Bioinformatics

Finding Genes in DNA with a Hidden Markov Model

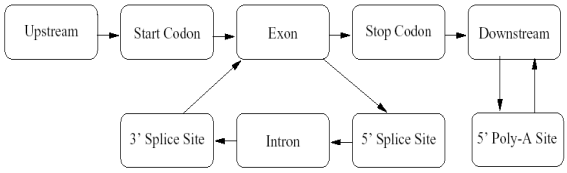
John Henderson* Steven Salzberg† Kenneth H. Fasman‡

Sami Khuri
Department of Computer Science
San José State University
San José, California, USA
khuri@cs.sjsu.edu
www.cs.sjsu.edu/faculty/khuri

VEIL

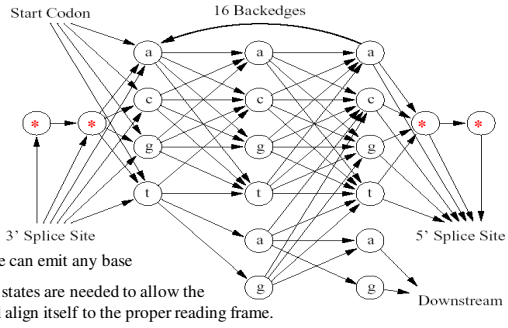
- **VEIL:** the **V**iterbi **E**xon-**I**ntron **L**ocator was developed by Henderson, et al. at Johns Hopkins University.
- **VEIL** has a modular structure:
  - It uses a HMM made up of sub-HMMs to describe different parts of the sequence:
    - exon, intron, start, stop, splice, upstream, etc..
- **VEIL** assumes test data starts and ends with noncoding DNA and contains exactly one gene.

VEIL: The Combined Model



The start codon model is very simple:

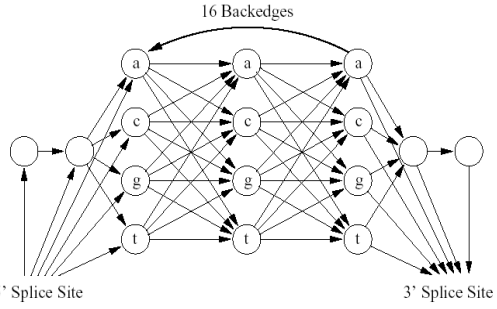
VEIL: Exon Model



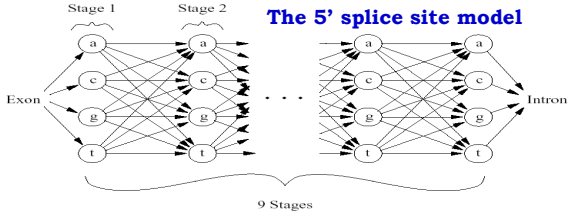
* State can emit any base

The 4 states are needed to allow the model align itself to the proper reading frame.

VEIL: Intron Model

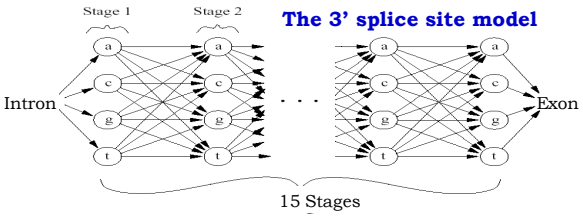


VEIL: Donor Model



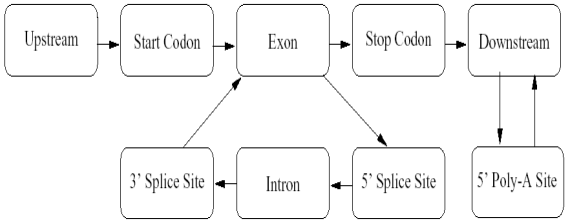
The **donor model** (5' splice site) has 9 stages because it is based on the length of the consensus sequence for donors which is of length 9.
Stages 1 to 3 are part of the exon and 4 to 9 of the intron.

VEIL: Acceptor Model



The **acceptor model** (3' splice site) has 15 stages because it is based on the length of the consensus sequence for acceptors which is of length 15. Stages 1 to 14 are part of the intron and 15 of the exon.

VEIL: Transitions and States



The **VEIL** hidden Markov model has 1003 edges and 241 states

VEIL: Preparing the Training Data

- They used a set of 570 vertebrate sequences from different species each containing exactly one gene.
- Quality control step: filter the data by removing pseudogenes, entries with no introns (from cDNA), entries with non-standard splice junctions (introns that do not begin with GT and end with AG).
- The 570 vertebrate sequences contain a total of 2649 exons and 2079 introns.

VEIL: Five-Fold Cross Validation

- A five-fold cross validation experiment was performed to estimate how well the system would perform when tested on data that was not in the training set.
  - The 570 sequences were randomly partitioned into five sets of 114 sequences each.
  - For each partition, the system is trained on 4 sets and tested on the fifth.
  - Combine the results from the five test sets.

Training and testing VEIL

The testing involved calculating sensitivity and specificity of both nucleotide labelling (coding/ noncoding) and exons exactly found.

Sensitivity (Sn) for nucleotides is the percentage of coding nucleotides correctly labeled as coding. Specificity (Sp) is the percentage of nucleotides labeled as coding that were actually coding. P(All) is the overall probability of predicting any base correctly. The right half of the table contains the corresponding values for whole exons; i.e., the accuracy at predicting the coding regions exactly. 1ME is the percentage of exons for which one or both edges was correct, and Ov is the percentage of true exons that overlapped a predicted exon. The Test-All line contains the combined results for all test data.

Partition	Nucleotides				Whole Coding Exons			
	Sn	Sp	CC	P(All)	Sn	Sp	1ME	Ov
Full Vertebrate Data Set								
Train-1	0.82	0.75	0.75	0.93	0.54	0.53	0.73	0.80
Test-1	0.80	0.76	0.74	0.93	0.51	0.49	0.71	0.80
Train-2	0.82	0.74	0.74	0.93	0.53	0.50	0.73	0.80
Test-2	0.80	0.75	0.73	0.93	0.52	0.52	0.70	0.78
Train-3	0.82	0.75	0.74	0.93	0.55	0.52	0.73	0.81
Test-3	0.75	0.70	0.68	0.92	0.45	0.44	0.64	0.72
Train-4	0.82	0.75	0.74	0.93	0.54	0.51	0.73	0.81
Test-4	0.79	0.72	0.71	0.93	0.50	0.46	0.69	0.76
Train-5	0.82	0.73	0.73	0.93	0.54	0.50	0.72	0.80
Test-5	0.87	0.70	0.74	0.92	0.53	0.47	0.77	0.86
Test-All	0.83	0.72	0.73	0.92	0.53	0.49	0.73	0.81