

Structural Biochemistry and Bioinformatics

Stockholm Bioinformatics Center

SU / KTH / KI

Teachers:

Arne Elofsson	course organiser (on leave until Dec)
Bob MacCallum	today & structure prediction
Erik Lindahl	sequence methods, modelling
Lotta Berglund	molecular evolution, phylogeny
Gunnar von Heijne	membrane proteins, protein interactions
Erik Sandelin	structure comparison
Olivia Eriksson	teaching assistant
Björn Wallner	teaching assistant

What is bioinformatics?

“Dealing with bio-data”

Some sources of data

- genes/genomes
- biomolecular structure
- high throughput technologies
- clinical data
- biomedical literature

Genes, genomes and gene finding

Today's topics:

What is a genome?

Mapping and sequencing genomes

History of genome sequencing

Uses for genomes

Gene finding

What is a genome?

- Simple: sequence of nucleotides in an individual (who is the individual?)
- Mitochondria and chloroplasts too
- Blueprint of life? maternal factors...

What is a gene?

Difficult question!

Region of genome corresponding to a functional unit
(usually protein or RNA)

But also genetic traits & diseases

Why study genomes?

Complete sets of genes, proteins and RNA can be used to:

- characterise gene families
- describe biological systems
- plan experiments (e.g. knockouts)

Genomic DNA → regulation of gene expression

Genome comparison: study evolution, pathogenicity, ...

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 – 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

Can be used to make genetic or physical maps

Linkage between (co-occurrence of) disease and polymorphic markers can help locate disease genes.

DNA Sequencing

Sanger's *chain termination method* 1977 is widely used and automated

Length of DNA limited to around 1-2kb

Shotgun approach: break up DNA into small fragments, reassemble whole sequence computationally via overlaps.

Whole-genome shotgun approach works for compact genomes

Eukaryotic genome assembly

Whole-genome shotgun is difficult due to repeats and low-complexity

Shotgun sequencing of large clones → contigs

Assemble chromosomes using genetic and physical maps

General rule: more overlaps → better assemblies

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	<i>S. cerevisiae</i> chr. III	315 kb
1994	<i>S. cerevisiae</i> chr. XI	666 kb
1995	<i>H. influenzae</i> genome	1.83 Mb
1996	<i>S. cerevisiae</i> genome	12 Mb
1997	<i>E. coli</i> genome	4.6 Mb
1998	<i>C. elegans</i> genome	97 Mb
2000	<i>D. melanogaster</i> genome	120 Mb
2001	Human genome	3300 Mb

Some currently available genomes

18 Archaea and 150 Bacteria

Eukaryotes at ensembl.org, NCBI:

Human, mouse, zebrafish, rat, chicken, mosquito, fugu, fruitfly, chimp, honeybee, pufferfish, cow, dog, *C. elegans*, *C. briggsae*...

Many only at “draft” stage

Finishing the human genome

Recommended reading: “Finishing the euchromatic sequence of the human genome”, *Nature* 431, 931 - 945 (21 October 2004)

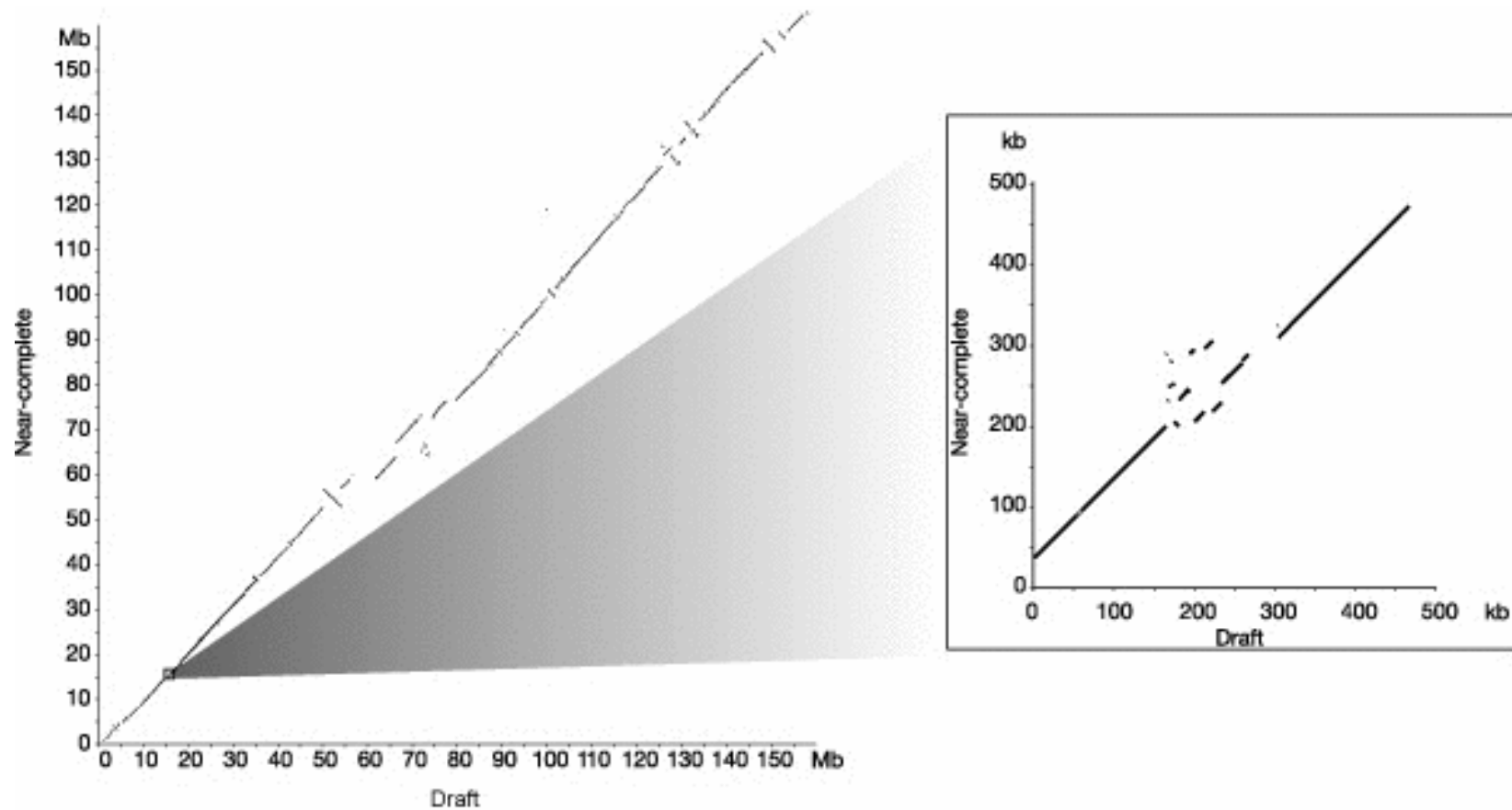
	2001 “Draft”	2004 “Near-complete”
gaps	147,821	341 (198 Mb)
gaps in euchromatin		308 (28 Mb)
finished euchromatic	1 Gb	2.85 Gb
fraction finished	~34%	~99%
fraction draft	~95%	
mean contig size	80 kb	40,000 kb
errors	1 per 10 kb	1 per 100 kb

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



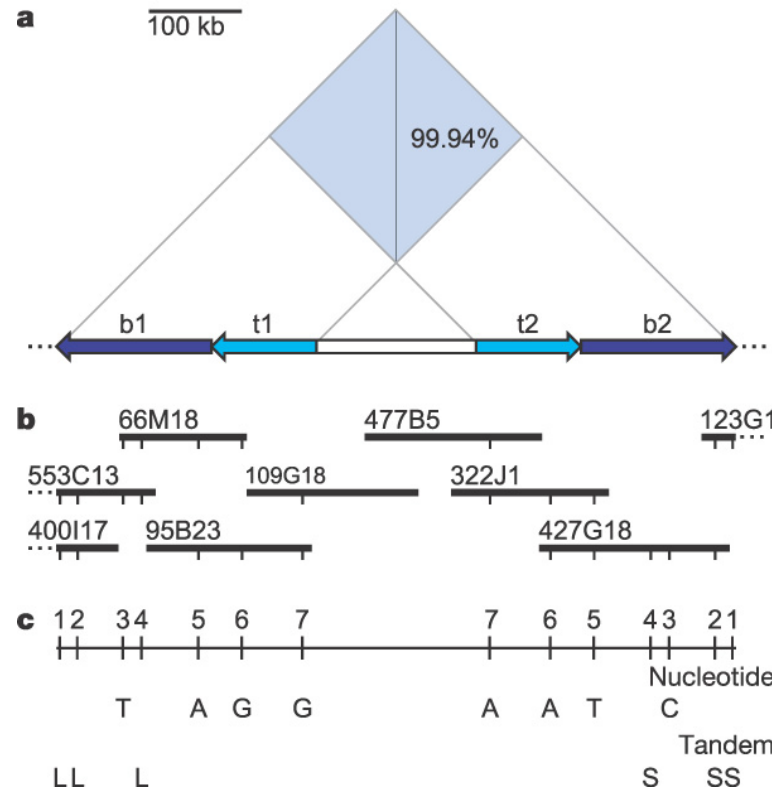
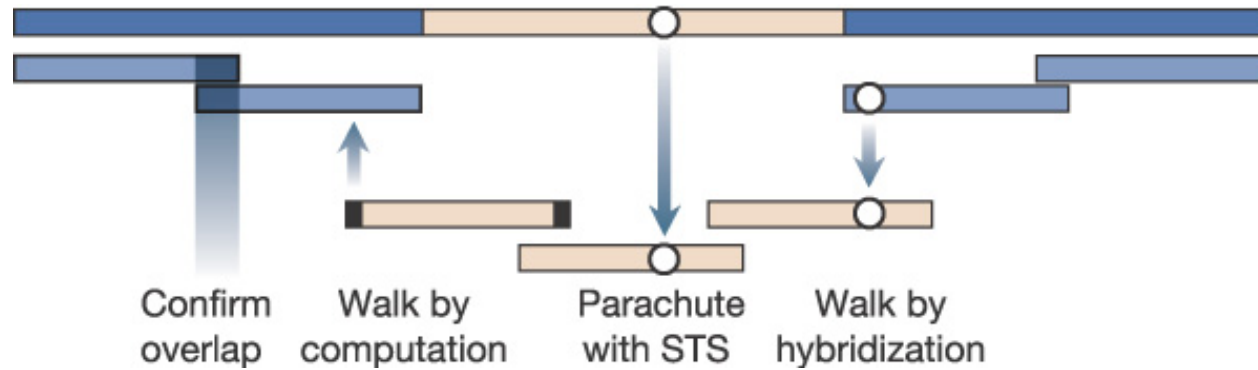
Challenges for finishing genomes

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...

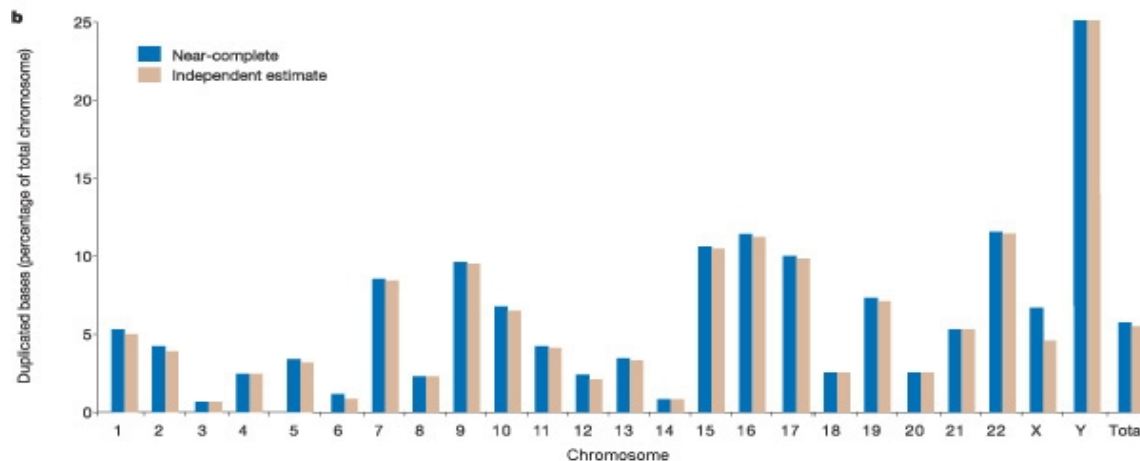


Uses for finished genomes

Detailed study of duplications, rearrangements, gene birth/death, pseudogenes.

Improved resource for all types of computational analysis.

Framework for other vertebrate/mammalian/primate genomes.



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.

19,599 “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

1. Prokaryotes

Compact genomes, no introns

Long (>300bp) *open reading frames* rarely occur by chance

ATG TAA

Genetic code and codon usage can differ between organisms

Short genes often missed

Non-protein coding genes don't look the same

Cross-species comparisons (>150 genomes) also help find genes

Gene prediction

2. Higher eukaryotes

Much more difficult...

- only 1.2% protein coding DNA
- transcription start sites (TATA) less clear
- very big genes
- weak intron/exon boundary signals

Two main approaches:

- using sequence similarity
- pure computational / *ab initio*

ab initio gene prediction

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

- C+G content

Similarity-based gene prediction

Example: GENewise (also used at Ensembl)

Alignment based method (protein \leftrightarrow DNA)

Also models splice sites

RNA gene prediction

tRNA, rRNA, miRNA, ...

Secondary structure can be used in search algorithms

Hot topic at the moment

MicroRNAs may control expression of 1000s of genes (in mammals)

Reading

Essential

Course book – Chapter 2 “Gene finding” from *Bioinformatics – Genes, Proteins and Computers* by Orengo, Jones and Thornton. (This also includes material on genetic and physical mapping.)

Optional further reading / study aids

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”