PROCEEDINGS OF SCIENCE

iDNA-BiProt: Predicting DNA-binding Proteins via Feature Extraction and Fuzzy K Neighbor Algorithm

Mengjuan Hui[1]
Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen , 333046, China
E-mail: huimengjuan@163.com

Xuan Xiao[2]
Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen, 333046, China;
Information School, Zhejiang Textile & Fashion College, Ningbo, 315211, China
E-mail: jdzxiaoxuan@163.com

Zi Liu[3]
Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen, 333046, China
E-mail: liuzi189836@163.com

DNA-binding proteins play a pivotal role in most of the biological reaction. They are defined as the amino acids directly involved DNA packaging, replication and activities. The knowledge of these properties are important for understanding the function of DNA-binding proteins (DNA-BPs). In the post-genomic era when the grown the of protein sequences appeared explosive, we urgently need a kind of method that can identify the DNA-binding protein in accordance with the sequences information in a fast and reliable manner. In order to resolve this problem, our team put forward a novel approach named iDNA-BiProt. In the process of the system, the protein sequence information was extracted by wavelets transforms and amino acid physicochemical properties. The overall accuracy obtained by Fuzzy K neighbor algorithm is 82.33%. It's expected this method can be an essential tool for the basic of drug design and further research.

CENet2015
12-13 September 2015
Shanghai, China

[1]Speaker
[2]Corresponding Author
[3]This work was supported by National Nature Science Foundation of China (No.31260273,31560316), the LuoDi plan of the Department of Education of JiangXi Province (KJLD12083) and the Graduated innovation found of Jingdezhen ceramic institute (JYC201427). The funders played no role in the study design, data collection and analysis, decision to publish the paper or preparation of the manuscript.

PoS(CENet2015)003

1. Introduction

DNA-BPs are closely related to the control of cellular processes and play an important role in the epigenetic gene regulation in both life development and disease formation, hence they are considered as a major epigenetic mark responsible for silencing of the cell-fate regulators. In mammals, DNA-binding proteins help regulate the gene expression, DNA replication and DNA damage repair [1]. DNA-binding proteins control the gene activity, also called transcription factors, binding to DNA and acting as switches, either activating or repressing transcription of particular gene[2, 3]; therefore, the knowledge of DNA-binding proteins is vitally important for both essential biomedicine research and the utility of drug development. In fact, during the last decade, some efforts have been made to use computational approach to identify the DNA-binding proteins. For instance, Fang incorporated Chou's PseAAC with other sequence information to predict DNA-BPs. Lin enveloped a method, called iDNA-Prot by mixing extracted traits information to the normal PSeAAC (pseudo amino acid composition) in order to foresee the DNA-BPs. The algorithm is a random forest. Lou et al. proposed a means, called DBPPred on the basis of the hybrid feature selection of DNA-binding proteins[4]. Liu et al. developed a novel predictor, named iDNA-Prot|dis, that the DNA-binding proteins sequence can be formulated by a usual PseAAC vector which combines the amino acid simplified alphabet profile and the distance-pair coupling information[5]. It should be shown that most of usual existing predictors spent a lot of time for a simply single prediction. As an efficient output tool in managing a large quantity of protein sequences, if the predictor costs little time in dealing with the query object, it will become a much more useful output tool. The present document was motivated to explore a newfangled forecaster in this regard by addressing the drawback mentioned above. A flowchart is a direct picture to state the flow of the **iDNA-BiProt** predictor to be observed as below in **Fig** 1**.**

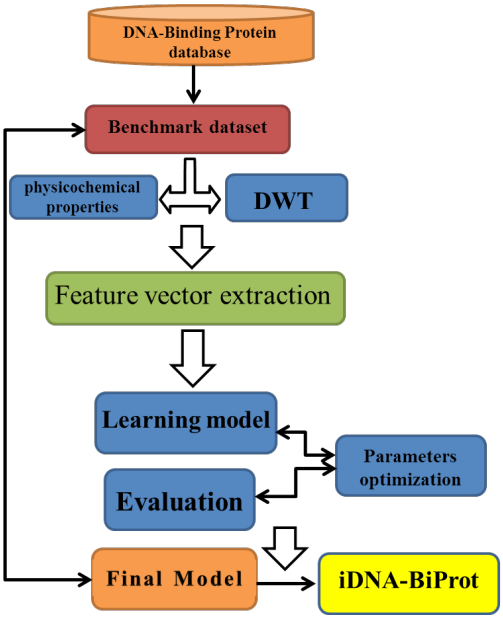

Figure 1: Flowchart of the Predictor iDNA-BiProt

2. Material and Method

2.1 Benchmark Datasets

The basic set $\mathbb{S}$ for the DNA-BPs and non DNA-BPs was taken from Liu et al.[5]. It possesses 525 positive samples of DNA-bind proteins and 550 negative samples of no DNA-binding proteins, which can be expressed as

$$\mathbb{S}=\mathbb{S}^{+}\cup\mathbb{S}^{-} \tag{2.1}$$

where $\cup$ stands for 'union'; the set $\mathbb{S}^{+}$ represents the positive sample and $\mathbb{S}^{-}$ represents the negative sample.

2.2 Representation of Protein Chains

In order to invent a powerful predictor for identifying DNA-BPs as to the sequence information, the crucial steps of formulating the peptide chains are a kind of utile and practical mathematical expressions that can describe the sequence information veridically. In order to achieve this goal, the PseAAC was proposed to avert the missing of essential sequential information.

With the pseudo amino acid composition that a query protein sequence can be described as a discrete numbers, a sequence P can be stated as:

$$P=[\Psi_1,\Psi_2,\cdots,\Psi\mu,\cdots,\Psi_\Omega]^T \tag{2.2}$$

where T stands for transpose operation and the symbol Ω is the subscript of Ψ_1, Ψ_2,…Ψ_Ω. To squeeze the information of the protein, **P** decides the value of Ω together with Ψ_1, Ψ_2,…Ψ_Ω while acquiring much more vital information, we do not feel hesitated to use physicochemical property and the Discrete Wavelet Transform (DWT).

2.2.1 Physicochemical

Each of the residues in relation to protein has many physicochemical properties. In this study, the following seven physicochemical properties have been adopted: (1)Ξ^1:hydrophobicity, (2) Ξ^2:hydrophicility, (3)Ξ^3:volumes of side chains of amino acids, (4) Ξ^4: polarity, (5) Ξ^5: polarizability, (6) Ξ^6: solvent accessible surface area (SASA) and (7) Ξ^7: net charge index(NCI) of the side chains of amino acids. Respectively, the original values of the seven descriptors for each amino acid are listed in **Table 1**. Thus, in accordance with the seven Ξ physicochemical properties, the protein P of **Eq.2.2** is able to be formulated with a 7×L physicochemical property matrix given as follows

$$P=\begin{bmatrix}
\Xi^1(R_1) & \Xi^1(R_2) & \cdots & \Xi^1(R_L)\\
\Xi^2(R_1) & \Xi^2(R_2) & \cdots & \Xi^2(R_L)\\
\Xi^3(R_1) & \Xi^3(R_2) & \cdots & \Xi^3(R_L)\\
\Xi^4(R_1) & \Xi^4(R_2) & \cdots & \Xi^4(R_L)\\
\Xi^5(R_1) & \Xi^5(R_2) & \cdots & \Xi^5(R_L)\\
\Xi^6(R_1) & \Xi^6(R_2) & \cdots & \Xi^6(R_L)\\
\Xi^7(R_1) & \Xi^7(R_2) & \cdots & \Xi^7(R_L)
\end{bmatrix} \tag{2.3}$$

where $\Xi^i(R_j)$ is the value of $\Xi^i(i=1,2,3,4,5,6,7)$ for residue $R_j(j=1,2,3,\cdots,L)$.

Before substituting these physiochemical values into **Eq. 2.3**, the initial values of the seven descriptors for the basic 20 native amino acids are shown in **Table 1** for $\Xi^i(R_j)(i=1,2,3,4,5,6,7;j=1,2,3,\cdots,L)$, they are all attached to a stand conversion, which can be formulated as

$$\Xi^i(R_i)=\frac{\Xi^i(R_i)-\langle\Xi^i\rangle}{SD(\Xi^i)} \tag{2.4}$$

where '< >' represents the mean values of amino acids and SD represents the seven different standard deviations. The elements of **Table 2** refer to the value of $\Xi^i(R_j)(i=1,2,3,4,5,6,7;j=1,2,3,\cdots,L)$ obtain via **Eq. 2.4** from **Table 1**

encoding	Ξ^1	Ξ^2	Ξ^3	Ξ^4	Ξ^5	Ξ^6	Ξ^7
A	0.620	-0.500	27.500	8.100	0.046	1.181	7.187×10^{-3}
C	0.290	-1.000	44.600	5.500	0.128	1.461	-3.661×10^{-2}
D	-0.900	3.000	40.000	13.000	0.105	1.587	-2.382×10^{-2}
E	-0.740	3.000	62.000	12.300	0.151	1.862	6.802×10^{-3}
F	1.190	-2.500	115.500	5.200	0.290	2.228	3.755×10^{-2}
G	0.480	0.000	0.000	9.000	0.000	0.881	1.791×10^{-1}
H	-0.400	-0.500	79.000	10.400	0.230	2.025	-1.069×10^{-2}
I	1.380	-1.800	93.500	5.200	0.186	1.810	2.163×10^{-2}
K	-1.500	3.000	100.000	11.300	0.219	2.258	1.771×10^{-2}
L	1.060	-1.800	93.500	4.900	0.186	1.931	5.167×10^{-2}
M	0.640	-1.300	94.100	5.700	0.221	2.034	2.683×10^{-3}
N	-0.780	2.000	58.700	11.600	0.134	1.655	5.392×10^{-3}
P	0.120	0.000	41.900	8.000	0.131	1.468	2.395×10^{-1}
Q	-0.850	0.200	80.700	10.500	0.180	1.932	4.921×10^{-2}
R	-2.530	3.000	105.000	10.500	0.291	2.560	4.359×10^{-2}
S	-0.180	0.300	29.300	9.200	0.062	1.298	4.627×10^{-3}
T	-0.050	-0.400	51.300	8.600	0.108	1.525	3.352×10^{-3}
V	1.080	-1.500	71.500	5.900	0.140	1.645	5.700×10^{-2}
W	0.810	-3.400	145.500	5.400	0.409	2.663	3.798×10^{-2}
Y	0.260	-2.300	117.300	6.200	0.298	2.368	2.360×10^{-2}

Table1: Initial Values for Each Amino Acid of the Seven Physicochemical Properties

In this sense, a query protein sequence with L residues can be formulated as a $7\times L$ physicochemical property matrix, as given in **Eq.2.3.**

encoding	Ξ^1	Ξ^2	Ξ^3	Ξ^4	Ξ^5	Ξ^6	Ξ^7
A	0.620	-0.189	-1.239	-0.084	-1.330	-1.382	-0.441
C	0.290	-0.440	-0.768	-1.050	-0.489	-0.775	-1.115
D	-0.900	1.573	-0.895	1.738	-0.725	-0.502	-0.918
E	-0.740	1.573	-0.290	1.477	-0.254	0.094	-0.447
F	1.190	-1.195	1.181	-1.161	1.171	0.887	0.026
G	0.480	0.063	-1.995	0.251	-1.801	-2.032	2.202
H	-0.400	-0.189	0.178	0.771	0.556	0.447	-0.716
I	1.380	-0.843	0.576	-1.161	0.105	-0.019	-0.219
K	-1.500	1.573	0.755	1.106	0.443	0.952	-0.279
L	1.060	-0.843	0.576	-1.273	0.105	0.244	0.243
M	0.640	-0.591	0.593	-0.976	0.464	0.467	-0.510
N	-0.780	1.070	-0.381	1.217	-0.428	-0.355	-0.469
P	0.120	0.063	-0.843	-0.121	-0.459	-0.760	3.132
Q	-0.850	0.164	0.224	0.808	0.044	0.246	0.205
R	-2.531	1.573	0.892	0.808	1.181	1.607	0.119
S	-0.180	0.214	-1.189	0.325	-1.166	-1.128	-0.481
T	-0.050	-0.138	-0.584	0.102	-0.694	-0.636	-0.500
V	1.080	-0.692	-0.029	-0.901	-0.366	-0.376	0.325
W	0.810	-1.648	2.006	-1.087	2.390	1.830	0.032
Y	0.260	-1.095	1.231	-0.790	1.253	1.191	-0.189

Table 2: Normalized Values for Seven Physicochemical Properties for Each Amino Acid

2.2.2 Discrete Wavelet Transform (DWT)

The DWT analysis can resolve the protein sequence into different dilation coefficients and then get rid of the Interference elements; therefore, it is capable of providing people with partial formation of sequences that can much effectively embody the sequential results. When DWT is used on any of the seven numerical series for protein $\mathbf{P}$ (cf. **Eq.2.2**), we can view it as a discrete time series with the 1st residue as $t=1$, the 2nd residue as $t=2$, and so forth. The discrete time series thus obtained is input into a high-pass filter and a low-pass filter. The coefficients thus obtained can be approximately used for the signal's high scale and low frequency components. As a matter of fact, such transform is to be used recursively on the low pass series with the Mallat algorithm[6] until the anticipated results of iterations are achieved.

In this paper, the decomposition level $\lambda=4$ has been selected to represent a sample, which is similar to the treatment[7]. Accordingly, we can obtain $(4+1)=5$ sub-bands when

the discrete series **P** is decomposed by DWT with level $\lambda = 4$. Each of the five sub-bands has four coefficients: (1) α_j, the max of the wavelet coefficients in the j-th sub-band; (2) β_j the mean wavelet coefficients in the j-th subband; (3) γ_j the min wavelet coefficients in the j-th sub-band; (4) δ_j the wavelet coefficient's standard deviation in the j-th sub-band $(j = 1,\ 2,\ \cdots, 5)$. Thus, in a way quite to the treatment[8], each of the components in **Eq.2.2**, we can get a feature vector $\Omega = 5 \times 4 = 20$ components by using each of the seven physicochemical properties of **Eq.2.3;** in other words, we have seven different modes of PseAAC as given below

$$P^{(k)} = \left[\Psi_1^{(k)} \quad \Psi_2^{(k)} \quad \Psi_3^{(k)} \quad \cdots \quad \Psi_\mu^{(k)} \quad \cdots \quad \Psi_{20}^{(k)} \right] \quad (k = 1,2,\cdots, 7) \tag{2.5}$$

Where

$$\Psi_\mu^k = \begin{cases} \alpha_\mu & when\ 1 \leqslant \mu \leqslant 5 \\ \beta_\mu & when\ 6 \leqslant \mu \leqslant 10 \\ \gamma_\mu & when\ 11 \leqslant \mu \leqslant 15 \\ \delta_\mu & when\ 16 \leqslant \mu \leqslant 20 \end{cases} \tag{2.6}$$

Note that when using each of the seven different physicochemical features (cf. Eq.2.3) in turn, we can generate seven different PseAAC vectors to represent the same protein pair, as formulated by

$$P^k = \begin{cases} hydrophobicity & k = 1 \\ hydrophicility & k = 2 \\ side-chain\ volume & k = 3 \\ polarity & k = 4 \\ polarizability & k = 5 \\ solvent-accessible\ surface & k = 6 \\ side-chain\ net\ charge & k = 7 \end{cases} \tag{2.7}$$

3. Learning Algorithm

The Fuzzy KNN algorithm is one of the meaningful classifiers for classification prediction. According to the algorithm rule, given a sample unknown of the labels, their labels are distributed the same the labels of their KNN neighbors in the traindata.

The above algorithm is a member of the KNN family. In this study, the following equation has been used for classification. Suppose $P_1, P_2, ..., P_N$ be the vector group which represents a proteins sequence P in the training dataset divided into two types: C_1 and C_2, where C_1 represents the DNA-binding protein and C_2 represents non DNA-binding protein. Consequently, given a protein sequence P, the value of the fuzzy degree of membership for the i-th class can be formulated as below:

$$\mu_i(p) = \frac{\Sigma_{j=1}^{K} \mu_i(P, P_j) d(P_j)^{\frac{-2}{(m-1)}}}{\Sigma_{j=1}^{K} d(P, P_j)^{\frac{-2}{(m-1)}}} \tag{3.1}$$

The letter K represents the amount of the neighbors, for a protein P belong to i-th class the $\mu_i(P_j)$ is fuzzy-valued, $d(P, P_j)$ is the distance between these inquired protein sample P its nearest neighbors in the train data; and m(>1) is the heavy weight. There are all kinds of

measurements can be chosen as $d(P,P_j)$, for example, the hamming distance, the Euclidean distance and the Mahalanobis distance. In this article, we select the Euclidean distance as $d(P,P_j)$. The parameters *K* and *m* will be given later. As t o a protein sequence, in order to calculate all the values of the members, the query protein belong to the highest value of $\mu_i(P)$ with its identify label shown as :

$$\mathfrak{M}_u = \arg\max_i \{\mu_i(P)\} \tag{3.2}$$

4. Discussion and Result

For the sake of predicting the potential DNA-binding protein and improving the accuracy, we've used the known DNA-binding protein as the train dataset. The predictor has been developed via the above procedures.

In term of statistical forecast, the jackknife validation independent dataset test and subsampling (five ,seven, ten fold) are often applied to inspect the performance properties of a designed predictor about its efficacy in real application; however, as confirmed and demonstrated by some experts within three verification methods, only the jackknife verification has yielded an accurate result. Thus, the jackknife verification method will be implemented on the basic dataset.

The jackknife validation rate achieved by iDNA-BiProt for the DNA-binding protein predictor system is given in Table 3, respectively. The 2D searching to optimize the overall rate gets the value of K and m utility in **Eq.3.1**. The result was obtained in **Fig.2**, through which we've found when m=1.21and k= 10 the predictor reached its optimized status.

As is known to all in the **Table 3**, the accuracy achieved by iDNA-BiProt for the DNA-binding predictor system was 82.33% in the benchmark dataset; at meanwhile, we can see that the corresponding MCC were 64.65%, (Sn=83.17%, Sp=81.58%).

In order to further approve its power, let's compare **iDNA-BiProt** with the existing predictor in this area. The best way to compare them is though practical application. Let's compare the accuracy in the benchmark dataset. Here we've compared **iDNA-BiProt** with the two predictor **iDNA-Prot|dis** [5]and **iDNA-Prot**[9] on the basis of the sequence predictor. As we can see from the **Table 3**, the accuracy of the predictor on the dataset achieved by **iDNA-BiProt** is remarkably higher than those by its counterparts. These results have clearly indicated that **iDNA-BiProt** is superior to its counterparts in predicting the DNA-binding protein.

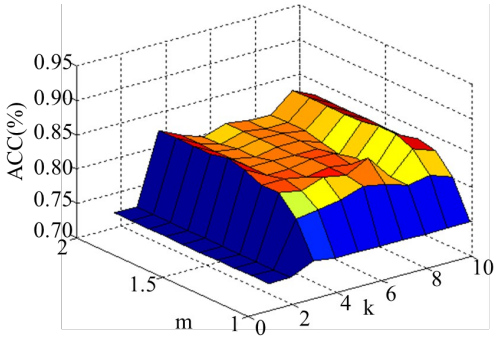

Figure 2: The 3D graph to show the success rate by jackknife test were different values of m and K in the FKNN. The results were obtained for the independent testing prediction.

Predictor	Acc(%)	MCC(%)	Sn(%)	Sp(%)
iDNA-Prot|dis[a]	0.7730	0.54	0.7940	0.7527
iDNA-Prot[b]	0.8019	0.6043	0.7955	0.8080
iDNA-BiProt	0.8233	0.6465	0.8317	0.8158

[a] From [5].
[b] From [9].

Table 3: Comparison of iDNA-BiProt with the Existing Predictor

5. Conclusion

In order to timely acquire the information of the iDNA-binding protein in DNA sequence, it is important to carry out in-depth study of DNA function and develop new drug. In my manuscript, we developed a novel method for the forecast of DNA-BPs via the comprising information from Chemical and Physical properties of DNA-proteins and nucleotides composition. The results were very promising and it is anticipated that our predictor should be also used to solve many other genome research problems.

References

[1] Akinori Sarai, Hidetoshi Kono. *Protein-DNA recognition patterns and predictions*.[J]. Annual Review of Biophysics & Biomolecular Structure, 2005, 34(1):379-398.

[2] Jennifer M. Hinerman1,John David Dignam ,Timothy C. Mueser1. *Models For The Binary Complex Of Bacteriophage T4 Gp59 Helicase Loading Protein: Gp32 Single-Stranded Dna-Binding Protein And Ternary Complex With Pseudo-Y Junction Dna.*[J]. Journal of Biological Chemistry, 2012, 287(22):18608-18617.

[3] Oscar Almarza, Daniel Nú?ez, Hector Toledo. *The DNA‐Binding Protein HU has a Regulatory Role in the Acid Stress Response Mechanism in Helicobacter pylori*[J]. Helicobacter, 2015, 20(1):29-40.

[4] Wang-Chao Lou, Xiao-Qing Wang, Fan Chen, Yi-Xiao Chen, Bo Jiang, Hua Zhang. *Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes.*[J]. Plos One, 2014, 9(1):e86703-e86703.

[5] Bin Liu, Jing-Hao Xu, Xun, Lan, Rui-Feng Xu, Ji-Yun Zhou, Xiao-Long Wang, Kuo-Chen Chou. *iDNA-Prot| dis: Identifying DNA-Binding Proteins by Incorporating Amino Acid Distance-Pairs and Reduced Alphabet Profile into the General Pseudo Amino Acid Composition*[J]. PloS one 2014;9(9):e106691-e106691.

[6] Stephane Mallat. *A Wavelet Tour of Signal Processing. 3 edition*[J]. Elsevier Ltd Oxford, 2008:xxii+805.

[7] Jian-Hua Jia, Zi Liu, Xuan Xiao, Bing-Xiang Liu, Kuo-Chen Chou. i*PPI-Esml: An ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC[J]*. Journal of theoretical biology 2015;377:47-56.

[8] Jian-Ding Qiu, Ru-Ping Lang, Xiao-Yong Zou,Jin-Yuan Mo. P*rediction of protein secondary structure based on continuous wavelet transform.*[J]. Talanta, 2003, 61(3):285-293.

[9] Wei-Zhong Lin,Jian-An Fang, Xuan Xiao, Kuo-Chen Chou. *iDNA-Prot: identification of DNA binding proteins using random forest with grey model*[J]. PLoS One 2011;6(9):e24756-e24756.