==========AUGUSTUS==========

Gene Prediction with a Hidden Markov Model (2004)

This paper introduces a probablistic model called AUGUSTUS for genomic analysis of DNA based on generalized Hidden Markov Model (GHMM). AUGUSTUS has different functional parts and aims to find protein-coding genes. Based on GHMM, AUGUSTUS also has submodels, whose focuses including the length distribution of certain structural parts, all possible triplets fromed, and other extermal information in DNA sequences (AUGUSTUS+).

AUGUSTUS has been tested on several DNA sequences with known annotation: human, fruit fly, and Drosophila melanogaster. The test results of AUGUSTUS are compared with existing methods GENSCAN and GENEID and show an obviously higher accuracy. This approach efficiently solves the problem faced by previous sequence modeling methods, which are either computational inefficient or insufficient for extracting reliable genomic information. Furthermore, this approach allows to make use of evidence about a range of the DNA sequence, is supported by underlying theory, and significantly improves accuracy.

AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints (2005)

This paper presents a WWW server for AUGUSTUS, which allows users to artificially restrict the generalized Hidden Markov Model based on prior knowledge about the DNA sequences. This program allows users to set constraints like the position of a splice site, a translation initiation site, stop codon, position of known exon and intron information. Then it makes predictions that are consistent with the contraints.

The program has been tested on sequence data from human and Drosophila. The prediction of program with constraints turns out to be much better than the prediction of program without constraints on a sequence of 5000bp. This improved version of AUGUSTUS provides a new gene finding program, which performs significantly better than previous methods: for long input sequences, the accuracy of this program is superior to that of existing ab initio gene finding approaches.

Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources (2006)

This paper illustrates AUGUSTUS+ and how it performs integration of external information. AUGUSTUS+ utilizes again generalized Hidden Markov Model yet takes both intrinsic and extrinsic information as input, as well as combines hints from EST, protein, and other databases into the model. In case that hints are not certain, the model also provides the probabilities associated with hints after analysis.

Previously, extrinsic evidence is often incomplete, uncertain, and insufficient to recover complete gene structure when applied to existing gene prediction methods. The improvement from AUGUSTUS to AUGUSTUS+ has enabled the integration of sequence-intrinsic evidence and extrinsic evidence which lead to relatively sufficient evidence to modeling. As a result, its accuracy supercedes its previous version and other existing methods including GENSCAN, GENEID, HMMGene, and GenomeScacn, and TWINSCAN. All methods are tested on both sag178 (easy to test false positives) and human chromosome 22 (more realistic and complicated).

==========GENEMARK==========

Article Navigation Effects of choice of DNA sequence model structure on gene identification accuracy

This paper introduces the GeneMark algorithm which first models coding and non-coding sequence based on fixed order (FO) Markov chain, then implements Bayesian formalism to calculate the posterior probability of a DNA sequence segment to be coding or non-coding. It also provides several variants on the Markov chain model including Variable-order, Chi-square confidence (CHI-I), and delete interpolation (DI).

As Markov-chain-based genome annotation algorithms arise, it is crucial to refine the existing Markov model and further improve genomic models overall. So this paper provides a step forward. In its experiment, it tests and compares the performance of four different models mentioned above on several artificial sequences and prokaryotic

organisms. The result turns out that delete interpolation performs the best, which indicates that we may replace the original fixed order method by delete interpolation.

Probabilistic methods of identifying genes in prokaryotic genomes: Connections to the HMM theory

Instead of comparing markov chain models as the previous paper does, this paper introduces two overall extensions on the original GeneMark method: the first is GeneMarkS, which uses more elaborate Hidden Markov Model; the second is GeneMark.fba, which implements forward-backward algorithm and modifies the Viterbi's algorithm. The forward-backward algorithm takes into account both previous and next several consecutive positions to determine the current state with given model.

Currently there is no methods to locate all protein-coding genes efficiently, so this paper provides an demostration about the three perspective methods mentioned above and analyses their efficiency. First the paper performs an experiment to illustrate the scoring of the forward-backward algorithm. Later experiments are again carried out by comparing the methods on several artificial sequences and prokaryotic organisms. The result turns out that GeneMark.fba has the best accuracy among all three methods.

GeneTack: Frameshift Identification in Protein-Coding Sequences by the Viterbi Algorithm

This paper provides a method named GeneTack which detects frameshift in protein-coding nucleotide sequences, including both natural mutations and sequencing errors. It implements Hidden Markov Model and is wrapped by GeneMarkS. GeneTack first takes fragments based on the gene prediction of GeneMarkS; then it separates the fragments into three scenarios: true overlapping genes, true non-overlapping adjacent genes, and adjacent genes; finally it filters the fragments to predict the type of frameshift.

Existing programs that identify programmed frameshifts rely on the signaling sequences and do not accurately identify frameshifts related to sequence errors. Thus GeneTack is the first algorithm that solve the

problem adequately well. The performances of GeneTack and other existing algorithms including FrameD, FSFind, FSFind-BLAST are compared in the fragement detection of 17 prokaryotic genomes with GC content raging from 28% to 75%. Although GeneTack is not always the best, it has the best average accuracy. It is also worth noting that GeneTack does not use external evidence.