Prediction of DNA-Binding Proteins from Structural Features

Andrea Szabóová[1], Ondřej Kuželka[1], Filip Železný[1], and Jakub Tolar[2]

[1] Czech Technical University, Prague, Czech Republic
[2] University of Minnesota, Minneapolis, USA

Abstract. We use logic-based machine learning to distinguish DNA-binding proteins from non-binding proteins. We combine previously suggested coarse-grained features (such as the dipole moment) with automatically constructed structural (spatial) features. Prediction based only on structural features already improves on the state-of-the-art predictive accuracies achieved in previous work with coarse-grained features. Accuracies are further improved when the combination of both feature categories is used. An important factor contributing to accurate prediction is that structural features are not Boolean but rather interpreted by counting the number of their occurences in a learning example.

1 Introduction

The process of protein-DNA interaction has been an important subject of recent bioinformatics research, however, it has not been completely understood yet. DNA-binding proteins have a vital role in the biological processing of genetic information like DNA transcription, replication, maintenance and the regulation of gene expression. Several computational approaches have recently been proposed for the prediction of DNA-binding function from protein structure.

Stawiski et al. investigated positively charged patches on the surface of DNA-binding proteins. They used a neural network with 12 features like patch size, hydrogen-bonding potential, the fraction of evolutionarily conserved positively charged residues and other properties of the protein [1]. Ahmad and Sarai trained a neural network based on the net charge and the electric dipole and quadrupole moments of the protein [2]. Bhardwaj et al. examined the sizes of positively charged patches on the surface of DNA-binding proteins. They trained a support vector machine classifier using the protein's overall charge and its overall and surface amino acid composition [3]. Szilágyi and Skolnick created a logistic regression classifier based on the amino acid composition, the asymmetry of the spatial distribution of specific residues and the dipole moment of the protein [4].

In the present work, we combine two categories of features to predict the DNA-binding function of proteins. The first category contains the above mentioned *coarse-grained features* which enabled [4] to achieve state-of-the-art predictive accuracies. The second category contains *structural* features representing characteristic spatial patterns in the unbound conformations of the protein residues. These features are formally described in first-order logic [5] and automatically discovered by our algorithm [7].

Nassif et al. [6] have previously used a first-order logic based approach in a similar context, in particular to classify hexose-binding proteins. The main differences of our approach from [6] are as follows. First, our fast feature-construction algorithm [7] enables us to produce features by inspecting much larger structures (up to tens of thousands of entries in a learning example) than those considered in [6] using the standard learning system Aleph. Second, our structural features acquire values equal to the number of occurrences of the corresponding spatial patterns, whereas [6] only distinguished the presence of a pattern in a learning example from its absence. Our results indicate that occurrence-counting indeed substantially lifts predictive accuracy. Third, rather than proposing an alternative classification method to state-of-the-art approaches, we elaborate its *augmentation* by the use of the structural features. Lastly, the approach of [6] resulted in classifiers that are more easily interpretable than state-of-the-art classifiers and comparable in predictive accuracy. Here we maintain the interpretability advantage but actually improve on the state-of-the-art predictive accuracies both by a purely structural approach (without the coarse-grained features) and even more so through the combination of structural and coarse-grained features.

2 Materials and Methods

Data. Both the protein and the DNA can alter their conformation during the process of binding. This conformational change can involve small changes in side-chain location, and also local refolding, in case of the proteins. Predicting DNA-binding propensity from a structural model of a protein makes sense if the available structure is not a protein-DNA complex, i.e. it does not contain a bound nucleic acid molecule. We decided to work with a positive data set (UD54) of 54 protein sequences in unbound conformation obtained from [4]. As a negative data set (NB110) we used a set of 110 non-DNA-binding proteins created by [2]. From the structural description of each protein we extracted the list of all contained residues with information on their type and the list of pairwise spatial distances among all residues. As for the coarse-grained features, we followed [4] and extracted features indicating the respective proportions of the Arg, Lys, Asp, Ala and Gly residues, the spatial asymmetry of Arg, Gly, Asn and Ser, and the dipole moment of the protein.

Method. We experimented with 7 state-of-the-art attribute-value classifier types listed in Table 1. The attributes correspond to the coarse-grained features as listed above and to the structural features constructed as follows. The feature construction method assumes that proteins are described by means of formal-logic assertions. For example, the assertion `res('1AJY', r1, 'CYS')` denotes that the protein 1AJY contains a residue r1, which is a cysteine. Similarly, the assertion `dist(r1,r2,10)` denotes that the distance between residues r1 and r2 is (*approximately*) 10 angstroms. A complete description of a protein is a logical conjunction of such statements, pertaining to all involved residues, and their all pairwise spatial distances that do not exceed 40 Angstroms (computed from coordinates of *alpha carbons*). The full description of a real protein corresponds to a conjunction containing up to tens of thousands of literals.

A *feature* F is a conjunction of first order literals. For a protein p and a feature F we define the *value* of feature F to be the number of groundings θ such that $p \models F\theta$. In other words, the *value* of a feature is the number of possible ways to match the feature against a given protein. For example, a feature $F = \mathsf{res(P,\ R,\ 'CYS')}$ counts the number of cysteines in a protein P. An example of a more complicated feature is the following feature

$$F = \mathsf{res(P,R1,'CYS'),\ res(P,R2,'HIS'),\ dist(R1,R2,8)}$$

which counts the number of pairs cystein-histidine, which are 8 angstroms apart from each other. Once we have a sufficiently rich set of features, we may feed the features into any attribute-value learning algorithm.A detailed description of the computational procedures used to accomplish the feature construction task is beyond the scope of this paper. In brief, we rely on the framework of inductive logic programming [5]. In particular, we employ our recently published algorithm [7] since it can scale to rather large structures corresponding to proteins, which would be prohibitively large for mainstream inductive logic programming algorithms. This feature construction algorithm exhaustively constructs a set of features which are not *redundant*, comply with a user-defined language bias and have frequency higher than a given threshold.

3 Results

As a result of structural pattern searching we obtained about 1500 patterns present in 54 unbounded DNA-binding proteins. We made two sets of trainings (accuracies are shown in Tab. 1): i) considering just the occurrence of the structural patterns - columns marked with (NC), ii) considering also the number of the occurrence of each pattern - columns marked with (C). We compare classifiers based on our structural patterns (F2) with classifiers based on 10 features (F1) from Szilágyi et al. [4]. We also trained classifiers based on both our features and features from Szilágyi et al. (F1+2). As we can see, we get better results for classifiers considering the number of the occurrence of each pattern. For the most classifiers the accuracy is higher when they are based on our features than on features of Szilágyi et al. However, we get the best results with combination of the two feature-sets. We show here three examples of unbounded DNA-binding proteins with the residues of the pattern which is the most informative according to the χ^2 criterion (Fig. 1).

4 Conclusion and Future Work

We have improved on the state-of-the-art accuracies in predicting DNA-binding proteins by combining previously used coarse-grained features with logic-based

Classifier	**F1**	**F2(NC)**	**F1+2(NC)**	**F2(C)**	**F1+2(C)**
Linear SVM	84.0 (2)	77.5 (5)	78.1 (4)	83.0 (3)	**84.2 (1)**
SVM with RBK	81.6 (3)	67.1 (4-5)	67.1 (4-5)	83.0 (2)	**85.4 (1)**
Simple log. regr.	81.6 (3)	73.9 (5)	78.8 (4)	**87.6 (1)**	82.3 (2)
L_2-regularized log. regr.	84.0 (2)	78.7 (5)	80.5 (4)	82.4 (3)	**84.2 (1)**
Ada-boost	77.4 (4)	73.2 (5)	83.0 (2)	79.3 (3)	**84.7 (1)**
Random forest	78.6 (4)	76.8 (5)	**83.6 (1)**	80.5 (2)	79.9 (3)
J48 decision tree	75.0 (3)	70.7 (4)	75.6 (2)	68.1 (5)	**76.2 (1)**
Average ranking:	3	4.79	3.07	2.71	**1.43**

Table 1. Accuracies obtained by stratified 10-fold crossvalidation using features of Szilágyi et al. (F1), our structural pattern features (F2) and combination of both of them (F1+2). The numbers in parentheses correspond to ranking w.r.t. the obtained accuracies.

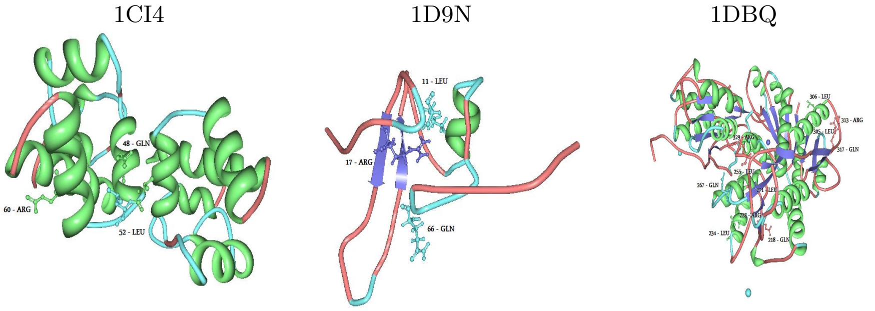

Fig. 1. Example proteins containing one discovered pattern shown using the protein viewer software [8]. Residues assumed by the pattern are indicated.

spatial protein features. It turns out that an important factor contributing to the high predictive accuracies is that the latter features are not Boolean but rather are assigned values counting the occurrences of the corresponding spatial pattern in the example protein. We are currently trying to further improve the predictions by incorporating further background knowledge.

Acknowledgement: This work was supported by project ME10047 granted by the Czech Ministry of Education and the Czech Technical University internal grant #10-801940.

References

1. Stawiski, Gregoret, and Mandel-Gutfreund. Annotating nucleic acid-binding function based on protein structure. *J Mol Biol. 2003*
2. Ahmad, Shandar, and Akinori Sarai. Moment-based prediction of DNA-binding proteins. *Journal of Molecular Biology* 341, no. 1 (July 30, 2004): 65-71.
3. Bhardwaj et al. Kernel-based machine learning protocol for predicting DNA-binding proteins. *Nuc. Acids Res. 2005*

4. Szilágyi A., Skolnick J.. Efficient Prediction of Nucleic Acid Binding Function from Low-resolution Protein Structures *Journal of Molecular Biology.* 2006; 358:922–933.
5. De Raedt L.. *Logical and Relational Learning.* Springer 2008.
6. Nassif, H., Al-Ali, H., Khuri, S., Keirouz, W., and Page, D. (2009a). An Inductive Logic Programming approach to validate hexose biochemical knowledge. In Proceedings of the 19th International Conference on ILP, pages 149-165, Leuven, Belgium.
7. Kuželka O., Železný F.. Block-wise construction of acyclic relational features with monotone irreducibility and relevancy properties in *ICML '09: 26th International Conference on Machine Learning* 2009.
8. J.L. Moreland and A.Gramada and O.V. Buzko and Qing Zhang and P.E. Bourne. The Molecular Biology Toolkit (MBT): A Modular Platform for Developing Molecular Visualization Applications. *BMC Bioinformatics.* 2005.