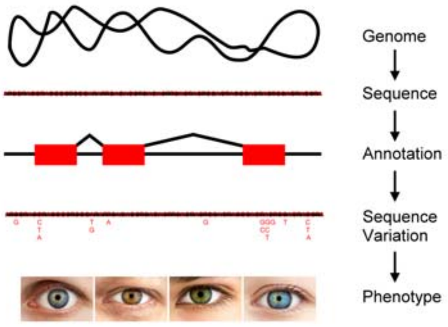
What is Genome Informatics?



Genome

Sequence

Annotation

Sequence Variation

Phenotype

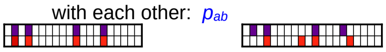
Alignment

- Scoring – match, mismatch, gaps
- Alignment algorithm

Scoring using log likelihoods

From a large set of high quality *ungapped* protein sequence alignments, for pairs of aligned sequences:

- measure background frequencies of residues *a* and *b*: q_a, q_b.
- measure frequency with which *a* and *b* are found aligned with each other: p_{ab}



Log likelihood: score(*a*,*b*) = log ($p_{ab} / q_a q_b$)

score 0 if aligned as often as expected
score positive if preferentially aligned
score negative if alignment is avoided

Rounded to nearest integer for computational efficiency

Where did the BLOSUM62 alignment score matrix come from?
Eddy SR.
Nat Biotechnol. 2004 Aug;22(8):1035-6. Review. PMID: 15286655

Gap penalties

Linear:

total penalty = -d * g

where *g* is length of gap
and d is the per-residue
gap penalty

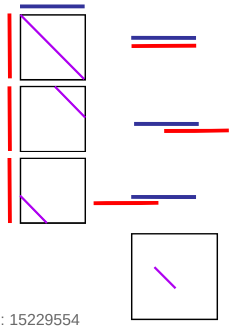
Affine:

total penalty = -d - e(g-1)

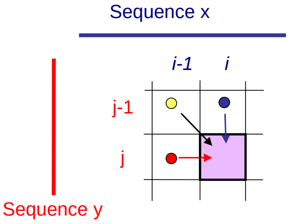
where *e* is the gap extension
penalty and *e < d*

Dynamic Programming

Given a scoring scheme for aligning residues and gaps, DP algorithm guarantees the best (sub)sequence alignment



What is dynamic programming? Eddy SR.
Nat Biotechnol. 2004 Jul;22(7):909-10. PMID: 15229554



F(*i,j*) is the score of the best alignment of subsequence $X_{1...i}$ and subsequence $y_{1...j}$

Recursive: F(*i,j*) depends on previously evaluated elements of F

$$F(i,j) = \max \begin{cases} F(i\text{-}1, j\text{-}1) + \text{score}(x_i, y_j) \\ F(i, j\text{-}1) - d \\ F(i\text{-}1, j) - d \end{cases}$$

where d = gap penalty
score() = score from aligning
two residues

For each cell, a *traceback pointer* records from which parent the best score was inherited

Short read sequence aligners: 1

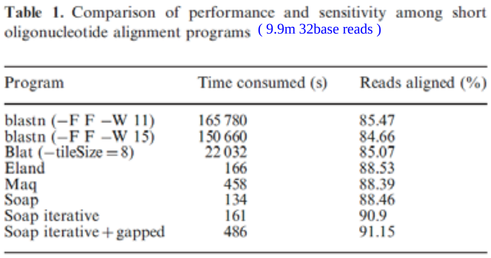


Table 1. Comparison of performance and sensitivity among short oligonucleotide alignment programs (9.9m 32base reads)

Program	Time consumed (s)	Reads aligned (%)
blastn (−F F −W 11)	165 780	85.47
blastn (−F F −W 15)	150 660	84.66
Blat (−tileSize = 8)	22 032	85.07
Eland	166	88.53
Maq	458	88.39
Soap	134	88.46
Soap iterative	161	90.9
Soap iterative + gapped	486	91.15

SOAP: short oligonucleotide alignment program.
Li R, Li Y, Kristiansen K, Wang J. Bioinformatics. 2008 Mar 1;24(5):713-4. PMID: 18227114

bowtie 200-600x faster than Soap

Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.
Langmead B, Trapnell C, Pop M, Salzberg SL. Genome Biol. 2009;10(3):R25. PMID: 19261174

Multiple sequence alignment

- Heuristic vs. global optimisation

- DP – v. v. slow
- Progressive alignment construction – e.g. Clustal family
- Iterative methods – e.g. MUSCLE
- Consensus methods

- HMMs e.g. HMMer
- Motif finding e.g. MEME – see Regulation lectures

- Not practical on large scale

From raw reads to annotation…

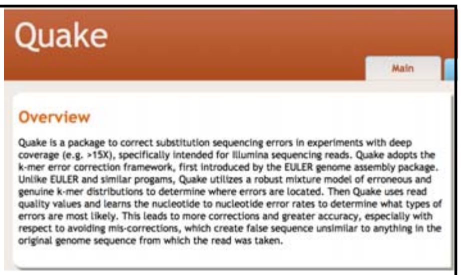
QC / Error Correction

- Removal of vector/adapter
- Quality trimming
- Correction of reads

Error Correction

- Improves assembly quality
- Reduces memory requirements (25% reduction in earlier example, >50% experimentally)

- Quality trimming
- Correction of reads



Quake

Overview

Quake is a package to correct substitution sequencing errors in experiments with deep coverage (e.g. >15X), specifically intended for Illumina sequencing reads. Quake adopts the k-mer error correction framework, first introduced by the EULER genome assembly package. Unlike EULER and similar progams, Quake utilizes a robust mixture model of erroneous and genuine k-mer distributions to determine where errors are located. Then Quake uses read quality values and learns the nucleotide to nucleotide error rates to determine what types of errors are most likely. This leads to more corrections and greater accuracy, especially with respect to avoiding mis-corrections, which create false sequence unsimilar to anything in the original genome sequence from which the read was taken.

http://www.cbcb.umd.edu/software/jellyfish/
Guillaume Marcais and Carl Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics (2011) 27(6): 764-770

http://www.cbcb.umd.edu/software/quake/
Kelley DR, Schatz MC, Salzberg SL. Quake: quality-aware detection and correction of sequencing errors. Genome Biology 11:R116 2010.

EMBnet.journal
Bioinformatics in Action

Cutadapt removes adapter sequences from high-throughput sequencing reads

Marcel Martin

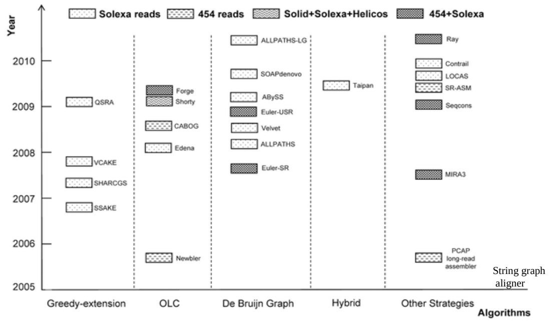
http://code.google.com/p/cutadapt/

http://www.cbcb.umd.edu/software/jellyfish/ (Jellyfish)

- Marcais G. *et al*. - A fast, lock-free approach for efficient parallel counting of occurrences of k-mers - Bioinformatics 2011

http://www.cbcb.umd.edu/software/quake/ (QUAKE)

- Kelley DR. *et al*. - Quake: quality-aware detection and correction of sequencing errors - Genome Biology 2010.



Zhang *et al*, PloS One 2011

Annotation

- Repeat Finding
- Gene Finding
- Regulatory regions

Existing Repeat Databases

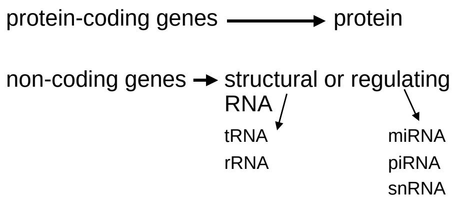
- RepBase
  - All types of repeats; actual sequence

  http://www.girinst.org/repbase/index.html

- Dfam
  - Alignments, HMMs and match lists of repeats

  http://dfam.janelia.org/

Types of gene



protein-coding genes → protein

non-coding genes → structural or regulating RNA

tRNA
rRNA

miRNA
piRNA
snRNA

(pseudo genes)

How to find human genes?

Via human cDNA or EST sequences
Via vertebrate cDNA or EST sequences

Finding similarity in genome to known proteins

Ab initio - using statistical gene finders.

Main genome annotation sites:
http://www.ensembl.org
http://genome.ucsc.edu

Integration of various evidence

- manually
- using statistics and computational methods
  - simple counting
  - hidden Markov models
  - Bayesian statistics
  - neural networks

- Always best to use many different finders and combine. Some frameworks try to keep this process as user-friendly as possible, e.g. Maker

Types of gene

protein-coding genes → protein

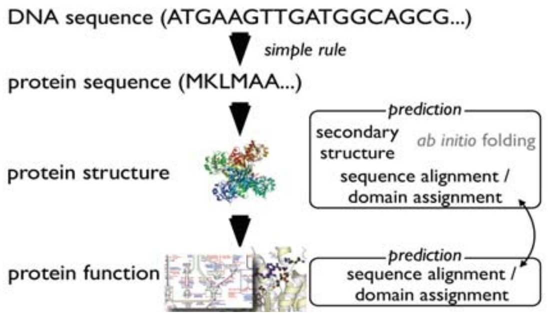
non-coding genes → structural or regulating RNA

tRNA
rRNA

miRNA
piRNA
snRNA

(pseudo genes)

From sequence to function


DNA sequence (ATGAAGTTGATGGCAGCG...)
simple rule
protein sequence (MKLMAA...)
prediction
secondary structure
ab initio folding
sequence alignment / domain assignment
protein structure
prediction
sequence alignment / domain assignment
protein function

Ab initio prediction of secondary structure from primary structure

- learning directly from X-ray structures
- consideration of environment
- neural network-based training

e.g. Dor *et al.* - Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training – Proteins 2007

Prediction by alignment

- Primary sequence similarity >30% can be assumed to have the same 3D structure (but not necessarily function - beware of details!)
- Any available structural or functional data on orthologues (the “same” protein in a different organism) can be of great relevance.

Functional annotation

- For enzymes: EC number
- 6 groupings – Oxidoreductases, Transferases, Hydrolases, Lyases, Isomerases, Ligases

Functional annotation

For other proteins: Gene Ontology, Reactome, KEGG

- GO is the *de facto* standard used by all major model organism databases
- Founded in 2000 by the GO Consortium lead by Michael Ashburner
- Database curators read the scientific literature and assign functional classifications along with an evidence code to proteins
- These annotations follow a controlled vocabulary that is organized into an ontology.

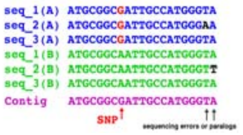
Sequence Variation

- Indels – Insertions or deletions
- CNV – Copy Number Variation
- SNPs – Single Nucleotide Polymorphisms

Slides adapted from Irene Paratheodorou

What are SNPs?

- DNA sequence variations occurring when a single nucleotide in the genome is altered
- Frequency of 1% or more
- Occur in both coding and non-coding regions
- Occur every 100-300 bases
- ~15 million in human genome



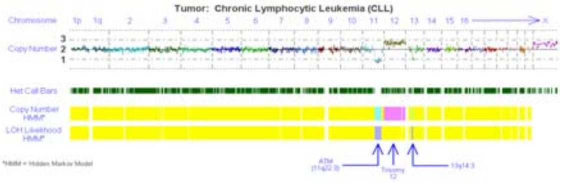
```
seq_1(A) ATGCGGCGATTGCCATGGGTA
seq_2(A) ATGCGGCGATTGCCATGGGAA
seq_3(A) ATGCGGCGATTGCCATGGGTA
seq_1(B) ATGCGGCAATTGCCATGGGTA
seq_2(B) ATGCGGCAATTGCCATGGGTT
seq_3(B) ATGCGGCAATTGCCATGGGTA
Contig   ATGCGGCGATTGCCATGGGTA
```

Figure from Alexander Kozik, Compositae Genome Project, UCDA

SNP resources

- dbSNP
  - Central repository for SNPs
  - Initial SNPs identified with PolyBayes
  - dbSNP build 138, human genome build 37.5 – 233M submissions at 63M loci
  - High false positive rate?
- HapMap
  - Database of haplotypes and ‘tag’ SNPs which identify them
  - Samples from 270 people from Nigeria, Japan, China, USA (of North and West European decent)

Virtual karyotype - SNP arrays


Tumor: Chronic Lymphocytic Leukemia (CLL)
Chromosome
1p
1q
X
Copy Number
Het Call Bars
Copy Number HMM*
LOH Likelihood HMM*
*HMM = Hidden Markov Model
ATM (11q22.3)
Trisomy 12
13q14.3