Beoung Hun Lee

CS 199

March 23, 2019

Edwin Solares

Markov Model has been frequently used as the computational method to determine genes in genome statistically. However, amongst the Markov Model, the optimal algorithm has not been determined yet. This paper mainly compares forward algorithm and interpolation algorithms to determine the most optimal algorithm to be used for the gene prediction. It is predicted that although forward algorithm is mainly used, there are algorithms that work better than forward algorithm in presence of sparse data. Other algorithms like Chi-I and Variable orders are considered as well. This differs from the AUGUSTUS, which uses both Markov Model and extrinsic information, as Genemark only utilizes intrinsic method. The study mainly used artificial prokaryote genomes with high and low GC contents as a sample and used different amount of data to examine the effect of size of data set on effectiveness of algorithm. Two different cases of genomes were used. GP case has coding and non-coding sequence that has a same size as actual genome, and TP case has non-coding sequence that was one-third the size of a coding sequence.

Overall, the result showed that interpolated model and forward algorithms were more accurate than other algorithms with the interpolated model performing better in some cases. The performance of each algorithms depended on data size, order of model, case and genome. In the lower G-C contents genome of B. subtilis, the interpolated model outperformed Chi-I and forward model, although FO model was as good as the interpolated model as training data set gets bigger. For the higher G-C contents genome of R. solanacearum, the forward model performed better in GP case, but interpolated model performed better on TP case. In medium G-C content genome of U. urealyticum, interpolated model performed better on TP case, while in GP case, the level of performance was similar to that of forward model. Throughout the experiments, forward algorithm showed lower error rate as the training data size increased while interpolated model was more accurate in smaller training data set. Interpolated model performed the best in medium G-C content genome, while forward algorithm performed equally or better in high or low G-C content genome. Based on the result, it seems that optimal algorithm for gene prediction is species specific, as the performance of the models depend on the percentage of G-C contents in the genome.

References

1. Azad, R. and Borodovsky, M. (2004). Effects of choice of DNA sequence model

structure on gene identification accuracy. *Bioinformatics.* Vol 20. 993-1005.