International Journal of Bioinformatics Research
ISSN: 0975-3087 & E-ISSN: 0975-9115, Volume 5, Issue 1, 2013, pp.-285-288.
Available online at http://www.bioinfopublication.org/jouarchive.php?opt=&jouid=BIJ0000001

ANALYSIS AND PREDICTION OF MAJOR BLOOD PROTEINS BASED ON THEIR AMINO ACID AND DIPEPTIDE COMPOSITION

MUTHUKRISHNAN S.[1,2], PURI M.[1,3]* AND LEFEVRE C.[3]

[1]Protein Biotechnology Laboratory, Department of Biotechnology, Punjabi University, Patiala- 147002, Punjab, India.
[2]Institute of Microbial Technology, Sector-39A, Chandigarh- 160036, India.
[3]Centre for Chemistry and Biotechnology, Deakin University, Victoria, Australia.
*Corresponding Author: Email- munish.puri@deakin.edu.au

Received: August 08, 2013; Accepted: September 18, 2013

Abstract- A method has been developed for predicting blood proteins using the SVM based machine learning approach. In this prediction method a two-step strategy was deployed to predict blood proteins and their subclasses. We have developed models of blood proteins and achieved the maximum accuracies of 90.57% and 91.39% with Matthews correlation coefficient (MCC) of 0.89 and 0.90 using single amino acid and dipeptide composition respectively. Furthermore, the method is able to predict major subclasses of blood proteins; developed based on amino acid (AC) and dipeptide composition (DC) with a maximum accuracy 90.38%, 92.83%, 87.41%, 92.52% and 85.27%, 89.07%, 94.82%, 86.31 for albumin, globulin, fibrinogen, and regulatory proteins respectively. All modules were trained, tested, and evaluated using the five-fold cross-validation technique.

Keywords- Major Blood Proteins, Amino Acid Composition, Dipeptide Composition, SVM, five-fold cross-validation technique

Citation: Muthukrishnan S., Puri M. and Lefevre C. (2013) Analysis and Prediction of Major Blood Proteins Based on Their Amino Acid and Dipeptide Composition. International Journal of Bioinformatics Research, ISSN: 0975-3087 & E-ISSN: 0975-9115, Volume 5, Issue 1, pp.-285-288.

Introduction

Determination of protein functions is one of the most challenging problems in the genomic era. Enormous amounts of protein sequences are available in the database as raw sequence data. Using various computational approaches, these data's are being processed into meaningful biological information's. The support vector machine (SVM) based prediction system is fully automatic and reliable. It has been used in many applications including sub-cellular localization, protein secondary structure prediction, and micro array data analysis of proteins [1-4]. However, no direct method is currently available to predict blood proteins. Thus analysis and prediction studies of blood proteins are important for researchers.

Blood proteins found in blood plasma are also called serum proteins. Major blood proteins are albumin, globulin, fibrinogen, and regulatory proteins [5,6]. Sixty percent of plasma proteins are made up of albumins [7], which are major contributors to the osmotic pressure of plasma and which assists in the transport of lipids and steroid hormones. Globulins make up thirty five percent of plasma proteins and are used in the transport of ions, hormones, and lipids thus assisting in immune function [8]. Four percent is fibrinogen and it is essential in the clotting of blood when converted into insoluble fibrin [9]. Regulatory proteins, which make up less than one percent of plasma proteins, are proteins such as enzymes, proenzymes and hormones. The main functions of the blood proteins are transporting lipids, hormones, vitamins and metal molecules. Thus, these proteins are playing an important role in the regulation of a cellular activity and many different functions in the immune system. Due to their great function of blood, a classification prediction system has been developed in order to facilitate better understanding of their roles. The superior facility of classification system using machine learning based approach rather than experimental techniques is apparent. Currently, there is no classification of blood proteins available based on amino (AC) and dipeptide composition. Support vector machine (SVM) is one of the promising kernel based machine learning for building effective model for predicting class labels of unknown protein data. Therefore, in this study, we have developed an integrative SVM based prediction with a two step approach to predict the blood proteins and further classify them into different classes. The method presented is highly specific and sensitive to predict the blood proteins.

Results

To develop a prediction platform for blood proteins, amino acid composition dataset of all blood-proteins were made using five fold cross-validation technique. The SVM module was developed using blood-protein and non-blood protein training sets. Dataset was divided into five equal sets randomly, four sets were used for training, and the remaining set was used for testing [10,11]. This process was repeated five times to test each protein at least once [12]. We labeled all blood-proteins as positive proteins and non blood-proteins were used as negative proteins. The results demonstrated that the method can differentiate blood-proteins from non blood-proteins with great accuracy of 90.57% in 0.89 of MCC at a default cutoff score of 0. The best result was obtained using an optimal RBF kernel with parameters g =3, C 375. The dipeptide composition method was also tested and achieved 91.39% accuracy with 0.90 of the MCC. In average, all amino and dipeptide composition analysis of blood-proteins were significantly different from non blood-proteins [Table-1]. We have applied the same method for predicting the classification of major blood-proteins. Here we took one class of blood proteins as positive and all other classes for negative examples. This was repeated to all other classes of blood proteins. We prepared models of each blood protein classes based on their amino acid as well as dipeptide composition with different optimized SVM kernels parameters. This indicates that each class of blood-proteins can be discriminated from other classes of proteins based on their amino acid and dipeptide composition [13-16]. In amino acid composition we achieve maximum accuracy as shown in the [Table-2], 90.38%, 92.83%, 87.41%, 92.52% with MCC of 0.88, 0.85, 0.63, and 0.43 of albumin, globulin, fibrinogen, and regulatory proteins respectively. Regulatory and globulin proteins are showing maximum accuracy (>92%) compared to other classes of blood proteins. Since the simple amino acid composition provides only information about frequency but not about local order of residues, additional SVM module based on dipeptide composition were developed. While in amino acid composition SVM were provided with an input vector of dimention 20 for amino acid composition, a vector of 400 dipeptide composition was used for the dipeptide module which was optimized by g =10, C=400. The results are shown in [Table-2] and we achieved the maximum accuracy of 85.27%, 89.07%, 94.82%, 86.31% to 0.87, 0.88, 0.87, and 0.70 of the MCC for albumin, globulin, fibrinogen, and regulatory proteins respectively. The fibrinogen protein shows the highest accuracy 94.82% among subclass proteins.

Table 1- Performance of various SVM modules of blood-protein predictions developed using various types of compositions; amino acids (AC) and dipeptides (DC).

Methods	ACC(%)	SN(%)	SP(%)	MCC	Parameters	
					γ	C
AC	90.57	97.16	84.69	0.89	3	375
DC	91.4	96.77	86.6	0.9	10	400

AC- Amino acid composition, DC- dipeptide composition, ACC- accuracy, SN- Sensitivity, SP- specificity, MCC- Matthews correlation coefficient, C- tradeoff value, γ- gamma factor (a parameter in RBF kernel)

Table 2- Performance of various SVM modules of blood-protein classifications (albumin, globulin, fibrinogen and regulatory) predictions developed using various types of compositions; amino acids (AC) and dipeptides (DC).

Proteins	Methods	ACC(%)	SN(%)	SP(%)	MCC	Parameters	
						γ	C
Albumin	AC	90.38	88.89	90.6	0.88	15	350
	DC	85.27	94.1	84	0.87	2	375
Globulin	AC	92.83	80.21	93.39	0.85	50	350
	DC	89.07	89.58	89.05	0.88	5	200
Fibrinogen	AC	87.41	97.94	49.4	0.63	100	75
	DC	94.82	99.92	76.03	0.87	3	450
Regulatory	AC	92.53	27.5	94.88	0.43	15	450
	DC	86.32	66.25	87.05	0.7	25	500

AC- Amino acid composition, DC- dipeptide composition, ACC- accuracy, SN- Sensitivity, SP- specificity, MCC- Matthews correlation coefficient, C- tradeoff value, γ- gamma factor (a parameter in RBF kernel)

The dipeptide prediction accuracy was improved significantly over single amino acid prediction. Therefore, the prediction accuracy can be increased using a wide range of information about a protein. The sensitivity and specificity also has been calculated for blood proteins and subclasses, shown in [Table-1] & [Table-2].

Analysis of Amino Acids

Determining the relative amino acid composition of a protein will give a characteristic profile for protein. This amino acid analysis profile provides enough information to identify major blood-proteins. Here, we used the total number of amino acid divided by the total number of amino acids in protein. The average amino acid composition of blood proteins has been calculated which shown in [Fig-1] and [Fig-2] with non blood proteins. In this analysis results shows that Cys, Pro, Ser and Tyr are higher in blood proteins than the non blood proteins. In sub classes of blood proteins Leu residues are higher in all classes, Cys, Asp, Gly and Asn are higher in fibrinogen proteins than other classes. Overall regulatory and globulin protein is having a similar percentage in all residues.

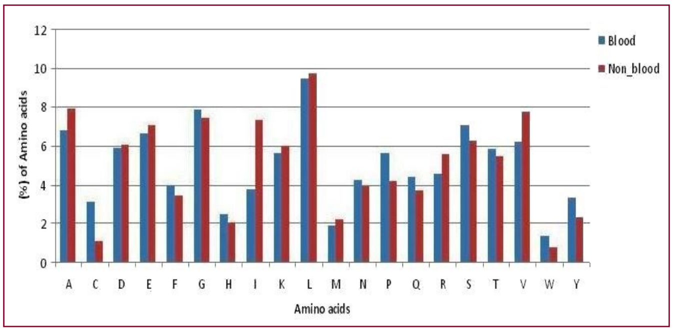

Fig. 1- Average Amino acid composition chart of blood vs. non-blood proteins

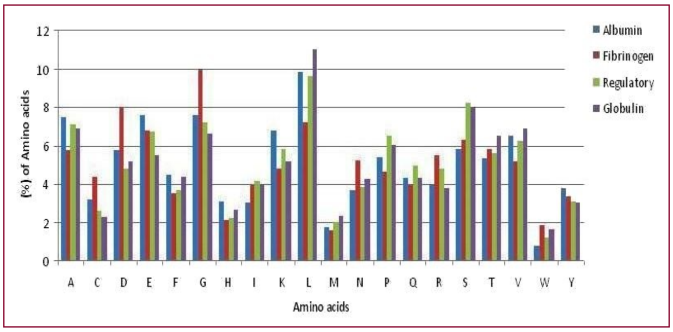

Fig. 2- Average amino acid percentage chart of subclasses blood proteins.

Discussion

Blood proteins serve many different functions, such as circulatory transport molecules for (lipid hormones, vitamins, and metals), protease inhibitors, and regulation of cellular activities, including the immune system. Based on their importance, we have decided to develop a method using SVM for prediction of these proteins. Here, we have described amino and dipeptide based method which is helpful in differentiating various blood proteins. This method is usefull to show whether a newly discovered protein sequence belongs to blood proteins and identify its subfamilies. This method is a highly accurate method and able to perform classification separation properly. Finally, our results demonstrate that using the concept of amino and dipeptide using SVM is a successful method for predicting these proteins.

Methods

Dataset

The final data set of blood proteins including subfamilies consist 717 (albumin 91, globulin 33, fibrinogen 564 and regulatory 29). As negative set 899 belonging to protease family were selected randomly. These protein sequences were obtained from Uniprot and Expasy server. In this dataset “fragments”, “isoforms”, “potentials”, “similarity”, or “probables” in comment field were removed, to avoid bias in the classifier. We have used 90% cutoff to generate non-redundant dataset of both blood and non-blood sequences.

Support Vector Machine (SVM)

In the present study, a free downloadable package of SVM, SVM_light has been used to classify major blood-protein sequences. This software enables the users to define a number of parameters as well as the choice of inbuilt kernel, such as a radial basis function (RBF) or a polynomial kernel (of given degree) [17]. In this study, all parameters of a kernel were kept constant, except for the regulatory parameter C. The experimentation was conducted by using various types of kernels such as polynomial and radial base function. The SVMs required a fixed number of inputs for training, thus necessitating a strategy for encapsulating the global information about the proteins of variable length in a fixed length format. The fixed length format was obtained from protein sequences of variable length using amino acid and dipeptide composition. It has been successfully applied to numerous classification and pattern recognition problems such as classification of microarray data, protein secondary structure prediction and sub cellular localization [3,4, 18].

Amino Acid Composition

Amino acid composition is the fraction of each amino acid in a protein. The fraction of all 20 natural amino acids was calculated by using [Eq-1] [19,20].

$$\text{Fraction of amino acid (i)} = \frac{\text{Total number of amino acid (i)}}{\text{Total number of amino acids in protein}} \quad (1)$$

where *i* can be any amino acid.

Dipeptide Composition

Dipeptide composition is used to encapsulate the global information about each protein sequence, which gives a fixed pattern length of 400 (20 X 20). The fraction of each dipeptide was calculated using [Eq-2] [19].

$$\text{Fraction of dipeptide (i)} = \frac{\text{Total number of dipeptides (i)}}{\text{Total number of all possible dipeptides}} \quad (2)$$

where dep (i+1) is one out of 400 dipeptides.

Evaluation of Performance

In this study, 5-fold cross-validation technique was adopted according to which dataset was partitioned randomly into five equal subsets. The training and testing were carried out five times, each time using one subset for testing and remaining 4 subsets for training. The performance of each classifier is measured in terms of accuracy (ACC), sensitivity (SN), specificity (SP) and Matthews correlation coefficient (MCC) by standard [Eq-3] to [Eq-6] [20].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (5)$$

$$MCC = \frac{TP \text{ X } TN - FP \text{ X } FN}{\sqrt{(TP + FP)\,(TP + FN)(TN + FP)\,(TN + FN)}} \quad (6)$$

where, TP, TN, FP, FN’s are the number of true positives, true negatives, false positives and false negatives respectively. TP and TN are number of correctly classified blood proteins and non blood proteins respectively. FP and FN are incorrectly classified as blood and non blood proteins.

Prediction System

The prediction of blood and blood related proteins are a multi-class classification problem. To handle this multi-class situation, we have to design a series of binary SVMs. For *N* class classification, *N* SVMs was constructed. The *i*th SVM will do training with all samples of the *i*th subfamily being labeled as positive, and the samples of all other subfamilies being labeled as negative. The SVMs trains in this way will reefer to as 1-v-r SVMs. In this classification approach, each of the unknown proteins will achieve four scores. An unknown protein will be classified into the subfamily that corresponds to the 1-v-r SVM with the highest output score.

Acknowledgements

The authors are thankful to the Department of Biotechnology, Punjabi University, Patiala and Institute of Microbial Technology, Chandigarh, India, for providing necessary facility to conduct research work. The authors acknowledge Deakin University, Australia for providing web server for this study.

Conflict of Interest: None declared.

References

[1] Park K.J., Kanehisa M. (2003) *Bioinformatics*, 19, 1656-1663.

[2] Yu C.S., Chen Y.C., Lu C.H., Hwang J.K. (2006) *Proteins*, 64, 643-651.

[3] Chang D.T., Ou Y.Y., Hung H.G., Yang M.H., Chen C.Y., Oyang Y.J. (2008) *BMC Res. Notes*, 1(1), 51.

[4] Watson M., Dukes J., Abu-Median A.B., King D.P., Britton P., Detecti V. (2007) *Genome Biol.*, 8(9), R190.

[5] Anderson N.L. and Anderson N.G. (1977) *Proceeding of the National Academy of Sciences*, 74, 5421-5425.

[6] Adkins J.N., Varnum S.M., Auberry K.J., Moore R.J., Angell N.H., Smith R.D., Springer D.L., Pounds J.G. *(2002) Molecular and Cellular Proteomics*, 1, 947-955.

[7] Zunszain P.A., Ghuman J., Komatsu T., Tsuchida E., Curry S. (2003) *BMC Structural Biology*, 3(1), 6 10.1186/1472-6807-3-6.

[8] Roux K.H. (1999) *Int. Arch Allergy Immunol.*, 120, 85-99.

[9] Laurens N., Koolwijk P., de Maat M.P. (2006) *J. Thromb. Haemost.*, 4(5), 932-939.

[10]Bhasin M., Garg A., Raghava G.P.S. (2005) *Bioinformatics*, 21, 2522-2524.

[11]Kumar M., Gromiha M.M., Raghava G.P.S. (2007) *BMC Bioinformatics*, 8, 463.

[12]Garg A., Bhasin M., Raghava G.P.S. (2005) *J. Biol. Chem.*, 280, 14427-14432.

[13]Yu C.S., Chen Y.C., Lu C.H., Hwang J.K. (2006) *Proteins*, 64, 643-651.

[14]Cai C.Z., Han L.Y., Ji Z.L., Chen X., Chen Y.Z. (2003) *Nucleic Acids Research*, 31, 3692-3697.

[15]Muthukrishnan S., Garg A., Raghava G.P.S. (2007) *Genomics, Proteomics & Bioinformatics*, 5, 250-252.

[16]Lata S., Sharma B.K., Raghava G.P.S. (2007) *BMC Bioinformatics*, 8, 263.

[17]Chou K.C., Shen H.B. (2007) *Biochem Biophys Res. Commun.*, 360, 339-345.

[18]Nair R., Rost B. (2003) *Proteins*, 53(4), 917-930.

[19]Bhasin M., Raghava G.P.S. (2004) *J. Biol. Chem.*, 279, 23262-23266.

[20]Kumar M., Verma R., Raghava G.P.S. (2006) *J. Biol. Chem.*, 281, 5357-5363.