Hidden Markov Model Based DNA-binding Proteins Prediction by Mining on Sequence and Structure Information

Wei-Jhih Chen, Po-Cheng Chuang and Hung-Yu Kao

Institute of Medical Informatics, National Cheng Kung University, Tainan, Taiwan, R.O.C.

E-mail: q5695109@mail.ncku.edu.tw, kailoven@yahoo.com.tw, hykao@mail.ncku.edu.tw

Abstract *In the post-genome period, the protein domain structures are published rapidly. For figuring out the cell function, the mechanism of protein-DNA interaction is an important subject in resent bioinformatics research and has not been comprehensively studied. Several machine learning based methods have been attempted to solve this issue. Until recently, few studies have been successful in translating the tertiary structure characteristics of proteins into appropriate features for utilizing the learning mechanism to predict DNA-binding proteins. In this work, a novel machine learning approach based on using HMMs (hidden Markov Models) to express the characteristics of DNA-binding proteins in the both aspects of amino acid sequence and tertiary structure are presented. Moreover, several helpful features of DNA-binding proteins are also utilized in the proposed method, such as residue composition, structure pattern composition and accessible surface area of residues.*

We develop a SVM (Support Vector Machine) based classifier to predict general DNA-binding proteins, and obtain the accuracy of 88.45% through 5-folds cross-validation. Furthermore, a response element specific classifier is constructed for predicting response element specific DNA-binding proteins, and is obtained the precision of 96.57% with recall rate as 88.83% in average.

Keywords: Machine learning, Hidden Markov Model, DNA-binding proteins, Support vector machines.

1. Introduction

The progress in genome analyses and structure genomic projects is solving the structures of protein-DNA complexes at an alarming rate. It's crucial to figure out the functions of genes and proteins. In molecular biology field, genomic processes act through the transcriptional regulations, nuclear translocations and binding on specific response elements, and then lead to the regulation of the expression of target genes. These regulation relationships control a lot of major cellular processes. As the beginning of genomic processes, many proteins which named as DNA-binding proteins might bind to specific response element sequences which locate among the promoter regions of target genes [11]. However, the mechanism of protein-DNA binding has not been completely understood yet. Several works have been published for analyzing the protein-DNA recognition mechanism [13] and several works have been devoted to predict DNA-binding proteins [4, 9].

In this work, we focus on DNA-binding proteins prediction by applying the structural and sequence information of proteins. According to previous analyses, a lot of DNA-binding proteins can be divided into several groups based on the sequence similarity, such as 146 protein families of Pfam [10] or 54 types of DNA-binding proteins based on structure similarity [12]. The DNA-binding proteins which belong to the identical groups are normally containing homologous relationships. For representing the homology information of each group of DNA-binding proteins, HMMs are the ideal probability models to describe the characteristics of each group of DNA-binding proteins. Moreover, these HMMs are able to predict un-known DNA-binding proteins based on the homologous relationships between sequences and HMMs. However, the DNA-binding proteins predicted are

certainly limited to constant types which homologous to at least one of the existing HMMs. An approach which combines HMMs with other characteristics for predicting DNA-binding proteins is presented in this work. Otherwise, to be universally known that the tertiary structures of proteins are more conserved than amino acid sequences. Consequently, the homologous information of protein structures are significantly expressed in structure based protein families. However, as our knowledge, there exist no HMMs which trained by using tertiary structures until recently. In this work, we train structure based and sequence based HMMs (SQ-HMM and ST-HMM) for expressing the characteristics of DNA-binding proteins in various aspects.

2. Method

Initially, the structure data as three dimension coordinates format are translated into the form of alphabet sequences. Then the features of alphabet sequences are gathered in the same way of generating the features of residues. The characteristics of DNA-binding proteins in the aspects of amino acid sequences and tertiary structures are expressed by using sequence-based and structure-based HMMs respectively. These sequence and structure based HMMs are employed to distinguish between DNA-binding proteins and non-DNA-binding proteins. Furthermore, several features of sequence and structure are also taken into account for training a SVM based classifier. Such as the composition of residues and structure alphabets, the accessible surface area (ASAs) information of residues in protein structures, the composition of structure and sequence patterns, and etc.

2.1. SQ-HMM and ST-HMM

The sequence-based HMMs are obtained from DBD [10]. Presently, there are 146 sequence-based HMMs existed in DBD. These HMMs are obtained via manually inspected all of the families from Pfam [3] and selected the models which confidently to represent the domains that recognize DNA sequences specifically. These 146 models are named as SQ-HMM in this paper.

Otherwise, a novel method is proposed for training structure-based HMMs in this study. Presently, there are 302 structure-based HMMs regarded to stand for the characteristic of DNA-binding domains in DBD. Even though that these 302 HMMs are constructed based on the structure classification information of SCOP [7]. Still, these HMMs are curated by applying amino acid sequences. For distinguishing the inconsistency of information between sequence-based and structure-based HMMs, a novel methods to construct structure-based HMMs is presented in this study.

In this method, SADB [19] is employed to construct HMMs through following principles: (1) Selected 302 alphabet sequences from SADB which corresponding to the 302 seed domain structures of structure-based HMMs built by DBD. (2) Divided 302 alphabet sequences into groups for which belong to the identical families in SCOP. (3) For the alphabet sequences of each group, ClustalW1.83 [15] is employed to execute the multiple sequence alignments. Due to substitution matrixes play a crucial role in affecting the performance of sequence alignments. The specific structural alphabet substitution matrix (SASM) is chosen in this step. (4) Use *hmmbuild* function of HMMER [6] to build HMMs for all of the multiple sequence alignment results. Subsequently, 89 structure-based HMMs named as ST-HMM are produced after executing this procedure.

2.2. Feature Selection

LIBSVM-2.82 [5] is employed to train SVM models in the approach of DNA-binding proteins prediction. For building an effective SVM classifier, it's crucial to construct a suitable feature vector. There are six types of features (totally 87 features: 40 for residue composition, 2 for hydrophobicity, 7 for alphabet composition, 18 for pattern, 8 for HMM, and 12 for HMM groups) employed in this study for constructing models to distinguish between DNA-binding proteins and non-DNA-binding proteins. These six types of features can also be allocated into two categories: $F_{\text{HMM-similarity}}$ and $F_{\text{Statistical}}$. The details of these features have been described in following sections. All of the features below are analyzed by using DBP-set and NDBP-set. DBP-set contains 107 DNA-binding domains and NDBP-set contains 248 non-DNA-binding domains. These domains are gathered from previous work [9] and filter out the domains which not exist in SADB.

For the features of the $F_{\text{HMM-similarity}}$ type, we use HMM and HMM groups to represent the characteristics of DNA-binding proteins. In this work, SQ-HMM and ST-HMM are employed for distinguishing between DNA-binding proteins and non-DNA-binding proteins, and the *hmmbuild* function of HMMER is employed for training HMMs. *Hmmbuild* is a function of HMMER

which can build a HMM by reading a multiple sequence alignment file. There are several parameters enable user to train proper models depend on various purposes. For using HMMs to distinguish between DNA-binding proteins and non-DNA-binding proteins, two types of HMMs are produced by choosing the option "-f" or not. If choosing the option "-f", the HMMs (HMM_ls) produced can support that searching for sequences by using local alignments algorithm. Through the HMM_ls, if a sequence contains a domain which matches only a part of the HMM_l, this domain of the sequence is able to be detected. Otherwise, when not choosing "-f" (HMM_g), the domains only can be detected if the domains are able to match the whole HMM_gs. The advantageous of HMM_l is to detect the domains which contain low homologous with HMMs, but in some cases, HMM_ls may detect several domains which could be noises. In order to discover more DNA-binding proteins with high reliability, both types of HMMs are taken into account in this step. For applying HMMs to distinguish DNA-binding proteins from non-DNA-binding proteins, *hmmpfam* function of HMMER is employed to calculate the scores to describe the similarity between sequences and HMMs.

Initially, we align both set (DBP-set and NDBP-set) of sequences to 146 models of SQ-HMM individually and then select the highest score for each sequence (*Big_Score*). The sequences of DBP-set which have *Big_Score*s superior to the lowest *Big_Score* of whole NDBP-set and the sequences of NDBP-set which have *Big_Score*s inferior to the highest *Big_Score* of whole DBP-set are defined belong to the "*difference region*" of DBP-set and NDBP-set. The number of sequences in *difference region* is bigger may imply that the quality of this type of HMMs are better. In the same way, structure alphabet sequences of DBP-set and NDBP-set are compared with 89 HMMs of ST-HMM.

As results, SQ-HMM and ST-HMM indeed distinguished between DNA-binding proteins and non-DNA-binding proteins in several cases. In the cases of low scores, HMM_ls are more sensitive than HMM_gs (Table 1).

Table 1. The coverage of the *difference region* for SQ-HMM and ST-HMM.

Difference Region	SQ-HMM_g	SQ-HMM_l	ST-HMM_g	ST-HMM_l
DBP-set	15	15	35	35
NDBP-set	1	5	1	5

Since certain DNA-binding domains have no significantly obvious similarity with SQ-HMM or ST-HMM. These ambiguous protein domains are not able to be detected by applying SQ-HMM or ST-HMM in the methods mentioned before. We assumed that the *Big_Score*s of several DNA-binding domains are not high enough to reveal the characteristics in the viewpoint of sequences and structures, but these domains may indeed similar to specific HMMs than others of SQ-HMM and ST-HMM. For detecting these ambiguous cases, the *Diff_Scores* are computed for extending the divisions between DNA-binding domains and non-DNA-binding domains.

The way to compute *Diff_Score*s is using the difference values which produced by employing the *Big_Score*s of sequences to subtract the others scores which produced by aligning the sequences to others HMMs, and then average these difference values:

$$Diff\_Score_{ij} = \sum_{i=1}^{N} \frac{Big\_Score_j - Align_{ij}}{N} \quad (1)$$

where j denoted the sequence number, i represented the HMM number, N is the quantity of HMM in each HMM-set, Big_Score_j is the *Big_Score* of sequence j, and the $Align_{ij}$ denoted the score produced by aligning the j sequence to i HMM. As shown in Table 2, *Diff_Score*s are beneficial to distinguish DBP-set from NDBP-set more effective in the viewpoint of amino acid sequences.

Table 2. The coverage of *difference region* for using *Diff_Score*.

Difference Region	SQ-HMM_g	SQ-HMM_l	ST-HMM_g	ST-HMM_l
DBP-set	26	15	35	36
NDBP-set	0	12	0	2

Due to there are several DNA-binding proteins swhich contain weak homology relationships with others DNA-binding proteins. The sequences or structures of these protein domains are probably not conserved sufficiently to express the similarity with SQ-HMM and ST-HMM by applying the alignment scores. For detecting these remote homologous relationships, the concepts of protein family grouping have been presented to use a group of similar protein families to represent a more extensive type of proteins, such as the clans of Pfam database [3]. Although that there are a lot of families grouped into clans. However, the greater part of HMMs in Pfam has never been allocated into at least one clan. Hence, a HMMs grouping algorithm is addressed in this work to allocate the HMMs with high domain similarity into a group. The traditional Expectation Maximization (EM) cluster algorithm is employed here to cluster these HMMs into several groups.

EM clustering algorithm is executed by computing the probabilities of each data (in this article, each data means every HMMs contained in SQ- and ST-HMM) on diverse probability distributions and then return the data with maximize likelihood and the corresponding clustered. Due to our demand to implement HMMs clustering is to allocate the HMMs which have least homologous distances between each other into the same groups and not need to define the number of groups. The EM algorithm is chosen here instead of others partitioned or hierarchical cluster algorithms, such as k-means, ROCK, and etc.

Initially, due to the number of sequences for training each HHM is critically inconsistent. For performing fairly HMMs comparisons to compute the similarity of each pairs of HMMs belong to SQ-HMM and ST-HMM. The *hmmemit* function of HMMER is employed to produce 10 sequences for representing each of HMM. After that, the similarity of a pair of HMMs are computed by summarizing the local alignment scores of each sequence pair produced from these two HMMs. Through this step, a 146*146 similarity matrix is produced corresponding to SQ-HMM and an 89*89 matrix for ST-HMM. The EM algorithm is executed by employing cluster package of WEKA [18] and the parameters are set as default. After executing EM algorithm, SQ-HMM are divided into 8 groups and 5 groups for ST-HMM. In SQ-HMM, there are 57 families of HMM covered by 15 clans of Pfam. For the clans which have high coverage in SQ-HMM, there are 5 of 6 HMMs which covered by “p53-like” clan are divided into identical group. Otherwise, in the aspect of ST-HMM, 5 of 6 HMMs belongs to "Winged helix" superfamily of SCOP are divided into identical groups. These cluster results are employed to detect the remote homology relationships through sequence and structure based HMMs. We totally use three types of scores of HMM groups for distinguishing between DNA-binding proteins and non-DNA-binding proteins:

$$Big\_Avg = Big\_Score - \sum_{G=1,G\neq j}^{N} \sum_{i=1}^{G_n} \frac{Align\_Score_i}{\sum_{G=1,G\neq j}^{N} G_n} \quad (2)$$

$$Big\_Big = Big\_Score - \sum_{G=1,G\neq j}^{N} \frac{Big\_Score_G}{(N-1)} \quad (3)$$

$$Avg\_Avg = \sum_{i=1}^{j_n} \frac{Align\_Score_i}{j_n} - \sum_{G=1,G\neq j}^{N} \sum_{i=1}^{G_n} \frac{Align\_Score_i}{\sum_{G=1,G\neq j}^{N} G_n} \quad (4)$$

where N denotes the number of groups, G_n is the number of HMM in group G, $Align\_Score_i$ is the hmmpfam score between each sequence and HMM , and Big_Score_G is the big score of a sequence for each HMM groups.

For the features of the $F_{\text{Statistical}}$ type, we consider residue and structure composition, pattern composition, and hydrophobicity. In the aspect of residue composition, previous research [2] has specified that the frequencies of several types of residues in DNA-binding proteins perform higher than non-DNA-binding one (e.g. Arg and Lys), but some of other types are lower. However, not all of the residues are exposed on the surface of protein structures, and the buried residues have low divergence between DNA-binding proteins and non-DNA-binding proteins. Consequently, residue composition of protein surface is probably more disposed than overall composition [4].

20 features (*Res_com(1)* to *Res_com(20)*) which represent the ratios of each type of amino acids for each instance are employed to training this classifier:

$$Res\_com(i) = \frac{R_i}{N}, 1 \le i \le 20 \quad (5)$$

where N denoted the number of residues in a protein domain sequence and R_i represented the number of each type of residues contained in a protein domain sequence.

For analyzing whether the composition of residues are different in exposed regions, DSSP program [8] are employed to calculate an ACC value for each residue of protein domains. ACC is the number of water molecules in contact with each residue. Hence, AD (accessible degree) value is produced by dividing each ACC value by the overall surface area of each amino acid defined previously [14]. After testing various AD thresholds, we find the best effect is produced when the AD threshold is equal 0.1. This phenomenon is owing to that the AD threshold as 0.1 can filter the buried residues with not to filter out residues too much. 20 features represent the ratios of each type of amino acids with AD value>0.1 are employed to train this classifier, too.

Otherwise, the structure composition is also considered in this work. As defined in 3D-Blast [19], all of five-residue long continuous protein tertiary structures have been divided into 23 groups of structure types and each group of structures is symbolized by an alphabet. After translating the protein tertiary structures into alphabet sequence formats, we can use the method of calculating residue composition to produce the features of protein structure composition. Result

shows that the difference of composition ratio is most significant in alphabet A, B, C, D, E, and F. These six kinds of alphabets are also the representative alphabet types of helix and strand. In structure composition, we only consider these six types of structures.

In several protein-DNA binding sites, the binding activity of a residue might be affected by adjacent residues [2]. The unity of the neighborhood of each type of residues is probably significant. In order to utilize the actually exists phenomenon to distinguish between DNA-binding proteins and non-DNA-binding proteins. Each two adjacent residues or alphabets are defined as a pattern, and each pattern is composed by a central residue C_i (or alphabet) and an adjacent residue A_i (or alphabet). Base on this definition, there are 400 kinds of residue patterns (20 types of central residues multiplied by 20 types of adjacent residues), and 529 kinds of patterns for alphabets (23 types of central residues multiplied by 23 types of adjacent alphabets). The *Pattern_com(i)* is calculated by using the appearance frequency of a residue type C_i to divide the frequency of P_i:

$$\text{Pattern}\_com(i) = \frac{P_i}{C_i}, 1 \le i \le 400(529) \quad (6)$$

The types of *Pattern_com(i)* with the difference ratios larger than 7% between DBP and NDBP-set are chosen for being employed as features for training SVM models. Through these criteria, 10 residue patterns and 8 alphabet patterns are picked.

For using the characteristics of hydrophobicity, previous study has observed that the hydrophobicity of several residues are critical for DNA binding intensity [16]. By applying the properties of hydrophobicity, amino acids are classified into three groups: Arg, Lys, Glu, Asp, Gln, and Asn are allocated to polar; Gly, Ala, Ser, Thr, Pro, His, and Tyr are neutral; and Cys, Val, Leu, Ile, Met, Phe, and Trp are hydrophobic. The distribution of these three types of residues is advantageous in separating the DNA-binding proteins from non-DNA-binding proteins [20]. Due to the composition of hydrophobic type residues is highly less and not significantly difference between DNA-binding proteins and non-DNA-binding proteins. In this type of features, only the compositions of polar and neutral are taken into account.

3. Experiments

The effect of each type of feature is tested by applying DBP-set and NDBP-set for training SVM models (totally 355 instances) and verified through 5-folds cross validation. Six types of DNA sequence specific DNA-binding proteins (SPDBP-set) are employed for verifying the performance of response element specific DNA-binding proteins prediction. These DNA-binding proteins are gathered by submitting the DNA sequences of antioxidant response element (ARE), glucocorticoid response element (GRE), cAMP-responsive element (CRE), p53 responsive element (p53RE), peroxisome proliferator response element (PPRE), and estrogen response element (ERE) to search TRANSFAC [17] by Patch function [1].

For predicting general DNA-binding proteins, the classifier trained in this work can achieve the sensitivity of 74.78% and the specificity of 91.13%. In overall accuracy, we reach to 88.45% which is better than the 86.62% of PSSM when using the same training and testing data. Moreover, Figure 1 shows that the performance of SVM models is most effective while combining HMM based features with amino acid and structure alphabet based features.

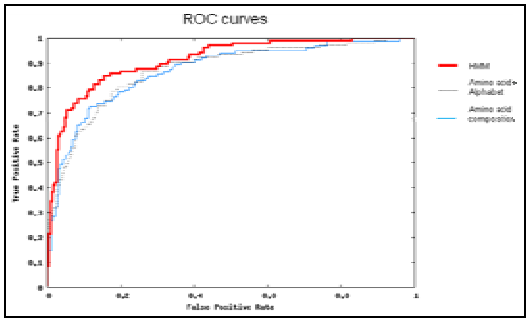

Figure 1. The ROC curves of the classifiers.

Table 3. The performances of predicting the response element specific DNA-binding proteins.

	ARE	*CRE*	*GRE*	*P53RE*	*RpRE*	*ERE*
Presicion	100	100	100	88.9	90.5	100
Recall	92.3	85.7	80	80	95	100
F-Measure	96	92.3	88.9	84.2	92.7	100

4. Conclusion and Future Work

In this work, a high precision DNA-binding proteins classifier is constructed for predicting general DNA-binding proteins by using features of $F_{\text{HMM-similarity}}$ and $F_{\text{Statistical}}$. Through integrating the sequence and structure information of proteins, several DNA-binding proteins less conserved in structure or sequence are also identified by our proposed classifier. The classifiers for specific

DNA-binding proteins is able to obtain the high precision and recall rates. These classifiers trained from specific DNA-binding proteins can help biology researchers to figure out un-known biology processes.

After predicting un-known DNA-binding proteins, the SQ-HMM and ST-HMM are able to analyze the sites of residues which could bind to DNA directly. Moreover, the response element specific classifiers could be trained by using the verified response element specific DNA-binding proteins to predict un-known DNA-binding proteins for each response element from the databases (e.g. TRANSFAC) automatically. Furthermore, a DNA-binding proteins prediction database could be constructed based on integrating existing databases with our response element specific classifiers.

References

[1] http://www.gene-regulation.com/.

[2] Ahmad, S., M.M. Gromiha, and A. Sarai, "Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information," *Bioinformatics*, vol. 20, pp. 477-86,2004.

[3] Bateman, A., et al., "The Pfam protein families database," *Nucleic Acids Res*, vol. 30, pp. 276-80,2002.

[4] Bhardwaj, N., et al., "Kernel-based machine learning protocol for predicting DNA-binding proteins," *Nucleic Acids Res*, vol. 33, pp. 6486-93,2005.

[5] Chang, C.C. and C.J. Lin, "LIBSVM: a library for support vector machines," Vol, 80: 604–611, 2001.

[6] Eddy, S.R., "Profile hidden Markov models," *Bioinformatics*, vol. 14, pp. 755-63,1998.

[7] Hubbard, T.J., et al., "SCOP: a structural classification of proteins database," *Nucleic Acids Res*, vol. 25, pp. 236-9,1997.

[8] Kabsch, W. and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, pp. 2577-637,1983.

[9] Kumar, M., M.M. Gromiha, and G.P. Raghava, "Identification of DNA-binding proteins using support vector machines and evolutionary profiles," *BMC Bioinformatics*, vol. 8, pp. 463,2007.

[10] Kummerfeld, S.K. and S.A. Teichmann, "DBD: a transcription factor prediction database," *Nucleic Acids Res*, vol. 34, pp. D74-81,2006.

[11] Latchman, D.S., "Transcription factors: an overview," *Int J Biochem Cell Biol*, vol. 29, pp. 1305-12,1997.

[12] Luscombe, N.M., et al., "An overview of the structures of protein-DNA complexes," *Genome Biol*, vol. 1, pp. REVIEWS001,2000.

[13] Paillard, G. and R. Lavery, "Analyzing protein-DNA recognition mechanisms," *Structure*, vol. 12, pp. 113-22,2004.

[14] Samanta, U., R.P. Bahadur, and P. Chakrabarti, "Quantifying the accessible surface area of protein residues in their local environment," *Protein Eng*, vol. 15, pp. 659-67,2002.

[15] Thompson, J.D., D.G. Higgins, and T.J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Res*, vol. 22, pp. 4673-80,1994.

[16] West, M., et al., "Functional mapping of the DNA binding domain of bovine papillomavirus E1 protein," *J Virol*, vol. 75, pp. 11948-60,2001.

[17] Wingender, E., et al., "TRANSFAC: a database on transcription factors and their DNA binding sites," *Nucleic Acids Res*, vol. 24, pp. 238-41,1996.

[18] Witten, I.H., et al., "Weka: Practical Machine Learning Tools and Techniques with Java Implementations," *ICONIP/ANZIIS/ANNES*, vol. 99, pp. 192–196,1999.

[19] Yang, J.M. and C.H. Tung, "Protein structure database search and evolutionary classification," *Nucleic Acids Res*, vol. 34, pp. 3646-59, 2006.

[20] Yu, X., et al., "Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines," *J Theor Biol*, vol. 240, pp. 175-84, 2006.