar groups of 2D structural elements are superimposed by minimizing the sum of the atomic distances between the aligned C_α atoms resulting in a common core structure.

Bases of Distance method

Based on the degree to which all of the matched elements can be superimposed. The score of a matching set of helices is the sum of the similarity scores of all the atom pairs as follows:

Protein A helices a and b interacting Protein B helices a' and b' are interacting The helices of protein B and the helices of protein A are superimposed taking

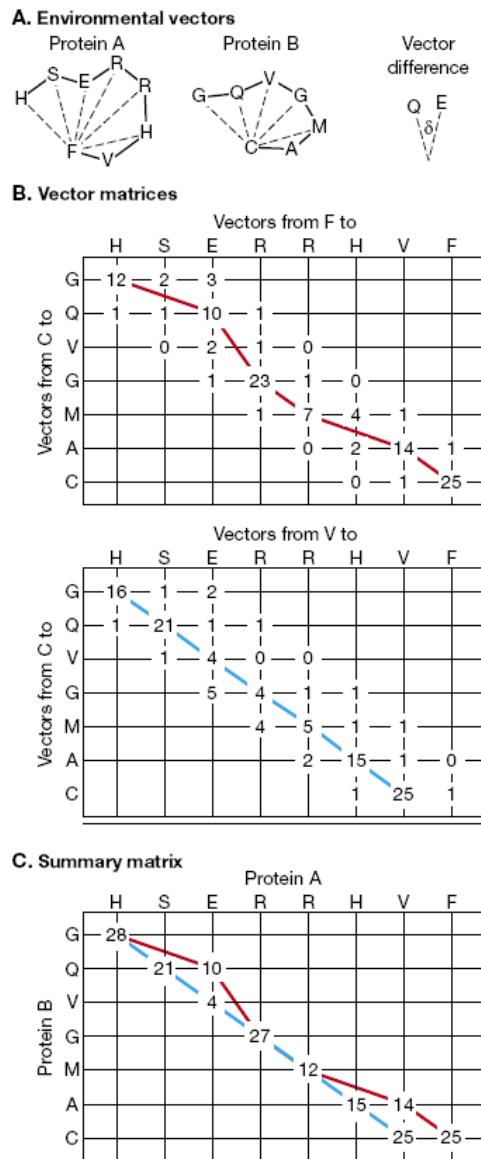


Figure 3.2: Double Dynamic Programming (DDP). (A) Step1: Two dimensional projections of vectors from one amino acid on a set of other close amino acids in each of two protein segments to be compared. Both fragments can be interchanged and placed in each others protein context in order to get the difference vector, since for both the same coordinate system is defined. The smaller the difference the more alike are the structures . (B) Step 2: Matrix in which the two vectors are plotted against each other; the differences between each C vector and the amino acids forming part of the environmental vector in each protein are measured. An optimal alignment (red path) is computed by a global form of the dynamic method. The procedure continues by calculating the vector differences for the next amino acid in one protein (in this case amino acid V for protein A) in an each against all comparison of the environmental components (blue path).(C) Summary matrix in which the resulting alignments are placed (red and blue paths) . (A) and (B) first dynamic programming alignment, (C) second dynamic programming alignment.

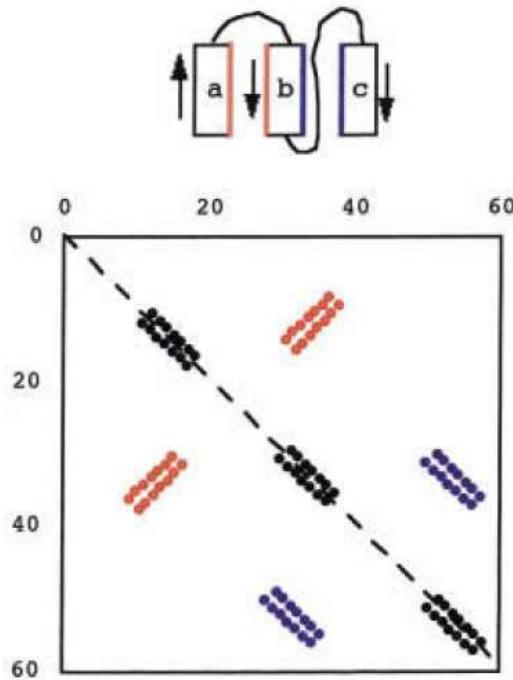


Figure 3.3: Distance Matrix. Distances between the C_α atoms of the amino acids in a hypothetical three helix structure; the known 3D structures of the proteins are plotted in the matrix. Consecutive and very close amino acids in the helix are represented as dashed lines; shortest C_α - C_α distances along the diagonal indicate positions of the a helix (red and blue dots). Helices a and b show opposite polarity (dots perpendicular to the diagonal) while b and c show the same one (dots parallel to the diagonal).

Sets of C_α atoms:

In helix a: i^A In helix b = j^A

In helix a': i^B In helix b' = j^B

Matched pairs correspond

$d_{ij}A$ = distance between i^A and j^A (distance between the C_α set of atoms in helix a and helix b of protein A)

$d_{ij}B$ = distance between i^B and j^B (distance between the C set of atoms in helix a and helix b of protein B)

$d_{ij}^* = \text{average of } d_{ij}A \text{ and } d_{ij}B$

The similarity score for this pair of atoms is calculated as

$$SS = \frac{|d_{ij}A - d_{ij}B|}{d_{ij}^*}$$

Threshold SS = 0.2: two atom pairs can be superimposed

Adjacent β strands (matching 1 Å)

Adjacent α helix and β strands (matching 2-3 Å)

Threshold $SS < 0.2$: two atoms pairs can not be superimposed.

3.2.2 SARF2, VAST, COMPARER

Both VAST and SARF are Structure prediction programs based on vector comparisons.

The secondary structural elements are converted into vectors based on their position, direction and length. This is computationally more simple than comparing the positions of all C_α and $C\beta$ atoms.

3.2.3 SARF2: Spatial Arrangement of Backbone Fragments

A method based on comparing the C_α of each residue in the secondary structural elements (SSEs) of each protein. The procedure is designed to find those SSEs which can form similar spatial arrangements but have different topological connections (Nickolai N Alexandrov 1998). First the SSEs are detected through comparison with common templates for α -helices and β -strands, and then larger assemblies of SSEs are constructed from the compatible pairs found.

In the first step pairs of SSEs are matched up, and from the middle point of the line drawn, the next points are measured.

- Shortest distance between their axes
- Closest point on the axes
- Minimum and maximum distances from each SSE

In the second step the largest possible ensembles are formed by

- Graph theory and maximum clique problem approximation

Finally, the alignment is extended by

- Including additional residues

The similarity score is calculated as a function of rmsd and the number of matched C_α atoms. The significance of the comparison is evaluated by comparing this score with the one obtained when comparing a model protein (leghemoglobin, Fischer et al, 1996) with a non redundant set of structures.

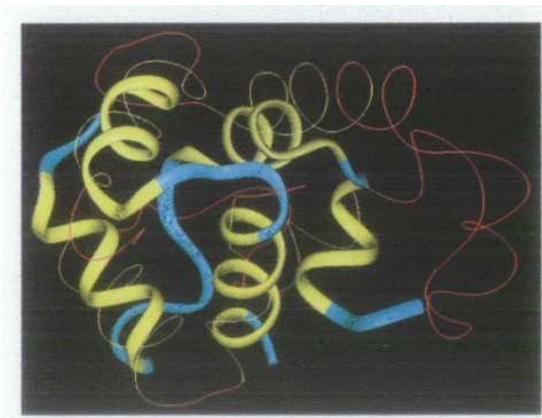


Figure 3.4: SARF superimposed result. Repressor 434 and calcium binding protein recoverin, superimposed structures. Only the C-terminal domain (residues 97-129 and 147-190) of recoverin is shown. Repressor 434 (Ir69) is shown as a blue ribbon and recoverin (lrec) as a red line. Yellow fragments can be superimposed with a small rmsd (2.61). Matches of 52 C_α were found. There is no evolutionary relationship between the two proteins but structural stability of the motif is apparent.

3.2.3.1 VAST: Vector Alignment Search Tool

Method based on the representation of structures as a set of vectors of secondary structural elements whose direction, type and connectivity infer the topology of the structure. Based on a SSEs-pair alignment, uses Gibbs sampling algorithm to examine alternative alignments. The VAST algorithm uses a statistical theory for calculating the probabilities of an alignment similar to that of the BLAST algorithm.

- BLAST: the probability or expected value that a sequence alignment score at least as high as that found between a test sequence and a DB sequence would also be found by an alignment of random sequences.
- VAST: the score is the number of superimposed secondary structure elements found by comparing two structures. The statistical significance (SS) is the likelihood that such a score would be the result of a random alignment of unrelated structures. This SS is the product of two numbers N1 and N2
 - N1 = probability that such a score would be found by picking elements randomly from each protein domain
 - N2 = number of alternative element pair combinations

$$SS = N1 \times N$$

The optimal alignment is that with the highest relation to the background distribution of C_α in the superimposed amino acid residues.

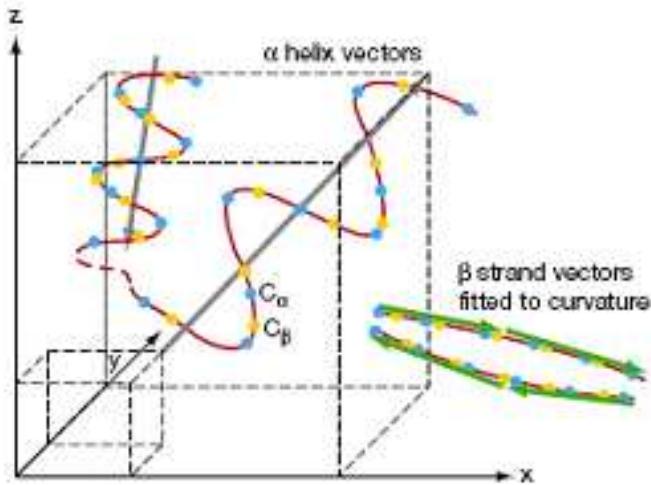


Figure 3.5: Structural vector alignment software SARF2 and VAST. Two α -helices and two β -strands in their vector representation are shown as 2 dimensional projections of the three-dimensional structure. Only the coordinates of the beginning and the end of the secondary structures, represented here by wide dashed lines, are needed to specify the location of the vector. Connectors of secondary structural elements like loops connecting α helices are also detected by the algorithm. The 3D structures of proteins are predicted to be similar if, once the representations of their vectors were compared, the type and arrangement are alike within a rational range.

If the elements in two structures are similarly arranged, the corresponding 3D structures are also expected to be similar. Once these elements are found, a clustering classification ensures that these elements are pooled in larger alignments groups of secondary structure elements. The most likely ones must be selected.

VAST and SARF are methods used for comparing new structures to the existing DB or for viewing structural similarities within the existing DB.

3.2.3.2 COMPARER

A method that uses equivalences between protein structures to define general topologies. The alignment procedure is based on the sequence alignment algorithm of Needleman & Wunsch. The modus operandi involves both the comparison of properties and of relationships through simulated annealing and dynamic programming

Fourteen properties and relationships are compared:

Properties:

Residues like identity and local conformation. Segments like secondary structure type and orientation relative to the center of gravity.

Relationships:

Relations between residues like hydrogen bonds and hydrophobic clusters.

Relations between segments like distance to one or more closer neighbors and the relative orientation of two or more segments.

For properties: A dynamic programming algorithm is used to find an optimal alignment.

For relations: A combinatorial simulated annealing technique (Sali and Blundell, 1990) is applied

The method uses two scores, the values E (residue equivalences) and A (gap penalties) for statistical analysis, comparing to the values obtained by using two unrelated proteins and two random sequence relationships.

3.2.4 CE, DALI, SSAP

3.2.4.1 CE: Combinatorial Extension of the Optimum Path

This is a method that uses distance matrices in which the distance between each C_α of each octamer fragment combination from both proteins is plotted and represented. The method is more robust and fast in finding an accurate 3D structure alignment and is not sensitive to the optimization protocol as Monte Carlo and other clustering algorithms are. CE sets up a empirical target function for the heuristics that assumes the continuity of the aligned path when including gaps, and the existence of one optimal path. CE will not solve non topological similarities. Assumed rules

- The rmsd (root mean square deviation) between two chains is $< 2 \text{ \AA}$
- The difference in length between two chains is $< 10\%$
- The number of gap positions in the alignment is $< 20\%$ of aligned residue positions
- At least 2/3 of the residue positions in the chains are aligned

The alignment of the octamers is based on heuristic measurements and every time the octamer fragments match in both proteins, is said to be an Aligned Fragment Pair (AFP). One fragment of a given length m (empirically 8) from the first protein and another fragment from the second protein form a pair if they satisfy a similarity criterion.

Three thresholds based on empirical comparison of intra-residue distances in known aligned proteins are employed:

1st threshold detecting AFP

2nd threshold detecting the correctness of a next candidate AFP relative to the current one

3rd threshold evaluating all alignments to find the optimal ones

Due to the restricted gap size, the bottleneck computational effect is eliminated but penalized with the loss of non topological alignments and insertions of more than 30 residues. As explained above when significant alignments are found former optimization must be performed like structure superposition. For statistical analysis, two distributions corresponding to both proteins, the root mean square deviations (rmsd) and gaps values for the non-redundant set are built and numerically tabulated. Assuming normality the final z-score is calculated by combining both z-scores. Two methods can be addressed as approximations to the comparison problem, one that detects homology for only structural information and another one that includes composite properties.

Method 1. For detecting structural homology from ONLY structural information

1. ALIGNMENT PATH

The alignment between two protein structures A and B with a given length is considered the longest continuous path P of AFPs in a similarity matrix S. The selection of starting point for the alignment path is determined by the length of points such that all starting points not leading to an alignment of length greater than the length of the longest alignment found thus far are discarded. Computational time is saved but limits the matches to one per polypeptide chain.

Protein A Length: n^A

Protein B Length: n^B

AFPs fixed size: m (8 has been shown empirically to be the practical choice)

Similarity matrix size: $(n^A - m)(n^B - m)$

The first AFP starting the path can be selected at any position within the similarity matrix S, but two consecutive AFPs i and i+1 in the alignment path are added only and only if the following conditions are satisfied:

- Condition (1): No Gaps between AFPs i and $i + 1$

$$P_{i+1}^A = P_i^A + m$$

and

$$P_{i+1}^B = P_i^B + m$$

- Condition (2): Gaps inserted in protein A

$$P_{i+1}^A > P_i^A + m$$

and

$$P_{i+1}^B = P_i^B + m$$

- Condition (3): Gaps inserted in protein B

$$P_{i+1}^A = P_i^A + m$$

and

$$P_{i+1}^B > P_i^B + m$$

P_i^A : the AFP starting residue position in protein A at the i -th position in the alignment path
 P_i^B : the AFP starting residue position in protein B at the i -th position in the alignment path

The search is limited to a gap of no more than 30 residues in both proteins A and B to be compared. Two conditions are included to enhance the condition (2) and (3)

- Condition (4): Gaps on protein A

$$P_{i+1}^A \leq P_i^A + m + G$$

- Condition (5): Gaps on Protein B:

$$P_{i+1}^B \leq P_i^B + m + G$$

G: Maximum allowable size of the Gap (30)

2. DISTANCE MEASURES FOR SIMILARITY EVALUATION

The evaluation of similarity is followed by three distance measures

- (i) Distance D_{ij} calculated using an independent set of inter-residue distances. Each residue participates once and only once in the selected distance set. This distance is used to evaluate combination of two AFPs, one already in the alignment path and other to be added

$$D_{ij} = \frac{1}{m} \left(\left| d_{P_i^A P_j^A}^A - d_{P_i^B P_j^B}^B \right| + \left| d_{P_i^A+m-1, P_j^A+m-1}^A - d_{P_i^B+m-1, P_j^B+m-1}^B \right| + \sum_{k=1}^{m-2} \left| d_{P_i^A+k, P_j^A+m-1-k}^A - d_{P_i^B+k, P_j^B+m-1-k}^B \right| \right) \quad (3.1)$$

- (ii) Distance D_{ij} calculated using a full set of inter-residue distances. All possible distances except those for neighboring are evaluated. This distance is used to evaluate a single AFP i.e., the accuracy of how well two protein fragments forming an AFP match each other.

$$D_{ij} = \frac{1}{m^2} \left(\sum_{k=0}^{m-1} \sum_{l=0}^{m-1} \left| d_{P_i^A+k, P_j^A+l}^A - d_{P_i^B+k, P_j^B+l}^B \right| \right)$$

- (iii) RMSD obtained from structures optimally superimposed using least-squares minimization (Hendrickson, 1977). This distance is used to select the best alignments and for the optimization of gaps in the final alignment

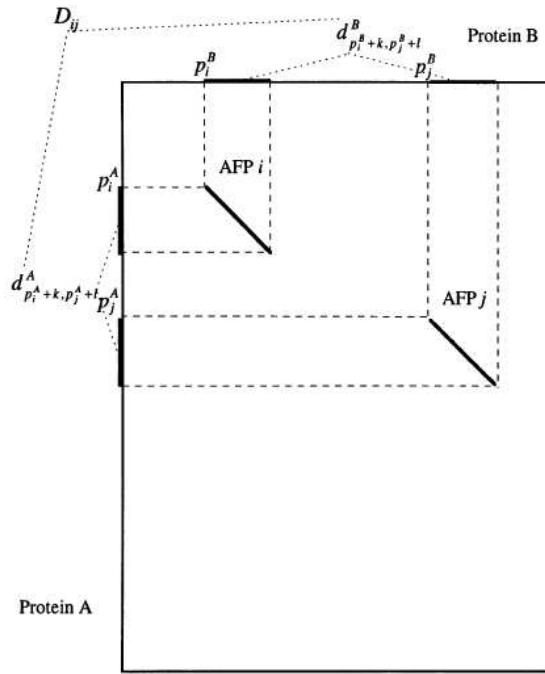


Figure 3.6: Calculation of distance D_{ij} for two AFPs.

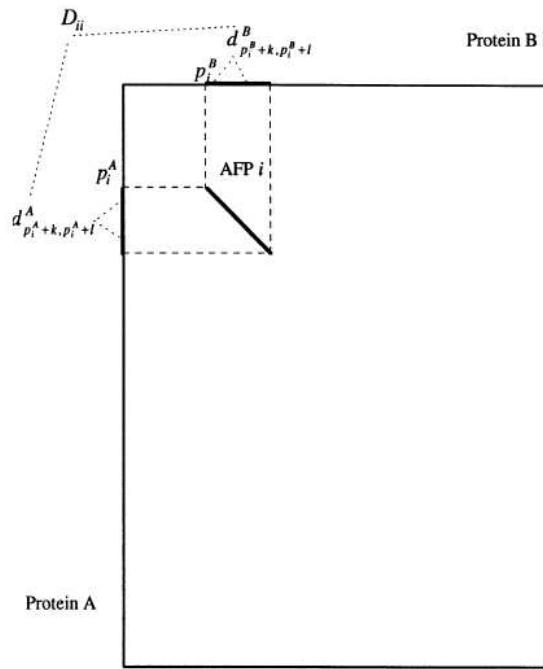


Figure 3.7: Calculation of distance D_{ij} for a single AFP.

D_{ij} : Distance between two combinations of two fragments from proteins A and B defined by two AFPs at positions I and j in the alignment path where $i \neq j$. In the case of a single AFP, $i = j$, the distance is denoted as D_{ii} .

d_{ij}^A : Distance between residues i and j in the protein A based on the coordinates of C_α atoms

d_{ij}^B : Distance between residues i and j in the protein B based on the coordinates of C_α atoms

When adding the next AFP to the alignment path, three strategies can be used

- (i) All possible AFPs which extend the path and satisfy the similarity criteria. Exhaustive combinatorial search for the optimal path
- (ii) Only the best AFP which extend the path and satisfy the similarity criteria. Limited search among the best paths
- (iii) Intermediate strategy

Three heuristics were used to decide whether a path should be extended and three conditions were resulted

1. Single AFP

$$\text{Condition(6)} D_{nn} < D_0$$

2. AFP against the path

$$\text{Condition (7)} \frac{1}{n-1} \sum_{i=0}^{n-1} D_{in} < D_1$$

3. Whole path

$$\text{Condition (8)} \frac{1}{i^2} \sum_{i=0}^n \sum_{j=0}^n D_{ij} < D_1$$

n: the next AFP to be considered for addition to the alignment path of n-1 AFPs in length.

D_0 : similarity threshold with value 3 Å.

D_1 : similarity threshold with value 4 Å.

The most accurate alignment is the one resulted by following the next steps: selected AFPs based on condition (6), the best AFP one based on condition (7) and to extend or terminate the path based on condition (8). An optimization of the final path is also applied contributing up to 2Å improvement in the rmsd between two proteins A and B but it is only valid for Z-scores above the threshold value of 3.5. The 20 best paths are evaluated and the best one selected. Gaps in this single alignment are evaluated for possible relocation in both directions up to m/2 positions and whenever the rmsd indicating superimposition structures is improved then the modified gap boundaries are adopted. Finally dynamic programming is executed on the distance matrix calculated using residues from the two superimposed structures.

Terminal gaps are not penalized.

Method 2. For detecting structural homology from ONLY structural information

After the initial superposition through the method above described is obtained, the similarity is calculated by adding the following properties. These properties are represented as scores P_{ij} being determined each score for each property.

P_{ij} measures the match between residues i and j from two proteins

Structure: **Property 1**, defined by coordinates of C_α atoms

$$P_{ij} = \begin{cases} c_1 - d_{ij}, & \text{if } c_1 - d_{ij} > c_2 \\ c_2, & \text{otherwise} \end{cases} \quad (3.2)$$

d_{ij} : distance between residues i and j in proteins A and B calculated from the C_α atomic coordinates after obtained the superposition with CE algorithm

C1: Constant to convert d_{ij} into a composite vector with value 7

C2: Constant to convert d_{ij} into a composite vector with value -2

Sequence: **Property 2**

P_{ij} is the value of the PET91 matrix for amino acids at positions i and j

Secondary Structure: **Property 3**

$$P_{ij} = \begin{cases} 1, & \text{if } s_i = s_j \\ 0, & \text{otherwise} \end{cases} \quad (3.3)$$

s_i : Secondary structure code for an amino acid defined by Kabsch and Sander (1983)

Solvent exposure: **Property 4**

$$P_{ij} = E_0 - |E_i - E_j| \quad (3.4)$$

E_i : solvent exposure defined by Lee and Richards (1971)

E_0 : constant

Conservation index: **Property 5**, BLAST and dynamic programming

$$P_{ij} = 20 - |I_i - I_j| \quad (3.5)$$

The calculus is done on residue-by residue basis and then dynamic programming to find the optimal alignment for the whole polypeptide chain. The composite property that measure structural similarity at residue level is defined as:

$$\tilde{P}_{ij} = \sum_k w_k * P_{ij}^k \quad (3.6)$$

P_{ij}^k : structural similarity for residues at positions i and j from proteins A and B calculated based on the k^{th} property

w_k : weight chosen empirically

Local dynamic programming was used with a gap initialization penalty of 10 and gap extension penalty of 1. The alignments obtained were compared as follows:

$$a^D = \sum_i a_i^D \quad (3.7)$$

$$a_i^D = \begin{cases} 1, & \text{if } a_i^1 \neq -1 \text{ and } a_i^1 \neq a_i^2 \\ 0, & \text{otherwise} \end{cases} \quad (3.8)$$

a^D : number of differences between the first and the second alignment

a_i^1 : residue position from the second sequence in the alignment matching a residue located at the i -th positions in the first sequence in first alignment.

a_i^2 : residue position from the second sequence in the alignment matching a residue located at the i -th positions in the first sequence in second alignment

a_i^1 and a_i^2 are assigned as -1 if position i is not aligned to any other position.

3.2.4.2 DALI: Distance Matrix Alignment

Distance Alignment program based on the use of distance matrices to represent each structure as a 2D array for aligning protein structures. It represents a general approach to align a pair of proteins represented by two dimensional matrices. The method allows gaps of any length, reversal of chain direction and free topological connectivity of aligned segments.

We can distinguish two categories of searches can be defined: the ones focused on finding occurrences of predefined structural pattern in a database (we have to define the object function that

minimizes dissimilarities) and the ones focused on finding the largest common structure between two proteins (we have to define a similarity measure that balances the contradictory requirements of maximizing the number of equivalences residues and of minimizing structural deviations) Plotted in the matrix are the distances between the C_{α} - C_{α} residues in the 3D whole structure of the protein. First the distance matrices are decomposed into sub-matrices of fixed size that contain elementary similar patterns as fragments of hexapeptides-hexapeptides contact patterns. Then the similarities in both matrices (for protein A and B) are paired and combined into larger combined sets of pairs, thus the overlapping occurs within the distance matrix and the regions (or sub-matrices) are then enclosed and compared together looking for similar contact patterns. The DALI program is originally based on Monte Carlo simulation, thus probabilistic method to improve previous found alignments. Monte Carlo approximation is used here to optimize the similarity score defined as equivalent intramolecular distances. The method can be summarized as follows:

Hexapeptides-hexapeptides contact patterns: equivalents fragments Identification of new matching contact patterns sharing the previous equivalent fragment: (a,b)-(b,c)-(c,d). Iterative improvement to maximize the similarity of the alignment built up

The significance outcome can be visually identified: matches substructures are patches or sub-matrices which represent small different in distances. A main diagonal is formed once the patches or points overlapping and are centered. This corresponds to locally similar backbone conformations, thus SSEs. Matches of short distances off the main diagonals, thus out of the diagonal do represent tertiary structure similarities. Common motifs structural motifs are represented as disjoint regions of the backbone To find out gap insertion or deletion is enough to move horizontally or vertically one structure representation to the other. Solution of branch and bound algorithm can be addressed under Holm and Sander (1996).

The similarity score is derived from all against all comparisons with less than 30% sequence identity and the DALI score is expressed as the number of standard deviations (z-score) from the average score derived from the DB distribution.

Method

The comparison between two protein structures A and B is given by the match of two substructures using the additive similarity score S. The larger the value of S, better set of residues equivalences

$$S = \sum_{i=1}^L \sum_{j=1}^L \Phi(i, j) \quad (3.9)$$

i and j: label pairs of matched residues e.g i=(iA, iB)

L: number of matched pairs or the size of each substructure

Phi: similarity measure here based on the C_{α} - C_{α} distances d_{ij}^A and d_{ij}^B

$$\Phi^R(i, j) = \Phi^R - |d_{ij}^A - d_{ij}^B| \quad (3.10)$$

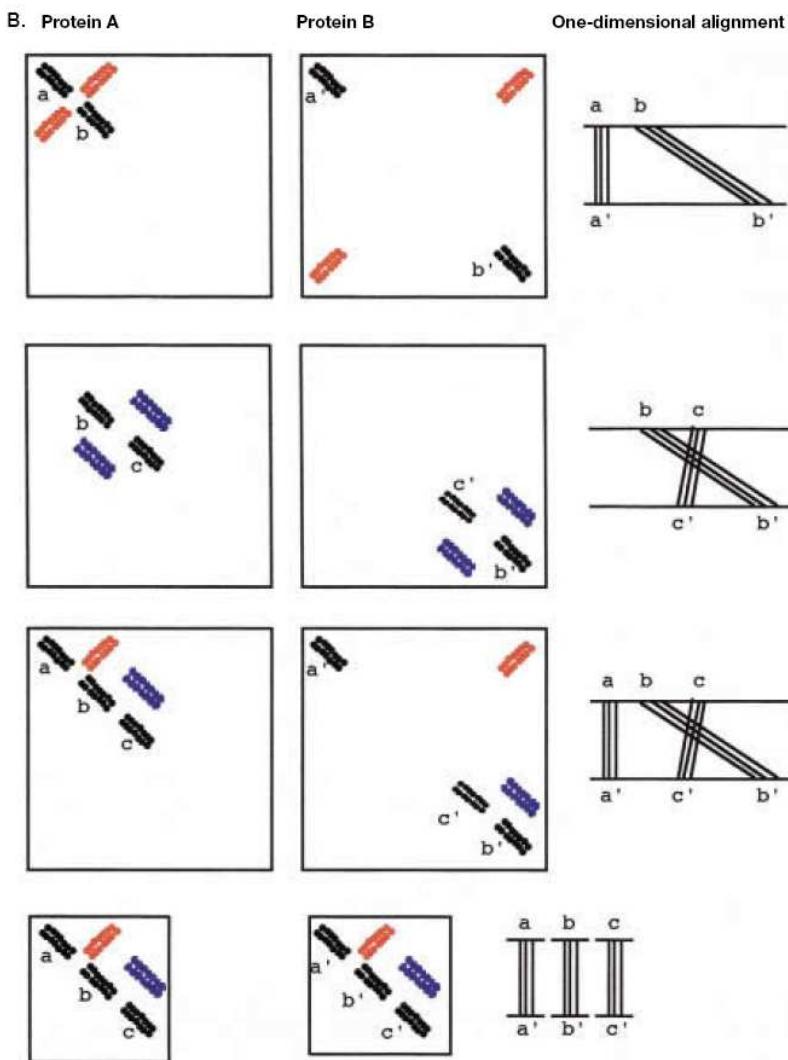


Figure 3.8: Hypothetical structural alignment by DALI. Distance Matrices for Helices a, b of the protein A and a_S, b_S of protein B are performed to search for distance patterns (first row, first and second columns respectively). The algorithm works overlapping sets of sub-matrices (6 x 6 amino acids) from the whole matrix for each protein and comparing them to place similar configurations and then combining them to build complete alignment. In order to produce a parallel alignment, DALI removes modifications as insertions and deletions.

R: rigid

d_{ij}^A and d_{ij}^B : Equivalence elements in the distance matrices of proteins A and B

Φ^R : 1.5 Å (zero level of similarity) or similarity threshold

The gradual geometrical distortions effect is more tolerant when including the elastic variant of the residue-pair score, Φ^E (superscript E) reflecting the relative deviation of equivalent distances

$$\Phi^E(i, j) = \begin{cases} \left(\Phi^E - \frac{|d_{ij}^A - d_{ij}^B|}{d_{ij}^*} \right) w(d_{ij}^*), & i \neq j \\ \Phi^E, & i = j \end{cases} \quad (3.11)$$

Envelope function:

$$w(r) = e^{\frac{r^2}{\alpha^2}} \quad (3.12)$$

d_{ij}^* : average of d_{ij}^A and d_{ij}^B :

Φ^E : similarity threshold

w: envelope function that weights the contribution of pairs in the long distance range ($\alpha=20$ Å calibrated on the size of a typical domain)

Note: the alignments reported are generated by using the elastic similarity $\Phi^E(i, j)$

The chosen value for Φ^E here is 0.20, thus 20% deviation.

Adjacent strands in a β -sheet which typical distance is 4-5 Å should match within 1 Å. For strands-helix or helix-helix with typical distance of 8-15 Å should match within 2-3 Å

The method accurately aligns related proteins pairs and detects common 3D folding motifs in database search. When a couple sets of coordinates are given, DALI is able to determine the maximal structures providing an alignment of the common residues.

The Program is fast enough to scan the entire PDB looking for Protein similar to a probe structure. As a newly structure is found, the DALI in web site should be used. When new crystallography structure or NMR coordinates are being obtained, the routine search, submits the coordinates to the DALI server for getting similarities to known proteins. The FSSP(For classification based on structure-structure alignment of Proteins) and the DALI domain dictionary are organizing according the results this program gives when compare the query structure with existing structures run against the entire PDB.

The drawback of this method is that there is not an algorithm for direct alignment because it should find the closest alignment of 2 sets of points in 3D space and that is computationally a difficult problem.

3.2.5 SSAP: Secondary Structure Alignment Program

Method that uses double dynamic programming (DDP) to obtain the optimal alignment.

The DDP is applied in terms of matrixes on:

- A first matrix to get the **selected matches**. The matrix is constructed by placing the differences in distances between C_β - C_β of positions i and j of the proteins A and B respectively, to all the other proteins positions.
- A second matrix to get the **scores** S_{ik} . For every pair of positions i and k of proteins A and B, vectors between C_β at position i and k are compared with the C_β atoms in the selected matches in the first matrix. Intra-protein C_β - C_β vectors comparison that provides directionality

The procedure is the following:

Each amino acid in each sequence is given a local environment.

Local environment = $\sum R + \text{bonds angles} + \text{interatomic distances} + \text{degree of burial in hydrophobic core} + \text{type of secondary structure.}$

Interatomic vectors between positions i and j of the protein n : \vec{v}_{nij}

Average vector

$$\vec{r}_i = \frac{1}{n} \sum_{n=1}^N \vec{v}_{nij} \quad (3.13)$$

N : number of total residues

v_{nij} : is the interatomic vector

Error associated

$$e_{ij} = \frac{1}{N} \sum_{n=1}^N (\vec{r}_{ij} - \vec{x}_{ij})^2 \quad (3.14)$$

Score: as the difference between the average vectors of the two pairs residues in the two proteins to be compared

$$S_{ijmn} = (\vec{r}_{ij} - \vec{r}_{mn})^2 \quad (3.15)$$

ij: residues in protein A whose interatomic vectors is given

mn: residues in protein B whose interatomic vectors is given

To build up a consensus vector for providing additional information, a shift vector S_{ij} is used

$$S_{ij} = \vec{A}_i - \vec{A}_j + \vec{r}_{ij} \quad (3.16)$$

2rhe00	87.7	86.9	83.9	90.7	79.8	80.4	80.0	86.9	78.5	78.5
1cd800		84.7	76.0	87.4	80.1	79.4	80.2	87.5	78.6	78.4
3fabH1			77.0	85.7	78.2	79.8	79.2	88.6	73.0	79.3
3fabH2				74.4	85.5	84.1	86.8	77.7	91.0	84.8
3fabL1					80.6	76.5	80.0	86.9	79.1	76.6
3fabL2						86.4	88.4	80.3	86.5	86.0
1fc1A1							87.7	80.4	85.0	86.2
1fc1A2								80.9	88.2	89.1
2fb4H1									78.2	79.7
2fb4H2										84.8
3hlaB0										

Table 3.1: Pairwise SSAP scores matrix for immunoglobulins fold. Values are above 75 showing the domains structures within these family. Names of immunoglobulines are given in PDB code.

\vec{a}_i : atomic coordinates for atom i

\vec{a}_j : atomic coordinates for atom j

Atoms were then shifted along the s vectors towards their new positions

And additional weight reflecting the conservation of the error associated vector (e_{ij}) is incorporated under:

$$\vec{A}_j^i = A_j + \frac{\vec{s}_{ij}}{e_j |j - i|^{\frac{s}{2}}} \quad (3.17)$$

The type of secondary structure defines the geometry of the protein: vectors from the C_β atoms of one amino acid to all other C_β of all amino acids of the other protein. The resulting geometry in both compared proteins should be similar. The local environment of a given amino acid of the first protein is compared with the local environment of the corresponding amino acids in the second protein. The goal is to match residues by comparing these structural environments.

A scoring matrix is then derived and the highest scoring region is the one that defines the optimal structural alignment of the two proteins. So each sequential pairs of amino acids are compared and residue selection is implemented in order to increase the accuracy of the method by reducing possible noise in the score matrix and also for a higher computational speed. Those residues must be the ones having similar buried areas and torsion angles (default value of 150).

In the method the structural environment of every residue is placed and compared taking the residues as set of vectors from C_β of one amino acid to the C_β atoms of all other amino acids of the other protein.

We can conclude there are two levels of dynamic programming when aligning environmental variables of each successive amino acid in two proteins structures:

1. Comparing residues environment between pairs residues

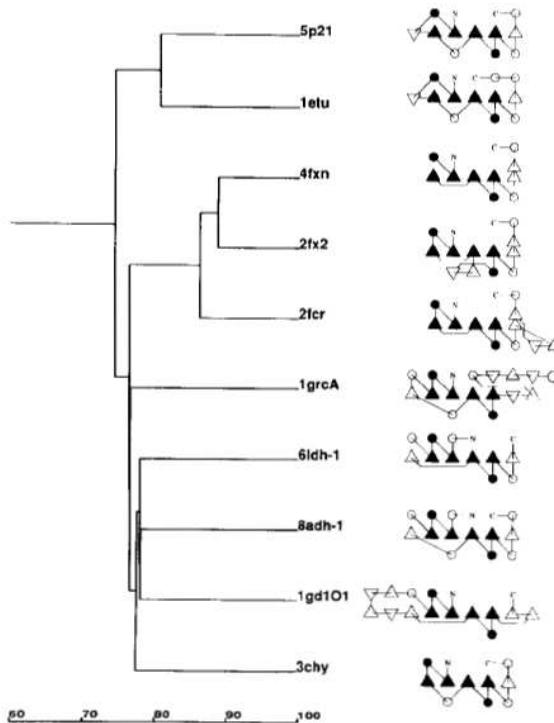


Figure 3.9: SSAP dendrogram. Structural relationship of immunoglobulins domains generated from the score on the above. Axis labeled with the SSAP scores from 0-100, PDB Brookhaven codes and TOPS representation (strands represented by triangles, helices represented by circles and lines penetrating these symbols designate at the front or behind secondary structure).

2. Obtaining an alignment from accumulated data on residue pairs.

This method can be applied to concatenate simple pair of alignment in order to get a multiple alignment (Multiple protein structure alignment) and has the advantage of being able to copy in a robust way with insertions and deletions between proteins.

The SSAP cut off

- Similarity of 70%: fold families 150
- Similarity of 70-80%: analogous folds with greater variations in loops and orientation of secondary structure
- Similarity of 80%: fold families 200
- Similarity of > 80: homologous fold so highly similar folds and related functions indicating a divergence from a common ancestor

It is used for clustering proteins in CATH DB.

```

2rhe00   : ESVLTOP PSASG  TPGQRVTISCTGSATDIG S NSVIWYQQVP GKAPKLLI
3fabL1   : XSVLIQP PSVSG  APGQRVTISCTGSSSNIG AGNHVKWYQQLP GTAPKLL
1cd800   : SQFRVSPLDRTW NLGETVELKCQVL LSN PTSGCSWLFQPRGAAASPTFLL
3fabH1   : XVOLEQSG PGLV  RPSQTLSLTCTVGTSF DDYYSTWVRQPP GRGLEWIG
2fb4H1   : EVQLVQSG GGVV  QPGRSRLSCSSS GFIF SSYAMYWVRQAP GKGLEWVA
1fc1A1   : PSVF LFPPKVKDTLMISRTPEVTCVVVDVSHEDPOVKENWYVD GVQVH
3fabL2   : QPKAAPSVTLFPPSSEELQA NKATLVCLISDFY PGAVTVAWKAD SSPV
1fc1A2   : EPQVYTLPPSREEMTK NQVSLTCLVKGFY PSDIAVEWESEN GQPE
3hlaB0   : IQRTPKIQVYSRHP AENG KSNFLNCYVSGFH PSDIEVDLLKN GERI
3fabH2   : PLAPSSKSTSG GTAALGCLVKDYF PEPVTWVN SGAL
2fb4H2   : PLAPSSKSTSG GTAALGCLVKDYF PQPVTVSWN SGAL
score    : +**##@#+*##*#*+*: :##:#@#/#@##*#*+: :*##:#@#/#*-* - +**##:#*:
           EE          EEEEEEE  EEEEEEE  EE
           A          B          C          D

2rhe00   : YYN      DLLPSGV DRFSASKS      GT SASLAISGLESEDEADYYCAAWN
3fabL1   :         IFHNN  ARESVSKS      GS SATILATGLOAEDEADYYCQS YD
1cd800   : YLS      QNKPKAAEGLDTQRFSGKRL  GDTFVLTLSDFRRNEGYFYFCALS
3fabH1   : YVFYHG TSDTD TPLRSRVTMLVNTS  KNOFSLRISSVTAADTA VYYCARNL
2fb4H1   : IIWDDGSDQHYA DSVKGRTFTISRND S KNTLFLOMDSLRPEDTGVYFCARDG
1fc1A1   : N        AKTKPR EQQY NSTYRVVSVLTVLHQNWLDGKEYKCKVSN
3fabL2   : K        AGVETTIPSKQSNKYAASSYLSITPEQWKSHKSYSQVTH
1fc1A2   : N        NYKTTTPVLDSDGSFFLYSKLTVDKSRWQQGNVFSCSVMH
3hlaB0   : E        KVEHSDSLFSKDWSFYLLYYTEFT PTEKDEYACRVNH
3fabH2   : T        SGVHTFPAVLQSSGGLYSLSSVVTVPSSLGT QTYICCNVNH
2fb4H2   : T        SGVHTFPAVLQSSGGLYSLSSVVTVPSSLGT QTYICCNVNH
score    : *--++: :+++:****:##*##*##*#**+*****##*##*#**#*##*#*##*#
           EEE      EE       E  EEEEEEE  EEEEEEE
           E          F          G

2rhe00   : DSDL      EPGFGGGTKLTVLGQPK
3fabL1   : R S       LRVFGGGTKLTVLR
1cd800   : NS        IMYFSHFBVFLPA
3fabH1   : I         AGC IDVWGQGSLSVTVSS
2fb4H1   : GHGFbSSASbFGPDYWGQGTPTVSSASTKGPSVF
1fc1A1   : KAL       PA PIEK TISKAK
3fabL2   : EGS       TVEK TVAPTECS
1fc1A2   : EAL       HNHYTQK SLSL
3hlaB0   : VTL       SQ PKIV KWDRDM
3fabH2   : KPS       N TKVDK KVEPKSC
2fb4H2   : KPS       N TKVDK RVEPKSC
score    : *--#*##*#*##*#*+- .
           EE
           H

```

Figure 3.10: Multiple sequence alignment derived from pairwise simple alignment concatenation generated by SSAP. β -Strands constant domains of immunoglobulins are shown (A, B, C, D, E, F, G, H) and residues in equivalent secondary structure regions derived from hydrogen bonds patterns are highlighted. The residues with the highest conserved structural environments are cysteines (C) due to their role holding the sheets (strands B and G) and tryptophan residue (W) because it is the one against they pack (strand C).

PDB code	Title	SSAP	Equivalent
1p11 452	Anthranilate isomerase	86.48 (76.9)	157
5timA 249	Triosephosphate isomerase	85.74 (100)	157
1wsyA 246	Tryptophan synthase	84.58 (68.8)	157
1ald 363	Aldolase A	84.25 (77.8)	157
5rubA 434	Rubisco	77.36 (68.5)	155
4enl 436	Enolase	75.75 (65.9)	141
2taaA 478	Taka-amylase	74.35 (62.9)	128
1ximA 392	Xylose isomerase	73.78 (70.8)	122
1dri 271	D-Ribose binding protein	69.76 (62.1)	139
1cseE 274	Subtilisin Carlsberg	69.23 (61.5)	122
2cmd 312	Malate dehydrogenase	68.78 (58.7)	133
2liv 344	Leucine binding protein	68.12 (60.6)	148
3grs 461	Glutathione reductase	66.53 (59.6)	133
1ldb 291	Lactate dehydrogenase	66.51 (59.9)	120
5p21 166	Ras p21 protein	65.85 (68.3)	122

Table 3.2: Pairwise SSAP scores obtained using a representative TIM structure consensus as template. The algorithm generates an overall normalized score in the range of 1-100 independently of the size of the proteins that are compared. Eight of the fifteen enzymes show high score values, from 70.8 for 1ximA 392 to 76.9 for 1p11 452 , meaning the TIM barrels folds are detected and as argued in the text above, related fold have more variation on the loops and orientation of secondary structures. Information regarding conformational preferences, functional significance, topological features such as solvent accessibility, residue preference, salt bridges and sequences similarity can be obtained by analyzing and searching for TIM-barrel folds

ALGORITHM	DESCRIPTION	FOCUS	STATISTICAL ANALYSIS	ADVANTAGES	DRAWBACKS
DALI	^a Distance Matrix Alignment	Complete sequence. Distances between all C _α atoms	Score derived from all against all comparisons. Z-score as the number of standard deviations from the average score derived from the DB distribution.	One single frame of representation. Speed of execution Ability to recognize distant relationships	Not algorithm for direct alignment. Statistical significance based on rmsd value which is considered suboptimal Non topological regions are not detected
CE	^b Combinatorial Extension of the Optimum Path	Distance between C _α of octameric fragments (combinatorial properties)	Tabulation of rmsd of the distributions of both proteins Z-score as result of combination of both z-scores	Computational Speed High percentage of homology detection	Reduction in the accuracy Domains miss recognition "Non topological" recognition or detection
SSAP	^c Sequential Structure Alignment Program	Domain level Intraprotein C _B -C _B vectors comparison	Scores compared against CATH database.	Dealing with internal domains Generation of multiple alignment by pairs alignments concatenation	Global alignment is missed once the whole structure is broken down into small SSEs Lost of details when just C _B -C _B are compared
VAST	^d Vector Alignment Search Tool	SSEs Vectors comparisons	P-value calculated for the best substructure superposition as if randomly obtained multiplied of alternative substructure alignments possible.	Computational time saved: SSEs converted into vectors	The whole 3D structure can not be used but just the predefined SSEs. Not complete SSEs C _α coordinates represented but just the beginning and the ends
SARF2	^e Spatial Arrangement of Backbone Fragments	Superimposable SSEs comparing typical α helix and β strands templates	Score as a function of rmsd and the number of matched C _α atoms Comparison of scores obtained from non redundant set of structures	Computational time saved: SSEs converted into vectors Difficult cases detection	The whole 3D structure can not be used but just the predefined SSEs. Not complete SSEs C _α coordinates represented but just the beginning and the ends
COMPARER	^f Comparer	Comparison of residues properties and relationships	Two scores E and A are contrasted, residues equivalences and gaps penalties respectively	Residues properties and relationships and segments relationships are studied at once	DP NOT applicable to relationships due to the dependence of the scores for a given relationship on the assignment of other relationships

Table 3.3: Methods for structural comparison and alignment. The web resources associated to these methods are of two types depending on the possible direct use of the method when comparing two proteins or an indirect way provided by a database of precalculated comparisons against all or a subset of PDB. a) Holm and Sander (1993a). b) Shindyalov and Bourne (1998). c) Taylor and Orengo (1989). d) Gibrat et al (1996). f) Mizuguchi et al (1998). g) Sali and Blundell (1990)SSEs Secondary Structure Elements

3.3 Conclusion

As seen along the whole text, significant differences in structural alignments do exist and one of the possible reasons to explain these differences could be due to the NP-hard problem nature. Since the amount of available information is higher and computer skills faster and more robust, structural comparison and alignment can be treated when are used with other methodologies as methods attempting to understand biological function and/or to provide a putative potential function. Results can be also used to perform molecular biology techniques as directed mutagenesis, blocking, knock out, monitoring enzyme activity and so on. The similarity comparison between two 3D structures can be defined as the superposition of the atoms being forming the structures following by a calculation of an optimal RMSD. Superposition is understood as, given correspondences, computing both the optimal alignment transformation and alignment score. Under this context, alignment is the necessary step to find such correspondences previous the superposition process. This process attempts to find the best matches between proteins in terms of residues position, geometry, and side chains contacts. Besides the fact itself, the goal also includes to figure out which biological function is performed and when possible which applications within medical and pharmaceutical field can be used. Even though it is still necessary to utilize hand alignment in order to achieve the closest and real biological functionality of target proteins. Features that are mainly compared can be summarized in just three: distances between coordinates of C_α - C_α , between coordinates of C_β - C_β and between secondary structure elements. The way to compare them, besides the mathematical statements of every particular method, is by moving the set of atoms or the chosen secondary structure, that does represent one rigid body over the other and looking for similar residues or context (VAST or COMPARER). Some methods filter the features to be compared and just keep the specific SSEs without loops or connectors or non topological regions therefore detecting just local similarities but not the whole structure comparison. This problem can be solved by combining the local and global criteria defining first a list of equivalent positions in the two structures (dynamic programming, distance matrices, etc...) and then by optimizing those equivalences (annealing, combinatorial extension of the optimal path, Monte Carlo alignment, etc).

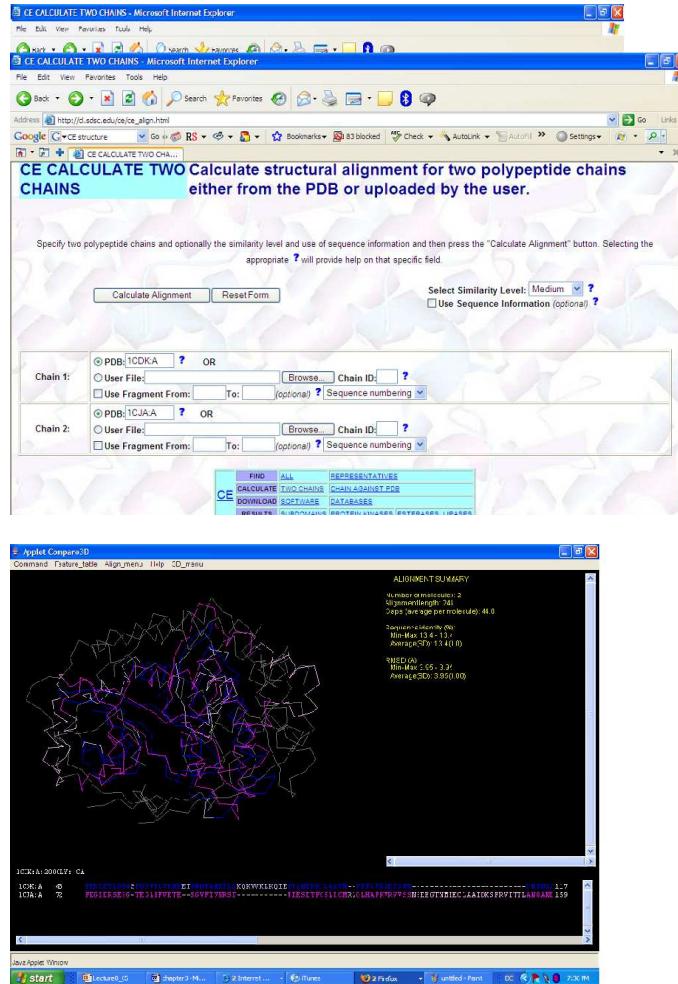
Evolutionary trees can be built up from analysis of proteins structures and family classifications based on molecular data from related molecules in different species. Thus, comparison methods can also perform phylogenetic approaches. As the databases are growth the comparison power also increases as well as advances in computer algorithms. Those advances make possible these algorithms to progressively approximate biology.

3.4 Exercises

CE to align cAMP-dependent protein kinase (1CDK: A) and an actin-fragmin kinase (1CJA: A)

1. Go to <http://c1.sdsc.edu>
2. Choose two chains structural alignment comparison and write down the PDB proteins codes into the boxes
3. Leave the default value for Similarity Level (Medium corresponding to heuristic D1)

4. Play around with the results both PDB file and 3D representation trying to find the gaps, the similarity sequences of both proteins



5. Compare the chain of 1CDK:A against the whole PDB with the default criteria and the following trying to figure out the differences

- Z-score: > 3.5
- Length Difference: < 20%
- Sequence Identity: > 70%
- RMSD: < 4.0 Å
- Gaps: < 30% S

3.5 Concepts

“Non topological” similarities: The order if polypeptide fragments in the structure alignment does not follow their order in the sequence.

Monte Carlo: Used for simulating the behavior of physical and mathematical systems. It differs from other simulation methods like molecular dynamics by being stochastic so non deterministic and using random numbers. There is no specification of which points within a set of possible choices will be taken. The method is used to find solutions to mathematical problems which have many variables. It is used also for modeling phenomena with significant uncertainty in inputs.

HMMs: Statistical model in which the system to be modeled has unknown parameters and the goal is to determine the hidden parameters from the observable parameters. The extracted parameters can then be used to perform further analysis as pattern recognition application. It could be considered as the simplest dynamic Bayesian Network. The state is not directly visible but variables influences by the state are visible. Each state has a probability distribution over the output. **Regular Markov Models:** The state is directly visible to the observer so the state transition probabilities are the only parameters. **Graph theory:** The study of graphs, mathematical structures used to model pairwise relations between objects from a certain collection.

Graph: (in the above context) refers to a collection of vertices and a collection of edges that connect pairs of vertices. Set of objects called points, nodes, or vertices connected by links called lines or edges. A graph may be undirected, meaning that there is no distinction between the two vertices associated with each edge, or its edges may be directed from one vertex to another.

Gibbs sampling: is an algorithm to generate a sequence of samples from the joint probability distribution of two or more random variables. The purpose of such a sequence is to approximate the joint distribution, or to compute an integral (such as an expected value). Gibbs sampling is an example of a Markov chain Monte Carlo algorithm.

Maximum clique problem approximation: The corresponding optimization problem to find the largest clique in a graph, including this clique in the NP problems. A clique in a graph is a set of pairwise adjacent vertices, or in other words, an induced subgraph which is a complete graph. Then, the clique problem is the problem of determining whether a graph contains a clique of at least a given size k . Once we have located k or more vertices which form a clique, it's trivial to verify that they do, which is why the clique problem is in NP.

Chapter 4

Protein Secondary Structure Prediction

4.1 Introduction

If a protein folds, first the secondary structures – especially the α -helix – are formed and build the major blocks for the 3D structure. That means the most local information in the primary sequence about the structure is given in the secondary structure. The prediction of secondary structure is important for

- design of de novo proteins, where the secondary structure sequences can be used as building blocks
- homology detection which can be enhanced [Ding and Dubchak, 2001]
- model building methods (e.g. “Modeller”) which rely on secondary structure prediction
- determining structures with 2D NMR
- first step of ab initio structure prediction

However it was found that also the secondary structure is influenced by 3D interactions [Ceronia et al., 2005]. For example, amino acid subsequences forming a β -sheet were cut out of a protein and inserted at another place where the sequence, surprisingly, formed a helix. Sometimes the exchange of single amino acids can change the secondary structure [Hofmann et al., 2004].

4.2 Assigning Secondary Structure to Measured Structures

Before predicting the secondary structure, training examples must be generated from known structures.

4.2.1 DSSP

The “Dictionary of Secondary Structure of Proteins” (DSSP) [Kabsch and Sander, 1983a] (<http://swift.cmbi.ru.nl/gv/dssp/>) assigns sheet and helical structures based on backbone-backbone

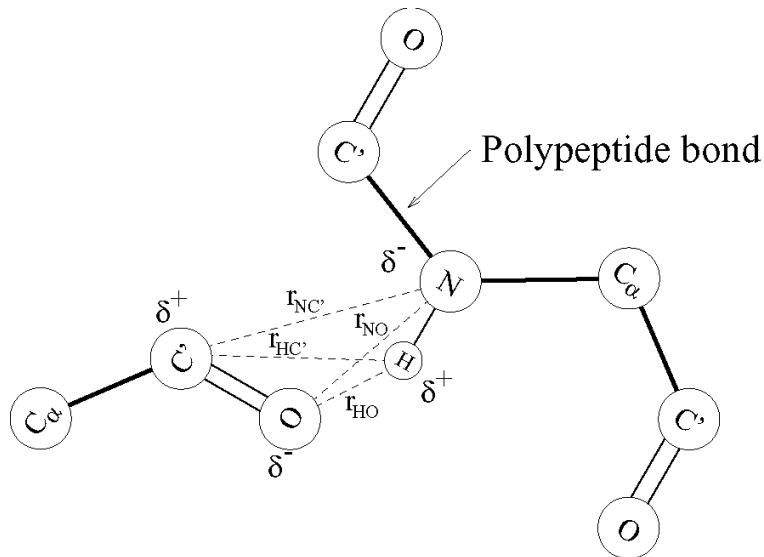


Figure 4.1: Distances used to compute the DSSP Coulomb hydrogen bond according to eq. (4.1). Copyright © 2003 John Wiley & Sons, Inc. from C.A.F. Andersen and B. Rost “Secondary structure assignment” in [Bourne and Weissig, 2003].

hydrogen bonds. A hydrogen bond is defined to be present if the bond energy is below -0.5 kcal/mol from a Coulomb approximation of the hydrogen bond energy according to eq. (4.1) where the distances are depicted in Fig. 4.1.

$$E = f \delta^+ \delta^- \left(\frac{1}{r_{NO}} + \frac{1}{r_{HC'}} + \frac{1}{r_{HO}} + \frac{1}{r_{NC'}} \right), \quad (4.1)$$

where $f = 332 \text{ \AA kcal}/e^2$ is a normalizing constant, δ^+ and δ^- are the polar charges given in electron charges e , and the distances are depicted in Fig. 4.1.

The structure is given to get unbroken structures and overlap is solved by giving α -Helices priority.

An α -helix is indicated by “H” if two consecutive amino acids have a $i \rightarrow (i + 4)$ hydrogen bonds and ends with two consecutive $(i - 4) \rightarrow i$ hydrogen bonds. An 3_{10} -helix is indicated by “G” with $i \rightarrow (i + 3)$ hydrogen bonds and a π -helix is indicated by “T” with $i \rightarrow (i + 5)$ hydrogen bonds.

Single helix hydrogen bonds are judged as turns and indicated by “T”.

β -sheets are indicated by “E” and have either two hydrogen bonds or are surrounded by hydrogen bonds.

Single amino acids with hydrogen bonds are labeled as β -bridge and are indicated by “B” (thus β -sheets consist at least of two “E”).

“S” indicates a bend in the chain of amino acids and a space an unassigned amino acid.

Symbol	Meaning
H	α -helix
G	3_{10} -helix
I	π -helix
T	turn
E	β -sheet
B	β -bridge
S	bend
—	unassigned

Table 4.1: The secondary structure symbols assigned by DSSP.

Symbol	conventional	newer
H	H	H
G	H	C
I	H	C
T	C	C
E	E	E
B	E	C
S	C	C
—	C	C

Table 4.2: The 8 secondary classes mapped to 3 classes. The second column gives the conventional mapping and the last column a newer kind of mapping which yields higher prediction accuracy.

Table 4.1 gives an overview over secondary structure symbols assigned by DSSP.

DSSP is sort of standard for secondary structure assignment.

Fig. 4.2 gives an example for the output of the DSSP program.

The 8 secondary structure classes of DSSP are often mapped to 3 classes, which is shown in Tab. 4.2.

PDB:lcrn	#	RESIDUE	AA	STRUCTURE	BP1	BP2	ACC	N-H-->O	O-->H-N	N-H-->O	O-->H-N
...	15	15	V	H >< S+	0	0	99	-4,-1.7	3,-1.3	2,-0.2	-2,-0.2
16	16	C	H	3<>S+	0	0	18	-4,-2.5	5,-0.8	1,-0.3	-2,-0.2
17	17	R	H	><5S+	0	0	94	-4,-2.0	3,-1.6	1,-0.2	-1,-0.3
18	18	L	T	<<5S+	0	0	144	-3,-1.3	-1,-0.2	-4,-0.6	-2,-0.2
19	19	P	T	3 5S-	0	0	107	0, 0.0	-1,-0.3	0, 0.0	-2,-0.1
20	20	G	T	< 5 +	0	0	53	-3,-1.6	-3,-0.2	1,-0.2	-2,-0.1
21	21	T		< -	0	0	37	-5,-0.8	-1,-0.2	1,-0.1	5,-0.1
22	22	P		>> -	0	0	81	0, 0.0	4,-2.2	0, 0.0	3,-0.7
23	23	E	H	3> S+	0	0	70	1,-0.2	4,-2.5	2,-0.2	5,-0.1
24	24	A	H	3> S+	0	0	63	1,-0.2	4,-1.7	2,-0.2	-1,-0.2
25	25	I	H	<> S+	0	0	99	-3,-0.7	4,-1.8	2,-0.2	-1,-0.2
26	26	C	H	X S+	0	0	0	-4,-2.2	4,-1.9	2,-0.2	6,-0.4
27	27	A	H	X S+	0	0	12	-4,-2.5	4,-2.7	-5,-0.2	5,-0.5
28	28	T	H	< S+	0	0	120	-4,-1.7	-1,-0.2	1,-0.2	-2,-0.2
29	29	Y	H	< S+	0	0	176	-4,-1.8	-1,-0.2	-5,-0.2	-2,-0.2
30	30	T	H	< S-	0	0	24	-4,-1.9	-2,-0.2	-3,-0.2	-3,-0.2
31	31	G	S	< S+	0	0	35	-4,-2.7	-3,-0.2	1,-0.4	-4,-0.1
32	32	b		-	0	0	5	-5,-0.5	-1,-0.4	-6,-0.4	2,-0.3
33	33	I	E	-A	3	0A	51	-30,-2.8	-30,-2.4	-3,-0.1	2,-0.5
34	34	I	E	-A	2	0A	78	-2,-0.3	-32,-0.2	-32,-0.2	3, 0.0
...											

The diagram illustrates how specific sections of the DSSP output relate to structural features. It shows arrows originating from the 'secondary structure synopsis' section at the bottom of the table and pointing to various annotations. These annotations include:

- beta-sheet label**: Points to the first two residues (V and H) which are both labeled as beta-sheets.
- beta-bridge 2 label**: Points to residue 16, which is part of a beta-bridge.
- beta-bridge 1 label**: Points to residue 21, which is also part of a beta-bridge.
- chirality assignments**: Points to the 'E' and 'A' residues, which represent chiral centers.
- bend assignments**: Points to the 'T' and 'G' residues, which are involved in bends.
- pi-helix hydrogen bonds**: Points to the 'I' residue, which is part of a pi-helix.
- alpha-helix hydrogen bonds**: Points to the 'C' residue, which is part of an alpha-helix.
- 3_10-helix hydrogen bonds**: Points to the 'P' residue, which is part of a 3_10-helix.
- secondary structure synopsis**: Points back to the 'secondary structure synopsis' section at the bottom of the table.

Figure 4.2: Output example for DSSP. Copyright © 2003 John Wiley & Sons, Inc. from C.A.F. Andersen and B. Rost “Secondary structure assignment” in [Bourne and Weissig, 2003].

The output from DSSP:

```

HEADER      HYDROLASE   (SERINE PROTEINASE)          17-MAY-76   1EST
...
240 1 4 4 0 TOTAL NUMBER OF RESIDUES, NUMBER OF CHAINS,
NUMBER OF SS-BRIDGES(TOTAL,INTRACHAIN,INTERCHAIN) .
10891.0  ACCESSIBLE SURFACE OF PROTEIN (ANGSTROM**2)
162 67.5  TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(J) ; PER 100 RESIDUES
0 0.0    TOTAL NUMBER OF HYDROGEN BONDS IN      PARALLEL BRIDGES; PER 100 RESIDUES
84 35.0   TOTAL NUMBER OF HYDROGEN BONDS IN ANTI PARALLEL BRIDGES; PER 100 RESIDUES
...
26 10.8  TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+2)
30 12.5  TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+3)
10 4.2   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+4)
...
# RESIDUE AA STRUCTURE BP1 BP2 ACC   N-H-->O  O-->H-N  N-H-->O  O-->H-N
2   17   V   B 3   +A  182   OA   8   180,-2.5 180,-1.9  1,-0.2 134,-0.1

                           TCO  KAPPA ALPHA PHI   PSI    X-CA   Y-CA   Z-CA
                           -0.776 360.0   8.1 -84.5 125.5 -14.7  34.4  34.8

....;....1....;....2....;....3....;....4....;....5....;....6....;....7..
--- sequential resnumber, including chain breaks as extra residues
|   --- original PDB resname, not nec. sequential, may contain letters
|   |   --- amino acid sequence in one letter code
|   |   |   --- secondary structure summary based on columns 19-38
|   |   |   |   xxxxxxxxxxxxxxxxxxxx recommend columns for secstruc details
|   |   |   |   .--- 3-turns/helix
|   |   |   |   .--- 4-turns/helix
|   |   |   |   ||.--- 5-turns/helix
|   |   |   |   |||.--- geometrical bend
|   |   |   |   ||||.--- chirality
|   |   |   |   |||||.--- beta bridge label
|   |   |   |   |||||.--- beta bridge label
|   |   |   |   ||||||.--- beta bridge partner resnum
|   |   |   |   ||||||| .--- beta bridge partner resnum
|   |   |   |   ||||||| .--- beta sheet label
|   |   |   |   |||| .--- solvent accessibility
|   |   |   |   |||| |
# RESIDUE AA STRUCTURE BP1 BP2 ACC
|   |   |   |   ||||| |
35  47   I   E   +   0   0   2
36  48   R   E > S-K  0   39C  97
37  49   Q   T 3  S+   0   0   86   (example from 1EST)
38  50   N   T 3  S+   0   0   34
39  51   W   E < -KL  36  98C   6

```

The number 2 under column “8” in line “residues per alpha helix” means: there are 2 alpha helices of length 8 residues in this data set.

- “# RESIDUE”: two columns of residue numbers.
- “AA”: one letter amino acid code, lower case for SS-bridge CYS.
- “S” (first column in STRUCTURE block): compromise summary of secondary structure,

- “BP1 BP2”: residue number of first and second bridge partner followed by one letter sheet label
- “ACC”: number of water molecules in contact with this residue *10. or residue water exposed surface in Angstrom².
- “N-H->O” etc.: hydrogen bonds; e.g. -3,-1.4 means: if this residue is residue i then N-H of I is h-bonded to C=O of I-3 with an electrostatic H-bond energy of -1.4 kcal/mol. There are two columns for each type of H-bond, to allow for bifurcated H-bonds.
- “TCO”: cosine of angle between C=O of residue I and C=O of residue I-1. For alpha-helices, TCO is near +1, for beta-sheets TCO is near -1. Not used for structure definition.
- “KAPPA”: virtual bond angle (bend angle) defined by the three C-alpha atoms of residues I-2,I,I+2. Used to define bend (structure code ‘S’).
- “ALPHA”: virtual torsion angle (dihedral angle) defined by the four C-alpha atoms of residues I-1,I,I+1,I+2.Used to define chirality (structure code ‘+’ or ‘-’).
- “PHI PSI”: IUPAC peptide backbone torsion angles
- “X-CA Y-CA Z-CA”: echo of C-alpha atom coordinates

4.2.2 STRIDE

STRuctural IDEntification method (STRIDE, http://www.embl-heidelberg.de/argos/stride/stride_info.html) [Frishman and Argos, 1995] utilizes an empirical hydrogen bond energy according to eq. (4.2).

$$\begin{aligned}
 E &= E_r E_t E_p && (4.2) \\
 E_r &= -2.8 \text{ kcal/mol} \left(\frac{43^6}{r_{\text{NO}}^6} - \frac{33^8}{r_{\text{NO}}^8} \right) \\
 E_p &= \cos^2 \theta \\
 E_t &= \begin{cases} (0.9 + 0.1 \sin(2 t_i)) \cos t_o & \text{for } 0^\circ < t_i \leq 90^\circ \\ K_1 (K_2 - \cos^2(t_i)) \cos t_o & \text{for } 90^\circ < t_i \leq 110^\circ \\ 0 & \text{for } 110^\circ < t_i \end{cases},
 \end{aligned}$$

where the θ , t_i , and t_o are given in Fig. 4.3. The energy E_r is similar to the Lennard-Jones potential and includes the optimal distances of 3 Å for the backbone hydrogen bond.

STRIDE uses additionally to the energy eq. (4.2) ϕ - ψ (phi-psi) torsion angles which are determined by α -helix and β -sheet propensities according to their distance to their typical regions in Ramachandran plots.

Often STRIDE assignments agree better with the expert assignments than DSSP.

STRIDE terminates α -helices according to ϕ - ψ angles.

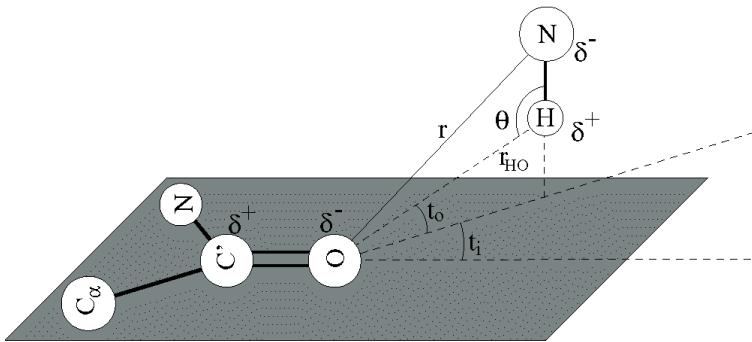


Figure 4.3: Distances used to compute the STRIDE hydrogen bond according to eq. (4.2). Copyright © 2003 John Wiley & Sons, Inc. from C.A.F. Andersen and B. Rost “Secondary structure assignment” in [Bourne and Weissig, 2003].

Symbol	Meaning
H	α -helix
G	3_{10} -helix
I	π -helix
T	turn
E	β -sheet
B	β -bridge
C	unassigned

Table 4.3: The secondary structure symbols assigned by STRIDE.

In contrast to DSSP, STRIDE does not assign the “S” symbol and classifies all unassigned amino acids to the coiled class “C”.

Table 4.3 gives an overview over secondary structure symbols assigned by STRIDE.

Fig. 4.4 gives an example for the STRIDE output.

4.2.3 DEFINE and P-Curve

DEFINE [Richards and Kundrot, 1988] performs an C_α -coordinate match with the optimal secondary structure. Like BLAST in sequence alignment, first perfect matches are found and then elongated.

Problems appear with sheets which can bend and may have high curvature if they are long. Therefore also bended sheets are allowed as ideal structure.

P-Curve [Sklenar et al., 1989] uses differential geometry to calculate a helicoidal axis based on series of peptide planes. Assignments are made according to pattern matching.

```

PDB:lcrn
REM |---Residue---|   |--Structure--| |-Phi-| |-Psi-| |-Area-| 1CRN
.....
ASG  VAL -  15  15  H  AlphaHelix  -69.24  -41.22  93.8  1CRN
ASG  CYS -  16  16  H  AlphaHelix  -56.67  -36.00  18.4  1CRN
ASG  ARG -  17  17  H  AlphaHelix  -77.07  -16.13  94.1  1CRN
ASG  LEU -  18  18  H  AlphaHelix  -53.21  -46.17  143.0 1CRN
ASG  PRO -  19  19  C  Coil      -77.19  -7.60  108.9  1CRN
ASG  GLY -  20  20  C  Coil      106.26  7.31  52.1  1CRN
ASG  THR -  21  21  C  Coil      -52.67  136.34  38.4  1CRN
ASG  PRO -  22  22  C  Coil      -56.98  146.62  81.9  1CRN
ASG  GLU -  23  23  H  AlphaHelix -56.41  -36.19  68.9  1CRN
ASG  ALA -  24  24  H  AlphaHelix -63.43  -34.86  61.3  1CRN
ASG  ILE -  25  25  H  AlphaHelix -74.77  -37.89  98.2  1CRN
ASG  CYS -  26  26  H  AlphaHelix -64.95  -31.69  0.0  1CRN
ASG  ALA -  27  27  H  AlphaHelix -62.04  -54.03  11.6  1CRN
ASG  THR -  28  28  H  AlphaHelix -68.78  -25.49  121.1 1CRN
ASG  TYR -  29  29  H  AlphaHelix -67.59  -36.30  174.0 1CRN
ASG  THR -  30  30  H  AlphaHelix -108.96  -18.47  23.4  1CRN
ASG  GLY -  31  31  C  Coil      91.82  -3.07  36.1  1CRN
ASG  CYS -  32  32  C  Coil      -69.52  164.38  4.6  1CRN
ASG  ILE -  33  33  E  Strand    -129.76  157.03  51.0  1CRN
ASG  ILE -  34  34  E  Strand    -111.56  129.59  78.0  1CRN
.....

```

Figure 4.4: Output example for STRIDE. Copyright © 2003 John Wiley & Sons, Inc. from C.A.F. Andersen and B. Rost “Secondary structure assignment” in [Bourne and Weissig, 2003].

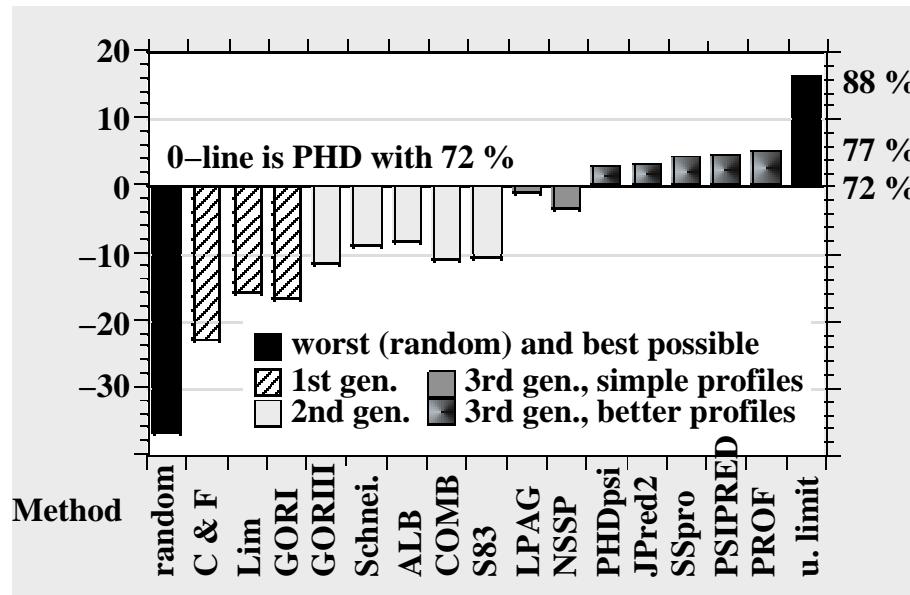


Figure 4.5: Comparison of different methods from [Rost, 2003b]. The compared methods are: “C+F” Chou & Fasman (1st generation) [Chou and Fasman, 1978]; Lim (1st) [Lim, 1974]; GORI (1st) [Garnier et al., 1978]; Schneider (2nd); ALB (2nd) [Ptitsyn and Finkelstein, 1983]; GORIII (2nd) [Gibrat et al., 1987]; COMBINE (2nd); S83 (2nd) [Kabsch and Sander, 1983b]; LPAG (3rd) [Levin et al., 1993]; NSSP (3rd) [Solovyev and Salamov, 1994]; PHDpsi (3rd) [Przybylski and Rost, 2001]; JPred2 (3rd) [Cuff and Barton, 2000]; SSPro (3rd) [Baldi et al., 1999]; PSIPRED (3rd) [Jones, 1999]; PROF (3rd) [Rost, 1996a]; PHD (3rd) [Rost and Sander, 1993, 1994, Rost, 1996b]. PHD served as baseline method.

4.3 Prediction of Secondary Structure

The protein secondary structure prediction methods can be divided into 3 generations according to Burkhard Rost:

- **1. Generation.** These methods use only *single residue statistics* and were developed in the 70's. The best performance in percentage of correct predicted positions was up to 60%. Methods include the Chou-Fasman approach, GORI, Lim's approach.
- **2. Generation.** These methods use a *window over the current position* were developed in the mid to end 80's. The best performance in percentage of correct predicted positions was up to 65%. Methods include GORIII, ALB, Schneider's approach,
- **3. Generation.** These methods use a profile or a position specific scoring matrix (PSSM) stemming from a multiple alignment and were developed in the early 90's. The best performance in percentage of correct predicted positions is up to 78%. Methods include PHD, Jpred2, SSPro and Porter, PSIPRED, PROF.

4.3.1 Chou-Fasman Method

The Chou-Fasman method [Chou and Fasman, 1978] computes the likelihood ratio of the joint distribution of amino acid a and structure s , where s is an α -helix, a β -sheet, or a turn. These likelihood ratios are called “propensities” P_α , P_β , and P_t :

$$P_s(a) = \frac{p(a, s)}{p(a) p(s)}. \quad (4.3)$$

$p(a, s)$ is estimated by the number of amino acid a in structure s divided by the number of all amino acids in the data base. $p(a)$ is estimated by the number of amino acid a divided by the number of all amino acids in the data base. $p(s)$ is estimated by the number of all amino acids in structure s divided by the number of all amino acids in the data base.

The secondary structure is assumed to start a nucleation for

- α -helices if four out of six residues have $P_\alpha > 1.03$.
- β -strands if three out of five residues have $P_\beta > 1.00$.

These nucleations are elongated in each direction until the mean propensity of four residues is below threshold. If both α -helices and β -strands are predicted the higher average propensity wins.

Turns are predicted based on four residues where the probability of a certain amino acid $p(a | t, i)$ at a certain position i in a turn is computed. The probabilities are multiplied (assuming independence) to obtain the joint probability of the four residues forming a turn. A turn is predicted if the first probability is larger than that of an α -helix and of a β -strand and if the second probability is larger than $7.5 \cdot 10^{-5}$.

This method reaches 50-60% of accuracy in predicting the secondary structure. Note that this is much higher than random guessing which would be correct in 1/3 of the prediction.

4.3.2 GOR Methods

The GOR (Garnier-Osguthorpe-Robson) methods are based on statistical principles.

The probability of a residue a participating at a certain secondary structure depends on the residue itself and on the neighboring residues,

From a data base of known structures a frequency matrix F^s 17 residue window is calculated for each secondary structure s . Let the frequency matrix be $F_{a,j}^s$ for amino acid a at the j th position, then

$$P_s(a_l) = \sum_{j=l-8}^{l+8} F_{a,j}^s. \quad (4.4)$$

The maximal value over the structures determines the predicted structure.

In [Garnier et al., 1996] the GOR method is explained by information theory.

The likelihood ratio was given in eq. (4.3) which can be estimated by

$$P_s(a) = \frac{p(s | a)}{p(s)} \approx \frac{f_{s,a}}{f_s}, \quad (4.5)$$

where $f_{s,a}$ is the fraction of amino acids of type a in structure s from all amino acids of type a and f_s is the fraction of amino acids in structure s from all amino acids.

The mutual information between residue a and structure s , that is how much information does a contain about s is

$$I(s, a) = H(s) + H(a) - H(s, a) \quad (4.6)$$

$$H(x) = - \sum_x p(x) \log p(x) \quad (4.7)$$

$$I(s, a) = \sum_{s,a} p(s, a) \log \frac{p(s, a)}{p(s) p(a)} = \mathbf{E}_{(s,a)} \log \frac{p(s, a)}{p(s) p(a)}. \quad (4.8)$$

For specific s and a the value

$$\log \frac{p(s, a)}{p(s) p(a)} = \log \frac{p(s | a)}{p(s)} \quad (4.9)$$

gives the local information.

A difference of the local information can be given as

$$\begin{aligned} I_{\text{loc}}(\Delta s; a) &= \log \frac{p(s | a)}{p(s)} - \log \frac{p(\neg s | a)}{p(\neg s)} = \\ &\log \frac{p(s | a)}{p(s)} - \log \frac{1 - p(s | a)}{1 - p(s)} = \\ &\log \frac{p(s | a)}{1 - p(s | a)} + \log \frac{1 - p(s)}{p(s)}. \end{aligned} \quad (4.10)$$

The $I_{\text{loc}}(\Delta s; a)$ can be estimated using

$$I_{\text{loc}}(\Delta s; a) = \log \frac{f_{s,a}}{1 - f_{s,a}} + \log \frac{1 - f_s}{f_s}. \quad (4.11)$$

Exponentiating gives

$$\frac{p(s | a)}{1 - p(s | a)} = \frac{p(s)}{1 - p(s)} \exp(-I_{\text{loc}}(\Delta s; a)). \quad (4.12)$$

Now the values $I_{\text{loc}}(\Delta s; a)$ can be estimated for the 17 positions. The probability ratios from eq. (4.12) can be computed and multiplied together under an independence assumption.

Here only the single position specific influence of amino acids on the secondary structure are considered.

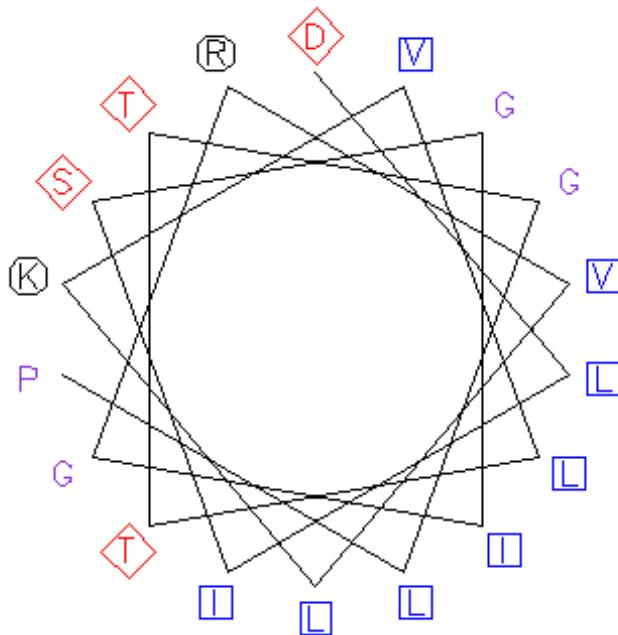


Figure 4.6: Helical wheel depicting the positions of amino acids in an α -helix. Aliphatic residues (blue squares), polar or negatively charged residues (red diamonds), and positively charged residues (black octagons).

In GORIII [Gibrat et al., 1987] the probability of the sequences was extended to consider also the probability of the structure conditioned on amino acid pairs, that is how to pairs of amino acids influence the structure.

To further extend this approach and assuming less independence is difficult. To include probabilities of structure s which are conditioned on n amino acids requires to estimate 20^n variables $p(s | a_1, \dots, a_n)$ – one for each combination of a_1, \dots, a_n . However the data sets are not large enough to estimate these values.

4.3.3 Lim's Method

The method of Lim [Lim, 1974] uses stereochemical rules for prediction the secondary structure.

For example the α -helix needs a hydrophobic side which faces internally of the protein.

This method uses advances biochemical insights and has high explanatory power. For example Fig. 4.6 and Fig. 4.7 show the hydrophobic side of a helix depicted in a helical wheel.

4.3.4 Neural Networks

In 1988 Qian and Sejnowski [Qian and Sejnowski, 1988] had excellent performance with a neural network derived from the The neural network of [Qian and Sejnowski, 1988] is shown in Fig. 4.9 was based on the NETTalk architecture (see Fig. 4.8) and achieved 64.3% accuracy.

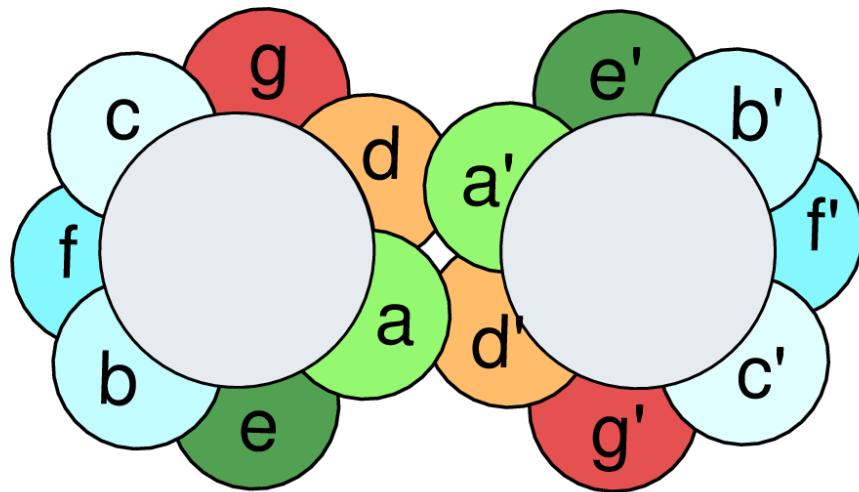


Figure 4.7: Helical wheel depicting for leucine zipper, a coiled-coil structure. The heptad repeat “a b c d e f g” corresponds to “H P P H P P P”, where “H” is hydrophobic and “P” is polar.

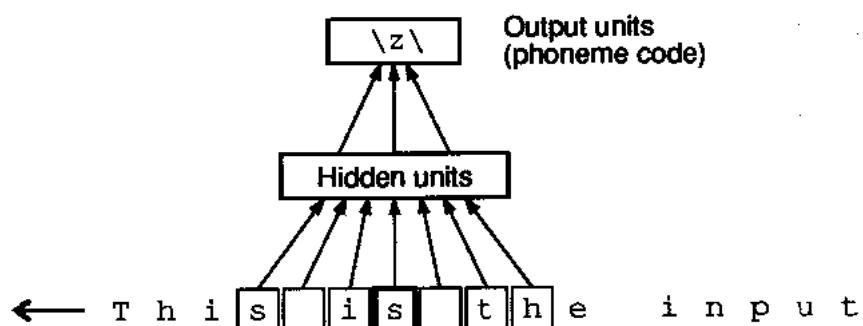


Figure 4.8: The NETTalk neural network architecture is depicted. A window scans the text and the network should predict the pronunciation of the letter in the center of the window. The same architecture was used for protein secondary structure prediction in [Qian and Sejnowski, 1988], where the input was a window of amino acids and the output the secondary structure of the center amino acid.

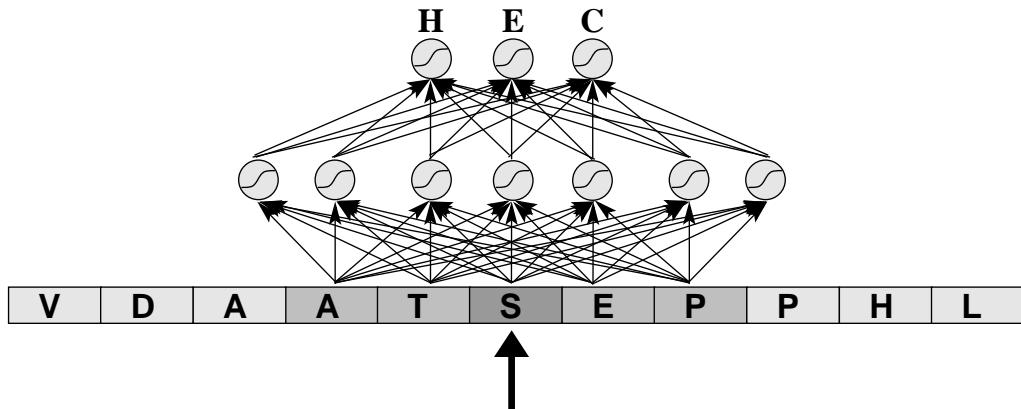


Figure 4.9: Neural network approach to secondary structure prediction (sequence-to-structure). The input is a window of the amino acid sequence over the position for which the structure has to be predicted.

4.3.5 PHD, PSIPRED, PREDATOR, JNet, JPred2, NSSP, SSPro

In 1993 with the “Profile network from Heidelberg” (PHD) [Rost and Sander, 1993] (see also [Rost and Sander, 1994, Rost, 1996b]) a breakthrough has been achieved. The accuracy jumped to 70.2%.

The most important novelty was to use profiles (multiple alignments) or position specific scoring matrices (PSSMs) instead of the primary sequence as input. Now the sequence is averaged over many sequences which are very similar to the query sequence. Helical or β -sheet regions are much better to recognize in the profile than in the primary sequence because deviations in the sequence are averaged over.

Another important fact is that implicitly long range information is included. Only if amino acids in the primary sequences which are far away from the current position are aligned in the profile then the current position is correctly aligned.

Alignment extracts homologs which have the same 3D structure. Alignment checks for suited interaction partners and therefore for possible long range information. For example, if a strand needs a certain partner strand to form a β -sheet then the alignment can detect the partner.

The alignment includes evolutionary information into the input. This is of advantage because the number of 3D structures is limited therefore alignment makes the relation to a known structure visible.

The PHD method works in 3 levels. First the sequence-to-structure network (see Fig. 4.9) predicts the secondary structure based on an input window which is now a profile. Then a structure-to-structure network as depicted in Fig. 4.10 predicts the secondary structure based on the predictions of the first network. Here sequences like “HHHHHCHHHH” can be corrected to “HHHHHHHH-HHH” or the length of β -sheets adjusted. In the final, 3rd level a voting procedure must be applied. Here it is possible to correct for biases like underestimating β -sheets. The 3 levels are depicted in Fig. 4.11.

The PHD method has an accuracy of 70.2% for profiles vs. 63% for primary sequence.

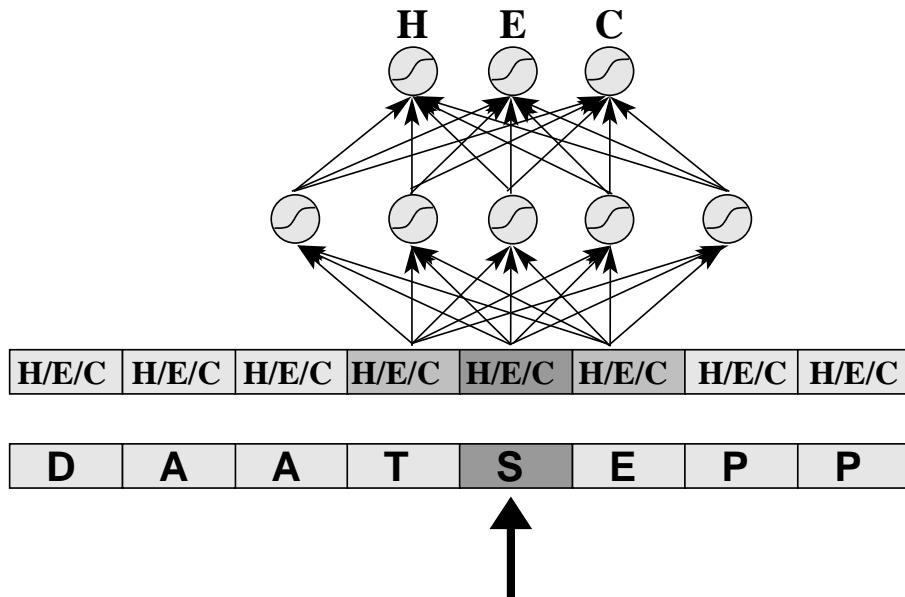


Figure 4.10: Neural network approach in the second level to secondary structure prediction (structure-to-structure). The input is a window of the predicted structure sequence over the position for which the structure has to be predicted.

The method PSIPRED [Jones, 1999] developed by David Jones is based on PSI-BLAST which is nowadays used in many secondary structure prediction methods.

Other methods followed the development of PHD like NSSP [Solovyev and Salamov, 1994], PHDpsi[Przybylski and Rost, 2001], JNet [Cuff and Barton, 2000], SSPro [Baldi et al., 1999], PROF [Rost, 1996b], and another PROF [Ouali and King, 2000]. JPred2 [Cuff and Barton, 2000] combines the results of other prediction methods.

It is interesting to note that SSPro is based on a recurrent neural network, but with a special architecture called “bi-directional recurrent neural network” (BRNN). For BRNN the amino acid sequence is scanned from left to the current position and from the right to the current position and at the current position both information are combined. The method “unfolding in time” for recurrent networks (see Bioinformatics II) can be used to unfold some units into the past and some units into the future and a top level combines both unfolded networks.

Currently the best performing method is an extension of SSPro which is called “Porter” [Pol-lastri and McLysaght, 2005]. Porter is reported to exceed 79% accuracy on large data sets on a 5-fold cross validation.

Fig. 4.12, Fig. 4.13, and Fig. 4.14 show comparisons of the different methods which are made in [Rost, 2003b].

Critic:

The danger is that through alignment an implicit structure classification is made. The local PSSM may be unique for a certain 3D structure. If the local PSSM of the query structure fits the local PSSM of a training structure then both structures are equal. The local PSSM is only equal if the same sequences are used for building the PSSM.

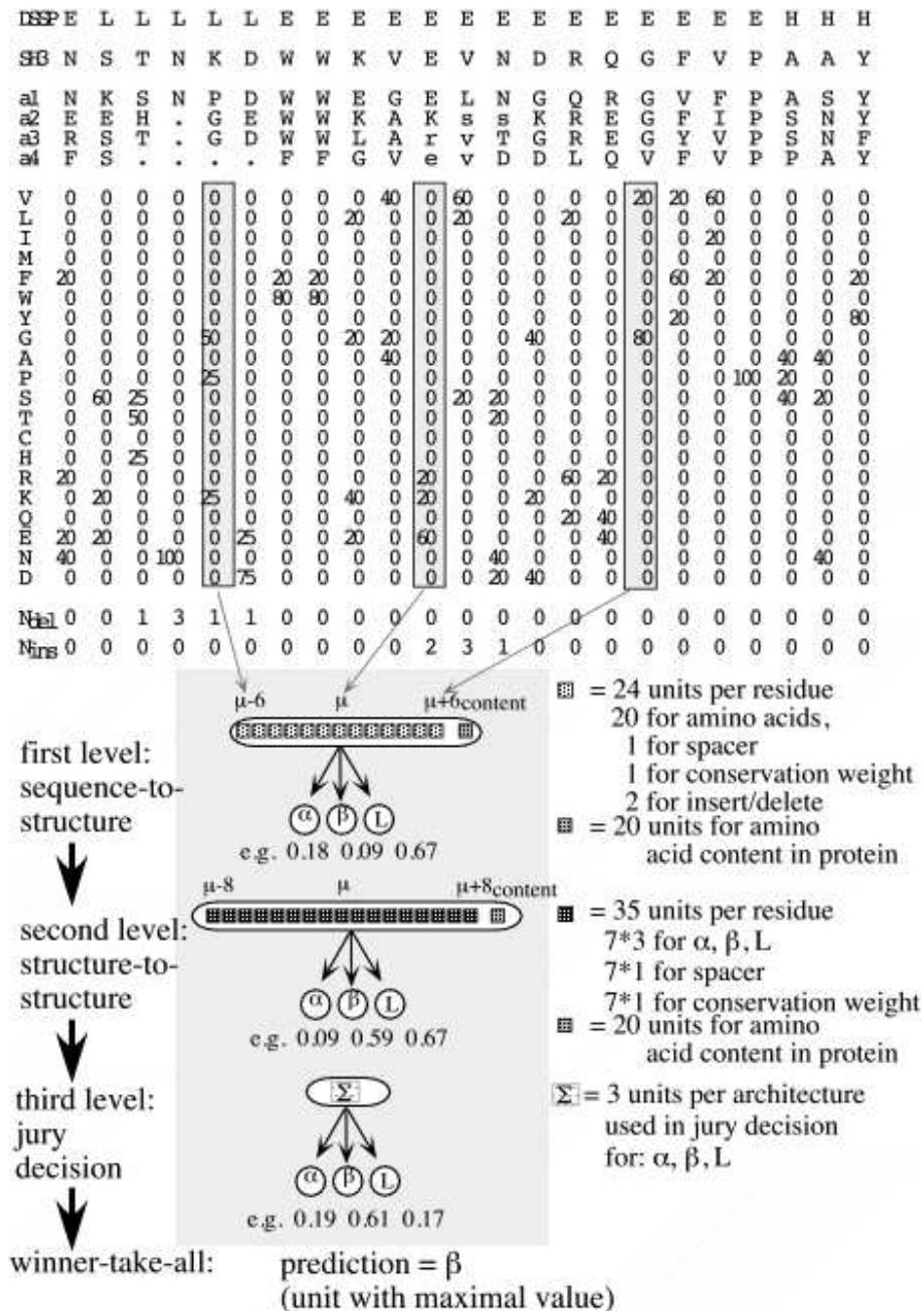


Figure 4.11: The PHD levels of prediction – taken from [Rost, 2003a].

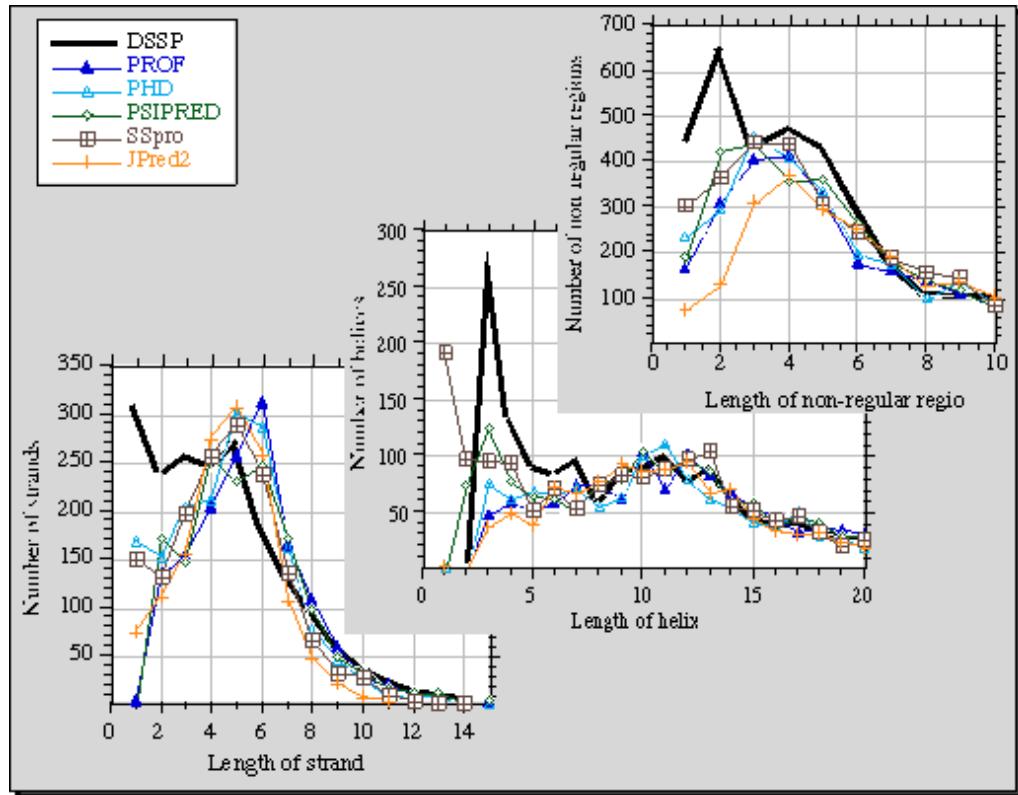


Figure 4.12: Comparison of different methods from [Rost, 2003b]. Distribution of segment length and predicted segment length.

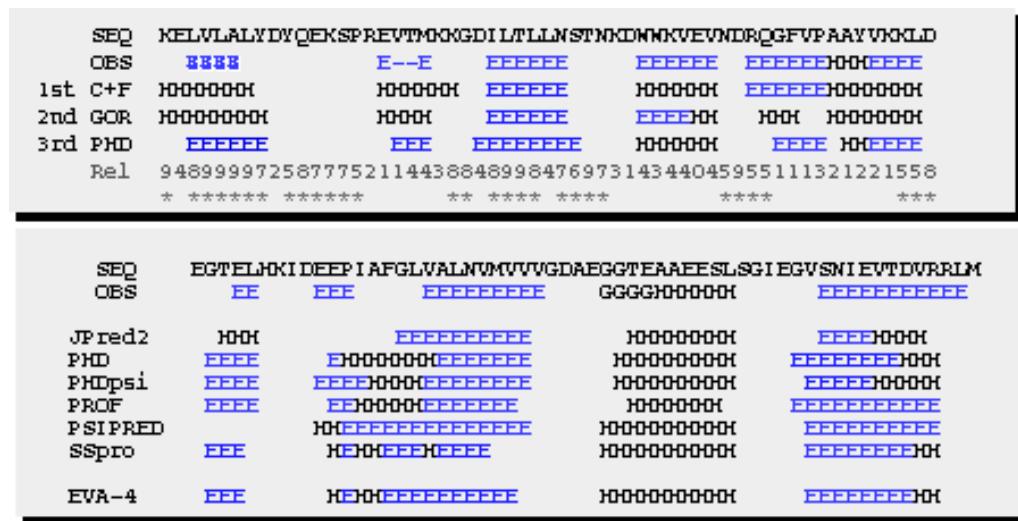


Figure 4.13: Comparison of different methods from [Rost, 2003b]. Secondary structure prediction for different methods from 1st generation to 3rd generation.

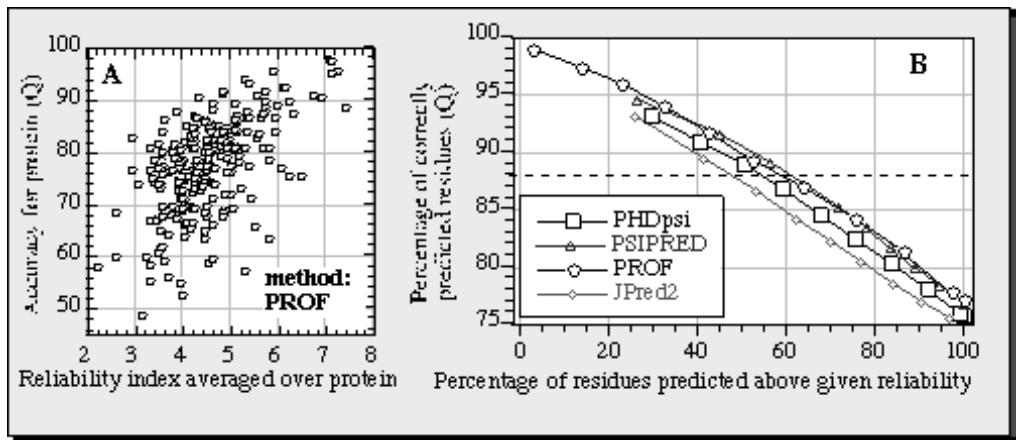


Figure 4.14: Comparison of different methods from [Rost, 2003b]. Reliability and accuracy are plotted against each other. If residues are predicted with higher reliability then the accuracy is higher.

That means the local PSSM may identify the correct structure (similar to profile-profile alignment) and the according SS of a training sequence with the same structure is used.

That would mean SS is 3D classification and thereafter back-projection to SS.

4.4 Evaluating Secondary Structure Prediction

The models must be constructed on training data and tested on test data. Both data must be based on known structures solved by x-ray crystallography or NMR. The data base storing this information is the Brookhaven Data Bank (PDB).

However it makes sense first to separate the proteins in blocks which fold separately and build separate structures – these blocks are called domains. Domain data bases which divide PDB into domains are for example SCOP and CATH.

4.4.1 Non-Homologous Test Sequences

In order to estimate the error on future data, the test error, in machine learning independent identical distributed (iid) data samples are assumed. However nature does not sample proteins iid in the space of possible proteins. It builds a new protein based on existing ones which means that new proteins depend on existing ones. Sometimes certain regions of the space of possible proteins are explored in more detail and other are not explored at all.

Another selection bias stems from the experimenter. Certain proteins are selected for resolving their 3D structure. This may be important proteins like drug targets and proteins which can be measured. For example membrane proteins are more difficult to measure. Also globular proteins are more often resolved.

If we sample the space of proteins iid: how many proteins with a certain sequence identity or with a certain similarity we would find? We are sure that we would not find as many as we will

Symbol	conventional	newer	with turn
H	H	H	H
G	H	C	C
I	H	C	C
T	C	C	T
E	E	E	E
B	E	C	C
S	C	C	C
-	C	C	C

Table 4.4: The 8 secondary classes of DSSP mapped to 3 or 4 classes. The second column gives the conventional mapping, second a recent mapping yielding higher accuracy, and the last column has an additional turn class

find if we randomly select proteins from PDB.

To correct for the non-iid sampling, the test set proteins which are similar to the training set must be removed from the test set. Typically a threshold between 30% to 40% of mutual identity is set to remove proteins from the test set. If a test sequence has higher identity than the threshold to a training sequence, then the test sequence is removed. However also the test set has to be corrected so that the identity between pairs of test sequences is below the threshold. The first case “train-test” corrects for the fact that a test sequence is only correctly predicted because a training sequence is memorized and the method would fail if another sequence appears. The second case “test-test” corrects for the fact that some sequences types are very often in the test and the performance on the test is governed by the performance of this sequence type. The threshold correction results in test examples in the space of proteins where no other protein (test or training) is present.

Whether the non-iid training set is of disadvantage for the method or not depends on the method. Some protein types may be over- or underestimated.

4.4.2 Secondary Structure Classes

First we have to consider how the secondary structure has been assigned to the measurement data. Was DSSP or STRIDE used? In general that should not make a large difference in the assignment except at the end of structural elements. Also the available classes are determined, e.g. DSSP has a bend class and STRIDE not.

If 3 classes H,E, and C are used, it must be considered how the mapping is done. In Tab. 4.4 the second and third column shows two possible mapping where the third column yields higher accuracy. The fourth column in Tab. 4.4 has an additional turn class which is useful for 3D structure inference.

true	predicted		total
	+1	-1	
+1	TP	FN	TP + FN
-1	FP	TN	FP + TN
total	TP + FP	FN + TN	N

Table 4.5: Confusion matrix. TP: true positive - positive correctly predicted; FN: false negative - positive incorrectly predicted; FP: false positive - negative incorrectly predicted; TN: true negative - negative correctly predicted.

4.4.3 Quality Measures

We consider a binary classification task with a positive class (+1) and a negative class (-1) with N test examples. To evaluate the prediction result of a method the predicted values as well as the true classes have to be known.

First we define: TP: true positive - positive correctly predicted; FN: false negative - positive incorrectly predicted; FP: false positive - negative incorrectly predicted; TN: true negative - negative correctly predicted. This is shown in the confusion matrix in Tab. 4.5.

With the definitions of Tab. 4.5 we can define

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{N} \quad (4.13)$$

$$\text{specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \quad (4.14)$$

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4.15)$$

$$\text{balanced error} = 0.5 (\text{specificity} + \text{sensitivity}) \quad (4.16)$$

$$\text{Matthews corr.} = \frac{\text{TP TN} - \text{FP FN}}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{FN} + \text{TN})(\text{FP} + \text{TN})}} \quad (4.17)$$

$$\text{weight of evidence} = \log \frac{\text{TP TN}}{\text{FP FN}} \quad (4.18)$$

Another measure is the area under the ROC curve. The ROC curve – short for Receiver Operating Characteristic curve – is a curve which plots sensitivity vs. (1 - specificity) for a binary classifier system as its discrimination threshold is varied. The ROC is also the plot of the fraction of true positives vs. the fraction of false positives.

ROC analysis allows to assess the quality of a binary classifier independently from the cost function or the class distribution.

Measures used in secondary structure prediction.

The accuracy given in percent over all predictions is called Q_3 . For each class we can evaluate an one against the rest classifier obtain from the original classifier. The accuracy for these sub-classifiers are denoted by Q_H , Q_E , and Q_C for three classes. The performance for the subclasses

is interesting because it informs about the strength of the classifier. Most methods have problems in predicting β -sheets and it is interesting comparing different methods on the β -sheets.

Another quality measure is the segment overlap (SOV for Segment OVerlap) [Park et al., 1997] which was improved by [Zemla et al., 1999]. If in the middle of a helix three coils are predicted then this is more serious than if there are 3 errors at helix end (helix too long or too short). The prediction error in the middle would suggest two helices and the coils amino acids may form a turn. 3D Structure prediction has more problems if two helical blocks are predicted instead of one block with wrong length.

$$\text{SOV}_s = \frac{1}{N_s} \sum_{S_1 \cap S_2 \in s} \frac{\min \text{ov}(S_1, S_2) + \delta(S_1, S_2)}{\max \text{ov}(S_1, S_2)} \text{length}(S_1), \quad (4.19)$$

where

- S_1 and S_2 are the observed and predicted secondary structure segments in state s , which can be either H, E or C;
- $\text{length}(S_1)$ is the number of residues in the segment S_1 ;
- $\min \text{ov}(S_1, S_2)$ is the length of actual overlap of S_1 and S_2 , i.e. the extent for which both segments have residues in state s , for example H;
- $\max \text{ov}(S_1, S_2)$ is the length of the total extent for which either of the segments S_1 and S_2 has a residue in state s ;
- $\delta(S_1, S_2)$ is a measure of disagreement, where either S_1 is in state s but not S_2 or vice versa. The disagreement is basically $\max \text{ov}(S_1, S_2) - \min \text{ov}(S_1, S_2)$ which is upper bounded by some other values:

$$\delta(S_1, S_2) = \min \left\{ \begin{array}{l} \max \text{ov}(S_1, S_2) - \min \text{ov}(S_1, S_2) \\ \min \text{ov}(S_1, S_2) \\ \text{length}(S_1)/2 \\ \text{length}(S_2)/2 \end{array} \right\} \quad (4.20)$$

- $S_1 \cap S_2 \in s$ means all the pairs of segments (S_1, S_2) , where S_1 and S_2 have at least one residue in state s in common
- N_s is the number of residues from sequence 1 in state s defined as follows:

$$N_s = \sum_{S_1 \cap S_2 \in s} \text{length}(S_1) + \sum_{S_1 \cap S_2 \notin s, S_1 \in s} \text{length}(S_1), \quad (4.21)$$

where $S_1 \cap S_2 \notin s, S_1 \in s$ means all segments S_1 which belong to s but no pair exists where they have a residue from s in common.

The overall segment overlap is

$$\text{SOV} = \frac{1}{N} \sum_{s \in \{H, E, C\}} \sum_{S_1 \cap S_2 \in s} \frac{\min \text{ov}(S_1, S_2) + \delta(S_1, S_2)}{\max \text{ov}(S_1, S_2)} \text{length}(S_1), \quad (4.22)$$

where

$$N = \sum_{s \in \{H, E, C\}} N_s = N_H + N_E + N_C. \quad (4.23)$$

In experiments it turned out that support vector machine approaches yield a higher SOV value with state of the art Q_3 values [Hua and Sun, 2001a,b].

We tested the support vector approach and obtained $Q_3=78.84\%$ and $\text{SOV}=77.85\%$ which was better than the results of PROFS ($Q_3=76.51\%$ and $\text{SOV}=75.69\%$) or PSIPRED ($Q_3=77.75\%$ and $\text{SOV}=67.36\%$) [Drescher, 2005].

4.4.4 Problems in Quality Comparisons

Often a new method is compared to the results of previous methods, however it is difficult to make a fair comparison because

- newer methods use newer versions of data bases like the NR database for generating a profile.
- newer methods use newer versions of software like PSI-BLAST – techniques to generate alignments and correct for high sequence similarity are continuously improved.
- newer methods use newer versions of the original data set where some sequences are corrected.
- newer methods know about the pitfalls other methods run into and can include certain prior knowledge – however previous methods may also be able to incorporate this knowledge.
- newer methods know the performance threshold to achieve in order to make a publication (perhaps the new method is one out of thousands which reached higher performance by chance)

Chapter 5

Homology 3D Structure Prediction

5.1 Introduction

Now we move on to the 3D prediction of proteins based on the primary structure.

In the first part, this chapter, we assume that the new sequence has a homolog with about the same structure which is already solved. In the second part, next chapter, we try to predict also new structures which were unknown so far or have been designed.

The second part would be the more exact and the more general approach to structure prediction but the process by which a amino acid sequence is folded into a protein complicated, poorly understood, and determined by many local effects. In principle quantum mechanics can be used for folding simulations but it is intractable with computers we have today – even ligand binding is often too complicated.

Protein folding can be based on classical mechanics by using atomic force fields which are used to find a minimum energy state of the amino acid sequence. However also molecular dynamics approaches have problems:

- they do not always model the forces correctly
- they have to compute many sums over all atoms or sets of atoms to compute
- they must simulate water and its temperature with many molecules
- they must down-scale macroscopic parameters like dielectric constant
- they do not simulate the context in the cell, e.g. chaperones are not considered
- they perform simulation steps in femtoseconds while folding takes milliseconds which gives a gap of 10^{12}
- they exceed for larger proteins a simulation time of 10^{12} CPU-years at current supercomputer speeds

Therefore methods not based on first principles are still necessary for protein folding.

First we consider homology search to detect a known structure. To find a homolog there are two conceptional different approaches:

- sequence comparisons without using the structure but possibly using other sequence data bases like the NR data base which may include **sequence-sequence**, **sequence-profile**, and **profile-profile** alignments.
- comparisons which also utilize the 3D structure of a solved protein which includes **sequence-structure** alignment (also called “threading”).

In the CASP6 (3D structure prediction competition) homepage the authors write

Fold recognition takes advantage of the fact that protein structure is much more strongly conserved than sequence. Increasingly, new structures deposited in the Protein Data Bank turn out to have folds that have been seen before, even though there is no obvious sequence relationship between the related structures. The goal is to identify these structural relationships in cases where the sequence signal is either weak or does not exist. Techniques for fold recognition include advanced sequence comparison, secondary structure prediction, tests of the compatibility of sequences with known three-dimensional folds (“threading”), and the use of expert human knowledge.

Robert Service wrote in *Science* [Service, 2005]

The Protein Structure Initiative has already come up with one surprise: Proteins apparently come in a limited variety of shapes.

Also other authors observed the fact that the number of new folds is low [Govindarajan et al., 1999, Wang, 1998, Brenner and Levitt, 2000, Zhang and DeLisi, 1998]. For example [Govindarajan et al., 1999] expects 2,000 different folds in total in nature whereas [Wang, 1998] predicts 650 folds in total. Currently (beginning 2007) the SCOP data base has 971 fold entries.

These arguments supporting that the number of folds is small and most folds are already known are the basis that comparative modeling through sequence-sequence or sequence-structure alignment can predict the structure of new sequences. For every new structure it is very likely that a similar structure already was resolved and only the homology to the resolved structure must be extracted.

In the comparative modeling field threading methods (sequence-structure alignments) were until recently the golden standard for structure prediction.

5.2 Comparative Modeling: Sequence-Sequence Comparison

For high sequence similarities pairwise alignment methods like the Smith-Waterman algorithm [Smith and Waterman, 1981] or its approximations like FASTA [Pearson and Lipman, 1988] or BLAST / PSI-BLAST [Altschul et al., 1990, Holm and Sander, 1999] are the best methods to find homologs. For high sequence similarities the homologs have the same structure.

Also to store a multiple alignment in hidden Markov models (HMMs) and then search for homologs, works well for highly homolog sequences [Krogh et al., 1994, Baldi et al., 1994, Eddy, 1998, 2004, Bateman et al., 2004].

However these methods are applicable for sequences which are very similar to one another. For remote homologous amino acid sequences the methods must be refined. For example, the sequences in the same fold of the SCOP data base are not always similar to one another if pairwise alignment is applied.

The alignment based methods were recently enhanced for remote homology search by discriminative methods like the support vector machine (SVM, [Vapnik, 2000, Schölkopf and Smola, 2002]). Alignment-based method only look at the positive examples but do not look what discriminates positive from negatives. That is what positive examples are best suited to detect other positives or which conservative regions are most suited to detect other positives.

SVM based protein homology detection methods rely on a kernel which is specially designed for protein sequences: the Fisher-kernel [Jaakkola et al., 1999, 2000] is based on HMMs and alignments, the mismatch kernel [Leslie et al., 2004b,a] is based on sequence identities, the SW-kernel uses the Smith-Waterman (SW) score and the local alignment kernel uses a local SW score [Vert et al., 2004]. The SVM-pairwise method [Liao and Noble, 2002] represents sequences by their SW scores with other training sequences.

In the following we describe some methods for protein remote homology detection:

- (a) PSI-BLAST [Altschul et al., 1997] which is known from Bioinformatics I; with more than one iteration PSI-BLAST can generate a profile for the sequence at hand, where a data base like the NR data base is used; the profile is used for comparisons;
 - a.1 a new sequence is compared to profiles of all known structures and the best match selects a template structure for the new structure;
 - a.2 a new sequence is compared to all members of a fold class and the matches over fold classes are combined to find the best matching fold from which a template is generated;
 - a.3 a multiple alignment of the members of a fold class is used as a start for the PSI-BLAST which generates a profile for the whole fold class;
- (b) Family Pairwise Search (FPS, [Grundy, 1998, Bailey and Grundy, 1999]) which is a method based on BLAST [Altschul et al., 1990] and is essentially the same as previous a.2 without profiles.
- (c) SAM-T98 to SAM-T06 [Park et al., 1998, Karplus et al., 1998] codes an alignment iteratively into a hidden Markov model (HMM).
- (d) the Fisher-kernel support vector machine [Jaakkola et al., 1999, 2000]. It represents the sequence through a vector which is the gradient of the sequence's likelihood with respect to the HMM parameters (e.g. an HMM produced by SAM or by HMMER). Here the sequence of variable length is through the unsupervised HMM model transformed into a vector of fixed (number of parameters) length. This vector can be used by the SVM.
- (e) SVMs using the mismatch-kernel [Leslie et al., 2004b,a]. The mismatch kernel measures sequence similarity through amino acid identities between the sequences where both the length of the identical subsequences and their frequency is taken into account; it is similar to the BLAT alignment method;
- (f) SVM-pairwise method according to [Liao and Noble, 2002], where the feature vector is the Smith-Waterman alignment score to all other training sequences. This is a straight

forward method to extract support vectors which are the most indicative sequences of the positive (fold) and negative (non-fold) class.

- (g) SVM using the SW-kernel, where the SW-pairwise scores is used as kernel matrix as in [Vert et al., 2004]. Note, that this is not a valid kernel because it is not ensured to produce positive semi-definite kernels and the SVM-optimization is not well defined. It may be that the kernel matrix has negative eigenvalues which would avoid a solution of the SVM optimization.
- (h) SVM with the local alignment (LA) kernel [Vert et al., 2004] which is similar to a local Smith-Waterman score and is based on gap-penalties and the BLOSUM similarity matrix.
- (i) SVM with oligomer based distance measures [Lingner and Meinicke, 2006], which explicitly constructs a feature space of indicative patterns. Also approaches which use a data base of motifs like BLOCKS or PROSITE are similar to this approach. Using feature selection indicative motifs can be selected from a dictionary of motifs.
- (j) SVM-HMMSTR [Hou et al., 2004]. Also this method combines like the Fisher-kernel SVMs and hidden Markov models however whole motifs are used. SVM-HMMSTR constructs a profile using the SwissProt data base.
- (j) SVM with the mismatch kernel applied to profiles [Kuang et al., 2005]. For each sequence a profile is constructed by PSI-BLAST applied to the NR data base
- (j) SVM with LA- and SW-kernels applied to profiles [Rangwala and Karypis, 2005] Following kernels based on profiles (“position-specific scoring matrix”, PSSM) and the BLOSUM matrix instead of the profile (“global scoring matrix”, GSM) are considered: “All Fixed-width ω -mers” (AF-PSSM, -GSM), “Best Fixed-width ω -mer” (BF-PSSM, -GSM) “Best Variable-width ω -mer” (BV-PSSM, -GSM) “Local Alignment-based Kernels” (SW-PSSM, -GSM).
- (k) the LSTM recurrent network (see Bioinformatics II).

In order to have an intuition about performance and running time (test sequences) we will give an overview of above mentioned methods used the widely used benchmark data set for remote homology detection from [Liao and Noble, 2002] which is available under <http://www.cs.columbia.edu/compbio/svm-pairwise>. The data set defines 54 superfamily recognition tasks from the SCOP data base. For each task the positive examples of the training set are taken from one superfamily from which one family is withhold. The task is to detect the examples from the withhold family. Negative training examples are chosen from outside the fold the family belongs to. The quality of a ranking of the test set examples was evaluated by the area under the receiver operating characteristics curve (ROC). The methods were evaluated through 54 ROC values, where the ROC value is between 0.5 (random guessing) and 1.0 (perfect prediction). As a more precise quality measure we also used the area under the ROC50 which is the area under the ROC up to 50 false positives. ROC50 essentially re-scales the false positive rate of the ROC. The average results are given in Tab. 5.1.

In Tab. 5.1 we see that the profile-based method outperform all other methods. Essentially that means profile-profile alignment is better suited for remote homology detection than sequence-sequence or sequence-profile alignments.

method	m	p	S	ROC	ROC50	time
(a) PSI-BLAST	+	-	-	0.693	0.264	5.5s
(b) FPS	-	-	-	0.596	-	6800s
(c) SAM-T98	+	-	-	0.674	0.374	200s
(d) Fisher	-	-	+	0.773	0.250	>2000s
(e) Mismatch	-	-	+	0.872	0.400	68 h
(f) Pairwise	-	-	+	0.896	0.464	>194h
(g) SW	-	-	+	0.916	0.585	>129h
(h) LA 1	-	-	+	0.923	0.661	550h
(h) LA 2	-	-	+	0.925	0.649	550h
(i) Oligomer	-	-	+	0.919	0.508	2000s
(j) HMMSTR	-	+	+	-	0.640	>500h
(j) Mismatch 1	-	+	+	0.974	0.756	>500h
(j) Mismatch 2	-	+	+	0.980	0.794	>500h
(j) AF-GSM	-	+	+	0.926	0.549	>620h
(j) BF-GSM	-	+	+	0.934	0.669	>620h
(j) BV-GSM	-	+	+	0.930	0.666	>620h
(j) SW-GSM	-	+	+	0.948	0.711	>620h
(j) AF-PSSM	-	+	+	0.978	0.816	>620h
(j) BF-PSSM	-	+	+	0.980	0.854	>620h
(j) BV-PSSM	-	+	+	0.973	0.855	>620h
(j) SW-PSSM	-	+	+	0.982	0.904	>620h
(k) LSTM	+	-	-	0.932	0.652	20s

Table 5.1: Results on the SCOP benchmark data set. The first column gives the method. The second column “m” denotes whether it is a model based method (“+”) or not (“-”). The third column “p” denotes whether a profile input is used (“+”) or not (“-”). The fourth column “S” denotes whether a support vector machine is used (“+”) or not (“-”). The fifth column reports the average area under the receiver operating curve (“ROC”). The sixth column shows the average area under the ROC50 curve (“ROC50”). The last column reports the average time needed to classify 20,000 new sequences into one class. CPU time is measured on an Opteron 165. The time for the oligomer method was based on the LA-kernel and computed according to the numbers in [Lingner and Meinicke, 2006]. For measuring the time of the mismatch kernel we used the software from <http://www.cs.columbia.edu/compbio/string-kernels/> according to [Leslie et al., 2004a]. The measurements of the LA and the SW kernel were based on the BLAST algorithm from <http://www.ncbi.nlm.nih.gov/Ftp/> according to [Altschul et al., 1997]. The PSI-BLAST, SAM-T98 and Fisher-kernel CPU times were computed from the CPU values given in [Madera and Gough, 2002] and [Tarnas and Hughey, 1998]. The classification results except for LSTM are taken from [Vert et al., 2004, Liao and Noble, 2002, Hou et al., 2004, Kuang et al., 2005, Rangwala and Karypis, 2005, Lingner and Meinicke, 2006].

There are many other machine learning methods to find homolog sequences. Many of these approaches extract chemical or physical features from the sequence or make some statistics on the sequence [Ding and Dubchak, 2001]. However it seems that the best performance is obtained by the methods given above.

The sequence-structure alignment methods (threading methods) discussed in the next section outperformed on the CASP challenges for protein structure prediction most sequence-sequence comparisons methods. However in [Cheng and Baldi, 2006] the profile-profile alignment methods combined with machine learning achieved performance comparable with threading methods.

5.3 Threading: Sequence-Structure Alignment

For sequence-to-structure alignment or “threading” we have a dictionary of structures which are resolved. As already mentioned at the end of the introduction Section 5.1 the number of folds occurring in nature is very limited. Therefore the chance is high that the structure of a new sequence is already known but must be detected in the dictionary.

Sequence alignments only use the primary sequence to compare the sequences. Here we go a step further and also include the structural information which is available in order to see if two sequences are similar.

The basic idea is to approximate the energy or parts of it of the native fold and compare this energy to the energy of the new sequence squeezed into this fold. The energy of the new sequence in the fold tells whether it is a suited fold for the sequence or not.

To know what is the range of energy values for a native fold and for sequences which do not fit to this fold, decoys, that is similar folds to the native fold, must be generated and evaluated. Then also the energies of decoys must be computed to separate the native fold from similar ones. Additionally to the decoys, the energy of native fold with the original sequence should be less than the energy of a random sequence.

The computations of thousands decoys and random sequences makes the physical energy as energy function computationally too complex to be feasible. Note that also water has to be simulated and the simulation be done over a certain time. Therefore threading relies on empirical energy functions which are fast to compute.

The energies – pseudo energies – are based on potentials which are values assigned to amino acids at certain positions or to pairs of amino acids etc.

In most cases the side chain is neglected and only the C_β carbon atom is considered. In most contact potential over the distances $C_\beta-C_\beta$ is averaged where Glycine (G) is given a virtual C_β atom. For the distances 3 Å to 13 Å is used.

For designing a threading methods following issues must be considered:

- A) the size and quality of the template dictionary, that is the number of known folds (also called “cores”, “structures”, “folding motifs/patterns”, “contact profiles”, etc.)
- B) the potential and energy function which is used and how it is optimized
- C) the alignment procedure, that is how the combinations of how a given sequence is imposed onto a structure are evaluated and generated; that is an optimization procedure

- D) the final selection of the template; different high scores cannot be compared directly because the energy is different for each fold

A fold is often called a core because the loops (the turns or the coils) have high variation. For example two helices may be connected by 4 or by 10 amino acids. Distances and neighbors are only computed in the core region which identifies the fold.

ad A):

Essential for threading is how many folds are in the template library because as more folds are there as higher is the probability of finding an existing fold.

In many cases not whole structures but domains are used for threading because they are the building blocks which fold individually. The SCOP or CATH data base supply such domains.

Another question is how good are the measurements given by x-ray crystallography or by NMR. This will influence how exact contacts can be measured.

Depending on the energy i.e. objective function the templates are preprocessed for an suitable representation. For example distances and weighting functions are pre-computed.

For many potential functions the potential must be deduced from 3D structures. However exactly these structures are used as templates. Therefore a bias is introduced which gives known structures a lower energy in their native folds than new sequences in their native fold. For many templates this bias is assumed to be very small.

ad B):

The energy function which is also called the objective function or the scoring function measures the fitness of a sequence in a certain structure (“the sequence is threaded through the structure”).

The energies are computed from the potentials which evaluate certain configurations in 3D. Potentials are used to describe core elements (the hydrophobic core), neighbor relations like contact order, number of contacts, distances (contact potential), environments (hydrophobic amino acids in the inside, hydrophilic amino acids on the outside), etc. The potentials for single residues are called *singleton* which counts whether the residue is buried (hydrophobic) or participates at certain secondary structures. Typical potentials are *contact potentials*, *knowledge-based potentials*, and *potential of mean force*. There exists potentials which only measure how many contacts makes a residue at a certain position and compare this value with an average measure of contacts for this residue – hydrophobic residues make more contacts as they are inside the protein than polar or charged residues.

The potentials are statistics computed by frequency counting the 3D configurations in a set of resolved structures. In A) we already mentioned that these structures often overlap with the templates, therefore, the statistics are biased towards low energy of the templates.

Using these statistics and the actual configuration of the query sequence, the energy is computed. The value computed from the statistics must be normalized in order to obtain the energy (the density is normalized to obtain a probability). The energy can be optimized (see below) in order to better discriminate the sequences corresponding to the native fold from random sequences.

ad C):

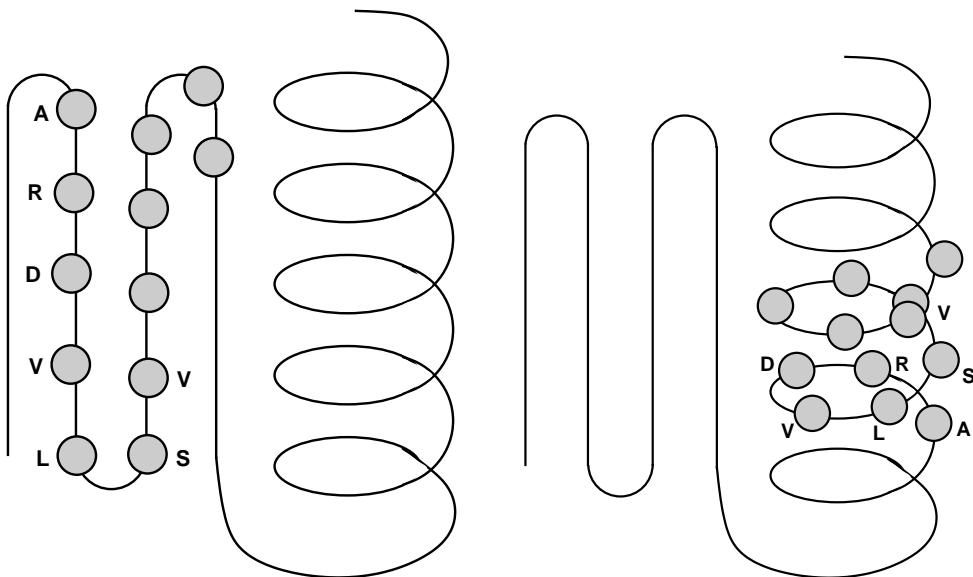


Figure 5.1: Threading alignment. The amino acid sequence ARDVLSV... is threaded through a structure with one helix and a sheet consisting of 4 strands.

In the sequence-structure alignment the sequence is threaded through the structure. Fig. 5.1 depicts how an amino acid sequence is threaded through a structure.

The alignment's objective is the energy function from B). As with sequence-sequence alignment the problem is difficult because of the gaps. Gaps typically result from the fact that the loops (the turns) have different length in similar structures.

For pairwise contact potentials the problem of finding the optimal alignment solution is NP-hard. For only singleton potentials alignment algorithms like for the sequence-sequence alignment can be applied – it is only an alignment of the new sequence to certain positions.

Exact methods are exhaustive search and methods which rely on a branch and bound algorithm. Approximative approaches to the general sequence-structure alignment problem involve

- “double dynamic programming” where the dynamic programming idea is iteratively applied to improve the current solution. A residue is placed in another position and all other residues are optimized for the residue in this place. [Jones et al., 1992]
- “frozen approximation”, where the template residues are kept and only a new query residue is inserted – this is done iteratively until convergence [Sippl, 1993, Wilman and Eisenberg, 1995, Godzik et al., 1992].
- sampling and search methods: Gibbs sampling, Monte Carlo sampling, simulated annealing; these methods are only practicable if the search space is restricted by forbidding gaps except in loops [Bryant and Lawrence, 1993, Madej et al., 1995].
- mean field approaches, where certain independence assumptions are made [Huber and Torda, 1999].

- branch and bound algorithms, where unlikely regions in the search space are eliminated [Xu and Xu, 2000, Lathrop, 1999, Xu and Uberbache, 1996, Lathrop and Smith, 1996]. These methods were recently made fast and were successful [Xu and Li, 2003].

ad D):

After the optimal energy for the sequence on each template (structure/fold) is computed a template must be chosen. However it is difficult to compare the energy on different folds with each other.

For a specific query sequence many folds will yield low energy if gaps are inserted and the hydrophobic amino acids are brought into contact.

To measure the fitness of a sequence independent of the fold z -scores can be used. The z -score measures how many standard deviations σ is the actual energy value separated from the mean value. To compute the z -score first the mean μ and the variance σ^2 must be computed.

To compute these values sequences of other folds are threaded through the fold and the values σ and μ are estimated.

Another approach computes decoys which are folds which are similar to the fold to investigate.

Decoys can be constructed either through (I) deviation of the native fold or (II) through minimizing (gradient descent) the energy of native fold with respect to the current potential function. Deviations are normally made by deviations of the torsional angles e.g they are perturbed by $-30^\circ \leq \phi \leq 30^\circ$.

However for the z -score a Gaussian distribution cannot be assumed because the score stems from an optimization procedure and follows a extreme value distribution (see Bioinformatics I).

Energy parameter optimization.

The energy is

$$E = \sum_i s(a_i, p(a_i)) + \sum_i \sum_{j:i < j} S_{ij} c(a_i, a_j), \quad (5.1)$$

where a is the amino acid sequence with amino acid a_i at position i , $p(a_i)$ is the position of the amino acid in the 3D structure, $s(a_i, p(a_i))$ gives the score of amino acid a_i in position $p(a_i)$, S is the contact matrix ($S_{ij} = 1$ if amino acids a_i and a_j are in contact and otherwise ($S_{ij} = 0$), and $c(a_i, a_j)$ is the contact potential for amino acids a_i and a_j .

Let us focus on a pure pairwise contact potential

$$E = \sum_i \sum_{j:i < j} S_{ij} c(a_i, a_j). \quad (5.2)$$

In Appendix B some global (not fold specific) contact potentials are given.

We number the folds (templates) and denote by 0 the native fold. The energy for fold p is denoted by E_p .

We want to ensure that

$$E_0 < E_p \quad (5.3)$$

for all $p \neq 0$.

The z -score is defined as

$$Z = \frac{E_0 - \mu}{\sigma}, \quad (5.4)$$

where μ is the mean energy for the fold and σ the standard deviation of the energy.

If decoys are generated then the amino acids remain the same and

$$\mu = \langle E \rangle = \sum_i \sum_{j:i < j} \langle S_{ij} \rangle c(a_i, a_j) \quad (5.5)$$

and

$$\sigma^2 = \sum_i \sum_{j:i < j} \sum_k \sum_{l:k < l} \text{cov}(S_{ij}, S_{kl}) c(a_i, a_j) c(a_k, a_l), \quad (5.6)$$

that is only the mean and covariance of the contact maps have to be computed.

The z score is therefore

$$Z = \frac{\sum_i \sum_{j:i < j} c(a_i, a_j) (S_{ij}^0 - \langle S_{ij} \rangle)}{\sqrt{\sum_i \sum_{j:i < j} \sum_k \sum_{l:k < l} \text{cov}(S_{ij}, S_{kl}) c(a_i, a_j) c(a_k, a_l)}} \quad (5.7)$$

or in vector notation if all pairs are put into one vector

$$Z = \frac{\mathbf{c}^T (\mathbf{s}^0 - \langle \mathbf{s} \rangle)}{\mathbf{c}^T \mathbf{S} \mathbf{c}}, \quad (5.8)$$

where \mathbf{c} is the vector with components $c(a_i, a_j)$ \mathbf{s} is the vector with components S_{ij} (analog for \mathbf{s}^0) and \mathbf{S} is the covariance matrix of \mathbf{s} .

Let us assume the vectors \mathbf{s}^p are normalized to zero mean and combined in a data matrix \mathbf{X} . The first element in \mathbf{X} is \mathbf{s}^0 . Let us further define a vector of dimension $(P + 1)$ $\mathbf{y} = (-1, \frac{1}{P}, \dots, \frac{1}{P})$, where P is the number of decoys.

In this case we have:

$$Z = \frac{\mathbf{y}^T \mathbf{X} \mathbf{c}}{\mathbf{c}^T \mathbf{X} \mathbf{X}^T \mathbf{c}}. \quad (5.9)$$

Let us assume we used the negative energy instead of the energy, therefore we want to maximize the value Z . This can be done by $\mathbf{c}^T \mathbf{X} \mathbf{X}^T \mathbf{c}$ and maximizing $\mathbf{y}^T \mathbf{X} \mathbf{c}$. To combine both optimization procedures we minimize

$$\frac{1}{2} \mathbf{c}^T \mathbf{X} \mathbf{X}^T \mathbf{c} - \mathbf{y}^T \mathbf{X} \mathbf{c}. \quad (5.10)$$

This is exactly the objective of the P-SVM from Bioinformatics II. Therefore the z -score optimization can be cast as a classification problem where the native fold is the only member of the positive class.

The energy E can be written as

$$E = \mathbf{c}^T \mathbf{s} . \quad (5.11)$$

We want to ensure

$$E_0 - E = \mathbf{c}^T (\mathbf{s}^0 - \mathbf{s}) > 0 . \quad (5.12)$$

This can be learned by the perceptron learning rule or by a one-class support vector machine (see Bioinformatics II).

If only different sequences are used then $c(a_i, a_j)$ can be replaced by c_{ij} , that is a value for each amino acid pair. S_{ij} counts how often amino acid i is in contact with amino acid j .

Then

$$E = \sum_{i=1}^{20} \sum_{j=i}^{20} S_{ij} c_{ij} \quad (5.13)$$

and

$$\mu = \langle E \rangle = \sum_i \sum_{j:i < j} \langle S_{ij} \rangle c_{ij} \quad (5.14)$$

as well as

$$\sigma^2 = \sum_i \sum_{j:i < j} \sum_k \sum_{l:k < l} \text{cov}(S_{ij}, S_{kl}) c_{ij} c_{kl} . \quad (5.15)$$

Performance. As can be seen at the CASP7 results from 2006 in Appendix C, threading methods or methods which are based on threading methods are still the best performing methods. Only Rosetta as an ab initio method can compete with threading methods.

Chapter 6

Ab Initio Prediction and Molecular Dynamics

6.1 Introduction

In this section we consider methods which only use the amino acid sequence to predict its 3D structure. Also physical forces or their approximations are used as are energy functions.

Sometimes experimental data is included to improve the methods like for the Rosetta method, where a library of 3 and 9 residue fragments is used.

Ab initio methods are applicable to proteins which have a novel structure so that threading methods would fail. These methods are also important for construction new proteins.

Ab initio method and molecular dynamics give insights into protein folding and protein stability in contrast to the methods considered so far.

6.2 Ab Initio Methods

Ab initio structure prediction only uses the amino acid sequence to find the 3D structure.

Rosetta

Protein folding starts locally with small fragments which build for example helices. If over many proteins is averaged these small fragments should be detected. Folds are build of these local fragments. To find a fold, Rosetta uses a library of 3 and 9 residue fragments from which a fold is constructed. A sequence and profile-profile method extracts the appropriate fragments from the library. For constructing the fold the hydrophobic residues should be in the inside of the protein, β -strands have to be paired, loops are at the outside of the protein, etc. Monte Carlo sampling is used to sample possible conformations.

The scoring function is based on hydrophobic burial, pairwise interaction like electrostatic and disulfide bonds and spherical packing, α -helix and β -strand packing and β -strand pairing. Important for the Rosetta method is to filter out non-plausible structure with poorly formed β -sheets, with low contact order, or poorly packed interior.

The next improvement of Rosetta is to use information from homologous sequences. These structures can be directly used and only insertions, loops, and extensions modeled by the fragments. Only if the homology is low the whole structure is modeled by the Rosetta method.

Rigid body models

Secondary structures are predicted and represented as rigid bodies where the torsion angles are only changeable at the junctions of these bodies. However they lack details which allow for strand twists and packing issues and therefore do not perform as well as Rosetta.

Lattice representations

The residues are restricted to points on a regular three-dimensional lattice. The state space can be very fast sampled and also advanced algorithmic optimization methods exist. On the other hand also these methods have their limitations in modeling more exact details.

Potential functions

Molecular mechanics and force fields may be used but they are computationally expensive if also water must be modeled. But also the potentials which are empirically derived for threading can be used. These potentials are especially important, if a reduced protein model (e.g. without side chains) is used because the molecular dynamics is no longer applicable.

Optimization techniques and search methods

The energy landscape of the current conformation must be sampled for which Monte Carlo Sampling, simulated annealing, evolutionary or genetic algorithms may be used.

Sampling is based on torsion angle variations, direct movement of the atoms, or on fragment insertion.

The most sampling strategies make multiple runs or perform a parallel search at different regions of the conformations space.

The candidate solutions are then filtered and checked for plausibility. In this phase a more detailed model may be used. In general, as fewer candidates have to be considered as more detailed the model can be.

6.3 Molecular Dynamics

In principle the whole folding process of a protein can be modeled by first principles, that is by the physical laws. If this modeling can be done then approximations are not necessary.

Modeling in most detail can be done with quantum mechanics. However to simulate on the quantum mechanic requires to solve many integrals and is computational very expensive – even to simulate the binding of a metal atom to a protein takes days.

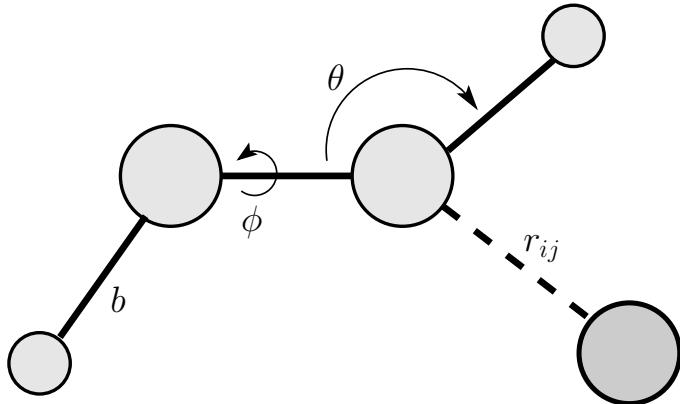


Figure 6.1: Schematic view of the angles and bond length which are used to compute the force field. The dashed line shows an interaction which is non-bonded.

The next level of abstractions is molecular dynamics or molecular mechanics. Assumptions is that atom movements are on a much slower time scale than electronic motions therefore averaging over the electronic motions is justified. For averaging the ground state energy is used.

With Molecular dynamics forces on individual atoms are computed using so called “*force fields*” where all atom positions are given as 3D coordinates. The computations of these forces allows for molecular dynamics and Monte Carlo simulations which compute the energy of the current state by sampling the energy landscape. This sampling also determines the next state which has with lower energy as the current state and into which the current state will move with high probability.

Applications of molecular dynamics range from modeling ligand binding, enzymatic reactions, denaturation, and refolding. Most importantly in bioinformatics, molecular dynamics can be used in structural prediction either to refine the final model or to compute the energies of a couple of candidates in more detail.

The most popular force field computes the energy as

$$\begin{aligned}
 V(r) = & \sum_{\text{bonds}} k_b (b - b_0)^2 + \sum_{\text{angles}} k_\theta (\theta - \theta_0)^2 + \\
 & \sum_{\text{torsions}} k_\phi (\cos(n \phi + \delta) + 1) + \\
 & \sum_{\text{nonbondpairs } ij} \left(\frac{q_i q_j}{r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^6} \right),
 \end{aligned} \tag{6.1}$$

where the first three sums are for bonds and the last sum is for non-bonded interactions. The first term penalizes energetic unfavorable *bond length*’ (bond stretching), the second term penalizes energetic unfavorable *angles* (angle bending), the third term penalizes energetic unfavorable *torsion angles* (dihedral angles), and the last term is a physical term. This last term involves *Coulomb’s law* using the partial charge q_i and q_j and the *Lennard-Jones (“van der Waals”)* potential. The angles and distances are depicted in Fig. 6.1. The parameters are k_b , b_0 , k_θ , θ_0 , k_ϕ , n , δ , A_{ij} , C_{ij} and the partial charges which are assigned to different molecules or molecule constellations.

Some force fields also include *improper torsion angle* terms in order to enforce planarity or include Urey-Bradley terms which model also other interactions between atoms separated by two bonds.

A very popular force field is the AMBER-ff99 force field which defines above mentioned parameters (see <http://amber.scripps.edu/> for more). The file, as used by the TINKER package, is available under <ftp://dasher.wustl.edu/pub/tinker/params/amber99.prm>.

Other popular force fields are the CHARMM (Chemistry at HARvard using Molecular Mechanics), e.g. the CHARMM19 force field, and the OPLS (Optimized Potentials for Liquid Simulations) force fields, or the MM2/MM3 and GROMOS force fields.

Besides the force fields for proteins the water must be modeled for which also force fields exist (ST2, SPC, TIP3P-TIP5P).

Molecular dynamics programs are TINKER <http://dasher.wustl.edu/tinker/>, which comes with a source code (in Fortran) and many helpful optimization and sampling programs. Another program is Moldy <http://www.earth.ox.ac.uk/~keithr/moldy.html> and many other programs can be found under <http://www.netsci.org/Resources/Software/Modeling/MMMD/index.html>.

The challenge for molecular dynamics programs is to derive fast methods to compute the forces on the single atoms or to sample the energy around the current state. Besides using parallel programs it is important how the sums are computed. For example distance geometry metricalization, Elber's reaction path, Scheraga's Straub's potential smoothing, multi-pole expansion, etc. can considerably speed up the simulations.

Part II

Genome Analysis

Chapter 7

Introduction

We will give a detailed overview over the DNA microarray technique in the first chapter of the second part of the course.

The DNA microarray technologies such as cDNA and oligonucleotide arrays provide means of measuring tens of thousands of genes simultaneously, therefore are a large scale high-throughput method for molecular biology experimentation.

One of the goals of microarray technology is the detection of genes that are differentially expressed in tissue samples like healthy and cancerous tissues to see which genes are relevant for developing the illness.

It has important applications in pharmaceutical and clinical research and helps in understanding gene regulation and interactions. The information obtained by recognizing genes that share expression patterns and hence might be regulated together are assumed to be in the same genetic pathway.

Issues addressed in this first chapter of the second part of the course will be among others, scanner image analysis, background correction, normalization, perfect match correction, summarization, machine learning applications (gene selection, clustering, classification).

In the second chapter an overview over genome anatomy and genome individuality (e.g. repetitions or single nucleotide polymorphism) are given.

In the next chapter some actual genomic research questions are considered like alternative splicing or nucleosome position.

Chapter 8

DNA Microarrays

8.1 Motivation

The microarray technique is a recent technique which allows to monitor the concentration of many kinds of messenger RNA (mRNA) simultaneously in cells of a tissue sample and provides a snapshot of the pattern of gene expression at the time of preparation. The DNA microarray technology is appealing because it is a high-density, high through-put method and gives a snapshot of the expression values of tens of thousands genes in a cell at a time and provides valuable information about whole genetic networks. To record the expression values of all genes at once gives a chance to detect relationships which were hidden from the researchers.

Before the microarray technology has been established, only the expression value of a small number of genes could be measured. Therefore, the experiments were hypothesis driven: first a hint to a dependency between a gene or a gene group and a condition or between genes or gene groups had to be present then the experiments verified or falsified this hint. Nowadays microarray measurements can be minded for discovering new hypotheses.

Conditions in cells can be externally induced through stimuli (toxics, chemicals which are potential drugs, viruses, temperature, or energy stress) or are macroscopic observed (cell division, growth state, tumor, etc.). With microarray it is possibly to systematically analyze the response or the state of the cell and to discover new regulatory dependencies.

In medical applications microarrays obtained a special attention because they are suited for diagnosis and prognosis. In a typical scenario the medical doctor takes a tumor sample from a cancer patient. Then from this sample a gene expression profile is made with the microarray technique. The expression state can determine the specific kind of cancer and its current status. The current status can also be used to predict the outcome of different treatments and whether metastasis are present. This prediction can be used to adjust the doses (if the outcome is positive predicted then the doses may be reduced) and to select the treatment which is best suited to the specific patient. Another important fact is that certain genes are indicative for the treatment outcome or for the diagnosis. These genes may serve to find new targets for drug design.

Microarrays are not only relevant for cancer but also for leukemia and many other diseases. Even the freely flooding mRNA in the blood gives hints to certain diseases.

Other medical applications involve genome-wide genotyping here tiling arrays are of interest where the genome is recorded piece by piece through mRNA expression. The other important field is SNP (single nucleotide polymorphism) data analysis. The first SNP arrays which detect certain predefined SNPs. The SNP genotype data is correlated with the clinical phenotype data.

Medical applications involve the prediction of the treatment outcome of childhood brain tumor [Pomeroy et al., 2002], Lymphoma cancer [Shipp et al., 2002], breast cancer [van't Veer et al., 2002]. In diagnosis gastrointestinal stromal tumors [Allander et al., 2001] was treated and many other cancer applications were published [Mukherjee et al., 2000, Furey et al., 2000, Brown et al., 2000, Cai et al., 2001].

The Netherlands Cancer Institute in Amsterdam aims at using microarray techniques for routine diagnostic and screening for treatment selection [Schubert, 2003].

8.2 DNA Microarray History and Current Status

With the *Southern blot* [Southern, 1988] the step from immunoassays – which used antibody-antigen to detect the presence of a certain genetic material – to direct DNA identification. At the same time there were many approaches towards extracting the expression level of a cell at various groups [Lysov et al., 1988, Drmanac et al., 1989, Bains and Smith, 1988].

In immunodiagnostics [Ekins and Chu, 1991] developed a microspotting technique (developed for practical use at Boehringer) called “multianalyte microspot immunoassay”. These techniques are still important for peptide and protein arrays.

The Southern blot technique brings denatured DNA fragments onto nitrocellulose filter on which hybridization between labeled *probes* and the fragments, the *targets*. Southern blot does not use antigen-antibody affinity but the affinity of complementary nucleotide sequences even between DNA and RNA. The probes can be attached by covalent bonds to a solid surface [Southern, 1975] after using gel electrophoresis. These days the array surface is of glass and does not use porous surfaces.

At Affymetrix, Inc. the first microarray chips, the GeneChip® were founded on [Fodor et al., 1991] and further developed [Lockhart et al., 1996]. Another technique has been developed at Stanford [Schena et al., 1995]. In the following we will report some Chips which can be bought from Affymetrix to give an idea of the current status of the microarray technology.

Affymetrix “Human Genome U133 Plus 2.0” array contains 61,000 probe sets over 47,000 transcripts and 45,500 human genes. Note that one probe set contains 11 to 20 probe-pairs (22 to 40 spots). The same array was constructed for formalin-fixed, paraffin-embedded (FFPE) samples and is called “X3P” (for this array the probes are selected from the 300 bases at the 3' end in contrast to 600 bases in the standard chip). The “Human Genome U95 Set” contains almost 63,000 probe sets interrogating approximately 54,000 UniGene clusters derived from Build 95 of UniGene. The “Human Genome U133 (HG-U133) Set” contains about 45,000 probe sets representing more than 39,000 transcripts and about 33,000 human genes. The sequence clusters are derived from the UniGene database (Build 133, April 20, 2001). The “Human Genome U133A 2.0 Array” contains 18,400 transcripts including 14,500 human genes on 22,000 probe sets.

Other Affymetrix array are the “Human Exon 1.0 ST Array” which contains 1.4 million probe sets comprising more than 1 million exon clusters. Here the probes are designed to mark each exon (about 4 probes per exon).

For DNA analysis Affymetrix supplies the “Human Tiling 2.0R Array Set” which is a set of seven arrays containing approximately 45 million oligonucleotide probes to analyze the human

genome. Probes are tiled at resolution of 35 base pair and measurement was done by 25-mer oligos. This results in a gap of 10bp between probes.

For SNP data analysis Affymetrix designed the “Genome-Wide Human SNP Array 5.0” SNPs on the array are present on 200 to 1,100 base pair fragments and are amplified. 440,794 SNPs are covered on the array.

Other array provided by Affymetrix focus on the mitochondrial genes and on diseases like SARS.

8.3 DNA Microarray Techniques

Fig. 8.1 depicts the basic microarray procedure for measuring the mRNA concentration.

The steps in Fig. 8.1 are described in the following.

Step 1: Messenger RNA is extracted from the samples

Step 2: mRNA is reversely transcribed to cDNA

Step 3: the cDNA from Step 2, the “target” cDNA, is then coupled to a fluorescent dye – sometimes a labeled cRNA is actually produced.

Step 4: the target cDNA (cRNA) is brought onto the chip with immobilized probes which had been synthesized and fixed to different locations (the spots) of the DNA chip during fabrication.

Step 5: the target is hybridized with the probes where targets (cDNA or cRNA) from the samples binds to their corresponding probes (the complementary sequences) on the chip.

Step 6: After cleaning, the chip is scanned with a confocal microscope and the strength of the fluorescent light is recorded. Genes which are predominantly expressed in the sample give rise to bright spots of strong fluorescent light. No expression is indicated by weak fluorescent light.

Step 7: After segmentation of the stained locations on the chip the intensity values are transformed to real numbers for every location.

After processing, the data from several experiments with different samples are collected and represented in matrix form, where columns correspond to tissue samples, rows correspond to genes, and matrix entries describe the result of a measurement of how strong a particular gene was expressed in a particular sample.

For an overview over the microarray technique see [Wang et al., 1998, Gerhold et al., 1999].

The major techniques can be divided into *oligonucleotide* arrays and *cDNA* arrays or *spotted* arrays.

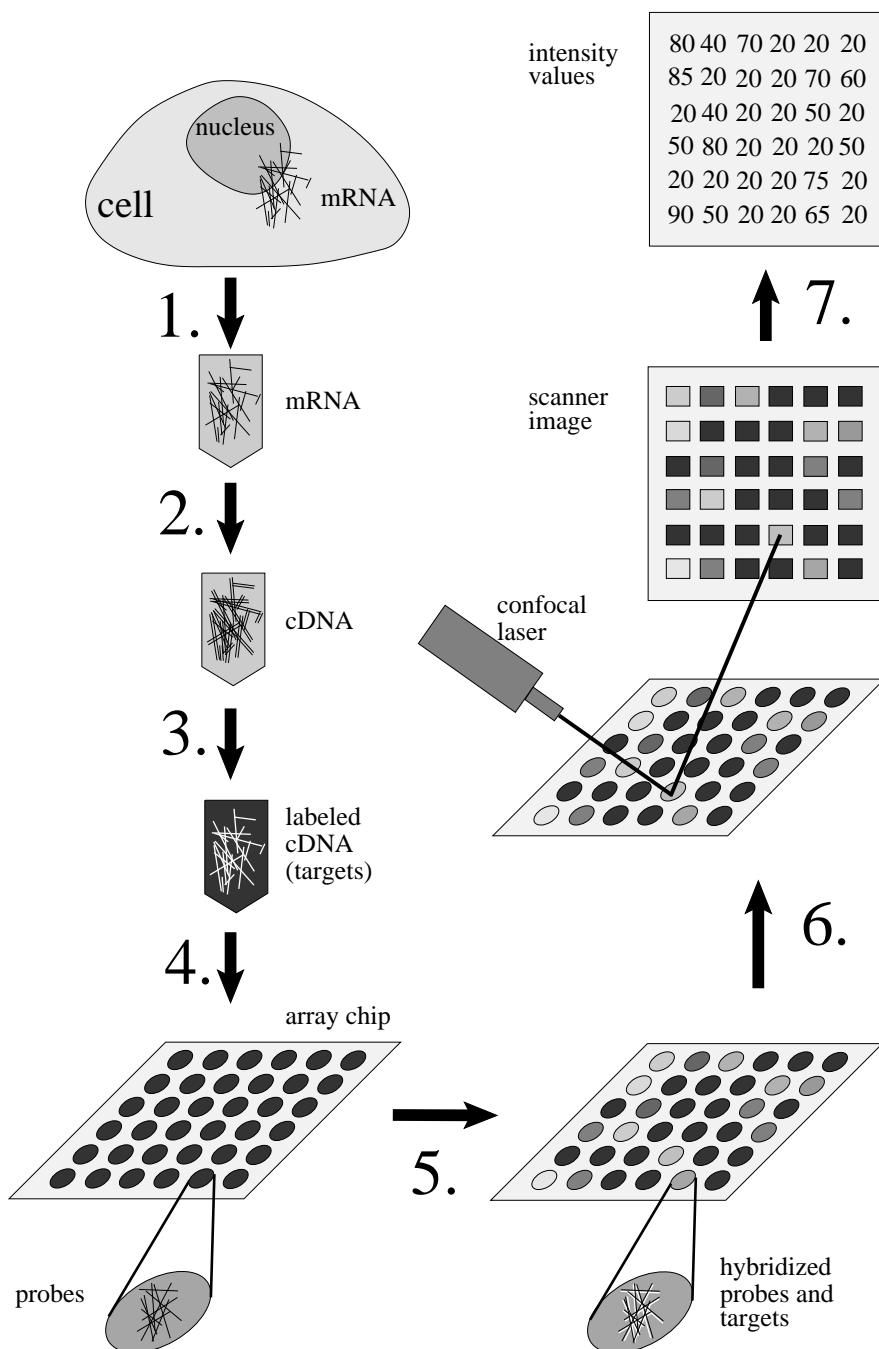


Figure 8.1: The microarray technique (see text for explanation).

8.3.1 Oligonucleotide Arrays

The oligonucleotide arrays use short sequences of 20 to 70 nucleotides (complementary to a subsequence in the mRNA) to measure the concentration of mRNA where fragments of the mRNA (mostly coded into cRNA) are used.

The Affymetrix GeneChip® technique belongs to the oligonucleotide arrays class. Its technique is shown in Fig. 8.2.

The steps of the Affymetrix oligonucleotide arrays:

Step 1: mRNA extraction from the cell

Step 2: reverse transcription of the mRNA to cDNA to conserve the mRNA information

Step 3: transcription of cRNA from the cDNA where together with transcription the cRNA is labeled with biotin (fluorescence substance)

Step 4: the cRNA is fragmented

Step 5: the fragmented cRNA, the target, is brought onto a chip with immobilized probes which consists of to the cRNA complementary nucleotide sequences; if target is available then probe and target hybridize

Step 6: wash and stain to remove superfluous fluorescence from the chip

Step 7: scan the chip

Fig. 8.3 shows a simpler procedure of the Affymetrix technology.

Fig. 8.4 shows the Affymetrix GeneChip® and the images obtained from these chips.

Fig. 8.5 shows the Affymetrix GeneChip® devices, the wash and stain station and the scanner-computer station.

With the Affymetrix technique 11 to 20 short oligonucleotides of 25 base pairs in length are distributed over the last 600 base pairs of a reference mRNA – see Fig. 8.6.

The perfect match and the mismatch are 25 bp sequences. The perfect match is the complementary of the corresponding mRNA sequence whereas the mismatch is identical to the perfect match except that the center nucleotide is exchanged:

TGTGATGGTGGGAATGGTCAGAAGGACTCCTATGTGGGTGACGAGGCC	mRNA reference
TTACCCAGTCTCCTGAGGATACAC	“perfect match”
TTACCCAGTCTGCTGAGGATACAC	“mismatch”

An example for actually scanned intensity values is shown in Fig. 8.7.

Affymetrix supplies different arrays, see Fig. 8.8.

For formalin-fixed paraffin-embedded (FFPE) samples a special array from Affymetrix exists, the GeneChip Human X3P Array, which marks the last 300bp (instead of 600) at the 3' end because the mRNA is degrading faster in paraffin.

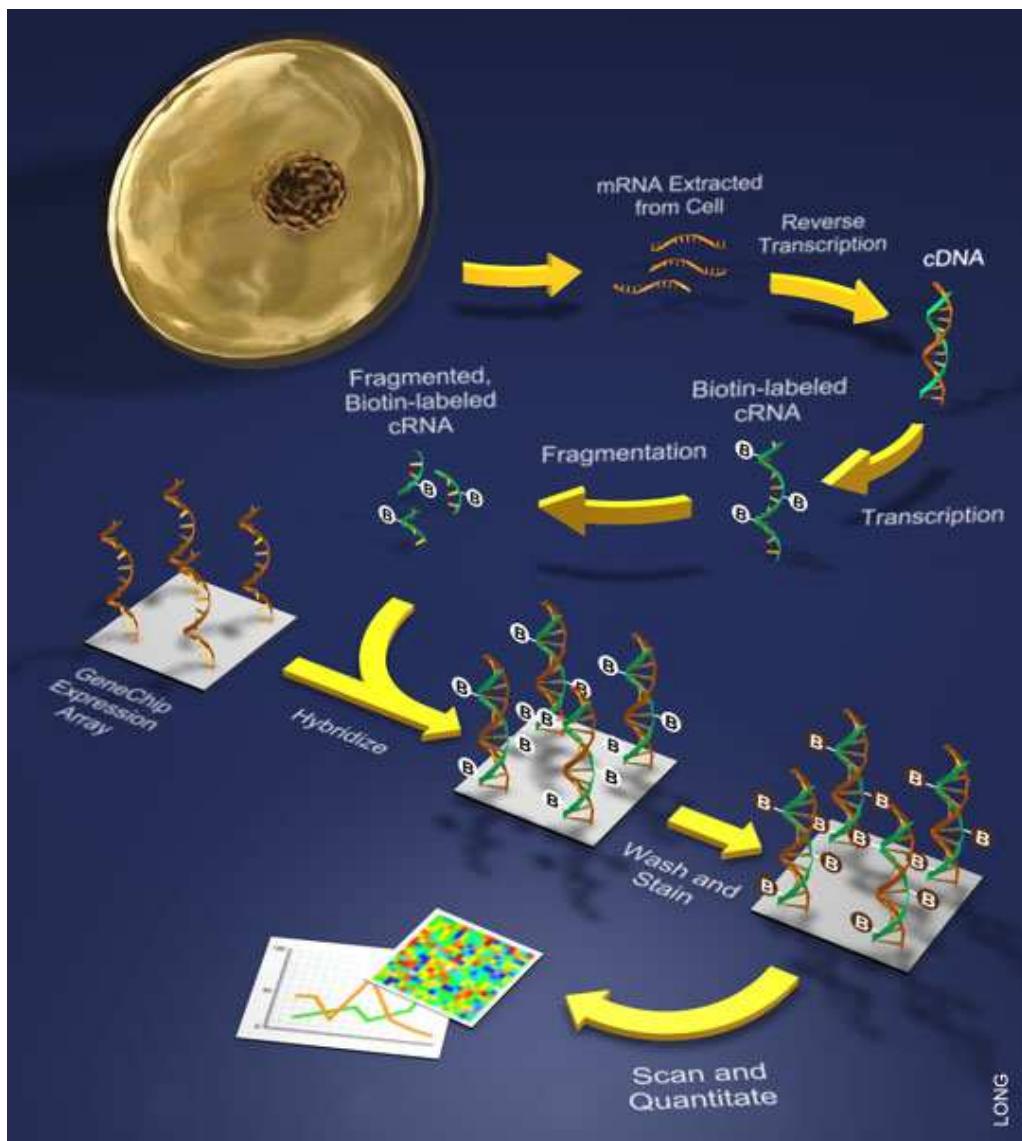


Figure 8.2: The Affymetrix microarray technology for oligonucleotide arrays.

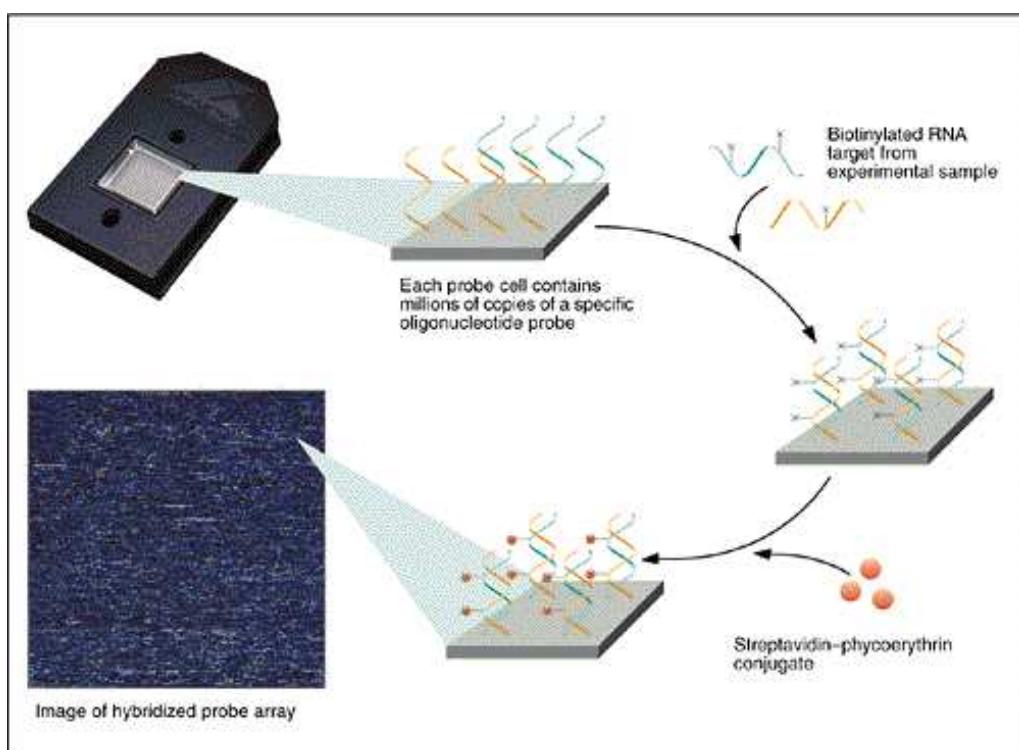


Figure 8.3: The Affymetrix technology in a simplified version.

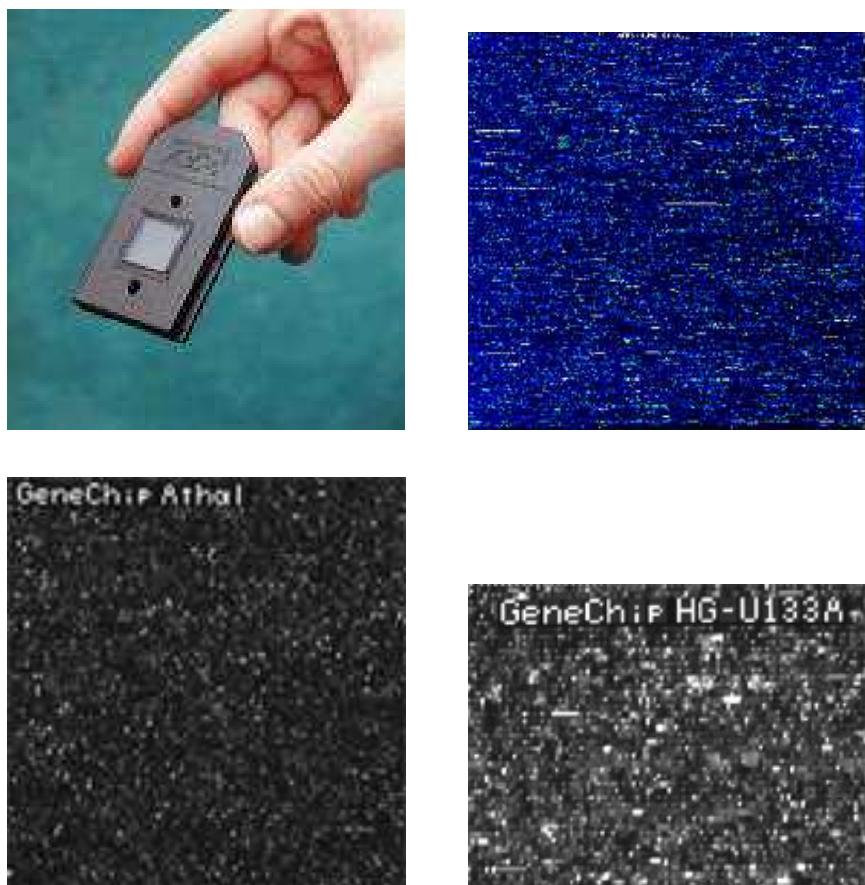


Figure 8.4: The Affymetrix GeneChip® and the scanner images.

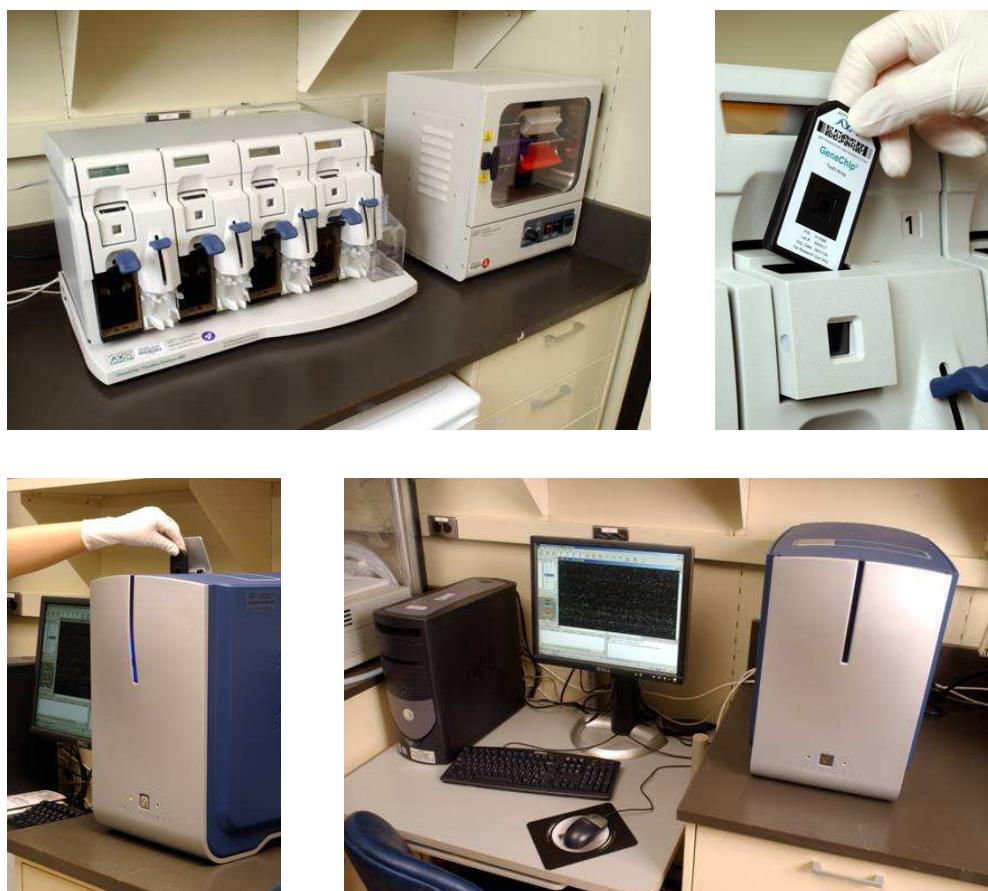


Figure 8.5: The Affymetrix GeneChip® devices. First line the wash and stain station (left) where the chip is inserted (right). The second line shows the scanner and computer station where the chip is inserted into the scanner (left).

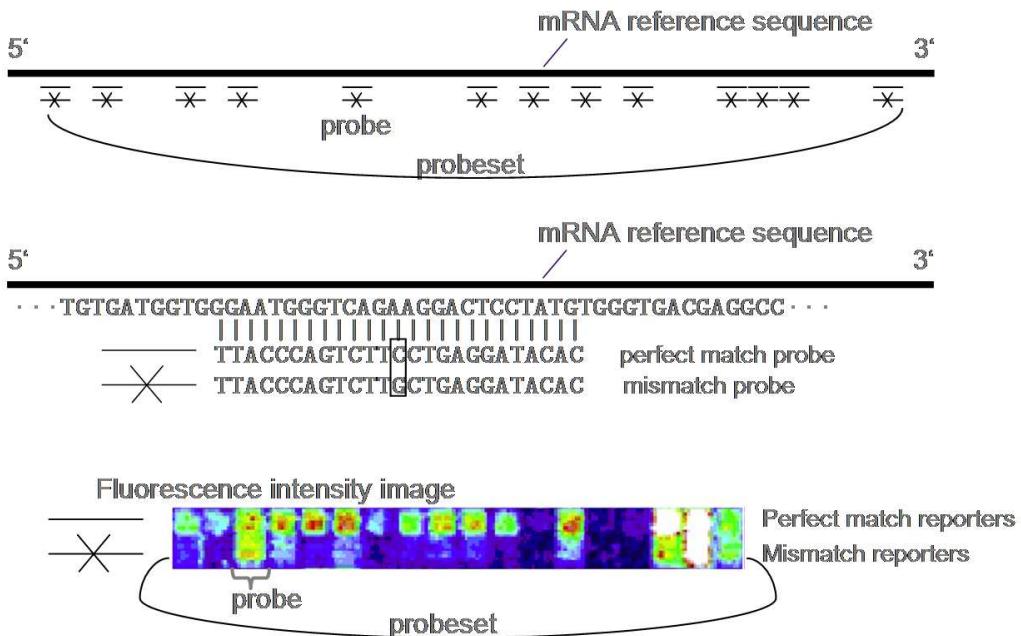


Figure 8.6: The Affymetrix technique to detect the expression of a gene. Probes (consisting of a perfect match and a mismatch) are distributed over the last 600 base pairs of an mRNA sequence.

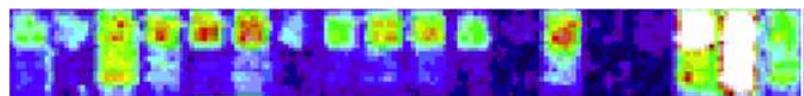


Figure 8.7: A probe set of an Affymetrix chip. Top are the perfect matches and bottom are the mismatches. The perfect matches should ideally have the same intensity values as should have the mismatches.

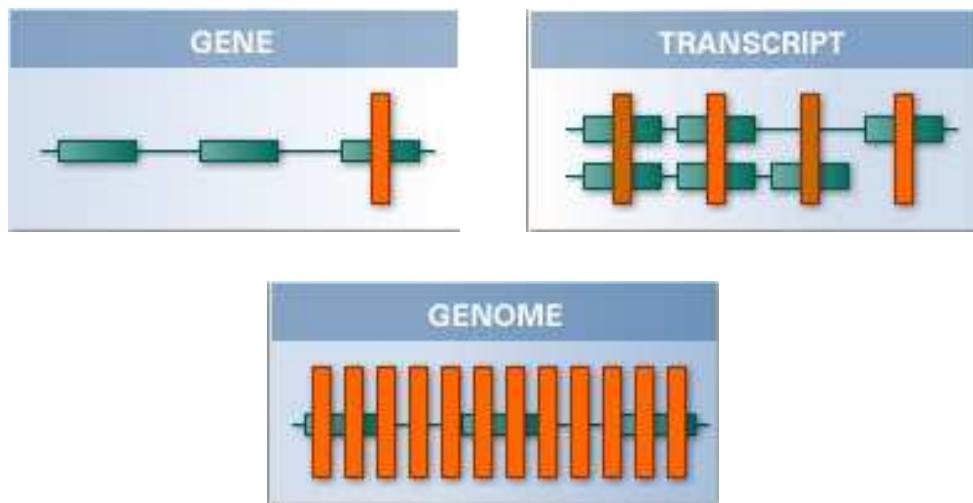


Figure 8.8: The different Affymetrix array techniques. The first image depicts the probe locations of a normal gene expression array, where the exons at the 3' end (last 600 bases) are marked by probe sets. The second image depicts the probe locations of an exon array, where each exon is marked by 3 to 4 probes (which is useful to detect alternative splicing). The third image depicts the probe locations of a tiling array, where the 25mers are shifted 35bp which means only 10bp are not covered between the probes (also other transcripts than coding can be detected).

8.3.2 cDNA / Spotted Arrays

The cDNA array usually use hundreds of complementary nucleotides for detecting mRNA.

The spotted array technique first separately extracts mRNA from two cell lines: condition 1 vs. condition 2 or healthy vs. disease tissue. These cell lines are amplified through PCR and then their mRNA reverse coded (or a complementary strand is produced) into cDNA. Either the transcripts of the cDNA or the mRNA is then marked with fluorescent dyes, where Cy3 and Cy5 are the standard dyes. Therefore mRNA of one cell line is marked with red and mRNA of the other one with green fluorescent dye. These cell lines are mixed and brought onto a glass chip. After hybridization the glass chip is scanned by excitations which leads to green emitting fluorescent and another excitation which leads to red emitting fluorescent. Each spot can emit either red or green or both fluorescent light. An average or ratio of green intensity and red corresponding to the different cell lines can be computed.

The spotted array are depicted in Fig. 8.9 to Fig. 8.12.

The images obtained from the spotted array technique if green and red are superimposed can be seen in Fig. 8.13 to Fig. 8.15.

These technique is called “spotted” arrays because a robot spotter brings small quantities of the probes onto a glass plate. The probes are fixed at the glass surface (e.g. through polylysine).

The glass chip manufacture involves

- prepare fixing regions on glass plate (polylysine)

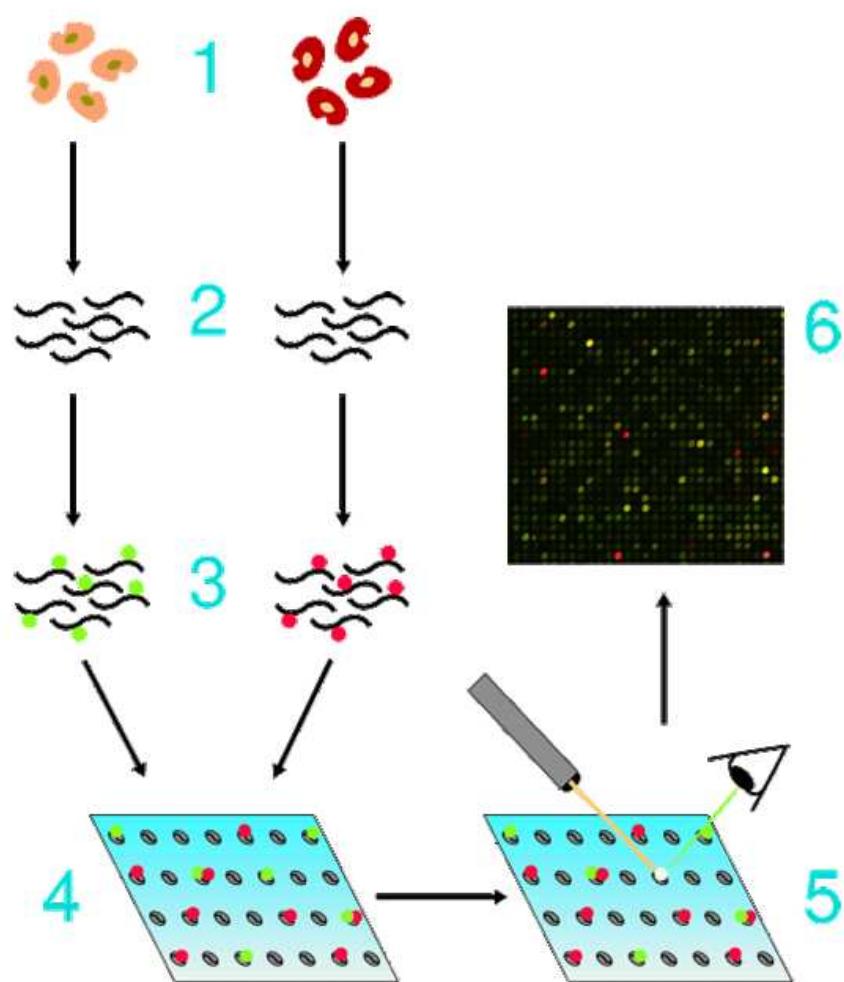


Figure 8.9: The steps of spotted arrays (cDNA arrays). Copyright ©1998–1999 Jeremy Buhler.

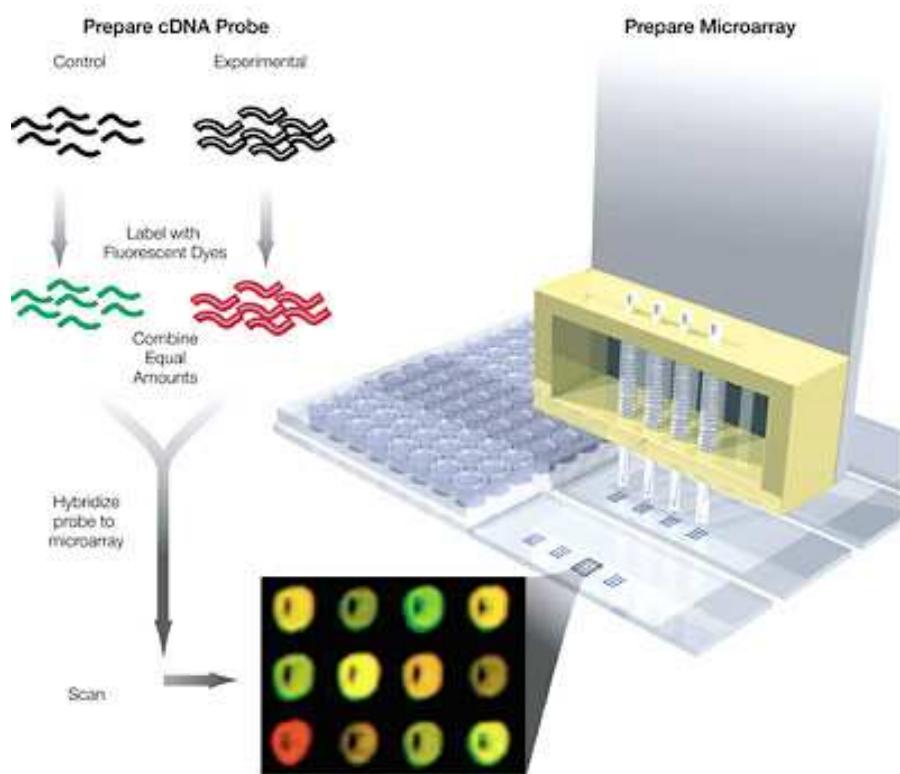


Figure 8.10: Spotted arrays (cDNA arrays).

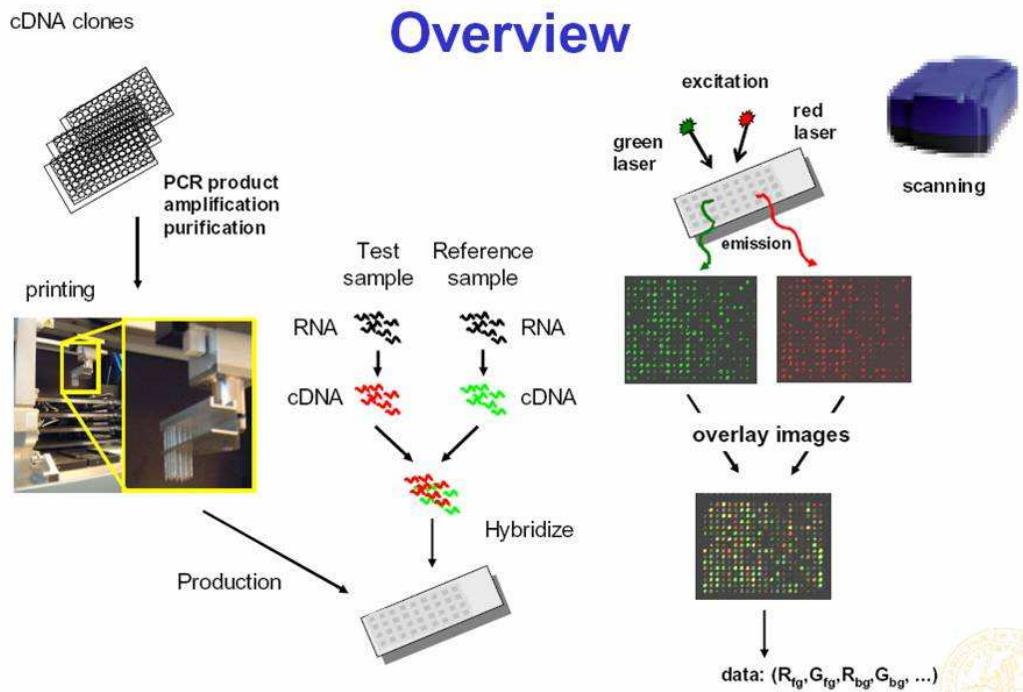


Figure 8.11: Spotted arrays (cDNA arrays).

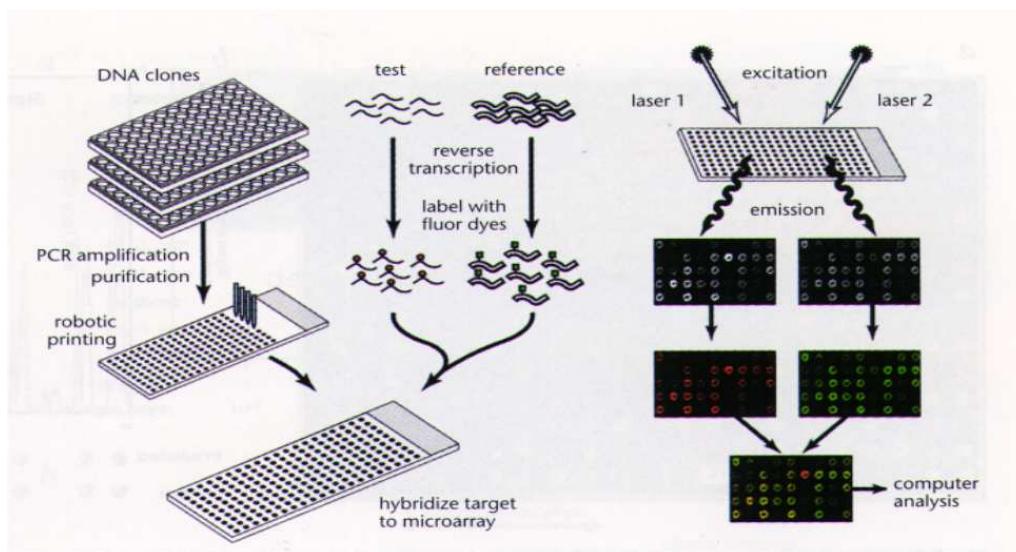


Figure 8.12: Spotted arrays (cDNA arrays).

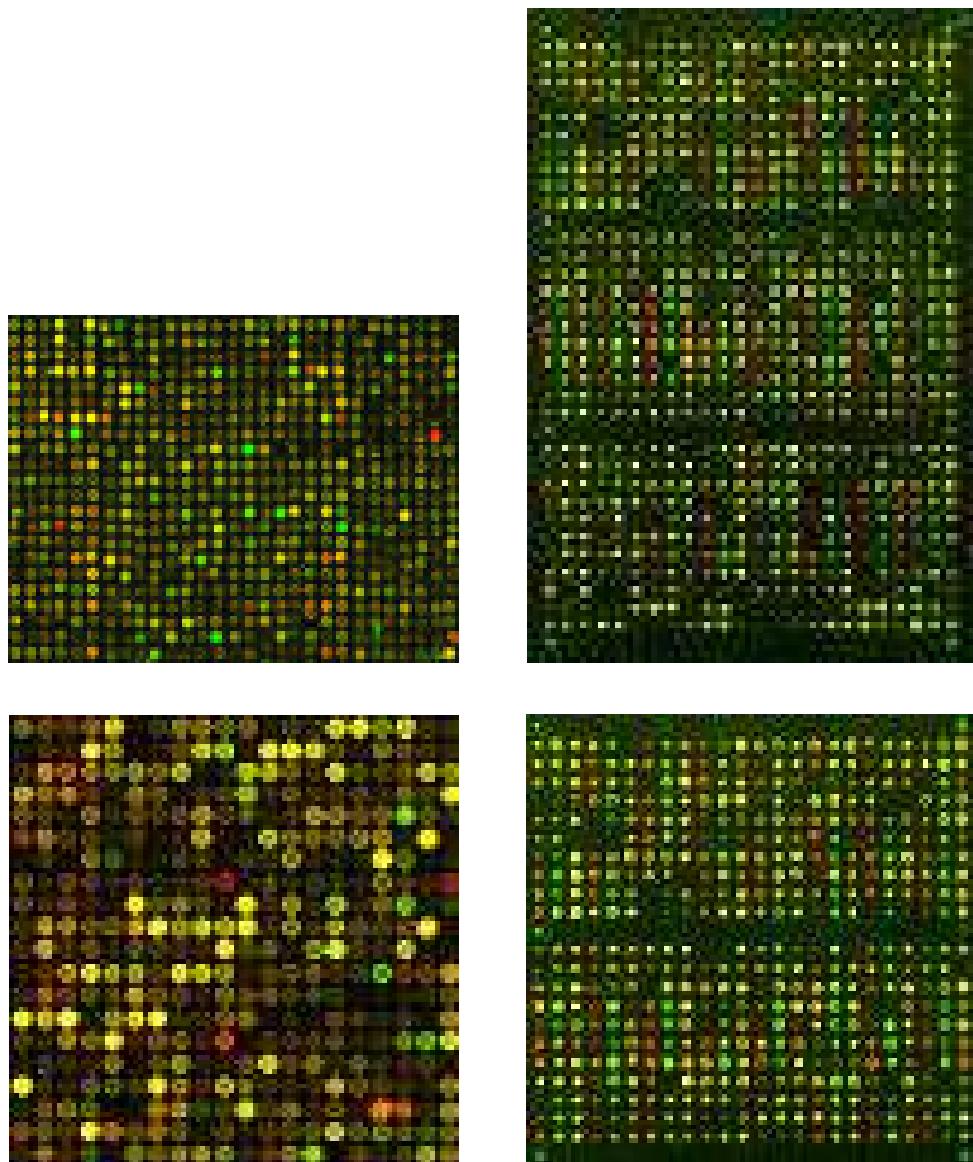


Figure 8.13: Examples of scanner images of red-green spotted arrays.

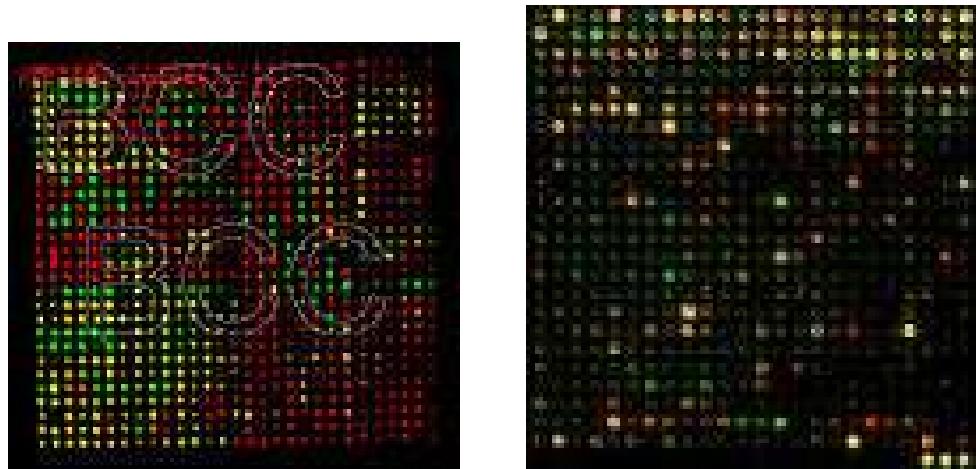


Figure 8.14: More examples of scanner images of red-green spotted arrays.

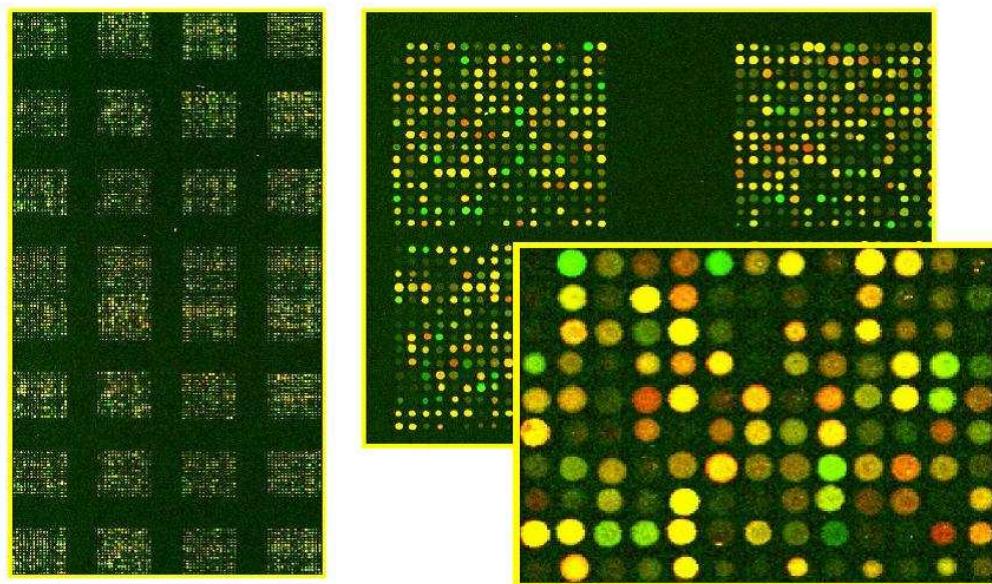


Figure 8.15: Different view on examples of scanner images of red-green spotted arrays.

- create probes and supply them in microtiter
- use robot to spot the microtiter probes on the given regions on the glass slide
- seal the glass slide (deactive the remaining polylysine)
- denature DNA to obtain single stranded probes

The steps of hybridization of a spotted array are

Step 1: Messenger RNA is extracted from the samples

Step 2: mRNA is reversely transcribed to cDNA

Step 3: the cDNA is labeled by Cy3 or Cy5 and transcription results in labeled cRNA

Step 4: the target cRNA is brought onto the glass chip with immobilized probes

Step 5: the target is hybridized with the probes where targets from the samples binds to their corresponding probes (the complementary sequences) on the chip NOTE: this is done twice for Cy3 and Cy5 (for sample and control which are marked with different fluorochromes).

Step 6: After cleaning, the chip is scanned with a confocal microscope and the strength of the green and red fluorescent light is recorded. Genes which are predominantly expressed in the sample give rise to a special color (strong fluorescent light of special color) and genes which are predominantly expressed in controls give rise to the other color.

Step 7: After segmentation of the stained locations on the chip the intensity values of the colors are transformed to real numbers for every location. In many cases a ratio red/green or a log-ratio is used.

8.3.3 Other Techniques

8.3.3.1 SAGE

The serial analysis of gene expression (SAGE) is based on sequencing short unique sequence tags (EST) where the presents of each tag indicates the transcription of a DNA subsequence.

The standard EST methods use transcription segments (tags) of 100 to 300 bases. In contrast SAGE only uses tags of 9 to 14 bases which are located within a gene. The presents of a tag shows that the corresponding gene has been transcribed.

One have to be careful when more genes share the same tag. Also the short tags do not guarantee that the whole gene has been transcribed.

8.3.3.2 Digital Micromirror Arrays

These oligonucleotide DNA arrays are read out by a CCD camera. The whole system is on a chip which is controlled by light beams through a micromirror. The light activates the probes and then the labeled target can be added.

8.3.3.3 Inkjet Arrays

The technique is adopted from standard inkjet printing from Hewlett Packard. Probes can be printed on a glass slide where the probes are pre-synthesized. The probes can also be created nucleotide by nucleotide on the glass slide.

8.3.3.4 Bead Arrays

Small glass beads with attached oligonucleotides are brought into a substrate which is put onto a array. Afterwards the beads have to be located and identified on the array. Then hybridization can be done.

8.3.3.5 Nanomechanical Cantilevers

Oligonucleotide probes are attached on cantilevers of silicon with a gold surface. If targets bind to the probes the cantilevers bend which can be detected by the deflection angle of a laser beam.

8.4 Microarray Noise

Expression values as measured by microarray technique are noisy. The noise has different origin. There exists biological noise, because samples do not show the same “expression state” and exactly the same levels of mRNA even if they belong to the same class or the same experimental condition.

Then there is noise introduced by the microarray measurement technique. Sources of noise include tolerances in chip properties which originate from the fabrication process, different efficiencies for the mRNA extraction and the reverse transcription process, variations in background intensities, nonuniform labeling of the cDNA targets (the dye may bind multiple times and with different efficiencies), variations in the dye concentration during labeling, pipette errors, temperature fluctuations and variations in the efficiency of hybridization, and scanner deviations.

Measurement noise is not always Gaussian. [Hartemink et al., 2001] for example found that the measurement noise distribution of the logarithmic expression values has heavy tails.

8.5 Image Analysis

Image analysis is the first computational step where tools from computer science are applied to improve the results.

Goal is to obtain an intensity value for each spot. The steps for image analysis are (see also Fig. 8.16):

Step 1: spot localization, also called “gridding” (if a grid is aligned to intensity peaks) or “addressing”. See Fig. 8.17 to see that these step can be difficult when spots are overlapping and have different size.

1. Addressing

Locate spot centers.

2. Segmentation

classification of pixels either as signal or background (using seeded region growing or fixed circles).

3. Information extraction

for each spot of the array, calculates signal intensity pairs, background and quality measures.

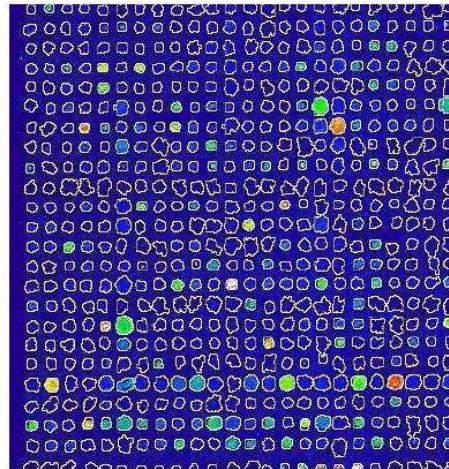


Figure 8.16: The steps of image analysis: 1. spot localization (gridding or addressing), 2. segmentation, and 3. intensity extraction - both background and spot.

Step 2: segmentation to separate spots from the background, where fixed circles, adaptive circles (radius is adjusted), or seeding and then growing can be used (see Fig. 8.18 and Fig. 8.19).

Step 3: intensity extraction - both background and spot.

The background correction which we consider next is sometimes also part of the image analysis.

8.6 Background Correction

As can be seen in Fig. 8.17 at some locations the array is brighter than at other locations. There may be regions where the intensity of the spot is equal to the intensity of the surrounding environment where no probes are attached. Such spots have zero signal as the background, however compared to other regions with zero signal their intensity is large. To obtain the same value for zero signal the background should be subtracted from the signal.

This is one way to perform background correction.

To obtain the background intensity the background can be measured in the image analysis step as depicted in Fig. 8.20.

The different background correction methods are:

- Affymetrix Microarray Suite (MAS) 5.0 (MAS5 [Aff, 2001, Hubbell et al., 2002]): The array divided into 16 rectangular “zones”. The local background is the lowest 2% intensities in the “zones”. This local background is subtracted both from perfect matches and mismatches. Perfect matches and mismatches are kept above a positive threshold

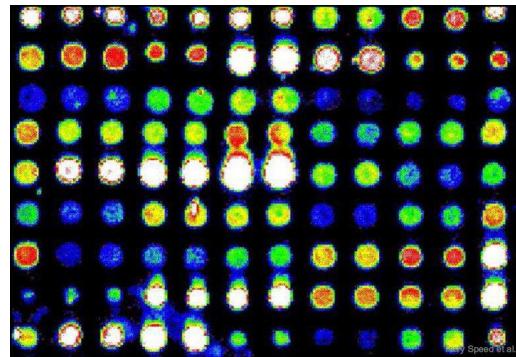


Figure 8.17: An image of a microarray from Terry Speed et al. It can be seen how different(size, shape, intensity) the single spots are – the spots are even overlapping. Sometimes intensities at the background at non-spot locations can be seen.

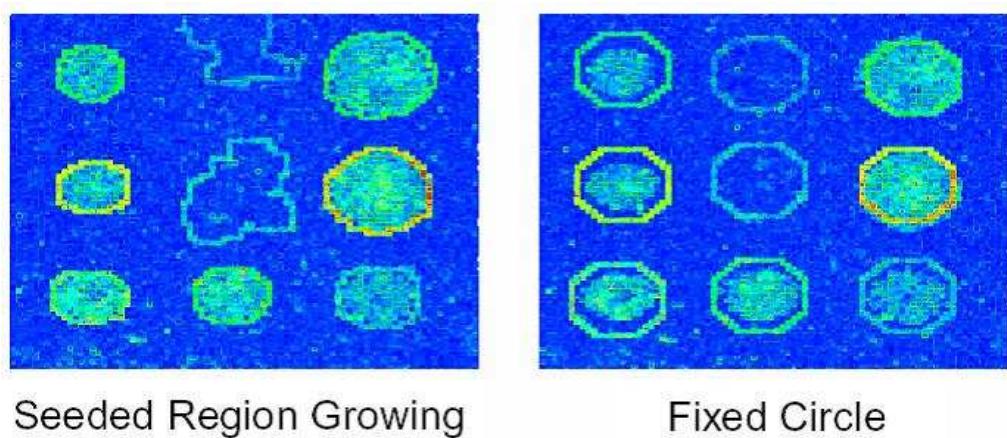


Figure 8.18: Examples of segmentation at the microarray image analysis. Left: growing seeding where a seed is grown until the intensity is decreasing. Right: Fixed circles per spot. The fixed circles do not always match the spot.

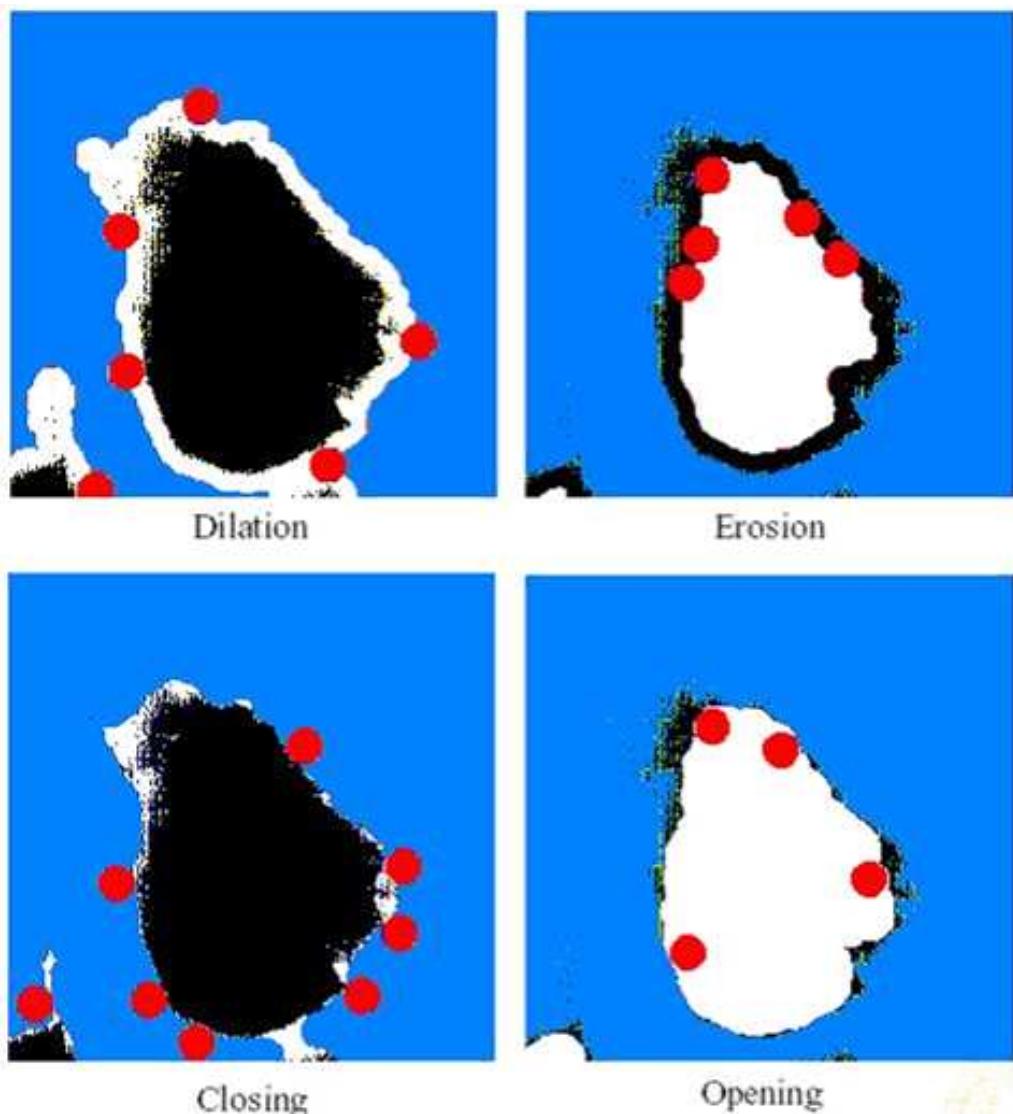


Figure 8.19: Different methods to access the shape of the spot.

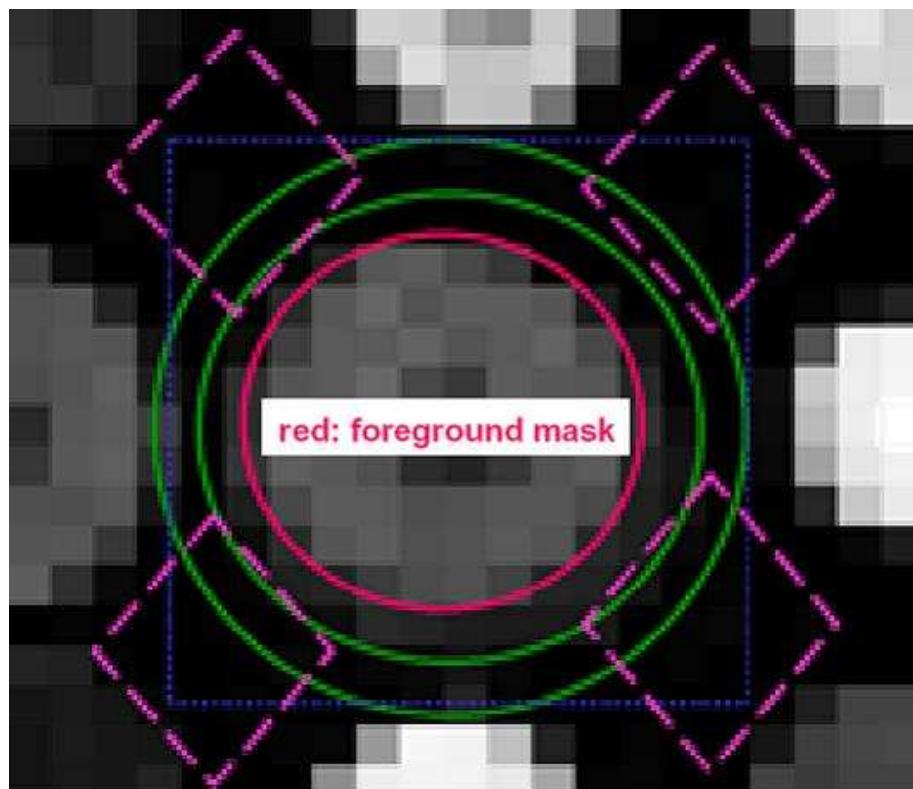


Figure 8.20: The background correction is shown. Red is the foreground mask and pink are the background masks.

- Robust Multi-array Average (RMA, [Irizarry et al., 2003b,a, Bolstad et al., 2003]): The assumption is that the signal S is distributed exponentially and the background B distributed normally. For the background it is assumed that only positive contributions exists, therefore a truncated Gaussian is used.

Let the signal density be

$$p_S(S) = \alpha e^{-\alpha S} \quad (8.1)$$

and the background density be

$$p_B(B) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(B-\mu)^2/(2\sigma^2)}. \quad (8.2)$$

We are now considering the joint density

$$p_{S,O}(S, O) = p_S(S) p_B(O - S). \quad (8.3)$$

Because of

$$\begin{aligned} \frac{(S - (O - \mu))^2}{2\sigma^2} + \alpha S &= \\ \frac{1}{2\sigma^2} (S^2 - 2(O - \mu - \alpha\sigma^2)S + & \\ (O - \mu - \alpha\sigma^2)^2 & \\ + (O - \mu)^2 - ((O - \mu)^2 - 2(O - \mu)\alpha\sigma^2 + \alpha^2\sigma^4)) & \\ = \frac{(S - (O - \mu - \alpha\sigma^2))^2}{2\sigma^2} + (O - \mu)\alpha - \frac{\alpha^2\sigma^2}{2} & \end{aligned} \quad (8.4)$$

the joint density can be written as

$$\begin{aligned} p_{S,O}(S, O) &= \\ \alpha e^{-(O - \mu)\alpha + \alpha^2\sigma^2/2} \frac{1}{\sqrt{2\pi}\sigma} e^{-(S - (O - \mu - \alpha\sigma^2))^2/(2\sigma^2)} &. \end{aligned} \quad (8.5)$$

To obtain the distribution $p_O(O)$ we now can integrate out the variable S . We use

$$a = O - \mu - \alpha\sigma^2 \quad (8.6)$$

$$c = \alpha e^{-(O - \mu)\alpha + \alpha^2\sigma^2/2} \quad (8.7)$$

and obtain (note we have to ensure $O - S \geq 0$)

$$\begin{aligned} p_O(O) &= \int p_{S,O}(S, O) dS = \\ c \frac{1}{\sqrt{2\pi}\sigma} \int_0^O e^{-(S - a)^2/(2\sigma^2)} dS &= \\ c \left(\Phi\left(\frac{O - a}{\sigma}\right) - \Phi\left(\frac{-a}{\sigma}\right) \right) &= \\ c \left(\Phi\left(\frac{O - a}{\sigma}\right) + \Phi\left(\frac{a}{\sigma}\right) - 1 \right), & \end{aligned} \quad (8.8)$$

where Φ is the cumulative distribution of the standard Gaussian ($\Phi(-x) = 1 - \Phi(x)$).

Now we can compute

$$\begin{aligned} E(S | O) &= \int S p_{S|O}(S | O) dS = \int S \frac{p_{S,O}(S, O)}{p_O(O)} dS = \\ &\frac{1}{p_O(O)} \frac{c}{\sqrt{2 \pi} \sigma} \int_0^O S e^{-(S-a)^2/(2 \sigma^2)} dS. \end{aligned} \quad (8.9)$$

We obtain

$$\begin{aligned} &\int_0^O S e^{-(S-a)^2/(2 \sigma^2)} dS = \\ &\sigma^2 \int_0^O \frac{(S-a)}{\sigma^2} e^{-(S-a)^2/(2 \sigma^2)} dS + \\ &a \int_0^O e^{-(S-a)^2/(2 \sigma^2)} dS = \\ &-\sigma^2 \left(e^{-(O-a)^2/(2 \sigma^2)} - e^{-a^2/(2 \sigma^2)} \right) + \\ &a \int_0^O e^{-(S-a)^2/(2 \sigma^2)} dS = \\ &-\sigma^2 \left(e^{-(O-a)^2/(2 \sigma^2)} - e^{-a^2/(2 \sigma^2)} \right) \\ &+ a \sqrt{2 \pi} \sigma p_O(O)/c. \end{aligned} \quad (8.10)$$

Finally we obtain

$$\begin{aligned} E(S | O) &= \\ &a + \frac{\sigma^2}{p_O(O)} \frac{c}{\sigma \sqrt{2 \pi}} \left(e^{-(O-a)^2/(2 \sigma^2)} - e^{-a^2/(2 \sigma^2)} \right) = \\ &a + \sigma \frac{\phi(\frac{a}{\sigma}) - \phi(\frac{O-a}{\sigma})}{\Phi(\frac{a}{\sigma}) + \Phi(\frac{O-a}{\sigma}) - 1}, \end{aligned} \quad (8.11)$$

where ϕ is the normal density.

The value α is estimated by the average distance of PM to their mean, μ is the mean of MM values, and σ^2 is the average squared distance of MM values which are below the mean to the mean (note that background is additive and positive).

- Background correction according to Felix Naef: First the difference PM-MM between perfect matches and mismatches which are smaller than 50 (100) are selected. Thereafter a Gaussian is fitted to estimate mean background intensity. These selected small differences PM-MM identify the PMs which have no signal so that the background can be easily extracted.

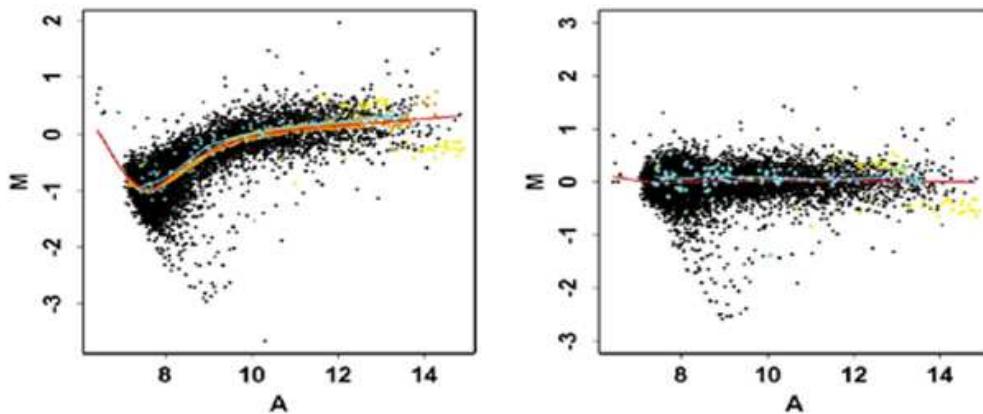


Figure 8.21: MvA plots. Left: MvA plot is not centered around $M = 0$ as the mean curve shows. Right: here the MvA plot is centered around $M = 0$ as desired. Copyright ©Oxford University Press; from [Yang et al., 2002].

8.7 Normalization

Most microarray applications include the comparison of more arrays. For example different array correspond to different conditions, to different time points in the development of the organism, or to different live cycles (cell multiplication).

The biologist or medical researcher is interested in differentially expressed genes (genes which are relevant for a certain condition) or in the change of certain expression levels.

Towards this goal the arrays have to be comparable. However different arrays have different intensity levels.

Also if the same condition is measured twice and the average intensity is the same, the distribution of intensity values may differ, e.g. one array has much more outliers in the high intensity region.

An ‘‘M vs. A’’ or MvA plot shows the difference between the chips. Here M is the difference of 2 log-expressions (difference), $M = \log p_1 - \log p_2$, and A is the average of 2 log-expressions (intensity), $A = 0.5(\log p_1 + \log p_2)$. The data points should be around $M = 0$ because then the average log-intensities are the same. The value A gives the intensity level. For each intensity level M should be centered around zero. That means for each intensity level some array one and some array two show high intensity and not always the same array. Fig. 8.21 shows an MvA plot occurring for two chips (left) and the desired MvA plot (right).

Normalization techniques are:

- **Affymetrix:** First, the highest and the lowest 2% probes per array are excluded. Then a baseline array is chosen. Then the average intensities of all arrays are globally scaled to this baseline array. Arrays are normalized to the median (over the arrays) mean index.
- **Invariant Difference Selection (IDS, [Schadt et al., 2001]):** First, probes pairs which have same order according to their intensity differences PM-MM in an array and in a base-

line array (the median) are said to be invariant. Invariant probe pairs are likely to be not differentially expressed.

$$R_i = \frac{(L(B_i + E_i) + H(2N - B_i - E_i))}{2N} \quad (8.12)$$

$$D_i = \frac{2|B_i - E_i|}{B_i + E_i}, \quad (8.13)$$

where L and H are the rank difference thresholds for low and high ends of the difference range; B_i and E_i are the ranks for the i -th difference (PM-MM) of baseline and array – N is the total number of differences. R_i is the threshold difference intensity i by linearly interpolating the threshold between L and H . D_i is the rank difference test statistic used to determine if the i -th difference should be included in the invariant set. For $D_i < R_i$ the i -th difference is viewed as invariant.

Next the relation of these invariant set of genes is fitted with smoothing splines with generalized cross-validation (GCVSS) along the intensity values (see colored curve in left subfigure of Fig. 8.21). Then the approximation is used to remap the intensities of the array to obtain a plot like in right of Fig. 8.21.

- **Quantile normalization (RMA):** First the PMs are sorted per array. Then each sorted array is put one in one line. This gives sequences above each other like for multiple alignment. Then the median per column is computed. Then all values in a column are to the median. Finally, each array has the same intensity distribution.
- **Cyclic loess:** Local regression (“loess”) fits the MvA data as the colored curve in left subfigure of Fig. 8.21). This curve is used to map the intensity values back to linear scale if compared to a reference array. The predicted loess value is subtracted from the data to decrease the standard deviation and place the mean log ratio at 0. See Fig. 8.22 for the back-mapping. Cyclic loess does this normalization for pairs of arrays. Finally averages are made for the resulting M and A values.

Loess or lowess [Cleveland, 1979, Cleveland and Devlin, 1988] fits simple models to localized subsets of the data. At each point in the data set a low-degree polynomial is fit to a subset of the data using weighted least squares, giving more weight to points near the point whose response is being estimated and less weight to points further away. The LOESS fit is complete after regression function values have been computed for each of the n data points. Often the polynomials are linear or quadratic so that one obtains a local linear or local quadratic model. The weighting function is

$$w(x) = \begin{cases} (1 - |x|^3)^3 & \text{for } |x| < 1 \\ 0 & \text{for } |x| \geq 1 \end{cases}. \quad (8.14)$$

Loess normalization is computationally very intensive.

Normalization like quantile normalization assume that almost all probes on the array show constant expression level. Few expression values change with the conditions.

Note, that for cDNA arrays with Cy3 and Cy5 the intensity may also be dye dependent.

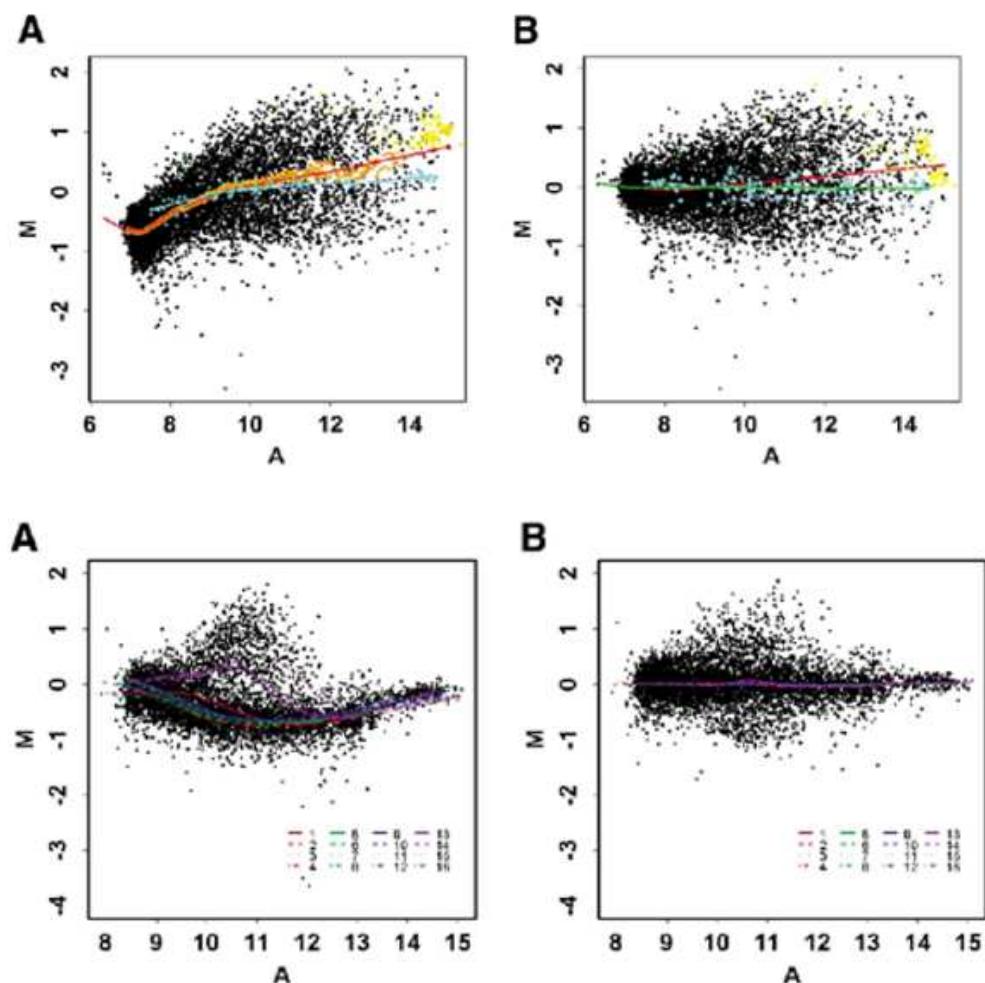


Figure 8.22: Per line MvA plots for original data (left) and data mapped to a linear scale (right) are shown. Before back-mapping a curve is fitted (see left). Copyright ©Oxford University Press; from [Yang et al., 2002].

8.8 PM Correction

At this step the perfect matches (PMs) and the mismatches are combined to obtain one value for each probe pair. This step is of interest especially for the Affymetrix technique.

The basic techniques are:

- **Simple Differences:** PM - MM
- **Ideal mismatch (IM) values (MAS5):**

$$IM_l = \begin{cases} MM_l & \text{for } PM_l > MM_l \\ \exp(-SB) PM_l & \text{for } PM_l \leq MM_l \end{cases} \quad (8.15)$$

$$SB = \frac{\tau}{1 + 0.1(\tau - SB_1)} \quad (8.16)$$

$$SB_1 = TB(\log(PM_j) - \log(MM_j), 1 \leq j \leq N), \quad (8.17)$$

where “TB” is the Tukey’s biweight estimation. Then the differences between PM and IM are computed.

Tukey’s biweight of \mathbf{x} with parameters c ($c = 5$) and ϵ ($\epsilon = 0.0001$) is computed as

$$m = \text{median}(\mathbf{x}) \quad (8.18)$$

$$s = \text{median}(|x_i - m|) \quad (8.19)$$

$$u_i = \frac{x_i - m}{c s + \epsilon} \quad (8.20)$$

$$w_i = (1 - u_i^2)^2 \quad (8.21)$$

$$TB(\mathbf{x}, c, \epsilon) = \frac{\sum_i w_i x_i}{\sum_i w_i}. \quad (8.22)$$

8.9 Summarization

In the summarization step an expression level per probe set should be produced, i.e. an expression level for each gene. Summarization supplies one value per probe set, where the measurement values of the probe pairs are summarized.

- **MAS5:** MAS 5.0 uses Tukey’s biweight function applied to $\log_2(\text{PM} - \text{IM})$.
- **Model Based Expression Index (MBEI, [Li and Wong, 2001]):** Least square fit the linear model

$$\text{PM}_{ij} - \text{MM}_{ij} = y_{ij} = \theta_i \phi_j + \epsilon_{ij}, \quad (8.23)$$

where θ_i is the expression index, ϕ_j is the probe pattern. Parameter estimation is done via the Li-Wong algorithm:

$$\hat{\theta}_i = \frac{\sum_{j=1}^J y_{ij} \phi_j}{\sum_{j=1}^J \phi_j^2} \quad (8.24)$$

$$\hat{\phi}_j = \frac{\sum_{i=1}^I y_{ij} \theta_i}{\sum_{i=1}^I \theta_i^2}. \quad (8.25)$$

These formulas can be derived from the squared error

$$R_{\text{emp}} = \sum_{ij} (y_{ij} - \theta_i \phi_j)^2 \quad (8.26)$$

which derivatives with respect to the parameters are set to zero:

$$\frac{\partial R_{\text{emp}}}{\partial \theta_i} = 2 \sum_j (y_{ij} - \theta_i \phi_j) \phi_j = 0. \quad (8.27)$$

Solving this equation for θ_i results in eq. (8.24). Analogously eq. (8.25) can be derived.

The solution is not determined up to scaling ϕ_j and θ_i , e.g. $\phi_j = \tau \phi_j$ and $\theta_i = \frac{1}{\tau} \theta_i$. This degree of freedom is used up by rescaling the parameters:

$$\hat{\phi}_j = \phi_j \sqrt{\frac{\sum_{j=1}^J \phi_j^2}{J}} \quad (8.28)$$

$$\hat{\theta}_i = \theta_i \sqrt{\frac{J}{\sum_{j=1}^J \phi_j^2}}. \quad (8.29)$$

Note, that eq. (8.24) and eq. (8.25) are not exact, because they have to be solved simultaneously. That is for computing θ_i the new ϕ_j must be used and vice versa. However if for computing θ_i the old ϕ_j are used, an iterative algorithm is obtained.

- **Robust Multi-array Average (RMA):** An additive model is fitted by median polish.
- **Factor Analysis for Robust Microarray Summarization (FARMS, [Hochreiter et al., 2006]):** The summarization problem is based on a linear model with Gaussian noise. The linear model is a factor analysis model with one hidden factor representing the mRNA concentration.

The zero mean normalized log-PMs are denoted by x . The log-RNA concentration is denoted by z .

The model is

$$x = \lambda z + \epsilon, \text{ where } x, \lambda \in \mathbb{R}^n \text{ and} \quad (8.30)$$

$$z \sim \mathcal{N}(0, 1), \epsilon \sim \mathcal{N}(\mathbf{0}, \Psi). \quad (8.31)$$

$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multidimensional Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ ($\mathcal{N}(0, 1)$ is the one-dimensional standard Gaussian). z is usually called a “factor”. $\boldsymbol{\Psi} \in \mathbb{R}^{n \times n}$ is the diagonal noise covariance matrix while ϵ and z are statistically independent. According to the model, the observation vector \mathbf{x} is Gaussian distributed:

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\lambda}\boldsymbol{\lambda}^T + \boldsymbol{\Psi}). \quad (8.32)$$

The model parameters are estimated by a maximum a posteriori estimation. The prior for $\boldsymbol{\lambda}$ is $p(\boldsymbol{\lambda}) = \prod_{j=1}^n p(\lambda_j)$ and for $p(\lambda_j)$ the rectified Gaussian distribution $\mathcal{N}_{\text{rect}}(\mu_\Lambda, \sigma_\Lambda)$ is used, which is given by

$$\lambda_j = \max\{y_j, 0\} \text{ with } y_j \sim \mathcal{N}(\mu_\Lambda, \sigma_\Lambda). \quad (8.33)$$

The Bayesian posterior $p(\boldsymbol{\lambda}, \boldsymbol{\Psi} | \{\mathbf{x}\})$ of the model parameters $(\boldsymbol{\lambda}, \boldsymbol{\Psi})$ given the data set $\{\mathbf{x}\} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ is proportional to the product of the observation's likelihood $p(\{\mathbf{x}\} | \boldsymbol{\lambda}, \boldsymbol{\Psi})$ of data $\{\mathbf{x}\}$ given the parameters $\boldsymbol{\lambda}, \boldsymbol{\Psi}$ multiplied by the prior $p(\boldsymbol{\lambda}, \boldsymbol{\Psi})$

$$p(\boldsymbol{\lambda}, \boldsymbol{\Psi} | \{\mathbf{x}\}) \propto p(\{\mathbf{x}\} | \boldsymbol{\lambda}, \boldsymbol{\Psi}) p(\boldsymbol{\lambda}, \boldsymbol{\Psi}). \quad (8.34)$$

The prior reflects the facts that:

1. the observed variance in the data is often low which makes high values of λ_j unlikely,
2. a chip typically contains many more genes with constant signal (λ_j approximately zero) than genes with variable signal (large value of λ_j),
3. negative values of λ_j are not plausible, because that would mean that increasing mRNA concentrations lead to smaller signal intensities.

The two hyperparameters ρ and μ_Λ allow to quantify different aspects of potential prior knowledge. For example, μ_Λ near zero assumes that most genes do not contain a signal, and introduces a bias for Λ -values near zero (items 1 and 2 from above).

The model parameters are estimated by an expectation-maximization (EM) algorithm of Dempster et al. [1977]. See for more details the course Bioinformatics II, where the algorithm is explained in more detail.

8.10 Different Combinations of the Processing Steps

The processing steps can be differently put together, e.g. different background correction, or normalization methods can be run with the same summarization method. Fig. 8.23 and Fig. 8.24 show how the different steps can be put together.

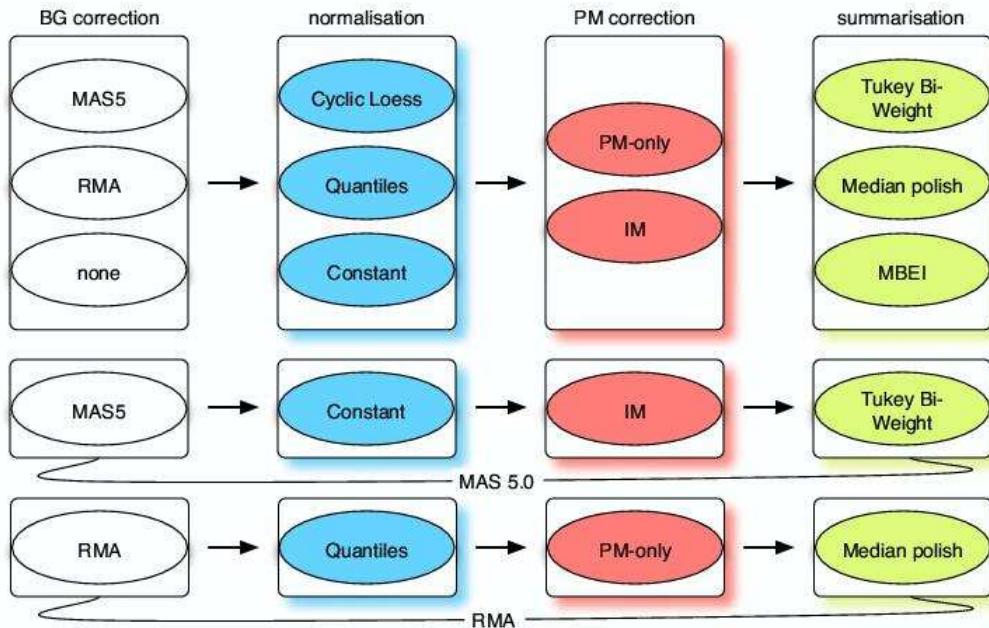


Figure 8.23: Different methods at different processing levels. RMA and MBEI are shown.

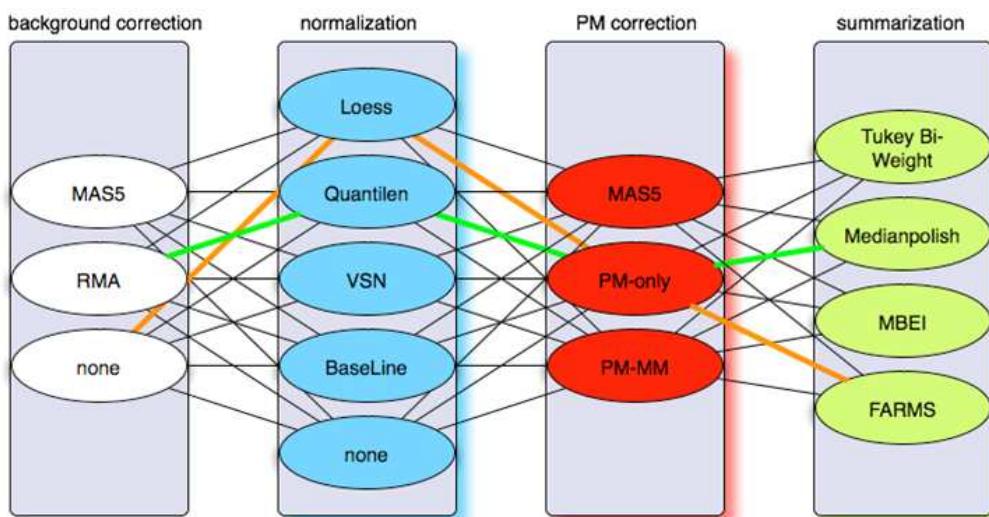


Figure 8.24: Different methods at different processing levels. RMA and FARMS are shown.

8.11 Microarray Gene Selection Protocol

In this section we describe the protocol for extracting meaningful genes from a given set of expression values for the purpose of predicting labels of the sample classes. The protocol includes data preprocessing, the proper normalization of the expression values, the feature selection and ranking steps, and the final construction of the predictor.

Note that our feature selection protocol requires class labels which must be supplied together with the expression values of the microarray experiment. For the following, however, we assume that the task is to select features for classification and that l labeled samples are given for training the classifier.

8.11.1 Description of the Protocol

1. **Expression values vs. log-ratios.** Before data analysis starts it is necessary to choose an appropriate representation of the data. Common representations are based on the ratio $T_j = \frac{R_j}{G_j}$ of expression values between the value R_j (red) of a gene j in the sample to analyze and the value G_j (green) in the control sample, and the log ratio $L_j = \log_2(T_j)$.

For arrays like the Affymetrix chips only one kind of expression level which indicates the concentration of the according mRNA in the sample. Here also the log expression value is a common choice for representing the expression levels.

2. **Normalization and Summarization.** Different normalization methods exist to make the measurements from different arrays comparable. The most famous ones are “Quantile normalization” and “Cyclic Loess”.

Some techniques like the Affymetrix GeneChip® makes more than one measurement for each gene. A so-called probe set makes 11 to 21 measurements for each gene. To obtain a single expression value these measurements must be “summarized”. Here also different summarization methods like RMA, GCRMA, MBEI, MAS5.0, or FARMS exist. Some methods like FARMS are able to supply a present call, which is discussed in next item.

3. **Present call.** The present call is usually the first step in the analysis of microarray data. During this step genes are identified for which the confidence is high, that they are actually expressed in at least one of the samples. Genes for which this confidence is low are excluded from further processing in order to suppress noise.

For this purpose an error model has to be constructed for the expression values or their ratios (sometimes before, sometimes after averaging across multiple measurements of the same sample — see [Tseng et al., 2001, Schuchhardt and Beule, 2000, Kerr et al., 2000, Hartemink et al., 2001]). This error model accounts for both measurement specific noise (for example background fluctuations), which affects all expression values in a similar way, and gene specific noise (for example the binding efficiency of the dye), which affects expression values for different genes in a different way. Using this error model one assigns a p -value, which gives the probability that the observed measurement is produced by noise, to every measurement of an expression level. If the P -value is smaller than a threshold q_1 (typical values are 5%, 2%, or 1%), the expression level is marked “reliable”. If this happens for a

minimum number q_2 (typical values range from 3 to 20) of samples, the corresponding gene is selected and a so-called present call has been made.

4. **Standardization.** Before further processing, the expression values are normalized to mean zero and unit variance across all training samples and separately for every gene. Standardization accounts for the fact that expression values may differ by orders of magnitudes between genes and allows to assess the importance of genes also for genes with small expression values.

Some summarization methods may already account for comparable variance and zero mean – in this case standardization is not necessary.

5. **Gene ranking and gene selection.** Here we assume that a feature selection method has been chosen where the size of the set of selected genes is controlled by a hyperparameter which we call ϵ in the following (according to the P-SVM).

In this step we perform two loops: An “inner loop” and an “outer loop” (the leave-one-out loop). The inner loop serves two purposes. It ranks features if only a subset method like the P-SVM is available and it makes feature selection more robust against variations due to initial conditions of the selection method. The outer loop also serves also two purposes. It makes the selection robust against outlier samples and allows to determine the optimal number of selected genes together with the optimal values of hyperparameters for the later prediction of class labels. In order to do this, a predictor must be constructed. Here we suggest to use a ν -SVM where the value of ν is optimized by the outer loop. In order to implement the outer (leave-one-out) loop, l different sets of samples of size $l - 1$ are constructed by leaving out one sample for validation. For each of the l sets of reduced size, we perform gene selection and ranking using the following “inner loop”.

Inner loop. The subset selection method is applied multiple times to every reduced set of samples for different values of ϵ . For every set of samples multiple sets of genes of different size are obtained, one for every value of ϵ . If the value of ϵ is large, the number of selected genes is small and vice versa. The inner loop starts with values of ϵ which are fairly large in order to obtain few genes only. Gradually the value is reduced to obtain more genes per run. Genes obtained for the largest value of ϵ obtain the highest rank, the second highest rank is given to genes which additionally appear for the second largest value of ϵ , etc. The values of ϵ are constant across sample sets. The minimal value should be chosen, such that the number of extracted genes is approximately the total number l of samples. The maximal value should be chosen such that approximately five to ten genes are selected. The other values are distributed uniformly between these extreme values.

Outer loop. The results of the inner loops are then combined across the l different sets of samples. A final ranking of genes is obtained according to how often genes are selected in the l leave-one-out runs of the inner loop. If a gene is selected in many leave-one-out runs, it is ranked high, else it is ranked low. Genes which are selected equally often are ranked according to the average of their rank determined by the inner loops. The advantage of the leave-one-out procedure is that a high correlation between expression values and class labels induced by a single sample is scaled down if the according sample is removed. This makes the procedure more robust against outliers.

The outer loop is also used for selecting an optimal number of genes and other hyperparameters. For this purpose, ν -SVMs are trained on each of the l sets of samples for different

values of the hyperparameter ν and the number F of high ranking genes (ranking is obtained by the inner loop). Then the average error is calculated on the left out samples. Since the leave-one-out error as a function of the number F of selected genes is noisy, the leave-one-out error for F is replaced by the average of the leave-one-out errors for F , $F + a$, and $F - a$. Then the values of the hyperparameter ν and the number of genes F which give rise to the lowest error are selected. This completes the feature selection procedure.

8.11.2 Comments on the Protocol and on Gene Selection

Normalization and Summarization of new arrays. If a new array has to be analyzed then all known arrays together with the new array must be normalized and used for summarization. Thereafter machine learning methods can be applied to the training set and the new array can be classified.

Corrections to the outer, leave-one-out loop. The samples which were removed from the data in the outer loop when constructing the l reduced subsets for the gene ranking should not be considered for the present call and for determining the normalization parameters. Both steps should be done individually for each of the l sets of sample, otherwise feature or hyperparameter selection may not be optimal.

Computational Costs. The feature selection protocol requires $l \times n_\epsilon$ feature selection runs, where n_ϵ is the number of different values of the ϵ parameter. However the computational effort is justified by the increased robustness against correlation by chance (see next item) and the elimination of single sample correlations.

Correlations by chance. “Correlations by chance” refers to the fact, that noise induced spurious correlations between genes and class labels may appear for a small sample size if the level of noise is high. If the number of selected genes is small compared to the total number of probes (genes) on the chip, spurious correlations may become a serious problem. Monte-Carlo simulations of van’t Veer et al. [2002] on randomly chosen expression values for a data set of 78 samples and 5000 genes resulted in 36 “genes” which had noise induced correlation coefficients larger than 0.3. In order to avoid large negative effects of above-mentioned spurious correlations the number of selected genes should not be too small, and one should extract a few tens of genes rather than a few genes only to decrease the influence of single spurious correlated genes. The random correlation effect can also be reduced, by increasing q_2 , the minimum number of “reliable” expression values for making a present call. This avoids the selection of genes for which too few samples contribute to the correlation measure. However as explained in the next paragraph too many genes should be avoided as well.

Redundancy. Redundant sets of genes, that is sets of genes with correlated expression patterns should be avoided in order to obtain good machine learning results [Jäger et al., 2003]. Selection of too many genes with redundant information may lead to low generalization performance. Another reason for avoiding redundancy is that not all causes which imply the conditions may be recognized. This may happen if the set has to be kept small while redundant genes are included (redundant genes indicate the same cause). Reducing redundancy does not preclude the extraction of co-expressed clusters of genes: co-regulated genes can be extracted in a subsequent processing step, for example based on classical statistical analysis.

Finally, one may wonder why redundant information does not help to decrease the noise level of otherwise informative genes. Empirically one finds that non-redundant feature selection meth-

ods (P-SVM and R2W2) outperform feature selection methods which include redundant genes (Fisher correlation and RFE). It seems as if the detrimental effects of a larger number of features are stronger.

8.11.3 Classification of Samples

In order to construct a predictor for the class labels of new samples a classifier is trained on all the l samples using the optimal set of genes and the optimal value of the hyperparameter (here: ν , cf. Step 5). The generalization performance of the classifier can again be estimated using a cross-validation procedure. This procedure must involve performing the full gene selection procedure including all preprocessing steps (for example normalization and feature selection) separately on all l cross-validation subsets. Otherwise a bias is introduced in the estimate. Note that this also requires to perform the “inner loop” of Step 5 on sets of $(l - 2)$ samples.

Before the classifier is applied to new data, the expression values for the new sample must be scaled according to the parameters derived from the training set. As a consequence we may observe expression values which are larger than the ones which occur in the training data. We set the expression values exceeding the maximal value in the training set to this maximal value. With this procedure we may underestimate certain expression levels but the robustness against unexpected deviations from the training data is increased.

Chapter 9

DNA Analysis

Genome sequencing resulted in whole genomes (all chromosomes) of some selected organisms (over 100) like yeast, rat, mouse, chimpanzee, wolf, fruit fly, bacteria, and human.

The DNA contains all information of life there are the genes stored which are the blueprints for building the nano-machines in the cell, the proteins and protein-RNA complexes. However the DNA also codes for small RNA strands which regulate the production of proteins. These RNA functions are currently investigated and are known as interfering RNA (iRNA), microRNA, or non-coding RNA (ncRNA).

The DNA contains highly repetitive and redundant sequences which function is not known.

The genes of eukaryotes are not coded as single sequences but as a sequences of coding (exons) and non-coding (introns) subsequences. After transcription the resulting preRNA is edited by spliceosomes which remove the introns from the sequences and glues together the exons. After this “splicing”, the product is the mRNA.

But the splicing process may splice out an intron or not at certain positions depending on the environment in the cell. That means the exons as building block are assembled according to the current needed protein. Therefore it is not true that one gene codes for one protein but one gene can code many proteins depending on which introns are spliced out. Actually splicing can be more complicated than only skipping some introns.

Some genes can move within the genome from one location to another. Mostly these locations are marked by highly repetitive sequences. Moving the genes on the chromosome can disrupt the genes or regulation at the position a gene is inserted but can have biological advantage as the genes is now regulated by other control mechanisms.

The genome is highly variable. For example the number of genes and where they are placed is different in a genome. Through duplications the genes are doubled and regulation mechanisms may have different effects. These rearrangements and shuffling the genomic content can be compared between different species. These addresses the task of whole genome comparisons and whole genome mapping.

The DNA varies among the individuals of a species. E.g. if the DNA of humans is compared then on average every 500 bases a nucleotide is different. This differences makes us humans individual from external features (how tall, shape of the face, etc.) to internal metabolic features (reaction to special food etc.). If these differences occur in at least one percent of the population then it is called “single nucleotide polymorphism” (SNP). Some of these polymorphism are one out of more causes for diseases, e.g. schizophrenia or alcohol dependence is related to SNPs. For

some people salty food leads to an increase of blood pressure but not for others. This effect is also due to SNPs where it is assumed that people living near the sea had an advantage if having a SNP which allows to cook with salt. Another SNP can be found which indicates how well milk sugar can be processed especially by elderly people. However in Africa the SNP is not found meaning that Africans cannot process as well milk. Here the opinion is that regions where cows gave a lot of milk the natives adjusted to this food by selecting the corresponding SNP.

Offsprings inherit the SNPs of both parents but as whole blocks, that means SNP are inherited together, i.e. block-wise. These blocks are the haplotype blocks of few thousands to hundreds of thousands of bases.

However there are other individualities in our genome like the repetitive sequences at certain positions differ from human to human.

The challenge for bioinformatics is to detect all these information on the DNA, compare the relevant information between species or within species, relate these informations with conditions or with diseases, analyze the informations on the DNA, and make prediction on the behavior of a individual to some external stimuli based in its information on the DNA. The later addresses for example in humans how an individual responds to special treatments or to special medication (how does the body consume the medication, how does the body react to it, how are the long term effects).

9.1 Genome Anatomy

- Prokaryotes: circular DNA in compact form
- Eukaryotes: chromosomes, in the nucleus, wrapped around protein complexes called nucleosomes

The first prokaryotic genomes to be sequences were *Hemophilus influenzae* and *E. coli* where it was found that 58% of the genes match in the two genomes.

The first whole genome which was sequenced was a viral genome, a bacteriophage, containing 11 genes in 1977 by Fred Sanger. In 1981 the human mitochondrion was sequenced by Anderson et al., which contains 16,568 base pairs coding 13 proteins, 2 ribosomal RNAs, and 22 tRNAs. Today of 400 mitochondrial genomes are sequenced. Plant chloroplast organelle gnomes were sequenced in 1986 with 120 to 200 kbp. The first eukaryotic chromosome was obtain in 1992 by Oliver et al. who sequenced the *S. cerevisiae* (yeast) of 315 kpb with 182 genes. The sequencing was enhanced by the whole-genome shotgun sequencing technique (Fleischmann et al.) in 1995 where *H. influenzae* with 1,830 kbp (1,83 Mbp) with 1,743 genes was sequenced.

The average gene density in the human genome is 1 gene per 80kbp which reduces to 1 gene per 40kbp if sequence repeats are skipped.

The genes in the human genome are clustered in 30 “ridges” which have high gene density, high GC content, and high SINE (see Subsection 9.5.1) repeat density whereas low LINE (see Subsection 9.5.1) repeat density.

Also the genomes of chicken, wolf (dog), chimpanzee, honey bee, etc. are available.

Organism	Group	Genome (Mbp)	Genes	kb containing one gene
<i>Methanococcus jannaschii</i> 1996	archaea	1.66	1,682	0.99
<i>Escherichia coli</i> 1997	bacteria	4.6	4,288	1.07
<i>Hemophilus influenzae</i> 1995	bacteria	1.83	1,743	1.05
<i>Mycoplasma pneumoniae</i> 1996	bacteria	0.82	676	1.21
<i>Bacillus subtilis</i> 1997	bacteria	4.2	4,098	1.02
<i>Aquifex aeolicus</i> 1998	bacteria	1.55	1,512	1.03
<i>Synechocystis sp.</i> 1996	bacteria	3.57	3,168	1.13
<hr/>				
<i>Arabidopsis thaliana</i>	plant	125	25,000	5.0
<i>Caenorhabditis elegans</i>	worm	100	18,424	5.43
<i>Drosophila melanogaster</i>	fruit fly	180	13,601	13.23
<i>Saccharomyces cerevisiae</i>	budding yeast	13.5	6,241	2.16
<i>Homo sapiens</i>	human	2900	> 30,000	96.67

Table 9.1: Genomes of different species. The empty line separates prokaryotes from eukaryotes. Given are the size of the genome in mega base pairs and the predicted number of genes.

Heterochromatin and Euchromatin

In eukaryotic cells the chromosomes have lightly and darkly stained regions called “heterochromatin” and “euchromatin”, respectively. Heterochromatin regions are not transcribed whereas euchromatin is. Heterochromatin regions are at the centromeres (where the two strands of the chromosome are attached to each other) and at the telomeres forming the chromosome ends. The dense packing in the heterochromatin regions barres access of the transcription factors.

Pseudo-genes

A mutation event is the duplication of genes (or duplications of DNA regions). The gene is then multiple present in the genome which allows for differentiation and fine adjustment of gene regulation. If the gene is only needed once then the other gene may mutate until it loses its function and is a *pseudo-gene*. Pseudo-genes are gene copies which lost their functions. A gene copy can also be made inactive by lacking a promoter which is also a pseudo-gene. Often these pseudo-genes without a promoter do not possess introns. Therefore they probably are created by mRNA reverse transcription and insertion into the DNA (e.g. by LINE1 reverse transcripts). Most pseudo-genes of the later class are housekeeping genes like genes coding ribosomal proteins (see <http://www.pseudogene.org>).

9.2 Gene Finding

Gene finding methods are based on hidden Markov models or on neural networks.

However the different organism have gene codon preferences and splice junctions, therefore each genome requires a model trained to its specific characteristics.

After whole genome sequencing, first “open reading frames” (ORFs) are identified. ORFs are sequences with a start codon (methionine), a sequence of possible codons, and a stop codon. Reading frames have to be controlled in the forward and the backward direction and with three starting positions. ORFs can be checked by homology search for a known gene, for codon usage specific for the organism, for codon statistics like pairwise codon frequency, the GC content gives a bias in the third (least important) codon position. The programs TESTCODE and CODONFREQUENCY check for these ORF characteristics.

Genes can especially identified in prokaryotes through ORFs because they have few introns. However in eukaryotes the problem is more difficult because of introns which can have extended sequences.

In eukaryotes first the promoter regions have to be identified then the introns have to be determined and removed. Thereafter the mRNA sequences (the ORFs) can be translated from first start codon to the first stop codon.

That means also computer models for intron recognition have to be constructed.

9.2.1 Hidden Markov models

Hidden Markov models (HMMs) for gene finding include GeneMark, GeneMark.hmm, GLIMMER, GRAIL, GenScan / GenomeScan, and Genie. These HMMs attempt at detecting the boundaries of the coding region of a gene which is indicated by splice sites, start and stop codons, transcription factors, protein binding sites like the TATA-box, transcription start points, branch points, transcription termination sites, polyadenylation sites (which prevent mRNA from degradation), ribosomal binding sites, topoisomerase I cleavage sites, topoisomerase II biding sites, and other. The HMMs are constructed hierarchically through region modules like exon modules, intron modules, and inter-genic modules. The exon module can be subdivided into initial, internal, and terminal exons. Other modules can account for repetitive regions. Fig. 9.1 (GLIMMER software) and Fig. 9.2 (GENEZILLA software) show HMMs for gene finding.

GLIMMER is an interpolated Markov Model which searches for long known patterns. Long frequent patterns are modeled by higher order Markov models whereas short patterns are modeled by low order Markov models. Longer matching pattern obtain higher probability than shorter patterns – these probabilities are combined in the final model. Because the model is adaptive according to the frequency of the occurrence of a pattern it avoids the computational complexity of high order Markov models which have to consider all combinations of the pattern they are conditioned on. GLIMMER is a hybrid model between pattern recognition (long frequent pattern) and probabilistic modeling (short pattern).

The HMM models often contain neural networks which approximate the transition probabilities $p(s_t^i \mid s_{t-1}, \dots, s_{t-\tau})$ by a function $f(s_{t-1}, \dots, s_{t-\tau}) = f(s)$ which may be a softmax

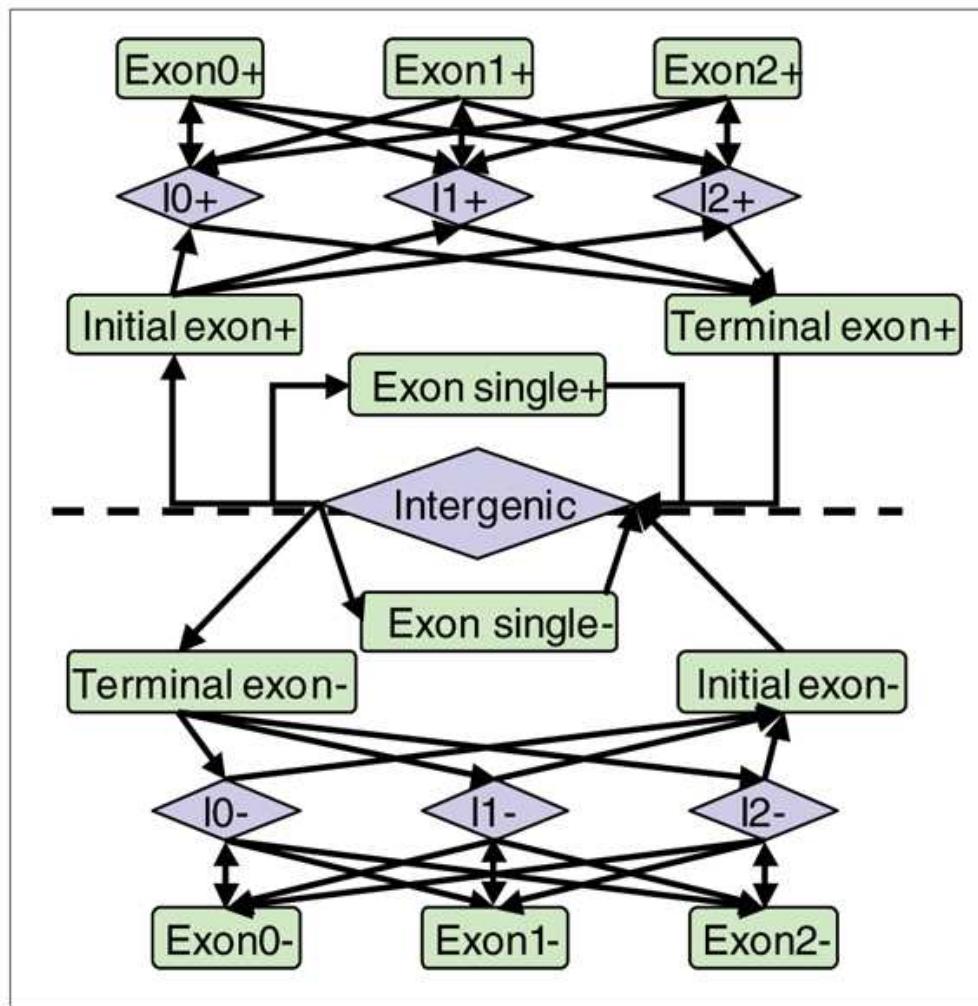


Figure 9.1: The GLIMMER hidden Markov model for gene finding. State-transition diagram. Each state in the HMM is implemented as a separate submodel, such as a weight array matrix or an IMM (interpolated Markov models). From Allen et al. Genome Biology 2006 7(Suppl 1):S9.

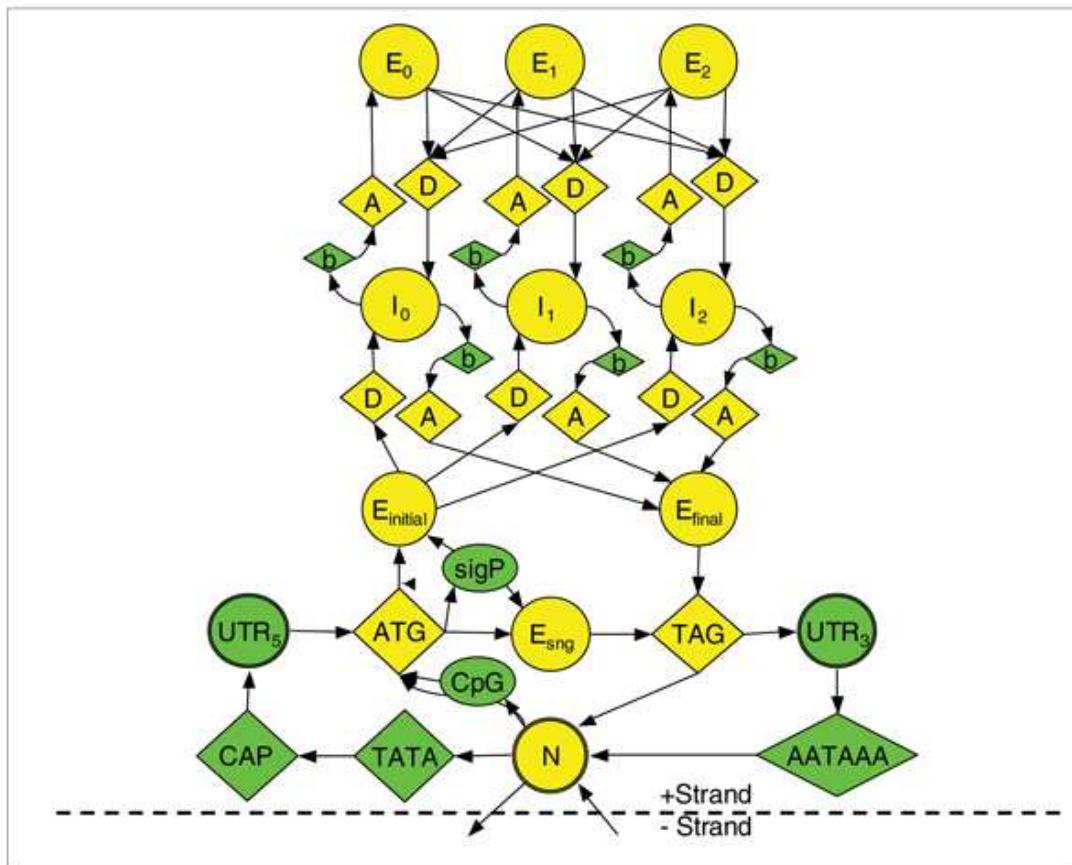


Figure 9.2: The GENEZILLA hidden Markov model for gene finding. State-transition diagram. A, acceptor site; AATAAA, polyadenylation signal (including ATTAAA); ATG, start codon; b, branch point; CAP, cap site; CpG, CpG island; D, donor site; E, exon; I, intron; N, intergenic; sigP, signal peptide; TATA, TATA box; TAG, stop codon (including TAA and TGA); UTR, untranslated region. From Allen et al. Genome Biology 2006 7(Suppl 1):S9.

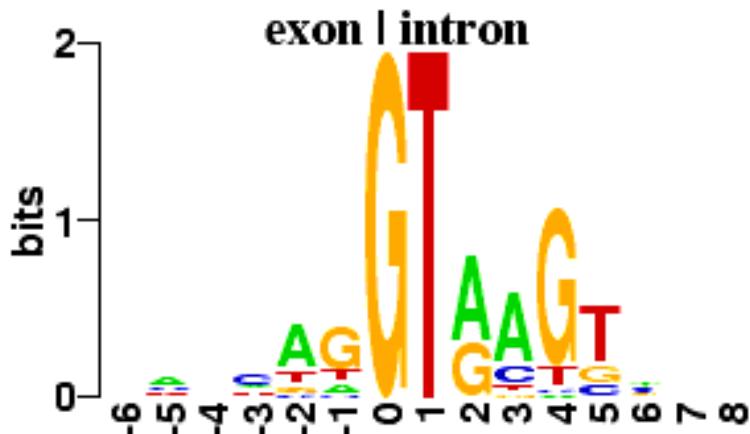


Figure 9.3: A pattern for the exon-intron boundary. This pattern is given in WebLogo format: size of the letter gives its frequency at this position.

function

$$f_i(\mathbf{s}) = \frac{\exp(-y_i(\mathbf{s}))}{\sum_i \exp(-y_i(\mathbf{s}))} . \quad (9.1)$$

9.2.2 Neural networks

A neural network based system for gene finding is GRAIL (<http://grail.lsd.ornl.gov>) which identifies coding-regions. The input for the GAIL neural network are different characteristics which are related to coding/non-coding regions. GRAIL identifies poly-A sites and promoter regions and constructs the protein sequence. Inputs to GRAIL include “score of 6-mers in candidate region”, “score of 6-mers in flanking regions”, “Markov model score”, “flanking region GC composition”, “candidate region GC composition”, “score for splicing acceptor site”, “score for splicing donor site”, “length of region”, etc. The scores are log-likelihood scores of some simple probabilistic models.

GeneParser is a splice site recognition system based on alignment of exon and intron starts and ends. The objective is a log-likelihood score. Splice site indicators are weighted by a neural network because the different alignment scores have to be combined into one score.

NetGene combines the prediction of splice sites with the prediction of coding/noncoding regions with neural networks. Three networks are combined which have an input window of 15, 41, and 301 bp, where the first two networks are donor and acceptor networks and the third is a global network.

The exon-intron and intron-exon boundaries show specific pattern as can be seen in Fig. 9.3 and Fig. 9.4

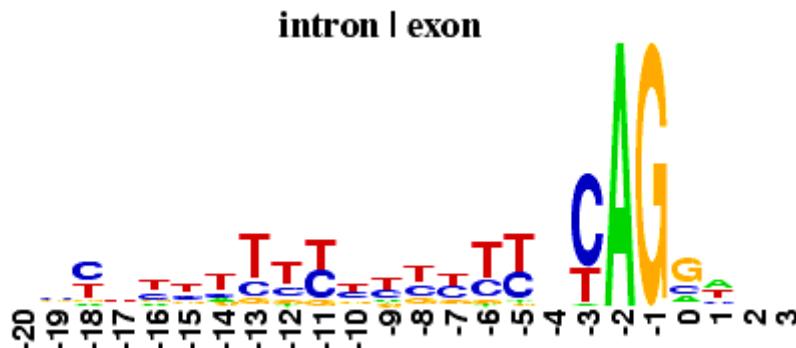


Figure 9.4: A pattern for the intron-exon boundary.

9.2.3 Homology Search

Another method to search for genes is to translate all possible ORFs into amino acid sequences. These sequences can be compared by alignment methods, e.g. BLAST or WU-BLAST, to known sequences. If a match with a low *e*-value (*p*-value) is observed then a gene has been found. There exist special alignment programs which include the translation like BLASTX or FASTAX or include the translation of both the query and the data base entry like TBLASTN or TFASTX.

Note, that local alignment methods may even work if the intron-exon boundaries are not recognized. If a exon is correctly translated then local alignment may find the corresponding exon in an amino acid sequence which is already known.

9.2.4 Promoter Prediction

Promoters are at the 5' end of genes and serve as indicator of the starting regions of genes. Therefore promoter region prediction is important to find genes.

9.2.4.1 Prokaryotes: *E. coli*

Alignment.

To find promoter regions often promoter sequences are aligned using the transcription start site as anchor point. After alignment specific RNA polymerase promoter pattern can be seen for *E. coli* in Fig. 9.5 and Fig. 9.6. Fig. 9.5 shows the TATAAT Pribnow box at position -10 and Fig. 9.6 shows the TTGACA pattern at position -35.

A conservative region exist at +1 and before position -35 there is an AT rich region.

New promoter regions can be found through building a scoring matrix according to the methods in bioinformatics I (e.g. like the position specific scoring matrices of PSI-BLAST).

Neural Networks.

Patterns can be detected by neural networks by using a local code, that is each nucleotide is coded by a vector with 4 components where all components are zero except one component which is one: A = (1,0,0,0), T = (0,1,0,0), G = (0,0,1,0), C = (0,0,0,1).

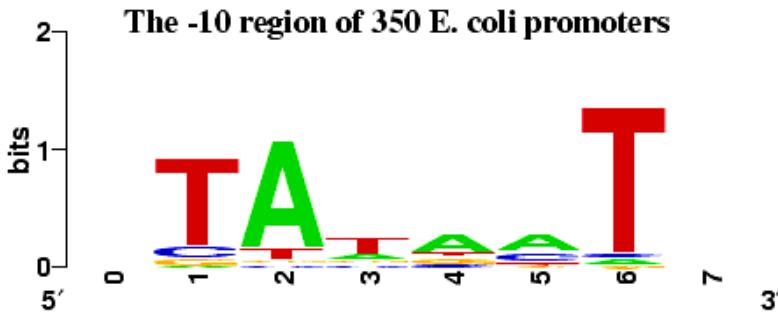


Figure 9.5: The pattern for *E. coli* RNA polymerase promoter at position -10.

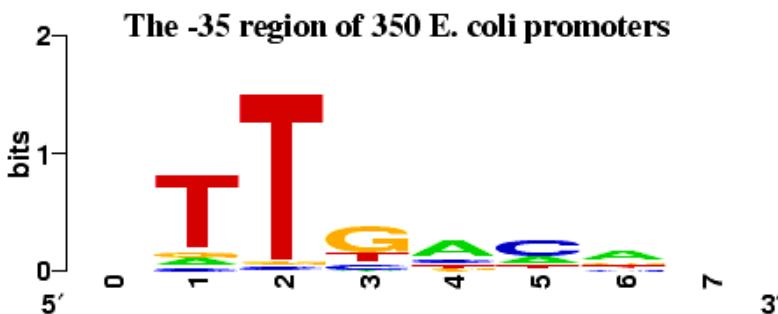


Figure 9.6: The pattern for *E. coli* RNA polymerase promoter at position -35.

If a window over the current position is used then the input weights of the neural network are equivalently to a scoring matrix. See Fig. 9.7 for representing the neural network ingoing weights through a scoring matrix.

Hidden Markov Models.

Either the alignment is coded into a hidden Markov model or they are trained on short promoter sequences by the EM algorithm (see Bioinformatics II).

9.2.4.2 Eukaryotes

For eukaryotes the RNA polymerase II (RNAPII) binding patterns are indicators for promoter regions upstream of genes.

Eukaryotic promoter regions have few short patterns which are conserved but many longer patterns which contain the short patterns have high variation. The short patterns are binding sites for transcription factors of which many exist in eukaryotes: TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH. Their position is given with respect to the transcription start site.

A very indicative pattern is the TATA-box with the consensus sequence “TATA[A, T]{C}[G, A]”, where “[]” denote alternatives and “{ }” exceptions. The other well known box is the GC-box

There are many other patterns which are mostly represented by profiles which are used to search for other promoter sites. Profiles are suitable for promoter search because the pattern are

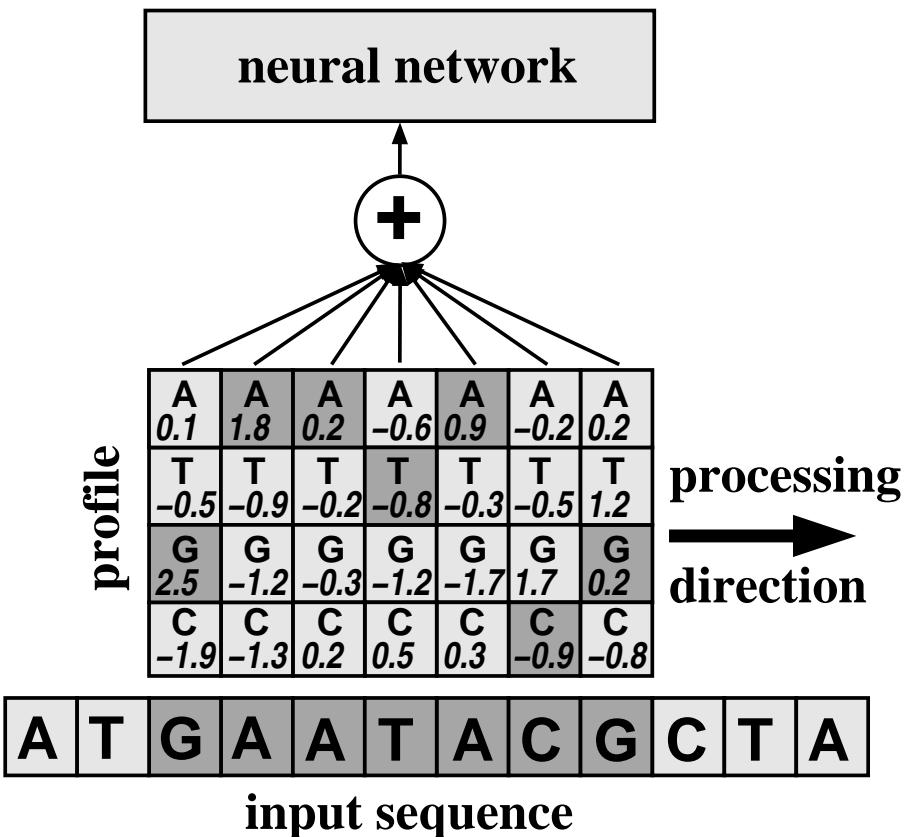


Figure 9.7: The input weights to a neural network which codes the nucleotides locally can be viewed as scoring matrix.

fuzzy so that only scoring can detect promoter sites. The function for some pattern is still unknown.

In a classification task cell cycle genes must be distinguished from other genes. Pentamers in the promoter region have been counted in both classes. The pattern ACCGGT was over-represented for the late G_1 phase cell cycle genes and the pattern CCCTT for the early G_1 phase.

Transcription factors can remodel the local nucleosome structure by acetylation and deacetylation of histones and making DNA regions accessible or avoiding access to them. Transcription factors are able to phosphorylate the RNAPII and so control transcription. They can activate or repress transcription.

Genes have different transcription factors in order to switch them on or off at a certain condition like developmental stage, external signal, heat shocks, nutrition deficiencies, or virus attack.

In order to identify co-regulated genes, the microarray technique may be used to find correlated expression values. Then upstream of the co-regulated genes one can search for common binding sites.

Prediction methods for promoter regions include

- **Neural Networks:** NNPP and PROMOTER2.0
- **Profiles:** weight matrices to identify promoter sites, e.g. PromoterScan; the weight matrices are extended to different organisms, e.g. TFSearch and TESS, and new matrices can be generated, e.g. MatInspector and ConsInspector.
- **Linear discriminant functions (LDA):** classifying promoter sequences and non-promoter sequences using as features TATA-box score, triple base statistics, hexamer frequencies, etc., e.g. TSSD and TSSW.
- **Quadratic discriminant analysis:** similar to LDA but with variable sequence length and with different and overlapping windows: CorePromoter.
- **Multiple pattern:** in the binding sites pattern are clustered which gives hints to the gene regulation, e.g. FastIM.

The eukaryotic promoter database (EPD) can be found under <ftp://ftp.epd.unil.ch/pub/databases/epd/views/>.

Web sites for gene and splice site recognition are Prediction methods for promoter regions include

- <http://linkage.rockefeller.edu/wli/gene/programs.html>
- <http://hto-13.usc.edu/software/procrustes/index.html>
- <http://cmgm.stanford.edu/classes/genefind/>
- <http://www1.imim.es/courses/SeqAnalysis/GeneIdentification/Evaluation.html>

true	predicted		total
	+1	-1	
+1	TP	FN	TP + FN
-1	FP	TN	FP + TN
total	TP + FP	FN + TN	N

Table 9.2: Confusion matrix. TP: true positive - positive correctly predicted; FN: false negative - positive incorrectly predicted; FP: false positive - negative incorrectly predicted; TN: true negative - negative correctly predicted.

9.2.5 EST Clusters

Another method to identify genes or to confirm predicted genes is to generate cDNA from mRNA and sequence these cDNA fragments. Here the fragments which identify individual sequences are called expressed sequence tags (ESTs). The ESTs which build a cluster of overlapping sequences are assumed to build a gene.

The new version of GRAIL namely GrailEXP provides EST data searches in order to confirm genes which were predicted.

9.2.6 Performance of Gene Prediction Methods

In Subsection 4.4.3 we defined

- TP: true positive - positive correctly predicted;
- FN: false negative - positive incorrectly predicted;
- FP: false positive - negative incorrectly predicted;
- TN: true negative - negative correctly predicted.

And the confusion matrix is given in Tab. 9.2.

In Subsection 4.4.3 we defined

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{N} \quad (9.2)$$

$$\text{specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \quad (9.3)$$

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (9.4)$$

$$\text{balanced error} = 0.5 (\text{specificity} + \text{sensitivity}) \quad (9.5)$$

$$\text{Matthews corr.} = \frac{\text{TP TN} - \text{FP FN}}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{FN} + \text{TN})(\text{FP} + \text{TN})}} \quad (9.6)$$

$$\text{weight of evidence} = \log \frac{\text{TP TN}}{\text{FP FN}} \quad (9.7)$$

Method	Sensitivity	Specificity	Matthews
GenParser	0.69-0.75	0.68-0.78	0.66-0.69
GeneID	0.65-0.67	0.74-0.78	0.66-0.67
Grail	0.48-0.65	0.86-0.87	0.61-0.72

Table 9.3: Test of finding the nucleotide ends of exons of gene prediction methods without data base search on three sets of human genes according to Snyder and Stormo 1993.

Method	Sensitivity	Specificity	Matthews
Grail	0.79	0.92	0.83
FGENEH	0.93	0.93	0.85
MZEF	0.95	0.95	0.89

Table 9.4: Test of Tab. 9.3 with other methods according to Zhang 1997.

Methods with data bases like GeneID+, GeneParser3, GrailEXP have considerably higher prediction performance.

In April 2000 the journal *Genome Research* gene prediction methods were compared and it turned out that more than 95% of the nucleotides in exons were found. However only a small fraction of the predicted gene models were correct.

9.3 Alternative Splicing and Nucleosomes

9.3.1 Nucleosomes

Eukaryotic DNA is wrapped around histone-protein complexes which are called “nucleosomes” giving a the chromatin.

Nucleosomes regulate gene expression because promoter sites may be not accessible at positions of nucleosomes.

Segal et al., “A genomic code for nucleosome positioning”, Nature, page 772-778, August 2006 build a Markov model for nucleosome positions. Because the nucleosome wrapping is done by a 147 bp sequences the model consists of 147 bp. Their model was

$$p(\mathbf{s}) = p_1(s_1) \prod_{i=2}^{147} p_i(s_i | s_{i-1}), \quad (9.8)$$

where the probability of finding a certain nucleotide in position i depends only on the nucleotide in position $(i-1)$. This model found a 10 base pair frequency of AA, TT, TA with alternates with GC (see Fig. 9.8). Similar models were found by hidden Markov models e.g. Baldi et al., 1996. 10

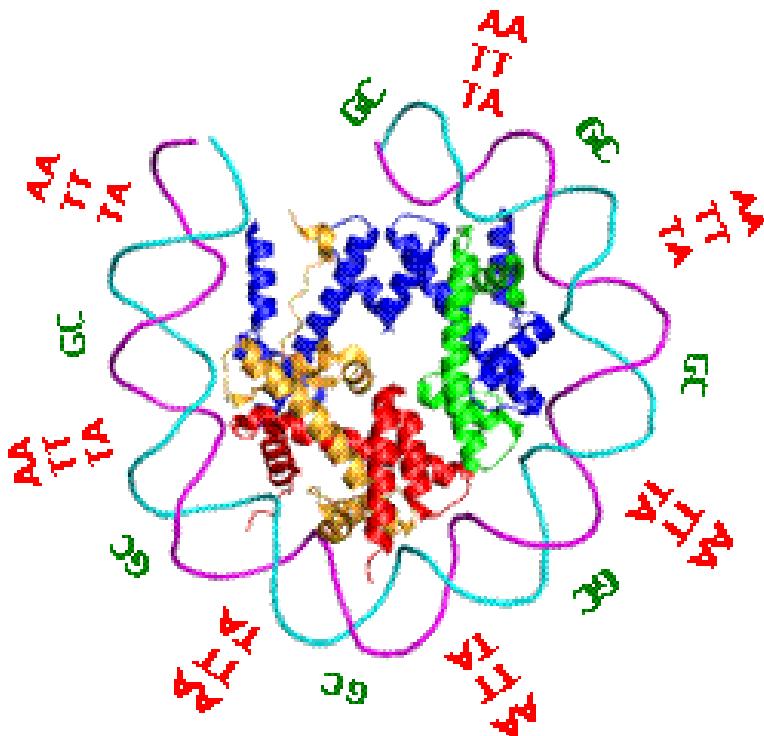


Figure 9.8: The nucleosome model found by Segal et al.

bp are the expected number of bp for a turn around the nucleosome, therefore HMM approaches are based on this frequency.

Using this model Segal et al. found that the highest density of nucleosomes was predicted in centromeres. The lowest density of nucleosomes was predicted where ribosomal RNA and transfer RNA is coded. However high nucleosome density was predicted at regions where ribosomal proteins are coded. Low nucleosome density was found at functional binding sites, i.e. transcription factor binding sites. The same holds true for transcription start sites.

9.4 Comparative Genomics

First, the proteins of two or more genomes that is the proteomes can be compared. Secondly, also gene locations, gene duplications, sequence repeats (location and length), single mutations (e.g. promoter), between genomes can be compared. See Fig. 9.9 and Fig. 9.10 for genome comparisons. Also a genome can be compared with itself or its chromosomes to one another – see Fig. 9.11 for the genome of *Arabidopsis*.

Another field is to determine the variability of the genome by analyzing SNPs, micro-satellites, gene expression etc. within one genome.

In genome comparisons we distinguish between:

- **Homolog genes:** Genes which are so similar that they share the same function and have a

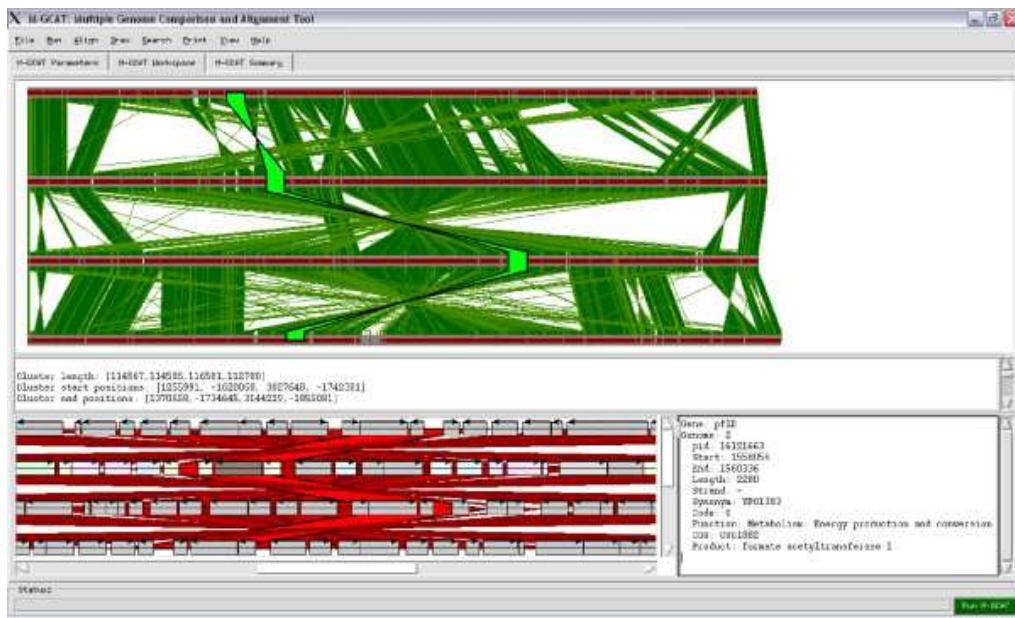


Figure 9.9: Genome comparison between species.

common ancestor are called *homolog*.

- **Ortholog genes:** Genes of different species which are similar so that they evolved from a common ancestor but which underwent speciation are called *ortholog*. *Speciation* is the process where a new species appears which has no longer genetic exchange with the parent or the sister species.
- **Paralog genes:** Genes of one species which are similar so that they evolved from a common ancestor – probably by gene duplication – and which often attained a new function are called *paralog*.

Gene duplication. Note, that most gene duplication events lead to pseudo-genes. In rare events the duplicated gene and the original gene both keep their function and are used to fine tune the regulation of these function by two genes. And in other rare events either the original gene or the duplicate develops a new function because mutations in one of the genes is not penalized as there is the backup gene.

Within one genome the proteins can be compared to one another in order to identify gene and protein families. These comparisons can be the basis of clustering the proteins and so figure out the families.

Comparisons between genomes can be made on the basis of protein sequences. If these sequences are not available because the genome has not been sequenced yet then the comparison can be made based on EST data. Homolog sequences can then be clustered by single-linkage analysis or other methods.

Another criterion is *synteny* which is the local gene order which is also conserved between close related species. To construct phylogenetic trees also the synteny and rearrangements of genes on the chromosomes can be used to estimate the evolutionary distance between species. The

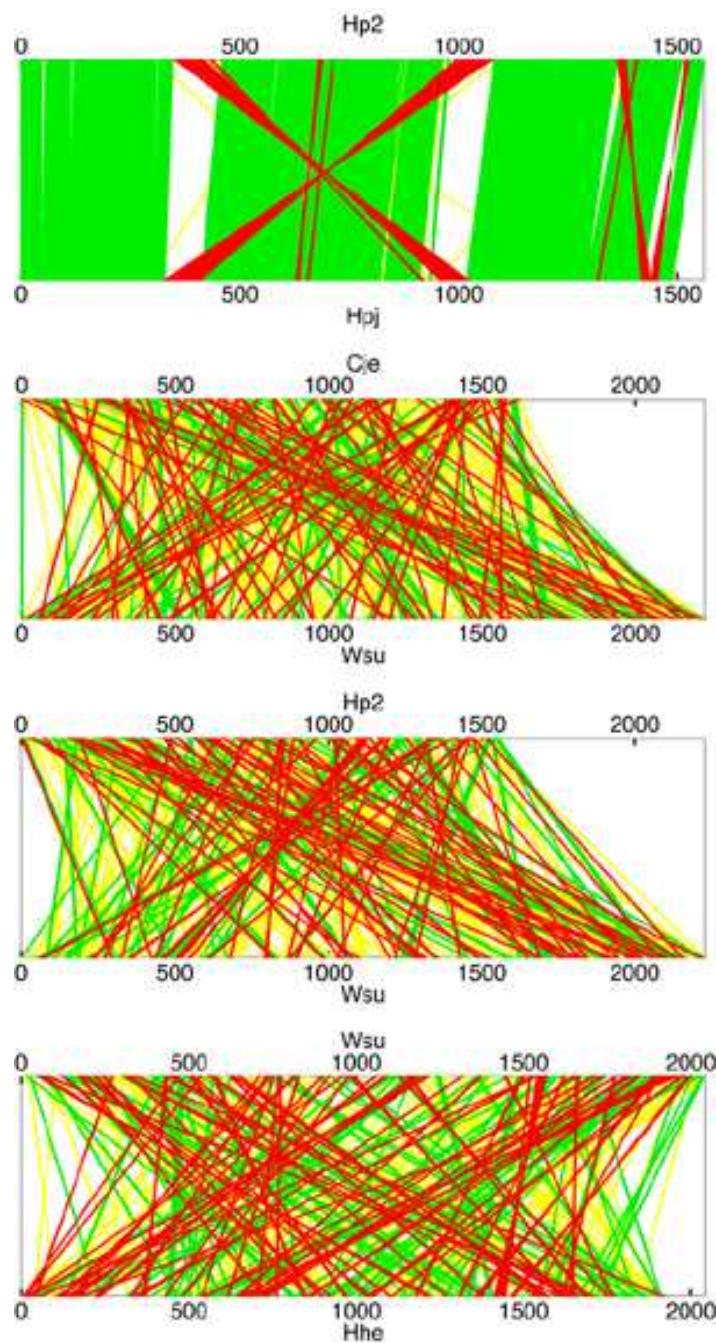


Figure 9.10: Genome comparison of campylobacterales. Within one species the genomic order is conserved (first panel) but not between species (second to fourth panel). Copyright © Max-Planck-Institut für Entwicklungsbiologie, Huson, Schuster.

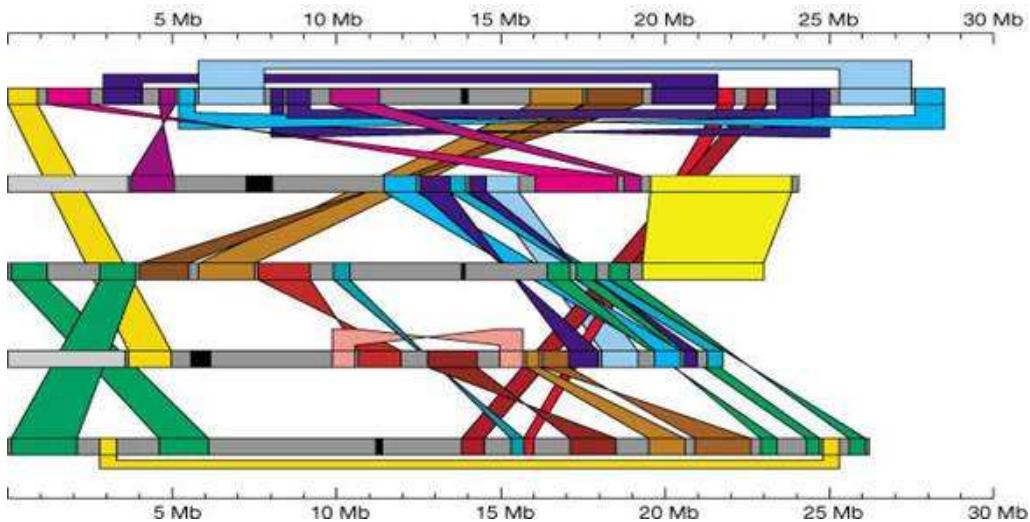


Figure 9.11: Segmental duplication between the chromosomes of *Arabidopsis*. Copyright © Arabidopsis Genome Initiative.

distance is computed by elementary rearrangement steps to transfer the genome of one species into the genome of another species.

In prokaryotes Horizontal Gene Transfer has to be kept in mind. Genomic material of one species is directly included into the genome of another species. For example organelle genomes like mitochondria or chloroplasts are assumed to be bacterial genome which has been incorporated into the eukaryotic genome. But horizontal gene transfer is more prominent in bacteria and is often detected by an deviation of base frequencies in a region of a genome. For example *E. coli* is thought to have acquired 12.8% of its genome form horizontal gene transfer whereas in other organism no horizontal gene transfer has been detected.

Fig. 9.12, Fig. 9.13, and Fig. 9.14 compare the mouse chromosomes and the according gene clusters with the human chromosomes. In contrast to the mouse genome the chimpanzee genome mapping is shown in Fig. 9.15. It can be seen chimpanzee and human gene cluster locations agree much more than the locations of human and mouse gene clusters. The experimental methods to identify corresponding location in the chromosomes of different species are fluorescent in situ hybridization (FISH) and Giemsa staining. Fig. 9.16 shows how the worm gene clusters are spread over the fruit fly chromosomes.

Fig. 9.17 shows a genome comparison of the mitochondrial genome (mtDNA) of the *Zea* family.

There are other way of representing the similarity of genomes like plots which are similar to dot plots but for gene matching instead of letter matching. Other comparison present the result as maps like in Fig. 9.9.

Genes with are interdependent or related to one another are clustered on the genome. These clusters are inherited as blocks. For the reason of this gene clustering it was speculated that alleles are interdependent and are inherited together. Genes participation in one pathway are often found within one of these clusters.

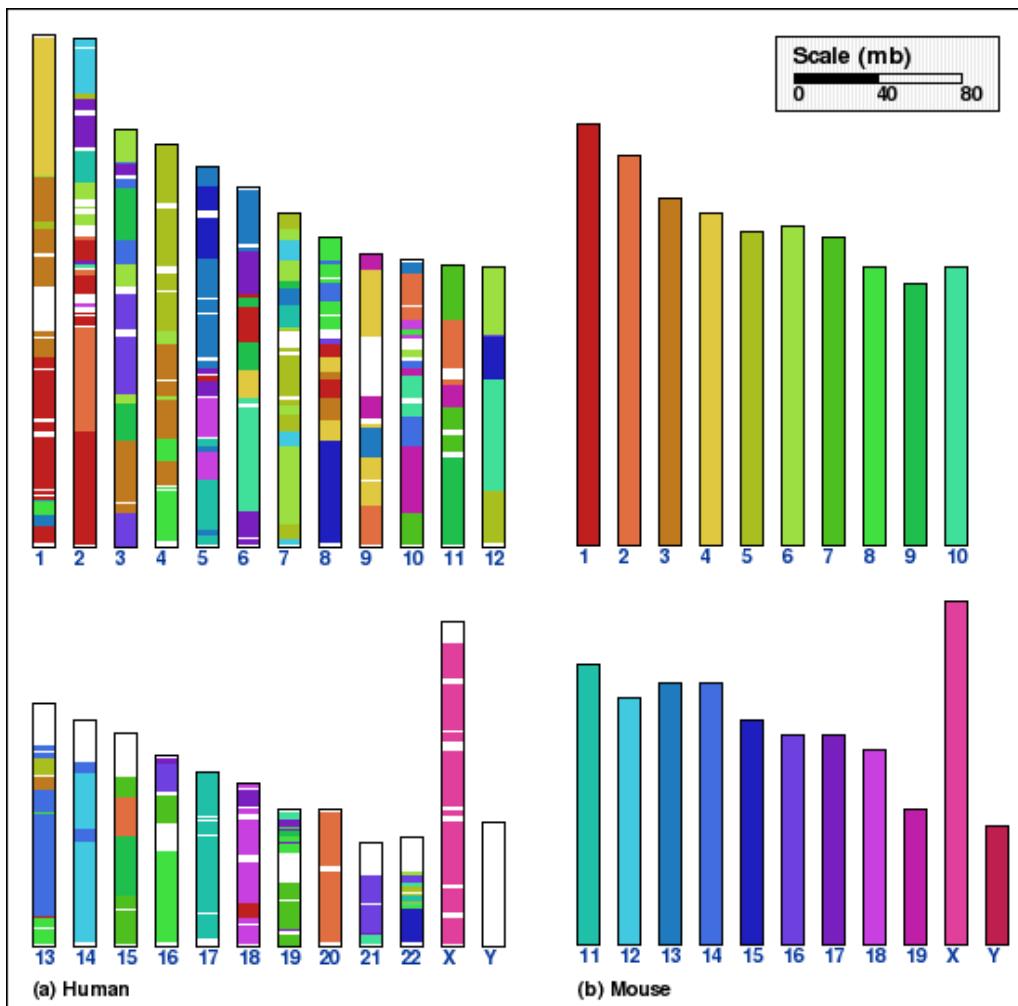


Figure 9.12: The mouse chromosomes mapped to the human chromosomes where the color gives the mouse chromosome number.

Mouse and Human Genetic Similarities

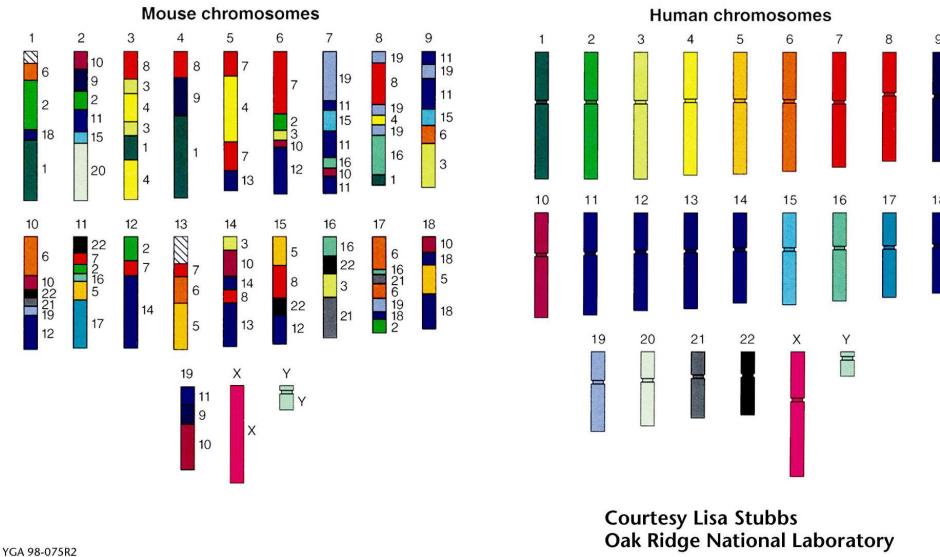


Figure 9.13: The human chromosomes mapped to the mouse chromosomes where the color gives the human chromosome number.

9.5 Genomic Individuality

9.5.1 Sequence Repeats

Eukaryotic genomes contain tandem repeats for example at movable genetic content where the moves multiply (even double) the repeats. These repeats also occur through mutations which multiply the regions containing repeats.

The repeats are a junction of the same sequence, therefore the base distribution is very characteristic for the repeats. This base pair distribution gives the repeats a characteristic mass per volume, the buoyant density. Measurement methods which can separate DNA fragments of different densities are able to identify these repeats as *satellite DNA* because of the characteristic mass per volume.

Satellites.

There are three types of satellite DNA *satellites*, *mini-satellites*, *micro-satellites*:

- **satellites:** repeats of one thousand to several thousand base pairs in tandem regions of up to 100 million bases long.
- **mini-satellites:** repeats of 15 base pairs in regions from 100kb up to some 1000kb. These regions vary in size so that individuals can be identified by these repeats. These “variable number of tandem repeats” is used in forensic science.

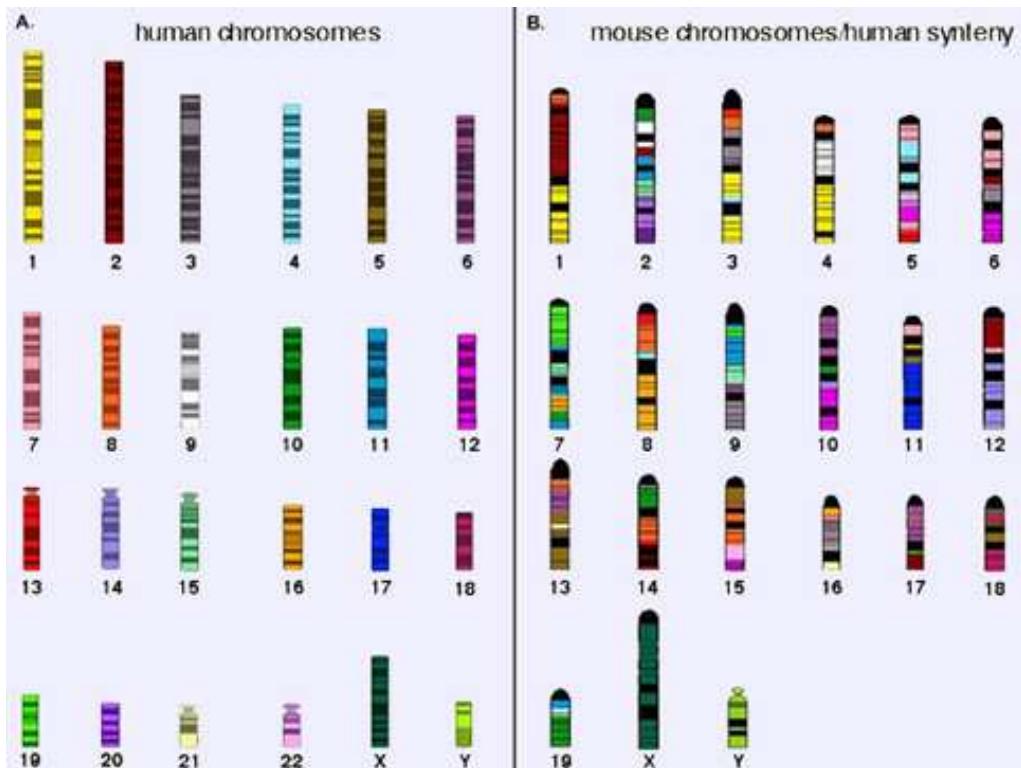


Figure 9.14: The human chromosomes mapped to the mouse chromosomes where the color gives the human chromosome number.

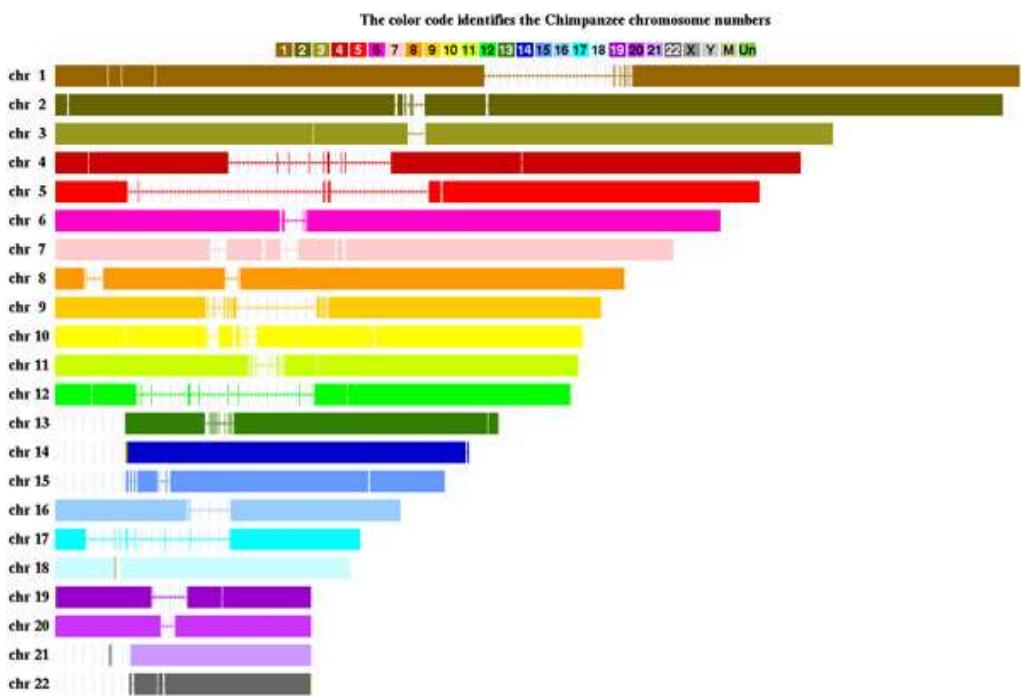


Figure 9.15: The chimpanzee chromosomes mapped to the human chromosomes where the color gives the chimpanzee chromosome number.

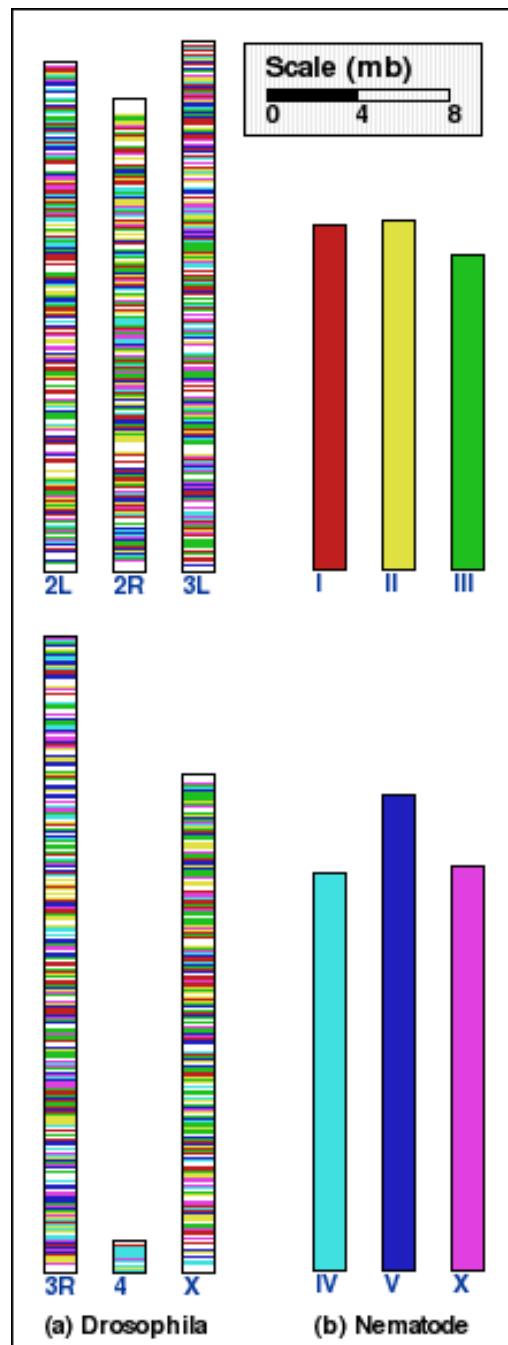


Figure 9.16: Worm chromosomes mapped to the fruit fly (drosophila) chromosomes where the color gives the worm (nematode) chromosome number.

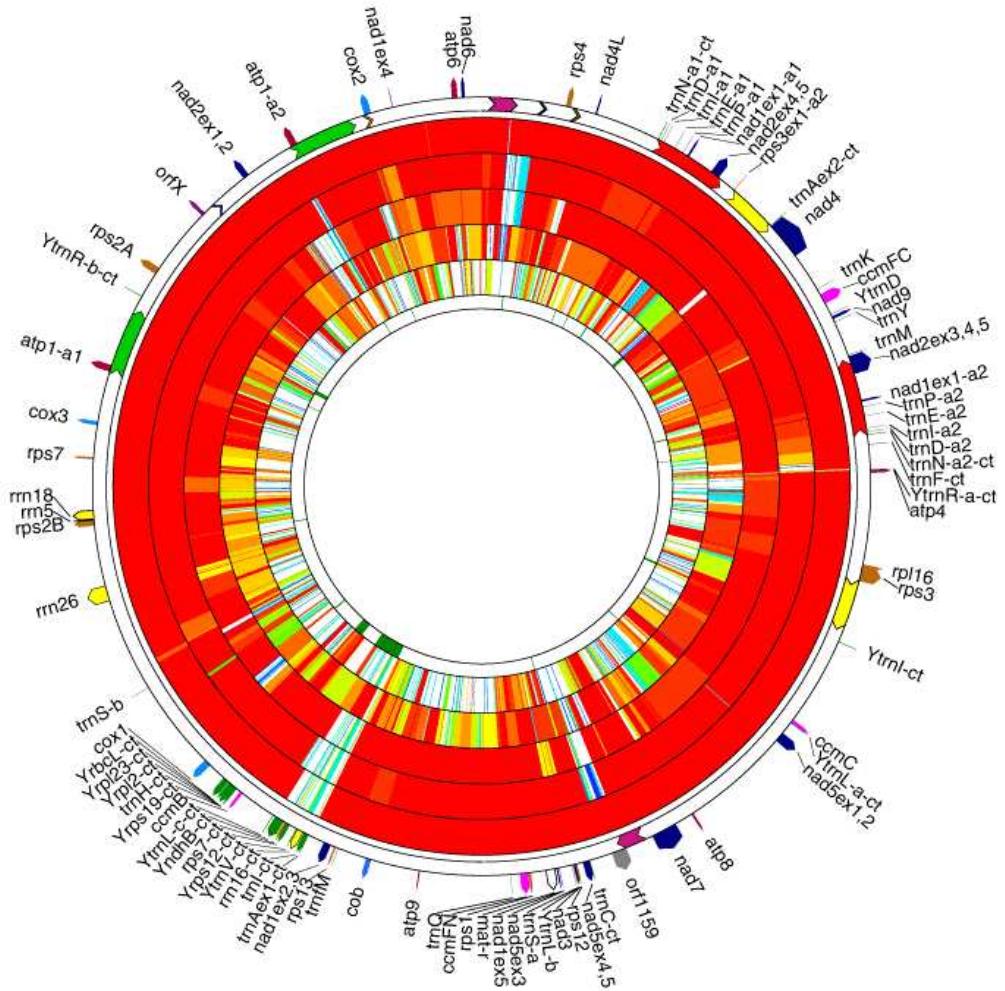


Figure 9.17: Mitochondrial genomes of the *Zea* family where *Zea mays* (maize) is the outermost circle whereas the other circles represent the mtDNA of other members of the *Zea* family. The innermost circle is the mtDNA of *Sorghum bicolor*.

Organism	% transposable
<i>H. sapiens</i> - human	35
<i>Z. mays</i> - maize	50
<i>D. melanogaster</i> - fruit fly	15
<i>A. thaliana</i> - plant	2
<i>C. elegans</i> - nematode	1.8
<i>S. cerevisiae</i> - budding yeast	3.1

Table 9.5: Percentage of the transposable DNA in the genome for different species.

- **micro-satellites:** short repeats of 2-6 base pairs in array of length 10 to 100 base pairs. The length are inherited from parents and are therefore used for evolutionary analysis or even as genetic markers. Micro-satellites are also called “simple sequence repeats” or “short tandem repeats” and are located in euchromatin, centromeres, and telomeres. A very prominent repeat is the TTAGGG near the telomeres with hundreds of copies.

Transposable Elements.

Transposable elements are DNA regions which can move from one location at the chromosome to another location, which is done faster than chromosomes replicate. Most of the repetitive sequences are in transposable elements. Tab. 9.5

The transposable elements either encode a reverse transcriptase and RNA-based transposition or DNA-based dynamics of transposition. The human genome contains 200,000 copies of the later class but the former is more dominant. There exist hybrids of these two classes the “miniature, inverted repeat transposable elements” (MITES) of 400 bp length.

The former class, the RNA-based, reverse transcriptase related transposable elements can be divided into *long terminal repeat retrotransposons*, *long terminal repeat retroposons*, and *long terminal repeat retrovirus-like*. The retroposons may be short interspersed nuclear elements (SINES) of 50 to 500 bp in length or long interspersed nuclear elements (LINEs) of 4 to 7 kbp in length.

10% of the human genome consists of a special family of SINES “Alu” with 1.2 million copies and 14.6% of the genome consists of a specific LINE namely LINE1 with 593,000 copies. Alu-sequences are found in non-coding regions like introns, promoter regions, or untranslated regions.

9.5.2 SNPs

Single Nucleotide Polymorphism (SNPs, pronounce “snips”) is a DNA sequence variation where a single nucleotide differs between members of a species. For example the sequences gt^agCccc and gt^agTccc differ in a single nucleotide and we say there are two *alleles* C and T. To be classified as SNP each allele must have at least 1% (sometimes 0.5%) frequency in the population.

SNPs may be in the exons or in the introns of genes but also in promoter sites or between genes. SNPs in the exons may be neutral to the translation but can change an amino acid in the

protein. SNPs in non-coding regions may influence splicing, transcription factor affinity, or other regulatory effects.

SNPs can affect humans affinity to diseases, responses to pathogens, chemicals, drugs, etc. Therefore SNPs are of interest for pharmacy because of tailoring drugs to individuals. E.g. drugs may be of optimal effect if a certain allele is present because of the individual metabolism. Here the future may be in the individualized medicine.

Formally SNPs are detected by specific enzymes which cut if a allele is present. Now SNPs can be detected by SNP arrays based on the microarray technique of Affymetrix or other companies.

In the following two specific SNPs for lactase and COMT are given as an example. The lactase SNP is responsible for how well milk is processed by the body. The COMT SNPs are associated with schizophrenia. Because whole regions are inherited, SNPs are also simultaneously inherited. Therefore SNPs are correlated and a SNP which is indicative for affinity to certain diseases may not be the cause of the disease but some other mutation which is inherited together with this SNP.

LACTASE SNP:

is a C/T SNP in the intron region of MCM6 gene. The -13910*T allele introduces a BsmFI restriction site.

Name	intron 13 C/T (-13910) SNP
UID	SI001784U / rs4988235
Locus Name	MCM6 minichromosome maintenance deficient 6
Locus Symbol	MCM6

Alleles:

Allele	Description
C	C 5' - atacagataagataatgttag C ccctggcctcaaaggaaactc - 3'
T	T 5' - atacagataagataatgttag T ccctggcctcaaaggaaactc - 3'

Reference: Coelho M, Luiselli D, Bertorelle G, Lopes AI, Seixas S, Destro-Bisol G, Rocha J., "Microsatellite variation and evolution of human lactase persistence". Hum. Genet. 117:329-339, 2005.

COMT SNPs:

The catechol-O-methyltransferase (COMT) gene, is a candidate gene for schizophrenia. COMT is one of the enzymes that degrade dopamine, a neurotransmitter. The polymorphism associated with the COMT gene show significant dependencies with schizophrenia:

- Val/Met polymorphism (SNP rs165688): Val/Val (G/G) genotype is associated with schizophrenia ($P=0.0074$).
- SNP rs165599: highly significant in women (G allele predisposing, $P=9.1 \times 10^{-6}$) and genotype level G/G ($P=6.8 \times 10^{-6}$)
- SNP rs737865: significantly associated with schizophrenia in men ($P=0.0011$) and women ($P=0.012$).

Table 1 Genotype and allele frequencies of the single nucleotide polymorphism g.-888G>C of the reelin gene in schizophrenic patients and controls

Polymorphism	Schizophrenia n = 279	Control n = 255	χ^2	df	P
g.-888G>C					
Genotype			2.99	2	0.22
GG	243 (87.1%)	209 (81.9%)			
GC	34 (12.2%)	42 (16.5%)			
CC	2 (0.7%)	4 (1.6%)			
Allele			3.16	1	0.08
G	520 (93.2%)	460 (90.2%)			
C	38 (6.8%)	50 (9.8%)			

Table 9.6: SNP associated with schizophrenia from “Identification of a single nucleotide polymorphism at the 5' promoter region of human reelin gene and association study with schizophrenia”, Molecular Psychiatry, 2002, 7, 447-448.

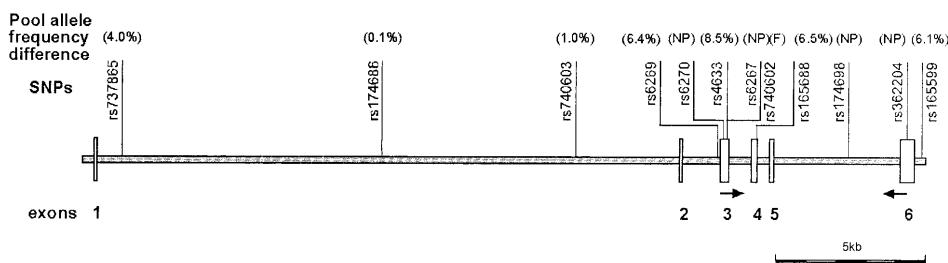


Figure 1 Location of SNPs studied in the COMT locus and allele-frequency differences observed between patients with schizophrenia and control individuals by means of analysis of the DNA pools. Four SNPs were not polymorphic (NP), and one amplification failed (F). Three of the SNPs studied (rs6270, rs6267, and rs165688) cause a nonsynonymous change.

Figure 9.18: The COMT SNPs associated with schizophrenia from Shifman, et al. “Highly Significant Association between COMT Haplotype and Schizophrenia”, Am. J. Hum. Genet. 71:1296-1302, 2002.

See Fig. 9.18 for the COMT SNPs in the gene (introns and exons).

In the chromosomal reproduction whole blocks are inherited, therefore SNPs are located and correlated within these blocks. These blocks are called *haplotype blocks*. Two haplotype sets per chromosome exist in each human. This means that the alleles come in pairs (one allele at each chromosome), e.g. CC, CT, or TT.

A SNP data base can be found under <http://www.ncbi.nlm.nih.gov/SNP/>. The SNP consortium's home page can be found under <http://snp.cshl.org>. The International HapMap Project (haplotype map project) can be found under <http://www.hapmap.org/>.

Chapter 10

DNA Sequence Statistics

This chapter focuses on statistical properties of DNA sequences.

10.1 Local Characteristics

The nucleotide frequency is given in Table 10.1.

This frequency differs in different regions, e.g. if we look at the human fetal globin gene in Table 10.2.

Next we consider pairs of nucleotides and compute the ratio

$$\frac{p_{ij}}{p_i p_j}, \quad (10.1)$$

where p_{ij} is the probability of the pair i and j and p_i is the probability of nucleotide i . Table 10.3 gives the ratio, where the low CG is obvious.

10.2 Long-Range Characteristics

Regions of uniform A,C,G, and T distribution are called *isochores*.

If DNA is considered as a double-strand, therefore often the C-G content matters.

10.2.1 Matching Probability of Subsequences

Repeats are identical subsequences within a sequence. Repeats may have evolutionary origin like duplication events. Such repeats may have biological relevance or may be indicative for certain evolutionary relationships.

A	C	G	T
0.31	0.31	0.25	0.13

Table 10.1: Nucleotide frequencies.

	length	A	C	G	T
5' flanking (2)	1000	0.33	0.23	0.22	0.22
3' flanking (2)	1000	0.29	0.15	0.26	0.30
Introns (4)	1996	0.27	0.17	0.27	0.29
Exons (6)	882	0.24	0.25	0.28	0.22
Inter-genic (1)	2487	0.32	0.19	0.18	0.31

Table 10.2: Nucleotide frequencies for human fetal globin gene.

first	second			
	A	C	G	T
A	1.15	0.84	1.16	0.85
C	1.15	1.18	0.42	1.26
G	1.04	0.99	1.15	0.82
T	0.65	1.00	1.29	1.07

Table 10.3: Nucleotide ratio $\frac{p_{ij}}{p_i p_j}$ of observed pairs p_{ij} and random pairs $p_i p_j$.

In bioinformatics 1 dot matrices have been introduced. These are matrices where at the most left column a sequences and in the first line a sequence is placed. The matrix contains a dot if the nucleotides are identical. An identical subsequence corresponds to a line of dots on a diagonal. We now use the same sequence in the first line and the left column. Let the length of the sequence be n , then the matrix has $l = \frac{n(n-1)}{2} \approx \frac{1}{2}n^2$ entries.

Now let us assume that we have sequences of length n and we want to know the probability of observing a subsequence of r matches.

Towards this end we write the dot matrix as a sequence of length l by writing down the main diagonals.

Now we want to know the probability of observing a subsequence of r dots in the sequence of length l . We denote this probability by $p_l(r)$ and its complement by $\bar{p}_l(r)$ (not observing a subsequence of r or more dots in a sequence of length l).

A match occurs with probability

$$p = \sum_{i=1}^4 p_i^2, \quad (10.2)$$

which is the probability that base i is in the first sequence multiplied by the probability that base i is also in the second sequence summed over all bases.

Let $q = (1 - p)$ be the probability that no match occurs.

Now let us assume that we have a sequence of length l and we want to compute the probability $\bar{p}_l(r)$ that we do not observe a subsequence of r matches in a sequence of length l . $p_l(r)$ is the probability of observing a subsequence of r matches in a sequence of length l .

The two sequences of length l are now divided into regions of matches and regions with mismatches. Let assume there are s matches, where the j -th match-sequence has length l_j . If we allow also $l_j = 0$ then we have exactly $(s - 1)$ mismatches, where one mismatch separates matches. We obtain

$$l = s - 1 + \sum_{j=1}^s l_j. \quad (10.3)$$

The probability these s matches is

$$q^{s-1} p^{\sum_{j=1}^s l_j}, \quad (10.4)$$

The probability $\bar{p}_l(r)$ requires that $l_j \leq r - 1$.

We now construct the probability generating function (see App. A):

$$\begin{aligned} \sum_{l=0}^{\infty} \bar{p}_l(r) z^l &= \sum_{s=1}^{\infty} \sum_{l_j; l_j < r} q^{s-1} p^{\sum_{j=1}^s l_j} = \\ &\sum_{s=1}^{\infty} (q z)^{s-1} \left(\sum_{l=0}^{r-1} (p z)^l \right)^s = \\ &\sum_{s=1}^{\infty} (q z)^{s-1} \left(\frac{1 - (p z)^r}{1 - p z} \right)^s = \\ &\left(\frac{1 - (p z)^r}{1 - p z} \right) \left(1 - \frac{q z (1 - (p z)^r)}{1 - p z} \right)^{-1} = \\ &(1 - (p z)^r) (1 - z + q z(p z)^r)^{-1} \end{aligned} \quad (10.5)$$

From first to second line all l_j of the same length are put into $(p z)^l$.

Let z_α a root of $(1 - z + q z(p z)^r) = 0$, that is $(1 - z_\alpha + q z_\alpha(p z_\alpha)^r) = 0$.

For a function $f(x) = (x - a)g(x)$ we have $f'(x) = g(x) + (x - a)g'(x)$ and therefore $f'(a) = g(a)$.

We set now

$$\begin{aligned} a_\alpha &= \lim_{z \rightarrow z_\alpha} \frac{(1 - (p z)^r)(z - z_\alpha)}{1 - z + q z(p z)^r} = \\ &\frac{1 - (p z_\alpha)^r}{-1 + q r (p z_\alpha)^r} = \frac{z_\alpha - \frac{z_\alpha - 1}{q}}{-z_\alpha + r(z_\alpha - 1)}, \end{aligned} \quad (10.6)$$

where we used the derivative of the denominator in the second line to factor out $(z - z_\alpha)$.

$$(1 - (p z)^r) (1 - z + q z(p z)^r)^{-1} = \sum_{\alpha=1}^r a_\alpha / (z - z_\alpha) = \sum_{\alpha=1}^r \sum_{l=0}^{\infty} \frac{-a_\alpha}{z_\alpha^{l+1}} z^l, \quad (10.7)$$

where we used the geometric series

$$\begin{aligned} \sum_{l=0}^{\infty} \left(\frac{z}{z_\alpha} \right)^l &= \frac{1}{1 - z/z_\alpha} = \frac{z_\alpha}{z_\alpha - z} \\ \Rightarrow \sum_{l=0}^{\infty} -\frac{z^l}{z_\alpha^{l+1}} &= \frac{1}{z - z_\alpha}. \end{aligned} \quad (10.8)$$

Through comparison of the coefficients in eq. (10.5) we obtain

$$\bar{p}_l(r) = \sum_{\alpha=1}^r \frac{-a_\alpha}{z_\alpha^{l+1}}. \quad (10.9)$$

If we assume that l is large enough then only the smallest absolute z_α , $z_* = \min_\alpha |z_\alpha|$, determines $\bar{p}_l(r)$.

$$\bar{p}_l(r) \approx \frac{z_* - \frac{z_* - 1}{q}}{-z_* + r(z_* - 1)} \frac{1}{z_*^{l+1}} \quad (10.10)$$

The root condition

$$1 - z_\alpha + q z_\alpha (p z_\alpha)^r = 0 \quad (10.11)$$

can be solved for z_α

$$z_\alpha = 1 + q z_\alpha (p z_\alpha)^r (1 + q z_\alpha (p z_\alpha)^r) \quad (10.12)$$

Inserting the last equation into itself gives

$$\begin{aligned} z_\alpha &= 1 + q (1 + q z_\alpha (p z_\alpha)^r) (p (1 + q z_\alpha (p z_\alpha)^r))^r = \\ &1 + q p^r + O(p^{2r}) \end{aligned} \quad (10.13)$$

Inserting

$$z_* = 1 + q p^r \quad (10.14)$$

into eq. (10.10) gives

$$\bar{p}_l(r) \approx \frac{1 + q p^r - p^r}{1 + q p^r - r p^r} \frac{1}{(1 + q p^r)^{l+1}} \approx \frac{1}{(1 + q p^r)^{l+1}} \quad (10.15)$$

We now set

$$\begin{aligned} \frac{1}{(1 + q p^r)^{l+1}} &= \exp(-(l+1) \ln(1 + q p^r)) \approx \\ &\exp(-(l+1) q p^r), \end{aligned} \quad (10.16)$$

where we used

$$\ln(1 + q p^r) \approx q p^r. \quad (10.17)$$

For large l we set $l+1 = l$ and obtain

$$\bar{p}_l(r) \approx \exp(-l q p^r). \quad (10.18)$$

$\bar{p}_l(r)$ is the probability that we do not observe a subsequence of r matches in a sequence of length l .

The probability $\bar{p}_l(r+1)$ differs from $\bar{p}_l(r)$ because subsequence of r matches in a sequence of length l are allowed but not longer subsequences.

Therefore the difference $\bar{p}_l(r+1) - \bar{p}_l(r)$ extracts the cases where there is exactly a match of a subsequence of length r but not longer match. That is the probability that the length of maximal matching subsequence is r .

More formally:

$$\begin{aligned} p_l(r = \max) &= p(r \text{ matches AND NO } (r+1) \text{ matches}) = \\ 1 - p(\text{NO } r \text{ matches OR } (r+1) \text{ matches}) &= \\ 1 - \bar{p}_l(r) - p_l(r+1) - p(\text{NO } r \text{ matches AND } (r+1) \text{ matches}) &= \\ \bar{p}_l(r+1) - \bar{p}_l(r), \end{aligned} \quad (10.19)$$

where we used $p(\text{NO } r \text{ matches AND } (r+1) \text{ matches}) = 0$ because $(r+1)$ matches imply r matches.

$$\begin{aligned} p_l(r = \max) &= \bar{p}_l(r+1) - \bar{p}_l(r) = \\ \exp(-l q p^{r+1}) - \exp(-l q p^r) &= \\ \exp(-l q p^r) (\exp(l q^2 p^r) - 1) &= \\ \exp(-l q p^r + \ln(\exp(l q^2 p^r) - 1)), \end{aligned} \quad (10.20)$$

where we used for the third “=” $p = (1 - q)$, therefore $-l q p^{r+1} = -l q p^r + l q^2 p^r$.

This function has a maximum with respect to r , where the maximum is peaked. This unimodal distribution is now approximated by a Gaussian. Towards this end we make a Taylor expansion of the exponent of $p(r = \max)$ around the maximum \bar{r} . Truncating this expansion at the second order term gives a quadratic term in r because the linear term vanishes at the maximum. The exponential function of a quadratic term is a Gaussian.

The exponent is

$$-l q p^r + \ln(\exp(l q^2 p^r) - 1) \quad (10.21)$$

and its derivative with respect to r is set to zero:

$$\begin{aligned} -l q (\ln p) p^r + l q^2 (\ln p) p^r \exp(l q^2 p^r) / (\exp(l q^2 p^r) - 1) &= \\ l q (\ln p) p^r (-1 + q / (1 - \exp(-l q^2 p^r))) &= 0 \end{aligned} \quad (10.22)$$

We have

$$-1 + q / (\exp(l q^2 p^r) - 1) = 0 \quad (10.23)$$

which can be solved for r

$$p^r = \frac{-\ln p}{l q^2} \quad (10.24)$$

which gives

$$\bar{r} = \ln\left(\frac{l q^2}{-\ln p}\right) / (-\ln p) \quad (10.25)$$

Now we compute the second order derivative for the Taylor expansion.

$$\begin{aligned} \frac{\partial}{\partial r} (l q (\ln p) p^r (-1 + q / (1 - \exp(-l q^2 p^r)))) &= \\ l q (\ln p) p^r (-\ln p + (q \ln p) / (1 - \exp(-l q^2 p^r)) - \\ l q^3 (\ln p) p^r \exp(-l q^2 p^r) / (1 - \exp(-l q^2 p^r))^2) &. \end{aligned} \quad (10.26)$$

Note, that $\frac{\partial}{\partial r} p^r = \ln p p^r$.

Now we can insert \bar{r} into the second order derivative, where we first use

$$q / (\exp(l q^2 p^r) - 1) = 1 \quad (10.27)$$

and obtain

$$\begin{aligned} \frac{\partial}{\partial r} (l q (\ln p) p^r (-1 + q / (1 - \exp(-l q^2 p^r)))) &= \\ l q (\ln p) p^r (-\ln p + \ln p - l q (\ln p) p^r \exp(-l q^2 p^r)) &= \\ -l^2 q^2 (\ln p)^2 p^{2r} \exp(-l q^2 p^r) &. \end{aligned} \quad (10.28)$$

Next we can insert

$$\begin{aligned} q / (\exp(l q^2 p^r) - 1) &= 1 \\ \Rightarrow \exp(l q^2 p^r) &= 1 - q = p \end{aligned} \quad (10.29)$$

and also

$$\begin{aligned} p^r &= \frac{-\ln p}{l q^2} \\ \Rightarrow p^{2r} &= \frac{(\ln p)^2}{l^2 q^4} \end{aligned} \quad (10.30)$$

and obtain

$$\begin{aligned} \frac{\partial}{\partial r} (l q (\ln p) p^r (-1 + q / (1 - \exp(-l q^2 p^r)))) &= \\ -\frac{p (\ln p)^4}{q^2}. \end{aligned} \quad (10.31)$$

Because the linear term of the Taylor expansion vanishes at the maximum and the constant term is independent of r , we obtain

$$p_l(r = \max) = Z \exp\left(-\frac{1}{2} \frac{p (\ln p)^4}{q^2} (r - \bar{r})^2\right), \quad (10.32)$$

where Z is a normalizing constant.

We have

$$\begin{aligned} \bar{r} &= E(r = \max) = \ln\left(\frac{l q^2}{-\ln p}\right) / (-\ln p) = \\ &= 2 \frac{\ln n}{-\ln p} + \frac{\ln(\frac{1}{2} q^2 / (-\ln p))}{-\ln p} \\ \sigma^2(r = \max) &= \frac{q^2}{p (\ln p)^4}, \end{aligned} \quad (10.33)$$

where we inserted $l \approx \frac{1}{2} n^2$.

Different Sequences.

The same line of arguments hold if the sequences of the dot matrix would have been different sequences with length l_1 and l_2 . Then $l = l_1 l_2$.

We can derive similar formulae with extreme value methods.

We again consider r matching dots. A subsequence of R matching dots was preceded by a failure and putting the failure and the R matching dots together we have

$$q p^R \quad (10.34)$$

as probability.

The probability of subsequences with matching length smaller R is

$$\sum_{r=0}^{R-1} q p^r = \sum_{r=0}^{R-1} (1-p) p^r = \sum_{r=0}^{R-1} p^r - \sum_{r=1}^R p^r = 1 - p^R. \quad (10.35)$$

Note that for a sequences of length l we have on average $q l$ mismatches. Each mismatch is the start of a matching subsequence. We have

$$r_{\max} = \max_{1 \leq i \leq l} R_i \quad (10.36)$$

if taking the maximum over the l q matching subsequences.

Now we obtain

$$\begin{aligned} p(r_{\max} \geq R) &= 1 - p(r_{\max} < R) = \\ 1 - \prod_{1 \leq i \leq l} p(R_i < R) &= 1 - (p(R_i < R))^{l q} = \\ 1 - (1 - p^R)^{l q} &\approx 1 - \exp(-l q p^R), \end{aligned} \quad (10.37)$$

where we used

$$\begin{aligned} (1 - p^R)^{l q} &= \exp\left(\ln\left((1 - p^R)^{l q}\right)\right) = \\ \exp(l q \ln(1 - p^R)) &\approx \exp(-l q p^R). \end{aligned} \quad (10.38)$$

Now we can compute the expectation

$$E(r_{\max}) = \int_0^\infty (1 - \exp(-l q p^R)) dR. \quad (10.39)$$

Setting $u := q l p^R$ gives $du = \ln p q l p^R dR = u \ln p dR$ and

$$E(r_{\max}) = \frac{1}{-\ln p} \int_0^{l q} \frac{1}{u} (1 - \exp(-u)) du. \quad (10.40)$$

Note, that

$$\begin{aligned} \text{Ein}(x) &= \int_0^x (1 - e^{-t}) \frac{dt}{t} = \sum_{k=1}^{\infty} \frac{(-1)^{k+1} x^k}{k k!} = \\ E_1(x) + \gamma + \ln(x) & \end{aligned} \quad (10.41)$$

(e.g. see http://en.wikipedia.org/wiki/Exponential_integral).

Here γ is the Euler gamma constant and $E_1(x)$ is the E-one function:

$$\begin{aligned} E_1(x) &= \int_1^\infty \frac{e^{-tx}}{t} dt = \int_x^\infty \frac{e^{-t}}{t} dt \\ \gamma &= 0.577215664901532860606512090082402431042 . \end{aligned} \quad (10.42)$$

$$\begin{aligned} E(r_{\max}) &= \frac{1}{-\ln p} (E_1(q l) + \gamma + \ln(q l)) \approx \\ &\frac{\ln(q l)}{-\ln p}, \end{aligned} \quad (10.43)$$

where the last approximation was made for large $(q l)$ for which $E_1(q l)$ is close to zero and $\ln(q l)$ is much larger than γ .

We obtain

$$E(r_{\max}) = \frac{\ln(q l_1 l_2)}{-\ln p} \quad (10.44)$$

It can also be derived that

$$\text{var}(r_{\max}) = \frac{\pi^2}{6 (\ln p)^2} . \quad (10.45)$$

10.2.2 Spectral Analysis

To identify repeated motifs or near repeated motifs Fourier transform is a well suited method - more exactly, the power spectrum to be shift invariant.

For each base s the *Fourier transform* is

$$a_k(s) = \frac{1}{N} \sum_{t=1}^N e^{2 \pi i k t / N} \delta_t(s) , \quad (10.46)$$

where $|k| < N/2$, i is the imaginary unit, and $\delta_t(s) = 1$ if at position t the nucleotide s appears and 0 otherwise.

The *power spectrum* is

$$\begin{aligned} F_k(s)^2 &= N |a_k(s)|^2 = \\ &\frac{1}{N} \sum_{t,r=1}^N e^{2 \pi i k (t-r) / N} \delta_t(s) \delta_r(s) = \\ &\frac{1}{N} \sum_{t=1}^N \delta_t(s) + \\ &\frac{1}{N} \sum_{n=1}^{N-1} \left(e^{2 \pi i k n / N} + e^{-2 \pi i k n / N} \right) \sum_{t=1}^{N-n} \delta_t(s) \delta_{t+n}(s) . \end{aligned} \quad (10.47)$$

The power should be averaged over all nucleotides:

$$F_k^2 = \sum_{s=1}^4 F_k(s)^2 . \quad (10.48)$$

Note, that the Fourier transform can be used to compute the autocorrelation

$$\begin{aligned} C_{ss}(n) &= \frac{1}{N-n} \sum_{t=1}^{N-n} \delta_t(s) \delta_{t+n}(s) - p_s^2 \\ p_s &= \frac{1}{N} \sum_{t=1}^N \delta_t(s) , \end{aligned} \quad (10.49)$$

where the correction with p_s is because autocorrelation is for zero centered data.

Using $\sum_{n=0}^{N-1} (N-n) \cos(2\pi k n / N) = 0$ we obtain for large N :

$$F_k^2 = 1 - 2 \sum_{s=1}^4 p_s^2 + 2 \sum_{n=1}^{\infty} \cos(2\pi k n / N) \sum_{n=1}^{N-1} C_{ss}(n) . \quad (10.50)$$

Random sequences will lead to

$$F_k^2 = 1 - 2 \sum_{s=1}^4 p_s^2 . \quad (10.51)$$

Further for

$$\sum_{s=1}^4 C_{ss}(n) \propto n^{-\alpha} \quad (10.52)$$

we obtain

$$F_k^2 \propto k^{\alpha-2} . \quad (10.53)$$

A Walsh transform may be more appropriate than a Fourier transform because Walsh transform is designed for discrete values and represent a discrete version of the sine function.

To extract patterns wavelet analysis may be more appropriate.

10.2.3 Entropy Analysis

Through entropic considerations the more frequent occurrence of a subsequence can be detected and regularities are found.

Consider words ω of length n , then we have the entropy

$$H_n = \sum_{\omega} -p_{\omega}^{(n)} \log_2 p_{\omega}^{(n)}. \quad (10.54)$$

For $n = 3$ the maximum entropy would be 6 but we obtain $H_3 \sim 5.9$.

The *excess entropy* is

$$h_n = H_{n+1} - H_n \quad (10.55)$$

and is a good regularity indicator for larger n .

If the word of length $(n + 1)$ is build from the word ω of length n by adding the nucleotide s then we have the word $\omega \cdot s$.

Therefore the excess entropy can be expressed as

$$\begin{aligned} h_n &= \sum_{\omega,s} -p_{\omega \cdot s}^{(n+1)} \log_2 \left(p_{\omega \cdot s}^{(n+1)} / p_{\omega}^{(n)} \right) = \\ &\text{KL} \left(p_{\omega \cdot s}^{(n+1)} \| p_{\omega}^{(n)} \right) = \left\langle \log_2 \left(p_{\omega \cdot s}^{(n+1)} / p_{\omega}^{(n)} \right) \right\rangle_{p_{\omega \cdot s}^{(n+1)}}, \end{aligned} \quad (10.56)$$

where KL denotes the Kullback-Leibler difference.

This is a measure of how similar the probabilities $p_{\omega \cdot s}^{(n+1)}$ and $p_{\omega}^{(n)}$ are. That is a measure of how well s can be predicted from ω . If s can be perfectly be predicted then both probabilities are identical.

For increasing n it is interesting when h_n gets into a saturation. Then the maximal dependency m over sequence intervals is detected as if we use m -order Markov models.

How h_n changes with n is an indicator for repetitive structures and their length.

Appendix A

Probability Generating Function

The text in this chapter is from the corresponding WIKIPEDIA page (http://en.wikipedia.org/wiki/Probability-generating_function). All text is available under the terms of the GNU Free Documentation License.

A.1 Definition

If N is a discrete random variable taking values on some subset of the non-negative integers, $\{0, 1, \dots\}$, then the probability-generating function of N is defined as:

$$G(x) = E(x^N) = \sum_{n=0}^{\infty} f(n)x^n,$$

where f is the probability mass function of N . Note that the equivalent notation G_N is sometimes used to distinguish between the probability-generating functions of several random variables.

A.2 Properties

A.2.1 Power series

Probability-generating functions obey all the rules of power series with non-negative coefficients. In particular, $G(1-) = 1$, since the probabilities must sum to one, and where $G(1-) = \lim_{z \rightarrow 1} G(z)$ from below. So the radius of convergence of any probability-generating function must be at least 1, by Abel's theorem for power series with non-negative coefficients.

A.2.2 Probabilities and expectations

The following properties allow the derivation of various basic quantities related to X :

1. The probability mass function of X is recovered by taking derivatives of G

$$f(k) = \Pr(X = k) = \frac{G^{(k)}(0)}{k!}.$$

2. It follows from Property 1 that if we have two random variables X and Y , and $G_X = G_Y$, then $f_X = f_Y$. That is, if X and Y have identical probability-generating functions, then they are identically distributed.
3. The normalization of the probability density function can be expressed in terms of the generating function by

$$E(1) = G(1^-) = \sum_{i=0}^{\infty} f(i) = 1.$$

The expectation of X is given by

$$\mathbb{E}(X) = G'(1^-).$$

More generally, the k -th factorial moment, $E(X(X - 1)\dots(X - k + 1))$, of X is given by

$$\mathbb{E}\left(\frac{X!}{(X-k)!}\right) = G^{(k)}(1^-), \quad k \geq 0.$$

So the variance of X is given by

$$\text{Var}(X) = G''(1^-) + G'(1^-) - [G'(1^-)]^2.$$

4. $G_X(e^t) = M_X(t)$ where X is a random variable, $G(t)$ is the probability generating function and $M(t)$ is the moment generating function.

A.2.3 Functions of independent random variables

Probability-generating functions are particularly useful for dealing with functions of independent random variables. For example:

If X_1, X_2, \dots, X_n is a sequence of independent (and not necessarily identically distributed) random variables, and

$$S_n = \sum_{i=1}^n a_i X_i,$$

where the a_i are constants, then the probability-generating function is given by

$$E(z^{S_n}) = E(z^{\sum_{i=1}^n a_i X_i}) = G_{S_n}(z) = G_{X_1}(z^{a_1})G_{X_2}(z^{a_2})\dots G_{X_n}(z^{a_n}).$$

For example, if

$$S_n = \sum_{i=1}^n X_i,$$

then the probability-generating function, $G_{S_n}(z)$, is given by

$$G_{S_n}(z) = G_{X_1}(z)G_{X_2}(z)\dots G_{X_n}(z).$$

It also follows that the probability-generating function of the difference of two random variables $S = X_1 - X_2$ is

$$G_S(z) = G_{X_1}(z)G_{X_2}(1/z).$$

Suppose that N is also an independent, discrete random variable taking values on the non-negative integers, with probability-generating function G_N . If the X_1, X_2, \dots, X_N are independent and identically distributed with common probability-generating function G_X , then

$$G_{S_N}(z) = G_N(G_X(z)).$$

This can be seen as follows:

$$\begin{aligned} G_{S_N}(z) &= E(z^{S_N}) = E(z^{\sum_{i=1}^N X_i}) = E(E(z^{\sum_{i=1}^N X_i}|N)) = \\ &E((G_X(z))^N) = G_N(G_X(z)). \end{aligned}$$

This last fact is useful in the study of Galton-Watson processes.

Suppose again that N is also an independent, discrete random variable taking values on the non-negative integers, with probability-generating function G_N . If the X_1, X_2, \dots, X_N are independent, but not identically distributed random variables, where G_{X_i} denotes the probability generating function of X_i , then it holds

$$G_{S_N}(z) = \sum_{i \geq 1} f_i \prod_{k=1}^i G_{X_k}(z).$$

For identically distributed X_i this simplifies to the identity stated before. The general case is sometimes useful to obtain a decomposition of S_N by means of generating functions.

A.3 Examples

The probability-generating function of a constant random variable, i.e. one with $Pr(X = c) = 1$, is

$$G(z) = (z^c) .$$

The probability-generating function of a binomial random variable, the number of successes in n trials, with probability p of success in each trial, is

$$G(z) = [(1 - p) + pz]^n .$$

Note that this is the n -fold product of the probability-generating function of a Bernoulli random variable with parameter p .

The probability-generating function of a negative binomial random variable, the number of trials required to obtain the r -th success with probability of success in each trial p , is

$$G(z) = \left(\frac{pz}{1 - (1 - p)z} \right)^r .$$

Note that this is the r -fold product of the probability generating function of a geometric random variable.

The probability-generating function of a Poisson random variable with rate parameter λ is

$$G(z) = (e^{\lambda(z-1)}) .$$

A.4 Example calculation: use of bivariate generating functions

The following example illustrates a very common technique the manipulation of PGFs: the use of bivariate super generating functions to compute the OGF of the PGFs of a sequence of random variables.

Suppose you sample a system that can assume two states, X and Y , X with probability p and Y with probability $1 - p$, e.g. a coin being flipped, obtaining the sequence of samples

$$S_1, S_2, S_3, \dots S_n ,$$

where the system was sampled n times and has no memory.

Define the random variable M_n to be the number of changes from one sample to the next in a sequence of n samples, i.e. how often S_m was different from S_{m-1} . For example, the sequence

$$X X X Y X X$$

has two changes, as does

$$Y X X X X X Y Y Y Y .$$

We want to calculate the PGF of M_n , which we will do by using bivariate generating functions.

We introduce the bivariate GF $G(z, u)$ given by

$$G(z, u) = \sum_{n \geq 1} E [u^{M_n}] z^n ,$$

i.e. $G(z, u)$ is the ordinary generating function of the PGFs of the M_n . This step is completely general and indeed the core of the method.

Now let $x_{n,k}$ be the probability of having k changes in a sequence of n samples, where the last sample was an X . Similarly, let $y_{n,k}$ be the probability of having k changes in a sequence of n samples, where the last sample was a Y , and put

$$X(z, u) = \sum_{n \geq 1, k \geq 0} x_{n,k} u^k z^n \quad \text{and} \quad Y(z, u) = \sum_{n \geq 1, k \geq 0} y_{n,k} u^k z^n$$

so that

$$G(z, u) = X(z, u) + Y(z, u) .$$

Now we clearly have

$$x_{n,0} = p^n \quad \text{and} \quad y_{n,0} = (1-p)^n ,$$

because having zero changes means getting a sequence of all X s or Y s.

For $k \geq 1$ we find

$$x_{n,k} = p y_{n-1,k-1} + p x_{n-1,k} \quad \text{and} \quad y_{n,k} = (1-p) x_{n-1,k-1} + (1-p) y_{n-1,k} ,$$

because e.g. to have k changes in a sequence of length n that ends in X , we either append an X to a sequence having $k-1$ changes and ending in Y , or append an X to a sequence having k changes and ending in X .

Summing these equations over n and k and writing X for $X(z, u)$ and Y for $Y(z, u)$, we obtain

$$X - \frac{pz}{1-pz} = puzY + pz \left(X - \frac{pz}{1-pz} \right)$$

and

$$Y - \frac{(1-p)z}{1-(1-p)z} = (1-p)uzX + (1-p)z \left(Y - \frac{(1-p)z}{1-(1-p)z} \right) .$$

The solution of this system is

$$X = -\frac{(-pz + puz + z - uz - 1) pz}{-z + 1 + pz^2 - p^2 z^2 - u^2 z^2 p + p^2 u^2 z^2}$$

and

$$Y = -\frac{z(-puz - 1 + p + pz - p^2z + p^2uz)}{-z + 1 + pz^2 - p^2z^2 - u^2z^2p + p^2u^2z^2}.$$

We may now use the general identity

$$\sum_{n \geq 1} E[M_n(M_n - 1) \dots (M_n - r)] z^n = \left(\left(\frac{d}{du} \right)^{r+1} G(z, u) \right)_{u=1}$$

to calculate the factorial moments of M_n . E.g. the OGF of the expectations is given by

$$\sum_{n \geq 1} E[M_n] z^n = \left(\frac{d}{du} (X + Y) \right)_{u=1} = -2 \frac{(-1 + p) z^2 p}{(-1 + z)^2},$$

from which we find (extracting coefficients) that

$$E[M_n] = 2p(1-p)(n-1).$$

An extensive discussion of this problem, as well as solutions by other methods, may be found on <http://les-mathematiques.net>.

Appendix B

Contact Potential for Threading

Tab. B.1 and Tab. B.2 show two examples for contact potentials used for threading.

$$\alpha = \begin{cases} \frac{M(d-U)^2(2d-3L+U)}{(U-L)^3} & \text{if } L \leq d \leq U \\ K_{\text{if}} & d < L \\ 0_{\text{if}} & d > U \end{cases} \quad (\text{B.1})$$

Tables B.3–B.8 show contact potential used by the threading programm PROSPECT [Xu and Xu, 2000].

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0.50	2.31	0.59	0.52	1.61	0.46	0.68	1.16	0.45	1.13	0.99	0.42	0.29	0.30	0.60	0.35	0.27	0.86	1.74	1.22
C	0.79	8.20	2.53	2.51	2.02	2.54	2.06	2.07	2.48	2.03	1.99	2.42	2.27	2.35	2.28	2.41	2.29	2.04	2.00	1.97
D	0.28	0.45	0.29	0.11	2.01	0.40	0.80	1.61	0.45	1.59	1.41	0.29	0.48	0.37	0.73	0.33	0.46	1.33	2.06	1.51
E	0.29	0.40	0.24	0.25	1.95	0.38	0.78	1.54	0.44	1.51	1.34	0.27	0.42	0.32	0.70	0.29	0.41	1.25	2.00	1.45
F	1.28	2.16	0.51	0.61	3.67	2.04	1.40	0.54	1.94	0.52	0.63	1.86	1.64	1.73	1.55	1.85	1.63	0.79	0.43	0.67
G	0.26	0.45	0.32	0.23	0.65	0.19	0.91	1.61	0.31	1.58	1.42	0.32	0.46	0.37	0.78	0.25	0.46	1.31	2.14	1.60
H	0.57	1.48	1.06	0.89	1.38	0.41	1.93	1.10	0.75	1.06	0.86	0.65	0.53	0.58	0.38	0.67	0.52	0.86	1.39	0.85
I	1.13	1.56	0.39	0.53	2.78	0.43	0.97	2.83	1.51	0.09	0.34	1.46	1.25	1.33	1.20	1.44	1.22	0.31	0.81	0.62
K	0.31	0.52	0.86	0.91	0.70	0.26	0.53	0.56	0.28	1.49	1.32	0.26	0.39	0.30	0.53	0.23	0.38	1.22	2.00	1.46
L	1.07	1.65	0.38	0.48	2.86	0.43	0.99	2.64	0.56	2.59	0.27	1.43	1.21	1.30	1.18	1.41	1.19	0.29	0.80	0.59
M	0.98	1.59	0.41	0.52	2.73	0.57	1.12	2.12	0.57	2.10	2.06	1.24	1.02	1.12	0.98	1.24	1.02	0.28	0.80	0.46
N	0.36	0.61	0.58	0.48	0.78	0.39	0.82	0.51	0.54	0.56	0.67	0.73	0.29	0.14	0.52	0.12	0.25	1.16	1.91	1.36
P	0.41	0.90	0.37	0.44	1.07	0.33	0.79	0.76	0.36	0.72	0.84	0.56	0.53	0.19	0.46	0.28	0.23	0.96	1.72	1.18
Q	0.40	0.73	0.48	0.45	0.91	0.37	0.75	0.71	0.55	0.72	0.84	0.61	0.56	0.63	0.47	0.15	0.14	1.03	1.80	1.25
R	0.50	0.87	1.37	1.34	1.22	0.50	1.02	0.93	0.92	0.89	0.85	0.79	0.81	0.86	0.56	0.44	0.95	1.55	1.01	
S	0.36	0.66	0.48	0.45	0.86	0.31	0.80	0.59	0.41	0.60	0.62	0.54	0.43	0.56	0.71	0.49	0.24	1.14	1.93	1.38
T	0.51	0.84	0.47	0.53	1.12	0.37	0.89	0.94	0.54	0.88	0.88	0.68	0.57	0.67	0.84	0.57	0.73	0.93	1.71	1.16
V	0.99	1.46	0.37	0.45	2.48	0.40	0.91	2.29	0.48	2.17	1.71	0.49	0.68	0.65	0.78	0.56	0.85	2.08	1.00	0.64
W	1.29	2.38	0.92	1.06	3.44	0.87	1.86	2.58	1.02	2.68	2.72	1.13	1.66	1.31	1.75	1.12	1.28	2.24	3.48	0.58
Y	1.10	1.83	1.00	1.03	2.54	0.72	1.64	2.06	1.05	2.01	2.06	0.99	1.47	1.11	1.58	0.87	1.01	1.73	2.89	2.12

Additional table 1: In the bottom half and diagonal, in bold, we have the contact propensities P_{ij} , normalised to average unity, derived from a database of 1073 non-redundant protein structures with contact cut-off set at 4.5 Å. The top half shows the corresponding inter-residue distances $D_{ij} = \sqrt{\frac{1}{20} \sum_k (P_{ik} - P_{jk})^2}$, again normalised to unity. The distance matrix is zero along the diagonal.

Table B.1: The contact potential from [Williams and Doherty, 2004].

Pair	<i>L</i>	<i>U</i>	<i>M</i>												
GLY-GLY	3.57	6.42	1.00	VAL-ASP	3.26	7.22	7.51	MET-PHE	0.44	11.20	13.80	HIS-LYS	1.38	8.47	19.00
GLY-ALA	3.58	6.42	1.00	VAL-ASN	3.23	7.84	7.15	MET-PRO	3.57	8.98	5.44	HIS-ARG	1.68	9.53	21.00
GLY-VAL	3.06	6.79	2.96	VAL-GLU	2.19	7.82	11.00	MET-TYR	2.98	10.69	10.22	HIS-ASP	-0.11	10.62	16.73
GLY-LEU	2.49	7.53	3.74	VAL-GLN	2.27	8.45	9.00	MET-HIS	3.48	9.86	8.02	HIS-ASN	1.86	9.73	12.54
GLY-ILE	2.86	7.13	3.81	LEU-LEU	0.45	10.46	9.50	MET-TRP	1.25	11.08	19.00	HIS-GLU	1.45	10.47	16.00
GLY-CYS	3.21	6.85	1.93	LEU-ILE	1.64	9.87	9.00	MET-SER	2.45	8.02	5.72	HIS-GLN	3.69	9.20	10.60
GLY-MET	2.71	7.75	3.45	LEU-CYS	1.02	8.93	6.57	MET-THR	2.24	8.40	8.00	TRP-TRP	1.07	12.23	23.01
GLY-PHE	2.59	8.11	5.05	LEU-MET	2.08	10.40	7.50	MET-LYS	3.36	8.85	6.77	TRP-SER	2.13	9.80	8.63
GLY-PRO	2.75	7.11	2.82	LEU-PHE	-0.43	11.15	15.58	MET-ARG	1.34	10.15	11.00	TRP-THR	1.93	9.48	12.00
GLY-TYR	2.58	8.50	5.20	LEU-PRO	0.63	8.93	9.00	MET-ASP	3.49	7.60	7.54	TRP-LYS	3.13	9.53	18.13
GLY-HIS	0.71	8.20	6.18	LEU-TYR	3.09	10.49	9.78	MET-ASN	0.83	8.96	12.00	TRP-ARG	3.61	10.72	18.34
GLY-TRP	-0.70	9.41	9.22	LEU-HIS	0.54	9.32	15.00	MET-GLU	3.69	8.64	7.32	TRP-ASP	2.49	10.48	11.00
GLY-SER	3.25	6.70	1.97	LEU-TRP	4.49	10.05	11.26	MET-GLN	0.16	9.91	12.00	TRP-ASN	1.68	10.10	17.00
GLY-THR	3.16	6.70	2.94	LEU-SER	1.84	8.13	6.53	PHE-PHE	2.89	11.68	15.21	TRP-GLU	2.91	10.30	14.27
GLY-LYS	2.60	7.50	3.77	LEU-THR	1.74	8.42	9.00	PHE-PRO	2.87	9.30	10.43	TRP-GLN	3.13	9.87	17.51
GLY-ARG	2.40	8.30	4.40	LEU-LYS	0.87	8.68	11.00	PHE-TYR	2.86	11.36	13.96	SER-SER	3.44	6.96	3.63
GLY-ASP	2.42	7.01	4.12	LEU-ARG	2.79	9.83	8.50	PHE-HIS	3.40	9.73	14.59	SER-THR	2.78	7.26	5.81
GLY-ASN	2.81	7.10	3.77	LEU-ASP	1.64	8.01	11.95	PHE-TRP	3.60	11.16	16.67	SER-LYS	1.28	8.30	7.50
GLY-GLU	2.09	7.49	4.69	LEU-ASN	2.93	8.33	8.53	PHE-SER	0.66	9.06	9.50	SER-ARG	1.89	8.59	8.00
GLY-GLN	2.29	7.40	4.72	LEU-GLU	0.32	8.79	14.00	PHE-THR	2.49	9.41	8.50	SER-ASP	2.51	7.83	7.43
ALA-ALA	3.61	6.41	1.00	LEU-GLN	1.14	8.99	13.00	PHE-LYS	0.71	9.40	17.50	SER-ASN	2.84	7.99	6.21
ALA-VAL	3.34	6.81	2.90	ILE-ILE	0.19	9.07	15.00	PHE-ARG	1.29	10.30	19.00	SER-GLU	1.82	8.56	8.19
ALA-LEU	2.76	8.00	3.10	ILE-CYS	2.03	8.37	6.50	PHE-ASP	0.38	8.71	17.00	SER-GLN	1.81	8.44	8.00
ALA-ILE	2.90	7.50	3.56	ILE-MET	0.92	9.60	11.00	PHE-ASN	1.59	8.83	15.25	THR-THR	2.74	7.52	8.15
ALA-CYS	3.04	7.20	1.89	ILE-PHE	3.25	10.00	11.01	PHE-GLU	0.91	9.74	16.00	THR-LYS	1.36	8.19	11.00
ALA-MET	3.02	8.10	2.84	ILE-PRO	1.99	8.58	7.69	PHE-GLN	2.16	9.57	15.88	THR-ARG	2.03	8.97	10.56
ALA-PHE	2.45	8.99	4.20	ILE-TYR	3.18	9.98	10.00	PRO-PRO	2.49	8.62	5.02	THR-ASP	2.57	7.89	9.49
ALA-PRO	2.86	7.00	2.65	ILE-HIS	2.48	8.66	12.21	PRO-TYR	4.61	9.28	8.75	THR-ASN	1.69	8.26	10.61
ALA-TYR	2.85	9.20	3.78	ILE-TRP	4.51	10.36	9.60	PRO-HIS	4.33	8.33	7.57	THR-GLU	2.69	8.33	9.77
ALA-HIS	2.67	8.00	3.98	ILE-SER	2.55	7.49	7.00	PRO-TRP	4.81	9.29	10.81	THR-GLN	2.07	8.57	11.00
ALA-TRP	2.35	9.60	4.62	ILE-THR	1.74	7.99	11.00	PRO-SER	0.94	8.15	6.10	LYS-LYS	1.35	7.70	13.00
ALA-SER	3.36	6.69	1.93	ILE-LYS	1.65	8.46	10.20	PRO-THR	2.40	8.00	6.50	LYS-ARG	1.64	9.31	12.00
ALA-THR	3.25	6.82	2.87	ILE-ARG	0.28	9.75	13.26	PRO-LYS	2.86	7.91	6.87	LYS-ASP	2.45	10.09	10.46
ALA-LYS	1.97	7.22	4.00	ILE-ASP	2.63	7.73	10.15	PRO-ARG	3.98	9.38	6.31	LYS-ASN	1.25	8.86	13.00
ALA-ARG	2.89	8.00	3.50	ILE-ASN	1.27	8.23	13.24	PRO-ASP	4.11	7.62	5.49	LYS-GLU	0.23	10.30	17.00
ALA-ASP	3.26	6.86	3.25	ILE-GLU	2.82	8.07	10.17	PRO-ASN	4.19	7.83	5.42	LYS-GLN	0.90	8.92	16.00
ALA-ASN	2.53	7.47	3.50	ILE-GLN	2.76	8.70	9.75	PRO-GLU	1.08	8.96	9.86	ARG-ARG	0.19	11.43	16.00
ALA-GLU	2.63	7.29	3.78	CYS-CYS	3.31	7.52	4.00	PRO-GLN	2.00	8.79	9.32	ARG-ASP	3.04	11.13	12.08
ALA-GLN	2.45	7.66	4.00	CYS-MET	2.70	8.59	5.40	TYR-TYR	3.59	10.37	15.00	ARG-ASN	3.30	9.35	10.04
VAL-VAL	2.21	7.94	8.50	CYS-PHE	1.32	9.38	8.50	TYR-HIS	4.41	9.83	14.43	ARG-GLU	-1.22	11.35	25.00
VAL-LEU	1.46	9.19	8.50	CYS-PRO	-0.33	8.72	6.32	TYR-TRP	4.05	10.54	15.21	ARG-GLN	3.98	10.12	10.02
VAL-ILE	1.56	8.35	11.00	CYS-TYR	-1.10	8.61	14.00	TYR-SER	0.02	9.61	10.00	ASP-ASP	1.71	8.57	11.00
VAL-CYS	2.60	7.90	5.00	CYS-HIS	3.49	8.92	5.06	TYR-THR	1.65	10.16	9.00	ASP-ASN	3.72	8.35	8.43
VAL-MET	2.34	9.25	7.00	CYS-TRP	4.45	9.61	7.08	TYR-LYS	3.70	9.53	13.38	ASP-GLU	1.07	8.80	12.35
VAL-PHE	0.67	10.34	12.00	CYS-SER	2.78	7.26	3.57	TYR-ARG	2.98	10.37	18.55	ASP-GLN	3.29	9.08	9.20
VAL-PRO	1.98	8.04	7.00	CYS-THR	2.86	7.34	5.18	TYR-ASP	-1.90	11.38	16.00	ASN-ASN	1.24	8.94	13.15
VAL-TYR	1.92	10.06	9.87	CYS-LYS	0.72	7.70	7.50	TYR-ASN	3.34	9.50	11.05	ASN-GLU	2.80	8.65	12.00
VAL-HIS	3.42	8.43	8.08	CYS-ARG	3.26	8.36	5.32	TYR-GLU	3.31	10.88	10.77	ASN-GLN	3.73	8.93	9.20
VAL-TRP	2.96	10.37	10.52	CYS-ASP	2.84	7.42	5.69	TYR-GLN	0.97	10.45	18.92	GLU-GLU	3.74	8.47	8.03
VAL-SER	2.27	7.39	6.11	CYS-ASN	1.43	7.88	7.00	HIS-HIS	0.65	10.45	23.00	GLU-GLN	2.15	9.30	12.00
VAL-THR	2.59	7.55	8.00	CYS-GLU	3.39	7.65	5.26	HIS-TRP	0.39	10.52	32.00	GLN-GLN	2.36	10.13	9.80
VAL-LYS	2.17	7.95	8.49	CYS-GLN	3.60	8.14	5.35	HIS-SER	1.04	9.09	10.00				
VAL-ARG	2.51	8.73	9.50	MET-MET	0.97	10.15	10.00	HIS-THR	2.31	9.02	10.00				

Table B.2: The contact potential from [Dombkowski and Crippen, 2000], where the potential value is computed by eq. (B.1).

ALA	-379
ARG	116
ASN	218
ASP	297
CYS	-295
GLN	305
GLU	158
GLY	-91
HIS	65
ILE	-440
LEU	-405
LYS	319
MET	-313
PHE	-229
PRO	132
SER	74
THR	-37
TRP	-126
TYR	-153
VAL	-436
ALA	-897
ARG	56
ASN	112
ASP	142
CYS	-695
GLN	217
GLU	66
GLY	-421
HIS	-163
ILE	-557
LEU	-413
LYS	216
MET	-459
PHE	-498
PRO	-286
SER	-259
THR	-322
TRP	-403
TYR	-448
VAL	-824

Table B.3: Overall threading potential.

ALA	-897
ARG	56
ASN	112
ASP	142
CYS	-695
GLN	217
GLU	66
GLY	-421
HIS	-163
ILE	-557
LEU	-413
LYS	216
MET	-459
PHE	-498
PRO	-286
SER	-259
THR	-322
TRP	-403
TYR	-448
VAL	-824

Table B.4: Potential for 5 Å.

ALA	-177
ARG	133 217
ASN	249 241 -85
ASP	344 21 184 522
CYS	-153 15 -26 310 -1568
GLN	331 -83 324 620 375 519
GLU	183 214 132 246 138 239 72
GLY	20 209 151 316 -171 485 274 -80
HIS	136 170 150 95 -352 219 242 94 -358
ILE	-406 -4 119 298 -485 218 61 95 -64 -988
LEU	-403 -20 214 386 -470 231 46 148 -75 -731 -743
LYS	349 577 373 70 216 -14 318 453 462 132 230 537
MET	-271 72 106 309 -463 330 115 36 -167 -568 -505 290 -696
PHE	-141 59 144 383 -462 332 111 13 -87 -577 -511 277 -530 -519
PRO	282 257 284 516 -146 499 335 243 200 52 93 575 59 45 163
SER	188 228 63 261 -98 349 178 111 104 36 86 384 82 47 268 103
THR	56 172 61 265 -219 260 99 102 44 -259 -96 279 -84 -1 161 108 -76
TRP	-36 -77 -15 307 -495 336 4 3 -148 -364 -348 237 -382 -385 -155 96 28 -478
TYR	-56 -58 45 327 -342 303 66 49 -96 -510 -419 58 -383 -357 -78 81 -11 -394 -354
VAL	-299 32 166 338 -526 274 114 69 -75 -902 -711 187 -511 -510 64 95 -239 -314 -393 -923

Table B.5: Potential for 7 Å.

ALA	-299
ARG	69 193
ASN	114 236 66
ASP	141 97 241 330
CYS	-252 -89 -161 70 -1238
GLN	258 170 467 549 141 591
GLU	110 277 244 355 -40 474 236
GLY	-171 59 9 119 -312 296 92 -212
HIS	-75 107 64 55 -437 194 159 -112 -333
ILE	-169 88 149 219 -344 294 175 -75 -68 -624
LEU	-122 70 231 300 -336 360 193 -36 -44 -537 -588
LYS	225 521 322 176 97 250 483 247 387 233 295 447
MET	-149 62 123 227 -371 307 134 -112 -164 -448 -364 297 -515
PHE	-252 -32 -6 99 -531 185 76 -114 -252 -522 -468 130 -470 -639
PRO	42 73 140 205 -149 319 160 -3 -3 120 74 306 -12 -45 16
SER	-33 150 84 225 -300 397 178 -64 -11 -8 61 291 11 -117 111 -38
THR	-70 143 119 233 -292 359 197 -75 -22 -86 -28 308 -80 -175 109 33 25
TRP	-213 -54 -85 60 -526 164 -28 -147 -263 -259 -291 187 -329 -525 -208 -97 -158 -653
TYR	-165 10 -31 -42 -411 137 63 -142 -232 -315 -267 148 -356 -498 -106 -110 -139 -421 -460
VAL	-243 50 152 174 -347 254 132 -144 -124 -416 -388 215 -360 -477 58 -35 -112 -287 -308 -361

Table B.6: Potential for 9 Å.

ALA	-248
ARG	14 84
ASN	106 238 77
ASP	135 146 249 310
CYS	-143 -37 2 78 -888
GLN	130 121 353 354 111 258
GLU	29 176 191 281 -30 289 129
GLY	-5 182 188 241 -121 343 197 28
HIS	-27 68 145 152 -231 189 140 57 -215
ILE	-274 -83 -41 5 -286 3 -27 -71 -139 -489
LEU	-255 -124 20 58 -262 -20 -88 -2 -133 -345 -388
LYS	163 355 288 296 131 246 317 353 300 -29 38 247
MET	-203 -41 33 57 -289 81 -32 -65 -178 -346 -298 120 -444
PHE	-141 -24 14 70 -287 106 9 3 -135 -351 -341 109 -314 -428
PRO	104 149 243 278 -35 350 228 185 102 64 43 366 41 53 152
SER	41 197 151 262 -133 334 169 140 87 -88 -44 312 -42 -42 225 76
THR	-39 133 124 183 -153 264 147 80 45 -145 -91 262 -63 -72 180 66 1
TRP	-136 -5 -23 40 -255 132 -43 -42 -258 -253 -262 191 -307 -287 8 -19 -95 -490
TYR	-89 3 -24 41 -190 91 10 -2 -121 -276 -213 127 -253 -294 3 -38 -64 -325 -287
VAL	-275 -73 12 37 -288 50 -34 -76 -147 -410 -331 53 -304 -282 24 -65 -132 -259 -235 -409

Table B.7: Potential for 11 Å.

ALA	-84
ARG	-34 -106
ASN	13 44 -111
ASP	-26 -13 -35 -48
CYS	139 107 110 114 -178
GLN	-10 -50 37 -14 106 -74
GLU	-6 3 -11 -1 117 29 -47
GLY	-55 -19 -23 -39 98 4 6 -81
HIS	8 -20 25 -9 86 -4 22 -4 -54
ILE	32 21 -23 -9 167 -5 40 4 31 -9
LEU	0 -33 34 5 135 -19 7 10 8 54 -5
LYS	45 55 -3 -1 126 -11 58 28 45 -3 46 -54
MET	-25 -28 -8 -36 90 -39 5 -34 -28 -10 11 -4 -146
PHE	22 -9 -29 -29 120 -7 17 -16 -10 24 25 10 -19 -34
PRO	-48 -62 -13 -35 47 -23 -15 -61 -43 2 -28 32 -44 -51 -114
SER	22 41 -37 -3 90 53 10 -4 40 27 34 41 27 3 -5 -34
THR	-10 18 -40 -21 107 32 7 -38 19 16 32 31 -8 2 -37 -19 -49
TRP	-20 -38 -73 -55 94 9 -47 -72 -52 47 4 32 -31 -52 -110 -36 -46 -272
TYR	10 -10 -84 -57 131 -12 -3 -40 -16 3 25 -20 -23 -49 -56 -25 -35 -120 -122
VAL	-0 3 21 0 155 10 37 -18 25 51 47 43 7 37 -25 36 11 19 20 33

Table B.8: Potential for 13 Å.

Appendix C

3D Prediction Challenge Results (CASP7)

In the following we list the results of the “Automated assessment of protein structure prediction” CASP7 from <http://zhang.bioinformatics.ku.edu/casp7/index.html>.

The leadint methods are

- Zhang: threading combined with clustering
- Tasser: threading
- FAMS: threading
- Baker: the Rosetta method and server; ab initio
- CHIMERA: comparative modelling by sequence-sequence comparison and then refinement by 3D modeling

In the following the legend for the results list afterwards.

- Predictors—Name of groups, A name with '*' indicates a server prediction.
- N——Number of targets used to calculate the cumulative score.
- Rank——Rank of the predictors on the basis of TM-score or GDT-score.
- TM_1——TM-score of the first models. A TM-score < 0.17 implies that the prediction is close to random structures. TM-score=0 means either that the target was not submitted or that there is no overlap between the predicted model and the solved structure.
- Zscore——Z-score derived from corresponding scores for each given target.
- MS_1——MaxSub-score of the first model. (The MaxSub-score is calculated based on the TM-score search engine).
- GDT_1——GDT-score of the first model. (The GDT-score is taken from official CASP7 assessors).
- RM_1——RMSD to native of the first model.

- cov——The percentage of residues in the predicted model relative to native.
- DGyr——Difference of radius of gyration between model and the native structure, i.e. $DGyr = |Gyr_{model} - Gyr_{native}|$, where Gyr_{model} and Gyr_{native} are the radius of gyration of the predicted model and the native structure, respectively. When the radius of gyration is calculated, the same set of residues in model and the native structure are used.
- NC——Number of clashed models. A clashed model is defined according to the Valencia's rule (Proteins Suppl 7: 27-45, 2005), i.e. the model with more than 4 severe clashes (Ca-Ca pair-wise distance < 1.9 Angstroms) or with more than 50 bumps (Ca-Ca pair-wise distance < 3.6 Angstroms).
- TM_B——TM-score of the best in top-five models.
- MS_B——MaxSub-score of the best in top-five models
- GDT_B——GDT-score of the best in top-five models.
- RM_B——RMSD to native of the model with the best TM-score in top-five models.
- L_seq——Length of the target sequences for modeling.
- L_native——Number of the residues in the solved structures.

The result list:

----- Cumulative Score of 113 targets (ALL), ranked by TM-score of the first model -----													
Predictors	(N)	Rank	TM_1(Zscore)	MS_1(Zscore)	GDT_1(Zscore)	RM_1(cov)	DGyr(NC)		TM_B(Zscore)	MS_B(Zscore)	GDT_B(Zscore)	RM_B(cov)	DGyr(NC)
Zhang(113)	1	77.69(105.2)	69.21(114.8)	72.39(109.6)	5.5(100)	0.8(0)		80.43(113.5)	72.02(119.9)	75.04(118.4)	4.8(100)	0.7(0)	
TASSER(113)	2	76.06(92.8)	67.09(95.2)	70.37(94.2)	5.6(100)	0.6(0)		77.70(86.6)	69.03(89.0)	71.83(86.4)	5.4(100)	0.7(0)	
Zhang-Server(113)	3	75.92(87.4)	67.54(92.8)	70.47(88.3)	5.8(100)	0.8(0)		79.05(101.4)	70.46(103.2)	73.55(104.0)	5.1(100)	0.7(0)	
CHIMERA(113)	4	75.16(80.8)	66.26(85.5)	69.69(83.7)	6.0(99)	0.9(0)		77.64(87.4)	68.96(90.8)	72.21(90.8)	5.6(99)	0.9(0)	
Baker(111)	5	74.80(108.4)	65.79(113.9)	69.43(110.9)	5.8(99)	0.8(0)		77.51(111.2)	69.07(119.9)	72.28(118.0)	5.4(99)	0.7(0)	
fams-ace(113)	6	74.66(73.4)	66.17(78.6)	69.50(76.9)	6.6(99)	1.2(0)		76.98(72.7)	68.71(76.7)	71.59(74.2)	5.9(99)	1.0(0)	
luethy(113)	7	74.45(77.8)	65.40(80.7)	68.62(78.4)	5.8(100)	0.6(0)		74.45(54.6)	65.40(53.9)	68.62(54.4)	5.8(100)	0.6(0)	
CIRCLE-FAMS(113)	8	74.31(78.4)	65.60(85.9)	69.18(82.5)	6.3(99)	0.9(0)		78.48(95.8)	69.84(101.2)	72.94(97.8)	5.5(99)	0.8(0)	
MQAP-Consensus(113)	9	74.21(75.6)	65.11(73.8)	69.02(79.3)	5.9(99)	0.8(0)		74.21(51.4)	65.11(47.0)	69.02(54.8)	5.9(99)	0.8(0)	
verify(113)	10	74.09(77.0)	64.95(82.6)	68.74(80.3)	5.9(100)	0.7(0)		74.09(52.9)	64.95(54.9)	68.74(55.9)	5.9(100)	0.7(0)	
hPredGrp(112)	11	73.15(66.2)	64.25(67.7)	67.49(66.3)	6.1(98)	0.9(0)		73.15(42.9)	64.25(41.6)	67.49(42.3)	6.1(98)	0.9(0)	
fams-multi(113)	12	73.12(62.2)	64.35(65.1)	67.90(65.8)	7.6(99)	2.1(0)		75.08(58.0)	66.53(60.6)	69.77(60.6)	5.9(99)	0.8(0)	
Bates(113)	13	73.04(71.9)	64.09(77.1)	67.57(72.4)	6.4(99)	0.9(0)		75.88(69.1)	67.38(73.3)	70.33(70.8)	6.1(99)	0.8(0)	
SBC(113)	14	72.49(76.2)	63.57(81.4)	67.02(79.4)	5.5(95)	0.7(0)		77.91(89.0)	69.49(93.4)	72.62(92.9)	5.3(99)	0.7(0)	
Jones-UCL(112)	15	72.19(67.2)	63.23(72.5)	66.70(69.2)	6.1(98)	0.8(0)		73.37(61.3)	64.28(62.6)	67.83(61.7)	6.1(98)	0.8(0)	
HHpred2(113)	16	72.16(52.3)	63.10(55.4)	66.93(54.9)	11.0(99)	5.4(0)		72.16(28.1)	63.10(28.4)	66.93(29.8)	11.0(99)	5.4(0)	
MetaTasser(113)	17	71.95(56.5)	61.42(47.5)	65.56(51.5)	6.3(100)	0.9(0)		73.90(51.2)	64.09(45.1)	67.57(46.7)	6.1(100)	0.9(0)	
Pmodeller6(113)	18	71.91(61.8)	62.52(61.5)	66.41(63.8)	6.7(97)	1.6(1)		75.48(68.5)	66.70(71.6)	69.96(70.0)	6.3(98)	1.3(0)	
HHpred3(113)	19	71.64(51.7)	62.69(55.6)	66.28(52.2)	8.8(98)	2.9(0)		71.64(27.1)	62.69(28.0)	66.28(27.1)	8.8(98)	2.9(0)	
ROBETTA(113)	20	71.57(58.1)	61.90(58.3)	66.29(60.4)	6.5(99)	0.9(0)		75.51(72.0)	66.43(74.6)	70.05(73.0)	6.6(99)	1.5(0)	
CIRCLE(113)	21	71.45(49.0)	62.37(50.2)	65.92(48.8)	7.2(99)	1.3(0)		73.97(48.6)	65.33(49.9)	68.57(48.8)	6.2(99)	0.8(0)	

BayesHH(113) 22 71.10(46.4) 61.82(45.0) 65.85(47.2) 9.1(100) 3.3(0) | 71.10(21.5) 61.82(17.9) 65.85(21.7) 9.1(100) 3.3(0)
 Pcons6(113) 23 70.93(49.8) 62.14(53.2) 65.48(50.7) 6.9(97) 1.3(0) | 73.64(48.6) 65.08(52.9) 68.35(51.6) 6.5(97) 1.3(0)
 HHpred1(112) 24 70.88(42.7) 61.80(45.1) 65.38(43.5) 11.4(99) 5.8(0) | 70.88(19.1) 61.80(18.6) 65.38(18.8) 11.4(99) 5.8(0)
 LEE(111) 25 70.73(56.0) 62.30(59.3) 65.99(60.6) 6.9(100) 1.0(0) | 73.20(58.4) 64.79(59.3) 68.28(62.6) 6.4(100) 0.9(0)
 SAMUDRALA(111) 26 70.72(57.1) 61.82(57.5) 65.87(61.4) 6.5(99) 1.0(0) | 74.58(67.3) 66.08(68.6) 69.58(71.8) 5.8(99) 0.8(0)
 Sternberg(112) 27 70.56(50.0) 60.62(43.4) 64.91(47.6) 6.4(99) 0.9(0) | 72.08(45.9) 62.01(37.8) 66.19(42.3) 6.1(99) 0.9(0)
 FAMSD(113) 28 70.50(40.9) 61.31(41.8) 65.26(43.5) 7.1(99) 1.3(0) | 72.65(39.8) 63.80(40.0) 67.16(39.0) 6.6(98) 1.1(0)
 keasar(113) 29 70.42(51.8) 60.05(44.4) 64.30(48.2) 6.5(98) 0.9(0) | 72.82(45.5) 62.86(39.4) 66.73(41.6) 6.4(99) 0.9(0)
 FAMS(113) 30 70.41(41.8) 61.23(40.5) 65.18(42.9) 7.2(99) 1.2(0) | 73.71(46.8) 65.12(48.6) 68.46(48.7) 6.5(99) 1.0(0)
 beautshot(113) 31 70.40(42.3) 61.14(42.0) 64.35(37.9) 6.9(99) 0.9(0) | 70.40(18.8) 61.14(15.8) 64.35(13.4) 6.9(99) 0.9(0)
 RAPTOR-ACE(113) 32 70.28(41.4) 61.08(40.9) 64.66(41.3) 7.2(100) 1.1(0) | 74.00(51.0) 65.43(54.0) 68.52(50.5) 6.7(100) 1.1(0)
 UNI-EID_expm(111) 33 70.09(39.1) 61.48(44.4) 64.37(37.9) 6.5(95) 1.0(30) | 70.09(16.5) 61.48(19.3) 64.37(14.0) 6.5(95) 1.0(30)
 SAMUDRALA-AB(111) 34 70.01(54.8) 61.06(55.0) 65.11(57.1) 6.5(99) 0.9(0) | 72.88(54.8) 64.40(56.8) 67.93(58.1) 6.1(99) 0.8(0)
 SP3(113) 35 69.89(36.4) 60.86(36.5) 64.45(36.6) 8.2(99) 2.0(0) | 72.76(41.4) 64.00(41.6) 67.35(41.6) 7.3(100) 1.3(0)
 SP4(113) 36 69.63(38.2) 60.24(37.4) 64.03(37.3) 7.8(99) 1.5(0) | 73.40(51.7) 64.29(50.7) 67.95(52.0) 6.6(99) 1.0(0)
 GeneSilico(107) 37 69.59(77.2) 61.17(82.1) 64.49(79.4) 6.1(99) 0.9(0) | 71.80(77.4) 63.31(77.9) 66.43(76.6) 5.7(99) 0.9(0)
 SAM-T06(113) 38 69.40(46.6) 59.57(40.5) 64.46(49.1) 6.8(100) 0.8(0) | 73.69(54.7) 64.66(52.8) 68.26(55.0) 6.2(99) 0.7(0)
 RAPTOR(113) 39 69.37(38.0) 59.72(34.1) 64.08(36.9) 6.9(100) 0.9(0) | 73.97(55.6) 64.82(55.1) 68.36(55.8) 6.8(100) 1.4(0)
 andante(111) 40 69.17(43.0) 59.99(41.0) 64.34(48.0) 6.9(99) 0.9(0) | 71.59(46.5) 62.65(43.8) 66.62(49.4) 6.5(99) 0.9(0)
 RAPTORESS(113) 41 68.89(35.0) 59.04(30.0) 63.16(31.4) 7.0(100) 0.9(0) | 73.47(51.5) 64.23(50.7) 67.67(48.8) 6.5(100) 0.9(0)
 Ma-OPUS(112) 42 68.83(38.6) 58.93(30.1) 63.72(41.0) 6.8(100) 0.9(0) | 71.60(40.2) 62.11(33.7) 66.45(41.5) 6.2(100) 0.8(0)
 shub(112) 43 68.67(26.5) 59.35(25.1) 63.06(24.8) 6.6(96) 0.9(1) | 68.67(2.4) 59.35(-1.4) 63.06(-0.3) 6.6(96) 0.9(1)
 UNI-EID_bnmx(113) 44 68.43(27.2) 60.75(42.1) 63.54(31.7) 6.0(88) 1.0(0) | 71.56(27.1) 64.21(41.4) 66.63(30.7) 5.8(91) 1.0(0)
 SPARKS2(113) 45 68.40(25.0) 59.20(24.8) 63.02(23.8) 8.6(99) 2.2(0) | 72.05(32.9) 63.37(34.0) 66.80(32.7) 6.8(100) 0.9(0)
 FUNCTION(113) 46 68.19(28.4) 58.84(23.8) 62.82(27.1) 7.2(97) 1.1(0) | 71.27(28.4) 62.42(27.4) 66.15(29.6) 7.2(99) 1.4(0)
 beautshotbase(110) 47 67.58(27.3) 59.07(30.3) 62.33(26.4) 6.5(94) 0.9(0) | 67.58(4.0) 59.07(4.8) 62.33(2.4) 6.5(94) 0.9(0)
 UCB-SHI(108) 48 67.37(33.7) 58.78(33.7) 62.71(35.4) 6.3(98) 0.7(0) | 70.05(38.1) 61.69(36.0) 65.29(38.0) 5.9(98) 0.7(0)
 GeneSilicoMetaServer(111) 49 67.16(33.5) 58.19(31.4) 62.01(33.5) 6.8(95) 1.0(0) | 70.70(32.3) 62.01(31.5) 65.47(33.7) 8.3(98) 2.7(0)
 CBSU(113) 50 66.90(20.6) 56.92(11.9) 61.87(20.2) 7.4(99) 1.1(0) | 69.31(12.9) 59.55(4.9) 64.08(12.5) 7.1(99) 1.0(0)
 FOLDpro(113) 51 66.89(15.7) 58.02(12.5) 62.08(17.2) 8.0(100) 1.3(0) | 69.22(13.2) 60.63(12.4) 64.23(12.1) 7.6(100) 1.2(0)
 Bilab(113) 52 66.80(28.1) 57.51(27.1) 61.51(25.5) 7.7(100) 1.1(0) | 71.07(36.5) 62.06(37.0) 65.56(34.7) 7.0(100) 0.9(0)
 3Dpro(113) 53 66.79(24.2) 58.18(23.1) 62.37(27.8) 7.8(99) 1.2(0) | 69.02(16.6) 60.39(15.6) 64.35(17.5) 7.3(100) 0.9(0)
 ROKKO(109) 54 66.64(42.6) 57.23(41.0) 61.48(42.9) 6.8(99) 0.8(0) | 69.76(46.3) 60.76(44.9) 64.55(47.3) 6.4(99) 0.9(0)
 Ma-OPUS-server(113) 55 66.36(19.3) 56.63(13.8) 61.76(20.6) 7.7(100) 1.2(0) | 71.74(35.8) 62.41(32.6) 66.36(35.2) 6.9(100) 0.9(0)
 Iwyrwicz(111) 56 66.08(20.1) 56.35(15.9) 60.83(20.7) 7.0(97) 1.0(0) | 66.08(-5.2) 56.35(-11.4) 60.83(-5.3) 7.0(97) 1.0(0)
 SAM_T06_server(113) 57 66.00(19.6) 55.84(13.2) 61.10(20.4) 7.3(100) 1.0(0) | 71.08(29.3) 63.15(39.2) 66.11(33.9) 5.6(92) 0.7(0)
 Pan(113) 58 65.62(7.1) 56.15(3.7) 61.05(9.0) 7.8(100) 1.3(1) | 69.34(14.2) 60.51(11.6) 64.54(14.4) 7.2(100) 1.1(1)
 Phyre-2(111) 59 65.45(15.8) 56.30(11.2) 60.42(14.7) 7.7(98) 1.6(0) | 67.24(9.4) 58.31(5.6) 62.27(9.2) 7.3(98) 1.2(0)
 PROTINFO-AB(112) 60 65.39(17.6) 56.15(14.6) 60.38(16.7) 7.7(99) 1.1(0) | 67.19(9.9) 58.04(6.4) 62.07(8.2) 7.2(99) 0.9(0)
 honiglab(100) 61 65.35(39.5) 56.63(37.1) 60.15(39.7) 6.0(99) 0.6(0) | 66.04(24.3) 57.35(20.0) 60.80(23.8) 5.8(99) 0.6(0)
 mGen-3D(112) 62 64.87(9.8) 56.40(15.6) 60.23(14.4) 6.4(87) 1.0(0) | 64.87(-17.2) 56.40(-12.7) 60.23(-13.0) 6.4(87) 1.0(0)
 UNI-EID_sfst(110) 63 64.83(4.9) 58.06(20.3) 60.21(8.8) 5.4(82) 0.9(0) | 68.97(12.9) 62.55(32.9) 64.32(17.6) 4.9(84) 0.8(0)
 PROTINFO(113) 64 64.73(19.3) 55.94(17.7) 60.02(22.6) 6.8(93) 0.9(0) | 70.13(21.8) 61.13(20.5) 65.08(23.9) 6.8(98) 1.0(0)
 ROKKY(111) 65 64.64(21.5) 56.17(25.7) 60.21(23.4) 9.0(100) 2.1(2) | 69.05(33.2) 61.31(40.8) 64.48(34.7) 8.0(100) 1.6(3)
 Bilab-ENABLE(112) 66 64.36(2.4) 54.62(1.5) 59.14(3.2) 9.0(100) 2.4(0) | 68.41(9.3) 58.98(8.1) 62.97(8.5) 7.6(100) 1.4(0)
 Chen-Tan-Kihara(108) 67 64.33(20.0) 55.34(18.8) 59.19(19.1) 8.5(100) 2.0(0) | 68.03(26.8) 59.30(25.6) 62.81(27.4) 7.5(100) 1.4(0)

```

*NN_PUT_lab*(111) 68 64.19( 4.6) 55.32( 7.3) 59.10( 3.5) 7.7( 96) 1.3( 0) | 64.19( -20.8) 55.32( -19.4) 59.10( -22.6) 7.7( 96) 1.3( 0)
Huber-Torda(112) 69 63.93( 3.0) 54.65( -3.3) 59.24( 1.4) 7.9( 96) 0.9( 0) | 65.32( -12.4) 56.05( -18.4) 60.44( -14.8) 7.7( 97) 0.9( 0)
NanoModel(113) 70 63.75( -1.3) 53.85( -6.7) 58.69( -3.3) 7.6( 99) 0.8( 0) | 72.09( 36.3) 62.58( 31.7) 66.35( 32.9) 6.2( 99) 0.7( 0)
*LOOPP*(113) 71 63.65( 4.3) 54.70( 4.9) 58.73( 3.7) 7.5( 94) 0.9( 0) | 68.34( 15.0) 59.41( 16.9) 63.32( 18.0) 6.4( 95) 0.7( 0)
LUO(103) 72 63.40( 49.1) 54.86( 48.0) 58.68( 50.2) 7.0(100) 1.0( 0) | 68.38( 68.9) 60.17( 68.8) 63.29( 69.3) 6.0(100) 0.8( 0)
AMU-Biology(105) 73 62.83( 33.9) 54.36( 30.7) 58.30( 34.7) 5.6( 93) 0.7( 0) | 68.76( 38.0) 60.26( 37.3) 63.86( 36.6) 5.8( 98) 0.7( 0)
KIST(108) 74 62.79( 15.0) 53.09( 6.1) 57.73( 15.4) 7.1( 98) 0.9( 0) | 67.81( 29.1) 58.63( 22.9) 62.64( 29.2) 6.3( 98) 0.8( 0)
*nFOLD*(113) 75 62.70( -15.2) 53.92( -13.6) 57.97( -14.1) 7.6( 93) 1.1( 0) | 67.43( 0.3) 59.01( 3.8) 62.55( 1.6) 6.8( 93) 1.2( 0)
Ligand-Circle(100) 76 61.98( 43.3) 54.26( 46.4) 57.53( 43.8) 7.3( 99) 1.3( 0) | 68.46( 72.4) 61.19( 77.5) 63.94( 74.9) 5.7( 99) 0.8( 0)
*keasar-server*(108) 77 61.84( 11.1) 52.83( 10.8) 56.55( 8.2) 8.0( 96) 1.4( 0) | 67.90( 25.3) 59.22( 23.7) 62.45( 23.3) 6.9( 98) 1.0( 0)
*CaspIta-POX*(113) 78 61.62( -12.4) 53.25( -8.0) 56.67( -15.4) 7.6( 89) 1.3( 3) | 68.65( 11.1) 60.32( 15.9) 63.58( 11.3) 6.9( 94) 1.2( 3)
FEIG(111) 79 61.60( 0.3) 49.76( -18.6) 55.06( -10.1) 7.9( 99) 0.8( 0) | 67.29( 16.8) 57.50( 7.7) 61.81( 14.7) 7.2( 99) 0.8( 0)
*FUGUE*(113) 80 61.57( -14.8) 53.73( -8.6) 57.22( -13.0) 6.9( 88) 1.1( 0) | 66.70( -7.7) 58.47( -1.4) 62.02( -4.0) 6.8( 92) 1.0( 0)
*SAM-T02*(111) 81 61.30( -15.9) 54.03( -4.7) 56.62( -13.2) 5.0( 77) 0.7( 0) | 67.00( -7.6) 60.03( 8.3) 62.31( -4.6) 5.0( 82) 0.7( 0)
TENETA(111) 82 60.51( -12.9) 50.98( -19.5) 55.64( -15.9) 8.5( 99) 1.3( 0) | 63.16( -16.7) 53.83( -22.9) 58.47( -17.4) 8.1( 99) 1.2( 0)
MLee(103) 83 60.29( 14.6) 52.47( 15.0) 55.86( 13.1) 7.2( 96) 0.8( 0) | 64.94( 23.4) 57.07( 22.4) 60.33( 22.1) 6.4( 97) 0.8( 0)
*Phyre-1*(104) 84 60.24( -12.7) 52.27( -6.8) 55.40( -13.1) 5.6( 84) 0.6( 0) | 60.24( -35.4) 52.27( -30.5) 55.40( -36.0) 5.6( 84) 0.6( 0)
*FUGMOD*(104) 85 60.14( 2.6) 51.85( 2.4) 55.61( 2.5) 7.4( 96) 1.0( 0) | 64.62( 13.0) 56.17( 10.3) 59.67( 11.9) 6.9( 98) 0.8( 0)
*karypis.srv*(111) 86 60.09( -16.2) 50.73( -16.3) 54.78( -17.1) 7.6( 93) 0.9( 0) | 63.88( -8.9) 54.48( -9.6) 58.11( -14.4) 6.8( 93) 0.7( 0)
jive(104) 87 59.28( 11.3) 50.68( 12.5) 54.41( 11.9) 7.9( 98) 1.6( 0) | 63.94( 26.9) 55.82( 30.2) 59.06( 28.2) 7.0( 98) 1.4( 0)
SHORTLIE( 96) 88 58.87( 29.5) 51.98( 34.3) 55.21( 33.3) 5.9( 95) 0.7( 0) | 60.40( 24.0) 53.63( 27.0) 56.82( 27.7) 5.3( 95) 0.6( 0)
*FORTE1*(113) 89 58.86( -42.0) 49.71( -38.3) 53.95( -40.0) 9.2( 93) 2.2( 0) | 64.02( -30.1) 55.35( -25.1) 59.08( -28.6) 7.7( 92) 1.4( 0)
Softberry(107) 90 58.84( -10.7) 49.58( -17.0) 54.10( -10.9) 8.3( 98) 1.3( 0) | 58.84( -37.9) 49.58( -44.7) 54.10( -38.5) 8.3( 98) 1.3( 0)
*3D-JIGSAW_POPULUS*(108) 91 58.61( -18.3) 49.90( -16.5) 53.97( -19.0) 8.5( 94) 1.9( 0) | 60.78( -28.0) 52.41( -24.9) 56.06( -27.2) 8.2( 95) 1.7( 0)
*karypis.srv.2*(113) 92 58.44( -29.5) 49.38( -30.6) 53.74( -31.0) 9.4( 96) 2.0( 0) | 62.11( -25.8) 53.10( -25.6) 57.12( -27.1) 8.8( 96) 1.8( 0)
*FORTE2*(113) 93 58.34( -43.9) 49.25( -38.8) 53.29( -43.2) 9.7( 92) 2.8( 0) | 63.87( -29.5) 55.04( -24.5) 58.79( -28.7) 7.8( 92) 1.5( 0)
UAM-ICO-BIB(111) 94 58.33( 11.2) 49.84( 11.7) 54.04( 12.4) 7.1( 90) 0.9( 0) | 64.59( -12.1) 55.22( -12.2) 59.64( -10.2) 7.5( 97) 1.0( 0)
*3D-JIGSAW_RECOM*(108) 95 56.88( -31.9) 49.25( -26.0) 52.55( -31.2) 7.8( 89) 1.2( 0) | 60.21( -34.8) 52.34( -30.2) 55.69( -33.1) 7.2( 92) 1.1( 0)
NanoDesign( 90) 96 56.62( 7.4) 49.42( 6.5) 53.15( 10.4) 5.8( 97) 0.6( 0) | 61.42( 21.3) 54.70( 20.9) 57.63( 24.0) 4.9( 97) 0.6( 0)
*3D-JIGSAW*(109) 97 56.41( -36.9) 48.06( -35.1) 51.97( -37.3) 7.7( 90) 1.2( 0) | 62.72( -15.5) 53.96( -15.4) 57.94( -13.9) 6.9( 93) 1.2( 0)
Akagi(107) 98 55.97( -32.4) 47.61( -29.0) 51.21( -33.7) 7.9( 89) 1.2( 0) | 55.97( -60.0) 47.61( -56.9) 51.21( -61.3) 7.9( 89) 1.2( 0)
*SAM-T99*( 87) 99 55.57( -6.2) 49.31( 2.6) 51.21( -5.0) 4.4( 85) 0.8( 0) | 58.15( -7.2) 52.33( 3.2) 53.87( -4.7) 4.2( 85) 0.6( 0)
panther( 75) 100 55.44( 20.1) 48.90( 19.2) 50.67( 19.5) 4.4( 99) 0.5( 0) | 57.08( 20.2) 50.90( 19.3) 52.41( 19.6) 3.9( 99) 0.4( 0)
*Huber-Torda-Server*(110) 101 54.66( -48.3) 47.88( -37.7) 51.26( -44.9) 7.6( 81) 1.0( 1) | 60.70( -37.5) 53.37( -28.2) 56.71( -35.0) 7.3( 87) 0.8( 0)
LTB-WARSAW( 90) 102 54.15( 21.2) 46.44( 20.9) 50.18( 21.0) 7.3(100) 1.1( 0) | 56.11( 16.9) 48.88( 17.0) 52.12( 16.3) 7.0(100) 1.2( 0)
forecast(109) 103 51.79( -57.1) 43.70( -56.2) 47.60( -60.3) 17.0( 99) 9.0( 0) | 59.33( -28.3) 50.49( -33.1) 54.50( -32.8) 11.3(100) 3.7( 0)
*forecast-s*(110) 104 50.84( -62.0) 43.89( -54.4) 46.88( -63.2) 7.4( 77) 1.4( 0) | 57.62( -59.2) 50.52( -48.2) 53.54( -58.9) 7.8( 85) 1.3( 0)
*Distill*(113) 105 50.76( -72.7) 37.95( -97.8) 44.43( -88.1) 10.0(100) 1.3( 1) | 53.39( -84.3) 40.46( -109.1) 46.80( -102.6) 9.5(100) 1.3( 1)
MTUNIC(108) 106 50.75( -51.9) 38.11( -73.4) 45.05( -61.8) 9.5(100) 1.2( 0) | 56.45( -42.8) 43.78( -64.0) 50.11( -54.6) 8.4(100) 1.1( 0)
Distill_human(113) 107 50.73( -72.1) 37.89( -97.3) 44.45( -88.5) 10.0(100) 1.2( 1) | 53.33( -84.9) 40.42( -109.6) 46.81( -102.8) 9.5( 99) 1.3( 1)
MIG( 94) 108 50.68( -24.7) 42.07( -33.2) 47.00( -24.5) 7.2( 94) 0.8( 0) | 53.77( -28.5) 45.09( -38.1) 50.03( -28.2) 6.9( 96) 0.8( 0)
karypis( 91) 109 48.16( -6.4) 40.53( -6.6) 44.42( -4.3) 7.8( 93) 0.9( 0) | 48.99( -22.0) 41.52( -34.6) 44.45( -30.6) 4.7( 91) 0.5( 0)
LMU( 74) 110 46.63( -25.0) 40.36( -26.2) 43.28( -23.3) 4.8( 89) 0.5( 0) | 48.05( -32.0) 41.52( -34.6) 44.45( -30.6) 4.7( 91) 0.5( 0)
HIT-ITNLP(110) 111 45.42( -108.2) 35.65( -122.2) 41.03( -117.7) 13.2(100) 3.3( 0) | 50.11( -105.3) 40.84( -115.4) 45.60( -112.9) 11.3(100) 2.2( 0)
fleil( 71) 112 44.13( -27.1) 36.40( -35.4) 39.53( -30.6) 6.9(100) 0.9( 0) | 47.93( -16.5) 40.95( -22.8) 43.72( -17.0) 6.2(100) 0.8( 0)
Ma-OPUS-server2( 72) 113 42.56( 7.8) 36.46( 5.6) 39.70( 7.8) 7.3(100) 1.1( 0) | 46.99( 25.1) 41.01( 21.7) 43.64( 24.3) 6.2(100) 0.7( 0)

```

fais(84) 114 42.13(-5.5) 34.64(-0.1) 38.92(-4.4) 9.2(100) 1.4(0) | 43.06(-10.6) 35.47(-17.4) 39.78(-10.8) 8.9(100) 1.3(0)
 ZIB-THESEUS(97) 115 41.57(-92.9) 33.40(-91.2) 38.41(-92.5) 10.0(89) 1.3(2) | 47.25(-78.9) 38.78(-80.9) 43.87(-77.9) 8.7(92) 1.1(2)
 BioDec(72) 116 38.79(-31.8) 32.59(-37.0) 35.24(-39.2) 7.9(95) 1.6(0) | 38.79(-50.5) 32.60(-55.7) 35.24(-58.5) 7.9(95) 1.6(0)
 Nano3D(66) 117 38.16(-0.7) 32.75(-1.3) 35.69(-0.6) 7.1(97) 0.7(0) | 42.88(-20.0) 37.46(-17.5) 39.86(-20.4) 5.5(97) 0.4(0)
 panther2(82) 118 37.76(-69.9) 33.11(-55.3) 35.02(-65.9) 7.6(71) 1.5(21) | 37.76(-91.0) 33.11(-75.6) 35.02(-87.4) 7.6(71) 1.5(21)
 CPHmodels(56) 119 37.39(-21.6) 33.63(-14.4) 34.81(-19.1) 3.9(84) 0.4(0) | 37.39(-32.9) 33.63(-25.5) 34.81(-30.3) 3.9(84) 0.4(0)
 CADCMILAB(107) 120 36.12(-153.3) 25.33(-166.3) 32.54(-157.2) 12.8(99) 1.8(0) | 42.09(-149.2) 30.71(-162.9) 38.12(-151.9) 11.4(99) 1.6(0)
 TsaiLab(51) 121 34.77(-7.3) 29.81(-11.4) 32.05(-10.0) 5.6(99) 0.9(0) | 35.51(-12.1) 30.92(-15.5) 32.87(-14.4) 5.3(99) 0.7(0)
 Frankenstein(61) 122 30.08(-28.4) 24.52(-33.7) 27.60(-31.2) 8.1(91) 1.6(1) | 34.26(-33.1) 28.45(-38.4) 31.45(-36.6) 8.8(99) 2.1(2)
 gtg(55) 123 29.24(-39.5) 25.46(-35.1) 26.82(-38.6) 6.0(77) 0.8(0) | 31.61(-39.2) 28.16(-32.6) 29.08(-37.7) 6.0(78) 1.1(0)
 PUT_lab(77) 124 29.16(-104.1) 24.69(-93.9) 27.84(-99.8) 14.2(89) 4.8(4) | 29.60(-125.2) 25.12(-113.6) 28.28(-120.3) 14.2(89) 4.8(4)
 ABIpro(112) 125 29.08(-199.9) 19.05(-198.0) 26.37(-195.6) 14.4(100) 1.7(0) | 34.26(-205.0) 23.37(-204.5) 30.70(-201.9) 13.3(100) 1.7(0)
 SEZERMAN(72) 126 27.34(-76.2) 23.46(-63.8) 25.75(-73.9) 9.0(70) 1.6(0) | 27.59(-98.4) 23.66(-85.0) 25.99(-95.8) 9.1(71) 1.6(0)
 Wymore(44) 127 26.48(-6.1) 21.89(-10.6) 24.04(-7.2) 7.5(99) 0.9(0) | 27.05(-11.7) 22.69(-15.7) 24.68(-12.5) 7.3(99) 0.9(0)
 CHEN-WENDY(30) 128 25.07(-9.0) 23.20(-8.2) 23.74(-8.6) 2.7(99) 0.3(0) | 25.51(-8.3) 23.90(-8.2) 24.29(-8.2) 2.5(99) 0.3(0)
 Bystroff(61) 129 24.40(-60.0) 20.20(-56.7) 22.91(-59.0) 9.4(85) 1.1(0) | 24.79(-77.8) 20.45(-74.0) 23.26(-76.8) 9.5(86) 1.1(0)
 MIG_FROST(48) 130 20.78(-61.7) 16.40(-63.5) 19.44(-60.9) 7.3(77) 0.9(0) | 20.78(-76.1) 16.40(-77.4) 19.44(-75.2) 7.3(77) 0.9(0)
 karypis.srv.4(102) 131 19.40(-230.7) 10.73(-229.5) 16.60(-233.1) 15.6(93) 3.1(3) | 22.13(-263.0) 12.44(-259.9) 18.63(-265.9) 14.6(95) 3.2(3)
 FPSOLVER-SERVER(106) 132 19.11(-269.9) 10.77(-263.4) 16.60(-270.4) 16.8(99) 2.6(0) | 21.12(-303.2) 11.98(-293.0) 18.43(-299.9) 16.2(99) 2.4(0)
 taylor(48) 133 17.77(-5.1) 13.64(-9.9) 16.94(-3.5) 10.6(96) 1.0(0) | 19.78(-0.4) 15.37(-7.1) 18.53(-0.8) 9.6(96) 0.9(0)
 YASARA(22) 134 16.83(-8.0) 15.66(-9.0) 15.70(-7.5) 3.6(97) 0.6(0) | 17.31(-8.1) 16.18(-8.6) 16.23(-8.0) 3.4(97) 0.5(0)
 Schomburg-group(21) 135 16.52(-10.4) 14.69(-11.2) 15.13(-10.6) 2.7(96) 0.3(0) | 16.68(-9.2) 14.86(-9.3) 15.28(-9.1) 2.8(97) 0.3(0)
 Brooks_casp(22) 136 14.86(-10.0) 13.46(-9.8) 13.63(-9.3) 4.9(99) 0.7(0) | 15.85(-12.9) 14.72(-13.6) 14.73(-13.4) 4.6(99) 0.5(0)
 LMM-Bicocca(32) 137 14.13(-0.1) 12.33(-2.6) 13.28(-0.2) 6.0(75) 0.6(0) | 18.61(-9.2) 16.03(-13.3) 17.37(-9.5) 8.4(100) 0.8(0)
 MUMSSP(15) 138 12.63(-5.6) 11.84(-5.8) 12.03(-6.0) 2.8(99) 0.3(0) | 12.63(-4.0) 11.84(-3.9) 12.03(-4.1) 2.8(99) 0.3(0)
 KORO(41) 139 12.52(-11.8) 9.81(-17.0) 12.07(-12.1) 15.2(100) 3.2(3) | 13.54(-7.6) 10.69(-11.1) 13.23(-11.2) 14.7(100) 3.3(2)
 PROTEO(71) 140 12.25(-168.6) 6.56(-171.8) 10.46(-177.5) 16.8(100) 3.3(0) | 12.28(-199.7) 6.61(-197.4) 10.49(-206.8) 16.8(100) 3.3(0)
 igor(50) 141 11.76(-55.9) 7.68(-61.1) 10.79(-59.2) 14.6(98) 1.5(0) | 11.82(-74.1) 7.73(-77.5) 10.83(-76.7) 14.6(98) 1.5(0)
 POMYSL(62) 142 11.66(-82.0) 8.06(-78.0) 10.99(-82.6) 15.0(87) 2.4(2) | 13.77(-96.7) 9.72(-90.8) 12.68(-98.7) 15.0(91) 2.4(0)
 Advanced-ONIZUKA(42) 143 11.41(-47.6) 9.29(-42.9) 11.55(-45.6) 15.9(100) 4.0(0) | 12.25(-55.2) 10.02(-50.1) 12.29(-52.7) 14.9(100) 3.2(0)
 Scheraga(42) 144 9.93(-49.1) 7.60(-46.4) 9.95(-47.6) 14.3(100) 2.4(0) | 11.79(-49.3) 8.96(-49.9) 11.44(-49.4) 13.2(100) 2.0(0)
 Cracow.pl(49) 145 9.29(-90.9) 6.23(-90.7) 8.91(-91.3) 14.5(95) 1.5(0) | 9.68(-120.7) 6.34(-117.5) 9.13(-119.6) 15.2(100) 1.5(0)
 Floudas(27) 146 9.04(-13.7) 7.71(-14.5) 9.39(-13.1) 10.3(100) 1.6(0) | 10.08(-12.9) 8.63(-15.4) 10.33(-12.7) 10.2(100) 1.4(0)
 POEM-REFINE(24) 147 8.93(-2.2) 7.61(-0.3) 9.23(-4.0) 9.4(100) 0.8(0) | 10.30(-7.0) 8.96(-4.0) 10.39(-8.7) 8.8(100) 0.8(0)
 Schulten(15) 148 8.82(-0.1) 7.32(-1.5) 8.13(-0.4) 6.8(97) 0.5(0) | 8.82(-3.1) 7.32(-4.8) 8.13(-2.8) 6.8(97) 0.5(0)
 tlbgroupp(13) 149 8.82(-0.1) 8.03(-0.1) 8.04(-0.4) 3.2(88) 0.3(0) | 9.83(-1.2) 9.00(-1.3) 8.98(-1.6) 3.3(96) 0.4(0)
 dokhlab(23) 150 8.80(-3.4) 7.80(-4.8) 8.93(-2.9) 9.4(100) 1.1(0) | 9.64(-2.5) 8.65(-4.4) 9.67(-2.0) 9.0(100) 1.2(0)
 panther3(18) 151 8.10(-22.4) 7.01(-19.4) 7.41(-22.1) 8.2(64) 1.2(6) | 8.10(-26.4) 7.01(-23.0) 7.41(-26.0) 8.2(64) 1.2(6)
 Peter-G-Wolynes(30) 152 7.82(-20.7) 5.81(-24.6) 7.86(-22.0) 13.5(100) 2.2(0) | 9.00(-23.6) 7.09(-26.2) 9.06(-24.0) 12.4(100) 1.8(1)
 chaos(15) 153 7.78(-13.8) 6.59(-13.6) 7.01(-14.3) 11.0(100) 2.5(0) | 7.78(-17.1) 6.59(-16.8) 7.01(-17.7) 11.0(100) 2.5(0)
 EBGM(15) 154 7.25(-10.2) 5.53(-12.4) 6.57(-10.2) 9.4(100) 1.3(0) | 7.61(-12.0) 5.81(-14.6) 6.77(-12.9) 8.8(100) 1.0(0)
 Tripes-Cambridge(9) 155 7.21(-0.9) 6.28(-0.6) 6.57(-1.1) 3.2(98) 0.2(0) | 7.21(-0.2) 6.29(-0.7) 6.58(-0.2) 3.2(98) 0.2(0)
 SSU(24) 156 6.99(-3.6) 5.49(-6.9) 7.21(-3.9) 11.1(100) 1.3(0) | 9.23(-11.1) 7.83(-10.1) 9.31(-11.2) 9.1(100) 1.2(0)
 Dlakic-MSU(8) 157 6.41(-1.0) 5.73(-0.4) 5.82(-0.6) 3.1(97) 0.3(0) | 6.41(-0.2) 5.73(-1.0) 5.82(-0.7) 3.1(97) 0.3(0)
 McCormack_Okazaki(11) 158 5.67(-0.7) 4.80(-1.0) 5.35(-1.5) 8.3(90) 2.0(0) | 5.67(-3.7) 4.80(-4.0) 5.35(-4.4) 8.3(90) 2.0(0)
 Deane(21) 159 5.61(-6.6) 3.84(-10.3) 5.08(-8.9) 14.0(98) 1.3(0) | 6.06(-8.4) 4.27(-10.6) 5.50(-9.4) 13.6(98) 1.3(0)

ShakSkol-AbInitio(14) 160 5.42(-7.3) 4.85(-6.8) 5.48(-6.7) 9.2(100) 0.9(0) | 6.09(-7.6) 5.80(-8.5) 6.25(-7.9) 8.2(100) 0.9(0)
 Hirst-Nottingham(18) 161 3.80(-22.4) 2.98(-24.9) 4.28(-22.6) 12.6(100) 1.8(0) | 3.80(-29.2) 2.98(-31.2) 4.28(-29.0) 12.6(100) 1.8(0)
 GSK-CMM(4) 162 3.40(-1.5) 3.29(-1.9) 3.24(-1.6) 1.8(96) 0.1(0) | 3.40(-1.0) 3.29(-1.4) 3.24(-1.0) 1.8(96) 0.1(0)
 EATORP(18) 163 3.29(-29.1) 2.30(-31.2) 3.54(-30.1) 12.0(94) 2.4(0) | 3.50(-36.3) 2.44(-38.3) 3.80(-37.1) 12.6(100) 2.4(0)
 Bristol_Comp_Bio(4) 164 3.22(-0.7) 3.01(-0.5) 3.06(-0.5) 2.7(100) 0.2(0) | 3.22(-0.3) 3.01(-0.0) 3.07(-0.1) 2.7(100) 0.2(0)
 osgdj(11) 165 2.56(-21.3) 1.88(-21.4) 2.44(-22.1) 15.0(100) 1.1(0) | 3.03(-22.3) 2.28(-22.3) 2.98(-22.0) 13.2(100) 0.8(0)
 Dlakic-DGSA(3) 166 2.27(-1.2) 2.02(-1.9) 2.16(-1.2) 3.4(100) 1.1(0) | 2.27(-1.8) 2.02(-2.5) 2.16(-1.8) 3.4(100) 1.1(0)
 ProteinShop(9) 167 2.10(-6.2) 1.67(-6.5) 2.34(-5.7) 12.3(100) 1.3(0) | 2.38(-6.6) 1.97(-7.0) 2.64(-5.8) 11.9(100) 1.4(0)
 ricardo(3) 168 1.93(-2.0) 1.24(-1.5) 1.53(-1.7) 5.1(100) 0.3(0) | 1.96(-1.8) 1.30(-1.2) 1.59(-1.8) 4.7(100) 0.2(0)
 Doshisha-Nagoya(9) 169 1.77(-10.8) 1.57(-10.9) 2.14(-10.4) 24.9(100) 14.6(0) | 1.88(-13.1) 1.68(-13.0) 2.17(-13.3) 25.0(100) 14.6(0)
 DILL-ZAP(6) 170 1.60(-4.4) 1.59(-3.8) 1.95(-4.0) 9.7(100) 1.7(0) | 1.77(-4.6) 1.75(-4.4) 2.09(-4.5) 10.6(100) 2.3(0)
 largo(2) 171 1.52(-1.9) 1.38(-2.0) 1.46(-1.8) 2.6(100) 0.3(0) | 1.52(-1.7) 1.38(-1.7) 1.46(-1.6) 2.6(100) 0.3(0)
 hu(2) 172 1.52(-0.4) 1.43(-0.3) 1.49(-0.3) 2.4(99) 0.2(0) | 1.52(-0.0) 1.44(-0.1) 1.49(-0.2) 2.3(99) 0.2(0)
 Avbelj(7) 173 1.51(-12.1) 1.16(-11.5) 1.52(-12.3) 15.5(100) 1.0(0) | 1.63(-13.5) 1.28(-12.6) 1.62(-13.7) 14.6(100) 1.1(0)
 MerzShak(4) 174 1.49(-2.6) 1.31(-2.8) 1.59(-3.0) 9.6(100) 0.8(0) | 1.88(-3.5) 1.66(-3.2) 1.81(-3.3) 7.1(100) 0.7(0)
 Struct-Pred-Course(2) 175 1.45(-0.9) 1.15(-0.9) 1.20(-0.8) 5.9(100) 0.4(0) | 1.45(-1.3) 1.15(-1.4) 1.20(-1.3) 5.9(100) 0.4(0)
 Oka(3) 176 1.28(-2.6) 0.87(-2.9) 1.05(-2.6) 10.3(91) 1.0(0) | 1.28(-3.2) 0.87(-3.4) 1.05(-3.1) 10.3(91) 1.0(0)
 UF_GATORS(4) 177 1.25(-6.1) 0.84(-6.1) 1.03(-6.0) 16.7(100) 3.8(0) | 1.25(-6.9) 0.84(-6.9) 1.03(-6.9) 16.7(100) 3.8(0)
 MIG_FROST_FLEX(3) 178 1.09(-2.8) 1.02(-1.2) 1.08(-2.2) 11.1(76) 3.4(0) | 1.09(-3.3) 1.02(-2.0) 1.08(-2.9) 11.1(76) 3.4(0)
 AMBER-PB(1) 179 0.85(-0.3) 0.78(-0.3) 0.78(-0.4) 2.0(100) 0.8(0) | 0.88(-0.4) 0.84(-0.4) 0.79(-0.3) 1.6(100) 0.8(0)
 Pushchino(3) 180 0.84(-4.8) 0.43(-5.3) 0.62(-5.1) 15.4(91) 1.8(0) | 0.84(-5.4) 0.43(-5.8) 0.62(-5.7) 15.4(91) 1.8(0)
 CDAC(4) 181 0.66(-8.6) 0.60(-8.2) 0.92(-8.3) 15.0(100) 5.3(0) | 0.66(-10.2) 0.60(-9.6) 0.92(-9.8) 15.0(100) 5.3(0)
 ROSETTA-late(3) 182 0.65(-1.3) 0.34(-0.4) 0.55(-1.2) 19.6(100) 5.2(0) | 0.80(-3.1) 0.49(-2.7) 0.69(-3.4) 19.1(100) 4.3(0)
 SCFBio-IIID(2) 183 0.55(-2.0) 0.60(-2.1) 0.66(-2.5) 16.3(100) 10.2(0) | 0.56(-2.6) 0.62(-2.6) 0.67(-3.4) 16.9(100) 10.6(0)
 Soeding(1) 184 0.54(-1.5) 0.36(-1.3) 0.46(-1.5) 6.1(100) 0.0(0) | 0.54(-1.3) 0.36(-1.0) 0.46(-1.2) 6.1(100) 0.0(0)
 ASSEMBLY(2) 185 0.42(-0.8) 0.34(-0.5) 0.48(-0.9) 14.8(100) 2.6(0) | 0.44(-0.2) 0.39(-1.0) 0.53(-1.1) 17.8(100) 2.2(0)
 CBIS(3) 186 0.39(-5.7) 0.23(-5.3) 0.35(-5.7) 13.8(48) 3.9(0) | 0.42(-6.3) 0.23(-5.9) 0.37(-6.3) 15.0(65) 3.7(0)
 Protofold(2) 187 0.23(-6.4) 0.21(-5.6) 0.28(-6.2) 37.6(100) 28.8(0) | 0.23(-6.9) 0.21(-6.0) 0.28(-6.9) 37.6(100) 28.8(0)
 BUKKA(1) 188 0.17(-0.9) 0.16(-0.8) 0.22(-1.1) 23.2(100) 10.4(0) | 0.18(-1.0) 0.18(-0.8) 0.23(-1.3) 21.7(100) 8.4(0)
 INFSRUCT(1) 189 0.14(-3.4) 0.11(-3.3) 0.16(-3.4) 14.1(100) 0.4(0) | 0.14(-3.7) 0.11(-3.5) 0.16(-3.6) 14.1(100) 0.4(0)

Bibliography

- Microarray Suite User Guide.* Affymetrix, version 5 edition, 2001. <http://www.affymetrix.com/support/technical/manuals.affx>.
- S. V. Allander, N. N. Nupponen, M. Ringner, G. Hostetter, G. W. Maher, N. Goldberger, Y. Chen, J. Carpten, A. G. Elkahloun, and P. S. Meltzer. Gastrointestinal stromal tumors with KIT mutations exhibit a remarkably homogeneous gene expression profile. *Cancer Research*, pages 8624–8628, 2001. Download: http://research.nhgri.nih.gov/microarray/gist_data.txt.
- S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. A basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.
- S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.
- T. L. Bailey and W. N. Grundy. Classifying proteins by family using the product of correlated p-values. In S. Istrail, P. Pevzner, and M. Waterman, editors, *Proc. 3rd Ann. Int. Conf. Computational Molecular Biology*, pages 10–14, 1999.
- W. Bains and G. Smith. A novel method for nucleic acid sequence determination. *Journal of Theoretical Biology*, 135:303–307, 1988.
- P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, 15:937–946, 1999.
- P. Baldi, Y. Chauvin, T. Hunkapiller, and M. A. McClure. Hidden Markov models of biological primary sequence information. *Proceedings of the National Academy of Sciences of the United States of America*, 91(3):1059–1063, 1994.
- A. Bateman, L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy. The pfam protein families database. *Nucleic Acids Research*, 32:D138–141, 2004.
- B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.
- P. E. Bourne and H. Weissig. *Structural Bioinformatics*. Wiley-Liss, Hoboken, New Jersey, USA, 2003.
- S. E. Brenner and M. Levitt. Expectations from structural genomics. *Protein Science*, 9:197–200, 2000.
- M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. Sugnet, T. S. Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1):262–267, 2000.
- S. H. Bryant and C. E. Lawrence. An empirical energy function for threading protein-sequence through the folding motif. *Proteins*, 16:92–112, 1993.
- J. Cai, A. Dayanik, N. Hasan, T. Terauchi, and H. Yu. Supervised machine learning algorithms for classification of cancer tissue types using microarray gene expression data. Technical report, Columbia University, 2001. <http://www.cpmc.columbia.edu/homepages/jict001/cs4995/project1.htm>.
- A. Cerboni, P. Frasconi, and G. Pollastri. Learning protein secondary structure from sequential and relational data. *Neural Networks*, 18(8):1029–1039, 2005.
- J. Cheng and P. Baldi. A machine learning information retrieval approach to protein fold recognition. *Bioinformatics*, 22:1456–1463, 2006.
- P. Y. Chou and G. D. Fasman. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol.*, 47:45–148, 1978.
- W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836, 1979.
- W. S. Cleveland and S. J. Devlin. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83:596–610, 1988.

- J. A. Cuff and G. J. Barton. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, 40:502–511, 2000.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39(1):1–22, 1977.
- C. Ding and I. Dubchak. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, 17(4):349–358, 2001.
- A. A. Dombkowski and G. M. Crippen. Disulfid recognition in an optimized threading potential. *Protein Engineering*, 13(10):679–689, 2000.
- S. Drescher. Sekundärstrukturvorhersage bei proteinen mit support-vektor-maschinen. Master's thesis, Technical University Berlin, 2005. Diploma Thesis in German.
- R. Drmanac, I. Labat, I. Brukner, and R. Crkvenjakov. Sequencing of megabase plus DNA by hybridization: theory of the method. *Genomics*, 4:114–128, 1989.
- S. R. Eddy. Profile hidden markov models. *Bioinformatics*, 14:755–763, 1998.
- S. R. Eddy. What is a hidden markov model? *Nature Biotechnology*, 22:1315–1316, 2004.
- R. P. Ekins and F. W. Chu. Multianalyte microspot immunoassay–microanalytical “compact disk” of the future. *Clinical Chemistry*, 37(1111):1955–1967, 1991.
- W. R. Fodor, J. L. Read, M. C. Pirrung, L. Stryer, A. T. Lu, and D. Solas. Light-directed, spacially addressable parallel chemical synthesis. *Science*, 251:767–773, 1991.
- D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. *Proteins*, 23:566–579, 1995.
- T. S. Furey, N. Duffy, N. Cristianini, D. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906–914, 2000.
- J. Garnier, J.-F. Gibrat, and B. Robson. GOR method for predicting protein secondary structure from amino acid sequence. *Meth. Enzymol.*, 266:540–553, 1996.
- J. Garnier, D. J. Osguthorp, and B. Robson. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.*, 120:97–120, 1978.
- D. Gerhold, T. Rushmore, and C. T. Caskey. DNA chips: promising tools have become powerful tools. *Trends in Biochemical Science*, 24(5):168–173, 1999.
- J.-F. Gibrat, J. Garnier, and B. Robson. Further developments of protein secondary structure prediction using information theory. New parameters and consideration of residue pairs. *J. Mol. Biol.*, 198:425–443, 1987.
- A. Godzik, A. Kolinski, and J. Skolnick. Topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol.*, 227:227–238, 1992.
- S. Govindarajan, R. Recabarren, and R. A. Goldstein. Estimating the total number of protein folds. *Proteins*, 35:408–414, 1999.
- W. N. Grundy. Family-based homology detection via pairwise sequence comparison. In *Proc. 2nd Ann. Int. Conf. Computational Molecular Biology*, pages 94–100, 1998.
- A. Hartemink, D. Gifford, T. Jaakkola, and R. Young. Maximum likelihood estimation of optimal scaling factors for expression array normalization. In M. Bittner, Y. Chen, A. Dorsel, and E. Dougherty, editors, *Proceedings of SPIE International Symposium on Biomedical Optics (BiOS01)*, volume 4266 of *Microarrays: Optical Technologies and Informatics*, pages 132–140, 2001.
- S. Hochreiter, D.-A. Clevert, and K. Obermayer. A new summarization method for affymetrix probe level data. *Bioinformatics*, 2006.
- M. W. Hofmann, K. Weise, J. Ollesch, P. Agrawal, H. Stalz, W. Stelzer, F. Hulsbergen, H. de Groot, K. Gerwert, J. Reed, and D. Langosch. De novo design of conformationally flexible transmembrane peptides driving membrane fusion. *Proc. Natl. Acad. Sci.*, 101(41):14776–14781, 2004.
- L. Holm and C. Sander. Protein folds and families: sequence and structure alignments. *Nucleic Acids Res.*, 27(1):244–247, 1999.
- Y. Hou, W. Hsu, M. L. Lee, and C. Bystroff. Remote homolog detection using local sequence-structure correlations. *Proteins: Structure, Function and Bioinformatics*, 57:518–530, 2004.
- S. Hua and Z. Sun. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *Journal of Molecular Biology*, 308:397–407, 2001a.
- S. Hua and Z. Sun. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, 17(8):721–728, 2001b.
- E. Hubbell, W.-M. Liu, and R. Mui. Robust estimators for expression analysis. *Bioinformatics*, 18(12):1585–1592, 2002.

- T. Huber and A. E. Torda. Protein sequence threading, the alignment problem and a two step strategy. *J. Comput. Chem.*, 20:1455–1467, 1999.
- R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31(4):1–8, 2003a.
- R. A. Irizarry, B. Hobbs, F. Collin, Y. Beazer-Barclay, K. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003b.
- T. Jaakkola, M. Diekhans, and D. Haussler. Using the fisher kernel method to detect remote protein homologies. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 149–158, 1999.
- T. Jaakkola, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 7 (1-2):95–114, 2000.
- J. Jäger, R. Sengupta, and W. L. Ruzzo. Improved gene selection for classification of microarrays. In *Biocomputing - Proceedings of the 2003 Pacific Symposium*, pages 53–64, 2003.
- D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292:195–202, 1999.
- D. T. Jones, W. R. Taylor, and J. M. Thornton. A new approach to protein fold recognition. *Nature*, 358:86–89, 1992.
- W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers*, 22: 2577–2637, 1983a.
- W. Kabsch and C. Sander. Segment83. unpublished, 1983b.
- K. Karplus, C. Barrett, and R. Hughey. Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, 14(10):846–856, 1998.
- M. K. Kerr, M. Martin, and G. A. Churchill. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7:819–837, 2000.
- A. Krogh, M. Brown, I. Mian, K. Sjölander, and D. Haussler. Hidden Markov models in computational biology: Applications to protein modeling. *Journal of Molecular Biology*, 235:1501–1531, 1994.
- R. Kuang, E. Ie, K. Wang, K. Wang, M. Siddiqi, Y. Freund, and C. Leslie. Profile-based string kernels for remote homology detection and motif extraction. *Journal of Bioinformatics and Computational Biology*, 3(3):527–550, 2005.
- R. H. Lathrop. An anytime local-to-global optimization algorithm for protein threading in $o(m^2 n^2)$ space. *J. Comput. Biol.*, 6:405–418, 1999.
- R. H. Lathrop and T. F. Smith. Global optimum protein threading with gapped alignment and empirical pair score functions. *J. Mol. Biol.*, 255:641–665, 1996.
- C. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble. Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4): 467–476, 2004a.
- C. Leslie, R. Kuang, and E. Eskin. Inexact matching string kernels for protein classification. In B. Schölkopf, K. Tsuda, and J.-P. Vert, editors, *Kernel Methods in Computational Biology*, pages 95–111. MIT Press, 2004b.
- J. M. Levin, S. Pasarell, P. Argos, and J. Garnier. Quantification of secondary structure prediction improvement using multiple alignment. *Prot. Engin.*, 6:849–854, 1993.
- C. Li and W. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences*, 98(1):31–36, 2001.
- L. Liao and W. S. Noble. Combining pairwise sequence similarity support vector machines for remote protein homology detection. In *Proceedings of the Sixth International Conference on Computational Molecular Biology*, pages 225–232, 2002.
- V. I. Lim. Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. *J. Mol. Biol.*, 88:857–872, 1974.
- T. Lingner and P. Meinicke. Remote homology detection based on oligomer distances. *Bioinformatics*, 22(18):2224–2236, 2006.
- D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14(13):1675–1680, 1996.
- Y. Lysov, V. Florent'ev, A. Khorlin, K. Khrapko, V. Shik, and A. Mirzabekov. DNA sequencing by hybridization with oligonucleotides. *Doklady Academy Nauk USSR*, 303:1508–1511, 1988.
- T. Madej, J. F. Gibrat, and S. H. Bryant. Threading a database of protein cores. *Proteins*, 23:356–369, 1995.

- M. Madera and J. Gough. A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res.*, 30(19):4321–4328, 2002.
- 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, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 2000.
- M. Ouali and R. D. King. Cascaded multiple classifiers for secondary structure prediction. *Protein Science*, 9:1162–1176, 2000.
- J. Park, K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard, and C. Chothia. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *Journal of Molecular Biology*, 284(4):1201–1210, 1998.
- J. Park, S. A. Teichmann, T. Hubbard, and C. Chothia. Intermediate sequences increase the detection of distant sequence homologies. *J. Mol. Biol.*, 273:349–354, 1997.
- W. Pearson and D. Lipman. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.*, 85:2444–2448, 1988.
- G. Pollastri and A. McLysaght. Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics*, 21(8):1719–1720, 2005.
- S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. H. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander, and T. R. Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–442, 2002.
- D. Przybylski and B. Rost. Psi-blast for structure prediction: plug-in and win. Columbia University, 2001.
- O. B. Ptitsyn and A. V. Finkelstein. Theory of protein secondary structure and algorithm of its prediction. *Biopolymers*, 22:15–25, 1983.
- N. Qian and T. J. Sejnowski. Predicting the secondary structure of globular proteins using neural networks. *J. Mol. Biol.*, 202:865–884, 1988.
- H. Rangwala and G. Karypis. Profile based direct kernels for remote homology detection and fold recognition. *Bioinformatics*, 21(23):4239–4247, 2005.
- F.M. Richards and C. E. Kundrot. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins*, 3:71–84, 1988.
- B. Rost. 1D structure prediction for Chameleon (IgG binding domain of protein G). EMBL Heidelberg, Germany, WWW document (<http://www.embl-heidelberg.de/~rost/Res/96C-PredChameleon.html>), 1996a.
- B. Rost. PHD: predicting one-dimensional protein structure by profile based neural networks. *Meth. Enzymol.*, 266:525–539, 1996b.
- B. Rost. Prediction in 1D: secondary structure, membrane helices, and accessibility. In P. E. Bourne and H. Weissig, editors, *Structural Bioinformatics*, pages 559–587. Wiley-Liss, Hoboken, New Jersey, USA, 2003a.
- B. Rost. Rising accuracy of protein secondary structure prediction. In D. Chasman, editor, *Protein structure determination, analysis, and modeling for drug discovery*, pages 207–249. New York: Dekker, 2003b.
- B. Rost and C. Sander. Prediction of protein secondary structure at better than 70 % accuracy. *Journal of Molecular Biology*, 232:584–599, 1993.
- B. Rost and C. Sander. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, 19:55–72, 1994.
- E. E. Schadt, C. Li, B. Ellis, and W. H. Wong. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry*, S37:120–125, 2001.
- M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences USA*, 93:10614–10619, 1995.
- B. Schölkopf and A. J. Smola. *Learning with kernels – Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, 2002.
- C. M. Schubert. Microarray to be used as routine clinical screen. *Nature Medicine*, 9(1):9, 2003.
- J. Schuchhardt and D. Beule. Normalization strategies for cDNA microarrays. *Nucleic Acids Research*, 28(10):e47, 2000.
- R. Service. A dearth of new folds. *Science*, 307(5715):1555, 2005.
- M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, R. C. T. Aguiar J. L. Kutok, M. Gaasenbeek, M. Angelo, M. Reich, T. S. Ray G. S. Pinkus, M. A. Koval, K. W. Last, A. Norton, J. Mesirov T. A. Lister, D. S. Neuberg, E. S. Lander, J. C. Aster, and T. R. Golub. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8(1):68–74, 2002.
- M. J. Sippl. Boltzmann's principle, knowledge-based mean fields and protein folding. *J. Comput. Aided Mol. Des.*, 7:473–501, 1993.

- H. Sklenar, C. Etchebest, and R. Lavery. Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins*, 6:46–60, 1989.
- T. Smith and M. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147:195–197, 1981.
- V. V. Solovyev and A. A. Salamov. Predicting α -helix and β -strand segments of globular proteins. *CABIOS*, 10:661–669, 1994.
- E. Southern. United Kingdom patent application GB8810400, 1988.
- E. M. Southern. Detection of specific sequences among dna fragments separated by gel electrophoresis. *J. of Molecular Biology*, 98:503–517, 1975.
- C. Tarnas and R. Hughey. Reduced space hidden Markov model training. *Bioinformatics*, 14(5):401–406, 1998.
- G. Tseng, M. Oh, L. Rohlin, J. Liao, and W. Wong. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research*, 29:2549–2557, 2001.
- L. J. van't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.
- V. N. Vapnik. *The Nature of Statistical Learning Theory*. Statistics for Engineering and Information Science. Springer Verlag, New York, 2nd edition, 2000.
- J.-P. Vert, H. Saigo, and T. Akutsu. Local alignment kernels for biological sequences. In B. Schölkopf, K. Tsuda, and J.-P. Vert, editors, *Kernel Methods in Computational Biology*, pages 131–154. MIT Press, 2004.
- D. G. Wang, J.-B. Fan, and C.-J. Siao. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280:1077–1082, 1998.
- Z. X. Wang. A re-estimation for the total numbers of protein folds and superfamilies. *Protein Engineering*, 11:621–626, 1998.
- G. Williams and P. Doherty. Inter-residue distances derived from fold contact propensities correlate with evolutionary substitution costs. *BMC Bioinformatics*, 5:153–157, 2004.
- M. Wilmanns and D. Eisenberg. Inverse protein folding by the residue pair preference profile method: Estimating the correctness of alignments of structurally compatible sequences. *Protein Eng.*, 8:627–639, 1995.
- J. Xu and M. Li. Assessment of RAPTOR's linear programming approach in CAFASP3. *Proteins*, 53:579–584, 2003.
- Y. Xu and E. C. Uberbache. A polynomial-time algorithm for a class of protein threading problems. *Comput. Appl. Biosci.*, 12:511–517, 1996.
- Y. Xu and D. Xu. Protein threading using PROSPECT: Design and evaluation. *Proteins: Structure, Function, and Genetics*, 40:343–354, 2000.
- Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed. Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4):e15, 2002.
- A. Zemla, C. Venclovas, K. Fidelis, and B. Rost. A modified definition of SOV, a segment-based measure for protein secondary structure prediction assessment. *Proteins*, 34:220–223, 1999.
- C. Zhang and C. DeLisi. Estimating the number of protein folds. *J. Mol. Biol.*, 284:1301–1305, 1998.

