Structural Biochemistry and Bioinformatics 2005

http://www.sbc.su.se/~arne/kurser/struktbiokemi_2005/

Teachers:

- Arne Elofsson
- Erik Lindahl
- Gunnar von Heijne
- Erik Sandelin
- Bengt Sennblad
- Olof Emanuelsson

Assistants:

- Björn Wallner
- Erik Granseth

Book:

Bioinformatics (Advanced Texts) C.A. Orengo, J.M. Thornton, D.T. Jones
Publisher: BIOS Scientific Publishers;
ISBN: 1859960545

Schedule

- Lecturs
  - Mornings
- Assignments
  - Afternoons
- Exam
  - Dec 22
- Project

Lectures (10-12)

- Genomes and genes* (AE)
- Sequence comparison methods] (EL)
- Amino acid comparison(EL)
- Multiple seqence alignments, HMMs PSI-BLAST (EL)
- Molecular evolution and Phylogeny (BS)
- Neural Networks and other machine learning methods* (OE)
- Membrane proteins: structure, prediction* (GvH)
- Protein Secondary Structure predictions etc*(AE=
- Fold Recognition* (AE)
- Comparative Modeling* (EL)
- Protein structure classification and comparison (ES)
- Databases in bioinformatics (AE)*
- From Protein structure to protein function (AE)
- Structured based genome annotations* (AE)
- Microarrays and Proteomics (AE)
- Protein-protein interactions (AE)

Goals

- Knowledge of bioinformatics tools
- Understanding of how tools work
  - Alignments
  - Structure predictions
  - Evolutionary tools
  - Analysis tools for postgenomic data
- Understanding limitations
- Knowledge of available data
  - Sequences, proteins, genomes
  - structures
  - Expression analysis

Background knowledge

- Sequences
  - DNA sequences and protein sequences
  - Translation and transcription
- Structure
  - Secondary structures
  - What is a protein domain ?
- Biology
  - What is the difference between:
  - Bacteria and yeast
  - Yeast and Humans

What is bioinformatics

Dealing with “bio-data”

What is “bio-data” ?

- genes/genomes
- Biomolecular structure
- Expression data
- Clinical data
- Biomedical literature
- Etc etc etc

From gene to folded protein

- Transcription (DNA-RNA)
  - Expression regulation
  - promotor region identification
  - Gene identification
  - Exon identification
  - Intron splicing (in nucleus)
  - Intron splicing (at spliceosomes)
- Translation (mRNA-Protein)
  - Translation at ribosomes
- Folding
  - Folding with/without chaperones
  - Post translational modification
  - Folded (functional) proteins
- Ubuiquitination and degradation



Information

- What is information in biology
  - The information in a genome is huge ($4^{3000000000}$)
  - Genomes can be copied **and read**
- **Storing and analyzing information**
  - The human genome is “small” for a computer today (3 gb vs >100gb)
  - The complexity is enorm ($3*10^9*3*10^9$)

Molecular evolution

Random mutations (genotypes)

Selection (phenotype)

Differential reproduction

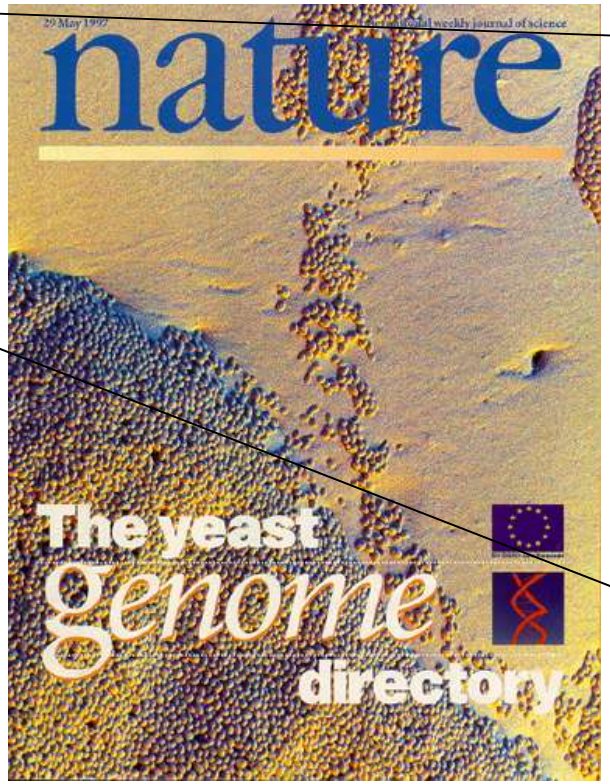
- Causes of genetic variation
  - Mutations
  - Duplications

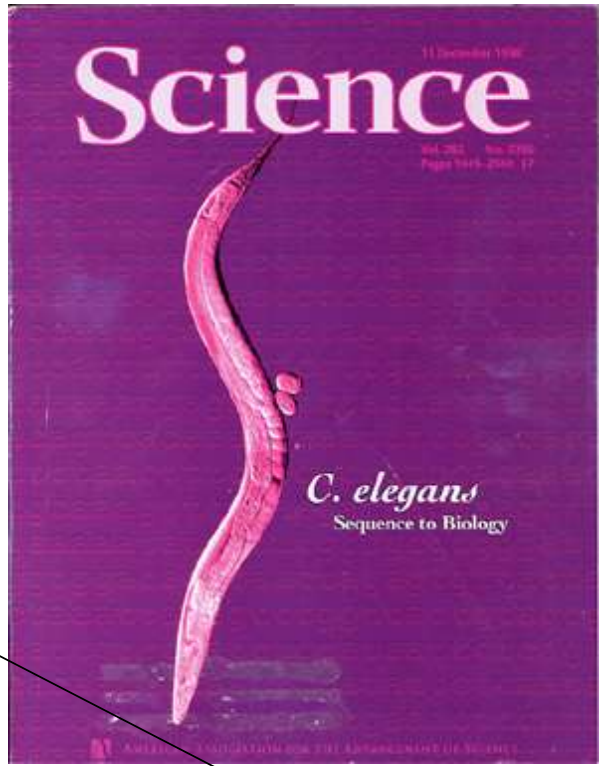
- Mutation rates:
  - Viruses: 10^{-3} to 10^{-5}
  - Bacteria: 10^{-10} to 10^{-11}
  - Humans: 10^{-10} to 10^{-11}
  - Mutation hotspots
    - Tandem repeats
    - CpG dinucleotides
  - Mutation types
    - Point mutations
    - Insertions and deletions

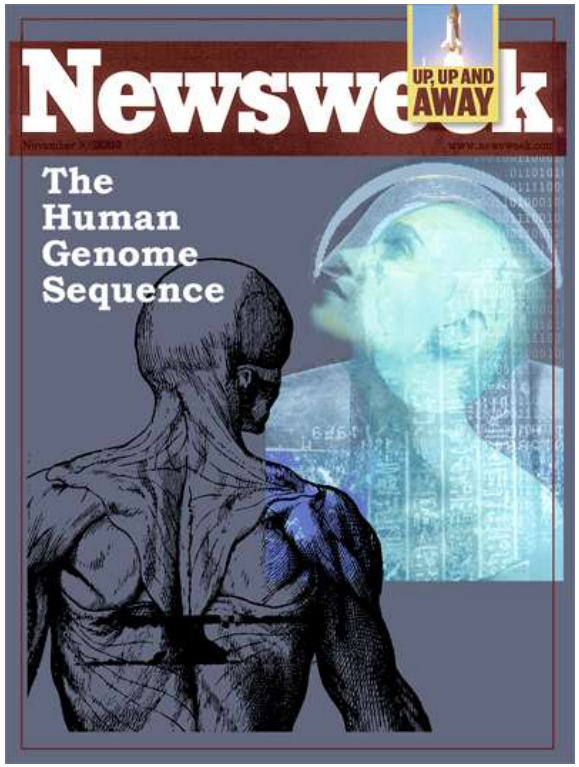
Genes, Genomes and finding

- Are genomes the sum of genes ?
  - No more..
- Sequencing a genome
  - Mapping
  - History
  - Small and Big
  - Price
  - When is complete complete ?
- Uses for genomes
- Gene Finding


AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE
SCIENCE
28 JULY 1995
VOL. 269 • PAGES 449 604
$7 00


29 May 1997
nature
The yeast
genome
directory


Science
C. elegans
Sequence to Biology


Newsweek
UP, UP AND AWAY
The Human Genome Sequence


Newsweek
THE RACE TO DECODE THE
HUMAN BODY
GENOME

What is a genome ?

- The sequence of nucleotides in an organism.
- But:
  - Vectors, Mitochondrias and Chloroplasts also have DNA
  - Differences between individuals
- Is it the sum of all genes ?

What is a gene ?

Possible answers:

a) Difficult question!

b) Region corresponding to expression of a protein

c) B + expression of RNA

d) C + control elements

But also genetic traits and diseases

Why study genomes ?

- *Complete* sets of genes, proteins and RNA:
  - Charactarize gene families
  - Describe biological systems
  - Plan experiments
  - Study expression
- Knowledge of what is missing
- Genomic DNA regulation of gene expression
- Genome comparison: Study evolution, pathogenicity....

Landmarks in genome sequencing

- *1978 SV40 5.2 kb*
- *1979 HPV 4.9 kb*
- *1981 mitochondrial DNA 16.6 kb*
- *1982 phage lambda 48 kb*
- *1984 EBV 172 kb*
- *1992 S. cerevisiae chr. III 315 kb*
- *1994 S. cerevisiae chr. XI 666 kb*
- *1995 H. influenzae genome 1.83 Mb*
- *1996 S. cerevisiae genome 12 Mb*
- *1997 E. coli genome 4.6 Mb*
- *1998 C. elegans genome 97 Mb*
- *2000 D. melanogaster genome 120 Mb*
- *2001 Human genome 3300 Mb*

DNA sequencing

- Sequencing techniques
  - Sanger, 1977. Chain termination method (Nobel prize in chemistry 1980)
  - Leroy Hood, 1985: Fluorescent sequencing
- Assembly:
  - Length of DNA limited to 1-2kb
  - Shotgun approach: break up DNA into small fragments. Use computers to reassemble whole sequences via overlaps.
  - Venter, 1995: Whole genome shotgun of *H. Influenzae*

Sequencing costs

- 1990 $10 per base ($30 000 000 000 for Human genome 1x)
- 2005 $0.05 per base (150 000 000)
- 2008 $10-100k per human genome
- 2010 $1000 per human genome

Genetic and physical maps of genomes

Genetic maps – based on recombination frequencies

- Non-random segregation of traits/genes or DNA markers (linkage)
- Need large family tree, or experimental organism
- Gives order of genes, but no actual distances

Physical maps – direct measurements, real distances

- **Restriction mapping**
- Fluorescent *in-situ hybridisation (FISH)*
- *Radiation-hybrid / STS mapping*

STS mapping

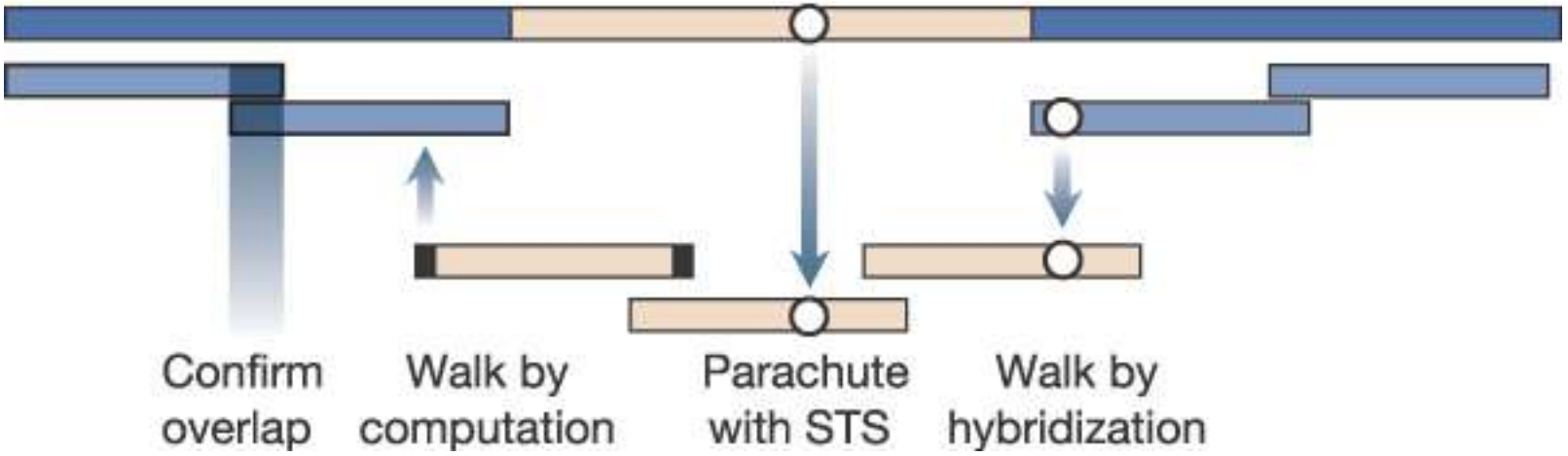
- STS = sequence tagged site = 100-500bp unique *sequenced DNA (ESTs, sequenced markers, database sequences)*
- *Probe genomic DNA fragments with PCR for STSs*
- *Close STSs will be detected on same fragments more often than distant STSs*
- *Compute map in similar way to genetic map*
- *Genomic fragments come from:*
  - *Radiation treated chromosomes, rescued by fusion to rodent cell lines*
  - *Random cloned fragments*

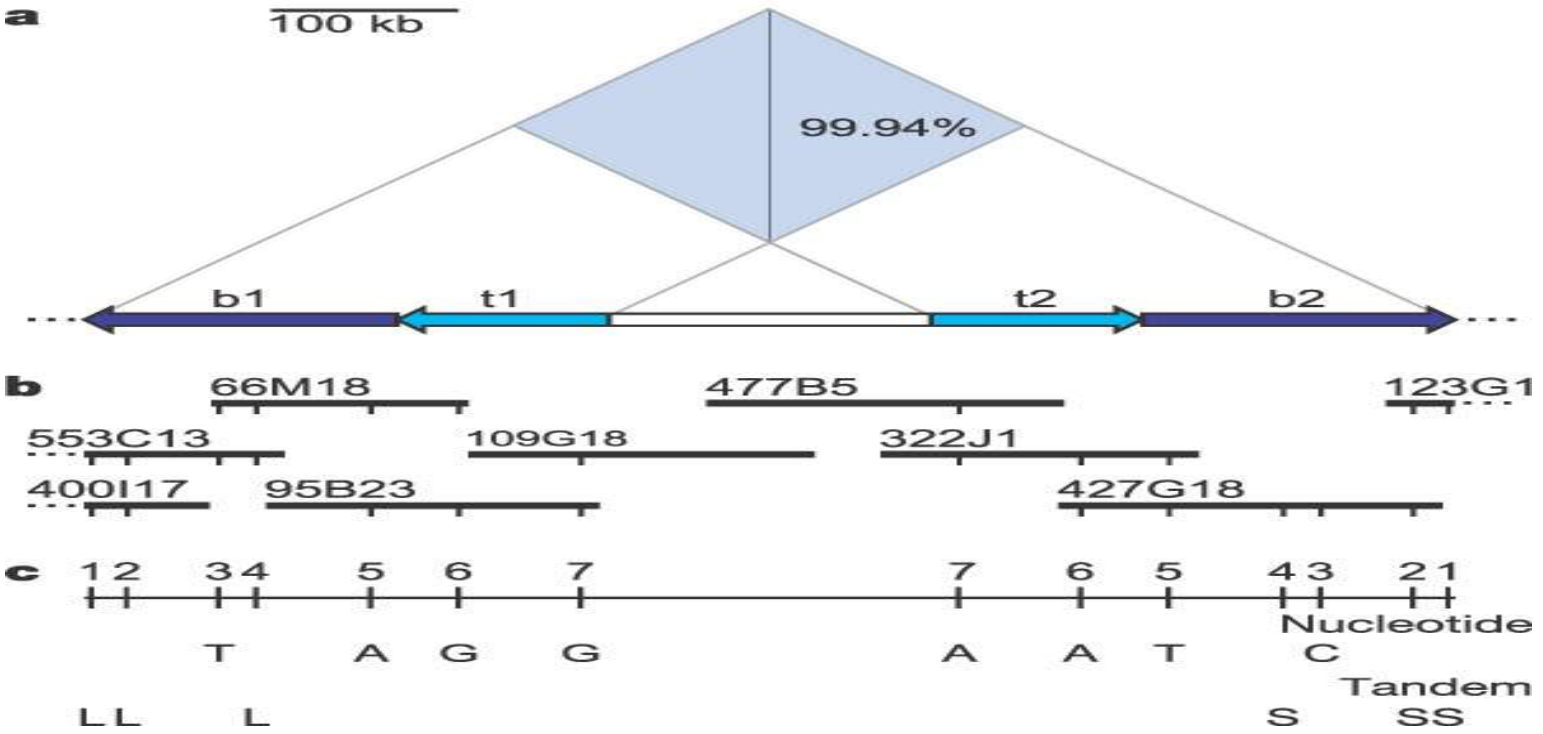
Polymorphic markers and human disease

- Some DNA markers differ between individuals
  - In length RFLPs and SSLPs
  - In content SNPs
  - In copy number (recent findings)
- Can be used to make genetic and physical maps
- Linkage between (co-occurrence of) disease and polymorphic markers can help locate disease genes

Clone-based sequencing

* First generate stable clones of rather large segments of DNA from the organism under study. The size of the segments depend on the technique used. The most commonly used are cosmids (max 45 kb), BACs (max 300 kb) and YACs (max 400 kb).
* The clones are selected so that they cover the genome (on a per-chromosome basis) in as complete but non-redundant fashion as possible, thus defining a so-called tiling path. This is done by so-called physical mapping.
* For each selected clone, sequence it by fragmenting the DNA in them randomly, sequencing about 500-700 bases, and reassembling the complete clone sequence from the data.
* The complete genome is then reassembled from the known tiling path and the clone sequences.


Confirm overlap
Walk by computation
Parachute with STS
Walk by hybridization


a
100 kb
99.94%
b1
t1
t2
b2
b
66M18
477B5
123G1
553C13
109G18
322J1
400I17
95B23
427G18
c
1 2 3 4 5 6 7 7 6 5 4 3 2 1
Nucleotide
T A G G A A T C
Tandem
L L L S S

Shotgun Based Sequencing

Obtain a pure sample of the entire genome of the organism (all chromosomes), and split it up into small fragments. The size is on the order of 500-1000 bases.

Clone these small fragments, thus creating a genomic library.

Sequence as many of the clones as necessary.

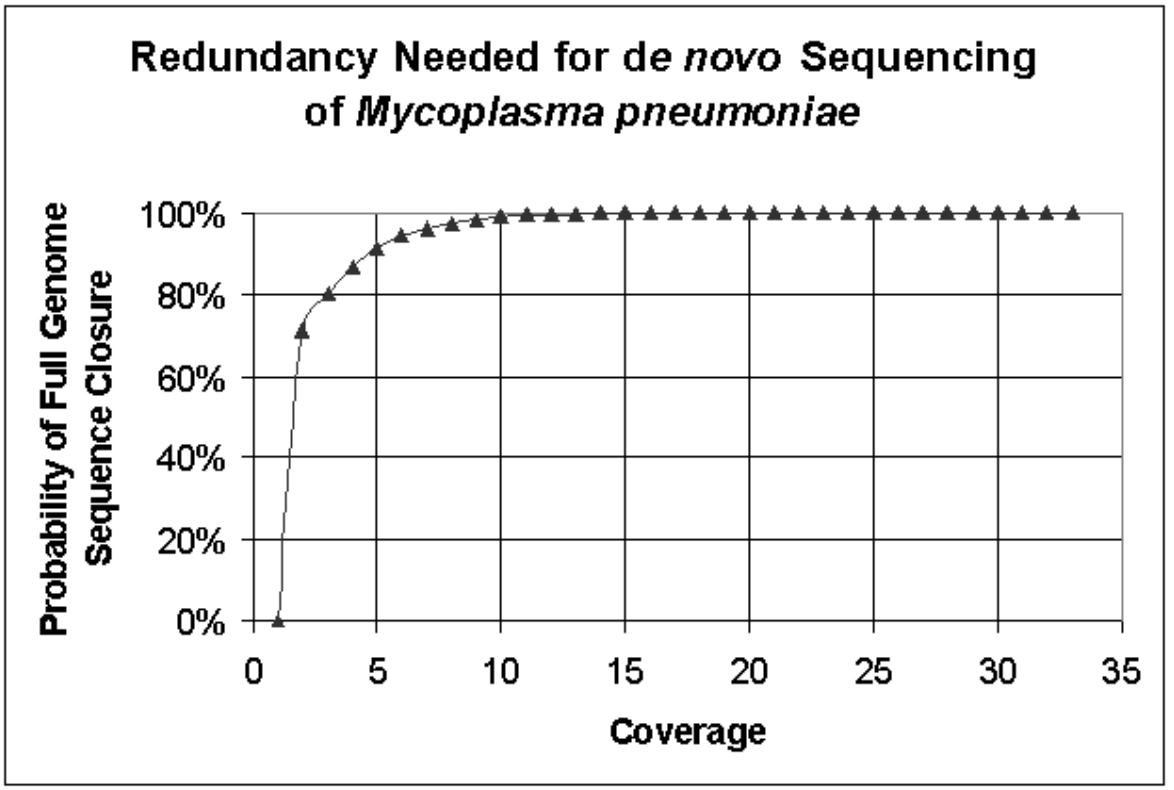
Reassemble the genome by computational analysis of the sequence fragments. For this to work, an oversampling of the genome is required, so that the residual number of unclosed gaps is as low as possible.

Problems with shotgun sequencing

The law of diminishing return:

There are always parts of a genome that are more difficult to sequence than others. For example, the DNA sequence may cause problems in the bacterial system used to maintain the clones, so that such clones are eliminated or strongly underrepresented in the genomic libraries. This also applies to gap-finishing: Some gaps are more difficult than others. The final 5% of sequence is more expensive to obtain than the first 50%.

Cost of finishing a genome



Problems II

The existence of intractable regions in the genomes.

In the higher eukaryotes, a very variable part of the genome consists of extremely repetitive sequence, which no known technology can handle. This part of the genome is called the heterochromatin, while the euchromatin are the more 'normal' parts of the genome. In Drosophila melanogaster the entire genome has a size of about 180 Mb, of which about 120 Mb is euchromatin, and was sequenced by Celera. The rest, 60 Mb or so of heterochromatin, has not been sequenced, and will not be, since it is impossible with the currently available techniques.

Challenges for finishing genomes

Resolving repeats, inversions, duplications: requires improved physical maps (use more markers)

Closing gaps: new clones to span gaps

Heterochromatin gaps, including telomeres: technology doesn't exist yet

Next we need to detect/sequence a complete set of polymorphisms...

Prokaryotic genome assembly

- *Traditional (mapping) approach*
  - *E.Coli*
  - *B. Subtilis*
- *Whole genome shotgun:*
  - *H. Influenzae (First complete genome)*
  - *All others*

Eukaryotic genome assembly

- *Classical approach:*
  - *Whole genome shotgun is diffucult due to repeats and low-complexity regions*
  - *Shotgun sequencing of large clones -> contigs*
  - *Assemble chromosomes using genetic and physical maps*
- *Celera:*
  - *Whole genome shotgun ?*
- *General rule: More overlap -> better assemblies*
- *Never really complete*

Some currently available genomes

(Nov 2005 at NCBI)

25 Archeae

280 Bacteria

20 Eukaryotic

- *4 vertebrates (Human, Mouse, Rat, Zebrafish)*
- *7 Plants*

Many almost complete (Dog, chimpanzee...)

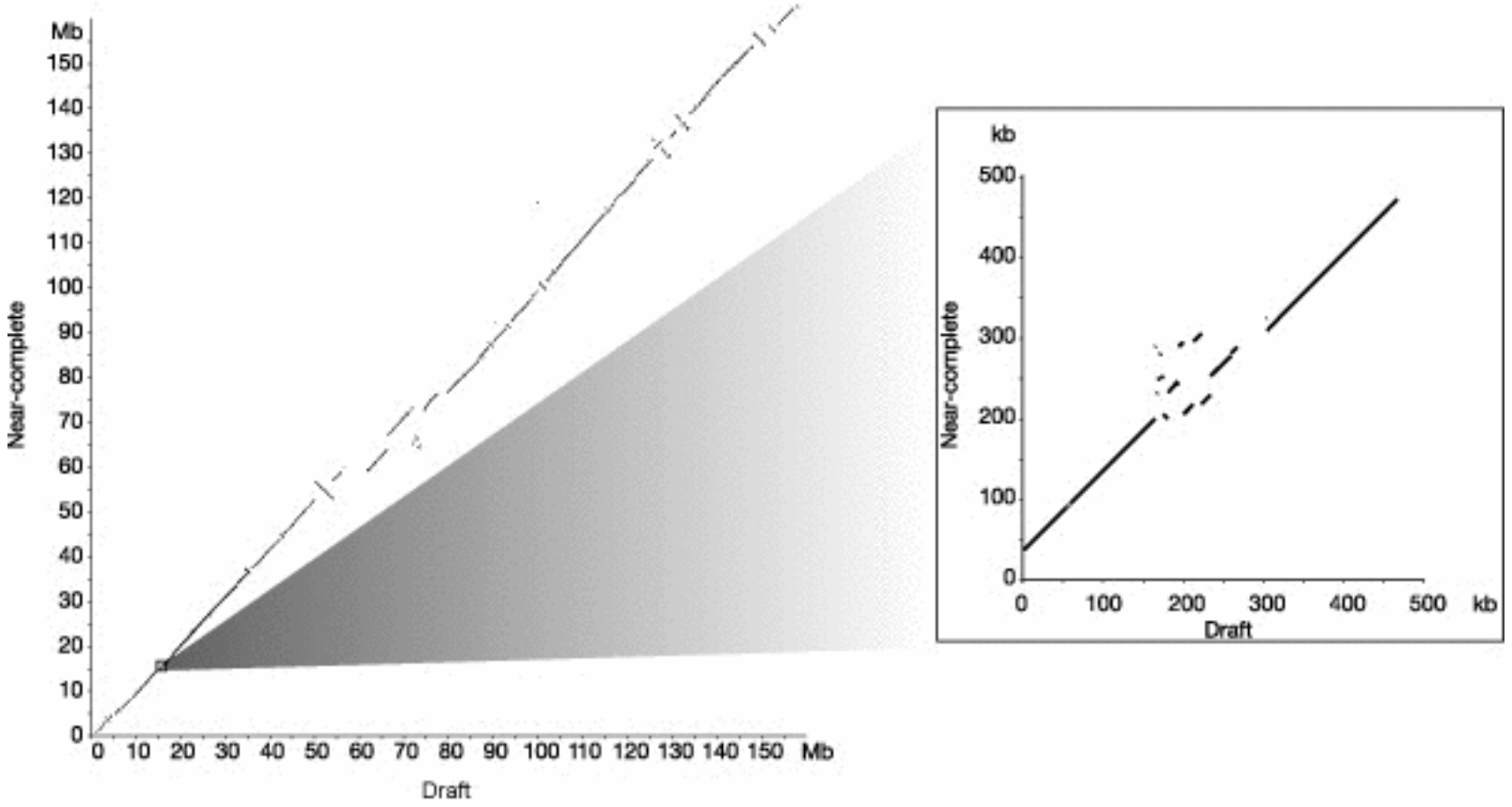
Finishing the human genome

*Recommended reading: ``Finishing the euchromatic sequence of the human genome", Nature **2004** 431, 931 – 945*

	2001 ``Draft"	2004 ``Near-complete"
gaps	147,821	341 (198 Mb)
gaps in euchromatin		308 (28 Mb)
finished euchromatic	1 Gb	2.85 Gb
fraction finished	34%	99%
fraction draft	95%	
mean contig size	80 kb	40,000 kb
errors	1 per 10 kb	1 per 100 kb

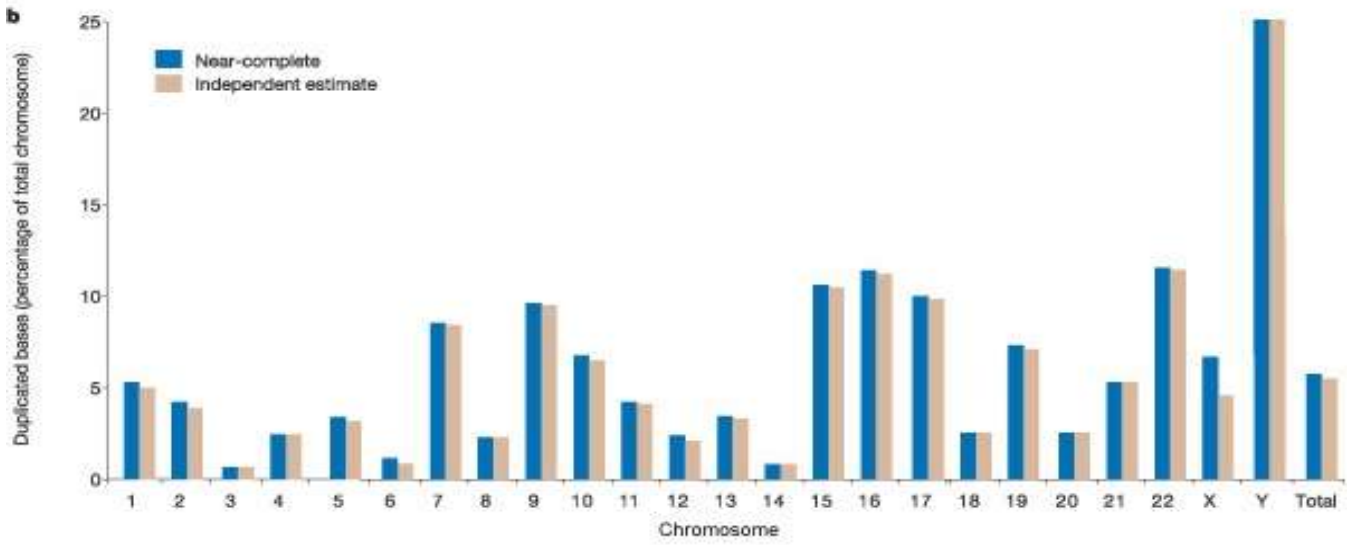
- Fewer erroneous duplications in finished sequence
- Exon structure was wrong in 39% of draft genes
- Producing draft and finishing took equal resources!

Finishing the human genome



Uses for finished genomes

- *Detailed study of duplications, rearrangements, gene birth/death, pseudogenes.*
- *Improved resource for all types of computational analysis.*
- *Framework for other vertebrate/mammalian/primate genomes.*



Location of genes

- # Gene prediction ab initio using software based on rules and patterns. Find Open Reading Frames (ORFs), with some additional criteria. Fairly simple for bacteria, very difficult for eukaryotes.
- # Gene identification through alignment with know proteins and EST sequences.
- # Gene prediction through similarity with proteins or ESTs in other organisms.
- # Gene prediction through comparison with other genomes; conserved regions may be coding or regulatory regions. Synteny.

Uses of genomes

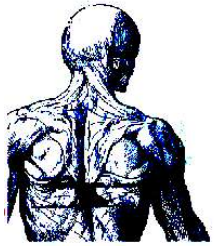
- # Annotation of the genes: Identify with known genes, similarity with genes in other organisms. Essentially: labelling the gene.
- # Functional classification. Broad groups of functional characterization, such as 'ribosomal proteins', 'nucleotide metabolism', 'signal transduction'.
- # Metabolic pathways.
- * Are any common pathways missing?
- * Are there 'gaps' (missing enzymes) in some pathways?
- * Compare identified pathways with the life style of the organism.

Uses of complete genomes

- Evolutionary history

- Internal genome duplications can sometimes be detected.
- Gene decay can sometimes be characterized: genes that are on their 'way out' after duplication, or because the life style of the organism has changed.

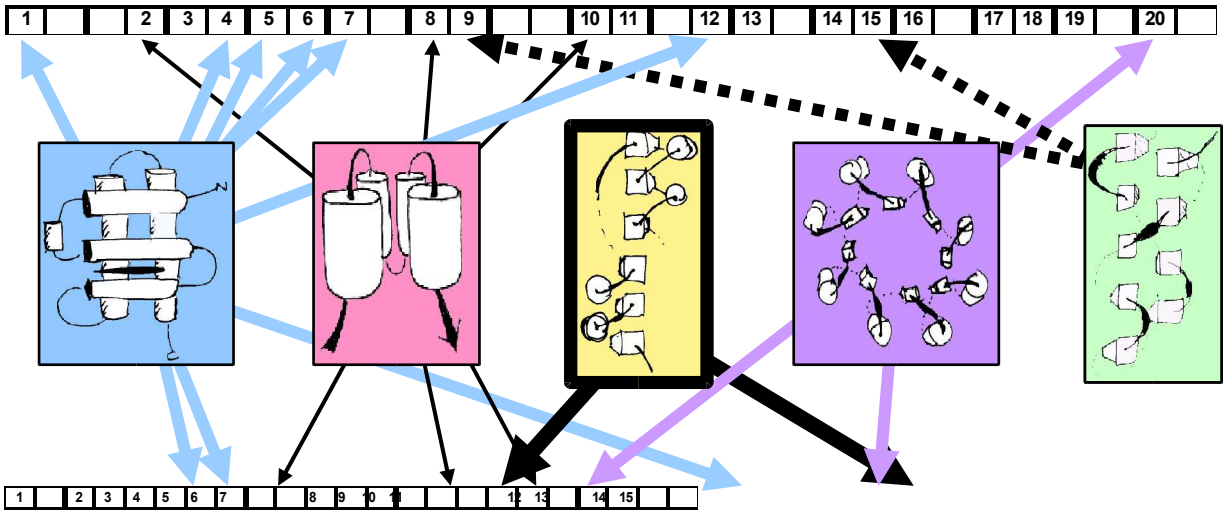
Global Surveys of a Finite Set of Parts from Many Perspectives



(human)

(*T. pallidum*)

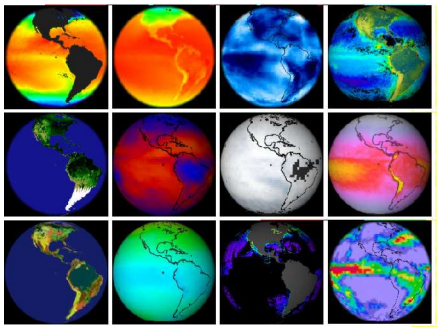




~30000 genes

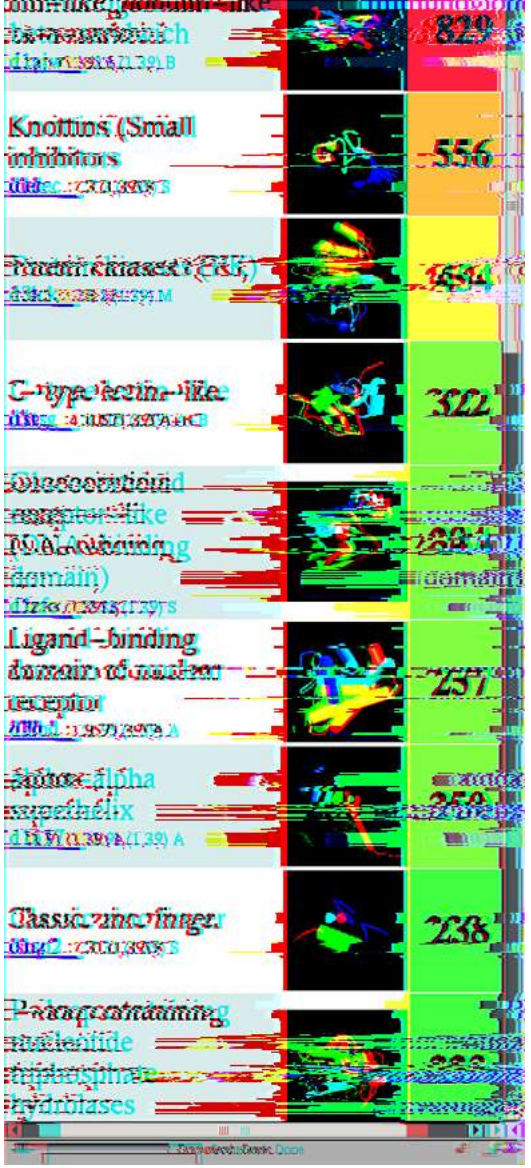
~1000 folds

~1000 genes



World of structures is even more finite than that of sequences, providing simplification…

Common Folds

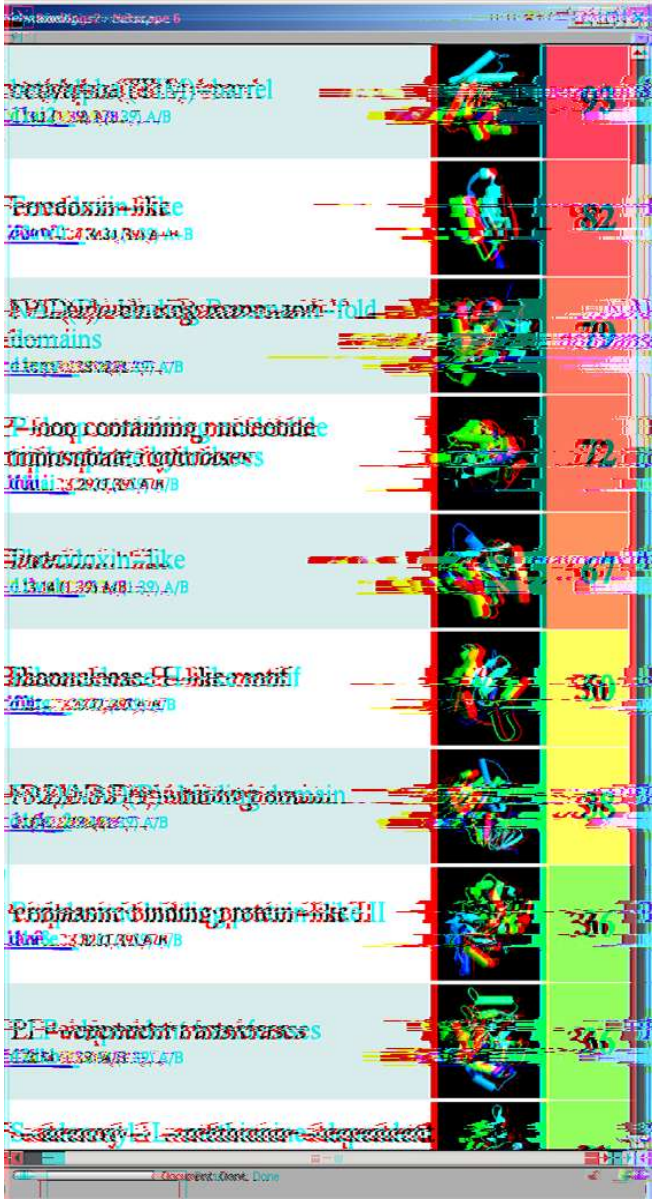


Worm

Extracellular, signaling and trans. factor folds (Ig, kinase, lectin, nuc. receptor, ZnF)

E. coli

Metabolic Folds (TIM, Ferrodoxin, Rossman, P-loop hydrolase, FAD-binding)



(partslist.org)

How many proteins do we have?

- Coding exons represent 1.2% of euchromatic genome
- Ignoring splicing and post-translational modifications:
  - estimated 20-25,000 protein-coding genes in human.
  - 24,194 (23/11 2005) ``known genes'' -- solid evidence (EST, cDNA, cloned, highly conserved in another species)
- Genes with low expression levels may be missed
- Computational methods for ``gene prediction'' needed (protein and RNA genes)

Gene prediction

- *Prokaryotes*
  - *Compact genomes, no introns*
  - *Long (>300 bp) open reading frames without STOP codons rarely occur by chance*
  - *Genetic code and codon usage can differ between organisms*
  - *Short genes often missed*
  - *Non-protein coding genes do not look the same*
  - *Cross species comparisons can aid*

Gene prediction II

- *Eukaryotes*
  - *Much more difficult*
  - *only 1.2 % protein coding DNA*
  - *transcript start sites (TATA) less clear*
  - *very BIG genes*
  - *weak intron/exon boundary signals*
- *Two main approaches*
  - *Using sequence similarity*
  - *pure computation / ab initio*

ab initio gene predictions

- *Example method: ``GENSCAN'' (used at www.ensembl.org)*
- *Predicts protein coding genes using models of*
  - *typical gene density*
  - *typical number of exons per gene*
  - *distribution of exon sizes for different types of exon*
  - *reading frame-specific hexamer composition of coding regions vs*
  - *the (reading frame-independent) hexamer composition of introns and*
  - *intergenic regions*
  - *position-specific composition of the translation initiation*
  - *(Kozak) and termination signals*
  - *TATA box, cap site, poly-adenylation signals*
  - *donor and acceptor splice sites*
- *Predicts 68,101 genes in the human genome (confirmed 24,194)*

Similarity based gene-prediction

- *Example: GENEWISE (also used at Ensembl)*
- *Alignment based methods (protein and DNA)*
- *Also models splice sites*
- *Problem:*
  - *Ultra Conserved elements between species (Science 28 May 2004:Vol. 304. no. 5675, pp. 1321 - 1325)*
  - *ORPhans*

RNA gene prediction

- *tRNA, rRNA, miRNA....*
- *Secondary structure can help search algorithms*
- *Similarity search algorithms differ*
- *Hot topic at the moment*
- *MicroRNAs may control expression of 1000s of genes (in mammals)*

Comparing genomes

is now possible to investigate which sets of genes are common to many different organisms, or groups of organisms. Is there a common core of genes necessary for all life? Is that core sufficient for life?

Are all the ribosomal proteins really similar between all known species, or have there been inventions during the course of evolution in this specific, but fundamental system?

Which genes are necessary for multicellular life forms; which set of genes are only found in multicellular organisms but not in unicellular ones?

Comparing genomes

The rate of horizontal gene transfer (genes that have jumped the species barrier) among bacteria can now be investigated. How often, and under what circumstances do bacteria exchange genes? Has anything similar happened with higher organisms?

Where and how have new genes emerged in evolutionary history? Can precursors of some gene families be found in distant relatives of a species?

Comparing genomes

The problem of identifying and characterizing orthologous genes versus paralogous genes becomes easier to address (but not necessarily solve).

  * Orthologues are genes that have diverged from a common ancestor because of a speciation event.

  * Paralogues are genes that have diverged as the result of a gene duplication event.

Warning: The term orthologous and paralogous are sometimes used to denote functional equivalence, and similarity without functional equivalence, respectively, but this is frowned upon by evolutionary biologists.

Reading

- Essential:
  - Course book – Chapter 2 “Gene Finding”
- Optional/additional
  - Article -- ``Finishing the euchromatic sequence of the human genome", Nature 431, 931 - 945 (21 October 2004)
  - Online books at http://www.ncbi.nlm.nih.gov/ -- e.g. ``Genomes" and ``Human Molecular Genetics *2*″

Non-genomic sequencing

- Venter:
  - Sequencing the Saragasso sea
- Others
  - Soil-sequencing, etc
- Results
  - Many incomplete genes
  - No complete genomes

Beyond the genome

- Description Time dependence Manipulation in the cell Measurement in the cell
- Genome The complete DNA sequence of the organism. First approximation: none. Genetic manipulation; site-directed mutagenesis, genetic screens. Routine. High accuracy and completeness.
- Transcriptome The complete set of mRNA transcripts in a cell or tissue of the organism. Reflects the state of gene expression. Indirectly via genetic manipulation. Routine, or nearly so. cDNA microarrays, SAGE, Northern blots. Quantification inaccurate. Approximate completeness.

Beyond the genome

- Proteome The complete set of proteins present in a cell or tissue, including variants due to covalent modifications. Sometimes also the associations with other molecular components (complexes). Responsive to all manner of influences throughout the life cycle. Indirectly through genetic manipulations. Small molecules as inhibitors of enzymes. Not routine. 2D-gels with mass spectrometry. Fusions with reporter proteins (GFP). Yeast two-hybrid system. Quantification very difficult. Completeness far away, and very difficult.

Beyond the genome

- Metabolome All metabolites and their concentrations. Defined as organic small molecules, not protein, RNA or DNA. Responsive to all manner of influences throughout the life cycle. Difficult. Not routine. NMR, mass spectrometry, a few other special methods for specific molecules. Quantification and resolution difficult. Completeness far away, but maybe reachable.