Acquisition of Rule-based Knowledge for Analyzing DNA-binding Sites in Proteins

Shinn-Jang Ho, Depart. of Automation Engineering, National Univ. Formosa University, Yunlin 632, Taiwan

Chia-Yun Chang, Institute of Bioinformatics, National Chiao Tung Univ. Hsinchu 300, Taiwan

Liang-Tsung Huang & Shiow-Fen Hwang, Depart. of Information Eng. and Computer Science, Feng Chia University, Taichung 407, Taiwan

Shinn-Ying Ho, Institute of Bioinformatics, National Chiao Tung Hsinchu 300, Taiwan

syho@mail.nctu.edu.tw

ABSTRACT

This study aims to analyze DNA-binding proteins via acquisition of interpretable knowledge which can accurately predict binding sites in proteins to understand DNA-protein recognition mechanism. For mining accurate and interpretable knowledge, a large-scale dataset consisting of 982 DNA-binding proteins is constructed. This study investigates a novel feature set consisting of 11 features, including solvent accessibility, secondary structure, charge information near the residue, amino acid group and neighbor property. The derived binding and non-binding rules reveal that besides the well-known solvent accessibility, the electric charge distribution near the residue and the amino acid groups also play important roles in prediction of binding sites. The interpretable and accurate knowledge is helpful for biologist to analyze DNA-binding proteins.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences – *biology and genetics.*

General Terms

Algorithms, Performance, Design, Verification.

Keywords

Knowledge acquisition, Binding site, Protein, Decision tree.

1. INTRODUCTION

DNA-binding proteins usually affect the regulation and expression of DNA in organisms. It is mainly controlled via binding of transcription factors to DNA for promoting or repressing gene expression levels. X-ray crystallographic and NMR spectroscopic analysis on DNA-binding proteins have provided valuable information about the general features of these complexes. However, the large diversity of amino acid and nucleotides complement combinations makes the recognition of DNA-binding residues obscure to decipher [1]. These researches reveal that the DNA-protein recognition mechanism is complicated and there is no simple rule for this recognition problem [1, 2].

Recently, various methods have been proposed to identify valuable features of DNA-binding proteins. Kono and Sarai [3] presented a structure-based method for prediction of DNA target sites by regulatory proteins. Pabo and Nekludova [4] developed geometrical models for characterizing protein-DNA interfaces. Selvaraj *et al*. [5] analyzed symmetric/ asymmetric and cognate/non-cognate binding by specificity of protein-DNA recognition. Luscombe and Thornton [6] investigated protein-DNA interactions based on amino acid conservation and the effects of mutations on binding specificity. Shanahan *et al*. [7] identified DNA-binding proteins using structural motifs and electrostatic potential. Ahmad *et al*. [8] analyzed and predicted DNA-binding protein based on composition, sequence and structural information using a neural network (NN) method. When the features evolutionary information of amino acid sequences in terms of their position specific scoring matrices (PSSMs) are used, the NN-based [9] and support vector machine (SVM) based [10] methods can enhance the net prediction (*NP*, an average of sensitivity and specificity) accuracy on the training dataset PDNA-62.

The methods [8-10] can fairly analyze and predict DNA binding sites in proteins, but suffer from obtaining human-interpretable knowledge. This study aims to analyze DNA-binding proteins from a large-scale dataset via acquisition of interpretable knowledge which can accurately predict binding sites in proteins to understand the DNA-protein recognition mechanism. The proposed method uses a data mining approach based on a decision tree method. To obtain creditable knowledge, we construct a large-scale dataset with 982 DNA-binding proteins from the Protein Data Bank (PDB) database, named PDNA-982, and present an informative feature set consisting of 11 features, including solvent accessibility, secondary structure, charge information near the residue, amino acid group and neighbor property. Note that the NN and SVM-based models are not interpretable. The derived binding and non-binding rules are helpful for biologist to analyze DNA-binding proteins.

2. MATERIALS AND METHODS

2.1 Datasets

For comparison, the same dataset PDNA-62 of protein-DNA complexes from PDB containing 62 proteins in previous studies [5, 8-10] is used to predict DNA-binding sites in proteins. The amino acid is defined as a binding residue if its side chain or backbone atoms fell within a cut-off distance 3.5 Å, which is the same as previous study from any atom in DNA sequences. Otherwise, the amino acid is a non-binding residue. This calculation result of DNA-Protein binding positions is highly consistent with that of the PDBsum database. This dataset consists

Conference: Infoscale 2007, June 6-8, 2007, Suzhou, China
Copyright number (LaTeX \crdata{}): 978-1-59593-757-5

of 7967 non-binding and 1792 binding residues and the protein structure resolution is 2.5 Å or better.

To obtain creditable knowledge, we additionally constructed a large-scale dataset with 982 DNA-binding proteins (PDNA-982). These data are acquired from the advanced search in the PDB database (The website is http://www.rcsb.org/pdb/advSearch.do). These items, "Contains Protein" and "Contains DNA", in "molecule or chain type" are chosen because we would like to get more data probably. And then, the fragments, the numbers of the residues being less than 10, are removed from the dataset. Finally, we obtained 982 DNA-binding proteins. Note that the identity in PDNA-62 is limited as identity < 25%. For increasing the number of proteins and faithfully responding the proportion of proteins in the whole dataset to mine representative knowledge, the protein identity in PDNA-982 is not limited.

2.2 Feature sets

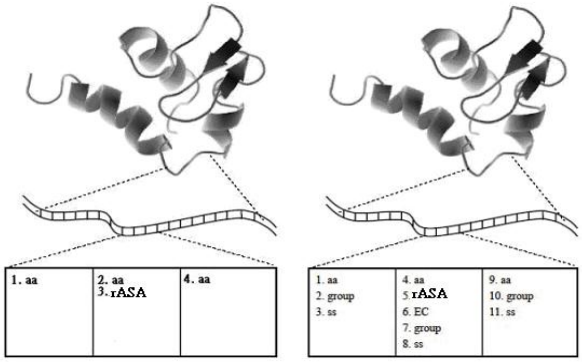

Figure 1. The used feature sets obtained from a protein structure. The same feature set with Ahmad et al. [8] is shown in the left part. The proposed feature set is shown in the right part. The symbol "aa" means the amino acid and "ss" indicates its secondary structure.

We present an informative feature set consisting of 11 features for each central residue of the protein segment with a window size 3, including solvent accessibility, secondary structure and information of the three residues.

1) Relative solvent accessibility or accessible surface area (*rASA*). The *ASA* values of these protein-DNA complexes are obtained by using DSSP program [11]. Absolute values of *ASA* are normalized to relative values as described in the work [12].

2) Electric charge distribution near the residue (*EC*). Based on the pI values of twenty amino acids, we get reference of the electric charge of each residue [13]. The electric charge is to subtract 5.0 from its pI value (e.g. the electric value of Glycine is 0.97). This shift process would result in that the residues with negative charge have negative values and those with natural charge have positive values. And the electric charge distribution near the residue (*EC*) is defined as

$$EC_i = \sum_{i-1,i,i+1} (pI_i - 5) \times (ASA_i / 10) \qquad (1)$$

where *i* represents a certain residue. And in sequence information, *i*-1 and *i*+1 show the closest left and right positions to the *i*-th residue, respectively. The equation through accumulation estimates the charge near the *i*-th residue.

The *EC* value has the opportunity of being shown obviously if the *i* residue lies near the surface of the protein. The larger *ASA* value means a larger exposed surface. And in case of being buried completely, *EC* value is near to zero. Note that the *ASA* values instead of *rASA* for the calculation of *EC* values are obtained by the DSSP program directly, because we want to response that the absolute larger surfaces of the residues make *EC* values affected more deeply and the relative larger surface in smaller residues will not have a great influence for *EC* values.

3) The amino acid group (*group*). The twenty amino acids are classified into five groups [13], listed in Table 1.

4) Secondary structure (*ss*): We get the secondary structures of the residues from the DSSP file.

5) Amino acid of residue (*aa*): The feature is acquainted by each residue in these protein sequences.

Table 1. The five groups of twenty amino acids.

Group	Amino Acid
Nonpolar, aliphatic R groups	G, A, P, V, L, I, M
Aromatic R groups	F, Y, W
Polar, uncharged R groups	S, T, C, N, Q
Positively charged R groups	K, H, R
Negatively charged R groups	D, E

2.3 Decision Tree based Method

Decision tree [14] is a popular machine learning method to classify the value of a discrete dependent variable with a finite set. Decision tree learning is a method for approximating discrete-valued target functions, in which the learned function is represented by a decision tree. Learned tree models can also be re-represented as sets of if-then rules to improve human readability [15]. A decision tree is constructed by looking for regularities in data. According to entropy calculation, we can select one with the minimum entropy from these features. The distributions of the levels of the tree are important and readable because we could analyze which feature is more significant than others.

2.3.1 The judgment for attributes of features

To select features which have more attributes for classifying samples is critical. What is a good quantitative measure of the worth of an attribute? We would define a statistic property that measures how well a given attribute separates training samples according to their target classification. The best feature choice to build trees generally leads to simple decision at the nodes. A variety of selection attributes measures have been proposed in past researches. The measure is simply the expected reduction in entropy caused by partitioning the samples according to this attribute.

2.3.2 Measure Criteria

In this work, we consider four criteria, *Sensitivity*, *Specificity*, net prediction (*NP*, an average of *Sensitivity* and *Specificity*) and *Accuracy* to evaluate the prediction performance using three-fold cross validation (3-CV). *Sensitivity* is the percentage of correctly predicted binding residues to total binding residues. *Specificity* is

the percentage of correctly predicted non-binding residues to total non-binding residues. *Accuracy* is the percentage of correctly predicted residues to total residues, which is not the first evaluation criterion here. In this study, *NP* is the first evaluation criterion considering the unbalanced distribution of binding and non-binding residues. Although total accuracy is commonly judged for the results of predictions, *NP* value is considered in the discussion of the result because binding and non-binding data sets are unbalanced in this work.

2.3.3 Parameters Setting

Overfitting is a practical difficulty for the bulk of machine learning methods. There are two major approaches to avoiding overfitting in decision tree methods. These approaches are to stop growing the tree earlier and post prune. Pruning a decision node consists of removing the subtree rooted at that node, making it a leaf node, and assigning it the most common classification of the training samples affiliated that node. In this study, the post pruning method with a pruning parameter, confidence factor (*cf*), is utilized.

Considering the unbalanced distribution of samples, the penalty is considered to avoid that accuracy of binding ones is sacrificed. The parameter setting is to increase the weight of binding influence for the classification results. Furthermore, the idea of adaptive boosting algorithm generating several decision trees is used in our method. Boosting is a technique for generating and combining multiple classifiers to improve predictive accuracy. When an unknown sample is to be classified, each decision tree votes for its predicted class and the votes are counted to determine the final class. In general, to predict the unknown data by using more decision trees will get a better accuracy than by only using one in this research.

2.3.4 Experimental processes

First, the same features and dataset PDNA-62 in the work [8] are applied. These features are the residue, its relative ASA and its two nearest neighbors in sequence composition. The models would consider these parameters, inclusive of boosting, *cf* values (the extent of the pruning) and diverse weight. Under the same condition, the proposed method is compared with the existing NN methods.

Beside PDNA-62, the dataset PDNA-982 with 11 features (shown in Fig. 1) is also applied for acquiring interpretable knowledge for analyzing DNA-binding sites in proteins. The extracted knowledge of the change using a large dataset with more features is more reliable.

3. RESULTS

3.1 Performance Evaluation

To evaluate the performance of the proposed decision tree based method, one existing NN-based method is conveniently compared using the same dataset PDNA-62. From the simulation results, we choose *cf* = 20 to do our later research. Decision trees are built by the same features with the research of Ahmad *et al.* [8]. The performance of our results (*NP*=65.6%) is better than the referenced method (*NP*=61.1%) [8].

According to the simulation performances, we suggest the classification problem should utilize the weight values between 2 and 4, and *cf* =20. Using the proposed 11 features, the rule-based system can enhance the prediction accuracy from *NP*=65.6% to 72.9%, which is better than the NN method with the PSSM features (*NP*=66.7%).

3.2 Knowledge Acquisition

Since the proposed feature set performs well, we would apply it to the large-scale dataset PDNA-982 to mine accurate and interpretable knowledge from DNA-binding proteins for biologist.

3.2.1 Feature Ranking

The features near the root of the decision tree have high ranks and are more informative. To further understand importance of the features, Table 2 shows the best features and their corresponding values with high gain ratio at each split in the top 3 tree levels.

Gain ratio, the ratio of information gain to potential information, is adopted by the decision tree system. For every tree split, this criterion will select an attribute with the highest gain ratio. The chosen attributes imply that they own the maximum distinct ability for each split. Features adopted from the dataset could be ranked by the contribution to predict. From Table 2, we could acquaint rASA and the new feature, e.g. *EC* and its group of amino acids, which really assist in classifying these data.

Table 2. The ranked importance of proposed features.

Tree level	Attributes	Value	Potential information	Information gain	Gain ratio
1	rASA	7.12251	0.857	0.026	0.030
2	rASA	1.27065	0.992	0.010	0.010
2	EC	1.866	0.797	0.014	0.018
3	Group	---------	1.264	0.001	0.001
3	rASA	2.897	0.975	0.003	0.003
3	EC	-10.891	0.989	0.004	0.004
3	EC	90.847	0.529	0.007	0.013

3.2.2 Rules of Mining

Once a decision tree model has been established, it is a simple and straightforward matter to convert it into an equivalent set of rules by traversing any given path from the root to any leaf. To discover binding and non-binding rules, it would supply the readability and understanding of data for humans.

Tables 3 and 4 represent the important rules for non-binding and binding proteins, respectively. The “Size” in these tables means the length of antecedent sentence. And “Support” of one decision rule refers to the proportion of records in the data set that conform to the rule. For example, there are 162490 samples support the non-binding rule 1 of size 1: if EC <= 5.217 and this rule has an accuracy of 96.5% in predicting non-binding sites.

The tendencies about non-binding and binding rules reveal that the major symbols after rASA or EC are “<=” and “>” in Tables 3 and 4, respectively. These rules support the fundamental biological knowledge, which mean the residues near the surfaces of the proteins with positive charge have a large opportunity of binding to DNA. These rules also verify the ranking importance of the attributes in Table 2 that rASA and EC are significant factor in the decision tree for classification. In Table 3, besides

the node 2 of the rule 4, most rules fit the above-mentioned common sense in biochemistry. The rule 4 in Table 3 means the residues do not prefer binding to DNA, even these neighbors of the residues having positively charged R groups because of the smaller surfaces exposed to solvent near the residues.

In Table 4, we find the aromatic R groups might raise the opportunity of binding. First, the aromatic R groups may exit some electronic effects. These effects might cause more attraction with DNA. Another reason is that these residues with these groups would be conserved or they are the sections of the domains in DNA-binding proteins. The inference about these reasons will need to do more experiments in biochemistry.

Table 3. The non-binding rules for DNA-binding proteins:

Decision rules	Size	Accuracy	Support
if EC <= 5.217	1	96.5%	162490
if rASA <= 1.271	1	99.9%	66836
if group = Nonpolar, aliphatic R groups and rASA <= 7.835 and nbr1_ss = H	3	99.8%	40149
if rASA <= 6.554 and nbr2_group = Positively charged R groups	2	99.3%	19718
if aa = L and rASA <= 47.782 and nbr1_ss = H	3	99.7%	19601

Table 4. The binding rules for DNA-binding proteins.

Decision rules	Size	Accuracy	Support
if ss = * and group = Aromatic R groups and rASA > 33.772 and nbr1_aa = R	4	83.09%	207
if group = Aromatic R groups and rASA > 37.422 and EC > 105.012	3	76.40%	267
if aa = I and ss = * and rASA > 1.271 and rASA <= 6.554 and EC > 5.252 and EC <= 12.6 and nbr2_group = Nonpolar, aliphatic R groups	7	73.56%	208
if ss = * and EC > 142.1 and nbr2_group = Polar, uncharged R groups	3	71.84%	309
if rASA > 7.123 and rASA <= 43.987 and nbr1_aa = G and nbr2_aa = S and nbr2_ss = H	5	64.19%	229

4. CONCLUSIONS

The DNA-binding protein prediction is an essential problem for studying gene regulation. Consequently, it concerns with metabolism and disease occurring in organisms indirectly. In this study, an interpretable decision tree based system and associated informative feature set are proposed. From interpreting the constructed decision tree model, we can extract some binding and non-binding rules with high prediction accuracy and support. The extracted rules can reveal what the significant features are for prediction of binding sites and realize the relationship between DNA and protein structures. Some significant factors in the decision tree model are relative accessible surface area (*rASA*), the electric charge distribution near the residue (*EC*), the amino acid group and neighbor information. For example, two non-binding rules are as follows: 1) if *EC* <= 5.217 then the residue is non-binding, where the rule has accuracy 96.5% and support 162490; and 2) if *rASA* <= 1.2171 then the residue is non-binding, where the rule has accuracy 99.9% and support 66836. The derived rule-based knowledge from a large-scale dataset can assist biologist to realize the DNA-binding recognition mechanism.

5. REFERENCES

[1] Sarai, A. and Kono, H. Protein-DNA Recognition Patterns and Predictions., *Annual Review of Biophysics and Bimolecular Structure*, 34, 379-398, 2005.

[2] O'Flanagan, R.A., Paillard, G., Lavery, R. and Sengupta, A.M. Non-additivity in protein-DNA-binding., *Bioinformatics*, 21, 2254-2263, 2005.

[3] Kono, H. and Sarai, A. Structure-based prediction of DNA target sites by regulatory proteins, *Proteins-Structure Function and Genetics*, 35, 114-131, 1999.

[4] Pabo, C.O. and Nekludova, L. Geometric analysis and comparison of protein-DNA interfaces: Why is there no simple code for recognition?, *Journal of Molecular Biology*, 301, 597-624, 2000.

[5] Selvaraj, S., Kono, H. and Sarai, A. Specificity of protein-DNA recognition revealed by structure-based potentials: Symmetric/asymmetric and cognate/non-cognate binding, *Journal of Molecular Biology*, 322, 907-915, 2002.

[6] Luscombe, N.M. and Thornton, J.M. Protein-DNA interactions: Amino acid conservation and the effects of mutations on binding specificity, *Journal of Molecular Biology*, **320**, 991-1009, 2002.

[7] Shanahan, H.P., Garcia, M.A., Jones, S. and Thornton, J.M. Identifying DNA-binding proteins using structural motifs and the electrostatic potential, *Nucleic Acids Res.*, 32, 4732-4741, 2004.

[8] Ahmad, S., Gromiha, M.M. and Sarai, A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information, *Bioinformatics*, 20, 477-486, 2004.

[9] Ahmad, S. and Sarai, A. PSSM-based prediction of DNA binding sites in proteins, *BMC Bioinformatics*, 6, 2005.

[10] S.-Y. Ho, F.-C. Yu, C.-Y. Chang and H.-L. Huang, "Design of Accurate Predictors for DNA-binding Sites in Proteins Using Hybrid SVM-PSSM Method," accepted by Biosystems, 2007.

[11] Kabsch, W. and Sander, C. Dictionary of protein secondary structure, Biopolymers, 22, 2577-2637, 1983.

[12] Ahmad, S. and Gromiha, M.M. NETASA: neural network based prediction of solvent accessibility, Bioinformatics, 18, 819-824, 2002.

[13] Nelson, D.L. and M.M., C. Lehninger Principles of Biochemistry. Worth Publisher, New York, 2004.

[14] Quinlan, J.R. Induction of decision trees, Machine Learning, 1, 81-106, 1986.

[15] Kohavi, R. and Quinlan, J.R. Decision-tree discovery.

Handbook of data mining and knowledge discovery. Oxford University Press, New York, 2002.