

A New Approach to Gene Prediction Using the Self-Organizing Map

Shaun Mahony¹, Terry J. Smith¹, James O. McInerney², Aaron Golden³

¹National Centre for Biomedical Engineering Science, NUI, Galway, Galway, Ireland

²Bioinformatics and Pharmacogenomics Laboratory, NUI, Maynooth, Co. Kildare, Ireland

³Department of Information Technology, NUI, Galway, Galway, Ireland

email: shaun.mahony@nuigalway.ie

Abstract

In this poster we present a gene prediction approach based on the Self-Organizing Map that has the ability to automatically identify all the major patterns of content variation within a genome. The genome may then be scanned for regions displaying the same properties as one of these automatically identified models. Even using a relatively simple coding measure (codon usage), this method can predict the location of protein-coding sequences with a reasonably high accuracy. We also show other advantages of the approach, such as the ability to indicate genes that contain frame-shifts. We believe that this method has the potential to become a useful addition to the genome annotation process.

1. Introduction

Computational gene prediction methods have yet to achieve a perfect accuracy rate, and many make a substantial number of false-positive predictions, even in prokaryotic genomes. One of the most obvious reasons for inaccurate gene predictions is the high degree of compositional variation that exists within most genomic sequences. For example, it has long been recognized that synonymous codon usage is highly variable, and under many evolutionary pressures (see [1] for a review). Many Markov model based gene-finding tools use only one model to represent protein coding regions in any given genome, and so are less likely to predict genes with an unusual composition. Indeed, it has been shown that using two or three models substantially increases the accuracy of a Markov-based gene-finder [2].

In this poster, we show how the Self-Organizing Map can be effectively used to overcome the problem of intra-genomic variation in composition, and we demonstrate this using codon usage as a coding measure.

2. Methods

2.1. RSCU Vectors

In our method, Relative Synonymous Codon Usage (RSCU) vectors are used as the measure of the protein-coding potential of a sequence. The RSCU value for a codon 'i' is defined as:

$$RSCU_i = \frac{\text{Observed } i}{\text{Expected } i}$$

The \log_{10} of each value is then found in order to center the value around 0. For each sequence, RSCU values are calculated for each of the 59 codons with synonymous alternatives. Similarity between two RSCU vectors can be found by measuring the cosine of the angle between them.

2.2. Self-Organizing Map

The Self-Organizing Map (SOM) is a popular unsupervised neural network algorithm, often used to visualize and cluster high-dimensional data. The SOM is based around the concept of a lattice of interconnected nodes, each of which contains a model (in our case, a RSCU vector). The models change during training to become similar to common or repeated patterns in the training set. Similar models are clustered together.

All SOMs used in our analyses were 15x15 nodes in a square lattice configuration. The full SOM training algorithm is given elsewhere [3], but we summarize our use here:

- (1) A gene's RSCU vector is loaded from the training dataset.
- (2) The lattice node is found whose RSCU model most closely resembles the input vector. This node is denoted the 'winning node'.
- (3) The winning node's model, W (as well as a certain number of 'neighbourhood bubble' node models)

is changed to be more similar to the input vector using:

$$W_{\text{new}} = W_{\text{old}} + \eta(X_i - W_{\text{old}}).$$

- (4) If all the vectors in the training dataset are processed, we say that an epoch has been completed. In this study, SOMs were trained for 3000 epochs.

SOMs have been previously used to study codon usage, but with a focus on attempting to identify horizontally transferred genes [4].

2.3. Gene Prediction

For each genome under test, SOMs are trained using all genes confirmed by homology which were also longer than 750 bp. A sliding window is then used to split the genomic sequence into short samples in each of the six reading frames. Each of these samples are passed into the SOM in turn, and we can easily find the most similar node model, and how similar it is to the sample.

Samples which score above a certain threshold are predicted to be protein coding. Concurrent high-scoring samples are strung together to form gene predictions. If a stop codon occurs within one of the samples, the gene prediction ends at that point, but no effort is made to find start codons, and predictions are never artificially lengthened in order to border the prediction with start and stop codons.

Some predictions are deleted if they overlap other predictions to the degree that they are obvious false-positive predictions, but the method tolerates many more overlapping genes than other methods. Unfortunately, our method only offered accuracy down to a lower prediction size of 75 codons.

3. Results and Discussion

An analysis of 8 genomes was carried out, the results of which are shown in Table 1. Gene predictions were counted as correct if they were in the correct frame and they predicted the majority of the actual gene.

While the results do suggest an accuracy that is less than the current state-of-the-art methods, we feel that the general approach is justified when the following points are taken into account. Firstly, codon usage is not the best indicator of protein-coding potential. Many previous studies have shown hexamer frequency to be more efficient [5], and using hexamers as the coding measure may increase accuracy. Also, genes that have evolutionary

origins in horizontal transfer would be expected to have codon usage patterns that are different from other genes in the same genome. This would make them unlikely to be predicted using a method that trains to recognize the typical codon usage patterns in a genome.

Finally, we recognize that post-processing of the predictions in our method is much less sophisticated than in other methods. However, we believe that since our method does not rely on modifying predictions in order to fit in likely start or stop codons, our predictions represent an accurate account of regions of the genome that display native codon usage patterns. This is especially useful in predicting frame-shifts in genes, as predictions will exist in separate frames each side of the frame-shift.

Organism	Known Genes	Correct Prediction	Sn (total)	Sn (>225bp)	Sp
<i>A. aeolicus</i>	1517	1453	95.78	96.54	87.8
<i>B. burgdorferi</i>	857	776	90.54	96.39	98.02
<i>E. coli</i>	4290	3835	89.39	92.85	89.04
<i>H. influenzae</i>	1754	1606	91.56	96.34	98.01
<i>H. pylori</i>	1593	1456	91.39	96.8	95.49
<i>M. genitalium</i>	483	432	89.44	91.52	92.32
<i>M. jannaschii</i>	1715	1516	88.39	91.82	96.5
<i>Synechocystis</i>	3169	2953	93.18	96.53	90.95

Table 1: Percentage Sensitivity (Sn) and Specificity (Sp) values in each of the 8 genomes tested.

Our method is included as a sub-function of the RescueNet codon usage exploration software program. The program and is available free of charge to academic users from <http://bioinf.nuigalway.ie/RescueNet>.

4. References

- [1] L. Duret, "Evolution of synonymous codon usage in metazoans," *Curr Opin Genet Dev*, vol. 12, pp. 640-9, 2002.
- [2] W. S. Hayes and M. Borodovsky, "How to interpret an anonymous bacterial genome: machine learning approach to gene identification," *Genome Res*, vol. 8, pp. 1154-71, 1998.
- [3] T. Kohonen, *Self-Organizing Maps*. Berlin: Springer-Verlag, 1995.
- [4] H. C. Wang, J. Badger, P. Kearney, et al., "Analysis of codon usage patterns of bacterial genomes using the self-organizing map," *Mol Biol Evol*, vol. 18, pp. 792-800, 2001.
- [5] J. W. Fickett and C. S. Tung, "Assessment of protein coding measures," *Nucleic Acids Res*, vol. 20, pp. 6441-50, 1992.