Extending the Ball-Histogram Method with Continuous Distributions and an Application to Prediction of DNA-Binding Proteins

Ondřej Kuželka†, Andrea Szabóová†, Filip Železný
Department of Cybernetics
Czech Technical University in Prague
Prague, Czech Republic
Email: {kuzelon2, szaboand, zelezny}@fel.cvut.cz

***Abstract*—We introduce a novel method for prediction of DNA-binding propensity of proteins which extends our recently introduced ball-histogram method (Szabóová et al. 2012). Unlike the original ball-histogram method, it allows handling of continuous properties of protein regions. In experiments on four datasets of proteins, we show that the method improves upon the original ball-histogram method as well as other existing methods in terms of predictive accuracy.**

***Keywords*-DNA-binding proteins, machine learning, data mining.**

I. 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.

Some of the recent approaches ([1], [2], [3], [4], [5]) rely exclusively on protein structure data (whether sequential or spatial). Szilágyi and Skolnick [6] created a method based on a logistic regression classifier with ten variables (physicochemical properties) to predict from sequence and low-resolution structure of a protein whether it is DNA-binding. To our knowledge, the predictive accuracy achieved by the lastly mentioned strategy [6] was only improved by incorporating an additional source of background knowledge, in particular, information on evolutionarily conserved domains. Nimrod et al. [7] presented a random forest classifier for identifying DNA-binding proteins among proteins with known 3D structures using detected clusters of evolutionarily conserved regions on the surface of proteins.

It is nevertheless important to continue improving methods that do not exploit evolutionary information. Such methods are valuable mainly due to their ability to predict DNA-binding propensity for engineered proteins for which evolutionary information is not available. Engineered proteins are highly significant for example in emerging gene-therapy technologies [8]. In [9] we were concerned with prediction of DNA-binding propensity from spatial structure information without using evolutionary information. To this end, we developed the *ball-histogram* method, which improved on most of the mentioned state-of-the-art approaches. A somewhat limiting property of the original ball-histogram method was that it could only work with discrete properties of proteins' regions such as numbers of amino acids of given types. In this paper we improve the ball-histogram method by developing an approach for dealing with continuous properties of proteins' regions which improves predictive accuracy w.r.t. the original ball-histogram method. The programs and data described in this paper can be downloaded from *http://ida.felk.cvut.cz/users/kuzelka/BIBM2012.zip*.

† These authors contributed equally to the presented work and therefore should be considered joint-first authors.

II. Ball-Histogram Method

In this section we describe the original *ball-histogram method* which has already been applied to prediction of DNA-binding propensity of proteins in [9]. Originally, the motivation for the method was the observation that distributions of certain types of amino acids differed significantly between DNA-binding and non-DNA-binding proteins. This suggested that information about distributions of some amino acids in local regions of proteins could have been used to construct predictive models able to classify proteins as binding or non-binding given their spatial structure. We developed an approach which was able to capture fine differences between the distributions. It consisted of four main parts. First, so-called *templates* were found. In the second step *ball histograms* were constructed for all proteins in a training set. Third, a transformation method was used to convert these histograms to a form usable by standard machine learning algorithms. Finally, a random forest classifier was learned on this transformed dataset and then it was used for classification. In the rest of this section we describe this method in detail.

A *template* is a list of some Boolean amino acid properties. A *bounding sphere* of a protein structure is a sphere with center located in the geometric center of the protein

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

structure and with radius equal to the distance from the center to the farthest amino acid of the protein plus the diameter of the *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 the geometric sense.

A ball histogram for a protein P is computed as follows. 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 α-carbon). The radius R_S of the sampling sphere for the protein structure 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 pre-defined number of samples containing at least one amino acid 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.

III. Extending Ball Histograms with Continuous Variables

We start by explaining why the existing ball-histogram approach is not suitable for work with continuous attributes. Then we introduce so-called *polynomial aggregation features* and after that we show how they can be used in a ball-histogram-based approach to predictive classification.

A drawback of the original ball-histogram method is that it is ill-suited for work with continuous variables. For example, it is possible to model the *distributions* of Arginines and Lysines using the ball-histogram method. However, if we tried to model distributions of e.g. *hydropathy* and *volume* of amino acids in a given protein structure in the very same way, we would face serious difficulties stemming from combinatorial explosion of the number of histograms' bins - attributes. We use *multivariate polynomial aggregation* - a strategy that we have recently introduced in the context of statistical relational learning.

A *monomial feature* M is a pair $(\tau, (d_1, \ldots, d_k))$ where τ is a template with k properties and $d_1, \ldots, d_k \in \mathbf{N}$. *Degree* of M is $deg(M) = \sum_{i=1}^{k} d_i$. Given a sampling ball B placed on a protein structure P, we define the *value* of a monomial feature $M = (\tau, (d_1, \ldots, d_k))$ as $M(B) = \tau_1^{d_1} \cdot \tau_2^{d_2} \cdot \ldots \cdot \tau_k^{d_k}$ where τ_i is the average value of the i-th property of template τ averaged over the amino acids contained in the sampling ball B. Sometimes, we will use a more convenient notation for monomial features motivated by this definition of *value*:

$$(\tau = (\tau_1, \ldots, \tau_k), (d_1, \ldots, d_k)) \equiv^{def} \tau_1^{d_1} \cdot \tau_2^{d_2} \cdot \ldots \cdot \tau_k^{d_k}$$

Example 1: Let us have a template $\tau = [hydropathy, volume]$, a monomial feature $M = hydropathy \cdot volume^2$ and a sampling ball containing two Leucines ($hydropathy = 3.8$, $volume = 124$) and one Arginine ($hydropathy = -4.5$, $volume = 148$). Then

$$M(B) = \frac{2 \cdot 3.8 - 4.5}{3} \cdot \left(\frac{2 \cdot 124 + 148}{3}\right)^2 \approx 1.8 \cdot 10^4$$

A *multivariate polynomial feature* is an expression of the form $N = \alpha_1 M_1 + \alpha_2 M_2 + \ldots + \alpha_k M_k$ where $M_1, \ldots, M_k$ are monomial features and $\alpha_1, \ldots, \alpha_k \in R$ (formally expressed as a pair of two ordered sets - one of monomials and one of the respective coefficients). *Value* of a polynomial feature $N = \alpha_1 M_1 + \ldots + \alpha_k M_k$ w.r.t to a sampling ball B placed on a protein structure P is defined as $N(B) = \alpha_1 M_1(B) + \alpha_2 M_2(B) + \ldots + \alpha_k M_k(B)$. *Degree* of a polynomial aggregation feature P is maximum among the degrees of its monomials.

Now, we extend the definitions of values of monomial and polynomial features for protein structures. Given a polynomial aggregation feature N and a sampling-ball radius R, we define the *value* $N(P)$ of a polynomial feature N w.r.t. a protein structure P as:

$$N(P) = \frac{\int_{\widehat{P}} N(B) dB}{\int_{\widehat{P}} dB} \quad (1)$$

where $\widehat{P}$ is the set of all sampling balls with radius R which contain at least one amino acid of the protein structure P. The integral $\int_{\widehat{P}} dB$ in the denominator is used as a normalization constant. Intuitively, the integral computes the average value of a polynomial feature N over balls located on a given protein structure.

It can be seen quite easily that polynomial aggregation features on protein structures share convenient properties with the discrete ball histograms. They are invariant to rotation and translation of the protein structures which is important for predictive classification tasks. Intuitively, a monomial feature $M = \tau_i$ corresponds to the average value of property τ_i (in sampling balls of a given radius) over a given protein structure. A monomial feature $M = \tau_i^2$ captures the *dispersion* of the values of property τ_i over a given protein structure. Indeed, let us have two proteins A and B and a monomial feature $M = charge^2$ and let us assume that A and B are composed of the same number of amino acids and that they contain the same number of positively charged amino acids and no negatively charged amino acids. Finally, let us also assume that the positively charged amino acids are distributed more or less uniformly over the protein structure A but are concentrated in a small region of the protein structure B. Then it is not hard to see that for the values $M(A)$ and $M(B)$ it should hold $M(A) \leq M(B)$. Analogically, a monomial feature $M = \tau_i \cdot \tau_j$ corresponds to *agreement* of values of properties τ_i and τ_j over a given protein structure but the *covariance* of these values is better captured by the following expression involving monomial features: $M_1(P) - M_2(P) \cdot M_3(P)$

where $M_1 = \tau_i \cdot \tau_j$, $M_2 = \tau_i$ and $M_3 = \tau_j$. Note that this expression is not a polynomial aggregation feature but only an expression composed of polynomial (monomial) aggregation features. This can be seen when we expand $M_1(P)$, $M_2(P)$ and $M_3(P)$ and obtain

$$M_1(P) - M_2(P)M_3(P) = \frac{\int_{\widehat{P}} \tau_i \cdot \tau_j dB}{\int_{\widehat{P}} dB} - \frac{\int_{\widehat{P}} \tau_i dB}{\int_{\widehat{P}} dB} \cdot \frac{\int_{\widehat{P}} \tau_j dB}{\int_{\widehat{P}} dB}$$

which is not a value of a polynomial aggregation feature. However, it can be easily constructed from some polynomial aggregation features.

Values of polynomial aggregation features can be further decomposed into so called *k-values* computed only from balls containing exactly k amino acids. Given a polynomial feature N and a positive integer k, the k-value of N w.r.t. a protein P is given as

$$N(P|k) = \frac{\int_{\widehat{P}_k} N(B) dB}{\int_{\widehat{P}_k} dB}$$

where $\widehat{P}_k$ is the set of all sampling balls which contain exactly k amino acids. The value of a polynomial feature can then be expressed using k-values as

$$N(P) = \sum_i \beta_i \cdot N(P|i)$$

where $\beta_i = \int_{\widehat{P}_i} dB / \int_{\widehat{P}} dB$.

When using polynomial features for construction of attributes for machine learning, we can rely solely on the k-values and the few proportions and let the machine learning algorithms compute the values of monomial or polynomial aggregation features from these values if needed.

Polynomial aggregation features can be used for predictive classification in a way completely analogical to discrete ball histograms. Given a template τ, sampling-ball radius R, a maximum degree d_{max} and a protein structure, we construct all monomials containing the continuous variables from τ and having degree at most d_{max}. After that we construct the attribute-table. The rows of this table correspond to examples and the columns (attributes) correspond to k-values of the constructed monomial features. There is an attribute for every k-value such that there is at least one protein structure in the dataset such that it contains a set of k amino acids which fit into a ball of radius R.

The integrals used in definitions of values (or k-values) of monomial aggregation features are difficult to evaluate precisely therefore we use a Monte-Carlo-based approach similar to the case of discrete ball histograms. The set of k-values of monomial aggregation features for a protein P is computed as follows. First, a bounding sphere is found for the protein structure (with geometric center located in the geometric center of the protein structure and with radius $R_S = \max_{Res \in P}(distance(Res, C)) + R$, where R is a specified sampling-ball radius). After that the method collects a pre-defined number of samples containing at least one amino acid from the bounding sphere. For each sampling ball B the algorithm computes k_B-values (where k_B is the number of amino acids contained in B) of all monomial features complying with a given template and with a given maximum degree and stores them. In the end, the collected k-values of sampling balls are averaged to produce *approximate* k-values for the protein structure P.

After the attribute-table is constructed, it can be used to train an attribute-value classifier such as random forest or support vector machine which can be then used for prediction on unseen proteins.

IV. Experiments

We used two datasets of DNA-binding proteins (PD138, UD54) and two datasets of non-DNA-binding proteins (NB110, NB843) in our experiments. The dataset PD138 was created using the Nucleic Acid Database (NDB) by [6]. It contains DNA-binding proteins in complex with DNA with a maximum pairwise sequence identity of 35% between any two sequences. We discard the information about DNA for the purpose of our experiments. However, both proteins and 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. Therefore it is important to assess any method for prediction of DNA-binding function also on DNA-binding proteins in unbound conformation. For this, we used the dataset UD54 of 54 DNA-binding proteins in unbound conformation. This dataset was also obtained from [6]. We used two datasets of proteins which do not bind to DNA: NB110 and NB843. The former dataset was created by Ahmad and Sarai [4] from an earlier dataset of Rost and Sandler [10] by removing proteins related to DNA-binding from it. The latter dataset was created by Nimrod et al. [7] by adding 733 non-DNA-binding proteins to the dataset NB110 in order to make the ratio of DNA-binding proteins more realistic.

We performed predictive experiments with the four combinations of datasets. We used monomial aggregation features with maximum degree 3 and the following basic chemical properties of amino acids: charge, Van der Waals volume, hydropathy index, isoelectric point (pI), dissociation constants pK1 and pK2 and the following three properties related to DNA-binding derived by Sathyapriya et al. [11], base-contact propensity, sugar-contact propensity and phosphate-contact propensity. We trained random forest classifiers using only the attributes having non-zero information gain-ratio on training set. When performing cross-validation, this attribute selection was performed separately on the respective training sets induced by cross-validation so that no information could leak from a training set to a testing set. We compared the *continuous ball-histogram method* presented in this paper with the original discrete ball-histogram method and with the method of Szilágyi and Skolnick [6], which

	Continuous ball histograms		Discrete ball histograms [9]		Szilágyi et al. [6]	
	Accuracy	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**
PD138/NB110	**0.89**	**0.95**	0.87	0.94	0.81	0.92
PD138/NB843	**0.89**	0.86	0.88	**0.87**	0.87	0.84
UD54/NB110	**0.87**	**0.90**	0.81	0.89	0.82	0.89
UD54/NB843	**0.95**	**0.83**	0.94	0.81	0.94	0.78

Table I
EXPERIMENTAL RESULTS ESTIMATED BY 10-FOLD CROSS-VALIDATION.

we reimplemented. The estimated accuracies and AUCs are shown in Table I. The new continuous ball-histogram method performed best in terms of accuracy in all cases and in terms of AUC in all but one case where the discrete ball-histogram method performed best. We also tested the original ball-histogram method with random forest classifiers enriched with attribute-selection but it did not improve performance.

In order to see whether the ball-histogram method, which uses only structural information, could come close to the results of methods which exploit also information about evolutionary conservation of regions on protein surfaces, we compared our results with the results of Nimrod et al. [7]. The AUC 0.96 and accuracy 0.90 reported in [7] for the datasets PD138 and NB110 differs only slightly (by 0.01) from our best results. The AUC 0.90 obtained for the datasets PD138 and NB843 differs by 0.04 from our best results. These results are encouraging given how important evolutionary information turned out to be according to experiments from [7]. When removing evolutionary information, Nimrod et al.'s misclassification error on the dataset PD138/NB110 increased by 0.035 which corresponds to lower predictive accuracy than obtained by our method.

V. CONCLUSIONS

We have extended our recently introduced *ball histogram method* by incorporation of polynomial aggregation features which are able to capture distributions of continuous properties of proteins' regions. The method achieved higher predictive accuracies than the original ball-histogram method as well as an existing state-of-the-art method. There are interesting future research directions regarding our novel approach. For example, it would be interesting to explore the possibility to use more chemical descriptors of amino acids or protein-regions.

ACKNOWLEDGMENT

This work was supported by the Czech Grant Agency through project 103/11/2170 *Transferring ILP techniques to SRL*, by project ME10047 granted by the Czech Ministry of Education and by the Czech Technical University in Prague through the student grant competition project SGS11/155/OHK3/3T/13.

REFERENCES

[1] E. Stawiski, L. Gregoret, and Y. Mandel-Gutfreund, "Annotating nucleic acid-binding function based on protein structure," *Journal of Molecular Biology*, 2003.

[2] S. Jones, H. P. Shanahan, H. M. Berman, and J. M. Thornton, "Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins," *Nucleic Acids Research*, vol. 31, no. 24, pp. 7189–7198, 2003.

[3] Y. Tsuchiya, K. Kinoshita, and H. 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*, vol. 55, no. 4, pp. 885–894, 2004.

[4] S. Ahmad and A. Sarai, "Moment-based prediction of dna-binding proteins," *Journal of Molecular Biology*, vol. 341, no. 1, pp. 65–71, 2004.

[5] N. Bhardwaj, R. Langlois, G. Zhao, and L. H, "Kernel-based machine learning protocol for predicting dna-binding proteins," *Nucleic Acids Research*, 2005.

[6] A. Szilágyi and J. Skolnick, "Efficient prediction of nucleic acid binding function from low-resolution protein structures," *Journal of Molecular Biology*, vol. 358, no. 3, pp. 922–933, 2006.

[7] G. Nimrod, A. Szilágyi, C. Leslie, and N. Ben-Tal, "Identification of dna-binding proteins using structural, electrostatic and evolutionary features," *Journal of molecular biology*, vol. 387, no. 4, pp. 1040–53, 2009.

[8] T. Cathomen and J. Joung, "Zinc-finger nucleases: The next generation emerges," *Molecular Therapy*, vol. 16, 2008.

[9] A. Szabóová, O. Kuželka, F. Železný, and J. Tolar, "Prediction of dna-binding propensity of proteins by the ball-histogram method using automatic template search," *BMC Bioinformatics*, vol. 13, no. Suppl 10, p. S3, 2012.

[10] B. Rost and C. Sander, "Improved prediction of protein secondary structure by use of sequence profiles and neural networks," *Proc. Natl Acad. Sci.*, p. 75587562, 1993.

[11] R. Sathyapriya, M. S. Vijayabaskar, and S. Vishveshwara, "Insights into proteindna interactions through structure network analysis," *PLoS Comput Biol*, vol. 4, no. 9, p. e1000170, 09 2008.