



Sequence analysis 2

Imperial College
MSc. Bioinformatics and
Computational Biology

Introduction

- Where does sequence data come from?
 - Always remember that any bioinformatic analysis is only as good as the data it is applied to.
- Where does sequence data go?
 - Always remember that any bioinformatic analysis is only as useful as the interface to its results.

Themes

- theory and practice
- empirical and modelled
- interfaces and algorithms
- *in vitro* versus *in silico*
- always get a third opinion



Raw genomic sequence

Output from sequencers

- limited read length
 - 200K bp cloned per experiment
 - 1000 bp per read (on a standard 3700 sequencer)
 - 8000-9000 reads
- limited quality
 - 4000-5000 of these are of sufficient quality

Remember the source

```
HUMXT 317 GCGTTGCTGGCGTTTTTCCATAGGCTCCGACCCCCTGACGAGCATCACAAAATCGACGCTCAA
*****
DINO1 1 GCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCTGACGAGCATCACAAAATCGACGC----
*****
DINO1 670 GCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCTGACGAGCATCACAAACAAGTCAGA----

HUMXT 234 GTCANAGGTGGCGGAAACCCGACAGGACTATAAAGATAACCAGGCGTTTCCCCCTTGGAGCTTCC
*****
DINO1 61 -----GGTGGCG-AAACCCGACAGGACTATAAAGATAACCAGGCGTTTCCCCCTGGAAGCTCCC
*****
DINO1 730 -----GGTGGCG-AAACCCGACAGGACTATAAAGATAACCAGGCGTTTCCCCCTGGAAGCGCTC
```

“Fact and Fiction in alignment”

Henikoff and Christensen

Nature **358** 271 (1992)

Destination of sequence data

The screenshot displays the Ensembl Genome Server interface within a Netscape browser window. The browser's address bar shows the URL: `http://www.ensembl.org/Homo_sapiens/contigview?clone=AL136096&geneid=ENSG00000118432&contig=AL136096.7.1.112212`. The page title is "Ensembl ContigView".

The main content area is titled "Chromosome" and shows a chromosome map for Chr6. Below this is an "Overview" section with a detailed view of the contig AL136096. The overview includes a scale from 96.49 Mb to 97.49 Mb. The contig is shown as a blue bar with several segments. Below the contig, various features are listed:

- DNA(contigs):** AL451126, AL049697, AL136096 (highlighted), AL121835, AL139042, AL139096.
- Markers:** WI-19695, WI-9825, AFMB236209, WI-3369.
- Genes:** ORC3L, SLC35A1, SPACR1, CNE1, and several other contig-specific genes like L43382110.5, L43382110.4, L43382110.3, L43382110.2, L43382110.1, L43382110.6, L43382110.7, L43382110.8, L43382110.9, L43382110.10, L43382110.11, L43382110.12, L43382110.13, L43382110.14, L43382110.15, L43382110.16, L43382110.17, L43382110.18, L43382110.19, L43382110.20, L43382110.21, L43382110.22, L43382110.23, L43382110.24, L43382110.25, L43382110.26, L43382110.27, L43382110.28, L43382110.29, L43382110.30, L43382110.31, L43382110.32, L43382110.33, L43382110.34, L43382110.35, L43382110.36, L43382110.37, L43382110.38, L43382110.39, L43382110.40, L43382110.41, L43382110.42, L43382110.43, L43382110.44, L43382110.45, L43382110.46, L43382110.47, L43382110.48, L43382110.49, L43382110.50, L43382110.51, L43382110.52, L43382110.53, L43382110.54, L43382110.55, L43382110.56, L43382110.57, L43382110.58, L43382110.59, L43382110.60, L43382110.61, L43382110.62, L43382110.63, L43382110.64, L43382110.65, L43382110.66, L43382110.67, L43382110.68, L43382110.69, L43382110.70, L43382110.71, L43382110.72, L43382110.73, L43382110.74, L43382110.75, L43382110.76, L43382110.77, L43382110.78, L43382110.79, L43382110.80, L43382110.81, L43382110.82, L43382110.83, L43382110.84, L43382110.85, L43382110.86, L43382110.87, L43382110.88, L43382110.89, L43382110.90, L43382110.91, L43382110.92, L43382110.93, L43382110.94, L43382110.95, L43382110.96, L43382110.97, L43382110.98, L43382110.99, L43382110.100.

The "Detailed View" section at the bottom includes navigation controls such as "Zoom", "Window", and "Navigation". The "Navigation" panel shows a scale from 96.94 Mb to 97.03 Mb, with a current position of 100.00 Kb. The "Navigation" panel also includes a "Centre on this scale interval" button.



Finishing sequence data

The bioinformatics of finishing

- reads overlaid
 - use sequence comparison algorithms to align reads
- ends re-sequenced
 - use primer design programs to prepare primers for sequencing off ends of contigs if needed
- programs used: Staden, PHRED, PHRAP

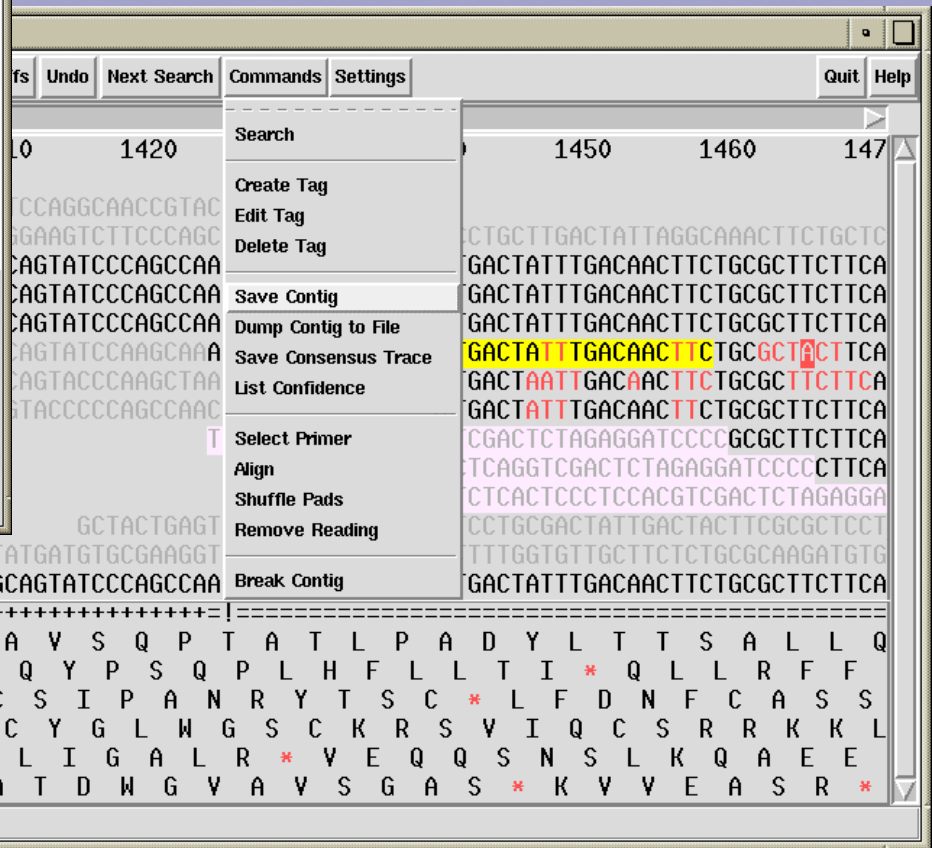
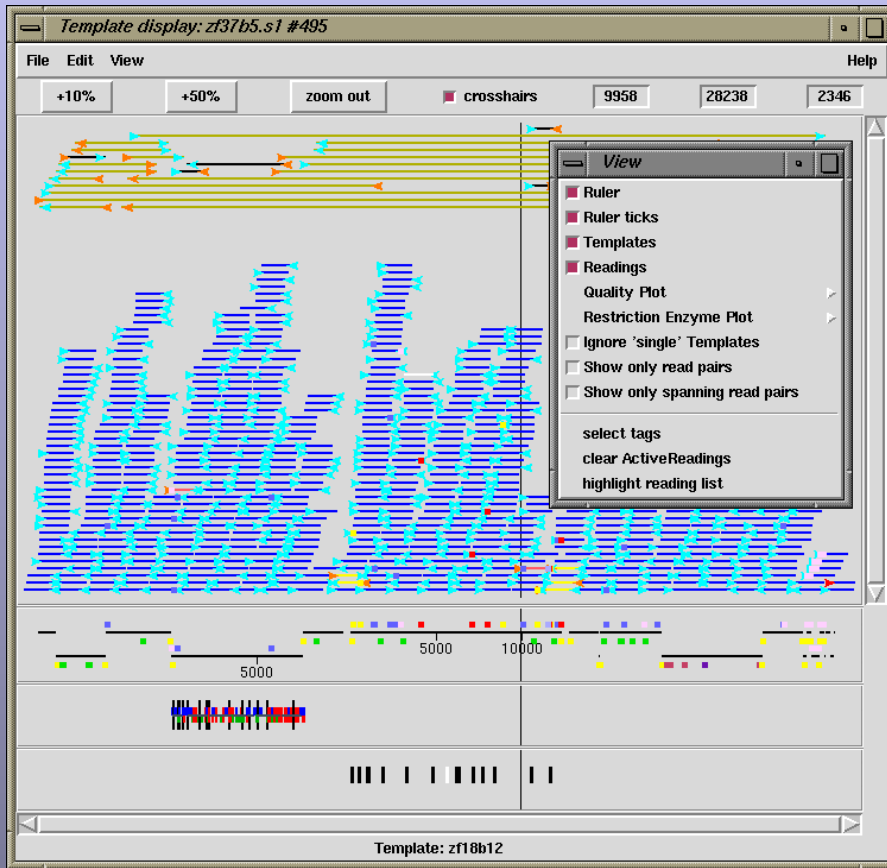
The bioinformatics of finishing *(cont.)*

- ends of completed contigs can then be joined by software (Staden again)
- remaining problem regions can be tackled like stunted ends
 - usually a problem of chemistry
 - different reagents can be used to deal with, for example, Hairpin DNA

What does the software do in the process?

- provides an intuitive interface to the sequence data which can be viewed at many levels
- allows examination of original dye data
- aligns reads
- flags up differences between reads
- matches up joins between contigs
- masks vector
- masks/highlights repetitive DNA

Source of sequence data



SSAHA

- different groups in different locations using different approaches can combine their reads
 - search with sequence from problem region
 - server returns matching reads from remote group
 - add new reads into assembly for further refinement of contigs



Removal of repeats

RepeatMasker

- removes simple, tandem and complex repeats
- compares sequence against a library of repetitive elements as well removing low complexity regions
- uses a more efficient Swiss Waterman variant (more later)
- returns masked query sequence
- server based---no need to download database or program

RepeatMasker

- Theoretically time-to-run should increase linearly with sequence size--- actually much faster with short sequences, much slower with long ones.
 - Cross_match stores queries in RAM

RepeatMasker

- Low complexity DNA screening
 - AT or GC rich sequence causes spurious matches (e.g. with mitochondria seqs)
 - Simple repeats can occur anywhere and may be interspersed
- Program has doubled running time now that repeat databases have increased.

Other programs...

- ...(e.g. sputnik) should be used to *find* polymorphic simple repeats
 - Interspersed repeats masked first and mask out simple ones
 - Only 2- 5-meric scanned for
 - Simple repeat regions often highly divergent, but not polymorphic

Vector or manual repeat removal

- BLASTN against database of repeat/vector sequences
- XBLAST against query (masks unwanted sequence with ambiguity characters)
- BLASTN against database

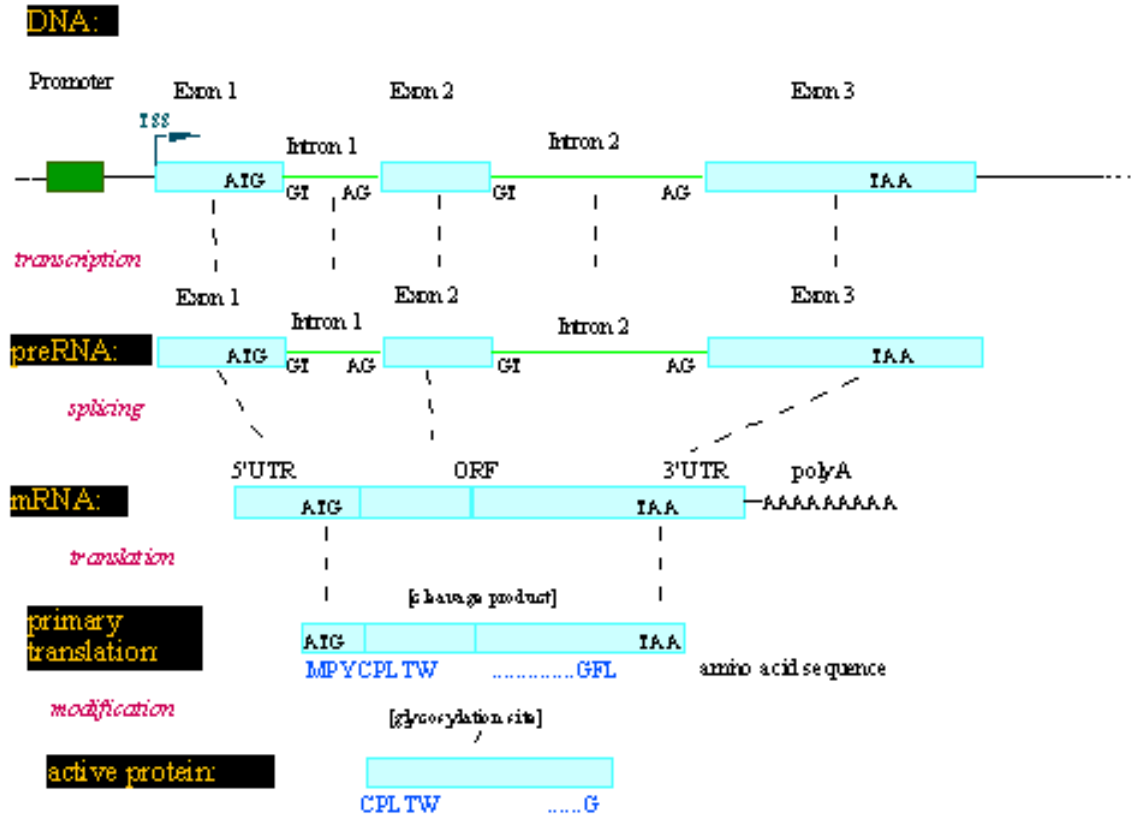


Gene identification

Gene identification

- find non-coding features of interest
- find genes
- identify genes
- remember: you can also find genes by doing some “old-fashioned” wet-lab work

Gene structure



Overview of gene identification/characterization

- identification of promoter sites
 - crucial flags for genes
- elucidation of gene/exon boundaries
- identification of CpG islands
 - found in most housekeeping genes
- identification of tRNA genes
- identification of repetitive elements
 - that is, ones not already removed

Other concerns (not covered today)

- SINE elements
- poly-A sites
- vector / E. coli contamination
- transcription factors or other control regions

Sanger Centre approach

- not formally set---exists in the minds of the sequencing management teams
- multi-step
- multi-program
- a human always makes the final pass.

Sanger detects genes with...

- BLASTX plus
- GeneWise (Birney)
 - uses HMMs---probabilistic profiles
 - LLRs as profile-scoring systems
 - compares genomic sequence to PFAM protein domain database
 - therefore must translate and maintain frame
 - tolerant of interruptions (exon/intron boundaries)

Accuracy of GeneWise (claimed)

- 90% identical protein
 - 99% accurate
 - 90% coverage
- 40% identical protein
 - 99% accurate
 - 30% coverage
- when identity falls accuracy stays high, coverage decreases



Gene feature elucidation

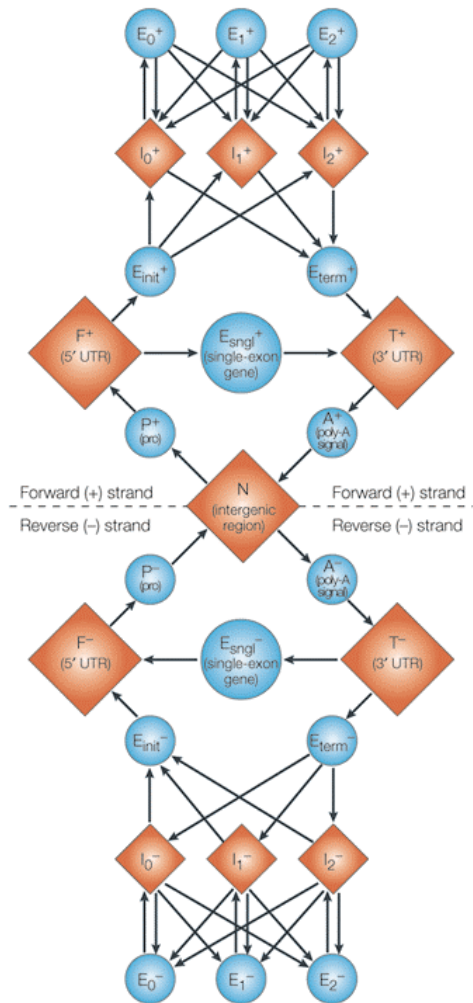
Promoter sites

- predicted promoter site upstream of a gene is good supporting evidence
- but prediction methods poor, with many false positives (low selectivity) and false negatives (low sensitivity)
- Many exon finding programs also attempt to locate promoters
- Best programs FGENES and TSSW

FGENES (Solovyev 97)

- Pattern-based variants of human gene structure prediction
 - large gene prediction only 70-80% accurate so suboptimal model could be correct
 - Mammalian genes often have alternative splice sites
- algorithm based on pattern recognition of different types of exons, promoters and poly-A signals
- optimal combination of these features is then found by dynamic programming

Sanger predicts gene structures with...



- Genscan

- (Burge and Karlin 97)

- HMMs again

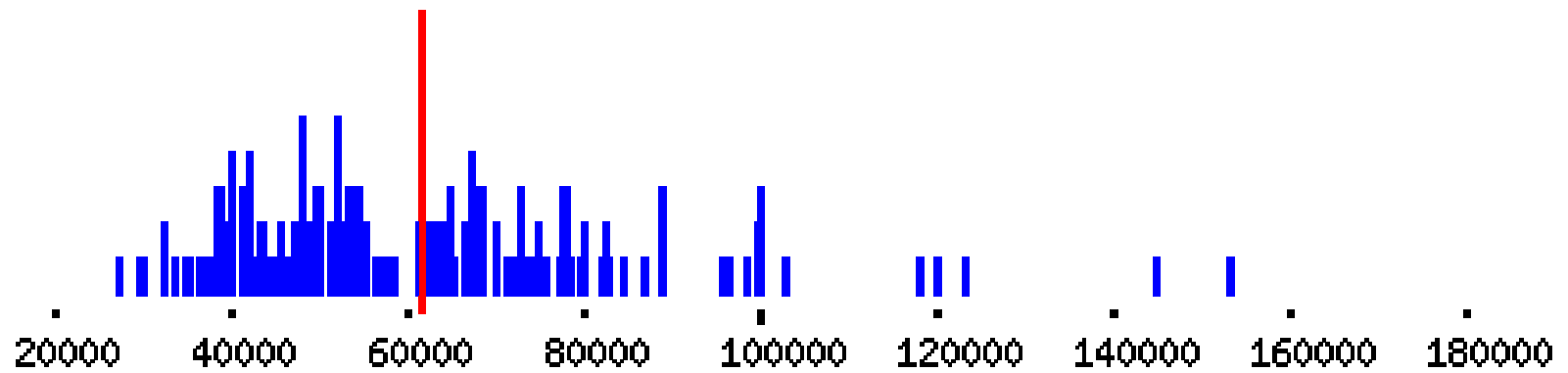
- Various states are various elements of gene structure

Accuracy of GENSCAN

range (bp)	Annotated exons				Predicted exons			
	No.	%Exact	%Part	%Miss	No.	%Exact	%Part	%Wrong
<= 24	89	38	8	52	44	77	11	11
25 - 49	163	58	15	25	124	76	6	18
50 - 74	248	70	12	16	204	85	9	6
75 - 99	382	85	8	6	389	84	6	10
100 - 124	351	84	9	7	366	81	8	11
125 - 149	425	88	8	4	460	81	10	7
150 - 174	261	88	9	2	283	81	11	7
175 - 199	167	91	7	2	188	81	12	7
200 - 299	353	90	8	1	390	82	8	8
>= 300	211	66	19	1	204	69	20	10
Total	2650	81	10	8	2678	81	10	9

Tested on 570 vertebrate genes

Genesweep



CpG islands

- adjacent C and G referred to as CpG to avoid confusion with opposite C and G
- C typically methylated
- susceptible to mutation to T
- methylation suppressed in short stretches of genome, such as near promoters
- therefore more CpGs here than elsewhere

CpG islands

- about half of mammalian genes have a CpG rich region
- thought that all mammalian housekeeping genes have a CpG island
- non-mammals also, but not so reliable in more distant species
- usually tackled with HMM methods

Identification of tRNA sequences

- tRNAscan-SE
 - Lowe & Eddy, *Nucleic Acids Research* **25**: 955-964 (1997)
 - identifies Pol III promoters in first pass
 - follows up with probabilistic RNA detection
 - detected almost perfectly
 - <1 false positive per 15bn nucleotides
 - detects 99% of all true tRNAs

Low complexity regions

- give spurious high scores due to compositional bias
 - Shuffle trick to show this
- *e.g.*
 - CA repeats
 - Poly A tails
- use DUST (for nucleotides)
 - Tatusov and Lipman (God knows)