CS426: Gene Prediction

Niranjan Nagarajan

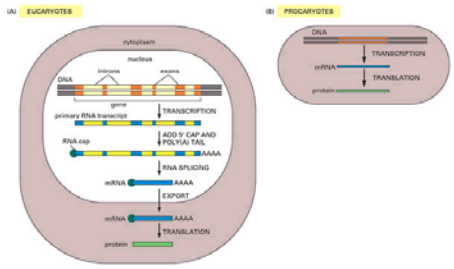
What is a Gene?

- Definition: An inheritable trait associated with a region of DNA that codes for a protein or specifies an RNA molecule which in turn has an influence on some characteristic phenotype of the organism.

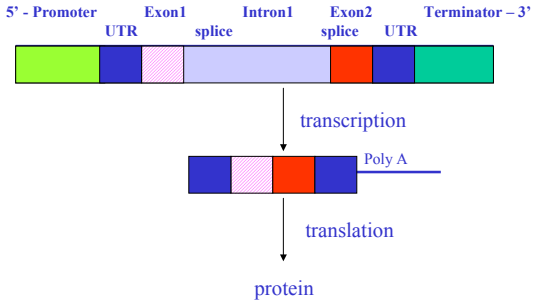
Motivation to find them!

- Need to know all the ***letters*** in order to understand the ***language*** of the cell.
- Experiments are expensive and time consuming.
- Computational techniques promise to leverage existing knowledge about genes to predict new genes at a fraction of the cost and time.
- The sequencing of numerous genomes (100 microbial and several eukaryotic) provides us the raw information to tackle this task.

Transcription and Translation



Eukaryotic Gene Structure



Gene finding approaches

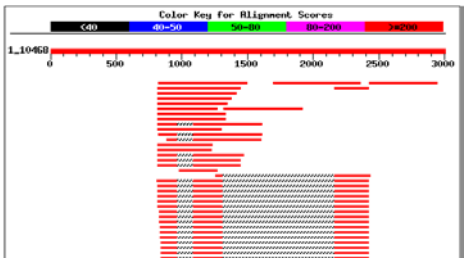
- Find genes by matching to known genes or proteins.
- Use information about transcribed elements in cDNA and EST databases.
- Model the various start and stop signals recognized by the transcription machinery to detect the signals in the genome.
- Study the difference in composition of intergenic regions and genes to train systems to distinguish them.

Homology

- Protein databases:
  - Good alignment using Smith-Waterman or BLAST can detect putative exons.
  - About 50% of the genes can be detected this way.
  - Problems with partial alignment and UTRs.
- Compare genomes:
  - Assumption that coding regions are more conserved than non-coding regions.
  - Sometimes conservation may not cover the entire exon or extend over to introns as well.

Homology continued …

Transcript databases: Wider coverage and gives hints about alternative splicing. However sometimes gives only partial information and is error prone and noisy.



Intrinsic Approaches

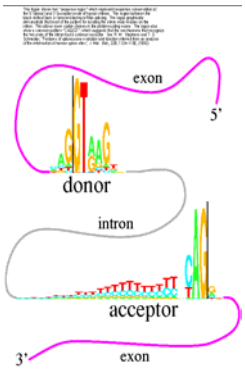
- In prokaryotic sequences it is enough to search for ORFs i.e. dna fragments between a start and a stop codon.
- HMMs are used to exploit differences in GC content, hexamer frequency and base occurrence periodicity.
- Three-periodic Markov models of order five work well for exons.
- Separate markov models for introns, UTRs and intergenic regions helps to boost performance.

Signal Sensors

- To detect promoter regions, splice sites, translation starts and stops.
- Can be used where homology based approaches fail.
- Usually the signals are weak and can only help in refining predictions from other methods.

Signal Sensors continued …

- Usually done by aligning "known" fragments and modeling them with
  - Positional Weight Matrices
  - Hidden Markov Models



Combining Them

- Homology based methods define boundaries poorly. Combining them with signal sensors improves predictions.
- Spliced alignment: local alignment of putative exons with gap opening and closing determined by the presence of splice signals.
- Intrinsic information about the putative exons as well as global information about their location can also be integrated into this setup.
- A HMM framework can also be used to combine various sources of information.

Pitfalls

- Long genes increase the complexity of the search process.
- Long introns weaken the assumptions made by alignment algorithms.
- Short exons are hard to detect.
- Presence of overlapping genes, non-canonical splice sites and alternative start sites.

Future directions

- Expert systems that combine predictions from various methods: biologically there is no single model for a gene.
- Signal sensors that can handle non-canonical cases.
- Improved identification of promoter sequences.
- Identification of functional RNA genes.

Credits

- *Current methods of gene prediction, their strengths and weaknesses,* Catherine Mathe et al., Nucleic Acids Research, 2002.
- *DNA Sequence Analysis* (presentation by Amir Mitchell.)