

Research proposal*

Alexandre Tchourbanov
University of Nebraska at Omaha
College of Information Science and Technology
achurbanov@mail.unomaha.edu

5th May 2003

Abstract

While genomes of many organisms were sequenced over the last few years, transforming the sequences into meaningful data remains a difficult task. The most important task is to identify genes and their structure, mainly exonic and intronic structure, CDS, 3'UTR, 5'UTR and splicing patterns. The document describes different approaches to build gene recognition software, discusses quality of available software, surveys literature and proposes further research activity aimed to Ph.D. degree in bioinformatics.

*The presentation is partially based on slides by Lorenzo Cerutti

Presentation structure

Problem introduction and methods

- Importance of the problem
- Problem definition
- Homology based methods
- Ab initio methods
- Signal sensors
- Integrating signal and compositional information

Plan

- What needs to be done
- Plan of work

What has been done

- Work on BLAST-assisted Spliced alignment
- Work on ESE and ESS splice site prediction enhancement

Genome annotation problem importance

Let us consider the importance of functional genome annotation on example of Severe Acute Respiratory Syndrome (SARS) genome.

The SARS virus genome [7] is 29,727 base pairs long (the largest known) as shown on the next slide.

The annotation helps to:

- Identify the organism as a novel coronavirus
- Understand the structure and functionality
 - ▷ *Find instructions for virulent protein production*
 - ▷ *Find the genes enabling the virus to infect cells and to reproduce*
- Compare the virus to other members of the family
- Identify the essential differences between the strains
- Design disease detection methods based on genomic tests
- Design vaccine treatment

The SARS genome annotation (example)

16

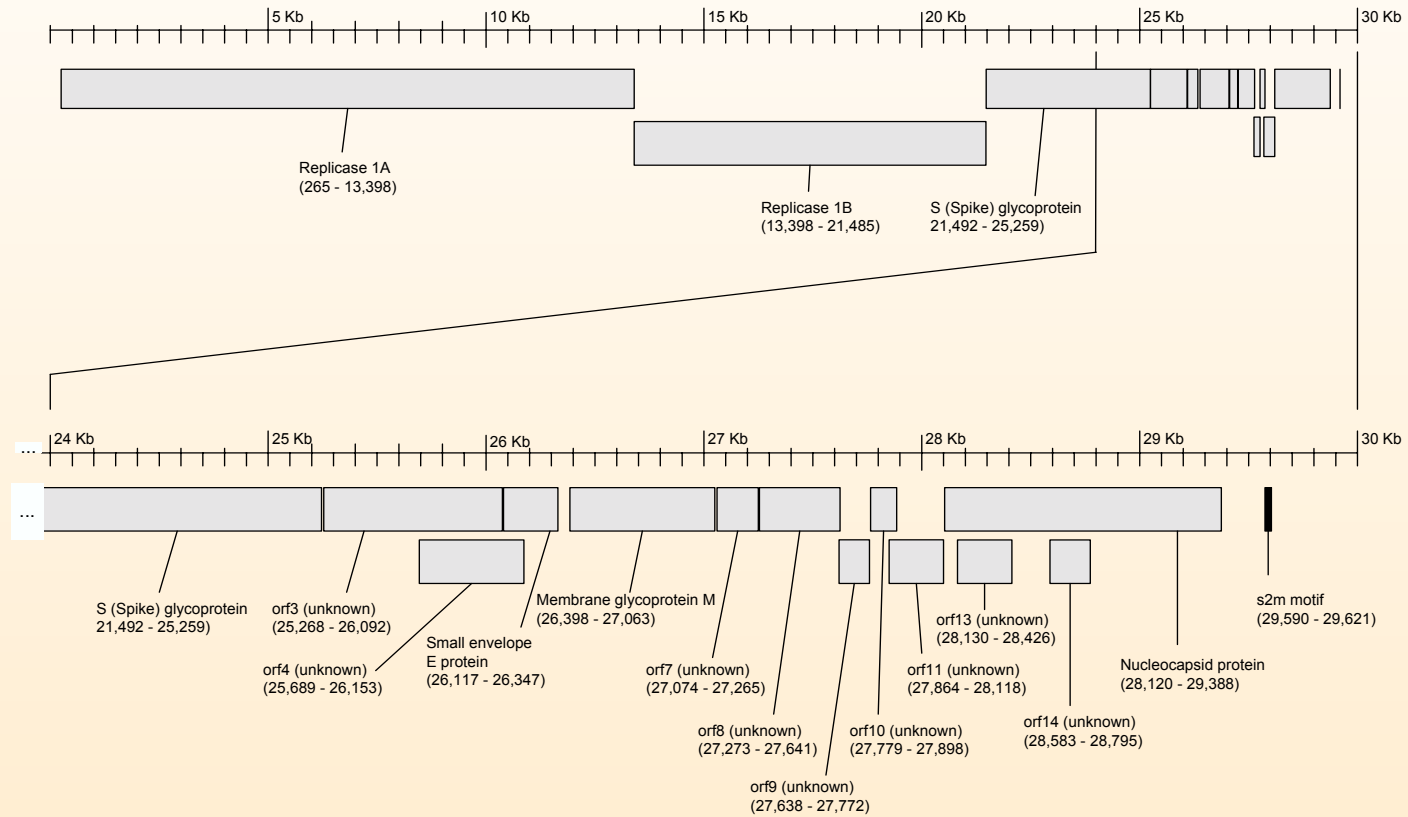


Figure 2.

Gene finding and annotation problem

Gene finding in **eukaryotes** is difficult

- DNA sequence signals have low information content (degenerate and highly unspecific)
- It is difficult to discriminate real signals from numerous copy regions
- Sequencing errors are still frequent

There are two methods available - **homology** and **ab initio**

Homology method

- Gene structure can be deduced by homology, i.e. alignment + signal information
- Requires a not too distant homologous sequence

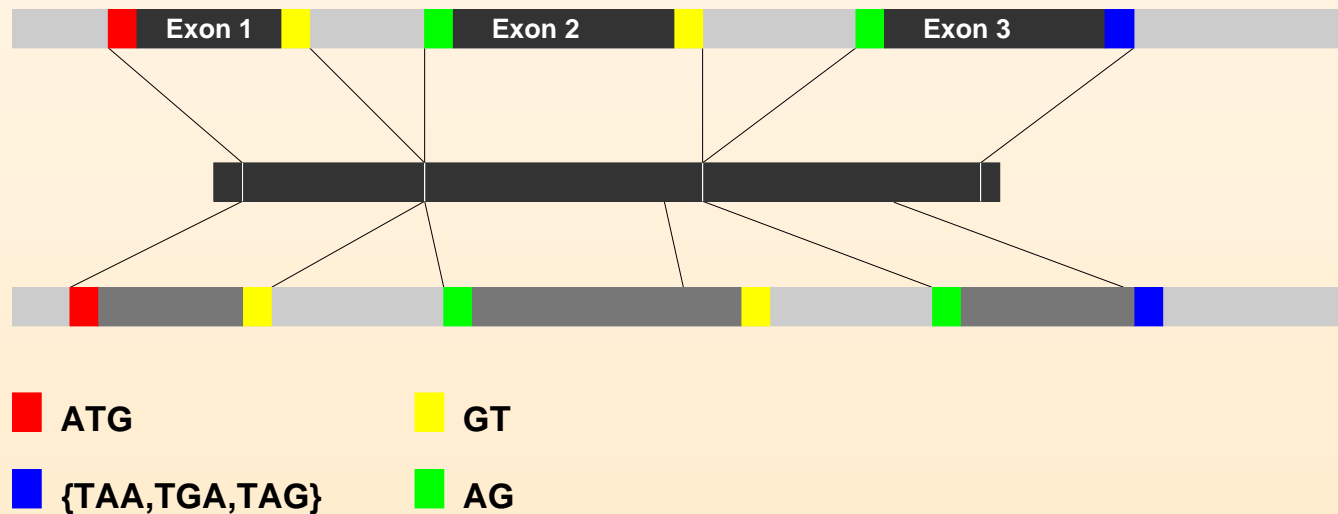
Ab initio method

- Requires two types of information
 - ▷ *compositional statistics information*
 - ▷ *signal information*

Homology method

Principles of the homology method

- Coding regions evolve slower than non-coding regions, i.e. local sequence similarity can be used as a gene finder
- Homologous sequences reflect a common evolutionary origin and possibly a common gene
- Standard homology search methods can be used (BLAST, Smith-Waterman, ...)
- Include "gene syntax" information (start/stop codons, ...)



Homology methods are also useful to confirm predictions inferred by other methods

Homology method (2)

Three types of homology used [8]

- DNA-DNA homology
- DNA-cDNA homology
- DNA-Protein homology

Advantages of the homology methods

- Successfully recognize short exons and exons with unusual codon usage
- Robustly deal with big introns
- May be used to connect several DNA sequences together [10]
- Usually obtain prediction of a high quality (more than 95% for DNA-cDNA homology)
- Assemble correctly complex genes (> 10 exons)

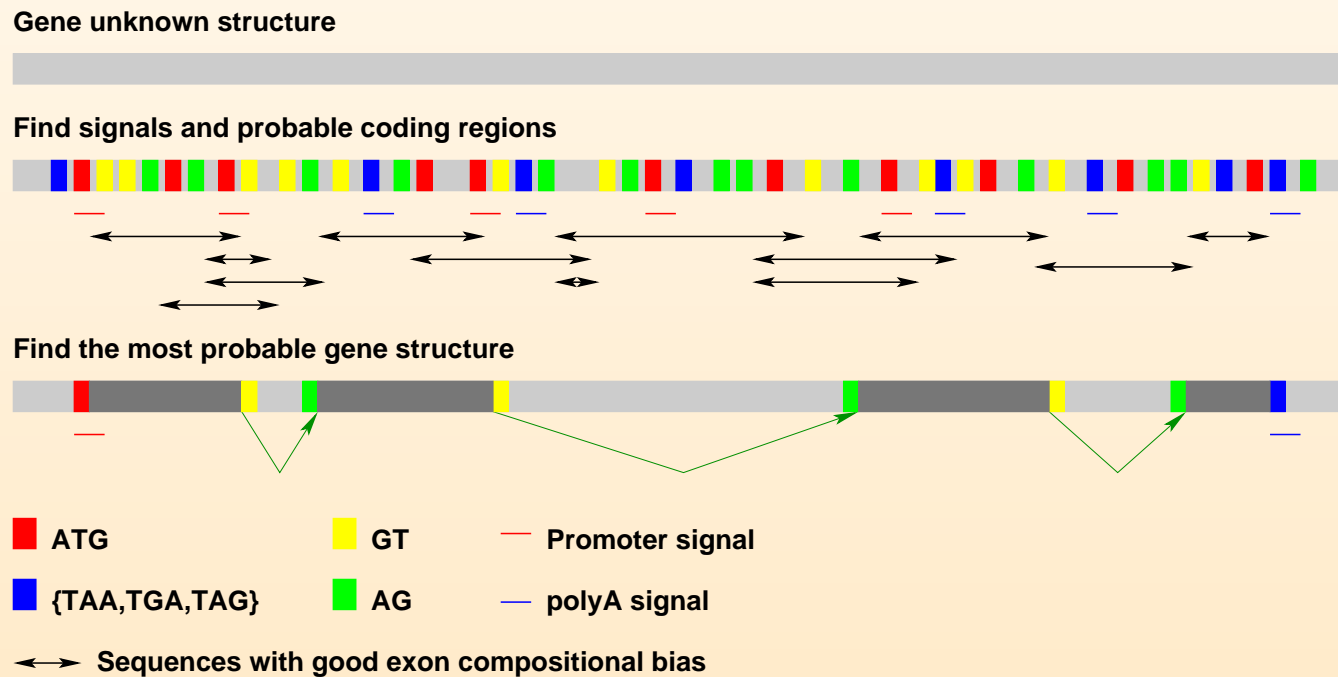
Problems of the homology methods

- Genes without homologous in the databases are missed (approximately 50%)
- Require close homolog to deduce gene structure
- Very sensitive to frame shift errors (in case of DNA-Protein homology)
- Methods are costly in terms of computing time

Ab initio methods

Principles of the ab initio methods

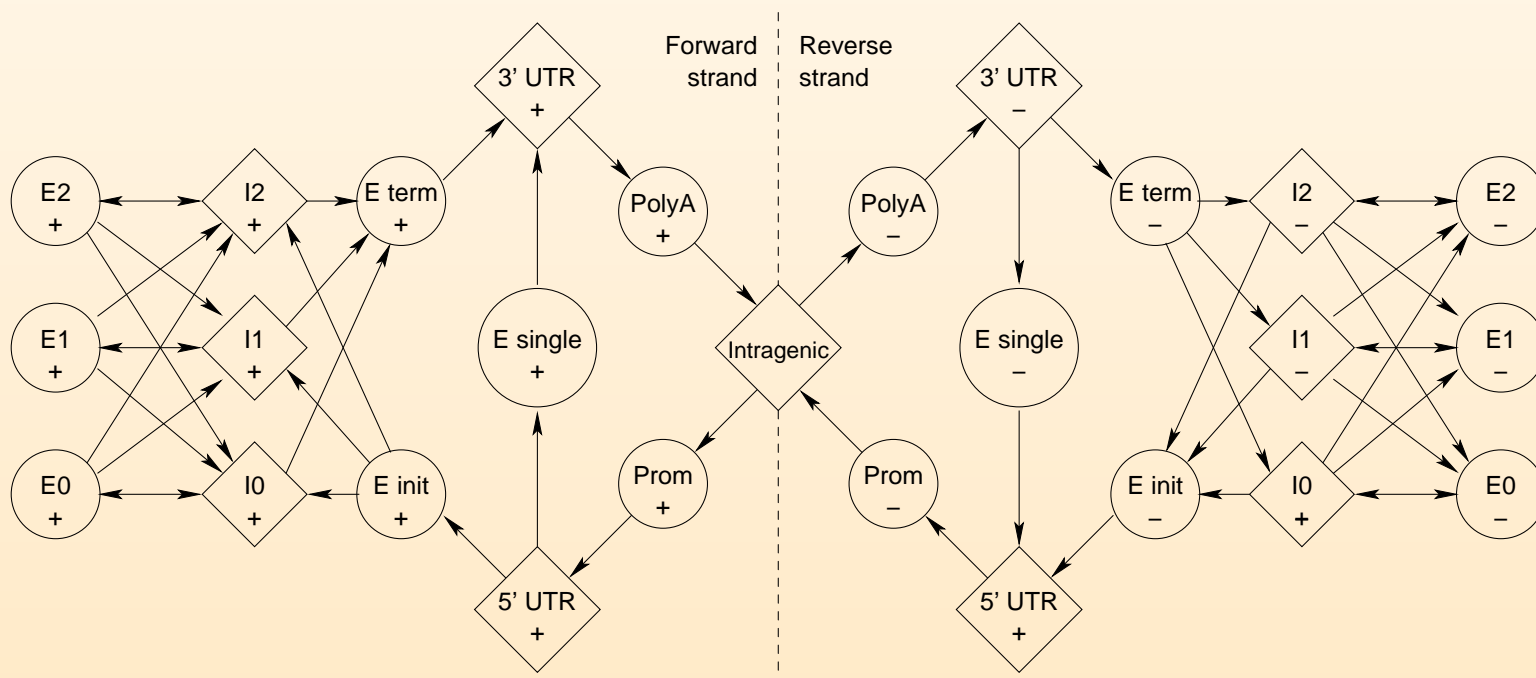
- Integration of signal detection and coding statistics
- Signal detection and coding statistics are deduced from a training set
- Probabilistic frameworks are used to infer a probable gene structure
- A solid scoring system can be used to evaluate the predictions



Integrating signal and compositional information for gene structure prediction

Example of **Generalized Hidden Markov Model** (GHMM) used in GenScan [1]

- Integration of signal detection and coding statistics
- Signal detection and coding statistics are deduced from a training set
- Probabilistic frameworks are used to infer a probable gene structure
- A solid scoring system can be used to evaluate the predictions



Signal detection

Signal detection problem

- DNA sequence signals have low information content
- Signals are highly unspecific and degenerated
- Difficult to distinguish between true and false positive

How improve signal detection

- Take context into consideration (ex. acceptor site must be flanked by an intron and an exon)
- Combine with coding statistics (compositional bias)

Signal detection(2)

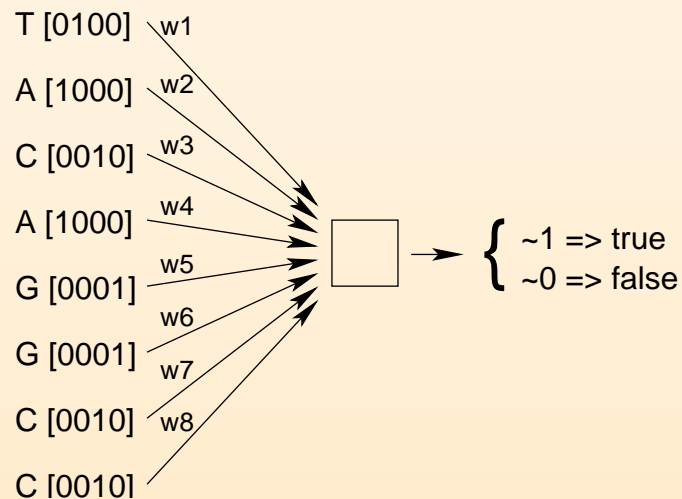
Sensor detects short DNA motifs (promoters, start/stop codons, splice sites, ...).

A number of methods are used in sensors for signal detection

- Consensus string
 - ▷ *Based on most frequently observed residues at a given position*
- Pattern recognition
 - ▷ *Flexible consensus strings*
- Weight matrices
 - ▷ *Based on observed frequencies of residues at a given position. Uses standard alignment algorithms. This method returns a score.*
- Weight array matrices
 - ▷ *Weight matrices based on dinucleotides frequencies. Takes into account the non-independence of adjacent positions in the sites.*
- Maximal dependence decomposition (MDD)
 - ▷ *MDD generates a bifurcation decision tree based on χ^2 statistics, which captures significant dependencies between non-adjacent as well as adjacent positions.*

Signal detection(3)

- Hidden Markov Models (HMM)
 - ▷ *HMM uses a probabilistic framework to infer the probability that a sequence correspond to a real signal*
- Neural Networks (NN)
 - ▷ *NN are trained with positive and negatives example and "discover" the features that distinguish the two sets.*
 - ▷ *Example: NN for acceptor sites, the perceptron [5]*



Coding statistics

Inter-genic regions, introns, exons, ... have different nucleotides contents.

This compositional differences can be used to infer gene structure.

Examples of coding region finding methods:

- ORF length
 - ▷ *Assuming an uniform random distribution, stop codons are present every $\frac{64}{3}$ codons (≈ 21 codons) in average*
 - ▷ *In coding regions stop codon average decreases*
 - ▷ *Method sensitive to frame shift errors*
 - ▷ *Can't detect short coding regions*
- Bias in nucleotide content in coding regions
 - ▷ *Generally coding regions are G+C rich*
 - ▷ *There are exceptions. For example coding regions of *P. falciparum* are A+T rich*
- Periodicity
 - ▷ *Plot of the number of nucleotides separating pairs of T is periodic in coding regions, but not in non-coding regions*

Coding statistics(2)

Codon frequencies

- Synonym codon usage is biased in a species dependent way
- 3rd codon position: 90% are A/T; 10% are G/C

How to calculate codon frequencies

- Assume $S = a_1b_1c_1, a_2b_2c_2, \dots, a_{n+1}b_{n+1}c_{n+1}$ is a coding sequence with unknown reading frame. Let f_{abc} denote the appearance frequency of codon abc in a coding sequence. The probabilities p_1, p_2, p_3 of observing the sequence of n codons in the 1st, 2nd and 3rd frame respectively are:

$$p_1 = f_{a_1b_1c_1} \times f_{a_2b_2c_2} \times \dots \times f_{a_nb_nc_n}$$

$$p_2 = f_{b_1c_1a_2} \times f_{b_2c_2a_3} \times \dots \times f_{b_nc_n a_{n+1}}$$

$$p_3 = f_{c_1a_2b_2} \times f_{c_2a_3b_3} \times \dots \times f_{c_n a_{n+1} b_{n+1}}$$

The probability P_i of the i th reading frame for being the coding region is:

$$P_i = \frac{p_i}{p_1 + p_2 + p_3}$$

where $i \in \{1, 2, 3\}$.

Coding Statistics(3)

In practice we use these computations in a search algorithm as follows:

- Select a window of size n (for example $n = 30$)
- Slide the window along the sequence and calculate P_i for each start position of the window

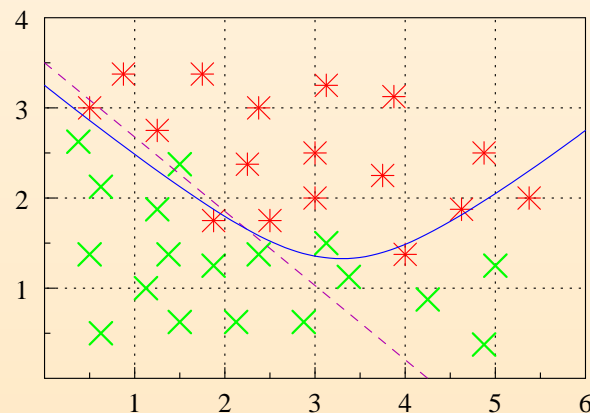
A variation of the codon frequency method is to use 6-tuple frequencies instead of 3-tuple (codon) frequencies. This method was found to be the best single property to predict whether a window of vertebrate genomic sequence was coding or non-coding [3].

The usage of hexamers frequencies has been integrated in a number of gene predictors.

Integrating signal information and compositional information for gene structure prediction

A number of methods exists for gene structure prediction which integrate different techniques to detect signals (splicing sites, promoters, etc.) and coding statistics

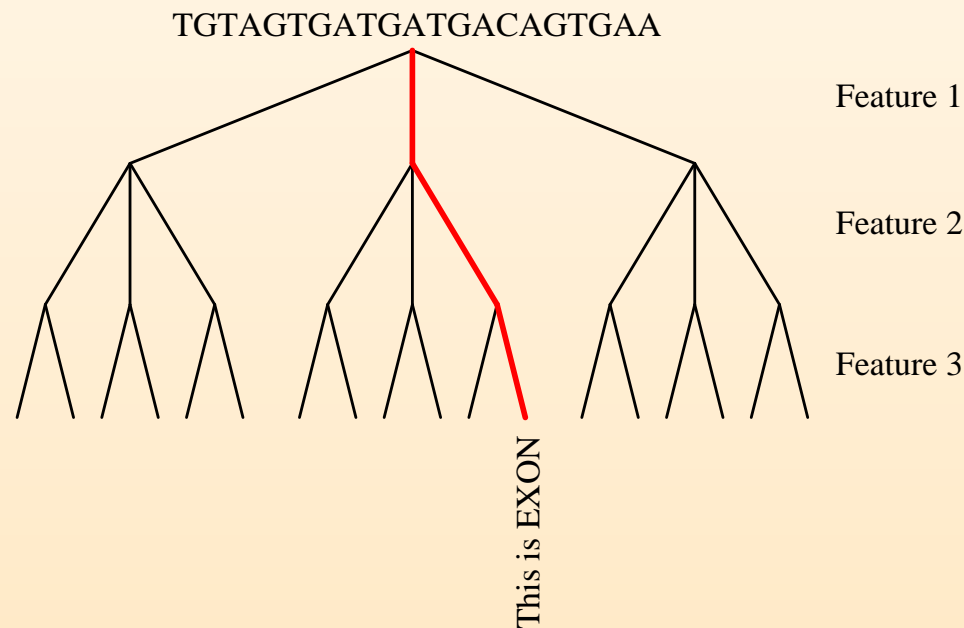
- Linear and quadratic discrimination analysis
 - ▷ *Linear discrimination analysis is a standard technique in multivariate analysis*
 - ▷ *Linear discrimination analysis is used to linearly combine several measures in order to perform the best discrimination between coding and non-coding sequences.*
 - ▷ *Quadratic discriminant analysis. Similar to linear discrimination analysis, but uses a quadratic discriminant function*
 - ▷ *Dynamic programming is used in to combine the inferred exons*



Decision tree

- Decision tree

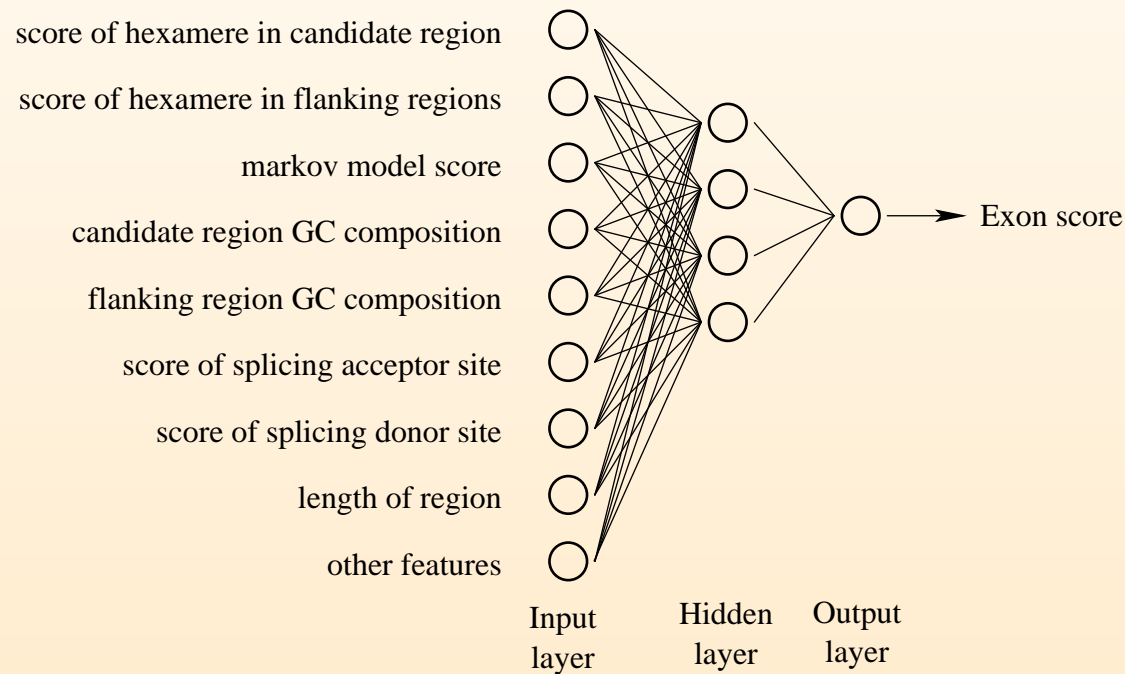
- ▷ *Internal nodes of a decision tree are property values tested for each subsequence passed to the tree*
- ▷ *Properties can be various coding measures (e.g. hexamer frequencies) or signal strengths*
- ▷ *Bottom nodes (leaves) of the tree contains class labels to be associated with the subsequences*
- ▷ *Dynamic programming is used to deduce the complete gene structure*



Neural Network

- Neural network

- ▷ *The neural network is trained with a set of true positives and true negatives examples*
- ▷ *For each training example, the neurons are tuned to return the right answer*
- ▷ *Dynamic programming is used to deduce the complete gene structure [11]*

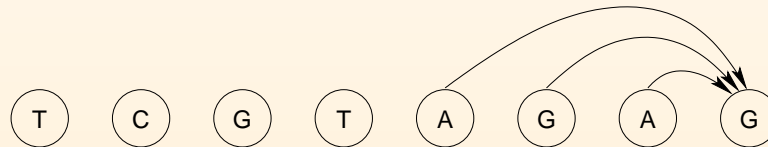


Markov model

- **Markov Model (MM)**

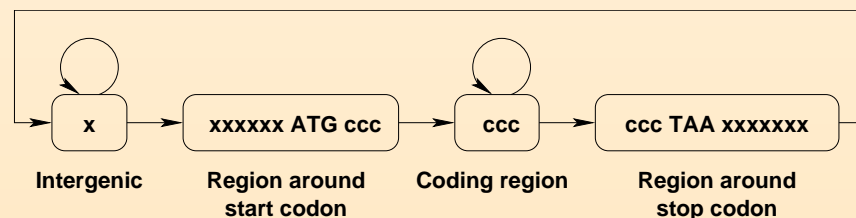
- ▶ *Biological sequences can be modelled as the output of a stochastic process in which the probability for a given nucleotide to occur at position p depends on the k previous positions. This representation is called k -order Markov Model.*

$$P(x_i | x_1, x_2, \dots, x_i) = P(x_i | x_{i-k}, x_{i-(k-1)}, \dots, x_i)$$



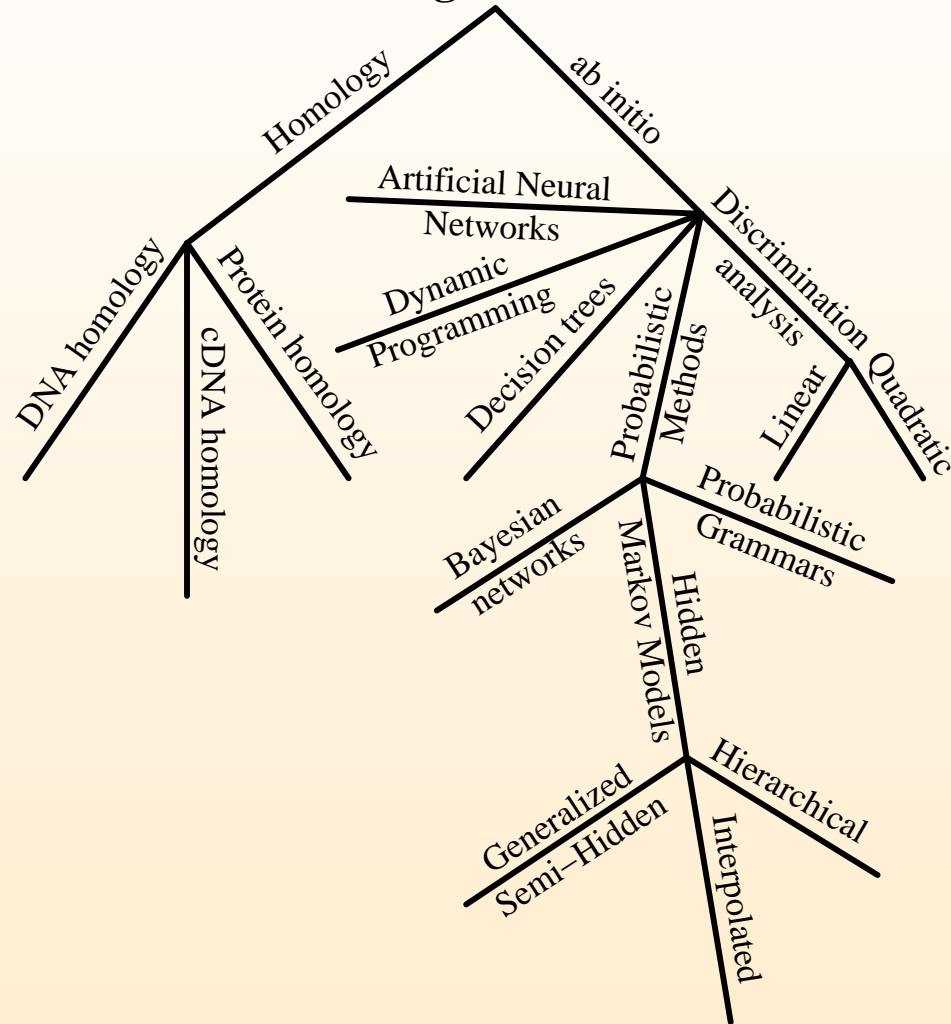
- **Hidden Markov Model (HMM)**

- ▶ *In a HMM the biological sequences are modelled as the output of a stochastic process that progresses through a series of discrete states. Each state model correspond to a Markov Model [6]*



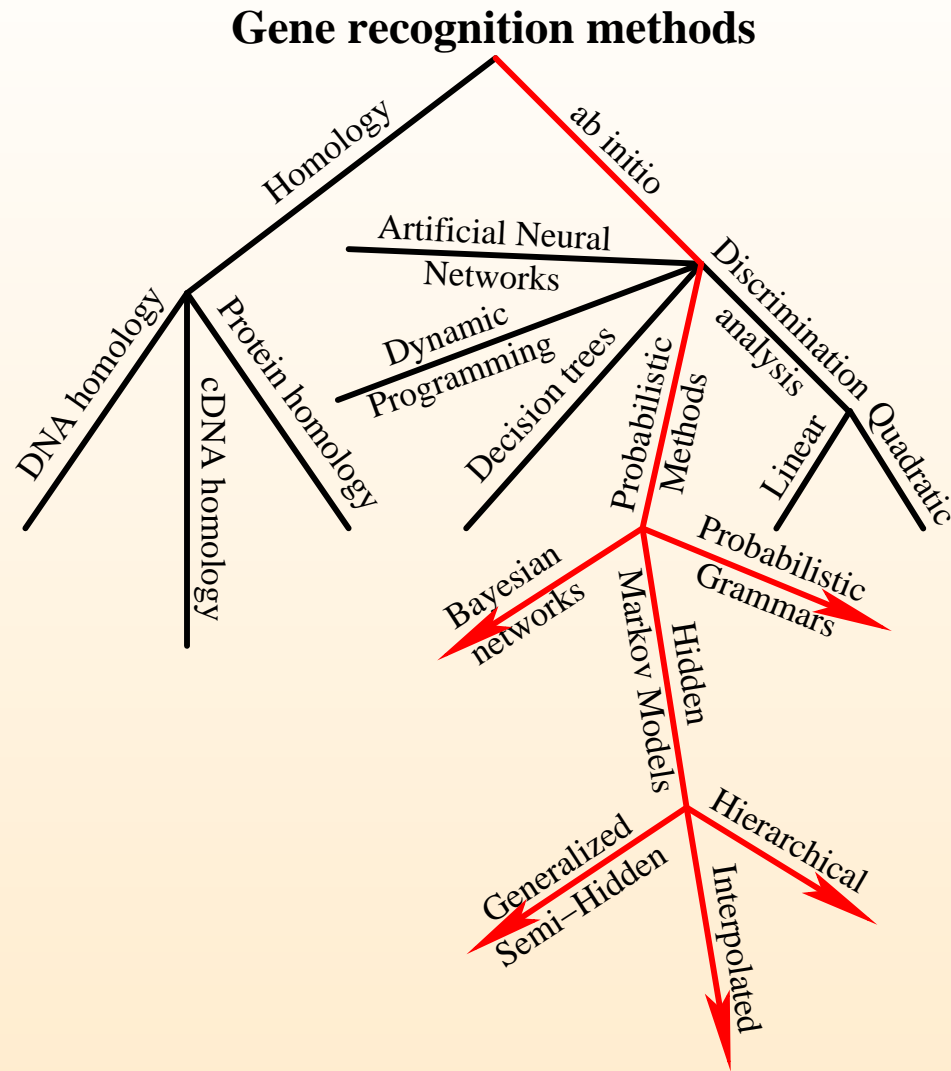
Ab initio Methods available

Gene recognition methods



The structure of methods available for gene structure annotation and recognition

The proposed class of methods to use



I use probabilistic methods based on Bayes' rule

What needs to be done

The prediction of protein encoding genes structure in higher **eukaryote** organisms needs significant improvement to gain industrial strength.

The general idea of the work proposed is simple - write a program to generate precise gene annotation for **human** and other organisms of similar evolutionary complexity.

The annotation should include precise prediction of all exon boundaries, information on alternative transcripts as well as numerous factors binding sites [4]. Other factor may also be present in annotation once they prove to be useful for the main purpose.

The program won't work as a scanner of a huge genomic sequences, it will rather emphasize on precision of a prediction on pre-mRNA structure.

The steps

Build data set of the whole human genes available at a given moment in RefSeq and periodically update it.

Study the signals available for the sequences through the biological literature.

Design a computational method to predict the combination of signals resulting into alternative splicing behavior. The method should highlight the signals characterizing the splice sites and other structural elements. Information theory methods could be useful at this stage.

Collect information on all probabilistic models available nowadays for sequence analysis, analyze them and extend the hypotheses to make them useful.

Based upon signals, their correlation and models discovered, build **probabilistic model** of the splicing mechanism.

Compare the results on a standard test set used to estimate other ab initio approach annotation engines [2].

The schedule

The analysis of spicing mechanism from biological literature, running and testing certain signal models, identifying correlation between the signals will take the Spring Semester 2003.

In the summer I start analyzing models available to combine the evidences from the first stage. In parallel I start building software implementing the probabilistic model.

In the fall/winter 2003-2004 the program will be finished and thoroughly tested.

In the Spring 2004 the dissertation will be written and defended at the end of summer 2004.

Spliced alignment package GIGOgene 1.0

The program combining BLAST search with spliced alignment was implemented.

The program runs on Human genome database in a batch mode.

The Spliced alignment algorithm was designed and implemented, as shown on Figure 1.

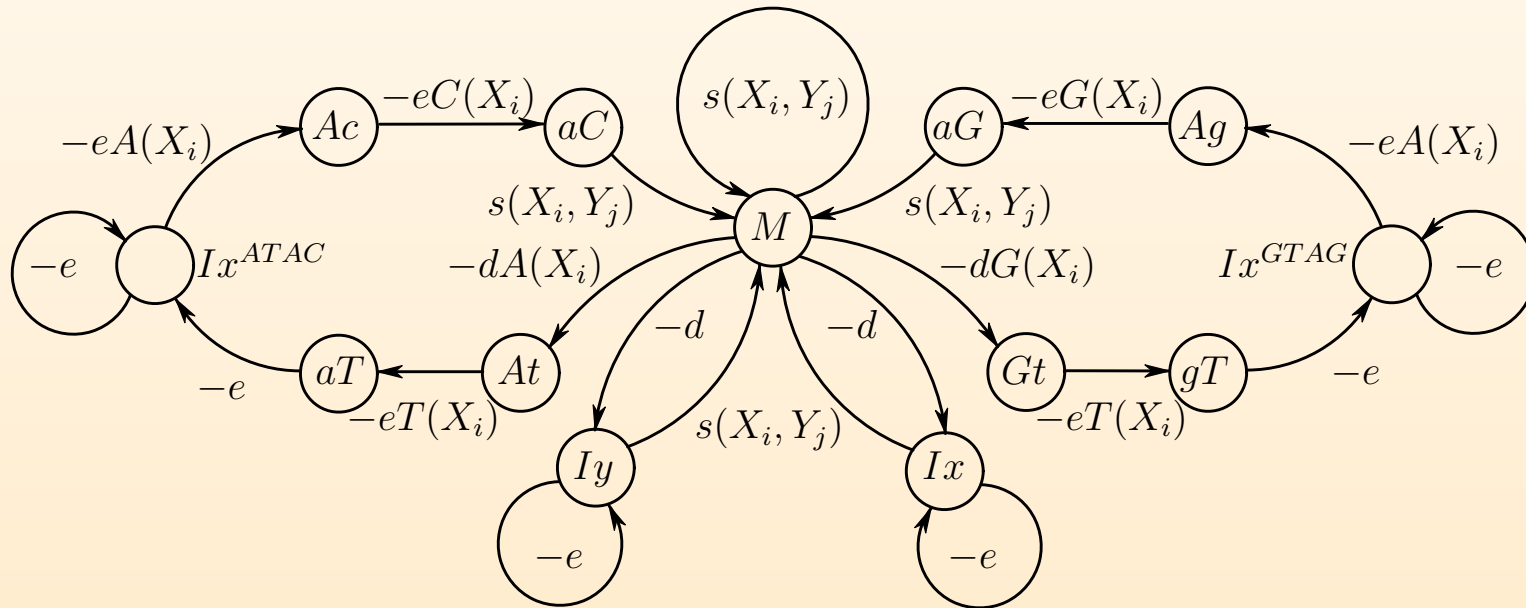


Figure 1: State diagram of our spliced alignment algorithm

Test run results

The program was tested on GENIE gene finding data set available at <http://www.fruitfly.org/sequence/human-datasets.html/>.

We estimated the **sensitivity** (ESn) and **specificity** (ESp) first according to the formulas

$$ESn = \frac{TE}{AE} \quad ESp = \frac{TE}{PE}$$

where

- TE - the number of exactly predicted exons (true exons),
- AE - number of annotated exons,
- PE - number of predicted exons.

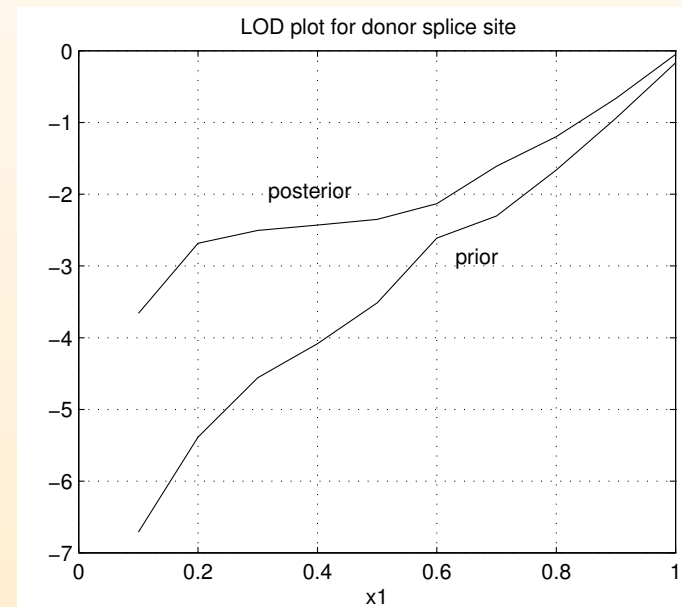
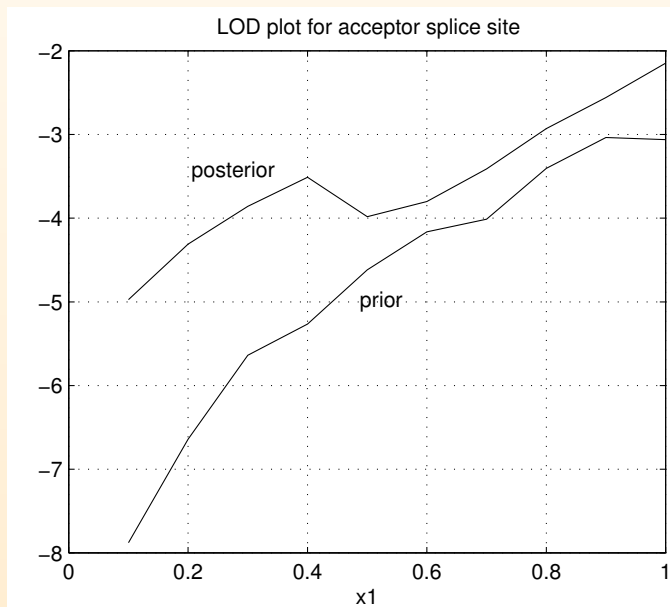
According to our test runs for the first 200 genes from the test set, the prediction quality was

$$ESn = 0.97 \quad ESp = 0.97$$

ESE and ESS prediction quality improvement

We implemented software to increase quality of splice sites prediction [9].

The diagram below shows quality improvement in terms of **Logarithm of Odds (LOD)** for donor and acceptor splice sites sites.



Visualization of the ESE and ESS

The ESE and ESS locations need to be correlated with the splice sites locations.

The visualization tool allows seeing certain dependencies between splice sites and ESE/ESS locations, as shown on Figure 2.

GIGOGene 1.0 predicted gene structure for AL359749.7

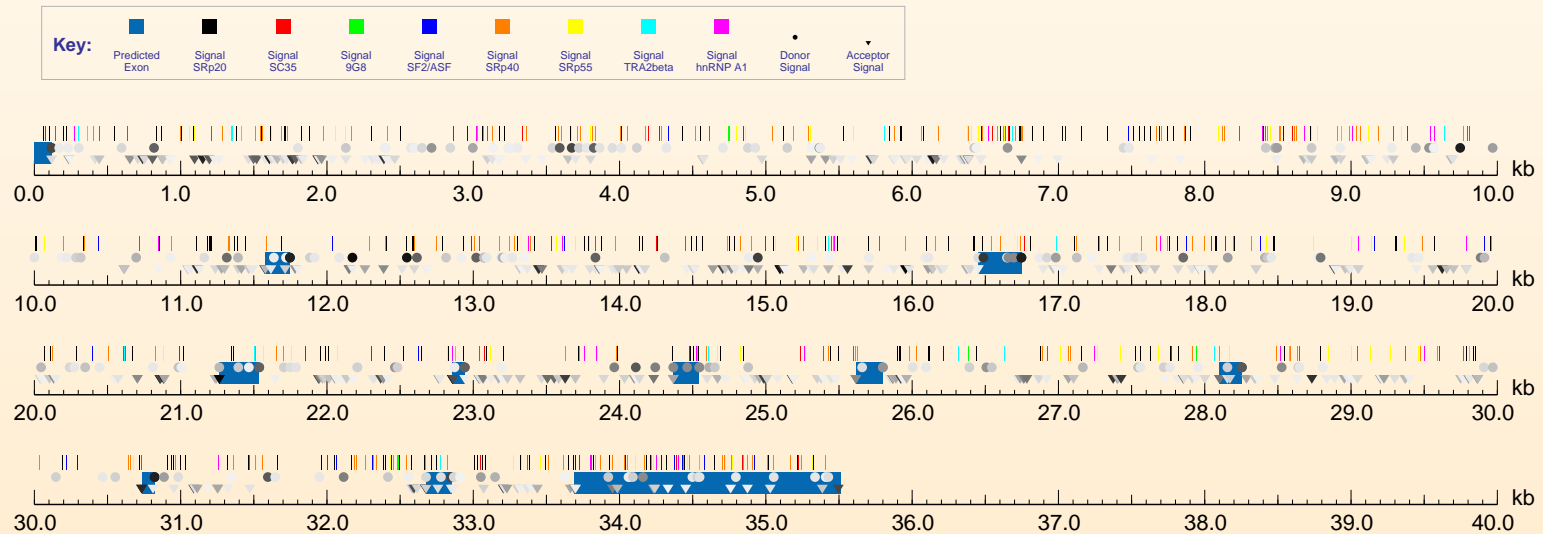


Figure 2: ESE/ESS locations visualization

References

- [1] Christopher B. Burge and Samuel Karlin, *Predictions of complete gene structures in human genomic dna*, *Journal of Molecular Biology* **268** (1997), 78–94.
- [2] Moisés Burset and Roderic Guigó, *Evaluation of gene structure prediction programs*, *Genomics* **34** (1996), 353–367.
- [3] J.M. Claverie and L. Bougueleret, *Heuristic informational analysis of sequences*, *Nucleic Acids Research* **14** (1986), 179–196.
- [4] William G. Fairbrother, Ru-Fang Yeh, Phillip A. Sharp, and Christopher B. Burge, *Predictive identification of exonic splicing enhancers in human genes*, *Science* **297** (2002), 1007–1013.
- [5] P. Horton and M. Kanehisa, *An assessment of neural network and statistical approaches for prediction of e.coli promoter sites*, *Nucleic Acid Research* **20** (1992), 4331–4338.

- [6] A. Krogh, S. L. Salzberg, D. B. Searls, and S. Kasif, *Computational methods in molecular biology: An introduction to hidden markov models for biological sequences*, ch. 4, pp. 45–63, Elsevier, Amsterdam, 1998.
- [7] Marco A. et al. Marra, *The genome sequence of the sars-associated coronavirus*, *Science Express* (2003), 1–13.
- [8] Catherine Mathé, Marie-France Sagot, Thomas Schiex, and Pierre Rouzé, *Current methods of gene prediction, their strengths and weaknesses*, *Nucleic Acids research* **30** (2002), 4103–4117.
- [9] Alexandre Tchourbanov, Hesham Ali, and Jitender Deogun, *Using bayesian network approach for splice sites recognition enhancement*, Available at: <http://csce.unl.edu/~tchourba>, Mar. 2003.
- [10] Alexandre Tchourbanov, Daniel Quest, Hesham Ali, Mark Pauley, and Robert Norgren, *A new approach for gene annotation using unambiguous sequence joint*, Available at: <http://csce.unl.edu/~tchourba>, Apr. 2003.
- [11] Xu Ying, Richard J. Mural, J. R. Einstein, Manesh Shah, and E. C. Uberbacher,

Grail: A multi-agent neural network system for gene identification, Proceedings of The IEEE **84** (1996), no. 10, 1544–1552.