DiRiboPred: A Web Tool for Classification and Prediction of Ribonucleases

Bhasker Pant[1] K.R. Pardasani[2]

GJCST Computing Classification
J. 3

***Abstract*-Ribonuclease [commonly abbreviated RNase] is a type of nuclease that catalyzes the degradation of RNA into smaller components and can be divided into endoribonucleases and exoribonucleases. All organisms studied contain many RNases of many different classes, showing that RNA degradation is a very ancient and important process. They are shown to have an important role in cancer, tumor and many neuro degenarative disorders for controlling which, in-silico drug designing can be a valuable tool. In the past machine learning has been used to classify other proteins like GPCRs but no attempt has been made for classification of Ribonucleases. Realizing their importance here an attempt has been made to develop an SVM model to predict, classify and correlate all the major subclasses of ribonucleases with their dipeptide composition. The method was trained and tested on 1857 proteins of ribonucleases. The method discriminated Ribonucleases from other enzymes with Matthew's correlation coefficient of 1.00 and 100% accuracy. In classifying different subclasses of Ribonucleases with dipeptide composition, an overall accuracy of 94.534% was achieved. The performance of the method was evaluated using 5-fold cross-validation. A web server DiRiboPred has been developed for predicting Ribonucleases from its amino acid sequence http://www.bif.manit.org/RiboPred2.**

Keywords-Classifier,Dipeptide,Composition,Ribonucleases, Support Vector Machine.

I. INTRODUCTION

All organism studied to date contain many kind of ribonucleases of different classes. They have a role to play not only in cleaning of cellular RNAs that is no longer required, but also in the maturation of
all RNA molecules, both messenger RNAs and non-coding RNAs that function in varied cellular processes. Besides, an active RNA degradation system are a first defense against RNA viruses, and provide the underlying machinery for more advanced cellular immune strategies such as RNAi [1]. RNases play a critical role in many biological processes, including angiogenesis and self-incompatibility in flowering plants [angiosperms]. Also, RNases in prokaryotic toxin-antitoxin systems are proposed to function as plasmid stability loci, and as stress-response elements when present on the chromosome [1].

A Endoribonuclease is a ribonuclease endonuclease which cleaves RNA molecule by attacking the internal bonds. Major types of endoribonucleases are RnaseA, RnaseH, RnaseI, RnaseIII, RnaseL, RnaseP, RnasePhyM, RnaseT1, RnaseT2, RnaseU2, RnaseV1 and RnaseV. An exoribonuclease is an exonuclease ribonuclease, which are enzymes that degrade RNA by removing terminal
nucleotides from either the 5' end or 3' end of the RNA molecule. Enzymes that remove nucleotides from the 5' end are called 5'-3' exoribonucleases and enzymes that remove nucleotides from the 3' end are called 3'-5' exoribonucleases. The major types of exoribonucleases are PNPase, RnasePH, RnaseIII, RnaseR, RnaseD, RnaseT, Oligoribonucleases, ExoribonucleaseI and ExoribonucleaseII [2].

These enzymes have an important role to play not only in normal body physiology but also in diseased conditions. Majority of ribonucleases have been implicated in tumors and cancers like ovarian cancer, melanoma and non-Hodgkin lymphomas. They can directly and indirectly influence cancer causation and spread. For ex. Ribonucleases have been shown to contribute significantly to telomerase inhibitory activity detectable in foregut cancer specimens [1].

These molecules with their immense capability to degrade RNA, on the contrary can be used for curing, treating and preventing many fatal diseases like cancer. Many variants of natural ribonucleases have been developed with minor changes in dipeptide composition with properties included but not limited to stability, cytotoxicity towards pathogenic cells, efficacy of degradation of pathogenic RNA of any origin including viral RNA, evasion of binding by RNase inhibitors, resistance to degradation by proteases, delivery to target cells, efficiency of import into the cell, dose response properties, pharmacokinetic properties, and longevity within the human body [2] . Due to their cytotoxic properties ribonucleases [RNases] can be potential anti-tumor drugs. Particularly members from the RNase A and RNase T1 superfamilies have shown promising results. Among these enzymes, Onconase, an RNase from the Northern Leopard frog, is furthest along in clinical trials [4].

Everyday many new ribonucleases are discovered the annotation and functional assignment of classes to these through wet lab techniques involve time consuming, laborious experiments, hence machine learning techniques like Support Vector Machines can be effectively used to complement them saving time, money and labor.

In this paper an attempt has been made to predict, classify and correlate these enzymes with their dipeptide composition by implementing SVM using SVM Light, a freely down loadable software [11]. This is a novel step where all the major classes of ribonucleases have been taken into consideration.

The typical strategies for identifying Ribonucleases and their types include similarity search based tools, such as BLAST, FASTA and motif finding tools. Although these

About-[1]Department of Bioinformatics, Manit, Bhopal, India
(telephone :+91-9425118244, email: pantbhaskar2@gmail.com)
About-[2]Department of Mathematics, Manit, Bhopal, India

tools are very successful in searching similar proteins, they fail when members of a subfamily are divergent in nature. Hence there arises a need for in-silico prediction which is not only quicker and economical but also accurate which has been proved in previous classifiers. Earlier SVM has been used for classification of biological data like GPCRs [5], Nuclear receptors [6] etc. but no attempt has been made for classification of these proteins.

On the basis of the above study, an online web tool 'DiRiboPred' has been made available at http:/www.bifmanit.org/RiboPred2. To the best of authors' knowledge, there is no web server that allows recognition and classification of Ribonucleases. Such kind of classifiers will help in annotation of piled up proteomic data and would complement the existing wet lab techniques.

II. Materials and Methods

A. Data Repository

Only non fragmented entries were obtained from SwissProt/Uniprot database of Expasy server [8]. The subclasses of Ribonuclease [Rnases] i.e. Endoribonucleases and Exoribonucleases have member families with much of the redundant data. A package of softwares called CD-HIT Suite [7] was used for removing the redundancy to 90% with sequence identity cutoff of 0.9. The final dataset has following number of proteins.

Endoribonucleases have following members with following number of instances after redundancy removal.

1] RNaseA [51]
2] RNaseH [682]
3] RNaseIII [230]
4] RNaseP [286]
5] RNaseT1 [15]
6] RNaseT2 [36]

Exoribonucleases have following

1] PNPase [338]
2] RNasePH [184]
3] ExoribonucleaseI [06]
4] ExoribonucleaseII [29]

For rest of the members the number of instances was very less hence they were not included in the classifier.

B. Proposed Methodology

The classifier, predictor works in 4 step methodology.

1] First checks whether the protein is ribonucleases or not.
2] If it is, then check whether it is exoribonuclease or endoribonuclease.
3] If endoribonucleases then to which subclass it belongs.
4] If exoribonucleases then to which subclass it belongs.

C. Recognition of Ribonucleases from rest

At first step, the main aim is the recognition of novel ribonucleases or discriminating ribonucleases from rest of the enzyme protein. A SVM was trained to discriminate the ribonucleases from other proteins. The training and testing was carried out on a dataset of 1857 proteins of ribonucleases. The training also required negative examples for discriminating ribonucleases from other proteins. Since ribonucleases can be divided into endoribonucleases and exoribonucleases, and comprise several sub-classes within the EC 2.7 [for the phosphorolytic enzymes] and 3.1 [for the hydrolytic enzymes] classes of enzymes, the dataset was extended by including 1857 enzymes other then class 3 and 2 to which ribonucleases belong . The final dataset has equal number of positive and negative examples, so that the performance of the method can be evaluated using single parameter, such as accuracy.

D. Classification of Endoribonucleases

Endoribonucleases has major six subclasses [RnaseA, RnaseH, RnaseIII,RNaseP, RnaseT1 and RnaseT2]. The dataset for classifying this class [endoribonucleases] consisted of 1300 sequences, of which 51 were RNaseA, 682 RNaseH, 230 RnaseIII, 286 RnaseP, 15 RnaseT1 and 36 for RnaseT2 type of enzymes. To achieve second step in our proposed methodology one SVM module with all the 1300 instances belonging to endoribonucleases labeled as positive and 611 sequences belonging to exoribonucleases marked as negative, was constructed. In the third step for classifying an unknown protein into one of the six types of endoribonucleases which is a multiclass classification problem, a series of binary classifiers were developed. Here, six SVMs were developed, one each for a particular type of endoribonucleases. The ith SVM was trained with all samples of the ith type enzymes with positive label and samples of all other types of enzymes as negative label. The SVMs trained in this way were referred as 1-v-r SVMs [5, 6, 9]. Explaining more explicitly, 6 datasets were prepared. The first dataset for subclass RnaseA comprised of 51 instances labeled positive and rest of the instances belonging to the other 5 were marked as negative, similarly in the second dataset for RnaseH 682 instances were labeled as positive and rest belonging to others as negative. The methodology was repeated for all the 6 subclasses.

In such classification, each of the unknown protein achieved six scores. An unknown protein was classified into the endoribonucleases type that corresponds to the 1-v-r SVM with highest output score.

E. Classification of Exoribonucleases

The same strategy was followed for exoribonuclease class which comprised of four major subclasses [PNPase, RnasePH, ExoribonucleasesI and ExoribonucleasesII]. The dataset for this class consisted of 557 sequences with 338 members of PNPase subclasse, 184 members of RnasePH subclass, 06 members of ExoribonucleaseI and 29 members of ExoribonucleaseII subclasses. Here also one SVM module comprising 557 positive instances and 1300 negatively labelled instances was made. Like above, classification of exoribonucleases is also a multiclass classification problem where 4 SVMs were developed for distinguishing 4 subclasses of exoribonucleases as indicated in step 4 of proposed methodology. The ith SVM was trained with all samples of the ith type enzyme with positive label and samples of all other types of enzymes as negative label. The classification was achieved in the manner indicated above. In the final dataset all the instances

belonging to 10 subclasses were clubbed together i.e. 51+682+230+286+15+36+338+184++06+29=1857, and equal amount of negatives were taken which were proteins other then ribonucleases.

F. Support Vector Machine

Kernel-based techniques [such as support vector machines, Bayes point machines, kernel principal component analysis, and Gaussian processes] represent a major development in machine learning algorithms. Support vector machines [SVM] are a group of supervised learning methods that can be applied to classification or regression. Support vector machines represent an extension to nonlinear models of the generalized portrait algorithm developed by Vladimir Vapnik. The SVM algorithm is based on the statistical learning theory and the Vapnik-Chervonenkis [VC] dimension introduced by Vladimir Vapnik and Alexey Chervonenkis. A Support Vector Machine [SVM] performs classification by constructing an N-dimensional hyperplane that optimally separates the data into two categories. SVM models are closely related to neural networks. In fact, a SVM model using a sigmoid kernel function is equivalent to a two-layer, perceptron neural network [10].

Support Vector Machine [SVM] models are a close cousin to classical multilayer perceptron neural networks. Using a kernel function, SVM's are an alternative training method for polynomial, radial basis function and multi-layer perceptron classifiers in which the weights of the network are found by solving a quadratic programming problem with linear constraints, rather than by solving a non-convex, unconstrained minimization problem as in standard neural network training [10].

G. SVM implementation: SVM Light and LibSVM

For implementing SVM, a software called SVMLight developed by Joachims et.al. has been used [11]. In this software there is inbuilt facility for choosing among many kernel types and their parameters. For our study RBF kernel was found to be the best. This kernel nonlinearly maps samples into a higher dimensional space so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear. Furthermore, the linear kernel is a special case of RBF the linear kernel with a penalty parameter C has the same performance as the RBF kernel with some parameters [C, γ]. In addition, the sigmoid kernel behaves like RBF for certain parameters [12], [13]. The second reason is the number of hyperparameters which influences the complexity of model selection. The polynomial kernel has more hyperparameters than the RBF kernel. Finally, the RBF kernel has fewer numerical difficulties. There are two parameters for an RBF kernel: C and γ. It is not known beforehand which C and γ are best for a given problem; consequently some kind of model selection [parameter search] must be done. The goal is to identify good [C, γ] so that the classifier can accurately predict unknown data [i.e. testing data] [14]. For obtaining the value of these two parameters software called LIBSVM by Chih-Chung Chang and Chih-Jen Lin was used. LIBSVM is an integrated software for support vector classification, [C-SVC, nu-SVC], regression [epsilon-SVR, nu-SVR] and distribution estimation [one-class SVM]. It supports multi-class classification [15]. The value of these two parameters was fed into SVMLight and analysis done.

H. The Attribute

For analyzing these protein enzymes attribute of dipeptide composition was used. Previously this characteristics has been used for solving other biological problems like classification of nuclear receptors, mycobacterial proteins [16], cytokines [17], GPCRs [5] and virulent proteins to name a few, but it has not been used till now for classifying and predicting ribonucleases.

One of the problems with protein classification is that proteins are variable in length. Dipeptide composition provides the information of protein in the form of a fixed vector of 400 dimensions. The dipeptide composition encapsulates the information about fraction of amino acids present in the protein but also local order in the proteins. Hence the information contained in the variable primary amino acid structure of the proteins is converted into a fixed length feature value vector, which is required by SVM and is used here for ribonucleases. The dipeptide composition is calculated by the following formula.

The dipeptide composition is calculated by the following formula.

Dep(i)=total no. of dep(i)/total number of all possible dipeptides

Where dep(i) is one out of 400 dipeptides.

I. PROCOS [protein composition server

For feeding the 400 dimensions of dipeptide composition into SVM, it is required to be converted into feature value vectors format eg. 1: 2: ----------400: etc. With servers like Copid, dipeptide composition can be easily found out but its manual conversion into feature vector form as indicated above is time consuming and laborious. ProCos is an integrated set of software with inbuilt facility of converting protein composition of any degree in the form required by the user. For our model, dipeptide composition of ribonucleases was calculated using the ProCos software. It is available at www.manit.ac.in/downloads/polycomp/ [18].

J. Evaluation of Performance

Here cross-validation was performed on the dataset. In limited cross-validation, a set of proteins is divided into M equally balanced subsets. The method was trained or developed on [[M − 1] N]/M proteins and then tested on the remaining N/M proteins. This process is repeated M times, once for each subset. In this study, the performance of dipeptide composition based classifiers was evaluated through 5-fold cross-validation [5], [6], [9]. The performance of the classifier developed at the first level [for recognizing proteins of ribonuclease] was evaluated using the standard threshold-dependent parameters, such as sensitivity, specificity, accuracy and Matthew's correlation coefficient [MCC]. The performance of classifiers for

classifying subclasses of ribonucleases was evaluated by measuring accuracy and MCC as described by Hua and Sun [9].

The LibSVM provides a parameter selection tool using the RBF kernel: cross validation via grid search. A grid search was performed on C and Gamma using an inbuilt module of libSVM tools as shown in figure1. Here pairs of C and Gamma are tried and the one with the best cross validation accuracy is picked.

K. Prediction System Assessment

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 [Q2], and the Matthews correlation coefficient [MCC]. The accuracy and the MCC for each subfamily of ribonucleases, was calculated as described by Hua and Sun [9] and shown below in equation 1 and 2.

$$Accuracy(x) = \frac{tp+tn}{tp+tn+fp+f} \quad \text{EQ.1}$$

$$MCC = \frac{(tp)(tn)-(fp)(fn)}{\sqrt{(tp+fp)}(tp+fn)(tn+fp)(tn+fn} \quad \text{EQ.2}$$

III. Result and Discussion

The performance of the method in distinguishing ribonucleases enzyme from other enzymes and also various subclasses is shown in Table 1 for training data set. The performance of the method is evaluated using a 5-fold cross-validation. This demonstrates that ribonucleases can be distinguished from other proteins on the basis of dipeptide composition with 100% accuracy. Prediction of endoribonucleases from exoribonucleases also reached higher accuracy of 99.89%.On using the RBF kernel with value of parameters [γ = 0.0078125 and C = 0.03125] an accuracy of 100% was obtained in distinguishing ribonucleases from rest of the proteins. The tabulated C and Gamma values for predicting various classes and subclasses of Ribonucleases for training dataset are given in Table1. The results are also consistent with our previous observation that dipeptide composition is better in classifying the proteins. The dipeptide composition is a better feature to encapsulate the global information about proteins as it provides information about fraction of amino acid contained in the protein as well as their local order. Furthermore, to classify different types of ribonucleases, a series of binary SVMs were constructed. The separate SVM modules have been developed for each type of nucleases of exoribonuclease and endoribonuclease family and the accuracy of classifying different subclasses for test dataset is indicated in Table2. The average accuracy of dipeptide composition based classifier is 94.534% for the above. This proved that dipeptide composition is an important feature not only for recognizing but also for classifying different types of the ribonucleases. This observation can also be extended to other types of enzymes by establishing good training data.

These results suggest that types of Ribonucleases are predictable to a considerably accurate extent with dipeptide composition. The development of such accurate and fast methods will speed up the identification of drug targets for curing various cancers and also will be helpful in formulation of newer ribonucleases with cytotoxic properties. For many of the classes like RnaseA, RnaseT1, RnaseT2, Exoribonuclease1 and ExoribonucleaseII since the dataset was small, quiet variation was seen in the values of Precision and MCC which were lesser then others.

Fig.1. Coarse Grid Search on C = 2-5, 2-4 … 210 and Gamma = 25, 24 … 2 - 10 [Adapted from Xu et al. 2004].

Table1. C and Gamma values for training set of Ribonucleases with accuracies

Modules	C	Gamma	Accuracy
Ribonucleases	0.03125000	0.00781250	100.00%
Endoribonucleases	8.00000000	2.00000000	99.89%
Exoribonucleases	32.00000000	0.00781250	99.96%
ExoribonucleaseI	128.00000000	0.00781250	99.96%
ExoribonucleaseII	0.03125000	2.00000000	99.86%
PNPase	8.00000000	0.00781250	100.00%
RnaseIII	8.00000000	2.00000000	99.72%
RnaseA	512.00000000	0.03125000	99.72%
RnaseH	32	0.0078125	99.97%
RnaseP	32.00000000	0.0078125	99.89%
RnasePH	0.5	0.50000000	100.00%
RnaseT1	512	0.000122	97.89%
RnaseT2	32	0.0078125	98.99%

Table2. Statistical detail and testing accuracies for various classes and subclasses in the testing dataset.

S.No	Ribonuclease	Accuracy	Precision	Tp	Tn	Fp	Fn	Specificity	MCC
1	Pnpase	95.00%	97.00%	120	60	5	4	0.92	0.89
2	Ribonuclease 1	93.12%	75.00%	6	170	11	2	0.94	0.49
3	Ribonuclease 11	93.93%	96.66%	29	147	12	1	0.92	0.8
4	Rnase111	90.00%	99.00%	100	70	10	9	0.88	0.79
5	RnaseA	95.76%	92.72%	51	130	4	4	0.97	0.9
6	RnaseH	95.00%	96.00%	130	50	4	5	0.93	0.89
7	RnaseP	96.29%	96.94%	127	55	3	4	0.95	0.91
8	FinalrnasePH	99.47%	99.29%	140	44	3	2	0.94	0.93
9	FinalrnaseT1	93.65%	88.88%	16	161	10	2	0.94	0.71
10	FinalrnaseT2	93.12%	90.62%	29	147	10	3	0.94	0.78

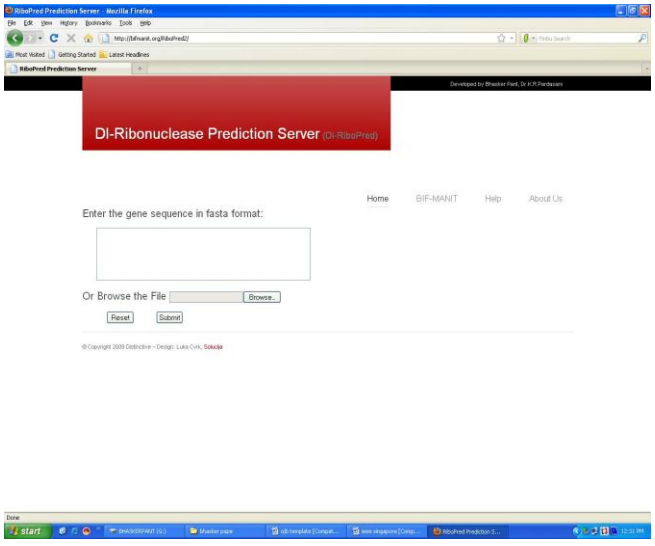

Fig.2. Coarse Grid Search on C = 2-5, 2-4 … 210 and Gamma = 25, 24 … 2 - 10 [Adapted from Xu et al. 2004].

IV. Description of Server

RiboPred is freely available at www.bifmanit.org/RiboPred2/ . RiboPred server is installed on a Windows Server environment. The user can provide the input sequence by cut-paste or directly uploading sequence file from disk. The server accepts the sequence in standard FASTA format. A snapshot sequence submission page of server is shown in Figure 2. User can predict the type of Ribonucleases based on dipeptide composition. On submission the server will give results in user-friendly format [Figure 2].

V. Conclusion

With dipeptide composition as evaluation parameter of ribonucleases, an overall average accuracy of 94.534% was obtained in classifying various subclasses. The tool DiRiboPred developed at www.bifmanit.org/RiboPred2 can be an efficient and time saving. These kinds of web servers can be an economical and time saving approach for annotation of piled up genomic data. They can be used to effectively complement the existing wet lab techniques. The author awaits discovery of more of theses proteins in the future so that more accurate classifiers and tools can be developed.

VI. Acknowledgement

We are highly thankful to Madhya Pradesh Council of Science and Technology and Department of Biotechnology, New Delhi for providing support in the form of Bioinformatics Infrastructure Facility

VII. References

1) Holzmann, J., Frank, P., Löffler, E., Bennett, K., Gerner C., & Rossmanith, W. (2008), RNase P without RNA: Identification and functional reconstitution of the human mitochondrial tRNA processing enzyme, Cell, 135(135), 462–474.
2) Alessio, D.G., & Riordan JF. (1997), Ribonucleases: Structures and Functions, Academic Press.
3) Gerdes, K., Christensen, S.K., & Lobner-Olesen, A. (2005), Prokaryotic toxin-antitoxin stress response loci, Nat. Rev. Microbiol., (3), 371–382.
4) Ardelta, B., & Darzynkiewicza, Z. (2009), Ribonucleases as potential modalities in anticancer therapy. European Journal of Pharmacology, (625), 1-3.
5) Bhasin, M., & Raghava, G. P. S. (2004), GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors, Nucleic Acids Research, 32[Web Server Issue], W383-W389.

6) Bhasin, M., & Raghava, G. P. S. (2004), Classification of Nuclear Receptors Based on Dipeptide composition and Dipeptide Composition, The Journal of Biological Chemistry, (279), 23262-23266.
7) Huang, Y., Niu, B., Gao, Y., Limin, F., & Weizhong, Li. (2010), CD-HIT Suite: a web server for clustering and comparing biological sequences, Bioinformatics.
8) Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Ron, D., & Bairoch, Amos. (2003), ExPASy: the proteomics server for in-depth protein knowledge and analysis, Nucleic Acids Research, 31(13), 3784-3788.
9) Hua, S., & Sun, Z. (2001), Support vector machine approach for protein subcellular localization prediction, Bioinformatics, 17, 721–728.
10) Ivanciuc, O. Applications of Support Vector Machines in Chemistry. (2007), Rev. Comput. Chem., 23, 291-400.
11) Joachims, T., (1999), Making large-scale SVM learning practical. In Scholkopf,B., Burges,C. and Smola,A. [eds], Advances in Kernel Methods Support Vector Learning. MIT Press, Cambridge, MA and London, 42–56. SVMlight; http://svmlight.joachims.org/
12) Keerthi, S. S., & Lin, S. S. (2003), Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Computation, 15(7), 1667–1689.
13) Lin, H.-T., & Lin, C.-J., (2003), A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods. Technical report, Department of Computer Science, National Taiwan University.
14) Wei, Hsu, C., Chung Chang, C., & Chih-Jen Lin, A., (2003), Practical Guide to Support Vector Classification.
15) Chang, C.-C., & Lin, C.-J., (2003), LIBSVM: a library for support vector machines.
16) Rashid, M., Saha, S., & Raghava, G.P.S. (2007), Support Vector Machine-based method for predicting subcellular localization of mycobacterial proteins using evolutionary information and motifs, BMC Bioinformatics, 8:337.
17) Lata, S., & Raghava, G.P.S. (2008), CytoPred: a server for prediction and classification of cytokines, Protein Engineering Design and Selection, 21(4), 279-282.
18) Protein composition server at MANIT,Bhopal, www.manit.ac.in/downloads/polycomp/.