BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 9(10)

SVM based model generation for binding site prediction on helix turn helix motif type of transcription factors in eukaryotes

Koel Mukherjee, Abhipriya, Ambarish Saran Vidyarthi & Dev Mani Pandey*

Department of Biotechnology, Birla Institute of Technology, Mesra, Ranchi-835 215, Jharkhand, India; Dev Mani Pandey – Email: dmpandey@bitmesra.ac.in; Tel: +91 651 2276223, Fax: +91 651 2276052; *Corresponding author

Received May 14, 2013; Accepted May 17, 2013; Published June 08, 2013

Abstract:
Support vector machine is a class of machine learning algorithms which uses a set of related supervised learning methods for classification and regression. Nowadays this method is vividly applied to many detection problems related with secondary structure, tumor cell and binding residue prediction. In this work, support vector machines (SVMs) have been trained on 90 sequences of transcription factors with HTH motif. Four sequence features were used as attribute for the prediction of interaction site in HTH motif. A web page was also developed so that user can easily enter the protein sequence and receive the output as interaction site predicted or not predicted. The generated model shows a very high amount of accuracy, sensitivity and specificity which proves to be a good model for the selected case.

Keywords: Support vector machine, machine learning algorithm, confusion matrix, helix turn helix motif.

Background:
Machine learning is a branch of artificial intelligence which deals with the construction and evaluation of algorithms that expedite pattern recognition, classification and prediction based on models derived from existing data **[1].** Sophisticated machine learning algorithms (MLA) have also been attempted **[2]** on biological aspects. These utilize techniques such as neural networks **[3],** logistic regression **[4]** and support vector machines (SVM) **[5].** SVM as a supervised machine learning technology is attractive because it has an extremely well developed statistical learning theory **[6, 7].** SVM is currently the best performers for a number of classification tasks ranging from text to genomic data. It has been gradually applied to protein secondary structure prediction **[8],** tumor cell prediction and DNA binding residues prediction **[9]** and many more pattern classification problems in biology related areas. Recently **[10]** SVM has been used as a new and promising technique for detecting DNA binding site prediction for TF proteins.

SVM based **[5, 11, 12]** method for binding site prediction depend on the sequence **[13]** and structure characteristics or both **[14]** of known protein binding sites. These characteristics such as sequence conservation **[15],** interface propensities **[16, 17],** the charge or dipole moment of the protein molecule, the presence of positively charged surface patches, secondary structure **[18],** accessible surface area **[19],** 3D motifs **[20]** and residue evolutionary information **[21]** are applied to train a set of data's for model generation.

The helix turn helix (HTH) motif is a major structural DNA binding motif. It is composed of two α-helices joined by short strand of amino acid and is found in many proteins that regulate gene expression **[22].** The two α helices of tri-helical HTH motif, one occupying the N-terminal end and the other occupying the C-terminal end of the helix are involved in recognition and binding to the DNA **[23].** The third helix is known as the recognition helix which has prime role in establishing the contacts with DNA major groove. The positively charged and hydrophobic amino acids residing in the

third helix increase the bonding capacity in DNA-protein complexes **[24, 25].** A very large class of transcription factors can be classified as HTH motif type of family according to Interpro database (http://www.ebi.ac.uk/interpro/).

In the present work, the characteristic features of interaction sites of DNA-binding proteins with HTH motif from the residue level were analyzed. The four sequence features include the evolutionary sequence conservation, positively charged residues, H-bond donor and acceptor residues and hydrophobic residues. A SVM based model was trained with 90 sequences (training set) using the above sequence features for the prediction of DNA-binding residues in HTH motif type of proteins. The total dataset was selected for training and test was 136 sequences and three fold cross validation was also done on the training set to get an average value of accuracy, sensitivity and specificity.

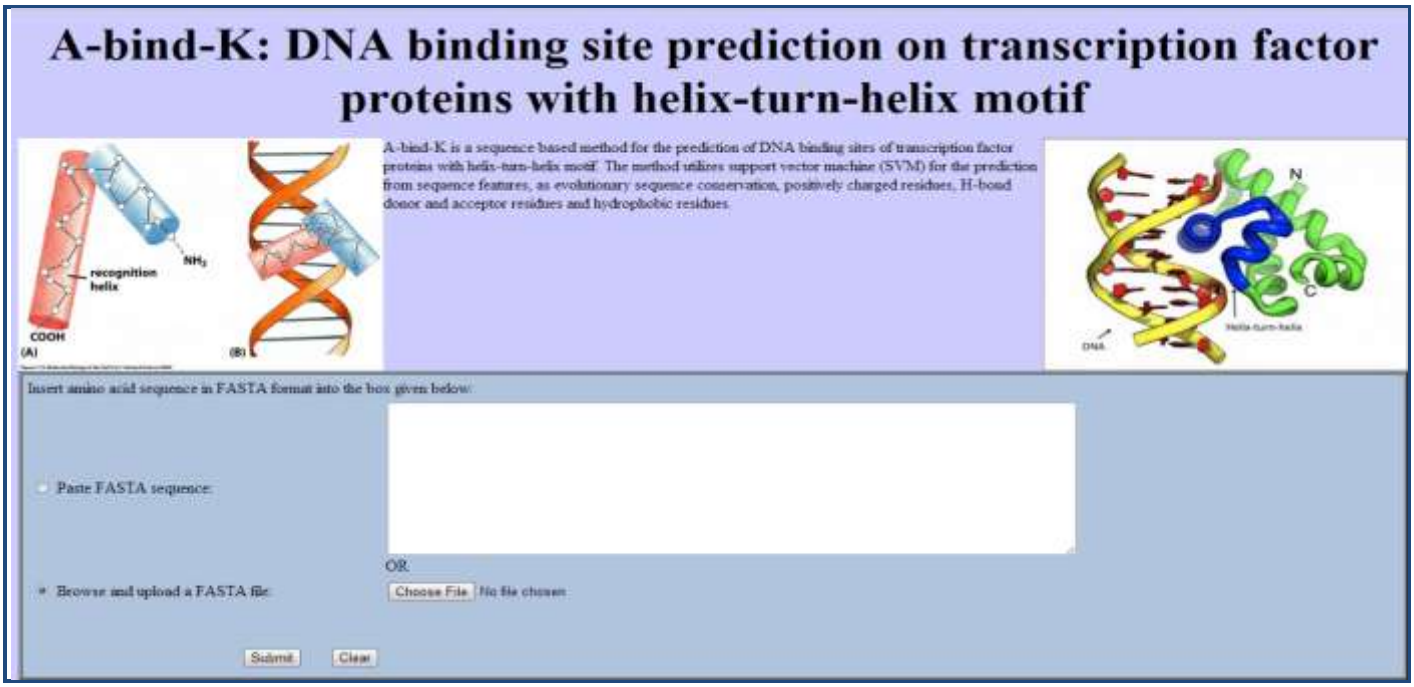

Figure 1: Snap shot of the web page designed for the DNA binding site prediction

Methodology:

Dataset selection

Sequence data set was retrieved from NCBI (http://www.ncbi.nlm.nih.gov/), 91 proteins (TFs) with HTH motif and 45 protein sequences without HTH motif of eukaryotes were collected. All these selected proteins are having DNA binding domain with HTH motif which was verified by Pfam database (http://pfam.sanger.ac.uk/) **[26].** Out of 136 total protein sequences 90 sequences were chosen randomly as training set and others as test set so that the ¾ of the total will remain as training set and ¼ as test set.

Features selection for training and test data sets

After the determination of crystal structures of C1 and Cro repressor proteins from bacteriophage lambda **[27],** the DNA binding HTH structural motif [22] has become one of the most important and studied examples of the interaction between proteins and DNA. The typical length of the HTH motif is around 25 residues. DNA interaction site was predicted through ScanProsite **[28]** available at http://prosite.expasy.org/scanprosite/. It was found that this site mainly compose of the maximum portion of HTH motif **Table 7 (see supplementary material).** For non HTH motif sequences interaction site with different motifs were found.

On the basis of these observations four features were chosen in this work for the interaction site. These are evolutionary sequence conservation, positively charged residues, H-bond donor/acceptor and hydrophobic residues **(Table 1).** Conservation of residues was determined from multiple sequence analysis **[10].** From the MSA result it was evident that in five columns (out of 25) the conservation score was very high. Thus, conditional probability was applied to find the value of each residue situated in that position. The value was calculated by,

$$P(AB) = P(A)\,P(B/A) \qquad (1)$$

Where A represents the column number to be selected and B represents the residue to be in that column.

As per many studies of DNA-protein interaction **[2, 29, 30]** the positively charges residues shows more prone of binding capacity with negatively charged residues of DNA strand. Side chain groups of positively charged residues as arginine, histidine and lysine within the interaction site were assumed to be capable of getting involved in H-bonding during the interaction **[10, 13].** Asparagine, arginine, glutamine, cysteine, histidine, serine, threonine, tryptophan and tyrosine were calculated as the H-bond donor, while aspartic acid and glutamic acid were calculated as the H-bond acceptor. The values were calculated and noted down for each feature. So, altogether four features were used as attributes for modeling the SVM classifier from the residue level of amino acids.

SVM implementation

The freely downloadable LIBSVM package was used for the implementation of SVM **[31]** with MATLAB interface. The widely used Radial Basis Function (RBF) kernel was used. In the present study our training and testing dataset was verified by simulating it in Matlab 7.0, Microsoft Windows 2007 Operating System and Intel Pentium D-2.80 GHz with 2 GM of Random Access Memory (RAM).

Web page designing

The concerned web page based on above selected SVM model has a simple user interface. It contains an input field where user can paste the FASTA format protein sequence. There is also a provision of uploading text file with protein sequence. The web page also gives a brief description of our model. The overall web page was designed in HTML codes with different color combination and pictures incorporated in it **(Figure 1).** When the user enters the data in the form, these values are passed to the matlab engine. Once the matlab engine is ready, matlab code is executed from within the java environment.

Results & Discussion:

Goal of the work was to characterize the interaction site in HTH motif as well as non-HTH motif of TF proteins and train the model based on SVM approach. The accuracy of the model was also checked and verified with existing models.

In this work a set of features were selected, studied and applied to distinguish the interaction site with non-interaction site. The interaction site of HTH motif and non HTH motif was retrieved from ScanProsite result. The outcome presents the type of motif, its length, interaction site with DNA and its length **(Figure 2).** DNA interaction site identified by ScanProsite mainly constitute of a large portion of HTH motif. The total length of the HTH motif was approximately 20 to 25 residues long.

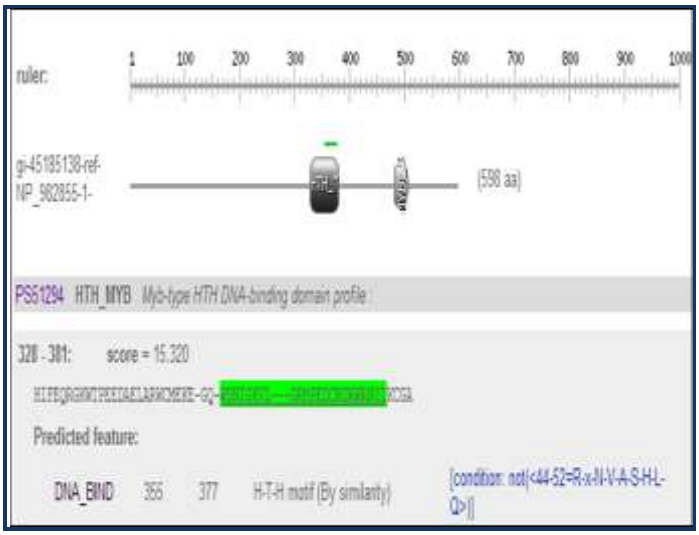

Figure 2: ScanProsite result of yeast protein showing the HTH motif as well as the interaction site

Due to the much specified selection criteria of the samples, the total data set used for the study were 136 which is a pretty small in size. But in spite of fewer samples the k-fold cross validation resampling technique was employed for training and testing of the classifier for the selected features. Considering the k=3, the whole dataset was randomly divided into three parts. Each an every parts represent the same proportion (approximately) as in the original data set. The procedure was repeated for total 3 times where two parts of the data was selected as training set and one part as test set. For all 3 fold confusion matrix was generated by the LIBSVM **Table 1, Table 2 & Table 3 (see supplementary material).**

Three performance matrixes were averaged to obtain an overall estimate of the classifier performance. Parallel with SVM approach the Multilayer perceptron (MLP) and K-nearest neighbor (KNN) method was also applied for comparative studies. The prediction outcome was classified and counted for the methods. Consistent with all other references **[13, 31]** the performance criteria used are accuracy **Table 4 (see supplementary material),** sensitivity **Table 5 (see supplementary material)** and specificity **Table 6 (see supplementary material).** These are defined as follows:

$Accuracy = TP+TN/(TP=TN=FP=FN)$
$Specificity=TN/(TN+FP)$
$Sensitivity=TP/(TP+FN)$

The numbers of true positive, true negative, false positive and false negative results are indicated by TP, TN, FP and FN respectively, and were calculated from the following confusion matrixes. The TP, TN, FP and FN are defined as follows: TP = Correct positive predictions / Total positives; FP = Incorrect negative predictions / Total negatives; TN = Correct negative predictions / Total negative; FN = Incorrect positive predictions / Total positive.

The above tables of accuracy **Table 4 (see supplementary material)**, sensitivity **Table 5 (see supplementary material)** and specificity **Table 6 (see supplementary material)** indicate that our model is showing the best result with average accuracy of 94.19%. Here we showed that this SVM model can predict interaction sites (DNA-binding residues) at 96.7% sensitivity and 89.16% specificity. The result of our study suggest that this SVM – based approach can be used as a good prediction tool for HTH motif type of interaction studies. In view of the previous knowledge **[32],** four web servers are available at the moment for sequence-based prediction of DNA-binding sites: DBS-PRED **[13],** DBS-PSSM **[33],** BindN **[34]** and DP-Bind **[9].**Our performance measures are higher than those reported by these four studies.

Conclusion:

In this work, we have described a new SVM-based approach for prediction of binding sites within HTH motif based on amino acid sequence data. The present work is the 1st SVM-based method for HTH motif interaction site prediction till date. The average accuracy of this work is 94.19%, which is a pretty good value for prediction purpose. K-fold cross re-sampling technique was used for validation. The sensitivity (96.7%) and specificity (89.16%) shows that this SVM – based approach can be used as a good prediction tool for HTH motif type of interaction studies. The output result of our model shows two outcomes in the web page. If HTH motif is present in the query protein, then interaction site is predicted otherwise, there is no prediction. In the future, prediction tool for other types of motifs can be developed as well.

Acknowledgement:
We thankfully acknowledge Mr. Suvrajeet Mahapatro Department of IT, BIT Mesra, Ranchi, India for extending the help and suggestion at the crucial time.

References:

[1] Tarca AL *et al*. *PLoS Comput Biol*. 2007 **3:** e116 [PMID: 17604446]
[2] Bhardwaj N & Lu H, *FEBS Lett*. 2007 **581:** 1058 [PMID: 17316627]
[3] Tjong H & Zhou HX, *Nucleic Acids Res*. 2007 **35:** 1465 [PMID:17284455]
[4] Szilagyi A & Skolnick J, *J Mol Biol*. 2006 **358**: 922 [PMID: 16551468]
[5] Koike A & Takagi T, *Protein Eng Des Sel*. 2004 **17**: 165 [PMID: 15047913]
[6] Vapnik V, *Springer-Verlag*. 1995 New York.
[7] Vapnik V, *John Wiley and Sons Inc.* 1998 New York.
[8] Mandle KA *et al*. *International Journal on Soft Computing*. 2012 **3:** 67
[9] Hwang S *et al*. *Bioinformatics*. 2007 **23:** 634 [PMID: 17237068]
[10] Dey S *et al*. *Nucleic Acids Res*. 2012 **40**: 7150 **[**PMID: 22641851]
[11] Bradford JR & Westhead DR, *Bioinformatics*. 2005 **21:** 1487 [PMID: 15613384]
[12] Li N *et al*. *BMC Bioinformatics.* 2008 **9:** 553 [PMID: 19102736]
[13] Ahmad S *et al*. *Bioinformatics*. 2004 **20:** 477 [PMID: 14990443]
[14] Kim WK *et al*. *PLoS Comput Biol*. 2006 **2:** e124 [PMID: 17009862]
[15] Zhou HX & Shan Y, *Proteins*. 2001 **44:** 336 [PMID:11455607]
[16] Dong Q *et al*. *BMC Bioinformatics*. 2007 **8:** 147 [PMID: 17480235]
[17] Liu B *et al*. *Comput Biol Chem*. 2009 **33:** 303 [PMID: 19646926]
[18] Neuvirth H *et al*. *J Mol Biol*. 2004 **338**: 181 [PMID: 15050833]
[19] Porollo A & Meller J, *Proteins.* 2007 **66:** 630 [PMID: 17152079]
[20] Henschel A *et al*. *BMC Bioinformatics*. 2007 **8**: S5 [PMID: 17570148]
[21] Res I *et al*. *Bioinformatics.* 2005 **21:** 2496 [PMID: 15728113]
[22] Brennan RG & Matthews BW, *J Biol Chem.* 1989 **264**: 1903 [PMID: 2644244]
[23] Ko S *et al*. *Nucleic Acids Res*. 2008 **36:** 2739 [PMID: 18367475]
[24] Nishikawa T *et al*. *Structure*. 2001 **9:** 1237 [PMID: 11738049]
[25] Wintjens R & Rooman M, *J Mol Biol*. 1996 **262**: 294 [PMID: 8831795]
[26] Finn DR *et al*. *Nucl Acids Res*. 2008 **36:** D281 [PMID: 18039703]
[27] Anderson WF *et al*. *Nature*. 1981 **290:** 754 [PMID: 6452580]
[28] De Castro *et al*. *Nucleic Acids Res.* 2006 **34:** W362 [PMID: 16845026]
[29] Burgoyne NJ & Jackson RM, *Bioinformatics.* 2006 **22**: 1335 [PMID: 16522669]
[30] Dong Q *et al*. *BMC Bioinformatics*. 2007 **8:** 147 [PMID: 17480235]
[31] Joachims T, *Dissertation*. 2002
[32] Qatawneh S *et al*. *International Journal on Bioinformatics & Biosciences.* 2012 **2:** DOI: 10.5121/ijbb.2012.2201
[33] Ahmad S & Sarai A, *BMC Bioinformatics*. 2005 **6**: 33 [PMID: 15720719]
[34] Wang L & Brown SJ, *Nucleic Acids Res*. 2006 **34**: W243 [PMID: 16845003]

Edited by P Kangueane
Citation: **Koel *et al.*** Bioinformation 9(10): 500-505 (2013)

BIOINFORMATION

open access

Supplementary material:

Table 1: Confusion table for k=1 fold in SVM based model

	Interaction site	Non-interaction site
Interaction site	30	2
Non-interaction site	1	13

Table 2: Confusion table for k=2 fold in SVM based model

	Interaction site	Non-interaction site
Interaction site	29	2
Non-interaction site	1	14

Table 3: Confusion table for k=3 fold in SVM based model

	Interaction site	Non-interaction site
Interaction site	30	1
Non-interaction site	1	14

Table 4: Comparative study of the accuracy of the present method (SVM) with other methods on the same dataset. Values are in percentage (%)

Approaches	FOLD (k)			Average
	1	2	3	
MLP	88.76	91.34	93.3	91.13
KNN	90. 32	92.45	94.23	92.33
SVM	93.47	93.47	95.65	94.19

Table 5: Comparative study of the sensitivity of the present method (SVM) with other methods on the same dataset. Values are in percentage (%)

Approaches	FOLD (k)			Average
	1	2	3	
MLP	75.22	84.33	88.76	82.77
KNN	85.89	86.45	90.90	87.746
SVM	96.77	96.66	96.67	96.7

Table 6: Comparative study of the specificity of the present method (SVM) with other methods on the same dataset. Values are in percentage (%)

Approaches	FOLD (k)			Average
	1	2	3	
MLP	79.84	83.45	87.67	83.65
KNN	84.55	88.97	90	87.84
SVM	86.66	87.5	93.33	89.16

Table 7: Selected feature values of the training set. Both HTH and non-HTH motif result is given

SL No.	Feature 1	Feature 2	Feature 3	Feature 4
1	0.25	0.25	0.58	0.43
2	0.21	0.30	0.60	0.84
3	0.30	0.34	0.56	0.57
4	0.21	0.34	0.60	0.84
5	0.12	0.37	0.54	0.61
6	0.20	0.29	0.58	0.83
7	0.25	0.37	0.56	0.83
8	0.21	0.34	0.60	0.84
9	0.21	0.30	0.65	0.85
10	0.21	0.30	0.65	0.85
11	0.21	0.34	0.60	0.84
12	0.21	0.34	0.65	0.85
13	0.25	0.30	0.65	0.85
14	0.21	0.34	0.60	0.84
15	0.21	0.34	0.60	0.84
16	0.21	0.30	0.60	0.84
17	0.21	0.30	0.60	0.84
18	0.21	0.30	0.65	0.85
19	0.17	0.21	0.43	0.57

20	0.21	0.34	0.65	0.85
21	0.21	0.30	0.60	0.84
22	0.30	0.34	0.56	0.83
23	0.17	0.26	0.65	0.85
24	0.21	0.26	0.52	0.57
25	0.21	0.26	0.56	0.83
26	0.25	0.35	0.42	0.61
27	0.25	0.35	0.39	0.45
28	0.32	0.32	0.39	0.45
29	0.32	0.28	0.46	0.61
30	0.28	0.32	0.50	0.65
31	0.32	0.35	0.46	0.43
32	0.25	0.37	0.55	0.83
33	0.25	0.37	0.50	0.88
34	0.25	0.37	0.50	0.88
35	0.23	0.37	0.58	0.55
36	0.28	0.36	0.56	0.83
37	0.33	0.37	0.58	0.83
38	0.33	0.32	0.54	0.43
39	0.33	0.32	0.56	0.43
40	0.33	0.32	0.46	0.45
41	0.33	0.32	0.46	0.57
42	0.33	0.32	0.46	0.45
43	0.32	0.32	0.42	0.57
44	0.32	0.32	0.42	0.43
45	0.28	0.32	0.50	0.88
46	0.32	0.32	0.42	0.45
47	0.28	0.32	0.54	0.88
48	0.28	0.32	0.46	0.43
49	0.37	0.33	0.54	0.61
50	0.21	0.34	0.60	0.85
51	0.23	0.34	0.42	0.43
52	00.32	0.38	0.53	0.61
53	00.33	0.38	0.53	0.88
54	0.23	0.34	0.46	0.43
55	0.23	0.34	0.46	0.54
56	0.23	0.34	0.42	0.43
57	0.23	0.34	0.42	0.54
58	0.21	0.34	0.65	0.85
59	0.21	0.34	0.69	0.88
60	0.21	0.30	0.65	0.85
61	0.21	0.32	0.65	0.88
62	0.21	0.33	0.56	0.83
63	0.29	0.33	0.62	0.85
64	0.21	0.34	0.56	0.83
65	0.29	0.33	0.62	0.88
66	0.22	0.27	0.72	0.88
67	0.31	0.31	0.63	0.88
68	0.18	0.43	0.54	0.43
69	0.27	0.42	0.55	0.31
70	0.22	0.42	0.63	0.45
71	0.34	0.43	0.47	0.30
72	0.21	0.14	0.60	0.44
73	0.34	0.32	0.65	0.45
74	0.17	0.33	0.69	0.44
75	0.26	0.30	0.65	0.45
76	0.34	0.26	0.60	0.44
77	0.26	0.21	0.65	0.45
78	0.26	0.33	0.69	0.44
79	0.33	0.43	0.39	0.26
80	0.26	0.34	0.65	0.45
81	0.27	0.18	0.77	0.55
82	0.27	0.22	0.72	0.55
83	0.27	0.47	0.54	0.43
84	0.31	0.27	0.68	0.45
85	0.36	0.27	0.63	0.45
86	0.31	0.36	0.59	0.43
87	0.08	0.31	0.59	0.43
88	0.31	0.36	0.54	0.35
89	0.31	0.41	0.54	0.35
90	0.27	0.27	0.59	0.40