International Research Journal of Computer Science (IRJCS)
Issue 6, Volume 2 (June 2015)
ISSN: 2393-9842
www.irjcs.com

PREDICTION OF TRANSCRIPTION FACTOR BINDING USING CLOUD COMPUTING AND RSCU VALUE FOR GENE CLASSIFICATION

Vishwanatha[1] Dr.K S JagadeeshGowda[2] Naganandini G[3] Hemavathy R[4]

[1]*Asst.Professor, Department of CS&E, Sri Krishna Institute of Technology, Bengaluru.*
[2]*Head of the Department-CS&E, Sri Krishna Institute of Technology, Bengaluru.*
[3]*Asst.Professor, Department of CS&E, Sri Krishna Institute of Technology, Bengaluru.*
[4]*Asst.Professor, Department of CS&E, Sri Krishna Institute of Technology, Bengaluru.*

Abstract— The prediction of the function of a novel gene remains to be a challenging problem. Given a piece of coding sequence, one can deduce its function by finding homologous genes using sequence or protein structural alignment. One can also perform gene expression measurements or gene knockout experiments to determine the function of a gene. However, a gene may not have homology with any known gene and experiments can be expensive. It would therefore be beneficial if the function of a gene can be predicted from the characteristics of its coding sequence. One such characteristic is synonymous codon usage. A novel approach for gene classification, adopting codon usage bias as feature inputs to support vector machines (SVMs) is proposed. The DNA sequence is first converted to a 59-dimensional feature vector, where each element corresponds to the relative synonymous usage (RSCU) frequency of a codon. Since the input to the classifier is independent of sequence length, the approach is especially useful when sequences to be classified are of differing lengths and homology-based methods tend to fail.

Keywords— HLA, Support Vector Machine, Synonymous codons, Cloud computing.

1. INTRODUCTION

In this paper we have presented the creation of gene/DNA to Protein sequence which contains the detailed list along with some of the major information regarding various amino acids and codon details including start and stop codon, possible physical and chemical properties of the each amino acids with the structure will be displayed in the new window Newly discovered stretch of DNA or amino acid is not informative sequence on its own. There must be a way to analyse these sequences by comparing them against existing databases to develop hypothesis about their relatives and functions. This method is called similarity searching by sequence alignment. Sequence alignment is a way of arranging the primary sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.

Codons are triplets of nucleotides that together specify amino acids in a polypeptide chain. Most organisms use 20 or 21 amino acids to make their polypeptides, which are proteins or protein precursors. It is sequence of Amino acids and we know that protein synthesis starts from the codon Methionine (AUG,Met), called as start codon. In the field of bioinformatics and computational biology, many statistical methods have been proposed and used to analyse codon usage bias. In the Standard Code there is a unique codon that always starts the coding region of a protein-coding gene. This codon is called the *start codon* , and in the Standard Code it is always the unique Met codon, AUG. In other genetic codes there may be more start codons in addition to this one. That means that, in organisms with the Standard Code, coding regions always start with AUG and end with one of the stop codons -- this forms one basis for automated methods that look for protein-coding genes in unannotated sequence. Synonymous codon usage biases are associated with various biological factors, such as gene expression level, gene length, gene translation initiation signal, protein amino acid composition, protein structure, tRNA abundance, mutation frequency and patterns, and GC compositions. Quantification of codon usgage biases are helps understand evolution of living organisms. A codon usage bias pipeline is demanding for codon usage biase analyses within and across genomes. Here we present a codon web server service as a user-friendly tool for codon usage bias analyses across and within genomes in real time system.

A gene is a unit of heredity in a living organism. It is normally a stretch of DNA that codes for a type of protein or for an RNA chain that has a function in the organism. All living things depend on genes, as they specify all proteins and functional RNA chains. Genes hold the information to build and maintain an organism's cells and pass genetic traits to offspring. A modern working definition of a gene is "a locatable region of genomic sequence, corresponding to a unit of inheritance, which is associated with regulatory regions, transcribed regions, and or other functional sequence regions " Colloquial usage of the term gene (e.g. "good genes, "hair color gene") may actually refer to an allele: a gene is the basic instruction, a sequence of nucleic acid (DNA or, in the case of certain viruses RNA), while an allele is one variant of that instruction. Thus, when the mainstream press refers to "having" a "gene" for a specific trait, the press is wrong.

2. CLOUD COMPUTING

Cloud computing provides scalable, real time, on demand computing services. This model fits with bioinformatics needs:On-Demand Scalability you are likely to require a huge computing infrastructure during the analysis phase of a big sequencing project, but may be not at all during the holiday season. So it's about being able to both grow and shrink your infrastructure.Real-Time Adjustment If you need it now an essential component of cloud computing is to adapt your computing infrastructure in real-time, programmatically (API calls with results in the range of minutes, if not seconds)Low-Cost Cloud computing adds lowering to all that You will see a lot of time by having to maintain your own infrastructure, and, so result, you end up concentrating on knowledge driven, value based activities.

3. SYNONYMOUS CODON

Synonymous codons are not generally used at equal frequencies, and this trend is observed for most genes and organisms. Several methods have been proposed and used to estimate the degree of the non random use of the different synonymous codons. The estimates obtained by these methods, however, show different levels of both precision and dispersion when coding regions of a finite number of codons are under analysis. Here, we present a study, based on computer simulation, of how the different methods proposed to evaluate the non random use of synonymous codons are affected by the length of the coding region analysed. The results show that some of these methods are heavily influenced by the number of codons and that the comparison of codon usage bias between coding regions of different lengths shows a methodological bias under different conditions of non random use of synonymous codons. The study of the dispersion of the estimates obtained by the different methods gives, on the other hand, an indication of the methods to be applied to compare values of codon usage bias among coding regions of equivalent length.

3.1 Computation of Relative Synonymous Codon Usage Frequency (RSCU) values.

To evaluate synonymous codon usage without confounding the influence of amino acid compositions of different sequence samples, the RSCU was adopted. For a given coding sequence, RSCU value rk of synonymous codon k is calculate

$$rk = nk * obsk / totk \qquad (1)$$

where obsk and totk are the observed number of codon k and the total observed number of codons coding for the amino acid coded by codon k, respectively, and nk denotes the number of synonymous codons of codon k. Though there are 64 codons in the genetic code, two codons, UGG and AUG, are unique codons for the amino acids Tryptophan (Trp, W) and Methionine (Met, M), respectively, and not considered as their RSCU values equal to unity; three stop codons (UGA, UAA, and UAG) are also not considered. Therefore, RSCU values of only 59 codons are used as input features for classification.

4. IMPLEMENTATION

4.1 Methodology

Approach for gene classification consists of two steps: Computation of relative synonymous codon usage (RSCU) frequency pattern as the feature vector representing each sequence and Classification using SVM (see Fig. 1). Binary SVM was used to classify a given HLA molecule into major classes , HLA-I and HLA-II , and multiclass SVM for the subclass classification of HLA-I and HLA-II molecules.

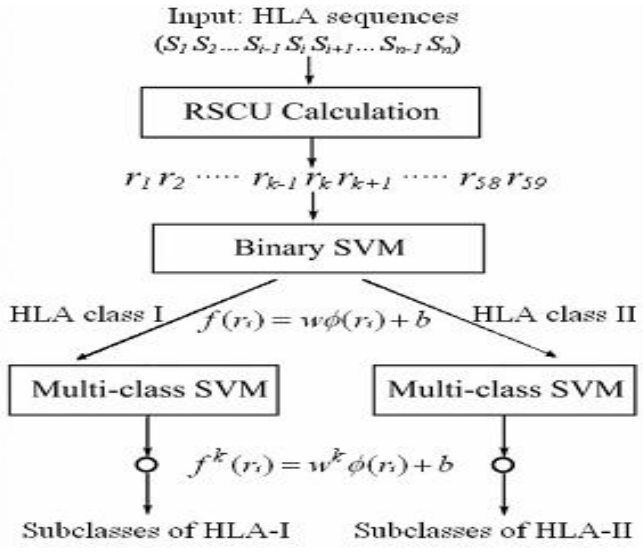

Fig. 1. Illustration of gene classification using codon usage as input features and SVM.

4.2 BioPerl

BioPerl is a collection of Perl modules that facilitate the development of Perl scripts for bioinformatics applications. It has played an integral role in the Human Genome Project. It is an active open source software project supported by the Open Bioinformatics Foundation.

BioPerl provides software modules for many of the typical tasks of bioinformatics programming. These include:

1. *Accessing nucleotide and peptide sequence data from local and remote databases*
2. *Transforming formats of database/ file records*
3. *Manipulating individual sequences*
4. *Searching for similar sequences*
5. *Creating and manipulating sequence alignments*
6. *Searching for genes and other structures on genomic DNA*
7. *Developing machine readable sequence annotations*

Bioperl is a large, object-oriented toolkit of interacting perl modules useful for building bioinformatics solutions in Perl. The collection of modules in the bioperl-live repository contains the core functionality. Additional packages for creating graphical interfaces (bioperl-gui), setting up persistent ORM storage in RDMBS (bioperl-db), running and parsing the results from hundreds of bioinformatics applications (bioperl-run), and software to automate bioinformatic analyses (bioperl-pipeline) are all available.

5. RESULTS

5.1. Comparison between RSCU and RAC for the Sequence:

The bellow graph shows the comparison between the RSCU value and RAC value of the codons of the input sequence which is bias for the classification of genes. It shows the vales with respect to the codons and graph has been constructed for each one by one.We produces each codon frequency, Relative Synonymous Codons Uses and Relative Adaptiveness of a Codon table and bar graph that will help you to calculate the Codon Adaptation Index (CAI) of a gene, to see the gene expression level. Relative Synonymous Codons Uses(RSCU) values are the number of times a particular codon is observed, relative to the number of times that the codon would be observed in the absence of any codon usage bias.In the absence of any codon usage bias, the RSCU value would be 1.00. A codon that is used less frequently than expected will have a value of less than 1.00 and vice versa for a codon that is used more frequently than expected.

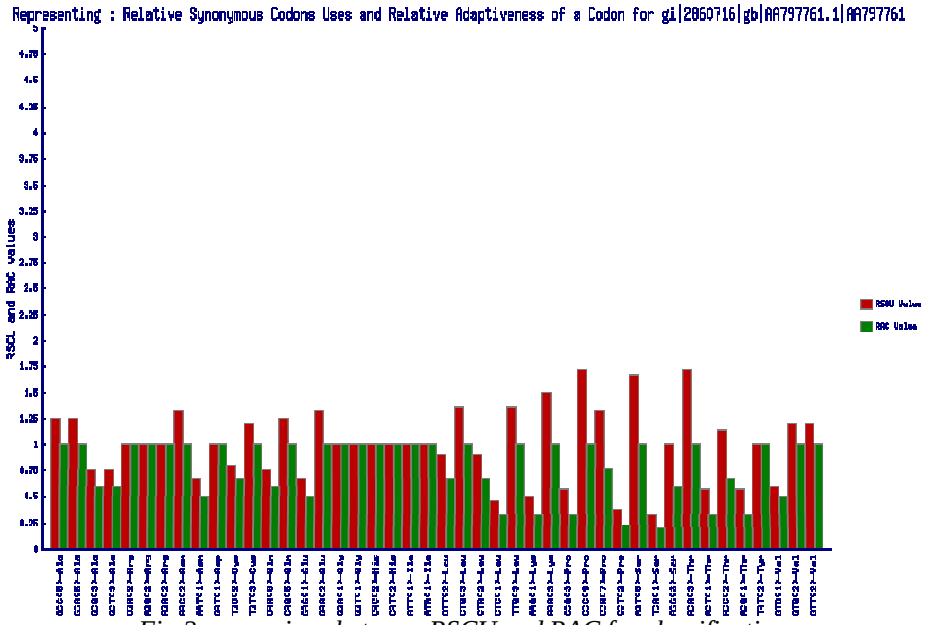

Fig 2 comparison between RSCU and RAC for classification.

Codon usage pattern-based approach for the classification of HLA sequences is relatively fast and efficient as the input sequences are transformed to a smaller dimensional RSCU space.

6. CONCLUSION

Large genomic sequencing projects of pathogens as well as human genome leads to immense genomic and proteomic data which would be very beneficial for the novel target identification in pathogens. Subtractive genomic approach is one of the most useful strategies helpful in identification of potential targets. The approach works by

subtracting the genes or proteins homologous to both host and the pathogen and identify those set of gene or proteins which are essential for the pathogen and are exclusively present in the pathogen.

GtoP Virtual converter is a project which gives the viewer complete details of the various types DNA sequence and functional gene, it's occurrence in the living organism, result's of various reactions on that functional part like gene. This tool gives a clear idea of various amino acid, codons and the protein occurs which comes under the given sequence of DNA from the different websites. This tool provides the converter for DNA to Protein through codon table and amino enhancements to the project can make the application more attractive acids. This tool also provides the structural details of the various amino acid of this family along with the detail protein description, their taxonomy, their mutation, and also the disease caused due to the mutation of sequences. This tool provides the details of amino acids along with the structure and all properties of each of them. In the future there is an intend to add references and details of other converter deficiency conversion. With increasing computing power such combined approaches become increasingly feasible, and can more efficiently utilize the information from a given set of experimental data.

REFERENCES

[1]. R. Grantham, C. Gautier, M. Gouy, R. Mercier, and A. Pave, "Codon Catalog Usage and the Genome Hypothesis," Nucleic Acids Research, vol. 8, pp. r49-r62, 1980.
[2]. T.C. Ghosh, S.K. Gupta, and S. Majumdar, "Studies on Codon Usage in Entamoeba histolytica," Int'l J. Parasitology, vol. 30, pp. 715- 722, 2000.
[3]. P.M. Sharp, E. Cowe, and D.G. Higgins, "Codon Usage Patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster, and Homo sapiens: A Review of the Considerable Within-Species Diversity," Nucleic Acids Research, vol. 16, pp. 8207-8211, 1988.
[4]. J.M. Ma, T. Zhou, W.J. Gu, X. Sun, and Z.H. Lu, "Cluster Analysis of the Codon Use Frequency of MHC Genes from Different Species," Biosystems, vol. 65, pp. 199-207, 2002.
[5]. J.M.Ma, N.M. Nguyen, G.B. Fogel, and J.C. Rajapakse, "Determination of the Relative Importance of Gene Function or Taxonomic Grouping to Codon Usage Bias Using Cluster Analysis and SVMs," Proc. IEEE Symp.Computational Intelligence in Bioinformatics and Computational Biology, Sept. 2006.
[6]. Build your own database using Php MySql 2003, 51-67.
[7]. Jenkins GM, Pagel M, Gould EA, de A Zanotto PM, Holmes EC. Evolution of base composition and codon usage bias in the genus Flavivirus. J Mol Evol 2001, 52: 383−390
[8]. Levin DB, Whittome B. Codon usage in nucleopolyhedroviruses. J Gen Virol 2000, 81: 2313−2325.
[9]. Codes in the codons: construction of a codon/amino acid periodic table and a study of the nature of specific nucleic acid-protein interactions Benyo, B.; Biro, J.C, 26th Annual International Conference of the IEEE
[10]. Jenkins GM, Holmes EC.The extent of codon usage bias in human RNA viruses and its evolutionary origin. Virus Res 2003, 92: 1−7
[11]. Di-codon usage for classification of genesnby Minh N. Nguyena, Jianmin Maa, Gary B. Fogelb, Jagath C. Rajapakse.

