You Ratanaksamrith

HBond-DPI: Hydrogen bond interactions in DNA-binding protein prediction

Abstract

HBond-DPI is a program that predict the possibility of hydrogen bond interactions that occur at DNA-binding site in protein. DNA-protein complexes used in this program were extracted from PDB and NPIDB. The residues are divided into data instances, each has 11 residues and each residue is translated into three features: (1) side chain pKa value, (2) hydrophobicity index, and (3) molecular mass of an amino acid. After standardization, all training data instances is fed into support vector machine model to build an estimator for predicting h-bond interactions in DNA-binding protein. The result of the program appears to be 90.73% accuracy with sensitivity of 65.82% and specificity of 95.10%, while the MCC is 0.62.

INTRODUCTION

Protein-DNA interaction is very essential in fundamental biological activities. As a result, protein-DNA recognition process has become a field of intense research for decades. Trying to identify those complex interactions between DNA and protein is one of the keys to understand the mechanism of gene regulation. This is what motivate the increasing research in determining DNA-protein interaction through traditional approach (e.g. X-ray crystallography, NMR, ...) and computational approach (e.g. algorithm based on individual descriptors, simple statistical methods, machine learning methods, hybrid learning and meta-prediction methods) (13). The latter method will be used in this study.

Unlike previous DNA-binding protein prediction programs, this project will study only hydrogen bond interactions that bind DNA to protein in DNA-protein complexes, and it does not take into account of any other forms of binding such as water-mediated bonds, van der Waals contacts or ionic interactions. This limited scope makes the method seems to be trivial compared to other methods. However, with this targeting extent, we can compartmentalize our study to a very focused aspect of interaction (i.e. hydrogen bonds), and this will help especially researchers who are interested in hydrogen bond in DNA-protein interaction.

There are 3 main reasons that only hydrogen bond is chosen to be examined in this project. First of all, recognition of a DNA sequence by a protein is achieved by interface-coupled chemical and shape complementation. This complementation between the two molecules is clearly directional and is determined by the specific chemical contacts including mainly hydrogen bonds (1). Secondly, hydrogen bonds, though not strong compared to covalent bonds or ionic bonds, provide more stable structure than other weaker bonds. Thirdly, when a prediction includes all types of binding, which consist of h-bond and other very weak bonds that is not stable that contribute to dynamic of the environment, the datasets will contain insufficiently reliable bindings. Therefore, it is favorable to discount other weaker and unstable forces other than hydrogen bonds.

The previous approach that inspired HBond-DPI is BindN which was one of the state-of-the-art prediction programs. BindN used the same 3 features (pKa, hydrophobicity index, and molecular mass) for constructing data instances for input vectors in SVMs model and the performance was at 69.40% sensitivity and 70.47% specificity (2). There are several other DNA-binding protein prediction web servers available such as DP-Bind (3), DBindR (4), bind-rf (5), DNABINDPROT (6), MetaDBSite (7), DR_bind (8), DNABR (9), and DNABind (10). Aforementioned, in the present work, HBond-DPI considers only h-bond in DNA-binding protein, while majority of programs consider any form of bindings in DNA binding protein. The support vector machine is trained using 3 sequence features because they are very efficient to compute for prediction.

DATA REPRESENTATION

Amino acid sequence chains listed in PDNA-62 (Supplementary Table 3) has been used to construct SVM model for predicting h-bond DNA-binding protein. PDNA-62 contains 62 DNA-binding protein complexes, 67 chains and had less than 25% identity among the sequences (12). The process of extracting the sequence chains began by downloading the PDB files of DNA-protein complexes from Protein Database (PDB). After that, the same amino acid sequences are extracted from Nucleic Acid-Protein Interaction DataBase (NPIDB) for identifying hydrogen bond residues between DNA and protein. The combination residue sequences from PDB and NPIDB according to PDNA-62 coupling with pKa value, hydrophobicity index and molecular mass, we can produce training dataset for the SVM classifier.

The training set is constructed by sliding window with the size of 11 across the amino acid sequences. The target residue is positioned in the center of the window. Each data instance from the slide window was labeled with 1 (positive) if the target residue was h-bond DNA-binding, or 0 (negative) if the target residue had no h-bond DNA-binding.

Next, three biochemical features are encoded correspond to each residue, including pKa value, hydrophobicity index and molecular mass of the amino acid (Supplementary Table 1). Consequently, the input vector has 33 feature values.

The three biochemical features are efficient and relevant for prediction of DNA-binding residues. The side chain pKa value determines the ionization state of a residue because phosphate groups of nucleic acids are negatively charged. Hydrophobicity is the cause of amino acid side chain packing and protein folding located inside globular proteins that indicate how well it interacts on DNA surface.

Table 1. Constructing SVMs input vectors. Using PDB ID: 1GAT, chain A as an example.

1GAT-A amino acid sequence 'KRAGTVCSNCQTSTTTLWRRSPMGDPVCNACGLYYKLHQVNRPLTMRKDGIQTRNRKVSS'
1GAT-A window sequence 'KRAGTVCSNCQ', 'RAGTVCSNCQT', 'AGTVCSNCQTS', '…', 'GIQTRNRKVSS'
1GAT-A labeled data instances represented with 3 biochemical ([11, -3.9, 146, 12, -4.5, 174, 7, 1.8, 89, 7, -0.4, 75, 7, -0.7, 119, 7, 4.2, 117, 8, 2.5, 121, 7, -0.8, 105, 7, -3.5, 132, 8, 2.5, 121, 7, -3.5, 146], **'0'**), (…), ([7, -0.4, 75, 7, 4.5, 131, 7, -3.5, 146, 7, -0.7, 119, 12, -4.5, 174, 7, -3.5, 132, 12, -4.5, 174, 11, -3.9, 146, 7, 4.2, 117, 7, -0.8, 105, 7, -0.8, 105], **'0'**)

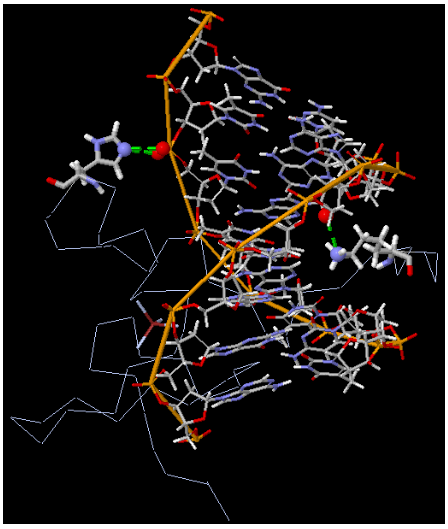

Figure 1. Three-dimensional structure of 1GAT complex. The green lines represent hydrogen bonds between DNA strand and protein residues.

Hydrophobicity scale developed by Kyte and Doolittle is used. Molecular mass represents volume of space that a residue occupies (2). The example and summary of steps to construct input vectors is in Table 1.

MACHINE LEARNING METHOD

To identify the h-bond interactions in DNA-protein complex, computational approach will be used because it is better than traditional approach in terms of time-consumption and cost (13). For the past decades, many machine learning methods have been developed and applied to DNA-binding protein prediction. Among them, support vector machines has been widely used due to its efficiency and speed. Support vector machines is a machine-learning method for classification, aims to identify a rule that correctly puts each member of a training set into corresponding classes. Using the kernel function, the SVMs could resolve nonlinear problems (14). SVMs algorithm is packaged into libraries and publicly accessible, including SVM^{light} and LIBSVM that use C/C++ language, and Scikit-learn that use Python for implementation.

In this project, the tool used for implementing SVMs is Scikit-learn (16) owing to the fact that it has a rapid prototyping nature. Radial basis function (RBF) kernel is used because it fits the data better than other kernels. Using this kernel requires 2 important parameters - regularization factor C and decision boundary smoother γ (gamma). After iterating test with different value of the parameters, C=1000 and γ=0.01 is selected for the optimal prediction accuracy. Before feeding the constructed input vectors into SVMs, the data need to be preprocessed through standardization because the vector values are a mixtures of scales (i.e. pKa value, hydrophobicity index, molecular mass); otherwise, the model will be skewed and behave badly.

PREDICTION RESULTS

To evaluate the classifier performance, 5-fold cross-validation was used. Each iteration takes 4 folds of the data instances as training sets and 1 fold as test set (Supplementary Table 2). The predictions produced by the SVMs each iteration are combined and used to compute the prediction results.

Predictions made for the test data instances are compared with the target labels (1 or 0). Prediction method is evaluated with respect to accuracy, specificity, sensitivity, positive predictive value (PPV), negative predictive value (NPV), and Matthews correlation coefficient (MCC). TP is the number of true positive, TN is the number of true negative, FP is the number of false positive, FN is the number of false negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$PPV = \frac{TP}{TP + FP}$$

$$NPV = \frac{TN}{TN + FN}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

TP: 734.0 FP: 313.0 TN: 6062.0 FN: 381.0

Accuracy	90.73%
Sensitivity	65.82%
Specificity	95.09%
PPV	70.10%
NPV	94.08%
MCC	0.62

As a result of higher number of negative residues compared to positive ones, HBond-DPI produces high specificity and NPV, while sensitivity and PPV tend to shrink relatively. The MCC is moderately good (coefficient of +1 is a perfect prediction, 0 is random prediction, -1 is a totally useless prediction).

COMPARISION WITH OTHER METHODS

As discussed, HBond-DPI method slightly differs from other existing methods. Comparing it to other methods directly seems to be inaccurate. However, because it shares resemblance with BindN due to the nature of feature representation and dataset used, comparing the two might be worth examining how well HBond-DPI do with other methods.

Table 2. Comparing prediction performance of HBond-DPI and BindN.

Program	Accuracy (%)	Sensitivity(%)	Specificity (%)
HBond-DPI	90.73	65.82	95.09
BindN (DNA-binding)	70.31	69.40	70.47

LIMITATIONS AND FURTHER EXTENSIONS

The evaluation of HBond-DPI prediction exhibits a rather good result. However, there is still space for improvement. Firstly, the features used are only sequence-based. Many previous studies have utilized the structural-based feature which proved to result in better performance. Moreover, spatial neighbors of a residue can accurately represent the residue’s environment (13). Secondly, more evaluations should be put into discussion by examining different datasets to make comparison with other previous methods possible. Different datasets may also produce either higher or lower performance. Third of all, due to the excess of negative training data, it may influence the performance of machine learning methods. The ratio of positive to negative training instances should be taken into account during the preparation of machine learning experiments, as it might significantly influence the performance of particular classifier (15). Therefore, balancing positive and negative training data should be considered when using machine learning methods. Last but not least, there is a possible extension to also cover the RNA-protein prediction.

CONCLUSIONS

In this study, we have constructed HBond-DPI that aims to predict hydrogen bond DNA-protein binding residues using SVMs model. HBond-DPI takes amino acid sequences as input and produce prediction of h-bond interacted residues in DNAs and proteins as output. The prediction performance is 90.73% accuracy with sensitivity of 65.82%, specificity of 95.10%, PPV of 70.10%, NPV of 94.80%, and MCC of 0.62. The program is open source and can be found at **https://github.com/samyrsd**

SUPPLEMENTARY DATA

Supplementary Table 1. Molecular mass, hydrophobicity index, and pKa value of amino acid.

Residues	Molecular mass	Hydrophobicity	pKa
Ile	131.1736	4.5	7
Leu	131.1736	3.8	7
Lys	146.1882	-3.9	10.5
Met	149.2124	1.9	7
Phe	165.19	2.8	7
Thr	119.1197	-0.7	7
Trp	204.2262	-0.9	7
Val	117.1469	4.2	7
Arg	174.2017	-4.5	12.0
His	155.1552	-3.2	6.08
Ala	89.0935	1.8	7
Asn	132.1184	-3.5	7
Asp	133.1032	-3.5	3.9
Cys	121.159	2.5	8.28
Glu	147.1299	-3.5	4.3
Gln	146.1451	-3.5	7
Gly	75.0669	-0.4	7
Pro	115.131	-1.6	7
Ser	105.093	-0.8	7
Tyr	181.1894	-1.3	10.1

Supplementary Table 2. Each iteration index used in 5-fold cross-validation.

Training Index	Test Index
[1499 1500 1501 …, 7492 7493 7494]	[0 1 2 … 1496 1497 1498]
[0 1 2 … 7492 7493 7494]	[1499 1500 1501 … 2995 2996 2997]
[0 1 2 … 7492 7493 7494]	[2998 2999 3000 … 4494 4495 4496]
[0 1 2 … 7492 7493 7494]	[4497 4498 4499 … 5993 5994 5995]
[0 1 2 … 5993 5994 5995]	[5996 5997 5998 … 7492 7493 7494]

Supplementary Table 3. List of non-redundant sequences in the PDNA-62 dataset (Ahmad et al., 2004. Bioinformatics, 20:477-486) for prediction of DNA-binding residues.

PDB ID chain	Structure resolution (Å)	Sequence annotation
1A02_F	2.7	Human proto-oncogene protein c-fos
1A02_J	2.7	Human transcription factor AP-1 (c-jun)
1A02_N	2.7	Human T cell transcription factor NFAT1
1A74_A	2.2	Intron-encoded endonuclease I-PpoI from *Physarum polycephalum*
1AAY_A	1.6	Mouse transcription factor Zif268 (zinc finger protein)
1AZQ_A	1.94	DNA-binding protein 7d from *Sulfolobus acidocaldarius*
1B3T_A	2.2	Human herpesvirus 4 nuclear antigen EBNA1 (DNA-binding domain)
1BF5_A	2.9	Human Stat-1 (signal transducer and activator of transcription)
1BHM_A	2.2	Endonuclease BamHI from *Bacillus amyloliquefaciens*
1BL0_A	2.3	*E. coli* transcriptional activator, multiple antibiotic resistance protein
1C0W_B	3.2	Diphtheria toxin repressor from *Corynebacterium diphtheriae*
1CDW_A	1.9	Human TATA-box binding protein (TBP)
1CF7_A	2.6	Transcription factor E2F-4
1CJG_A	NMR	*E. coli* lactose operon repressor
1CMA_A	2.8	*E. coli* Met repressor (metJ)

1D02_A	1.7	Type II restriction enzyme Muni from *Mycoplasma*
1D66_A	2.7	Yeast GAL4 transcriptional activator
1DP7_P	1.5	Human MHC class II regulatory factor RFX1
1ECR_A	2.7	*E. coli* replication terminator protein
1FJL_A	2	*Drosophila* homeodomain protein paired
1GAT_A	NMR	Erythroid transcription factor GATA-1 from *Gallus gallus*
1GCC_A	NMR	Ethylene-responsive transcription factor 1A (ERF1A) from *Arabidopsis thaliana*
1GDT_A	3	*E. coli* recombinase, gamma delta resolvase
1HCQ_A	2.4	Human estrogen receptor DNA-binding domain
1HCR_A	1.8	DNA invertase hin from *Salmonella typhimurium*
1HDD_C	2.8	*Drosophila* Segmentation polarity homeobox protein engrailed
1HLO_A	2.8	Human transcription factor Max (Myc-associated factor X)
1HRY_A	NMR	Human sex-determining region Y protein (SRY)
1HWT_D	2.5	Yeast activatory protein CYP1 (HAP1)
1IF1_A	3	Interferon regulatory factor 1 from *Mus musculus*
1IGN_A	2.25	Yeast DNA-binding protein RAP1
1IHF_A	2.5	*E. coli* integration host factor (DNA-binding, bacterial histone-like)
1IHF_B	2.5	*E. coli* integration host factor (DNA-binding, bacterial histone-like)
1J59_A	2.5	*E. coli* catabolite gene activator protein (CAP)
1LMB_4	1.8	Bacteriophage lambda repressor protein CI
1MDY_A	2.8	Mouse MyoD bHLH domain
1MEY_F	2.2	Designed consensus zinc finger
1MHD_A	2.8	Human Smad3 transcriptional activator
1MNM_A	2.25	Yeast Mcm1 transcriptional regulator
1MNM_C	2.25	Yeast Mat alpha-2 transcriptional repressor
1MSE_C	NMR	Mouse Myb proto-oncogene protein
1OCT_C	3	Human Oct-1 (POU domain)
1PAR_B	2.6	Bacteriophage P22 transcriptional repressor arc
1PDN_C	2.5	Prd paired domain from *Drosophila melanogaster*
1PER_L	2.5	Bacteriophage 434 repressor
1PNR_A	2.7	*E. coli* HTH-type transcription repressor purR (purine repressor)
1PUE_E	2.1	Transcription factor Pu.1 (Ets domain) from *Mus musculus*
1PVI_B	2.8	Type II restriction enzyme PvuII from *Proteus vulgaris*
1PYI_A	3.2	Yeast pyrimidine pathway regulator 1 (PPR1)
1REP_C	2.6	*E. coli* replication initiation protein
1SRS_A	3.2	Human serum response factor (SRF)
1SVC_P	2.6	Human nuclear factor NF-kappa-B p105 subunit (NFKB1)
1TC3_C	2.45	Transposable element Tc3 transposase from *Caenorhabditis elegans*
1TF3_A	NMR	Transcription factor IIIA from *Xenopus laevis*
1TRO_A	1.9	*E. coli* Trp operon repressor
1TSR_A	2.2	Human p53 tumor suppressor
1UBD_C	2.5	Human Yy1 protein zinc finger domain
1XBR_A	2.5	Brachyury transcription factor (T domain) from *Xenopus laevis*
1YRN_A	2.5	Yeast mating-type protein A1 (MATA1)
1YRN_B	2.5	Yeast mating-type protein ALPHA2 (MATALPHA2)
1YSA_C	2.9	Yeast transcription factor GCN4
1YUI_A	NMR	Transcription factor GAGA from *Drosophila melanogaster*
2BOP_A	1.7	Regulatory protein E2 from bovine papillomavirus type 1
2DRP_D	2.8	*Drosophila* Tramtrack protein beta isoform (Fushi tarazu repressor protein)
2GLI_A	2.6	Human zinc finger protein Gli1
2HDC_A	NMR	Rat forkhead box protein D3
3CRO_L	2.5	Bacteriophage 434 Cro protein

ACKNOWLEDGEMENTS

This study would not have been possible without the guidance of Prof. Kyungsook Han.

REFERENCES

1. Stavroula AC, Diomidis GP, Kostas AP, Athanasios GP. Hydrogen bonds in protein–DNA complexes: Where geometry meets plasticity. *Biochimie*. Volume 89, Issue 11, November 2007, Pages 1291-1303, ISSN 0300-9084, http://dx.doi.org/10.1016/j.biochi.2007.07.020. (http://www.sciencedirect.com/science/article/pii/S0300908407001964)
2. Wang L, Brown SJ. BindN: A web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res*. 2006, 34, W243–W248.
3. Hwang S, Gou Z, Kuznetsov IB. DP-Bind: A web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. *Bioinformatics*. 2007, 23, 634–636.
4. Wu J, Liu H, Duan X, Ding Y, Wu H, Bai Y, Sun X. Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. *Bioinformatics* 2009, 25, 30-35.
5. Wang L, Yang MQ, Yang JY. Prediction of DNA-binding residues from protein sequence information using random forests. *BMC Genomics*. 2009, 10, doi:10.1186/1471-2164-10-S1-S1.
6. Ozbek P, Soner S, Erman B, Haliloglu T. DNABINDPROT: Fluctuation-based predictor of DNA-binding residues within a network of interacting residues. *Nucleic Acids Res*. 2010, 38, W417–W423.
7. Si J, Zhang Z, Lin B, Schroeder M, Huang B. MetaDBSite: A meta approach to improve protein DNA-binding sites prediction. BMC Syst. Biol. 2011, 5, doi:10.1186/1752-0509-5-S1-S7.
8. Chen YC, Wright JD, Lim C. DR_bind: A web server for predicting DNA-binding residues from the protein structure based on electrostatics, evolution and geometry. *Nucleic Acids Res*. 2012, 40, W249–W256.
9. Ma X, Guo J, Liu HD, Xie JM, Sun X. Sequence-based prediction of DNA-binding residues in proteins with conservation and correlation information. *IEEE/ACM Trans. Comput. Biol. Bioinform*. 2012, 9, 1766–1775.
10. Liu R, Hu J. DNABind: A hybrid algorithm for structure-based prediction of DNA-binding residues by combining machine learning- and template-based approaches. *Proteins*. 2013, 81, 1885–1899.
11. Zhang Y, Xu J, Zheng W, Zhang C, Qiu X, Chen K, Ruan J. newDNA-Prot: Prediction of DNA-binding proteins by employing support vector machine and a comprehensive sequence representation. *Comput. Biol. Chem.* 2014, 52, 51–59.
12. Ahmad S, Gromiha MM, Sarai A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*. 2004, 20, 477–486.
13. Si J, Zhao R, Wu R. An Overview of the Prediction of Protein DNA-Binding Sites. Christov C, ed. *International Journal of Molecular Sciences*. 2015;16(3):5194-5215. doi:10.3390/ijms16035194.
14. Cortes C, Vapnik V. Support-vector networks. Mach. Learn. 1995, 20, 273–297.
15. Kurczab R, Smusz S, Bojarski AJ. The influence of negative training set size on machine learning-based virtual screening. *Journal of Cheminformatics*. 2014. 6:32. doi:10.1186/1758-2946-6-32.
16. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.