Bio-Medical Materials and Engineering 26 (2015) S1883–S1889
DOI 10.3233/BME-151491
IOS Press

Predicting a DNA-binding protein using random forest with multiple mathematical features

Changge Guan[a], Xiaohui Niu[b], Feng Shi[c], Kun Yang[a] and Nana Li[c,*]

[a]*College of Life Science and Technology, Huazhong Agricultural University, Wuhan, 430070, P.R. of China*

[b]*College of Informatics, Huazhong Agricultural University, Wuhan,430070, P.R. of China*

[c]*College of Science, Huazhong Agricultural University, Wuhan, 430070, P.R. of China*

Abstract. DNA-binding proteins are involved and play a crucial role in a lot of important biological processes. Hence, the identification of the DNA-binding proteins is a challenging and significant problem. In order to reveal the intrinsic information correlated to DNA-binding, nine classes of candidate features based on different mathematical fields are applied to construct the prediction model with random forest. They are fractal dimension, conjoint triad feature, Hilbert-Huang Transformation, amino acid composition, dipeptide composition, chaos game representation, and the corresponding information entropies. These mathematical expressions are evaluated with 5-fold cross validation test. The results of numerical simulations show that the mathematical features consisted of amino acid composition, fractal dimension and information entropies of amino acid and chaos game representation achieve the best performance. Its accuracy is 0.8157, and Matthew's correlation coefficient (MCC) achieves 0.5968 on the benchmark dataset from DNA-Prot. By analyzing the components of top combination of the nine candidate features, the concepts of fractal dimension and information entropy are the effective and vital features, which can provide complementary sequence-order information on the basis of amino acid composition.

Keywords: DNA-binding proteins, random forest, information entropy, fractal dimension, Hilbert-Huang transformation

1. Introduction

DNA-binding proteins are the most important protein families that possess the specific affinity with corresponding DNA strands [1]. DNA-binding proteins take part in a lot of vitally requisite biological processes. As previous research goes, in recognition of specific nucleotide sequences, transcription regulation, gene expression and gene repair, DNA-binding proteins play a very crucial role [2]. Due to the importance of DNA-binding proteins, identifying these proteins can provide a significant cue to further understand biological function. At present, there are several specific experimental techniques (including filer binding assays, X-ray crystallography, genetic analysis and so forth) [3]. Although experimental techniques can distinctly demonstrate the binding pattern, these experimental techniques

[*] Address for correspondence: Nana Li, College of Science, Huazhong Agricultural University, Wuhan, 430070, P.R. of China. Tel.: 86-27-87282425; Fax: 86-27-87282425; E-mail: nnl1010@126.com.

face two bottlenecks. On the one hand it is expensive and time consuming in the application. On the other hand, there are about 6-7% genes which can encode DNA-binding proteins according to inference. Facing the avalanche of the DNA-binding proteins, it is thirsted for developing the effective and rapid method to identify DNA-binding proteins.

With the rapid development of bioinformatics, there are many successful computational methods to predicting DNA-binding proteins. The tertiary structure was employed to predict DNA-binding proteins in the early attempts [4-8]. However, lack of detailed structural information restricts the generalization of this kind of methods. All of these current situations motivate the development of the computational method to predict DNA-binding proteins based on protein primary sequence information [9-24].

Among the methods based on protein primary sequence information, Bhardwaj, et al. set up the Support vector machine (SVM)-based prediction model with three feature classes (including overall charge, electrostatic patch, amino acid composition) [9]. Kumar, et al. had developed the SVM-based approach to identify the DNA-binding proteins with the evolutionary information as the mathematical expression. Polysaccharide Storage Myopathy (PSSM) was employed in this paper [10]. Kumar, et al. combined several classes of features, such as amino acid composition, peptide composition, physic chemical property and PSSM, to contrast the classifier with random forest [15]. Lin, et al. developed an effective and rapid predictor to distinguish the DNA-binding and non DNA-binding proteins using random forest as classifier and feature extraction with grey model [18]. It is worthy to mention that all of these progresses have proved that the primary sequence information hides the cue to predict and understand DNA-binding proteins. Furthermore the sequence-based prediction models are rapid and effective methods with satisfied performance.

As everyone knows, feature extraction is a very important step to construct the prediction model. Feature extraction aims to use various mathematical methods to extract mathematical expression from protein sequences. Different feature extraction can lead to widely divergent performance. In this paper, several classes of features are extracted from primary sequence of proteins. Through a great deal of numerical experiments, the important features have been selected, and the optimal model has achieved the satisfactory performance.

2. Material and methods

2.1. Benchmark dataset

In order to evaluate the effectiveness of this method, one benchmark dataset was selected from a previous study [15]. This dataset consists of a non-redundant training set and three test dataset with different scales. This dataset can be downloaded from the link: http://www3.ntu.edu.sg/home/EPNSugan/index_fles/dnaprot.htm.

2.2. Feature extraction

Feature extraction is important for a helpful predictor. The suitable features can truly reflect their intrinsic correlation with the target to be predicted. In order to extract comprehensive information from the primary sequence of a protein, nine classes of features were extracted. They are Amino Acid Composition (AAC), Chaos Game Representation (CGR), Fractal Dimension, Dipeptide composition, Information Entropy of CGR (En_CGR), Information Entropy of AAC (En_AAC), Information

Entropy of dipeptide composition (En_dipeptide), Conjoint Triad Feature (CTF) and Hilbert Huang Transformation (HHT). Among them, AAC is widely used in the prediction of protein function. Through the small scale of pre-experiment, its importance for DNA-binding protein prediction was proved. Hence, AAC served as a necessary feature in this paper. The detailed introduction of these features is shown as follows.

2.2.1. Chaos game representation (CGR)

Chaos Game Representation (CGR) is an iterative mapping technique. Following this iterative rule, a given sequence can be transformed into a picture. This CGR technique was firstly proposed to study DNA sequence by Jeffrey in 1990 [25]. Then, it was generalized to study protein sequence by Basu, et al. [26]. Basu clustered 20 native amino acids into 12 groups, which stand for 12 vertexes of 12-sided regular polygon, based on conservative substitutions. Subsequently the CGR graph can be divided by grid lines. The occurrence frequency of each grid is calculated as Chaos Game Representation.

2.2.2. Fractal dimension (FD)

Fractal Dimension was proposed by Mandelbrot [27]. It is a quantitative descriptor to measure the complexity of geometric objects which should satisfy three important properties of the fractal geometry. However, fractal objects always have non-integer fractal dimension. There are many methods used to estimate the fractal dimension. Here box-counting dimension was adopted to represent protein sequence [28].

2.2.3. Three information entropies

● Information Entropy

Information Entropy was firstly introduced by Claude E. Shannon [29]. It is one of the two fundamental concepts in the information theory. Information Entropy stands for the expected value of the information contained in a message. From a statistical view, a 'message' can be perceived as a specific realization of the random variable. In this paper, the information entropy was deemed as a measure of the uncertainty for a specific distribution. It has previously been applied to predict the protein solubility [29].

● Information Entropy of AAC

Information entropy can measure the uncertainty of a given distribution. The distribution is the normalized occurrence frequencies in the primary sequence of a given protein.

● Information Entropy of Dipeptide Composition

As mentioned above, the distribution is the normalization of the occurrence frequencies of dipeptide in the primary sequence.

● Information Entropy of CGR

The frequencies of all the grid in the chaos game representation is considered as a given distribution of a random variable.

2.2.4. Conjoint triad feature

The conjoint triad feature (CTF) representation is extracted from the primary sequence of a specific protein [30]. Firstly, the 20 native amino acids were clustered into 7 groups based on their dipole moments and the volume of their side chains. Then, each protein sequence can be represented by the 7-letter reduced alphabet. Thus, each component of CTF corresponds to the normalized occurrence frequency of the corresponding 3-mer in the sequence. That is to say that each protein sequence is represented by a 343 (7 × 7 × 7) dimensional vector.

2.2.5. Hilbert Huang transformation

Hilbert Huang Transformation (HHT) is indeed a rigorous and robust method for analyzing data from nonlinear and no stationary processes [31]. For its entirely empirical way, it cannot be limited by the uncertainty principle. Following this method, nonlinear and no stationary data can be mapped into time-frequency-energy space for feature extraction.

According to Hilbert Huang Transformation, the primary sequence is encoded with a hydrophobicity index [32]. Secondly, it is decomposed through empirical mode decomposition (EMD) process. The primary sequence is transformed into finite number of intrinsic mode functions (IMFs). Thirdly, the energy value is derived as the sum of the squares of IMFs. Finally, the mathematical expression of energy ratio is calculated. The 2th and 3th energy ratios (IMF2, IMF3) are retained. The first IMF just reflects the random error and the last one is just the trend of the processes, therefore both of them are excluded.

2.3. Random forest

Random Forest (RF) has been proposed by L. Breiman in 2001 [33]. It has been adopted in various scientific research fields, because of its outstanding prediction performance, especially in bioinformatics. Random Forest is an ensemble machine learning method based on a decision tree. Its basic idea is that the prediction result is voted by a certain number of decision trees. Each tree is independently constructed with a bootstrap sample of the training set. The random forest MATLAB toolbox which was applied in this research paper is available at http://code.google.com/p/randomforest-matlab/.

2.4. Performance evaluation

As is well-known, the following two methods are the most general and objective way to evaluate the effectiveness of a predictor: K-fold cross-validation and jackknife test. In this research, 5-fold cross validation and jackknife test were applied. In the cross validation test, the performance of the prediction method was usually measured by sensitivity (SE), specificity (SP), accuracy (ACC) and the Matthew's Correlation Coefficient (MCC) value.

3. Results

3.1. Comparison with DNA_prot

As mentioned in the former section, there are nine kinds of candidate features. Each given protein sequence can be converted into nine kinds of candidate features, i.e., AAC, CGR FD, Dipeptide, En_AAC, En_CGR,En_dipeptide, CTF and HHT.

In order to find the optimal combination of candidate features, a 5-fold cross validation test was adopted. The in-depth numerical experiments were executed based on the following procedure. Firstly, the preliminary numerical experiments are carried out for reducing the computation complexity. The combinations of nine features were randomly picked out to assess. The result of preliminary experiments shows that AAC was an important feature to identify the DNA-binding protein. So AAC was regarded as the required feature of the input vector. This trick cuts down the amount of calculation in half. Then, each combination of the candidate features was evaluated with the 5-fold cross

Table 1

Results of Comparison with DNA-Prot

Dataset	DNA-Prot		Our method	
	ACC	MCC	ACC	MCC
Training Dataset	0.8031	-	0.8157	0.5968
Test Dataset	0.8437	0.7000	0.8646	0.7503
Independent dataset1	0.8183	0.6400	0.8098	0.6197
Independent dataset2	0.9346	0.8300	0.9408	0.8495

Note: ACC means Accuracy; MCC represents Matthew's Correlation Coefficient.

Table 2

The occurrence times of eight features in top combinations

Features	FD	CGR	Dipeptide	En_CGR	En_AAC	En_dipeptide	CTF	HHT
Top 30	20	9	0	20	13	19	12	1
Top 50	29	25	0	30	26	29	23	5
Top 100	56	58	8	54	49	52	48	34

Note: FD represents Fractal Dimension; CGR means Chaos Game Representation; AAC represents Amino Acid Composition; En_ indicates the entropy calculated from the following normalized occurrence frequencies, for example, En_AAC means the entropy of Amino Acid Composition; CTF is Conjoint Triad Feature; Lastly, HHT means Hilbert Huang Transformation

validation test. The optimal combination, i.e., AAC, FD, En_AAC and En_CGR was selected according to the result of numerical experiments.

Taking the optimal combination as the input vector, the performance of this model surpassed the one of DNA-Prot. The better performance was achieved for the training dataset via the jackknife test. It was superior to DNA-Prot, for an independent dataset 1 and independent dataset 2. The sole deficiency was slightly lower accuracy than DNA-Prot for test dataset with the smallest size (92 DNA-binding proteins and 100 non DNA-binding ones). The limited size may lead to incorrect result. The results of comparison are listed in Table 1.

3.2. Analysis of feature selection

Nine kinds of candidate features were derived from primary sequence of protein. Except for AAC, which was regarded as an indispensable feature, the other eight candidate features should be confirmed whether they are effective ones. There are 2^8 (256) combinations which were evaluated with the 5-fold cross validation. It is far from being enough just considering the optimal combination. Hence, the occurrence times of the eight features in top combinations, which are listed in Table 2, were counted to measure their effectiveness.

4. Conclusions

Firstly, the results of the preliminary numerical simulations showed the prediction models, which contain amino acid composition, can achieve satisfactory performance. The former studies demonstrated that amino acid composition was a fundamental component for predicting DNA-binding protein.

Secondly, by analyzing the components of the optimal combination, it verifies that the fractal dimension and entropies of chaos game representation and amino acid composition are the promising quantitative indexes for bioinformatics research. They can excavate the information hidden in the primary sequence which may be ignored by amino acid composition. The occurrence times of eight candidate features in the top 100 and top 50 combinations show the efficiencies of fractal dimension, chaos game representation, three information entropies and conjoint triad feature. Among these six classes of features, the times in the top 30 combinations can also manifest that the concepts of fractal dimension and entropy are even more efficient and essential for a DNA-binding protein. Above all, the analysis of numerical simulation results can show the importance of the fractal dimension and entropies.

Finally, all of these nine candidate features can be derived from the primary sequence of protein without any extra knowledge. Additionally, this kind of method is able to gain the encouraging results. All of these prove that the sequence-based approaches have a strong versatility and effectiveness. Especially, the concepts of fractal dimension and information entropy are advantageous and essential feature indexes for identifying a DNA-binding protein.

Acknowledgment

Li Nana and Niu Xiaohui were supported by the National Natural Science Foundation of China (NSFC) (Grant No. 81202289). Changge Guan and Kun Yang were supported by the National Creative Innovation Plan of College Students of China (Grant No. 201310504038).

Reference

[1] N.M. Luscombe, S.E. Austin, H.M. Berman and J.M. Thornton, An overview of the structures of protein-DNA complexes, Genome Biology **1** (2000), 1-37.

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

[3] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne, The protein data bank, Nucleic Acids Research **28** (2000), 235-242.

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

[5] S. Ahmad and A. Sarai, Moment-based prediction of DNA-binding proteins, Journal of Molecular Biology **341** (2004), 65-71.

[6] E. Nordhoff, A.M. Krogsdam, H.F. Jorgensen, B.H. Kalliporlitis, B.F. Clark, P. Roepstorff and K. Kristiansen, Rapid identification of DNA-binding proteins by mass spectrometry, Nature Biotechnology **17** (1999), 884-888.

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

[8] M. Pellegrini-Calace and J.M. Thornton, Detecting DNA-binding helix–turn–helix structural motifs using sequence and structure information, Nucleic Acids Research **33** (2005), 2129-2140.

[9] N. Bhardwaj, R.E. Langlois, G. Zhao and H. Lu, Kernel-based machine learning protocol for predicting DNA-binding proteins, Nucleic Acids Research **33** (2005), 6486-6493.

[10] M. Kumar, M.M. Gromiha and G.P. Raghava, Identification of DNA-binding proteins using support vector machines and evolutionary profiles, BMC Bioinformatics **8** (2007), 463.

[11] R.F. Xu, J.Y. Zhou, B. Liu, L. Yao, Y.L. He, Q. Zou and X.L. Wang, enDNA-Prot: Identification of DNA-binding proteins by applying ensemble learning, Biomed Research International (2014), 294279.

[12] W.C. Lou, X.Q. Wang, F. Chen, Y.X. Chen, B. Jiang and H. Zhang, Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive bayes, PLoS One **9** (2014), e86703.

[13] W. Ka-Chun, C. Tak-Ming, P. Chengbin, L. Yue and Z. Zhaolei, DNA motif elucidation using belief propagation, Nucleic Acids Research **41** (2013), e153.

[14] L. Nanni and A. Lumini, Combing ontologies and dipeptide composition for predicting DNA-binding proteins, Amino Acids **34** (2008), 635-641.

[15] K.K. Kumar, G. Pugalenthi and P.N. Suganthan, DNA-Prot: Identification of DNA binding proteins from protein sequence information using random forest, Journal of Biomolecular Structure & Dynamics **26** (2009), 679-686.

[16] X. Shao, Y. Tian, L. Wu, Y. Wang, L. Jing and N. Deng, Predicting DNA- and RNA-binding proteins from sequences with kernel methods, Journal of Theoretical Biology **258** (2009), 289-293.

[17] Y. Fang, Y. Guo, Y. Feng and M. Li, Predicting DNA-binding proteins: Approached from Chou's pseudo amino acid composition and other specific sequence features, Amino Acids **34** (2008), 103-109.

[18] W.Z. Lin, J.A. Fang, X. Xiao and K.C. Chou, iDNA-Prot: Identification of DNA binding proteins using random forest with grey model, PLoS One **6** (2011), e24756.

[19] N. Bhardwaj and H. Lu, Residue-level prediction of DNA-binding sites and its application on DNA-binding protein predictions, FEBS Letters **581** (2007), 1058-1066.

[20] G. Nimrod, A. Szilagyi, C. Leslie and N. Ben-Tal, Identification of DNA-binding proteins using structural, electrostatic and evolutionary features, Journal of Molecular Biology **387** (2009), 1040-1053.

[21] X. Yu, J. Cao, Y. Cai, T. Shi and Y. Li, Predicting rRNA-, RNA-, and DNA binding proteins from primary structure with support vector machines, Journal of Theoretical Biology **240** (2006), 175-184.

[22] L. Nanni and A. Lumini, An ensemble of reduced alphabets with protein encoding based on grouped weight for predicting DNA-binding proteins, Amino Acids **36** (2009), 167-175.

[23] K.S. Leung, K.C. Wong, T.M. Chan, M.H. Wong, K.H. Lee, C.K. Lau Terrence and K.W. Tsui Stephen, Discovering protein–DNA binding sequence patterns using association rule mining, Nucleic Acids Research **38** (2010), 6324-6337.

[24] K.C. Wong, C.B. Peng, M.H. Wong and K.S. Leung, Generalizing and learning protein-DNA binding sequence representations by an evolutionary algorithm, Soft Computing **15** (2011), 1631-1642.

[25] H.J. Jeffrey, Chaos game representation of gene structure, Nucleic Acids Research **18** (1990), 2163-2170.

[26] S. Basu, A. Pan, C. Dutta and J. Das, Chaos game representation of proteins, Molecular and Modelling **15** (1997), 279-289.

[27] B.B. Mandelbrot, The Fractal Geometry of Nature, Freeman, San Francisco, 1982.

[28] X.H. Niu, X.H. Hu, F. Shi and J.B. Xia, Predicting DNA binding proteins using support vector machine with hybrid fractal features, Journal of Theoretical Biology **343** (2014), 186-192.

[29] X.H. Niu, N.N. Li, J.B. Xia, D.Y. Chen, Y.H. Peng, Y. Xiao, W.Q. Wei, D.M. Wang and Z.Z. Wang, Using the concept of Chou's pseudo amino acid composition to predict protein solubility: An approach with entropies in information theory, Journal of Theoretical Biology **33** (2013), 211-217.

[30] K.J. Falconer, Techniques in Fractal Geometry, Wiley, New York, 1997.

[31] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li and H. Jiang, Predicting protein-protein interactions based only on sequences information, Proceedings of the National Academy Sciences USA **104** (2007), 4337-4341.

[32] F. Shi, Q.J. Chen and N.N. Li, Hilbert Huang transform for predicting apoptosis proteins types, The 1st International Conference on Bioinformatics and Biomedical Engineering, Wuhan, 2007, 102-106.

[33] L. Breiman, Random forests, Machine Learning **45** (2001), 5-32.