Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102
Volume 26, Issue Number 6, (2009)

DNA-Prot: Identification of DNA Binding Proteins from Protein Sequence Information using Random Forest

http://www.jbsdonline.com

K. Krishna Kumar[1]
Ganesan Pugalenthi[2]
P. N. Suganthan[2,*]

[1]Institute for Neuro- and Bioinformatics
University of Lübeck
23538 Lübeck, Germany
[2]School of Electrical
and Electronic Engineering
Nanyang Technological University
Block S2, 50 Nanyang Avenue
Singapore 639798, Singapore

Abstract

DNA-binding proteins (DNABPs) are important for various cellular processes, such as transcriptional regulation, recombination, replication, repair, and DNA modification. So far various bioinformatics and machine learning techniques have been applied for identification of DNA-binding proteins from proteins structure. Only few methods are available for the identification of DNA binding proteins from proteins sequence. In this work, we report a random forest method, DNA-Prot, to identify DNA binding proteins from protein sequence. Training was performed on the dataset containing 146 DNABP and 250 non DNABP. The algorithm was tested on the dataset containing 92 DNABP and 100 non DNABP. We obtained 80.31% accuracy from training and 84.37% accuracy from testing. Benchmarking analysis on the independent of 823 DNA-binding proteins and 823 non DNA-binding proteins shows that our approach can distinguish DNA-binding proteins from non DNA-binding proteins with more than 80% accuracy. We also compared our method with DNAbinder method on test dataset and two independent datasets. Comparable performance was observed from both methods on test dataset. In the benchmark dataset containing 823 DNA-binding proteins and 823 non DNA-binding proteins, we obtained significantly better performance from DNA-Prot with 81.83% accuracy whereas DNAbinder achieved only 61.42% accuracy using amino acid composition and 63.5% using PSSM profile. Similarly, DNA-Prot achived better performance rate from the benchmark dataset containing 88 DNA-binding proteins and 233 non DNA-binding proteins. This result shows DNA-Prot can be efficiently used to identify DNA binding proteins from sequence information. The dataset and standalone version of DNA-Prot software can be obtained from http://www3.ntu.edu.sg/home/EPNSugan/index_files/dnaprot.htm.

Introduction

DNA binding proteins (DNABPs) play key roles in various cellular processes. Interaction between DNA and protein is essential for many biological process including transcriptional regulation, recombination, genome rearrangements, replication, repair, and DNA modification (1). DNA-binding proteins are composed of DNA binding domains that include transcription factors, which modulate the process of transcription, nucleases, which cleave DNA molecules, and histones, which are involved in chromosome packaging in the cell nucleus. Generally, DNA binding proteins show substantial variation in sequence and structure. On the basis of amino acid sequence, DNA-binding proteins are divided into various families such as helix-trun-helix, Zinc finger, Leucine zipper, C2-H2, et cetera (2). Similarly, DNA binding proteins are grouped into various structural families based on the DNA recognition motif (1).

In recent years, DNA binding protein has received much attention because of its functional roles. Although many efforts have been made to study the role of DNA-binding proteins, more data are required for the complete understanding of the DNA-binding proteins. There have been many methods reported to identify DNA

[*]Fax: +65 6793 3318
Email: EPNSugan@ntu.edu.sg

binding proteins from protein three dimensional structures (3-7). However, for most proteins, such structural information is not available. With the rapid increase in newly found protein sequences entering into databanks, an efficient method is needed to identify DNA-binding proteins from the sequence databases. Profile-based sequence similarity search methods such as PSI-BLAST (8) and Hidden Markov Models (HMM) (9) have been successfully used to identify DNA binding proteins from sequence databases (10, 11). However, the performance of these methods drops when the sequence similarity is insignificant. Few machine learning methods have been reported to predict DNA binding proteins using sequence derived features such as physico-chemical properties, amino acid composition and pseudo-amino acid composition (12, 13). Recently, M. Kumar *et al*. (2007) reported DNAbinder method that uses amino acid composition and PSSM profiles to discriminate DNA-binding proteins from non-binding proteins (14). High prediction accuracy of 86.62% was reported by DNAbinder using PSSM profiles (14). However, the performance of this method mainly depends on PSSM profile which needs sufficient number of sequence homologs to derive sequence alignment. Although the sensitivity of DNAbinder is high, the specificity remains low. In this work, we report a method, DNA-Prot that use random forest algorithm to identify DNA-binding proteins from sequence derived properties such as frequency of amino acid/amino acid groups, secondary structure, composition of hydrophobic/hydrophilic/neutral amino acids, physico chemical properties, et cetera.

Materials and Methods

Dataset

Training Dataset: The positive and negative dataset for training was obtained from M. Kumar *et al*. (2007) and Stawiski *et al*. (2003) (14, 15). The positive set consists of 146 DNA-binding proteins in which the sequence identity between any two proteins is not more than 25%. The negative dataset has 250 non DNA-binding proteins that have <= 25% sequence identity between any two proteins. In addition, each protein in the negative dataset has approximate size and electrostatics that are similar to DNA-binding proteins. Feature selection was performed on the training dataset.

Test Dataset: The test dataset was obtained from Wang and Brown, 2006 (16). This dataset contains 92 DNA-binding proteins from Protein Data Bank (17) and 100 non DNA-binding proteins from Swiss-Prot database (18).

Independent-dataset1: We created large independent dataset that contains 823 DNA-binding domains and 823 non DNA-binding domains. DNA binding domains were obtained from Pfam database (2). Those sequences having >= 40% sequence similarity with training and testing dataset were removed using CD-HIT program (19). Finally, 823 DNA-binding proteins were retained for independent testing. Similarly, 823 non DNA-binding domains were randomly selected from Pfam protein families that are unrelated to DNA-binding protein family.

Independent-dataset2: This dataset was obtained from N. Bhardwaj *et al*., 2005 (6). We employed CD-HIT program to remove sequences that have > 40% sequence similarity. Finally, we retained a non redundant set of 88 DNA-binding proteins and 233 non DNA-binding proteins. The datasets used in this study are summarized in Table I.

Table I
Datasets used for training, testing, and benchmarking study.

Dataset	No of DNA-binding proteins	No of Non DNA-binding proteins
Training dataset	146	250
Test dataset	92	100
Independent-test1	823	823
Independent-test2	88	233

Input Features

In this work, each sequence is encoded by 116 features. We categorized 20 amino acids into 10 functional groups based on the presence of side chain chemical group such as phenyl (F/W/Y), carboxyl (D/E), imidazole (H), primary amine (K), guanidino (R), thiol (C), sulfur (M), amido

(Q/N), hydroxyl (S/T), and nonpolar (A/G/I/L/V/P) (20). The composition of 20 amino acids (number of occurrences of amino acid x divided by length of the protein), composition of 10 amino acid group and composition of hydrophobic, hydrophilic, and neutral amino acids were computed for each sequence. Twenty-seven tripeptides were derived from all possible combination of hydrophobic, hydrophilic, and neutral amino acid groups. The frequency of 27 tripeptides was calculated for each sequence.

Additionally, we incorporated information about the short peptides (10 residue length, in this case) that are rich in hydrophobic, hydrophilic, or neutral amino acids. First, the sequence was split into sliding 10 residue windows. For each short peptide, the composition of hydrophobic, hydrophilic, or neutral amino acids was calculated. If the short peptide contains more than 6 hydrophobic amino acids, then we considered this peptide as hydrophobic rich peptide. The frequency of hydrophobic rich peptide was calculated from the number of hydrophobic rich peptides divided by total number of short peptides. Similarly, the frequency of hydrophilic rich peptides and neutral amino acid rich peptides were calculated.

Secondary structure information for each sequence was assigned using PSIPRED (21). PSIPRED method provides two options to predict secondary structure. The first option uses homologous sequence information and the second option predicts secondary structure from the query sequence without using homologous sequence information. We employed PSIPRED method that can predict secondary structure from sequence. The overall composition of helix (H), beta sheet (E), coil (C), and the frequencies of 10 amino acid group, hydrophobic, hydrophilic, and neutral amino acids at helix, sheet, and coil regions were calculated.

Matrices containing quantitative values for amino acid physico-chemical properties scaled between 0 and 1 were obtained from the UMBC AAIndex database (22). We considered 14 physico-chemical properties that include molecular weight, hydrophobicity, hydrophilicity, refractivity, average accessible surface area, flexibility, melting point, side chain volume, side chain hydrophobicity, normalized frequency of beta-sheet and alpha helix, polarity, heat capacity, and isoelectric points. The average value for each physico-chemical property was calculated from the sum of physico chemical property of all the amino acids in the sequence divided by total number of amino acids in the sequence

Random Forest Classification

The Random Forest (RF) is a popular classification technique and it has been successfully employed in various biological problems (23-28). The details of random forest method can be found elsewhere (29-32). The brief description of random forest method is provided here. Random forest is a collection of trees, where each tree is grown using a subset of the possible attributes in the input feature vectors. It has been shown that combining multiple trees produced in randomly selected subspaces can improve the generalization accuracy (33). Random forest constructs an ensemble of decision trees from randomly sampled subspaces of the input features, and final classification is obtained by combining results from the trees via voting. Random subspace method is used to avoid over fitting on the training set while preserving the maximum accuracy when training a decision tree classifiers (34). RF performs a type of cross-validation by using out-of-bag (OOB) samples. In training, each tree is constructed using a different bootstrap sample from the original data. Since bootstrapping is sampling with replacement from the training data, some of the sequences will be 'left out' of the sample, while others will be repeated in the sample. The 'left out' sequences constitute the OOB sample. On average, each tree is grown using about $1 - e^{-1} \approx 2/3$ of the training sequences, leaving $e^{-1} \approx 1/3$ as OOB. The RF algorithm was implemented by the randomForest R package (32)

Feature Selection

We used a correlation-based feature subset selection method (CFSS) for the feature selection (35-37). Recently, CFSS method has been successfully used in various bioinformatics problems to reduce the feature dimensionality and potentially improve the prediction accuracy (38-42). Rather than scoring and ranking individual features, the CFSS method scores and ranks the worth of subsets of features. CFSS evaluates the value of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. The strategy for searching subsets was the best-first-search method. Best-first-search explores the space of attribute subsets by using the greedy hill-climbing augmented with the backtracking. The method was implemented using WEKA 3.5 (43).

Prediction Assessment

The prediction system is evaluated using sensitivity, specificity, accuracy, and Mathew's Correlation Coefficient (MCC). These measurements are expressed in terms of true positive (TP), false negative (FN), true negative (TN), and false positive (FP). The measurements are defined as follows:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)} \quad [1]$$

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} \quad [2]$$

$$\text{Specificity} = \frac{TP}{(TN + FP)} \quad [3]$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TN + FN)(TP + FN)(TN + FP)}} \quad [4]$$

Classification by DNA-Prot

We trained our random forest model on the training dataset containing 146 DNA-binding proteins and 250 non DNA-binding proteins. DNA-Prot achieved 80.31% training accuracy using all 116 features. To identify the prominent features, we carried out feature selection using correlation-based feature subset selection method. The features were substantially reduced from 116 to 20.

In order to examine the performance of the newly developed model, we tested our training models on test dataset containing 92 DNA-binding protein chains and 100 non-binding proteins. Table II shows the performance of our method on the test dataset using different feature subsets. As seen Table II, feature selection (reduction) generally does not deteriorate the classification performance much until the number of features decreases to 20. Before that, the usage of smaller number of features only leads to a very little decrease in the sensitivity and specificity rates.

Table II
Performance of Random Forest model on testing dataset containing 92 DNA-binding proteins and 100 non DNA-binding proteins using different feature subsets.

Feature subset	Sensitivity (%)	Specificity (%)	MCC	Accuracy (%)
20	79.35	79.00	0.58	79.17
50	81.52	83.00	0.65	82.29
75	77.17	85.00	0.62	81.25
100	77.17	91.00	0.69	84.38
All	79.30	89.00	0.70	84.37

DNA-Prot achieved 84.37% accuracy using all the features and 79.17% accuracy using 20 features. We also investigated the influence of the feature reduction by plotting Receiver Operating Characteristic (ROC) curves (Figure 1) derived from the sensitivity (true positive rate) and specificity (false positive rate) values for the classifiers using all the features and the 20 best performing features, respectively.

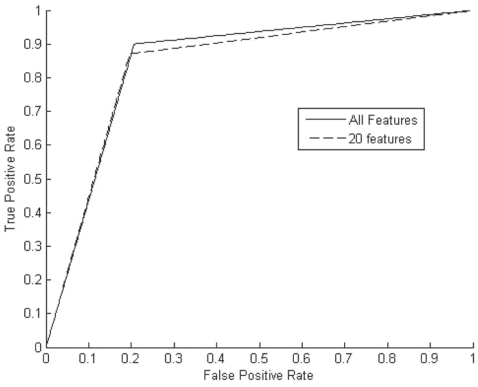

Figure 1: Receiver Operating Characteristic (ROC) curves. ROC curves were plotted utilizing the fractions of true positives and false positives values derived using top 20 features and all features.

It is important to evaluate the performance of new model on the independent dataset that is not part of training and testing dataset. We evaluated our approach on the large independent dataset containing 823 DNA binding proteins and 823 non DNA binding proteins obtained from Pfam database (2). As shown in Table III, DNA-Prot achieved better performance with high sensitivity and specificity. Using all the features, DNA-Prot obtained 81.83% accuracy with 82.02% sensitivity and 81.65% specificity.

Table III
Performance of DNA-Prot on the independent dataset containing 823 DNA-binding proteins and 823 non DNA-binding proteins.

Feature subset	Sensitivity (%)	Specificity (%)	MCC	Accuracy (%)
20	84.93	74.12	0.59	79.53
50	84.80	77.52	0.62	80.80
75	82.87	79.83	0.63	81.35
100	84.08	79.34	0.64	81.71
All	82.02	81.65	0.64	81.83

Comparison of DNA-Prot with SVM Method

In recent years, Support Vector Machine (SVM) (44) has successfully been applied to many bioinformatics problems. Therefore, the SVM method was selected as an alternative method to compare with DNA-Prot. We evaluate the performance of DNA-Prot with SVM using the same feature subsets that are mentioned in Table II. SVM model was trained on the training dataset containing 146 DNABP and 250 non DNABP. Table IV shows the classification result obtained from SVM models on the testing dataset containing 92 DNA-binding protein chains and 100 non DNA-binding proteins. As seen Tables II and IV, DNA-Prot shows better performance than SVM in differentiating DNA-binding proteins from non DNA-binding proteins. With all features, DNA-Prot achieved 84.37% accuracy whereas SVM model achieved 71.87%.

Table IV
Performance of SVM on the test dataset (92 DNA-binding proteins and 100 non DNA-binding proteins) using different feature subsets that are selected by DNA-Prot.

Feature subset	Sensitivity (%)	Specificity (%)	MCC	Accuracy (%)
20	54.35	85.00	0.415	70.31
50	66.30	82.00	0.490	74.47
75	73.91	78.00	0.519	76.04
100	75.00	76.00	0.509	75.52
All	61.96	81.00	0.438	71.87

Performance Comparison of DNA-Prot with DNAbinder

The performance of DNA-Prot was compared with DNAbinder method using three datasets (Table V). There are two reasons for the selection of DNAbinder method for comparison. First, among the few available methods for the prediction of DNA-binding proteins from the sequence information (12, 13), DNAbinder reported better performance rate (14). Second, same training dataset was used in both DNA-Prot and DNAbinder method.

The DNAbinder method was applied on the three datasets using amino acid composition and PSSM profile. The DNAbinder obtained 70.31% accuracy using

amino acid composition and 82.81% using PSSM profile on the test dataset. The performance of DNA-Prot (84.37% accuracy) is comparable to the PSSM based DNAbinder prediction and significantly better than amino acid composition based DNAbinder prediction. Similarly, better performance was observed for DNA-Prot in the independent dataset2. DNA-Prot achieved 93.46% accuracy where as DNAbinder reported 75.39% accuracy using amino acid composition and 82.55% accuracy using PSSM profile.

Although DNAbinder method achieved moderate accuracy with better sensitivity, the specificity is very low. As shown in Table V, DNAbinder obtained more than 80% sensitivity in the independent dataset1. But the overall accuracy dropped to less than 65% due to the poor specificity, which is 33.09% using amino acid composition and 30.61% using PSSM profile. It can be observed that DNA-Prot achieved 81.83% accuracy with better sensitivity (82.02%) and specificity (81.65%) rate.

Table V
Performance comparison of DNA-Prot and DNAbinder using Test dataset (92 DNA-binding proteins and 100 non DNA-binding proteins), Independent-dataset1 (823 DNA-binding proteins and 823 non DNA-binding proteins), and Independent-dataset2 (88 DNA-binding proteins and 233 non DNA-binding proteins ind-set2).

Method	Dataset	Sensitivity (%)	Specificity (%)	MCC	Accuracy (%)
DNA-Prot	Test dataset	79.35	89.00	0.69	84.37
DNAbinder (AA)	Test dataset	95.65	47.00	0.48	70.31
DNAbinder (PSSM)	Test dataset	76.09	89.00	0.66	82.81
DNA-Prot	Independent-dataset1	82.02	81.65	0.64	81.83
DNAbinder (AA)	Independent-dataset1	88.94	33.90	0.28	61.42
DNAbinder (PSSM)	Independent-dataset1	96.47	30.61	0.36	63.50
DNA-Prot	Independent-dataset2	81.82	97.85	0.83	93.46
DNAbinder (AA)	Independent-dataset2	94.32	68.24	0.56	75.39
DNAbinder (PSSM)	Independent-dataset2	87.50	80.69	0.63	82.55

Table VI
List of 20 best performing features selected by DNA-Prot.

No	Name of the feature
1	Frequency of Aspartic acid
2	Frequency of Glycine
3	Frequency of Arginine
4	Frequency of Histidine in strand
5	Frequency of Arginine in helix
6	Frequency of Arginine in coil
7	Frequency of Methionine in strand
8	Frequency of Methionine in coil
9	Frequency of AGILVP in coil
10	Frequency of hydrophobic amino acids in coil
11	Frequency of hydrophilic amino acids in coil
12	Tripeptide containing Neutral-Neutral-Hydrophobic amino acid
13	Tripeptide containing Neutral amino acids
14	Tripeptide containing Neutral-Hydrophilic-Neutral amino acid
15	Physico chemical property (Hydrphobicity)
16	Physico chemical property (hydrophilicity)
17	Physico chemical property (isoelectric point)
18	Physico chemical property (molecular weight)
19	Physico chemical property (accessible surface area)
20	Physico chemical property (side chain volume)

Analysis of the Selected Features

In order to study the influence of the features, we performed brief analysis of the analyzed the 20 best performing features obtained from CFSS. The top 20 selected features are listed in Table VI. Most of the features are correlated well with the previously reported features that differentiate DNA binding protein and non DNA binding protein (5, 45). For instance, it has been reported that the positively charged residues are often clustered near the DNA and negatively charged residues are scattered in the non binding regions (5, 45, 46). It can be seen from Table VI that the frequency of positive charged residues (Arginine and Histidine) and frequency of negatively charged residue (Aspartic residue) occurs in the top 20 features. The frequency of glycine was found in the list. Glycine may provide to conformational flexibility needed during the process of binding (5). Another important selected feature is accessible surface area that has been shown to play role in differentiating DNA binding residues from non binding residues (5). Hydrophobic and hydrophilic amino acids have significant influence in protein-DNA interaction. Previous studies pointed out that hydrophobic residues are critical for DNA binding activity (47).

Execution Time

The execution time for our algorithm is reasonably faster. The prediction procedure involves secondary structure prediction, feature formulation and prediction using DNA-Prot model. In order to provide a flavor of the computation time for DNA-Prot, we applied DNA-Prot on 192 proteins with varying lengths and measured the

user CPU time spent for the feature generation followed by prediction on a Pentium4 machine having 3 GHz CPU and 2GB memory (Table VII).

Table VII
Average time taken for the prediction by DNA-Prot

Length of protein	Average prediction time (in seconds)
40-200	6.87
201-400	7.22
401-600	7.63
601-800	8.17
801-1000	8.33
Above 1000	9.14

Conclusion

DNA-binding proteins play important roles in various cellular processes. Identification of DNA-binding proteins from protein sequence is essential and also difficult task. We implemented random forest approach to predict DNA binding proteins using sequence derived properties. Validation of the DNA-Prot on the test dataset obtained 84.37% accuracy with high sensitivity (79.30%) and specificity (89.00%). The performance of DNA-Prot was compared with DNAbinder on three datasets. DNA-Prot reported significantly better performance than DNAbinder method. The computation time for the DNA-Prot prediction process was assessed using 192 proteins with varying in lengths. Our analysis shows that DNA-Prot method can be used to identify DNA binding proteins from protein sequence. The dataset and standalone version of DNA-Prot software can be obtained from http://www3.ntu.edu.sg/home/EPNSugan/index_files/dnaprot.htm.

Acknowledgment

G. P. and P. N. S. acknowledge the financial support offered by the A*Star (Agency for Science, Technology and Research). K. K. K. acknowledges Mr. Rajeev Gangal, CEO, Insilico division, Systems Biology India Pvt Ltd, Maharashtra, India for providing various resources to accomplish this research work and Prof. Thomas Martinetz and Dr. Stefen Moller, Institute for Neuro- and Bioinformatics, University of Luebeck, Germany for their support and critical reading of the manuscript.

Reference and Footnotes

1. N. M. Luscombe, S. E. Austin, H. M. Berman, J. M. Thornton. *Genome Biol 1* (2000).
2. L. Sonnhammer, S. R. Eddy, R. Durbin. *Proteins 28*, 405-420 (1997).
3. S. Jones, J. A. Barker, I. Nobeli, J. M. Thornton. *Nucleic Acids Res 35*, 2811-2823 (2003).
4. H. P. Shanahan, M. A. Garcia, S. Jones, J. M. Thornton. *Nucleic Acids Res 32*, 4732-4741 (2004).
5. S. Ahmad and A. Sarai. *J Mol Biol 341*, 65-71 (2004).
6. N. Bhardwaj, R. E. Langlois, G. Zhao, H. Lu. *Nucleic Acids Res 33*, 6486-6493 (2005).
7. J. V. Ponomarenko, P. E. Bourne, I. N. Shindyalov. *Bioinformatics 18*, S192-S201 (2002).
8. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman. *Nucleic Acids Res 25*, 3389-3402 (1997).
9. S. R. Eddy. *Bioinformatics 14*, 755-763 (1998).
10. G. Pugalenthi, A. Bhaduri, R. Sowdhamini. *Nucleic Acids Res 33*, D252-D255 (2005).
11. J. Gough, K. Karplus, R. Hughey, C. Chothia. *J Mol Biol 313*, 903-919 (2001).
12. Y. D. Cai, S. L. Lin. *Biochim Biophys Acta 1648*, 127-133 (2003).
13. X.Yu ,J.Cao,Y.Cai, T.Shi, Y.Li, J Theor Biol 240,175-184(2006).
14. M. Kumar, M. M. Gromiha, G. P. Raghava. *BMC Bioinformatics 8*, 463 (2007).
15. E. W, Stawiski, L. M.Gregoret, Y. Mandel-Gutfreund. *J Mol Biol 326*, 1065-1079 (2003).
16. L. Wang, S. J. Brown. *Nucleic Acids Res 34*, W243-W248 (2006).
17. G. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne. *Nucleic Acids Res 28*, 235-242 (2000).
18. E. Gasteiger, E. Jung, A. Bairoch. *Mol Biol 3*, 47-55 (2001).
19. W. Li, L. Jaroszewski, A. Godzik. *Bioinformatics 17*, 82-283 (2001).
20. G. Pugalenthi, K. K. Kumar, P. N. Suganthan, R. Gangal. *Biochem Biophys Res Commun 367*, 630-634 (2008).
21. L. J. McGuffin, K. Bryson, D. T. Jones. *Bioinformatics 16*, 404-405 (2000).
22. S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa. *Nucleic Acids Res 36*, D202-205 (2008).
23. B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, H. Zhao. *Bioinformatics 19*, 1636-1643(2003).
24. S. Dudoit, J. Fridlyan, T. P. Fridlyan. *J Am Stat Assoc 97*, 77-87 (2002).
25. J. W. Lee, J. B. Lee, M. Park, S. H. Song. *Comput Stat Data Anal 48*, 869-885 (2005).
26. R. D. Uriarte and S. A. Andres. *BMC Bioinformatics*, 3 (2006).
27. Y. Qi , J. K. Seetharaman J, Z. B. Joseph. *Pac Symp Biocomput*, 531-542 (2005).
28. A. Statnikov, L. Wang, C. F. Aliferis. *BMC Bioinformatics 9*, 319 (2008).
29. L. Breiman. *Machine Learning 45*, 5-32 (2001).
30. Ho, T. Kam. *Pattern Analysis and Applications 5*, 102-112 (2002).

31. L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone. Chapman & Hall, New York (1984).
32. A. Liaw and M. Wiener. *R News 2*, 18-22 (2002).
33. T. K. Ho, J. J. Hull, S. N. Srihari. *IEEE Trans on Pattern Analysis and Machine Intelligence 16*, 66-75 (1994).
34. Ho, T. Kam. *IEEE Trans on Pattern Analysis and Machine Intelligence 20*, 832-844 (1998).
35. M. Hall. *Correlation based feature selection for machine learning*, Ph.D. dissertation, University of Waikato, Dept of Computer Science (1999).
36. M. Hall and L. A. Smith. *International Conference on Neural Information Processing and Intelligent Information Systems*, p. 855-858. Berlin: Springer (1997).
37. L. Yu H. Liu. *Proceedings of the Twentieth International Conference on Machine Learning* (ICML-2003), Washington DC (2003).
38. K. Chen, L. A. Kurgan, J. Ruan. *BMC Struct Biol 7*, 25 (2007).
39. F. Zeng, R. H. Yap, L. Wong. *Genome Inform 13*, 192-200 (2002).
40. M. Osl, S. Dreiseitl, B. Pfeifer, K. Weinberger, H. Klocker, G. Bartsch, G. Schäfer, B. Tilg, A. Graber, C. Baumgartner. *Bioinformatics 24*, 2908-2914 (2008).
41. C. H. Ooi, M. Chetty, S. W. Teng. *BMC Bioinformatics 7*, 320 (2006).
42. H. Liu, J. Li, L. Wong. *Genome Inform 13*, 51-60 (2002).
43. E. Frank, M. Hall, L. Trigg, G. Holmes, I. H. Witten. *Bioinformatics 20*, 2479-2481 (2004).
44. V. N.Vapnik. New York, Wiley (1998).
45. S. Ahmad, M. M. Gromiha, A. Sarai. *Bioinformatics 20*, 477-486 (2004).
46. K. Nadassy, S. J. Wodak, J. Janin. *Biochemistry 38*, 1999-2017 (1999).
47. M. West, D. Flanery, K. Woytek, D. Rangasamy, V. G. Wilson. *J Virol 75*, 11948-11960 (2001).

Date Received: November 1, 2008

Communicated by the Editor Ramaswamy H. Sarma