Chinmay Raut (93345289)

Summaries of Augustus and Gene map papers

Mario Stanke “Gene Prediction with a Hidden Markov Model”:

1. The problem:

Genome data was being created and we need a way to make sense of the sequences of nucleotides. We know that some sequences of nucleotides make proteins and in some cases we know the exact sequence of nucleotides, called genes, which make up some proteins. The thesis discusses a strategy for annotating nucleotide data to identify these genes.

1. The action (what was done):

A probabilistic model (specifically HMM) was proposed to model the nucleotides that were present in the DNA sample. The model tries to predict the probability of the presence of a gene starting at a strand of DNA given the observed strand as the observed states. Normally a naïve implementation would be rather time inefficient, but to improve the efficiency the detecting regions were further divided into subdivisions (such as splicing regions, introns, and extrons) reducing the search space. Once these regions are identified they get stitched together to see if individual genes can be identified using the same technique. The regions that the model looks for are fed in through a generic library for that class of organism. These regions fed in act like the training for the model so as new regions get discovered the program can be updated to detect them.

1. The results:

Augustus, the program had a space complexity linear in the number of nucleotides and the number of markov chain states. It has a time complexity linear to the number of nucleotides, but quadratic to the number of states. Augustus was compared to the other gene identification programs at the time (GENESCAN and GENEID). In terms of specificity and sensitivity it outperformed the other 2 programs, but in terms of raw accuracy which was calculated by number of genes correctly identified in a test dataset, it had similar levels of success.

1. The impact:

Augustus has both high exon identification and gene assembly accuracies. This was a big deal as before programs that identified exons properly were not as adept at stitching them together. The program had a linear time complexity which made scaling much easier to grasp and comprehend and the space complexity was rather low, making this program both light and fast. Also, the fact that higher order markov models not leading to improvements in accuracy (against theoretical beliefs) hints that straight-forward probabilistic calculations are not enough to understand the biological transcription and translation patterns of a cell and biological and environmental signals need to be incorporated in some way.

Mario Stanke “Gene prediction in eukaryotes with a generalized hidden Markov Model that uses hints from external sources”:

1. The Problem:

Genome-genome comparisons, EST sequences, and protein alignment are useful pieces of information about a gene sequence. While identifying genes in a nucleotide sequence sometimes this additional information about the gene or the sequence can be helpful. The question remained whether and how we can incorporate this information during gene detection.

1. The action (what was done):

The program Augustus was modified to incorporate protein database information or EST data to improve its ability to identify genes. Sites on the DNA sequence were updated to accept “hints” or locations of EST’s along with a type of that EST like start, stop codon, etc.… Hints are also given probabilities to reflect the user’s confidence then the HMM takes into account the proposal of the hint and its probability as it calculates the likelihood of observing a specific gene given the sequence DNA.

1. The results:

The new Augustus+ (one making use of extrinsic data motivators) performed much better than the other models (old Augustus, GENESCAN, GENEID, HMMGene) once incorporating all the information arriving at a sensitivity of 82% and a specificity of 79% for gene identification in the sag178 test set compared to the previous bests being 42% and 38% respectively. In a real world example Augustus+ tests on chromosome 22 in humans achieved 41% and 22% sensitivity and specificity.

1. The impact:

The predictions improved greatly with the incorporation of these “hints” as was expected. Careful consideration had to be made with hints that contradicted strong probabilistic evidence, so the model was updated to also remain skeptical of hints that contradicted with what was observed. This led to the model still being able to predict genes even if the wrong hints are given making the model robust in terms of incorporation of extrinsic evidence. It also sped up the program as unsuccessful exon hint searching led to no searching for exons in a region. The hints proved to be useful pieces of information worth factoring into gene prediction models, so we should try to incorporate more pieces of information into these models to improve their scores.

Mario Stanke “AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints”:

1. The Problem:

Gene prediction in eukaryotic genome sequences using Augustus is difficult to install and use.

1. The action (what was done):

Augustus was placed on a web server with an open web portal and website interface, so that people do not have to install software or deal with versioning issues. The site is located here: <http://augustus.gobics.de/submission>.

1. The results:

Anyone who has internet access can now use Augustus to identify genes in their sequence data. The analysis occurs within minutes and the report is also produced making sense of the prediction.

1. The impact:

Augustus is now easier to use and try. This “paper” is kinda like an ad…

Rajeev K. Azad and Mark Borodovsky “Effects of choice of DNA sequence model structure on gene identification accuracy”:

1. The Problem:

How can we choose the markov chain model properly to maximize gene identification in genomes? How and where do the fixed order (FO) models start to outperform the models built by deleted interpolation (DI)?

1. The action (what was done):

The performance of these models were done under the GeneMark identification tool. This tool operated previously under the FO model technique in which the segment of DNA would be coding or non-coding using markov chain based off of DNA for coding, shadow, and non-coding sections of a genome. The DI method divides the training dataset into 2 parts. The first is used to build a preliminary markov chain and the second is used to refine the transition parameters by maximizing the probability that the chain would correctly predict the 2nd set results. The analysis was done by using artificial DNA and then repeated using prokaryotic DNA.

1. The results:

Higher order models ended up performing better across both datasets. Also the FO model was better at predicting coding sequences whereas the DI model was better at predicting non-coding sequences. When training data was sparse DI performed better. When GC content was either high or low the FO model ended up predicting better.

1. The impact:

Many different markov chain producing models were attempted. None of them were really remarkable; however, 2 were identified to be promising. The FO and the DI model.

Rajeev K. Azad and Mark Borodovsky “Probabilistic methods of identifying genes in prokaryotic genomes: Connections to the HMM theory”:

1. The Problem:

How can the Bayesian method implemented in GeneMark be augmented and reintroduced to fit HMM theory to improve analysis on prokaryotic genomes?

1. The action (what was done):

A different HMM was proposed for the GeneMark algorithm that had the nucleotides as the observed states and the coding states as the hidden states. So the program goes through and predicts whether each nucleotide is part of a coding state or an observed state using the markov model for the transition probabilities.

1. The results:

For the prokaryotic genomes analyzed GeneMark boasted above 95% sensitivity and specificity values. The new program did take extra time to run, but the paper claims that it wasn’t too much extra and can be safely ignored.

1. The impact:

Statistical methods have much promise in detecting genes in genome data. False positives have not been completely eradicated, so there is still work to be done in this field but we are hitting a block pretty soon. We have not perfectly identified all the genes, so until then we can only estimate the performance and accuracies of the models and these techniques will have to evolve as well.

Ivan Antonov and Mark Borodovsky “GENETACK: FRAMESHIFT IDENTIFICATION IN PROTEIN-CODING SEQUENCES BY THE VITERBI ALGORITHM”:

1. The Problem:

Is it possible to identify frame-shift mutations in a gene? Frame shift mutations are when a nucleotide gets inserted or deleted by accident which causes the entire frame for the gene to shift resulting in a different gene than what was expected.

1. The action (what was done):

The GeneTack identification software uses a HMM and Viterbi acceleration to identify genes which also allows it to detect predictions made by frameshifts. They grabbed 17 prokaryotic genomes and induced a few frameshifts and tested GeneTack and other gene identification software on them.

1. The results:

GeneTack-GM and FSFind-BLAST both seemed to perform equally well with average sensitivity and specificity ranging from about 60-80%. GeneTack can also accurately predict frameshifts in about half the genes and is biased to predict frameshifts at the end better than frameshifts at the beginning.

1. The impact:

The GeneTack model has proven useful at analysis in other genomic sequences without introns. For short sequences this algorithm is better as longer into the sequence the predictions become worse. Alignment of the EST sequences is currently used to correct most of the frameshifts, but GeneTack can be used to predict and identify some more.