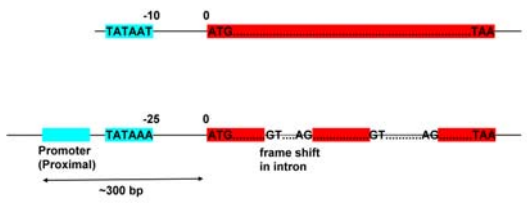
Annotation

- Gene Finding
  - Ab initio

Prokaryotic gene finding

- ORF e.g. GLIMMER
- Standard promoter sequence
  - Pribnow box: TATAAT

Prokaryotic gene finding



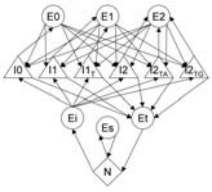
GenScan

Simultaneous forward and reverse genes
Nested genes missed/ no alternate splicing
Probabilistic model: includes
first, internal, last exon sub-models
O(Sequence Length x Model States)
So ~ O (M) and very fast.

What is happening in real life ?

SNAP - Semi-HMM-based Nucleic Acid Parser

- Each strand separately – allows nested genes <u>BUT</u> allows overlapping exons
- Bootstrap for parameter estimation

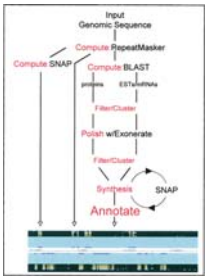


Integration of various evidence

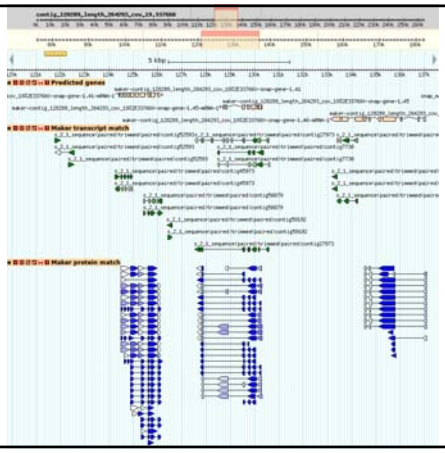
- manually
- using statistics and computational methods
  - simple counting
  - hidden Markov models
  - Bayesian statistics
  - neural networks
- Always best to use many different finders and combine. Some frameworks try to keep this process as user-friendly as possible, e.g. Maker

Maker genome annotation

gmod.org/wiki/MAKER

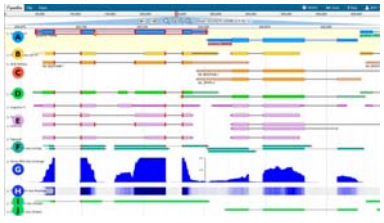


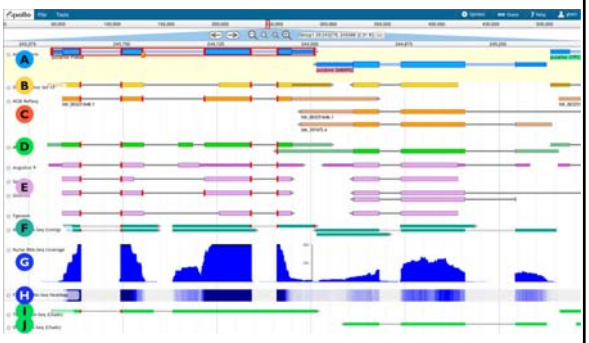
- Takes genome, EST and protein data
- Identifies repeats
- Aligns ESTs and proteins to a genome
- Makes gene predictions
- Integrates these data into protein-coding gene annotations.



Community annotation

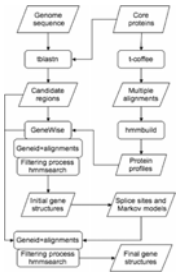
- Share annotations between groups
- WebApollo / Apollo
  - Jbrowse



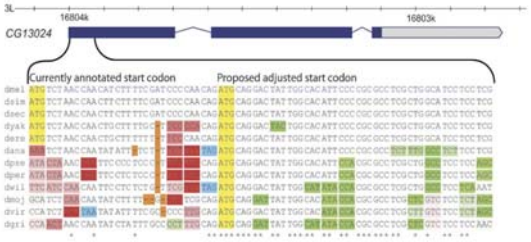


Core Eukaryotic Genes (CEGs)

- CEGs required for eukaryotic life
- 458 proteins/models in CEGMA set
- CEGMA aligns using HMM and DP



Making use of multiple genomes



Stark et al Nature. 2007 Nov 8;450(7167):219-232. PMID: 17994088
Lin et al Genome Research 2007 PMID: 17989253

Covered in Lecture 10

References

GenScan

- Burge, C. and Karlin, S. - Prediction of complete gene structures in human genomic DNA. - J. Mol. Biol. 1997

Doublescan

- Meyer and Durbin - Comparative ab initio prediction of gene structures using pair HMMs – Bioinformatics 2002

SNAP

- Korf, I. - Gene finding in novel genomes – BMC Bioinformatics 2004

References

- Lee *et al.* - Web Apollo: a web-based genomic annotation editing platform – Genome Biology 2013
- Parra *et al.* - CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. – Bioinformatics 2007

Perl

http://perldoc.perl.org/perlintro.html

- Practical 4 examples will be mostly in Perl (Python/Ruby/Java supported)
- Don't need to able to write it but will be very useful to be able to read it

Groups for Assignment 2

- MPhil Computational Biology (18 students)
  - 6 groups of 3
- Other…?
  - Let me know ASAP

Practical

- Get the transcripts FASTA for gene *trh* (*trachealess*) from FlyBase web site.
- Use the ORF finder (http://www.bioinformatics.org/sms2/orf_find.html) on each transcript. Use BLAT on the UCSC Genome Browser web site to map the longest ORFs to the genome. View the region.
- Do the FlyBase transcripts correspond with other cDNA/EST information available in the browser? Get cDNA sequence NM_001103991 from GenBank and repeat your analysis.
- A further analysis has identified chr3L:366538-366558 as potential target region of a miRNA. Use the 'add custom tracks' functionality to visually highlight the region in the browser. For help see http://genome.ucsc.edu/goldenPath/help/customTrack.html

Practical

- Get exons FASTA for gene *sim* (*single minded*) from FlyBase web site.
- The sequence ID will contain a substring like 'loc=3R:8898124..8898272'. Use this the genomic coordinate to order the exons ("from left to right").
- Using a simple longest-ORF search, piece together possible exon combinations (Start = ATG; Stop = TAA, TAG, TGA) and deduce all possible coding sequences.
- Modify your script's CDS output to be visualised in the UCSC Genome Browser. Have a look at http://genome.ucsc.edu/goldenPath/help/customTrack.html to help. Try it out using 'add custom tracks' in the browser view.