Comparative ab initio prediction of gene structures using pair HMMs

Irmtraud M. Meyer and Richard Durbin

Bioinformatics Vol. 18 no. 10 (2002)
pp. 1309-1318

Introduction

This paper describes Doublescan, a pair HMM approach for gene structure prediction. Doublescan makes simultaneous structure predictions for two homologous nucleotide sequences.

This paper also describes a new HMM-traversal algorithm – the stepping stone algorithm – as a lower-cost alternative to the well known Viterbi algorithm.

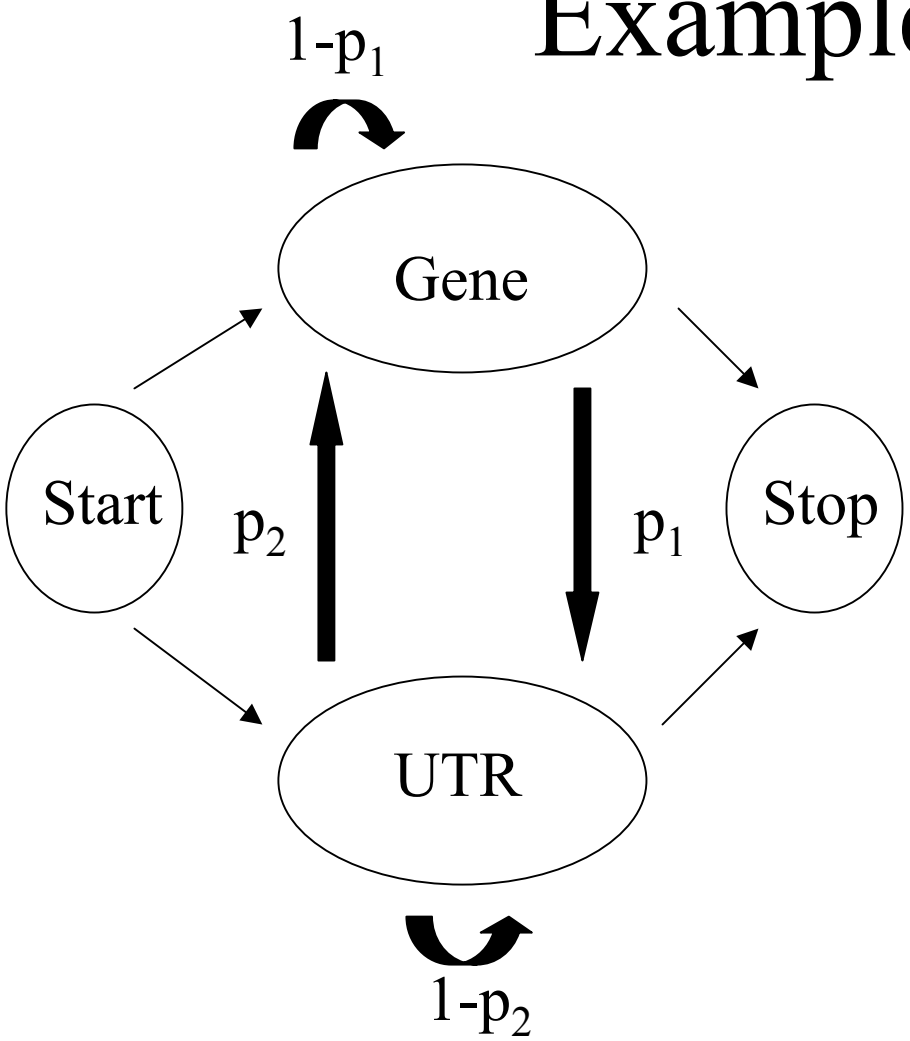
Background

- Gene Structure prediction is an important problem.
- Most attempts to date have focused on predictions on a single nucleotide sequence.
- Little has been done with comparative gene structure prediction – the necessary data has only become available recently (Human Genome project, etc.).

What are Hidden Markov Models?

- Hidden Markov Models (HMMs) are a statistical modeling tool, similar to a finite state machine.
- Each state produces output, which follows a defined probability distribution.
- The transitions from state to state are also defined by probability distributions.
- Several gene prediction algorithms (e.g. Genie) use HMMs. It is easy to make an association between the output of HMM states and a desired nucleotide sequence.

Example HMM



- Each state has transition probabilities p_i and $(1-p_i)$.
- There are also emission probabilities for each state (not shown), reflecting the nucleotide distribution in that region.

HMM Gene Prediction from a Single Nucleotide Sequence

- Genscan [Burge 1997]
- Genie [Kulp 1996]
- HMMgene [Krogh 1997]
- A combination of Genscan and HMMgene [Rogic 2002]

Why Use Multiple Sequences?

- Around 20 years of work has gone into single-sequence prediction.
- Some methods involve pure statistical models, others also attempt to use known protein data.
- Comparative genomics represents potential source of new information for prediction systems.

HMM Gene Prediction involving Multiple Nucleotide Sequences

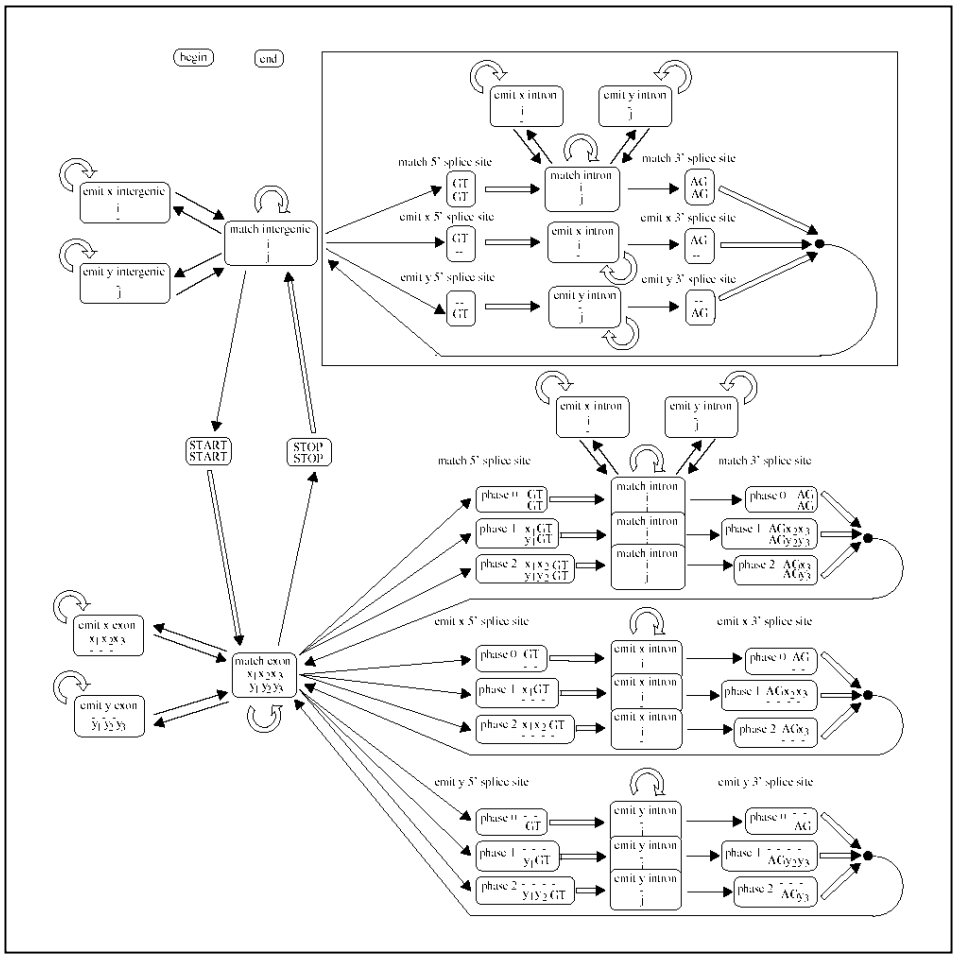
- Twinscan (an extension of Genscan) [Korf 2001]
- Rosetta [Batzoglou 2000]

- Doublescan innovates in having an actual comparative HMM, making simultaneous predictions for each sequence, and retrieving conserved subsequences.

Structure of Doublescan

- The basic structure of Doublescan is an HMM. Each state emits a codon.
- There are separate sections for introns, exons, and intergenics.
- Multiple states to express complexity within these sections (matches, insertions in each sequence, transition states).

The Doublescan HMM



Model Weights

- The emission probabilities were derived from matches between the sequences (specifically, from equal-length orthologous genes in the test set).
- The transition probabilities were estimated, and then tuned by hand for optimum performance with the training set.

The Scoring Algorithm

- The Viterbi algorithm is a well known way for taking sequences and computing the best-scoring path through an HMM.
- The time and space requirements of the Viterbi algorithm were too large for Doublescan.
- As a replacement, the authors wrote the Stepping Stone algorithm – a variant of the Viterbi algorithm that uses alignments to simplify the search.

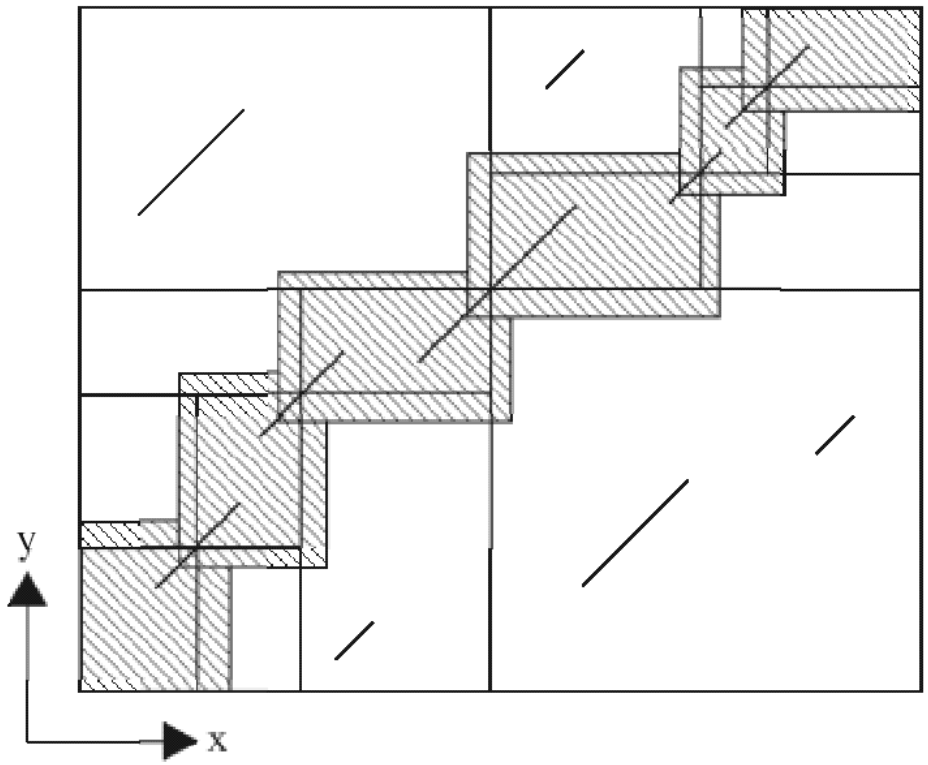
The Problem with Viterbi

- The Viterbi algorithm computes the best path through an HMM by scoring, in parallel, each path through the HMM and keeping track of the best ones.
- This can be memory intensive when it's time to re-construct the path!

The Stepping Stone Algorithm Part 1

- The two sequences are aligned with BLASTN.
- We do the following to each locally aligned portion, in order of score.
  - Find the midpoint of the alignment
  - Attempt to add the midpoint to a list of constraints
- At the end of this loop, we have a set of nucleotide pairing constraints for the path through the HMM.

A Visual Example



Diagonal lines represent local alignments. The hashed portion is the area defined by the constraints.

The Stepping Stone Algorithm Part 2

- For each pair of adjacent constraints, find the best scoring path between them with a variant of the Viterbi algorithm.
- Reconstruct the complete path through the HMM, and find its score.

Other Optimization of Doublescan

- "Doublescan including UTR-splicing still has a 14% rate or wrong genes corresponding to 30 genes which are predicted in addition to those that overlap the annotated gene in each DNA sequence" [Meyer 2002, pp 1313-1314].
- The authors were able to raise the specificity of Doublescan substantially by removing "all predicted genes with introns of less than or equal to 50 base pairs length and or a total coding length of less than or equal to 120 base pairs length" [Meyer 2002, pp 1317].

Results and Comparison with Genscan – Key Figures

	Doublescan	Genscan
Gene		
Sensitivity	0.57	0.47
Specificity	0.50	0.46
Start Codon		
Sensitivity	0.75	0.73
Specificity	0.78	0.91
Stop Codon		
Sensitivity	0.89	0.88
Specificity	0.86	0.97

Results and Comparison with Genscan - Analysis

- Overall, Doublescan gets better results at low resolution, but the lack of model detail restricts it's accuracy at more specific levels.
- Loss of codon specificity
  - Genscan had explicit states for promoters, the 5' and 3' UTR regions, and the final poly-A signal
  - Doublescan has fewer distinctions – UTR, intron, or exon

Conclusion

- Doublescan provides a contrast with Genscan for gene structure detection – low-resolution versus high-resolution accuracy.
- Interestingly, Genscan and Doublescan had a tendency to fail on different genes. It is not clear why they were complementary in this fashion.

Future Work Areas

- Twinscan [Korf 2001] is a pair-HMM improvement on Genscan. A comparison of Twinscan and Doublescan could contain more useful information.
- Scoring algorithms that find something besides the best path (e.g. the forward-backward algorithm)
- Exploration of how the alignment affects prediction quality

References

- Meyer, M., R. Durbin (2002) Comparative ab initio prediction of gene structures using pair HMMs. *Bioinformatics* **Vol. 18 no. 10** pp 1309-1318
- Korf I., *et al*. (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics* **Vol 17 Suppl. 1** pp S140-S148
- Batzoglou S., *et al.* (2000) Human and Mouse Gene Structure: Comparative Analysis and Application to Exon Prediction. *Genome Research* **Vol. 10 Is. 7** pp 950-958
- Burge, C., S. Karlin (1997) Prediction of Complete Gene Structures in Human Genomic DNA. *J. Mol. Biol.* **268** pp 78-94
- Krogh, A. (1997) Two methods for improving performance of an HMM and their application for gene finding. *Proc. Fifth Int. Conf. Intelligent Systems for Molecular Biology*, Eds T. Gaasterland *et al.* pp 179-186
- Kulp, D., *et al.* (1996) A Generalized Hidden Markov Model for the Recognition of Human Genes in DNA. *ISMB-96* pp 134-141
- Rogic, S. *et al.* (2002) Improving gene recognition accuracy by combining predictions from two gene-finding programs. *Bioinformatics* Vol 18 no. 8 pp 1034-1045