Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Prediction of enzymes and non-enzymes from protein sequences based on sequence derived features and PSSM matrix using artificial neural network

Pradeep Kumar Naik[1,*], Viplav Shankar Mishra[1], Mukul Gupta[1], Kunal Jaiswal[1]

[1]Department of Bioinformatics and Biotechnology, Jaypee University of Information Technology, Waknaghat, Distt.-Solan, 173 215, Himachal Pradesh, India; Pradeep Kumar Naik* - Phone: 91 1792 239227; E-mail: pknaik73@rediffmail.com; * Corresponding author

received September 08, 2007; revised November 06, 2007; accepted November 09, 2007; published online December 05, 2007

Abstract:
The problem of predicting the enzymes and non-enzymes from the protein sequence information is still an open problem in bioinformatics. It is further becoming more important as the number of sequenced information grows exponentially over time. We describe a novel approach for predicting the enzymes and non-enzymes from its amino-acid sequence using artificial neural network (ANN). Using 61 sequence derived features alone we have been able to achieve 79 percent correct prediction of enzymes/non-enzymes (in the set of 660 proteins). For the complete set of 61 parameters using 5-fold cross-validated classification, ANN model reveal a superior model (accuracy = 78.79 plus or minus 6.86 percent, Q(pred) = 74.734 plus or minus 17.08 percent, sensitivity = 84.48 plus or minus 6.73 percent, specificity = 77.13 plus or minus 13.39 percent). The second module of ANN is based on PSSM matrix. Using the same 5-fold cross-validation set, this ANN model predicts enzymes/non-enzymes with more accuracy (accuracy = 80.37 plus or minus 6.59 percent, Q(pred) = 67.466 plus or minus 12.41 percent, sensitivity = 0.9070 plus or minus 3.37 percent, specificity = 74.66 plus or minus 7.17 percent).

Key words: enzymes; non enzymes; neural network; sequence derived features; PSSM

Background:
It is generally accepted that protein structure is determined by its amino acid sequence **[1]** and that the knowledge of protein structures plays an important role in understanding their functions. To understand the rules relating amino acid sequence to three-dimensional protein structure is one of the major goals of contemporary molecular biology. A priori knowledge of protein as enzymes and non-enzymes has become quite useful from both an experimental and theoretical point of view.

One of the fundamental problems in post-genome era is the prediction and classification of proteins given only their primary sequence. **[2]** The number of proteins that are being made available to public and private databases is growing exponentially, and new methods must be found to understand and classify that information. The enormous task of function determination for every entry in GenBank has prompted the development of more sophisticated methods for protein automatic classification. **[3, 4]** A computational method allowing for the automatic determination of protein function from its sequence alone is one of the prevailing problems in bioinformatics. **[5]** Determination of three-dimensional structure is the traditional approach to functional classification of proteins. This is a very time-consuming process, and the need for a faster method of classification is obvious. **[6]**

It has been reported that structural classes of proteins correlate strongly with amino acid composition, marked the onset of algorithm developments aimed at predicting the structural class of a protein from its amino acid composition alone. **[7]** In addition to amino acid composition, considering the sequence order along the primary structure of a protein into account would result in the improvement of prediction accuracy. **[8]** Hence, in this study we have develop two different neural networks which extract valuable information from protein sequence only for prediction into enzymes/non-enzymes. The first network used sequence derived features derived from PEPSTAT (EMBOSS suite) and the second network used PSSM profile obtained from PSI-BLAST, which would be useful for the systematic analysis of small or medium size protein sequences. Results are discussed, assessing the benefits of using this methodology in binary prediction of enzymes / non-enzymes. The preliminary results suggest that sequence derived feature can be used as a fast and effective classification methodology for proteins.

Methodology:
Training data
A data set of 660 proteins, consisting of 330 non redundant enzymes and the same number of non redundant non-enzymes, were used for training and testing. The enzyme data set used in

this study is obtained from the BRENDA database http://www.brenda.uni-koeln.de. **[9]** The pairwise sequence identities in the datasets are less than 54 percent for enzyme class and 45 perent for non-enzyme class.

Sequence derived parameters calculation and selection
To build a binary ANN model enabling effective prediction of enzymes/non-enzymes we initially calculated 61 parameters (Table 1 in supplementary material) from the protein sequence alone using PEPSTAT (EMBOSS suite) ftp://emboss.open-bio.org/pub/EMBOSS **[10]** for all 660 protein sequences. The normalized values (varying from 0 to 1) have been then used to generate ANN models for binary prediction.

Fivefold cross-validation
Fivefold cross-validation technique has been used for training and testing the ANN model, in which the dataset is randomly divided into five subsets, each containing equal number of enzyme sequences. Each set is a balanced set that consist of 50 percent of enzymes and 50 percent non-enzymes. The data set has been divided into training and testing set. The training set consists of five subsets. The network is validated for minimum error on testing set to calculate the performance measure for each fold of validation. This has been done five times to test for each subset. The final prediction results have been averaged over five testing sets.

ANN model for prediction of enzyme/non-enzyme using sequence derived features
Stuttgart neural network simulator package (SNNS version 4.2) **[11]** with standard back propagation was used to implement the ANN model. ANN configuration consists of 61 inputs and 1 output node. Whereas the number of nodes in the hidden layer was varied from 0 to 6 in order to find the optimal network that allows most accurate separation of enzymes/non-enzymes in the training sets (Figure 1a). During the learning phase, a value of 1 was assigned for the enzyme sequence and 0 for non-enzyme. The corresponding counts of the false/true positive and negative predictions were estimated using 0.1 and 0.9 cut-off values for non-enzymes and enzymes respectively. Thus, an enzyme from the testing set was considered correctly predicted by the ANN only when its output value ranged from 0.9 to 1.0. For each non-enzyme of the testing set the correct prediction was assumed if the corresponding ANN output lies between 0 and 0.1.

ANN model for prediction of enzyme/non-enzyme using PSSM matrix
In this module of the developed tool, the position-specific scoring matrix generated by PSI-BLAST has been used as input to the neural network. The matrix has 20 x M real-number elements, where M is the length of the sliding window (M = 7). Each element represents the likelihood of that particular residue substitution at that position. Thus 20 real numbers rather than binary bits encode each residue. A standard back-propagation ANN configuration consisting of 140 inputs and 1 output node was used. The number of nodes in the hidden layer was varied from 0 to 6 in order to find the optimal network that allows most accurate separation of enzymes/non-enzymes in the training sets (Figure 1b). The training and validation methods are similar as mentioned above. The corresponding counts of the false/true positive and negative predictions were estimated using 0.4 and 0.9 cut-off values for non-enzymes and enzymes respectively. Thus, an enzyme from the testing set was considered correctly predicted by the ANN only when its output value ranged from 0.9 to 1.0. For each non-enzyme of the testing set the correct prediction was assumed if the corresponding ANN output lies between 0.1 and 0.4.

Performance measures
The prediction results of ANN model developed in the study were evaluated using the equations given in the supplementary material.

Results and Discussion:
The two different ANN models developed in this study are based on sequence derived features and PSSM matrix method. Applying a fivefold cross-validation using five testing subdata sets, we found that the network reached an overall accuracy of 78.79 ± 6.86% based on sequence derived features. The network has achieved an MCC of 0.596 ± 0.135. The other performance measures are: Qpred = 67.466 ± 17.084%, sensitivity = 90.70 ± 6.73% and specificity = 74.66 ± 13.39%. The vast majority of the predictions have been contained within 0.0 to 0.1 for non-enzymes and 0.9 to 1.0 for enzymes in case of sequence derived module. This illustrates that 0.1 and 0.9 cut-off values provide very adequate separation of two bioactive classes using ANN. To further enhance the prediction performance, the PSSM matrix is used for prediction. The network 7(20)-4-1 is trained on PSI-BLAST generated position-specific matrices (PSSM). The prediction results for both the networks are presented in Table 2 (supplementary material). It is clear from the results that the performance is improved slightly when PSI-BLAST-generated scoring matrices are used as input, compared with sequence derived features. The prediction accuracy is improved from 78.79% to 80.37%. However, most dramatic improvement is achieved in other parameters like Qpred, sensitivity and specificity. This is because it uses improved searching tool for multiple sequence alignment such as PSI-BLAST. PSI-BLAST searches the homologs against a larger database such as a nonredundant database and use multiple sequence information to generate PSSM matrix. From this study, it is clear that a combination of neural network and evolutionary information contained in multiple sequence alignment has improved the performance of prediction method.

The results demonstrate that the developed ANN-based binary prediction of enzymes/non-enzymes is adequate and can be considered an effective tool for *in silico* screening. The results also demonstrate that the sequence derived parameters as well as PSSM matrix readily accessible from the protein sequences only, can produce a variety of useful information to be used *in silico*; clearly demonstrates an adequacy and good predictive power of the developed ANN model. There is strong evidence that the introduced sequence features do adequately reflect the structural properties of proteins. The structure of a protein is an important determinant for the detailed molecular function of proteins, and would consequently also be useful for prediction of enzymes and non-enzymes. Based on the analysis of limited

sequence features from protein sequences, differences in the parameters between enzymes and non-enzymes have previously been shown to exist and used for prediction of enzymes/non-enzymes in archaeal. **[12]** This agrees well with our result that sequence derived features can be used for predicting enzymes.

Presumably, accuracy of the approach operating by the sequence derived features can be improved even further by expanding the parameters or by applying more powerful classification techniques such as Support Vector Machines or Bayesian Neural Networks. Use of merely statistical techniques in conjunction with the sequence parameters would also be beneficial, as they will allow interpreting individual parameter contributions into "enzymes/non-enzymes-likeness".

The results of the present work demonstrate that both the sequence derived features and PSSM matrix with ANN appear to be a very fast protein classification mechanism providing good results, comparable to some of the current efforts in the literature.

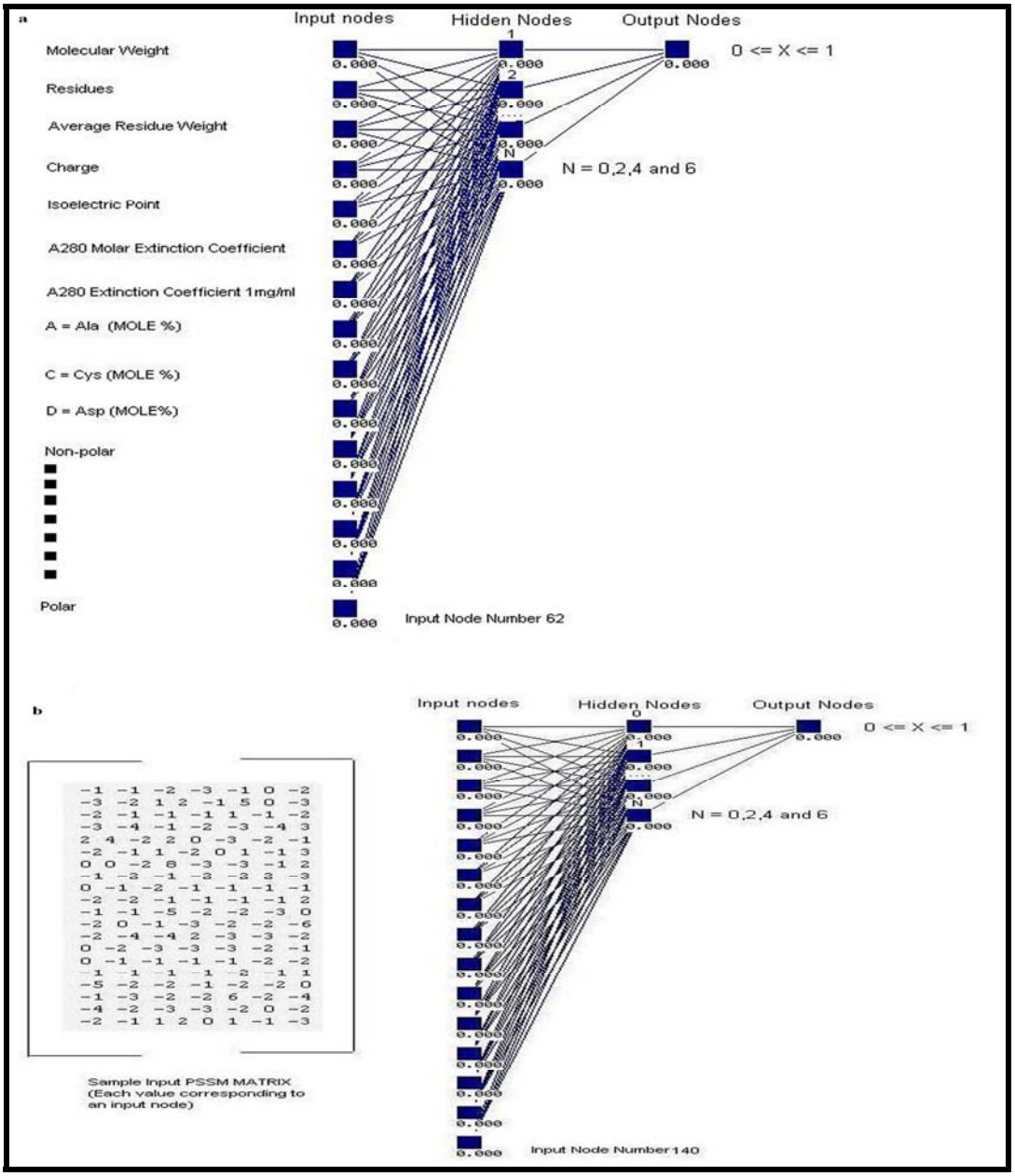

Figure 1: Configuration of artificial neural network used to develop binary primary sequence descriptor model for enzyme or non-enzyme proteins based on (a) sequence derived features and (b) PSSM matrix

Availability:
The program is implemented on the web server EnzymePred, available at http://www.juit.ac.in/enzyme/tool.html by using CGI/Perl script. The SNNS-generated network is converted into C program and is used as an interface. Users can enter primary amino acid sequence in fasta or free format. The protein sequence can be predicted as enzyme or non-enzyme.

Conclusion:
We have demonstrated the feasibility of combining ANN with sequence derived features and PSSM matrix for prediction of enzymes/non-enzymes from protein sequence only. Even as prototype, both the ANNs we implemented have shown practical performance. With appraisal tests, we have found clues to improve prediction accuracy of ANN further. Expanding the sequence derived features, use of merely statistical techniques in conjunction with the sequence parameters and an adequate and low-noise training set, are critical to the success of ANN. Apparently, the more specifically an enzyme is to predict, thus the more definite a training set can be assembled, and the higher predicting power the corresponding ANN can acquire. In the future, we envisage an array of ANNs being trained to predict different classes and sub-classes of enzymes and to parse genomic sequence data in parallel, complementing current methods to achieve more reliable, high-throughput gene function prediction.

References:
[01] C. B. Anfinsen, *Science*, 181: 223 (1973) [PMID: 4124164]
[02] D. Eisenberg, *et al.*, *Nature*, 405: 823 (2000) [PMID: 10866208]
[03] C. Z. Cai, *et al.*, *Math Biosci.*, 185: 111 (2003) [PMID: 12941532]
[04] R. D. King, *et al.*, *Bioinformatics*, 20: 1110 (2004) [PMID: 14764546]
[05] P. Bork & E. V. Koonin, *Nat Genet.*, 18: 313 (1998) [PMID: 9537411]
[06] E. N. Baker, *et al.*, *Applied Bioinformatics*, 2: s3 (2003) [PMID: 15130810]
[07] H. Nakashima, *et al.*, *J. Biochem.*, 99: 152 (1986) [PMID: 3957893]
[08] W. S. Bu, *et al.*, *Eur. J. Biochem.*, 266: 1043 (1999) [PMID: 10583400]
[09] I. Schomburg, *et al.*, *Nucleic Acids Res.*, 32: D431 (2004) [PMID: 14681450]
[10] P. Rice, *et al.*, *Trends in Genetics*, 16: 276 (2000) [PMID: 10827456]
[11] http://www-ra.informatik.uni-tuebingen.de/SNNS/
[12] L. J. Jensen, *et al.*, *Protein Sci.*, 11: 2894 (2002) [PMID: 12441387]

Edited by P. Kanguenae
Citation: Naik *et al.*, Bioinformation 2(3): 107-112 (2007)

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

Supplementary material

Sequence derived parameters	Enzyme		Non-Enzyme		Sequence derived parameters	Enzyme		Non-Enzyme	
	Max	Min	Max	Min		Max	Min	Max	Min
Molecular Weight	0.207588	0.00182	0.20947	0.00419	N_DayhoffStat	0.1671	0.0987	0.2114	0.1078
Average Residue	0.11811	0.09159	0.1209	0.09186	P_Mole %	0.9572	0.3450	3.6556	0.5680
Isoelectric Point	0.104656	0.0427	0.1288	0.03857	P_DayhoffStat	0.1841	0.0089	0.703	0.02908
Extinction Coefficient	0.29032	0.019	0.33257	0.027	Q_Mole %	0.585	0.0871	1.5106	0.1098
Extinction Coefficient (1 mg/ml)	0.275	0.024	0.376	0.036	Q_DayhoffStat	0.15	0.0098	0.3873	0.0129
Improablity / Proability inclusion bodies	0.928	0.494	0.979	0.41	R_Mole %	1.0682	0.0088	2.1256	0.0187
A_Mole %	0.18828	0.02881	0.21186	0.03	R_DayhoffStat	0.218	0.02389	0.434	0.0452
A_DayhoffStat	0.2189	0.0335	0.2464	0.045	S_Mole %	0.9035	0.1796	2.2034	0.0012
B_Mole %	0.1989	0.0017	0.0902	0.0011	S_DayhoffStat	0.1291	0.0257	0.3148	0.0389
B_DayhoffStat	0.0292	0.001	0.0109	0.0009	T_Mole %	1.0497	0.3091	1.4352	0.1203
C_Mole %	1	0.00659	2.0339	0.0089	T_DayhoffStat	0.1721	0.0507	0.2353	0.0092
C_DayhoffStat	0.3448	0.02154	0.7013	0.0154	V_Mole %	0.15	0.04484	0.17647	0.0289
D_Mole %	0.8147	0.0154	1.206	0.0015	V_DayhoffStat	0.2273	0.0679	0.2674	0.0546
D_DayhoffStat	0.1481	0.0152	0.2193	0.0652	W_Mole %	0.4598	0.00245	0.4839	0.0254
E_Mole %	1.018	0.0147	1.8615	0.0254	W_DayhoffStat	0.3537	0.0021	0.3722	0.0215
E_DayhoffStat	0.1697	0.0215	0.3102	0.0145	X_Mole %	0.4562	0.025	0.3262	0.0254
F_Mole %	0.9195	0.1277	1.0044	0.0596	X_DayhoffStat	0.5263	0.0562	0.3215	0.025
F_DayhoffStat	0.2554	0.0355	0.279	0.0101	Y_Mole %	0.6135	0.0159	2.4615	0.0521
G_Mole %	0.25	0.00769	0.36923	0.00503	Y_DayhoffStat	0.1804	0.0154	0.724	0.00987
G_DayhoffStat	0.2976	0.0092	0.4396	0.006	Z_Mole %	0.2222	0.0089	0.3262	0.0154
H_Mole %	0.6513	0.00894	1.0271	0.021	Z_DayhoffStat	0.894	0.1256	0.265	0.03652
H_DayhoffStat	0.3257	0.0456	0.5136	0.0598	Tiny Mole %	0.6	0.15569	0.6389	0.16239
I_Mole %	1	0.2077	1.0377	0.0089	Small Mole %	0.75	0.4012	0.77119	0.32479
I_DayhoffStat	0.2222	0.0462	0.2306	0.0564	Aliphatic Mole %	0.31481	0.14808	0.32903	0.02542
K_Mole %	1.018	0.0591	2.0455	0.00115	Aromatic Mole %	0.24521	0.04918	0.29231	0.08541
K_DayhoffStat	0.1542	0.00213	0.3099	0.0002	Non-polar Mole %	0.85	0.45521	0.86154	0.31818
L_Mole %	0.19444	0.03139	0.19101	0.0321	Polar Mole %	0.54479	0.15	0.68182	0.13846
L_DayhoffStat	0.2628	0.0424	0.2581	0.0021	Charged Mole %	0.33533	0.05	0.46986	0.01389
M_Mole %	0.5169	0.0456	1.2346	0.0268	Basic Mole %	0.17365	0.05	0.31624	0.00926
M_DayhoffStat	0.3041	0.0154	0.7262	0.0158	Acidic Mole %	0.16168	0.00897	0.25	0.0154
N_Mole %	0.7186	0.1200	0.9091	0.2300					

Table 1: 61 'Pepstat (EMBOSS)' primary sequence descriptors used in the study. The parameters are scaled down by appropriate scaling values.

5-fold cross validation	Accuracy	Specificity	Sensitivity	MCC	Q(Pred)	Prediction range (enzymes)	Prediction range (non-enzymes)
(a) using sequence derived features (PEPSTAT)							
C1	0.8947	1.00	0.8271	0.8072	100	0.9626-1.00	0.00-0.5340
C2	0.7969	0.7671	0.8333	0.5979	74.62	0.9579-1.00	0.00-0.6758
C3	0.7142	0.6794	0.7636	0.4364	76.68	0.9257-1.00	0.00-0.8786
C4	0.7443	0.6666	0.9495	0.5490	52.28	0.9692-1.00	0.00-0.8586
C5	0.7894	0.7934	0.8545	0.5891	70.14	0.9048-1.00	0.00-0.8236
Mean	0.7879 ± 0.0686	0.7713 ± 0.1339	0.8448 ± 0.0673	0.5959 ± 0.1345	74.734 ± 17.084		
(b) using PSSM matrix (PSI BLAST)							
C1	0.8230	0.7641	0.9158	0.6628	71.15	0.9237-0.9559	0.2180-0.2205
C2	0.8717	0.8148	0.9538	0.7560	78.13	0.9357-0.9443	0.3921-0.6006
C3	0.8521	0.8072	0.9123	0.7118	77.91	0.9061-0.9156	0.1626-0.7521
C4	0.7567	0.6988	0.8624	0.5368	61.09	0.9255-0.9272	0.3239-0.5133
C5	0.7153	0.6485	0.8911	0.4821	49.05	0.9123-0.9343	0.3005-0.4183
Mean	0.8037 ± 0.0659	0.7466 ± 0.0717	0.9070 ± 0.0337	0.6299 ± 0.1164	67.466 ± 12.411		

Table 2: Results of enzymes / non-enzymes prediction methods, using five fold cross validation

Equations used in this article:

Accuracy of the prediction methods

$$Q_{ACC} = \frac{P+N}{T} \text{ (where T = (P+N+O+U))} \rightarrow \textbf{(1)}$$

Matthews correlation coefficient (MCC)

$$MCC = \frac{(P \times N) - (O \times U)}{\sqrt{(P+U) \times (P+O) \times (N+U) \times (N+O)}} \rightarrow \textbf{(2)}$$

Sensitivity (Q_{sens})

$$Q_{sens} = \frac{P}{P+U} \rightarrow \textbf{(3)}$$

specificity (Q_{spec})

$$Q_{spec} = \frac{N}{N+O} \rightarrow \textbf{(4)}$$

Q_{Pred} (Probability of correct prediction)

$$Q_{pred} = \frac{P}{P+O} \times 100 \rightarrow \textbf{(5)}$$

where *P* and *N* refer to correctly predicted enzymes and non-enzymes, and *O* and *U* refer to over and under predictions, respectively.