Prediction of DNA-binding Propensity of Proteins by the Ball-Histogram Method

Andrea Szabóová[1], Ondřej Kuželka[1], Sergio Morales E.[2], Filip Železný[1], and Jakub Tolar[3]

[1] Czech Technical University, Prague, Czech Republic
[2] Instituto Tecnológico de Costa Rica ITCR
[3] University of Minnesota, Minneapolis, USA
szaboand@fel.cvut.cz

Abstract. We contribute a novel, *ball-histogram* approach to DNA-binding propensity prediction of proteins. Unlike state-of-the-art methods based on constructing an ad-hoc set of features describing the charged patches of the proteins, the ball-histogram technique enables a systematic, Monte-Carlo exploration of the spatial distribution of charged amino acids, capturing joint probabilities of specified amino acids occurring in certain distances from each other. This exploration yields a model for the prediction of DNA binding propensity. We validate our method in prediction experiments, achieving favorable accuracies. Moreover, our method also provides interpretable features involving spatial distributions of selected amino acids.

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.

In the early 80's, when the first three-dimensional structures of protein-DNA complexes were studied, Ohlendorf and Matthew noticed that the formation of protein-DNA complexes energetically driven by the electrostatic interaction of asymmetrically distributed charges on the surface of the proteins complement the charges on DNA [1]. Large regions of positive electrostatic potentials on protein surfaces has been suggested to be a good indication of DNA-binding sites.

Stawiski et al. proposed a methodology for predicting Nucleic Acid-binding function based on the quantitative analysis of structural, sequence and evolutionary properties of positively charged electrostatic surfaces. After defining the electrostatic patches they found the following features for discriminating the DNA-binding proteins from other proteins: secondary structure content, surface area, hydrogen-bonding potential, surface concavity, amino acid frequency and

composition and sequence conservation. They used 12 parameters to train a neural network to predict the DNA-binding propensity of proteins [2].

Jones et al. analysed residue patches on the surface of DNA-binding proteins and developed a method of predicting DNA-binding sites using a single feature of these surface patches. Surface patches and the DNA-binding sites were analysed for accessibility, electrostatic potential, residue propensity, hydrophobicity and residue conservation. They observed that the DNA-binding sites were amongst the top 10% of patches with the largest positive electrostatic scores [3].

Tsuchiya et al. analysed protein-DNA complexes by focusing on the shape of the molecular surface of the protein and DNA, along with the electrostatic potential on the surface, and constructed a statistical evaluation function to make predictions of DNA interaction sites on protein molecular surfaces [4].

Ahmad and Sarai trained a neural network based on the net charge and the electric dipole and quadrupole moments of the protein. It was found that the magnitudes of the moments of electric charge distribution in DNA-binding protein chains differ significantly from those of a non-binding control data set. It became apparent that the positively charged residues are often clustered near the DNA and that the negatively charged residues either form negatively charged clusters away from the DNA or get scattered throughout the rest of the protein. The entire protein has a net dipole moment, because of the topological distribution of charges. The resulting electrostatic force may steer proteins into an orientation favorable for binding by ensuring that correct side of the protein is facing DNA [5].

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 positive potential surface patches, the protein's overall charge and its overall and surface amino acid composition [6]. In case of overall composition, noticeable differences were observed (in binding and non-binding cases) with respect to the frequency of Lys and Arg. These are positively charged amino acids, so their over-representation in DNA-binding proteins is evident.

Szilágyi and Skolnick created a logistic regression classifier based on ten variables to predict whether a protein is DNA-binding from its sequence and low-resolution structure. To find features that discriminate between DNA-binding and non-DNA-binding proteins, they tested a number of properties. The best combination of parameters resulted in the amino acid composition, the asymmetry of the spatial distribution of specific residues and the dipole moment of the protein. When ranking these parameters by relative importance, they found out that the Arginine content was the strongest predictor of DNA-binding, followed by the Glycine and Lysine. The dipole moment was the fourth most important variable [7].

Here we contribute a novel approach to DNA-binding propensity prediction, called the *ball-histogram* method, which improves on the state-of-the-art approaches in the following way. Rather than constructing an ad-hoc set of features describing the charged patches of the proteins, we base our approach on a systematic, Monte-Carlo-style exploration of the spatial distribution of charged amino

acids (under normal circumstances, Arg and Lys are positively charged, whereas Glu and Asp are charged negatively). For this purpose we employ so-called ball histograms, which are capable of capturing joint probabilities of specified amino acids occurring in certain distances from each other. Another positive aspect of our method is that it provides us with interpretable features involving spatial distributions of selected amino acids.

The rest of the paper is organized as follows. Section 2 describes the protein data sets we use in the study. In Section 3 we explain the ball histogram method. Section 4 exposes the results of applying the method on the protein data. In Section 5 we conclude the paper.

2 Data

DNA-binding proteins are proteins that are composed of DNA-binding domains. A DNA-binding domain is an independently folded protein domain that contains at least one motif that recognizes double- or single-stranded DNA. We decided to investigate structural relations within these proteins following the spatial distributions of certain amino acids in available DNA-protein complexes.

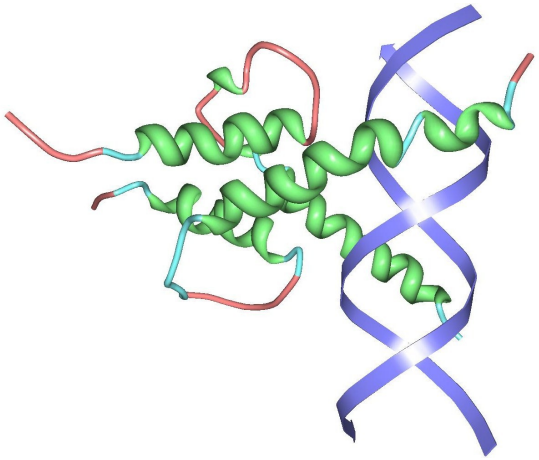

Fig. 1. Exemplary DNA-binding protein in complex with DNA shown using the protein viewer software [8]. Secondary structure motifs are shown in green (α-helices), light blue (turns) and pink (coils); the two DNA strands are shown in blue.

We decided to work with a positive data set (PD138) of 138 DNA-binding protein sequences in complex with DNA. It was created using the Nucleic Acid Database by [7] - it contains a set of DNA-binding proteins in complex with DNA strands with a maximum pairwise sequence identity of 35% between any two sequences. Proteins have $\leq 3.0\mathring{A}$ resolution. An example DNA-binding protein in complex with DNA is shown in Fig.1.

Rost and Sander constructed a dataset (RS126) for secondary structure prediction. Ahmad & Sarai [5] removed the proteins related to DNA binding from it, thus getting a final dataset of non-DNA-binding proteins. We used this set of non-DNA-binding proteins as our negative dataset (NB110).

From the structural description of each protein we extracted the list of all contained amino acids with information on their type and spatial structure.

3 Method

In this section we describe our novel method for predictive classification of proteins using so-called *ball histograms*. The classification method consists of three main steps. First, *ball histograms* are constructed for all proteins in a training set and then a transformation method is used to convert these histograms to a form usable by standard machine learning algorithms. Finally, a random forest classifier [9] is learned on this transformed data and then it is used for classification. Random forest classifier is known to be able to cope with large numbers of attributes such as in our case of *ball histograms* [10].

3.1 Ball Histograms

In this section we describe *protein ball histograms*. We start by defining several auxiliary terms. A *template* is a list of amino acid types or amino acid properties. An example of a template is (Arg, Lys) or $(Positive, Negative, Neutral)$. We say that an amino acid *complies with a property* f_1 if it has the corresponding property. For example, if A is an Arginine then it complies both with property Arg and with property $Positive$. A *bounding sphere* of a protein is a sphere with center located in the geometric center of the protein and with radius equal to the distance from the center to the farthest amino acid of the protein plus the diameter of *sampling ball* which is a parameter of the method. We say that an amino acid *falls* within a sampling ball if the alpha-carbon of that amino acid is contained in the sampling ball in geometric sense.

Given a protein, a template $\tau = (f_1, \ldots, f_k)$, a sampling-ball radius R and a bounding sphere S, a *ball histogram* is defined as:

$$H_\tau(t_1, \ldots, t_k) = \frac{\int\int\int_{(x,y,z)\in S} I_{T,R}(x, y, z, t_1, \ldots, t_k)dxdydz}{\sum_{(t'_1,\ldots,t'_k)} \int\int\int_{(x,y,z)\in S} I_{T,R}(x, y, z, t'_1, \ldots, t'_k)dxdydz} \quad (1)$$

where $I_{T,R}(x, y, z, t_1, \ldots, t_k)$ is an indicator function which we will define in turn. The expression $\sum_{(t'_1,\ldots,t'_k)} \int\int\int_{(x,y,z)\in S} I_{T,R}(x, y, z, t'_1, \ldots, t'_k)dxdydz$ is meant as a normalization factor - it ensures that $\sum_{(t_1,\ldots,t_k)} H_T(t_1, \ldots, t_k) = 1$. In order to define the indicator function $I_{T,R}$ we first need to define an auxiliary indicator function $I'_{T,R}(x, y, z, t_1, \ldots, t_k)$

$$I'_{T,R}(x,y,z,t_1,\ldots,t_k) = \begin{cases} 1 & \text{if there are exactly } t_1 \text{amino acids complying} \\ & \text{with } f_1, t_2 \text{ amino acids complying with} f_2 \text{ etc.} \\ & \text{in a sampling ball with center } x,y,z \text{ and radius } R \\ 0 & \text{otherwise} \end{cases}$$

Notice that $I'_{T,R}(x,y,z,0,\ldots,0)$ does not make any distinction between a sampling ball that contains no amino acid at all and a sampling ball that contains some amino acids none of them complying with the parameters in the template T. Therefore if we used $I'_{T,R}$ in place of $I_{T,R}$ the histograms would be affected by the amount of empty space in the bounding spheres. Thus, for example, there might be a big difference between histograms of otherwise similar proteins where one would be oblong and the other one would be more curved. In order to get rid of this unwanted dependence of the indicator function $I_{T,R}$ on proportion of empty space in sampling spheres we define $I_{T,R}$ in such a way that it ignores the empty space. For $(t_1,\ldots,t_k) \neq 0$ we set

$$I_{T,R}(x,y,z,t_1,\ldots,t_k) = I'_{T,R}(x,y,z,t_1,\ldots,t_k).$$

In the cases when $(t_1,\ldots,t_k) = 0$ we set $I_{T,R}(x,y,z,t_1,\ldots,t_k) = 1$ if and only if $I'_{T,R}(x,y,z,t_1,\ldots,t_k) = 1$ and if the sampling ball with radius R at (x,y,z) contains at least one amino acid.

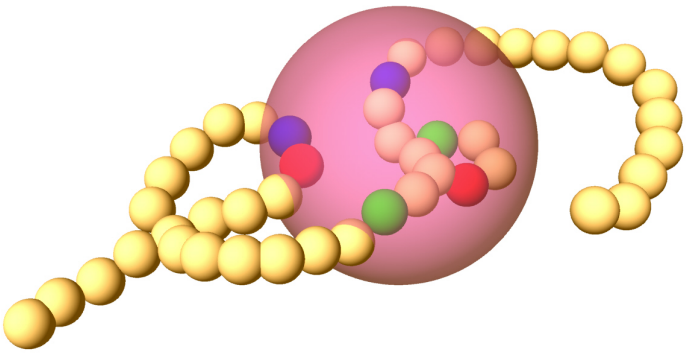

Fig. 2. Illustration of the Ball Histogram Method - Amino acids are shown as small balls in sequence forming an amino acid chain. A *sampling ball* is shown in violet. Some of the amino acids which comply with properties of an example template are highlighted inside the *sampling ball* area. They have different colors according to their type.

Ball histograms capture the joint probability that a randomly picked *sampling ball* (See Fig. 2) containing at least one amino acid will contain exactly t_1

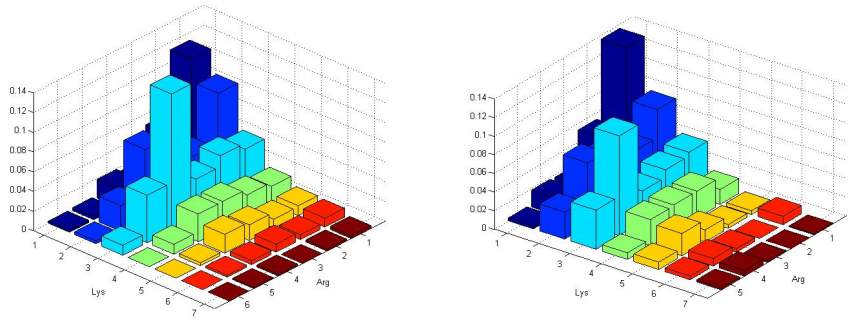

Fig. 3. Example ball histograms with template (Arg, Lys) and sampling-ball radius $R = 12\mathring{A}$ constructed for proteins 1A31 and 1A3Q from PD138.

amino acids complying with f_1, t_2 amino acids complying with f_2 etc. They are invariant to rotation and translation of proteins which is an important property for classification. Also note that the histograms would not change if we increased the size of the bounding sphere.

The indicator function $I_{T,R}$ makes crisp distinction between *amino acid falls within a sampling ball* and *amino acid does not fall within a sampling ball*. It could be changed to capture a more complex case by replacing 1 by the fraction of the amino acid that falls within the sampling ball, however, for simplicity we will not consider this case in this extended abstract.

3.2 Ball-Histogram Construction

Computing the integral in Eq. 1 precisely is infeasible therefore we decided to use a Monte-Carlo method. The method starts by finding the bounding sphere. First, the geometric center C of all amino acids of a given protein P is computed (each amino acid is represented by coordinates of its alpha-carbon). The radius R_S of the sampling sphere for the protein P is then computed as

$$R_S = \max_{Res \in P} (distance(Res, C)) + R$$

where R is a given sampling-ball radius. After that the method collects a predefined number of samples from the bounding sphere. For each sampling ball the algorithm counts the number of amino acids in it, which comply with the particular properties contained in a given template and increments a corresponding bin in the histogram. In the end, the histogram is normalized.

Example 1. Let us illustrate the process of histogram construction by a small example. Let us have a template (Arg, Lys). We assume that we already have

a bounding sphere. The algorithm starts by placing a sampling ball randomly inside the bounding sphere. Let us assume that the first such sampling ball contained the following amino acids: *2 Arginins and 1 Leucine* therefore we increment a counter in the histogram associated with vector $(2, 0)$. Then in the second sampling ball, we get *1 Histidine and 1 Aspartic acid* so we increment a counter associated with vector $(0, 0)$. We continue in this process until we have gathered a sufficient number of samples. In the end we normalize the histogram. Examples of such histograms are shown in Fig. 3.

3.3 Predictive Classification using Ball Histograms

In the preceding sections we have explained how to construct ball-histograms but we have not explained how we can use them for predictive classification. One possible approach would be to define a metric on the space of normalized histograms and then use either a nearest neighbour classifier or a nearest-centroid classifier. Since our preliminary experiments with these classifiers did not give us satisfying predictive accuracies, we decided to follow a different approach inspired by a method from relational learning known as *propositionalization* [11] which is a method for transferring complicated relational descriptions to attribute-value representations.

The transformation method is quite straightforward. It looks at all histograms generated from the proteins in a training set and creates a numerical *attribute* for each vector of property occurrences which is non-zero at least in one of the histograms. After that an attribute vector is created for each training example using the collected attributes. The values of the entries of the attribute-vectors then correspond to heights of the bins in the respective histograms. After this transformation a random forest classifier is learned on the attribute-value representation. This random forest classifier is then used for the predictive classification.

In practice, there is a need to select an optimal sampling-ball radius. This can be done by creating several sets of histograms and their respective attribute-value representations corresponding to different radiuses and then selecting the optimal parameters using an internal cross-validation procedure[4].

4 Results

In this section we present experiments performed on real-life data described in Section 2. We performed two types of experiments. First, we decided to study distribution of charged amino acids (represented by *ball histograms*). We constructed histograms with template (Arg, Lys, Glu, Asp) and three different

[4] When evaluating the classifiers' performance in Section 4 using 10-fold cross-validation, we optimize the sampling-ball radius parameter always on the nine training folds and then use it for the remaining testing fold, which is a standard way to obtain an unbiased estimate of the predictive performance of a classifier with tunable parameters.

sampling-ball radiuses: 6, 8 and 10$\mathring{A}$. We trained random forest classifiers selecting optimal sampling-ball radius and an optimal number of trees for each fold by internal cross-validation. The estimated accuracy is shown in Table 1. As we can see, the accuracy of our method is comparable with the accuracy obtained by the method used in [7]. They used properties of the following amino acids: Arg, Lys, Gly, Asp, Asn, Ser and Ala, whereas we used only the distribution of the charged amino acids. A natural question is whether our results could be improved by taking into account also this set of amino acids. Therefore, we performed the second set of experiments. In this case, the accuracy obtained by our method exceeded the accuracy of [7].

Method	**Accuracy** [%]
Szilágyi et al.	81.4
Ball Histogram using Charged Amino Acids	80.2
Ball Histogram using the Second Set of Amino Acids	84.7

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

In addition to improved accuracy, our method provides us with interpretable features involving spatial distributions of selected amino acids. We show the three most informative *features* according to the χ^2 criterion for the first set of experiments in Table 2 and for the second set of experiments in Table 3. Given a protein each feature captures the fraction of sampling balls, which contain the specified numbers of amino acids of given types. For example: the first feature from Table 2 denotes the fraction of sampling balls, which contain exactly one Arginine, one Lysine, no Glutamic acid and no Aspartic acid.

	Arg	Lys	Glu	Asp
1^{st} feature	1	1	0	0
2^{nd} feature	2	0	0	0
3^{rd} feature	1	0	0	0

Table 2. The three most informative features according to the χ^2 criterion using the distribution of the charged amino acids.

A quick look at the most informative features in the presented Tables 2 and 3 suggests that the major role is played by the amino acids: Arginine and Lysine. These amino acids are known to often interact with the negatively charged backbone as well as with the bases [12–14]. In order to get a global view on the differences between spatial distributions of these amino acids in DNA-binding and non-DNA-binding proteins, we computed the average ball histograms for these two classes of proteins. They are shown in Fig. 4. The histogram obtained

	Arg	Lys	Gly	Asp	Asn	Ser	Ala
1^{st} feature	1	1	0	0	0	0	0
2^{nd} feature	1	0	0	0	0	0	0
3^{rd} feature	2	0	0	0	0	0	0

Table 3. The three most informative features according to the χ^2 criterion using the distribution of the selected set of amino acids.

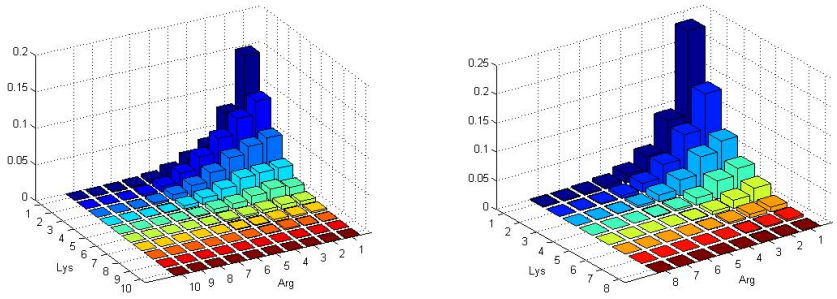

Fig. 4. Ball histograms with template (Arg, Lys) and sampling-ball radius $R = 12\mathring{A}$ averaged for all proteins from PD138 (left panel) and all proteins from NB110 (right panel).

by subtracting the average ball histogram for DNA-binding proteins from the average ball histogram for non-DNA-binding proteins is shown in Fig. 5. We can notice a remarkable difference of spatial distribution of Arginine and Lysine between DNA-binding and non-DNA-binding proteins.

5 Conclusion

We contributed a novel, *ball-histogram* approach to the prediction of DNA-binding propensity of proteins. We validated the method in prediction experiments with favorable results. Observing only the distribution of charged amino acids we achieved accuracies comparable to the state of the art [7]. Importantly though, the results reported by [7] had been achieved using the results of a prior systematic search for the best combination of parameters. In particular, the following amino acids were identified as critical: Arg, Lys, Gly, Asp, Asn, Ser and Ala. Using this set of amino acids instead of the original set of all charged amino acids, our accuracies in fact exceeded those of [7]. Moreover, our method provides us with interpretable features involving spatial distributions of selected amino acids.

Acknowledgement: Andrea Szabóová and Filip Železný were supported by external project ME10047 granted by the Czech Ministry of Education. Andrea Szabóová was further supported by the Czech Technical University internal

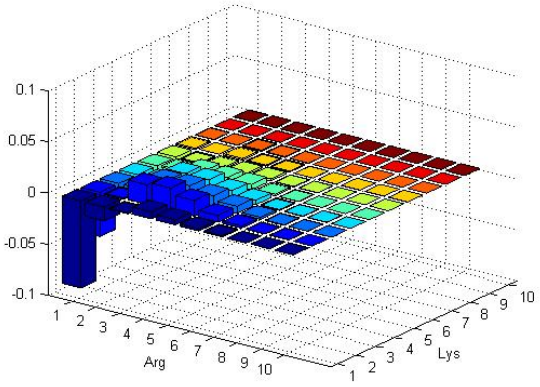

Fig. 5. Difference between histograms from Fig. 4.

grant #10-801940. Ondřej Kuželka was supported by the Czech Technical University internal grant OHK3-053/11. Sergio Morales was supported by Costa Rica Council for Scientific and Technological Research.

References

1. Ohlendorf, D. H. & Matthew, J. B. Electrostatics and flexibility in protein-DNA interactions. *Advances in Biophysics* Volume 20, 1985, Pages 137–151.
2. Stawiski, Gregoret, and Mandel-Gutfreund. Annotating nucleic acid-binding function based on protein structure. *J Mol Biol. 2003*
3. Jones, Shanahan, Berman, Thornton. Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucleic Acid Research, 2003, Vol. 31, No. 24, 7189–7198.
4. Tsuchiya, Kinoshita, Nakamura. Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. *Proteins: Structure, Function, and Bioinformatics*, 2004, Vol. 55, Issue 4, 885–894.
5. Ahmad, Shandar, and Akinori Sarai. Moment-based prediction of DNA-binding proteins. *Journal of Molecular Biology* 341, no. 1 (July 30, 2004): 65–71.
6. Bhardwaj et al. Kernel-based machine learning protocol for predicting DNA-binding proteins. *Nuc. Acids Res. 2005*
7. Szilágyi A., Skolnick J.. Efficient Prediction of Nucleic Acid Binding Function from Low-resolution Protein Structures *Journal of Molecular Biology.* 2006, 358:922–933.
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.
9. Breiman, Leo. Random Forests. *Machine Learning*. Vol. 45, 2001, 5–32.
10. Rich Caruana and Nikos Karampatziakis and Ainur Yessenalina. An empirical evaluation of supervised learning in high dimensions. *International Conference on Machine Learning (ICML)*. 2008, 96–103.

11. Nada Lavrač and Peter Flach. An Extended Transformation Approach to Inductive Logic Programming. *ACM Transactions on Computational Logic*. 2001, Vol. 2, 458–494.
12. Pabo, C. O. & Sauer, R. T. Transcription factors: structural families and principles of DNA recognition. *Annual Review of Biochemistry*. 1992. Vol.20, 137–151.
13. Mandel-Gutfreund, Y., Schueler, O. & Margalit, H. Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *Journal of molecular biology*. 1995. Vol. 253, 370–382.
14. Jones, S., van Heyningen, P., Berman, H. M. & Thornton, J. M. Protein-DNA interactions: a structural analysis. *Journal of molecular biology*. 1999. Vol. 287, 877–896.