Digest Journal of Nanomaterials and Biostructures Vol. 5, No 1, March 2010, p. 191 – 200

PREDICTION AND CLASSIFICATION OF DNA BINDING PROTEINS INTO FOUR MAJOR CLASSES BASED ON SIMPLE SEQUENCE DERIVED FEATURES USING ANN

AMIYA KUMAR PATEL*, SEEMA PATEL[a], PRADEEP KUMAR NAIK[b]
Division of Biotechnology, Majhighariani Institute of Technology and Science (MITS), At- Sriram Vihar, Bhujbala, Po- Kolnara, Rayagada, (Pin – 765017), Orissa, India
[a]*Department of Computer Science, Sundargarh Engineering College, At/Po-Kirei, Distt.-Sundargarh, (Pin- 770073), Orissa, India.*
[b]*Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Waknaghat, Distt.-Solan, (Pin- 173 215), Himachal Pradesh, India*

The problem of predicting the different classes of DNA binding protein from the protein sequence information is still an open problem in bioinformatics. We implemented a two-layered artificial neural network (ANN) of predicting the DNA binding proteins and their classification into four major classes from their amino-acid sequences. Using 61 sequence derived features we are able to achieve 72.99% correct prediction of proteins into DNA binding/non-DNA binding (in the dataset of 1000 proteins). For the complete set of 61 parameters using 5-fold cross-validated classification, ANN model revealed a superior model (accuracy = 72.99 ± 6.86%, Q_{pred} = 73.952 ± 13.12%, sensitivity = 81.53 ± 6.73% and specificity = 72.54 ± 6.39%). The classification accuracy for predicted DNA binding protein into four sub-classes was 70.73% (on average) using five fold cross validation, indicating that multi-class ANN classification system (61-11-4) may have certain level of unique prediction capability.

(Received February 14, 2010, accepted March 4, 2010)

Keywords: DNA binding proteins, Classification, Artificial neural network, Sequence derived features

1. Introduction

The prediction of protein structure from amino acid sequence has become the Holy Grail of computational molecular biology. The information necessary for protein folding resides completely within the primary structure; molecular biologists have been fascinated with the possibility of obtaining a complete three-dimensional picture of a protein by simply applying the proper algorithm to a known amino acid sequence [1]. The development of rapid methods of DNA sequencing coupled with the straightforward translation of the genetic code into protein sequences has amplified the urgent need for automated methods of interpreting these one-dimensional, linear sequences in terms of three-dimensional structure and function. Advanced and specialized databases are needed to facilitate the retrieval of relevant information from the deluge of sequence data and to provide insight into the protein structure and function. Further, it is clear that rational classification of proteins encoded in sequenced genomes is critical for making the genome sequences maximally useful for functional and evolutionary studies [2].

The family of DNA binding proteins is one of the most populated and studied amongst the various genomes of bacteria, archea and eukaryotes. Most of these proteins, such as the eukaryotic

*Corresponding author: amiya_gene@yahoo.com

and prokaryotic transcription factors, contain independently folded units (domains) in order to accomplish their recognition with the contours of DNA. It is now clear that the majority of these DNA-binding scaffolds which are in general relatively small, less than 100 amino acid residues, belong to a large number of structural families with characteristic sequences and three-dimensional designs or conformations [3]. Computational biology applying fast and sensitive algorithms strives to extract the maximum possible information from these sequences by classifying them according to their homologous relationships, predicting their likely biochemical activities and/or cellular functions, three-dimensional structures and evolutionary origin. There have been studies to detect [4, 5], design [6] and predict them using a probabilistic recognition code [7]. There have also been works towards analyzing protein–DNA recognition mechanism [8] and binding site discovery [9]. DNA binding proteins represent a broad category of proteins, known to be highly diverse in sequence and structure. Structurally, they have been divided into 54 protein-structural families [10]. With such a high degree of variance, using conventional annotation methods rooted in database searching for sequence similarity [11], profile or motif similarity [12] and phylogenetic profiles [13] may not lead to reliable annotations. In this context, a DNA binding protein prediction protocol that takes into account the structural information and does not depend on sequential or structural homology to proteins with known functions will be very useful.

Previously, there have been a few bioinformatics methods developed towards automated identification and prediction of DNA binding proteins. The pseudo-amino acid composition is used to identify proteins that bind to RNA, rRNA and DNA [14]. Structural information was integrated with the neural network approach for the prediction of DNA binding proteins [15]. Electrostatic features of proteins were also characterized through an automated approach for DNA binding protein and DNA binding site prediction [16, 17]. Further, the overall charge and electric moment can be used to identify DNA binding proteins [18]. Accuracy rates achieved in these methods varied from 65% to 86% depending on both the features used and the validation method adopted.

Strategically, we have used a neural network, two-layer, fully automated computational method capable of recognizing DNA binding proteins first, and then classifying them into their different classes based on their sequences derived features.

2. Methodology

Data set for prediction of DNA binding/non-DNA binding

A dataset of 500 DNA binding protein sequences were extracted from PDB. A non-redundant treatment was applied to eliminate the sequences which share a high degree of similarity (>90%) with others in order to avoid overtraining. The treatment was carried out using the program BLASTCLUST (http://www.ncbi.nlm.nih.gov/BLAST/), which used the BLAST algorithm to systematically cluster protein sequences on the basis of pair-wise matches. The default values were used for all BLAST parameters: matrix BLOSUM62, gap opening cost of 11, gap extension cost of 1, E-value threshold of $1e^{-6}$. These sequences were used as positive examples for prediction as DNA binding proteins. The sequences data on negative examples were obtained from the SWISSPROT database (http://expasy.org/sprot/). DNA binding proteins were removed from the original dataset. A non-redundant treatment was applied (same as for positive datasets) such that no sequence had similarity higher than 25% to any others. Thus, 500 non-DNA binding sequences were optimized as negative examples.

Dataset for classification of DNA binding proteins into four major classes

The above mentioned 500 protein sequences of DNA binding protein were then grouped into four major classes: class 1 (Homeo box domain) consist of 125 sequences, class II (Zinc finger) consist of 125 sequences, class III (Leucine zipper) having 125 sequences and class IV (Helix-Turn-Helix) with 125 sequences. They were used for construction of neural networks

training and validating the model for classification of predicted DNA binding proteins into four classes.

Neural network architecture

The implementation of ANN was realized using the software package SNNS (Version-4.2) from Stuttgart University [19]. We have used two feed-forward back-propagation neural networks with a single hidden layer. First layer of neural network is used for prediction of DNA binding/non-DNA binding proteins from the protein sequence, whereas, the second layer is used for classifying the predicted DNA binding protein into out of four major classes. The 1st neural network consisting of 61 inputs, 7 hidden nodes and 1 output node. The number of nodes in the hidden layer was varied from 1 to 11 in order to find the optimal network that allows most accurate separation of DNA binding and non-DNA binding proteins in the training sets. The 2nd neural network consisting of 61 inputs, 11 hidden nodes and four output nodes (each node is specified for each class of DNA binding protein) (Figure 1). The number of nodes in the hidden layer was varied from 1 to 15 in order to find the optimal network that allows most accurate classification of DNA binding protein in the training sets. For each sequence in the training and testing sets, we have transformed 61 network input parameters into the normalized values varying from 0 to 1. Similarly, the output parameters from the ANN were in the range of 0 to 1. During the learning phase, a value of 1 was assigned for the DNA binding protein and 0 for non-DNA binding. For configuration of the ANN, 100 independent training runs were performed to evaluate the average predictive power of the network. The corresponding counts of the false/true positive and negative predictions were estimated using 0.1 and 0.9 cut-off values for non-DNA binding and DNA binding proteins respectively. Thus, a protein sequence from the testing set was considered correctly predicted as DNA binding protein by the ANN only when its output value ranged from 0.9 to 1.0. For each non-DNA binding protein of the testing set the correct prediction was assumed if the corresponding ANN output lies between 0 and 0.1. Thus, all network output values ranging from 0.2 to 0.9 have been ultimately considered as incorrect predictions (rather than undetermined or non-defined). If the input protein sequence is predicted as enzyme than it is parsed into the second layer and is classify into its particular class based on the maximum value obtained from the defined out put node for each class. For example to classify the predicted DNA binding into class 1 (Homeo box) the predicted output value is 1, 0, 0, 0 and so on. The input to second filtering network is the same input values used for the first layer.

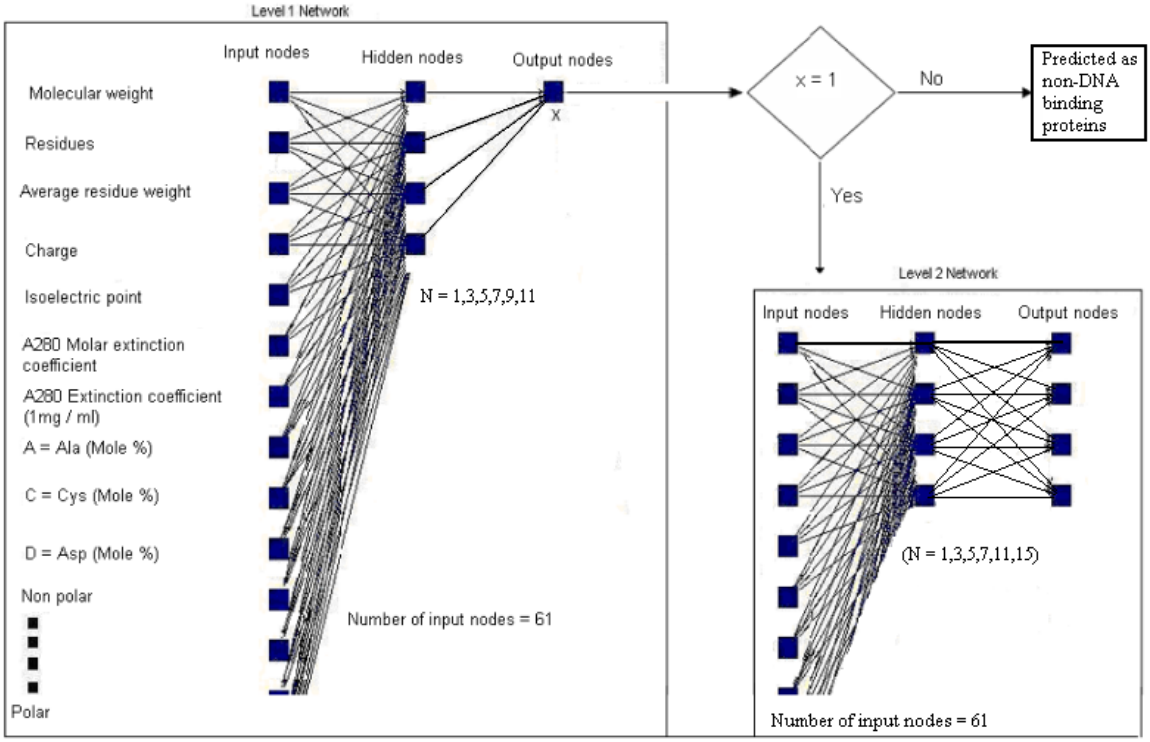

Fig. 1. Configuration of artificial neural network used to develop binary primary sequence descriptor model for DNA binding/non-DNA binding proteins.

Sequence derived parameters calculation

A set of 61 parameters were calculated from the protein sequence alone using PEPSTAT (EMBOSS suite) ftp://emboss.open-bio.org/pub/EMBOSS [20] for all 1000 protein sequences. The average values of these 61 parameters were independently calculated for DNA binding and non-DNA binding proteins as well as for each class of DNA binding protein and used as input values to the ANN model.

Fivefold cross-validation

A prediction method is often developed by cross-validation or jack-knife method [21]. Because of the size of the dataset, the jack-knife method (individual testing of each enzyme in the data set) was not feasible. So a more limited cross-validation technique has been used, in which the dataset is randomly divided into five subsets, each containing equal number of DNA binding proteins. Each set is a balanced set that consist of 50 percent of DNA binding and 50 percent non-DNA binding proteins. 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.

Table 1. 61 'Pepstat(EMBOSS)' primary sequence descriptors used in the study.

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

All the results reported for 2nd layer of ANN were obtained by performing a modified five-fold cross-validation procedure [22]. First, a given number of proteins (80) were randomly drawn from the dataset for each of the four families. The sum of these samples constituted the training set (320). All the other proteins were allocated to the evaluation set (180). Then, the neural network was trained and later evaluated using this partition. The accuracy rate on the evaluation set was

computed as the ratio of the number of correctly classified proteins to the total number of proteins, as is standard in the literature. Next, a new sampling was taken from the dataset to form another training set and evaluation set, and the training and evaluation process were repeated. This procedure was repeated five times and the final results were reported as the averaged accuracy rate over these five runs.

Table 2. 61 'Pepstat (EMBOSS)' primary sequence descriptors used in the study.

Parameters	Class 1 Homeo box		Class 2 Zinc finger		Class 3 Leucine zipper		Class 4 Helix-turn-helix	
	Max	Min	Max	Min	Max	Min	Max	Min
Mol. Weight	0.208	0.002	0.177	0.015	0.067	0.004	0.045	0.038
Average Residue	0.118	0.091	0.120	0.101	0.117	0.104	0.116	0.104
Isoelectric Point	0.105	0.043	0.110	0.046	0.086	0.045	0.101	0.045
Extinction Coefficient	0.290	0.016	0.181	0.001	0.085	0.004	0.067	0.020
Extinction Coefficient (1 mg/ml)	0.275	0.017	0.225	0.006	0.172	0.031	0.172	0.048
Improablity/Proability inclusion bodies	0.928	0.494	0.881	0.497	0.871	0.497	0.848	0.503
A_Mole %	0.188	0.029	0.241	0.027	0.161	0.028	0.160	0.022
A_DayhoffStat	0.219	0.034	0.280	0.032	0.187	0.032	0.186	0.025
B_Mole %	0.313	0.027	0.164	0.034	0.133	0.098	0.920	0.013
B_DayhoffStat	0.165	0.010	1.132	0.264	0.455	0.027	0.255	0.001
C_Mole %	1.000	0.092	0.345	0.013	0.400	0.024	0.299	0.049
C_DayhoffStat	0.345	0.022	0.119	0.001	0.138	0.028	0.103	0.017
D_Mole %	0.815	0.062	0.735	0.197	0.903	0.175	0.818	0.279
D_DayhoffStat	0.148	0.025	0.134	0.036	0.164	0.032	0.149	0.051
E_Mole %	1.018	0.013	1.310	0.317	1.132	0.416	0.938	0.299
E_DayhoffStat	0.170	0.001	0.218	0.053	0.189	0.069	0.156	0.050
F_Mole %	0.920	0.128	0.725	0.098	0.556	0.013	0.591	0.166
F_DayhoffStat	0.255	0.036	0.201	0.027	0.154	0.001	0.164	0.046
G_Mole %	0.250	0.008	0.112	0.024	0.100	0.053	0.118	0.050
G_DayhoffStat	0.298	0.009	0.133	0.028	0.119	0.063	0.141	0.059
H_Mole %	0.651	0.012	0.455	0.049	0.635	0.160	0.609	0.134
H_DayhoffStat	0.326	0.065	0.227	0.017	0.318	0.080	0.305	0.067
I_Mole %	1.000	0.208	1.215	0.117	1.148	0.172	1.317	0.307
I_DayhoffStat	0.222	0.046	0.270	0.026	0.255	0.038	0.293	0.068
K_Mole %	1.018	0.032	1.089	0.110	0.833	0.080	1.532	0.136
K_DayhoffStat	0.154	0.065	0.165	0.017	0.126	0.012	0.232	0.021
L_Mole %	0.194	0.031	0.167	0.036	0.139	0.034	0.140	0.060
L_DayhoffStat	0.263	0.042	0.226	0.049	0.188	0.047	0.189	0.081
M_Mole %	0.517	0.015	0.448	0.046	0.556	0.248	0.365	0.103
M_DayhoffStat	0.304	0.081	0.264	0.027	0.327	0.069	0.215	0.061
N_Mole %	0.719	0.103	0.611	0.045	0.862	0.258	0.887	0.140
N_DayhoffStat	0.167	0.061	0.142	0.495	0.201	0.060	0.206	0.033
P_Mole %	0.957	0.140	0.840	0.164	0.862	0.160	0.679	0.166
P_DayhoffStat	0.184	0.033	0.162	0.032	0.166	0.031	0.131	0.032
Q_Mole %	0.585	0.166	0.817	0.068	0.874	0.013	0.855	0.059
Q_DayhoffStat	0.150	0.002	0.210	0.136	0.224	0.001	0.219	0.015
R_Mole %	1.068	0.024	1.525	0.193	0.774	0.214	0.868	0.134
R_DayhoffStat	0.218	0.048	0.311	0.040	0.158	0.044	0.177	0.027
S_Mole %	0.904	0.180	1.250	0.280	0.935	0.248	0.893	0.357
S_DayhoffStat	0.129	0.026	0.179	0.040	0.134	0.069	0.128	0.051

Parameters	Class 1 Homeo box		Class 2 Zinc finger		Class 3 Leucine zipper		Class 4 Helix-turn-helix	
	Max	Min	Max	Min	Max	Min	Max	Min
T_Mole %	1.050	0.309	0.851	0.117	0.874	0.175	0.867	0.224
T_DayhoffStat	0.172	0.051	0.140	0.019	0.143	0.029	0.142	0.037
V_Mole %	0.150	0.045	0.129	0.038	0.152	0.041	0.105	0.040
V_DayhoffStat	0.227	0.068	0.196	0.058	0.230	0.062	0.160	0.061
W_Mole %	0.460	0.128	0.364	0.351	0.255	0.070	0.253	0.033
W_DayhoffStat	0.354	0.867	0.280	0.186	0.196	0.218	0.195	0.166
X_Mole %	0.512	0.142	0.957	0.097	0.519	0.051	0.853	0.002
X_DayhoffStat	0.265	0.105	0.184	0.076	0.332	0.148	0.198	0.024
Y_Mole %	0.614	0.160	0.597	0.032	1.035	0.254	0.544	0.207
Y_DayhoffStat	0.180	0.253	0.176	0.010	0.304	0.075	0.160	0.061
Z_Mole %	0.155	0.195	0.545	0.065	0.519	0.306	0.519	0.006
Z_DayhoffStat	1.231	0.022	0.335	0.208	0.332	0.044	0.332	0.497
Tiny Mole %	0.600	0.156	0.408	0.173	0.364	0.139	0.387	0.175
Small Mole %	0.750	0.401	0.519	0.387	0.597	0.389	0.590	0.368
Aliphatic Mole %	0.315	0.148	0.332	0.165	0.306	0.172	0.276	0.207
Aromatic Mole %	0.245	0.049	0.175	0.039	0.167	0.076	0.166	0.071
Non-polar Mole %	0.850	0.455	0.688	0.460	0.649	0.525	0.659	0.512
Polar Mole %	0.545	0.150	0.540	0.312	0.475	0.351	0.488	0.341
Charged Mole %	0.335	0.050	0.344	0.170	0.278	0.186	0.323	0.168
Basic Mole %	0.174	0.050	0.202	0.084	0.143	0.097	0.188	0.090
Acidic Mole %	0.162	0.000	0.173	0.057	0.151	0.076	0.147	0.071

Performance measures

The prediction results of 1st layer of ANN model developed in the study were evaluated using the following statistical measures.

1. Accuracy of the methods: The accuracy of prediction for neural network models were calculated as follows:

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

Where *P* and *N* refer to correctly predicted DNA binding and non-DNA binding proteins, and *O* and *U* refer to over and under predictions, respectively.

2. The Matthews correlation coefficient (MCC) is defined as:

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

3. Sensitivity (Q_{sens}) and specificity (Q_{spec}) of the prediction methods are defined as:

$$Q_{sens} = \frac{P}{P+U}$$

$$Q_{spec} = \frac{N}{N+O}$$

4. Q_{Pred} (Probability of correct prediction) is defined as:

$$Q_{pred} = \frac{P}{P+O} \times 100$$

3. Results and discussion

The 1st layer of ANN model develop in this study (61-7-1) is trained with the sequence derived features (61 parameters) calculated using PEPSTAT. The number of nodes in the hidden layer was varied from 1 to 11 in order to find the optimal network that allows most accurate separation of DNA binding/non-DNA binding proteins in the training sets (Table 3). When applying a fivefold cross-validation test using five data sets, we found that the network reached an overall accuracy of 72.99 ± 6.86%. The prediction results are presented in Table 4. The other performance measures were: Qpred = 73.95 ±13.12%, sensitivity = 81.53 ± 6.73% and specificity = 72.54 ± 6.39%. The value of the learning parameter was set to 0.1.

Table 3. Parameters of specificity, sensitivity, accuracy and positive predictive values for prediction of DNA binding and non-DNA binding from the protein sequence by the 1st layer of artificial neural networks with the varying number of hidden nodes. The cut-off values of 0.1 and 0.9 have been used for negative and positive predictions respectively.

Hidden Nodes	Accuracy	Specificity	Sensitivity	Q(Pred)
1	0.5869	0.6523	0.7423	65.23
3	0.6213	0.6452	0.5013	72.13
5	0.5522	0.5864	0.5123	55.23
7	0.6976	0.6878	0.7535	68.32
9	0.6435	0.6020	0.7632	65.18
11	0.6235	0.6425	0.7123	69.25

Table 4. Performance measure of 1st neural network for the prediction of DNA binding/non-DNA binding proteins using five fold cross validation based on sequence derived features.

Fivefold cross validation	Accuracy	Specificity	Sensitivity	Q(Pred)	Prediction range (DNA binding	Prediction range (Non-DNA binding)
C1	0.8.20	0.8632	0.7271	85.12	0.6726 – 1.00	0.00 – 0.5240
C2	0.7430	0.7791	0.8580	70.61	0.5079 – 1.00	0.00 – 0.5658
C3	0.7002	0.6024	0.8001	71.61	0.4257 – 1.00	0.00 – 0.5386
C4	0.7140	0.6567	0.8901	62.28	0.3592 – 1.00	0.00 – 0.6486
C5	0.6906	0.7259	0.8015	80.14	0.4748 – 1.00	0.00 – 0.5836
Mean	0.7299 ± 0.0686	0.7254 ± 0.0639	0.8153 ± 0.0673	73.952 ± 13.123		

By applying a modified fivefold cross-validation test using five data sets, we found that the second layer of network (61-11-4) is a superior model for classification of predicted DNA binding proteins into their suitable classes. The number of nodes in the hidden layer was varied from 1 to 15 in order to find the optimal network that allows most accurate classification system of DNA binding proteins in the training sets (Table 5). Out of 500 DNA binding proteins (125 proteins from each class) in each cross validation set 220 to 343 DNA binding proteins were correctly classified. However, the network was more efficiently classify the proteins belonging to Leucine zipper and Helix-turn-helix in compare to other classes (Table 6). The classification accuracy for DNA binding proteins from 4 families is in the range of 73.34% to 80.06% using five fold cross validation with an overall accuracy of 76.74% using a sequence derived features, indicating that multi-class ANN classification system (61-11-4) may have certain level of unique prediction capability.

Table 5. Result for classification of DNA binding proteins into four major classes using 2nd neural network based on protein sequence derived features with the varying number of hidden nodes.

Number of hidden nodes	Number of DNA binding protein taken	Correctly predicted DNA binding protein				Total
		Homeo box	Zinc finger	Leucine zipper	Helix-turn-helix	
1	500	65	74	61	65	265
3	500	56	61	51	52	220
5	500	55	70	61	65	251
7	500	65	80	75	78	298
11	500	86	87	81	89	343
15	500	84	83	77	81	325

Table 6. Predicted result of 2nd layer of neural network for classification of DNA binding proteins into corresponding classes using the fivefold cross validation sets.

Family	Accuracy rate (%) using five fold cross validation					
	1	2	3	4	5	Mean ± sd
Homeo box	66.7	70.8	66.7	79.2	83.3	73.34 ± 7.55
Zinc finger	69.2	73.3	78.3	71.7	82.5	75.0 ± 5.35
Leucine zipper	85.0	83.3	80.8	80.0	71.2	80.06 ± 5.34
Helix-turn-helix	70.8	73.2	80.6	85.8	82.4	78.56 ± 6.33
Average						76.74 ± 6.14

The classes of newly found DNA binding proteins are usually determined either by biochemical analysis of eukaryotic and prokaryotic genomes or by microarray chips. These experimental methods are both time-consuming and costly. With the explosion of protein entries in databanks, we are challenged to develop an automated method to quickly and accurately determine the enzymatic attribute for a newly found protein sequence: is it a DNA binding or a non-DNA binding protein? If it is, to which class does it belongs? The answers to these questions are important because they may help deduce the mechanism and specificity of the query protein, providing clues to the relevant biological function. Although it is an extremely complicated problem and might involve the knowledge of three-dimensional structure as well as many other physicochemical factors, some quite encouraging results have been obtained by a bioinformatical method established on the basis of amino acid composition alone [23]. Since the amino acid composition of a protein does not contain any of its sequence-order information, a logical step to further improve the method is to incorporate the sequence-order information into the predictor. To realize this, the most straightforward way is to represent the sample of a protein by its entire sequence, the so-called sequential form.

The results demonstrate that the developed ANN-based model for binary prediction of DNA binding/non-DNA binding proteins and classification of predicted DNA binding proteins into four major classes is adequate and can be considered an effective tool for *'in silico'* screening. The results also demonstrated that the sequence derived parameters 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 DNA binding proteins and for their classification. This observation is not surprising considering that the calculated parameters should cover a very broad range of proprieties of bound atoms and

molecules related to their size, polarizability, electronegativity, compactness, mutual inductive and steric influence and distribution of electronic density, etc. As it can be seen that the average value for different classes of DNA binding proteins were clearly separated (Table 1 & 2) and, hence, the selected 61 parameters should allow building an effective ANN model for binary prediction as well as their classification further.

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 "DNA binding/non-DNA binding-likeness".

References

[1] C. B. Anfinsen, Science **181**, 223 (1973).
[2] C. Wu, M. Berry, S. Shivakumar, J. McLarty, J. Mach. Learn. **21** N(1-2), 177 (1995).
[3] C. Branden, J. Tooze, Introduction to Protein Structure, Garland Publishing Co., New York (1991).
[4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, J. Mol Biol. **215,** 403 (1990).
[5] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Nucleic Acids Res. **25,** 3389 (1997).
[6] A. Krogh, M. Brown, I. S. Mian, K. Sjolander, D. Haussler, J. Mol. Biol. **235,** 1501 (1994).
[7] S. Vinga, J. Almeida, Bioinformatics **19,** 513 (2003).
[8] J. K. Vries, R. Munshi, D. Tobi, J. Klein-Seetharaman, P. V. Benos, I. Bahar, Appl. Bioinformatics **3,** 137 (2004).
[9] K. Harpreet, G. P. S. Raghava, Protein Science **12,** 923 (2003).
[10] E. C. Webb, Enzyme Nomenclature, Academic Press, San Diego CA, (1992).
[11] A. Narayanan, E. C. Keedwell, B. Olsson, Appl. Bioinformatics **1(4),** 191 (2002).
[12] M. Bhasin, G. P. S. Raghava, Nucleic Acids Research **32,** W383 (2004a).
[13] M. Bhasin, G. P. S. Raghava, Nucleic Acids Research **32,** W414 (2004b).
[14] Y. Cai, S. L. Lin, Biochimica et Biophysica Acta. **1648,** 127 (2003).
[15] S. Ahmad, M. M. Gromiha, A. Sarai, Bioinformatics **20**, 477 (2004).
[16] E. W. Stawiski, L. M. Gregoret, Y. Mandel-Gutfreund, J. Mol. Biol. **326**, 1065 (2003).
[17] S. Jones, P. Van Heyningen, H. M. Berman, J. M. Thornton, J. Mol. Biol. **287**, 877 (1999).
[18] S. Ahmad, A. Sarai, J. Mol. Biol. **341**, 65 (2004).
[19] A. Zell, G. Mamier, Stuggart Neural Network Simulator (Version- 4.2), University of Stuttgart, Stuttgart, Germany (1997).
[20] P. Rice, I. Longden, A. Bleasby, Trends in Genetics **16 (6),** 276 (2000).
[21] K. C. Chou, C. T. Zhang, Crit. Rev. Biochem. Mol. Biol. **30**, 275 (1995).
[22] D. J. Hand, 'Construction and assessment of classification rules'. New York: John Wiley and Sons (1997).
[23] K. C. Chou, D. W. Elrod, J. Proteome Res. **2,** 183 (2003).