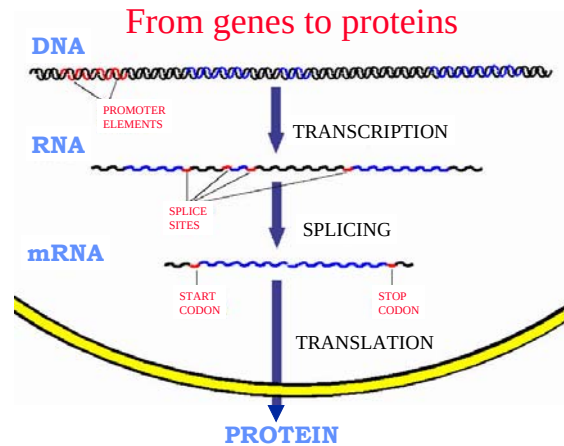


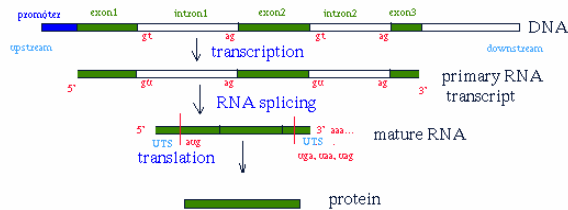
Introduction to Bioinformatics

Iosif Vaisman

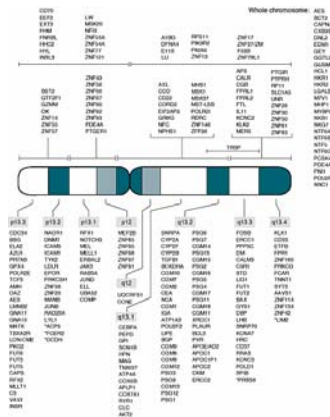
Email: ivaisman@gmu.edu



From genes to proteins



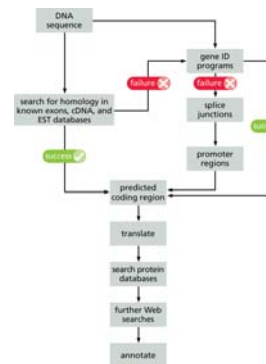
Chromosome 19 gene map



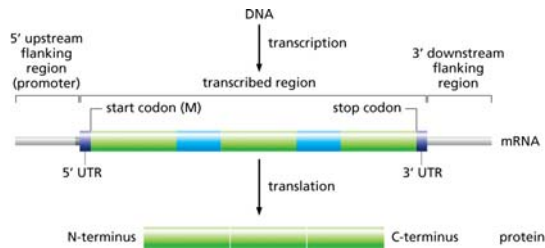
Computational Gene Prediction

- Where the genes are unlikely to be located?
- How do transcription factors know where to bind a region of DNA?
- Where are the transcription, splicing, and translation start and stop signals?
- What does coding region do (and non-coding regions do not)?
- Can we learn from examples?
- Does this sequence look familiar?

Computational Gene Prediction

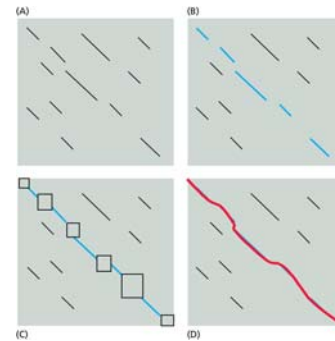


Computational Gene Prediction



Zvelebil & Baum, 2007

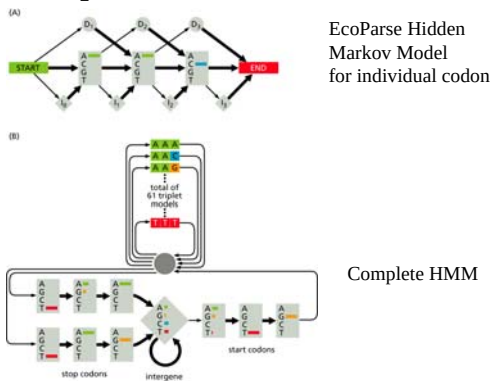
Computational Gene Prediction



LAGAN algorithm

Zvelebil & Baum, 2007

Computational Gene Prediction



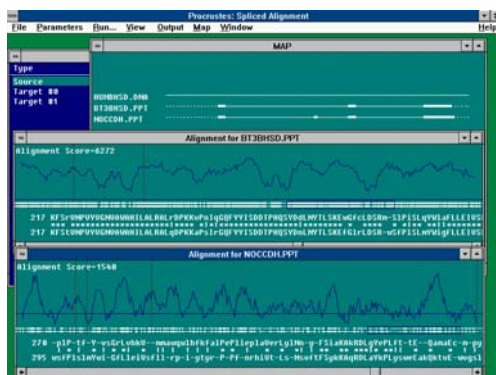
Complete HMM

Zvelebil & Baum, 2007

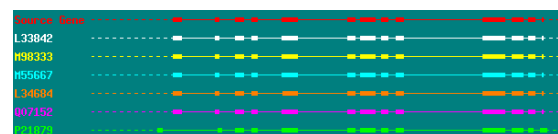
Spliced Alignment (Procrustes)

- New genomic sequence
- Selection of candidate exons
 - AUG --- GU initial exons
 - AG --- GU internal exons
 - AG --- UAA or UAG or UGA terminal exons
- Filtration (based on the codon usage statistics)
- Construction of all possible chains of candidate exons
- Finding a chain with the maximum global similarity to the target protein

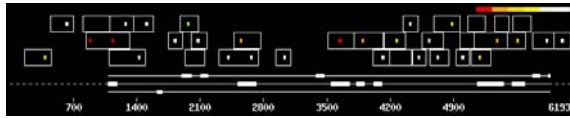
Spliced Alignment (Procrustes)



Predicted Exon Assembly (Procrustes)

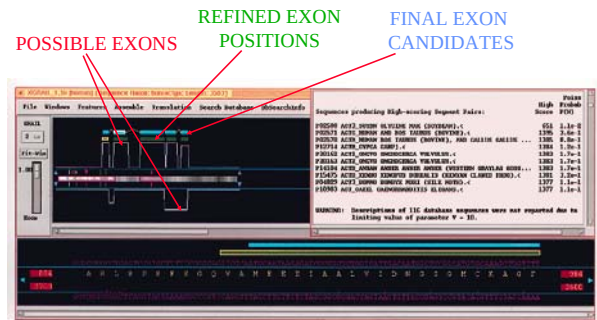


PCR Primers Prediction (GenePrimer)

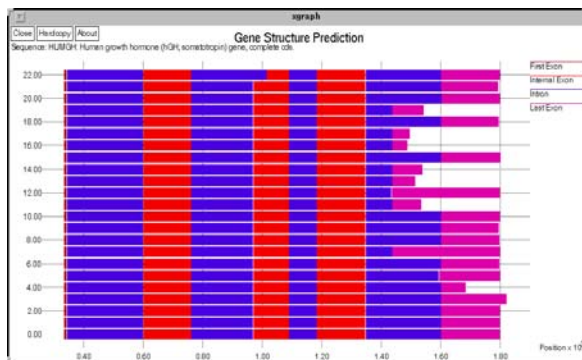


Exon 1085..1182 (98) hit using first 2 primers
 Exon 1628..1676 (49) missed
 Exon 1900..2001 (102) hit using first 8 primers
 Exon 2110..2184 (75) missed
 Exon 2516..2722 (207) hit using first 4 primers
 Exon 3385..3472 (88) missed
 Exon 3546..3746 (201) hit using first primer
 ...

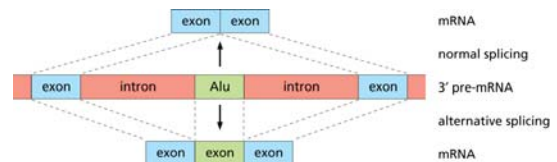
GRAIL gene identification program



Suboptimal Solutions for the Human Growth Hormone Gene (GeneParser)



Transposons



Zvelebil & Baum, 2007

Measures of Prediction Accuracy

Nucleotide Level



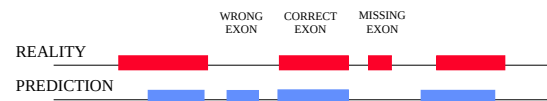
		REALITY	
		c	nc
PREDICTION	c	TP	FP
	nc	FN	TN

$$\text{Sensitivity } S_n = \frac{TP}{(TP + FN)}$$

$$\text{Specificity } S_p = \frac{TN}{(TN + FP)}$$

Measures of Prediction Accuracy

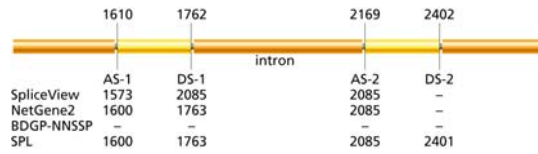
Exon Level



$$\text{Sensitivity } S_n = \frac{\text{number of correct exons}}{\text{number of actual exons}}$$

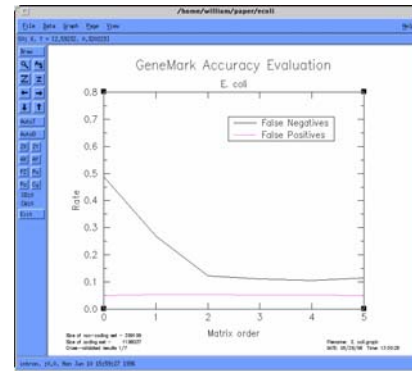
$$\text{Specificity } S_p = \frac{\text{number of correct exons}}{\text{number of predicted exons}}$$

Computational Gene Prediction

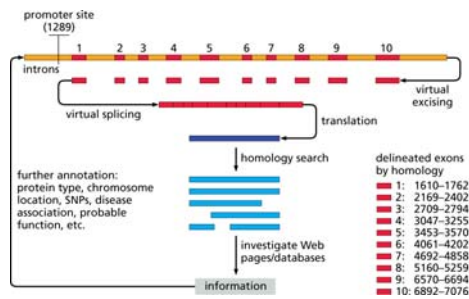


Zvelebil & Baum, 2007

GeneMark Accuracy Evaluation



Gene Annotation



Zvelebil & Baum, 2007