Journal of Advanced Bioinformatics Applications and Research
ISSN 0976-2604.Online ISSN 2278 – 6007
Vol 3, Issue 3, 2012, pp 357-363
http://www.bipublication.com

SVM MODEL FOR CLASSIFICATION OF GENOTYPES OF HCV USING RELATIVE SYNONYMOUS CODON USAGE

***C.M.Nisha, Bhasker Pant, and K. R. Pardasani**

[1]Department of Bioinformatics, MANIT, Bhopal, India
*Corresponding author: Email: nishacm2001@gmail.com

[Received-19/06/2012, Accepted-03/07/2012]

ABSTRACT:
The genotype of the HCV strain appears to be an important determinant of the severity and aggressiveness of liver infection, as well as patient response to antiviral therapy. HCV genotypes display significant differences in their global distribution and prevalence, making genotyping a useful method for determining the source of HCV transmission in an infected localized population. A novel approach for genotype classification is proposed, which adopts codon usage bias pattern as feature vector for the subsequent classification using Support Vector Machines (SVMs). A given DNA sequence is first converted to 59-dimensional feature vector, each element corresponding to the relative synonymous usage frequency of a codon. Therefore, the input to the classifier is independent of the size of the DNA sequences. Therefore, our approach is useful when the genes to be classified are of different length, where the homology-based methods are inapplicable due to the difficulty in the alignment of sequences having different lengths. The applicability and usage of the present method is demonstrated by a classification of genotype of HCV selected from the database of Los Alamos National Laboratory (LANL) HCV to train the SVM which is implemented using freely downloadable software Lib SVM-2.89. The results show that the proposed method is capable of accurately classifying Genotypes of HCV. Also, the results of gene classification according to the codon usage bias pattern are consistent with the molecule structures and biological functions, further validating our approach.

Keywords: Codon usage bias; genotype classification; Relative Synonymous Codon Usage (RSCU); Support Vector Machine; Kernel function.

[I] INTRODUCTION

Hepatitis C virus (HCV), the sole member of the *Hepacivirus* genus within the *Flaviviridae*, poses a global health burden, with an estimated 170 million infected individuals (according to the WHO) [4]. Over 80% of acutely infected individuals progress to a chronic carrier state that can lead to liver cirrhosis and hepatocellular carcinoma [1]. In acute HCV infection, an early HCV-specific cellular immune response is associated with viral clearance and recovery [12], whereas in chronically infected individuals, cellular immune responses are generally low and unable to eliminate the virus [5]. No prophylactic HCV vaccine is currently available and the only accepted therapy thus far, interferon-a, is successful in only 20% of the

cases. Increased efforts are therefore needed in the development of an effective vaccine against HCV [13].

HCV constantly changes and mutates as it replicates more than 1 trillion hepatitis C virions replicate each day. During the replication process, the hepatitis C virus will make 'bad' copies or errors in the genetic make-up of the newly replicated viruses. The process of constant mutation helps the virus evade the body's immune response---when the dominant quasi-species is eradicated, another quasi-species emerges [6]. This requires the immune system to constantly identify and kill the newly emerged variants. This is one of the reasons why so many people develop chronic disease. Scientists believe there are literally millions of different HCV quasispecies in everyone infected with hepatitis C, which are unique to everyone because of the individual's immune response to HCV and quasispecies constantly change over time [8]. In addition, it has been suggested that quasispecies play a role in disease progression and treatment response, but this is still controversial and more studies are needed to fully appreciate the role of quasi-species.

This variability (genotype, subtypes and quasispecies) of hepatitis C [9] has made it difficult to treat and to develop a vaccine that will protect against all HCV strains although recent advances in vaccine development have been encouraging [14].

Genetic information transfers from nucleic acids to proteins in the form of *codons*. S*ynonymous* codons are used at different frequencies during the process of translation [4]. This phenomenon, referred to as *codon usage bias*, was found to be highly variable among different species. The function of a gene is closely correlated with codon usage bias and when the function of the gene is confirmed, species determine further differentiation in the codon usage bias [15]. Further, codon bias pattern is also closely related to protein tertiary structure. The study of codon usage is important as it is closely related to translation, which bridges the gap between the languages of nucleic acids and proteins and also very useful in mutation studies. When a synonymous mutation occurs, the coded protein remains unchanged, yet the codon usage pattern varies.

Therefore, the codon usage pattern is a good indicator for the studies of mutation and molecular evolution. In this study, we use codon usage patterns to represent DNA sequences – a vector of 59 elements represents a given DNA sequence. Therefore, the input to the classifier is of the same dimension irrespective of the length of the DNA sequence [15]. Since codon usage patterns are diverse in different gene families, this feature input is a good indicator for discriminating different gene families.

Support Vector Machines (SVMs) have been earlier shown to perform well in multiple areas of biological analysis, which have strong foundations in statistical learning theory; as shown by Vapnik [10], SVMs implement a classifier that is capable of minimizing the structural risk and offer several associated computational advantages such as the lack of local minima and a solution completely encompassed by the set of support vectors. Its ability to scale well in large-scale problems is particularly attractive for predicting structures of protein sequences as well as gene expression [2]. The generalization capability of SVMs makes it an effective approach in clustering and well suitable for gene classification, so we adopt it in the classification of genes with codon usage pattern as inputs.

Realizing the importance of Genotype in predicting HCV medical treatment response, treatment duration and the dose of ribavirin [17], we have chosen different Genotype of HCV for our study, hence develop a SVM Model for the Classification of HCV isolates of unknown genotype [18]. Altogether 1864 sequences are obtained from the LANL HCV database [16] and the RSCU values of respective sequences are calculated by using our own toolkit for RSCU Calculation developed using PERL, where we took 59 RSCU feature vector corresponding to different codon except Methionine, Tryptophan and Stop Codons and

gave as an input in SVM to get a Model. For sub-class classification of different Genotypes of HCV in major 6 classes, multi-class SVM are implemented using SVMlight [7], which usually lead to faster convergence in large optimization problems.

The Gaussian kernel is selected $K(x, y) = e^{-\$|x-y|2}$ for multi-class SVM methods. The Gaussian kernel shows superior performance over the linear and polynomial kernels as the Gaussian kernel results in complex (but smooth) decision functions and therefore has the ability to better fit the data where a simple discrimination by using a hyper plane or a low-dimensional polynomial surface is impossible [3].

This paper is a step in the direction where machine learning and computational biology techniques can be used to compliment existing wet lab techniques.

[II] MATERIALS AND METHODS

2.1 (Data set)

To achieve our goal and develop our methodology we obtained sequences through the **Los Alamos National Laboratory (LANL) HCV database** [16]. The following data sets were used.

Finally, 1864 sequences were chosen and the numbers of sequences chosen in each sub category are shown.

[Table -1]

Main Class	Sub Class	Number of sequences
	Genotype 1	844
	Genotype 2	29
HCV	Genotype 3	9
	Genotype 4	15
	Genotype 5	3
	Genotype 6	32

Table 1: Datasets for each Genotype

2.2 (Data Analysis)

Here a novel approach for gene classification is proposed, which adopts codon usage bias pattern as feature vector for the subsequent classification using Support Vector Machines (SVMs).

Our tool developed in Perl will prepare an input format in form of feature vector for SVM where a given DNA sequence is first converted to 59-dimensional feature vector, each element corresponding to the relative synonymous usage frequency of a codon. Therefore, the input to the classifier is independent of the size of the DNA sequences and RSCU of each sequence is calculated.

2.3 (Calculation of codon usage)

The purpose of this function is to calculate the Number (***N***) of times a particular codon is observed in a gene or set of genes and also to calculate the **Relative Synonymous Codon Usage (RSCU)** values for the dataset [4].

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.

The program adds up the total number of times that the codons for a particular amino acid are observed (depending on the amino acid, this might be 2, 3, 4 or 6 different codons). It then divides this number by the number of codons for the amino acid (2, 3, 4 or 6); this gives the expected number of times that the codons should be observed. Then for each codon, the frequency of observation is divided by the expected frequency. Sometimes the observed frequency will be greater than the expected frequency (RSCU value greater than 1.00), and sometimes it will be less (RSCU value less than 1.00).

The **relative synonymous codon usage** (**RSCU**) is the number of times a codon appears in a gene divided by the number of expected occurrences under equal codon usage [15].

$RSCU = X_{ij} / (1/n_i * S \{X_{ij}; j=1, n_i\})$

Where Xij is the number of occurrences of the j^{th} codon for the i^{th} amino acid, and n_i is the number (from 1 to 6) of alternative codons for the i^{th} amino acid.

i.e.

For each chosen coding sequence *s*, the RSCU value of each synonymous codon *rk*, was calculated using the formula:

$$R_k = n_k \times obs_k / tot_k$$

Where the index ***k*** refers to the specific codon, obs_k is the number of codon ***k*** in the sequence ***s***, tot_k is the number of codon ***k*** and all its synonymous codons in sequence ***s***, and n_k is the number of synonymous codons of the amino acid coded by codon ***k***.

If the synonymous codons of an amino acid are used with equal frequencies, their *RSCU* values will equal 1.

Converting codon usage values to RSCU values has the effect of 'normalising' comparisons across genes. It makes the codon usage value independent of amino acid composition of the sequences and identifies when a codon is being used more frequently than expected and when it is being used less frequently than expected.

2.4 (Feature Space)

There are 64 different codons. However, since methionine (AUG) and tryptophan(UGG) have only 1 corresponding codon, they are not considered and are removed from analysis as their RSCU values are always equal to 1; three stop codons (UGA, UAA, UAG) are also not included. Therefore, the number of codons considered is 59: i.e., $k = 1, 2 \ldots 59$. Therefore, irrespective of the size, the DNA sequence is converted to a feature vector of 59 elements.

2.5 (Support Vector Machine (Binary classification))

SVM is a supervised machine learning method which is based on the statistical learning theory [18]. When used as a binary classifier, an SVM will construct a hyper plane, which acts as the decision surface between the two classes. This is achieved by maximizing the margin of separation between the hyper plane and those points nearest to it. The SVMs were implemented using freely downloadable software, libSVM [3]. In this software there is a facility to define parameters and choose among various inbuilt kernels. They can be radial basis function (RBF) or a polynomial kernel (of given degree), linear, sigmoid. The generalization capability of SVMs makes it an effective approach in clustering and well suitable for gene classification, so we adopt it in the classification of Genotypes of HCV with codon usage pattern as inputs.

2.6 (SVM software; LIBSVM)

Simulations were preformed using LIBSVM version 2.89 (a freely available software package) [2, 3]. For our study RBF Kernel was found to be the best. The SVM training was carried out by the optimization of the value of the regularization parameter and the value of RBF kernel parameter.

2.7 (Cross-validation)

In order to evaluate the generalization power of each of the classification methods and to estimate their prediction capabilities for unknown samples, we used a standard 10- fold cross-validation technique and split the data randomly and repeatedly into training and test sets. The training sets consisted of randomly chosen subsets containing 90% of each class; the remaining 10% of the samples from each class were left as test sets. In order to keep computing times reasonable, we reported accuracy and standard deviation estimates over 100 runs. More runs are required if more accurate estimates are desired. We also reported the accuracy of prediction using the prediction set which are never used for model training.

[III] RESULTS AND DISCUSSIONS

3.1 (Prediction System Assement)

True positives (TP) and true negatives (TN) were identified as the positive and negative samples, respectively. False positives (FP) were negative samples identified as positive. False negatives (FN) were positive samples identified as negative. The prediction performance was tested with sensitivity (TP/ (TP+FN)),

specificity (TN/ (TN+FP)), overall accuracy (x), and MCC (Mathew's correlation coefficient).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$MCC = \frac{TP \times TN - FP \times FN}{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$$

Realizing the importance of Genotype in predicting HCV medical treatment response, treatment duration and the dose of ribavirin, we have chosen different Genotype of HCV for our study, hence develop a SVM Model for the Classification of HCV isolates of unknown genotype [19]. Altogether 1864 sequences are obtained from the LANL HCV database and the RSCU values of respective sequences are calculated by using our own toolkit for RSCU Calculation developed using PERL, where we took 59 RSCU feature vector corresponding to different codon except Methionine, Tryptophan and Stop Codons and gave as an input in SVM to get a Model. For sub-class classification of different Genotypes of HCV in major 6 classes, multi-class SVM are implemented using SVMlight, which usually lead to faster convergence in large optimization problems.

The Gaussian kernel is selected $K(\mathbf{x}, \mathbf{y}) = e^{-\$|x-y|2}$ for multi-class SVM methods. The Gaussian kernel shows superior performance over the linear and polynomial kernels as the Gaussian kernel results in complex (but smooth) decision functions and therefore has the ability to better fit the data where a simple discrimination by using a hyper plane or a low-dimensional polynomial surface is impossible.

The multilevel support vector classification procedure uses a discriminative classifier (a linear support vector machine with cross validation to build models, for RSCU and Amino acid composition). The discriminative classifier is a computational tool that is designed to classify an unknown sample as belonging to one of the classes. The purpose for building the classifier is accurately classifying input strains, a problem commonly referred to as the feature selection problem. For the initial step RSCU and amino acid composition of complete genome of HCV versus negative sequences of other virus was found that separates the two classes (HCV and non-HCV).Further HCV was divided into classes by using a dataset in which one class was considered as positive and others were treated as negative and the same procedure was repeated for each and every class thus separating HCV Genotypes into six subclasses.

The results of multi class SVM approach to classify its Genotype using Relative Synonymous Codon usage showing its Specificity, Sensitivity, Accuracy and MCC.

[Table-2]

Table 2: MCC, Accuracy, Sensitivity, Specificity, Recall and Precision for subclasses of Genotypes of HCV using Relative Synonymous Codon Usage.

SNo	Genotype	MCC	Accuracy	Sensitivity	Specificity	Recall	Precision
1.	Genotype1	0.0012686	99.41%	1	0.9882	94.43%	95.11%
2.	Genotype2	0.08944	100%	1	1	20.69%	13.95%
3.	Genotype3	0.19245	100%	1	1	22.22%	12.50%
4.	Genotype4	0.125	100%	1	1	13.33%	6.67%
5.	Genotype5	1	100%	1	1	70.6%	75.84%
6.	Genotype6	0.04419	100%	1	1	37.50%	24.49%

The best Accuracy and MCC of the RSCU and amino acid composition-based classifier is obtained for Genotype 5 which was 100% and 1, respectively. It proved that our model can better classify this class with high accuracy and precision.

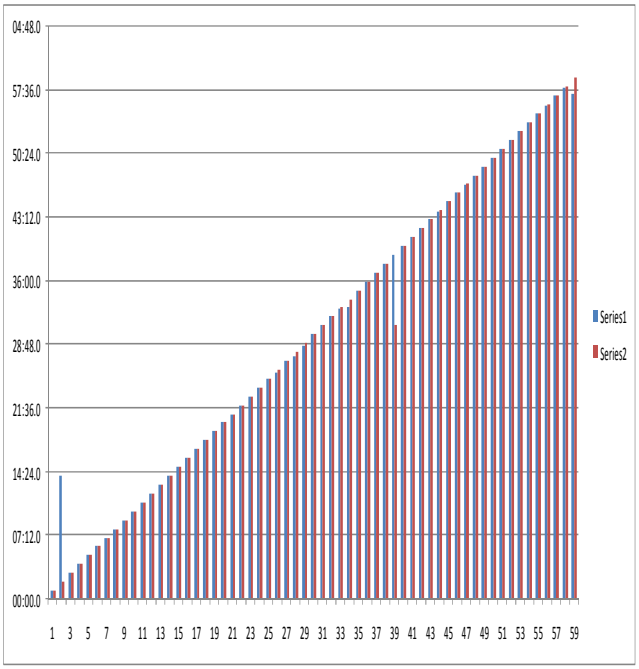

Figure 1 :

Also, when analyzed graphically it was found that RSCU when taken as parameter for classification gives better results than the amino acid composition. For example in Figure 1 RSCU values for 2, 39, 59 corresponding to Serine, Lysine and Glycine are showing more variations as compared to other codons and are responsible for classifying the sequences in a better way, whereas most of the instances in case of Amino Acid Composition are not showing much variation resulting in the lower accuracy of the results so found as compare to RSCU.

[IV] CONCLUSION

Gene classification has an important basis for the prediction of the functions of unknown genes, which is useful to classify large amount of genes into a few biologically meaningful groups to infer valuable information from genome data. The early methods include homology-based methods that require alignments of the sequences. Because of the time and space complexity of multiple sequence alignment of large sequences, it is difficult to classify large number of sequences through multiple sequence alignment. Further, if the given sequences are of significantly different lengths, the appropriate alignment is impossible. The use of codon usage pattern analysis simplifies the input into a 59-dimention vector, and avoids the increasing of computing complexity that happens to multiple sequence alignment. So, classification of HCV sequences by using codon usage pattern as the parameter is a better solution for comparing the homology based approaches.

.

This model can be used to analyze other proteins, such as entire proteomics data. Such type of prediction systems can be very useful for understanding the different proteins in a better way so as in conclusion, a novel computational method for classifying Genotypes and Structural proteins is presented here. This method will nicely complement the existing wet lab methods. It will assist in assigning the correct class to which these proteins belong. The prediction method presented here may be useful for the annotation of the piled-up proteomic data.

The author awaits discovery of more of these proteins in the future so that accuracy of the prediction model can be increased further and a server developed for public use.

ACKNOWLEDGEMENT

The authors are highly thankful to the Department of Bioinformatics MANIT, Bhopal, M.P, and India for providing support in the form of Bioinformatics infrastructure facility to carry out this work.

REFERENCES

[1] Alter H. J., Seeff L. B., Recovery, persistence, and sequelae in hepatitis C virus infection: a perspective on long-term outcome. Semin. Liver Diseases, 20: 17 – 35, 2000.

[2] Cai Y.D., et al, “Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect,” J. Cell. Biochem, vol. 84, pp. 343–348, 2002.

[3] Chang C.-C. and Lin C.-J., "LIBSVM: a library for support vector machines. Software,"available http://www.csie.ntu.edu.tw/~cjlin/libSVM, 2001.
[4] Charif D , Thioulouse **J**, Lobry J. R and Perrière G,"Online synonymous codon usage analyses with the ade4 and seqinR packages", Bioinformatics Oxford Journal,2005 ,21(4):545-547.
[5] Choo Q. L., Kuo G., Weine A.J.r, Overby, Isolation of a cDNA clone derived from a blood-borne non- A, non-B viral hepatitis genome, Science ,pp244: 359 – 362,1988.
[6] Choo Q.L., et al, Genetic organization and diversity of the hepatitis C virus, Proceedings of National Academic Science. USA , 88: 2451 – 2455,1991.
[7] Chou K.C.et al, "Using functional domain composition and support vector machines for prediction of protein subcellular location," J. Biol. Chem, vol. 277, 45765–45769, 2002.
[8] Cooper S., et al., Analysis of a successful immune response against hepatitis C virus, Immunity,10: 439 – 449,1999.
[9] Combet C, Penin F, Geourjon C, Deleage G: HCVDB: hepatitis C virus sequences database. Appl Bioinformatics 2004, 3:237-240.
[10] Cristianini N, Shawe-Taylor J: An Introduction to Support Vector Machines. Cambridge University Press, Cambridge, UK; 2000.
[11] Grakoui A.,et al Expression and identification of hepatitis C virus polyproteincleavage products. Journal of Virology. 67: 13851395, 1993.
[12] Gruner N. H., et al, Association of hepatitis C virus-specific CD8+ T cells with viral clearance in acute hepatitis C, Journal Of Infective. Diseases, 181: 1528 – 1536 ,2000.
[13] Habersetzer F., A. Fournillier and J. Dubuisson, Characterization of human monoclonal antibodies specific to the hepatitis C virus glycoprotein E2 with in vitro binding neutralization properties, Virology, 249:32– 41,1998.
[14] Hoofnagle JH, Wahed AS, Brown RS Jr, Howell CD, Belle SH;, Virahep- C Study Group: Early changes in hepatitis C virus (HCV) levels in response to peginterferon and ribavirin treatment in patients with chronic HCV genotype 1 infection. J Infect Dis 2009, 199:1112-20.
[15] Jianmin Ma, Minh N. Nguyen, Gavyn W. L. Pang, and Jagath C. Rajapakse, "Gene Classification using Codon Usage and SVMs", IEEE, 2005.
[16] Kuiken C, Yusem K, Boykin L, Richardson R: The Los Alamos hepatitis C sequence database. Bioinformatics 2005, 21:379-384.
[17] McHutchison J.G., et al, Interferon alpha-2b alone or in combination with ribavirin as initial treatment for chronic hepatitis C: Hepatitis Interventional Therapy Group, New England Journal of Medicine , 339: 1485 – 1492 ,1998.
[18] Ping Qiu, Xiao-Yan Cai,Wei Ding, Qing Zhang, Ellie D Norris,and Jonathan R Greene, "HCV genotyping using statistical classification approach", J Biomed sci, 16(1), 2009.
[19] Tokita H, et al: Hepatitis C virus variants from Thailand classifiable into five novel genotypes in the sixth (6b), seventh (7c, 7d) and ninth (9b, 9c) major genetic groups. J Gen Virol. 1995, 76(Pt 9):2329-2335.