

Using Machine Learning to Improve Gene Finding

Sethu Laxmi Sebastian¹, Chun-Yin Liu², Gary Livingston²

Keywords: gene finding, genome annotation, machine learning

1 Introduction.

GLIMMER (versions 1.0 and 2.0) is a computational gene finder that is able to find 97-98% of all genes in a prokaryotic genome. The gene learning system has four components mainly the gene generator, performance system, critic component and learning component. While GLIMMER has a recall rate, finding nearly all of the identified genes, they often have a high frequency of false predictions, sometimes as high as 40%. We propose the use of machine learning to learn a model for predicting when these programs make incorrect predictions. These models may be used to improve GLIMMER's predictions and could be used to identify improvements to GLIMMER. This method should be general and applicable to most gene finding systems.

2 Gene finding and GLIMMER.

Genome is the complete DNA sequence containing all the genetic information and the supporting proteins. DNA consists of four bases, adenine (A), guanine (G), cytosine (C) and thymine (T). The total genome size of a mammal is $\sim 3 \times 10^9$ bp [2]. The information in DNA is stored as a triplet codon. The codon codes for a specific amino acid. Genome annotation is the technique by which all meaningful information in the DNA sequence is decoded. The position of the gene determines its function. The sequence of the bases determines the information available for building and maintaining an organism. Gene finding, the identification of gene sequence, can provide a sense of relief from uncertainty and help people make informed decisions about managing their health care. GLIMMER is the primary microbial gene finder at TIGR (The institute for Genomic research), and has been used to annotate the complete genomes of over 80 bacterial species, including *Borrelia burgdorferi*, *Treponema pallidum*, *Chlamydia trachomatis* and *Thermotoga maritime* [1] at TIGR and elsewhere.

The gene learning system has 4 components: gene generator, performance system, critic component and learning component. The "gene generator" takes identified rules from learning system and generates a new test gene to improve the learning. This will collect the genes from public database and passes these genes to the second component. The "performance system" will use learned knowledge to identify genes and generate the solution trace to the critic component. The "critic component" will analyze the performance and identify the genes. The training examples are then given to the "learning component". The learning component used here is GLIMMER. The training genes are fed into GLIMMER, which will identify and classify the genes.

GLIMMER is a Gene Locator and Interpolated Markov Modeler. It is a system for finding genes in microbial DNA, although it can also be used to find eukaryotic genes. It uses interpolated Markov models (IMMs) to identify the coding regions and to distinguish them from non-coding regions. The basic concept in GLIMMER is Markov Model. The probability of a given base b is modeled

¹ Department of Biological Sciences, University of Massachusetts Lowell

² Department of Computer Science, University of Massachusetts Lowell

by the Markov model, in DNA sequence analysis, as depending only on the k bases immediately prior to b in the sequence. Since the training data available for building the model is limited, k must be limited. For a given position b , IMM finds a suitable position to cutoff.

GLIMMER 1.0 builds three separate IMM's and it can find ~97% of all the genes in a genome when compared to the published annotation. The IMM approach uses a combination of Markov models from the first to the eighth order, weighting each model according to its predictive power. GLIMMER 1.0 and 2.0 use three-periodic non-homogenous Markov models in their IMM's.

Training GLIMMER for all genes, ranging in size from 0.5 to 4.7 MB, takes less than 1 minute on a Pentium 400 PC running the Linux operating system. The gene finding step takes an additional 1 minute or less.

3 Learning to improve predictions

After applying GLIMMER to a training genome, we will extract test cases of GLIMMER's predictions. We are still developing the features, which we will use to describe cases. Next, several machine learning methods will be used to develop models for classifying GLIMMER's predictions. These models will be evaluated using similar, but new testing genomes. This method is general and may be easily adapted to other gene finding methods.

4 References

[1] Delcher, A., Harmon, D., Kasif, S., White, O., and Salzberg, S. 1999. Improved microbial gene identification with GLIMMER. pp. 4636-4641, *Nucleic Acids Research*, Vol. 27, No. 23.

[2] Lewin, 2004, *GENES*, 6th ed., Prentice-Hall.