Searching for Important Amino Acids in DNA-binding Proteins for Histogram Methods

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
szaboand@fel.cvut.cz

Abstract. We develop a method capable to identify important amino acids for histogram-based methods predicting DNA-binding propensity. This method can be used both for prediction from sequence information (*Tube Histograms*) and prediction from structural information (*Ball Histograms*). We validate our method in prediction experiments using only proteins' primary structure, achieving favourable accuracies. Moreover, the histogram-based methods equipped with this new searching method also provide interpretable features involving distributions of amino acids.

Keywords: Feature construction, Proteomics, DNA-binding proteins

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 been proposed for the prediction of DNA-binding function from protein structure.

In this paper we will be concerned with prediction of DNA-binding propensity from sequence information. Previously developed methods for DNA-binding-propensity prediction can be divided into two main groups: alignment-based approaches [5] and physicochemical-property-based approaches [8, 10]. Gao and Skolnick [5] developed a threading-based method for the prediction of DNA-binding domains and associated DNA-binding protein residues. Ofran et al. [8] used only protein sequence information, without requiring any additional experimental or structural information. Their method relies on sequence environment, evolutionary profiles and predicted structural features (secondary structure, solvent accessibility and globularity). Patel et al. [10] used artificial neural network for prediction from amino acid sequences. Yan et al. [4] started with a Naive Bayes classifier trained to predict whether a given amino acid residue is a DNA-binding residue based on its identity and the identities of its four sequence neighbours on each side of the target residue.

Recently, we have introduced histogram-based methods which are able to predict DNA-binding propensity either from sequence information (*Tube Histograms*) or from structural information (*Ball Histograms* [11]). In this paper we develop a method capable to identify important amino acids for histogram-based methods predicting DNA-binding propensity.

2 Method

We propose the following approach to predict DNA-binding propensity. It consists of four main parts. First, so-called *templates* are found, which determine amino acids whose distributions should be captured by *tube histograms*. In the second step *tube histograms* are constructed for all proteins in a training set. Third, a transformation method is used to convert these histograms to a form usable by standard machine learning algorithms. Finally, a random forest classifier [3] is learned on this transformed dataset and then it is used for classification.

A *template* is a list of names of some Boolean amino acid properties. Given a template and a location in the primary structure of a protein, we infer a list of binary values indicating the truth values of the respective properties in the template for the amino acid at the position. For example, the template (Arg, Lys, $Positive$, $Negative$, $Neutral$) acquires the value $(1, 0, 1, 0, 0)$ if the amino acid at the inquired position is an Arginine. A *tube* of size s represents a part of an amino-acid sequence containing s consecutive amino acids (see Fig. 1).

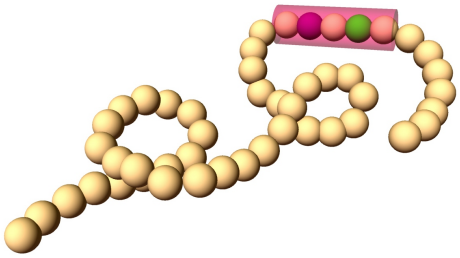

Fig. 1. Illustration of the Tube Histogram Method - Amino acids are shown as small balls in sequence forming an amino acid chain. They have different colors according to their type.

Given a protein, a template $\tau = (f_1, \ldots, f_k)$ and a sampling-tube size s, a *tube histogram* is a k-dimensional histogram constructed as follows. Starting by placing the sampling tube on the first s amino acids we get the first *sample* for the histogram. When a sample is collected the numbers of amino acids complying with the particular properties listed in the given template are extracted from it and stored. In further steps the tube is moved by one amino acid at time along the protein sequence and the samples are continuously stored for subsequent histogram construction. This process ends when the last amino acid is reached.

Finally, the histogram constructed from the collected samples is normalized. Intuitively, tube histograms capture the joint probability that a randomly picked *sampling tube* (Fig. 1) will contain exactly t_1 amino acids complying with f_1, t_2 amino acids complying with f_2 etc.

In our previous study [11] we used pre-fixed templates with charged amino acids selected according to [9, 7, 6]. Here, we introduce a method for automatic selection of templates which are sufficiently discriminative to distinguish DNA-binding proteins from non-DNA-binding proteins. The basic idea of the method is to find templates which maximize *distance* between average histograms from the two classes (DNA-binding and non-DNA-binding proteins). Intuitively, such templates should allow us to construct classifiers with good discriminative ability.

We construct the templates in a heuristic way using best-first search algorithm to maximize Bhattacharyya distance [2] between the average histograms from the two classes. In order to avoid repeated construction of histograms from the whole datasets, we construct a histogram corresponding to the biggest possible template (containing all amino acid properties), then, during the search, we construct histograms for the other templates by marginalising this biggest histogram.

3 Results

In this section we present experiments performed on real-life data (PD138 [12]/ NB110 [1]). We decided to study distribution of amino acids (represented by *tube histograms*). We constructed histograms with automatically discovered templates (with maximum length 5) and three different sampling-tube sizes: 5, 10 and 15. We trained random forest classifiers selecting optimal sampling-tube size and an optimal number of trees for each fold by internal cross-validation. The estimated accuracy and area under ROC is shown in Table 1. As we can see, the accuracy of our method exceeds the accuracy obtained by the method used in [12].

Method	**Accuracy**	**AUC**
Szilágyi et al.	81.4	0.92
Tube Histogram	**86.3**	**0.94**

Table 1. Accuracies and AUCs estimated by 10-fold cross-validation on PD138/NB110.

The four most informative automatically selected templates are: *(Arg, Cys, Lys, Gly, Ala)*, *(Arg, Cys, Lys, Gly, Asp)*, *(Arg, Cys, Lys, Gly, Glu)*, *(Arg, Cys, Lys, Gly, Leu)*. It is noteworthy that each charged amino acids (under normal circumstances *Arg* and *Lys* are positively charged, whereas *Glu* and *Asp* are charged negatively) is contained at least one of these templates.

4 Conclusions

We developed a method capable to identify important amino acids for histogram-based methods predicting DNA-binding propensity. We validated our method in prediction experiments using only proteins' primary structure, achieving favourable accuracies. In future work we plan to validate this method in prediction experiments using proteins' structural information (*Ball Histograms*).

Acknowledgement: Andrea Szabóová and Filip Železný were supported by project ME10047 granted by the Czech Ministry of Education. Andrea Szabóová was further supported by the Czech Technical University internal grant #10-801940. Ondřej Kuželka was supported by the Czech Technical University internal grant #10-811550.

References

1. Shandar Ahmad and Akinori Sarai. Moment-based prediction of dna-binding proteins. *Journal of Molecular Biology*, 341(1):65 – 71, 2004.
2. A. Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society 35*, pages 99–109, 1943.
3. L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.
4. C. Yan C, M. Terribilini, F. Wu abd R. L. Jernigan, D. Dobbs D, and V. Honavar. Predicting dna-binding sites of proteins from amino acid sequence. *BMC Bioinformatics*, 19(7):262, 2006.
5. M. Gao and J. Skolnick. A threading-based method for the prediction of DNA-binding proteins with application to the human genome. *PLoS Computational Biology*, 5(11), 2009.
6. S. Jones, P. van Heyningen, H. M. Berman, and J. M. Thornton. Protein-DNA interactions: a structural analysis. *Journal of Molecular Biology*, 287(5):877–896, 1999.
7. Y. Mandel-Gutfreund, O. Schueler, and H. Margalit. Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *Journal of Molecular Biology*, 253(2):370–382, 1995.
8. Y. Ofran, V. Mysore, and B. Rost. Prediction of dna-binding residues from sequence. *Bioinformatics*, 23(13):347–53, 2007.
9. C. O. Pabo and R. T. Sauer. Transcription factors: structural families and principles of DNA recognition. *Annual review of biochemistry*, 61:1053–1095, 1992.
10. A. K. Patel, S. Patel, and P. K. Naik. Binary classification of uncharacterized proteins into DNA binding/non-DNA binding proteins from sequence derived features using ann. *Digest Journal of Nanomaterials and Biostructures*, 4(4):775–782, 2009.
11. A. Szabóová, O. Kuželka, S. Morales E., F. Železný, and J. Tolar. Prediction of dna-binding propensity of proteins by the ball-histogram method. In *ISBRA 2011, LNBI 6674*, pages 358–367, 2011.
12. András Szilágyi and Jeffrey Skolnick. Efficient prediction of nucleic acid binding function from low-resolution protein structures. *Journal of Molecular Biology*, 358(3):922 – 933, 2006.