Diving into DNA-Protein Interaction with Machine Learning

Jun Wang

May 10, 2016

1 Introduction

1.1 Background

DNA-protein interaction is an important component for gene regulation system. Finding the interaction sites on protein and DNA molecules would be a critical step for drug design and researches on the metabolic network [11]. Nowadays, there are many different ways to investigate the binding sites, including experimental identification from lab, analysis from 3-d structure of the molecules, or inferrence from some known data such as amino acid sequences[7].

However, there are different disadvantages for these methods. For example,experimental identification is very expensive and relatively slow in practice. Especially, when dealing with a large amount of proteins and DNA molecules, it is not a good choice in fact. Analysis with the 3-d structure is a good way to determine the DNA-protein binding residues. But until now, it is still not easy to obtain the accurate 3-d structure of molecules for most proteins. Thus, inferring the DNA-protein binding sites from the protein sequences is a relatively more promising way, especially with the development of the machine learning theory.

1.2 Previous Research

Usually, there are at least 5 models available for predicting the binding sites on the protein molecule: Support Vector Machine(SVM), Artificial Neural Network(ANN), Decision Tree, Random Forest, Naive Bayes Model(NBM)[8]. For example, Wenyi developed an application for predicting the binding sites of protein based on the SVM[1]. Yan utilized the Naive Bayes Model to predict the binding sites with an accuracy equal to around 77%[12].

The features utilized for predicting the binding sites are also different in different research group. Jones utilized the electrostatic potential, accessible surface area, etc to predict the binding sites[3], while Yan only used the neighboring residue sequence to predict the binding sites[12]. In fact, Ofran indicated that the prediction based on the residue sequence is highly reliable in his research[7].

These previous research based on different models and different features provides helpful advice for the implementation in this final project.

PDB code	Protein name	Resolution	Residue overlap patch ranked 1	Random prediction value	Number of residues in DNA-binding interface
1qnaA	Transcription initiator factor TFIID-1	1.80	10	0.04	42
1mjoB	Methionine repressor	2.10	10	0.09	17
1d02A	Restriction endonuclease MUNI	1.70	10	0.10	27
1bg1A	STAT3β	2.25	10	0.03	23
1gdtA	Gamma-delta resolvase	3.00	10	0.15	33
1hcrA	HIN recombinase	1.80	10	0.40	26
1ewnA	AAG DNA repair glycosylase	2.10	9	0.06	25
1qpzA	Purine repressor	2.50	9	0.06	29
1fokA	Restriction endonuclease FOKI	2.80	9	0.09	67
1am9A	Sterol regulatory element binding protein 1A	2.30	9	0.11	17
1azpA	SAC7D	1.60	9	0.14	20
1dp7P	RFX-DBD	1.50	9	0.14	21
1a73A	Endonuclease I	1.80	9	0.15	36
1dctA	DNA (cytosine-5) methylase	2.80	9	0.15	46
1dmuA	Restriction endonuclease BGLI	2.20	9	0.15	50
1vasA	Endonuclease V	2.75	9	0.16	43
1bp7A	Endonuclease I-CREI	3.00	9	0.21	47
1tupB	Tumour supressor P53	2.20	8	0.06	18
2bop	E2 DNA-binding domain	1.70	8	0.08	23
1crxA	Cre recombinase	2.40	8	0.10	65
1dfmA	Restriction endonuclease BGII	1.50	8	0.14	45
3htsB	Heat shock transcription factor	1.75	8	0.14	21
1gd2E	BZIP transcription factor RAPI	2.00	8	0.15	15
1qpiA	Tetracycline repressor	2.50	7	0.02	23
1eqzA	Histone H2A	2.50	7	0.03	27
1emhA	Uracil-DNA glycosylase	1.80	7	0.06	20
1qrvA	Endonulcease V	2.20	7	0.06	27
2irfJ	Interferon regulatory factor-2	2.20	7	0.10	24
6mhtA	HHA1 methyltransferase	2.05	7	0.10	40
3pviA	Endonulcease PVUII	1.59	7	0.12	33
1bdtA	Arc transcription regulator	2.50	7	0.14	15
1ecrA	Replication terminator protein (TUS)	2.70	7	0.15	69
1pdnC	PRD paired domain	2.50	7	0.15	36
1b3tA	EBNA-1 nuclear protein	2.20	7	0.19	34
1ignA	RAP1	2.25	7	0.19	56
1au7A	PIT-1 POU domain	2.30	7	0.22	44
1sknP	SKN-1 transcription factor	2.50	7	0.25	21
1a1hA	QGSR zinc finger	1.60	7	0.30	36
1hwtC	HAP1	2.50	6	0.00	13
1dizA	3-Methladenine DNA glycosylase	2.50	6	0.03	27
1a36A	DNA topisomerase	2.80	5	0.01	71
1qumA	Endonuclease IV	1.55	5	0.02	31
2cgpA	Catabolic gene activator protein	2.20	5	0.02	17
2hmi	HIV-1 reverse transcriptase	2.80	5	0.02	59
1xbrA	Transcription factor T domain	2.50	5	0.03	30
1eonA	Type II restriction enzyme ECORV	1.60	5	0.07	37
1a3qA	NF-KAPPA-B	2.10	4	0.00	21
1c9bA	Transcription factor IIB	2.65	4	0.02	22
1mhdA	SMAD MH1 domain	2.80	4	0.02	17
2bdpA	DNA polymerase I	1.80	4	0.04	65
1ihfA	Integration host factor	2.50	4	0.07	30
6croA	Lambda CRO	3.00	4	0.35	20
1lmb3	Lambda repressor	1.80	1	0.07	23
1tauA	DNA polymerase	3.00	0	0.02	46
1zqfA	DNA polymerase β	2.90	0	0.02	28
2dnjA	Deoxyribonuclease 1	2.00	0	0.02	22

Figure 1: 56 different types of DNA-protein complexes[3].

2 Method

2.1 Data

2.1.1 Obtaining Dataset

The dataset is directly used from Jones' research[3], which includes 56 different types DNA-protein complexes(See Figure 1). And all of them are downloaded from the online dataset NPIDB[6], which stores the information of many different DNA-protein complexes.

2.1.2 Selecting Features

After obtaining the dataset, it is also critical to determine the features for the instance in the models. Here, two different types of the features are considered: 1. the neighboring residue sequence. 2. the electrostatic potential of the residues in the neighboring residues.

X_0	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
0	1	0	1	1	0	0	0	1

Figure 2: The structure of the neighboring residue sequence. The middle residue with yellow color is the target residue.

<s>	<s>	<s>	<s>	X_4	X_5	X_6	X_7	X_8
-1	-1	-1	-1	1	0	0	0	1

X1	X2	X3	X4	X_4	</s>	</s>	</s>	</s>
0	1	0	0	1	-1	-1	-1	-1

Figure 3: The structure of the neighboring residue sequence at the beginning and the end of the protein sequence. $< s >$ and $< /s >$ are the fake residues.

Neighboring residue sequence

In previous research, many researchers implementing the model based on the sequence of the neighboring residues and the performance is very good[12][3][1][10]. Thus, it is reasonable to use the neighboring residue sequence as the critical features when predicting the binding sites of the protein molecules. The most proper number of neighboring residues used for predicting is not clear, but Yan indicated that the length of 9 for a neighboring residue sequence is proper in the Naive Bayes Model (NBM)[12]. Thus, the length of the neighboring residue sequence will be 9. In this neighboring residue sequence, the first four residues are the residues before the target residue and the last four residues are the residues after the target residue. The middle residue is the target residue itself(See Figure 2). However, in each protein molecule, at the beginning and the end of the residue sequence, there are no enough neighboring residues. Thus, it is necessary to add the fake residue at the beginning and the end of the residue(See Figure 3).

Electrostatic Potential

There are many ways to calculate the electrostatic potential. Jones used the CHARMM force field to calculate the relative charge of each atom in a single residue and added the relative charges of all the atoms within the residue together[3]. But Kauffman indicated that it is proper to simplify the electrostatic potential to -1, 0, or 1 according to the specific type of the residues[4](See Table 1). Here, in order to simplify the calculation, it is proper to use the Kauffman's method to determine the electrostatic potentials. 9 neighboring residues will be considered for the electrostatic potential here.

Discrete Value of electrostatic potential	residues
Positive	Arg, Lys, His
Negative	Asp, Glu
Neutral	All others

Table 1: The electrostatic potential of different residues[4].

2.1.3 Determining Tag

There are two ways to determine whether the residue is the binding site on the protein molecule. The first way is to calculate difference of the accessible surface area of the residue between the DNA-protein complex and the protein alone[3]. The second way is to compare the distance of the hydrogen bond or the water bridge with the threshold value such as 3.5 Angstrom[8][5][10].

Since the dataset NPIDB have already stores all the distance of the hydrogen bond and the water bridge within 3.7 Angstrom, it is very convenient to use the second method and set 3.7 Angstrom as the cutoff distance.

2.2 Model

Three different models were implemented for predicting the binding sites between protein and DNA on the residue sequences.

2.2.1 Naive Bayes Model

Naive Bayes Model(NBM) is a simple and efficient model with relatively high accuracy when treating with small dataset. It is widely used in some machine learning area such as Natural Language Processing for sequence labeling. Since predicting the binding residues on the protein molecules is also a type of sequence labeling problem, it is reasonable to utilize this model. In the introduction section of previous research, some researchers have implemented the relatively sophisticated model by naive bayes model.

The process of naive bayes model can be divided into three different stages:

Stage 1: Building the dictionary

This model is implemented according to the Bayess' equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where A and B are two different events:

- $P(A)$ and $P(B)$ are the probabilities that A and B happens independently.
- $P(A|B)$ is the probability that A will happen when observing B has happened. So is $P(B|A)$. These probabilities are called conditional probability.

Using Bayesian probability terminology, the equation above can be rephrased as:

$$posterior = \frac{prior \times likelihood}{evidence}$$

Since the probability of evidence are usually the constant for a specific problem, we can convert the equation to:

$$posterior \propto prior \times likelihood$$

Considering that all the features are independent in the naive bayes model, the final equation will be:

$$P(c|X = x_1x_2...x_n) = P(c)\prod_{i=1}^{n} P(x_i|c)$$

where c is the tag of the instance and X is a vector of n features (x_1, ... x_n).

All the probabilities in the equation above will be calculated according to the counts of the instance with the corresponding tag and the feature value. Thus, it is necessary to build several dictionaries to store the corresponding counts before calculating the probability.

Stage 2: Training

In most naive bayes classifier, the predict tag will be selected according to the prediction with the largest probability. However, since this is a binary classification problem, it is possible to utilize a more accurate way to predict the tag. A threshold value named θ is defined to determine whether the tag of the corresponding residue is positive or negative [12] below:

$$\frac{P(c = 1|X = x_1x_2...x_n)}{P(c = 0|X = x_1x_2...x_n)} = \frac{P(c = 1)\prod_{i=1}^{n} P(x_i|c = 1)}{P(c = 0)\prod_{i=1}^{n} P(x_i|c = 0)} > \theta$$

where c is the tag for the instance and X is the vector of features. When this ratio is larger than θ, the predicted tag will be 1, otherwise, the predicted tag will be 0.

It is obvious that θ is between 0 and 1. An iteration of θ from 0.01 to 1 with stepsize = 0.01 will run on the same naive bayes model. In the iteration, the correlation coefficient will be calculated every time. Since the correlation coefficient can be used to described the performance of the model, the θ value will be selected with the largest correlation coefficient within the whole iteration.

Stage 3: Predicting

According to the sequence of the selected residues of the instance, the corresponding probabilities can be calculated by referring the dictionary and determine the predicted tag comparing to the threshold θ.

2.2.2 Support Vector Machine

Besides the naive bayes model, another very useful model in sequence labeling is the support vector machine especially it is a binary classification problem here[8][9][1].

In Support Vector Machine(SVM), the model will build a hyperlane according to the vector of the feature values for each instance in the training set to divide the instance and predict its class. Thus, it is necessary to convert the features of each residue which are residue sequence into the float numbers. Some researchers indicated that using Postition-specific scoring matrix(PSSM) to represent the neighboring residue sequence is a promising approach[10][1]. Thus, the implementation of the svm approach can be divided into two stages:

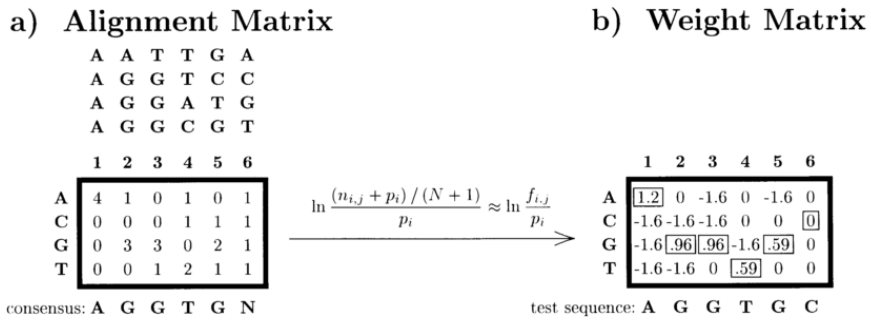

Figure 4: The calculation of PSSM of the DNA sequence with length 6[2].

Stage 1: Converting the feature values

PSSM represents the probabilities of the residues appeared at a specific position(See Figure 4). Thus, for a residue sequence with length N, it is a matrix of $N \times 20$ and each cell represents a corresponding probability of the residue x_i appeared at the j_{th} position. The corresponding probability should be smoothed and multiply with the corresponding probability of the residues in the whole dataset. The corresponding equations of the value in the cell x_{ij} is below[2]:

$$f_{ij} = \frac{n_{ij} + p_i}{\sum_{r=1}^{A} n_{rj} + 1}$$

where f_{ij} is the relative frequency of residue i at position j, n_{ij} s the occurences of residue i at position j, p_i is the prior residue probability of residue i, A is the size of dataset.

$$x_{ij} = ln(\frac{f_{ij}}{p_i})$$

where use the logarithmic value of probability to store in the cell.

After building the PSSM for the training set, we convert the neighboring residue sequence of each residue to a number sequence by this matrix. This number sequence represents a vector of features for a instance.

Stage 2: Training and Predicting

After converting the residue sequence into number sequence, the sklearn module of python was utilized to call the SVM implementation. Using the SVM module within sklearn to train the dataset and predict the dataset.

2.2.3 Neural Network

Neural Network(NN) approach is another way to predict the binding sites on protein molecules. Comparing to the support vector machine (SVM), it can investigate the interaction between the features during the training[8] and researchers have implemented with this model successfully[10].

The preprocess of the data for the NN are very similar to that of the SVM. After converting the neighboring residue sequence of the instance into the number sequence by the PSSM, the new feature vector is used to feed the neural network implemented by the python module named tensorflow developed by Google. Thus, the whole process of this model is very similar to the SVM.

Model	Average Test Accuracy
Naive Bayes Model without electrostatic potential	80.41%
Naive Bayes Model with electrostatic potential	78.07%
Support Vector Machine without electrostatic potential	90.07%
Support Vector Machine with electrostatic potential	90.07%
Neural Network without electrostatic potential	84.35%
Neural Network with electrostatic potential	84.35%

Table 2: The average test accuracy for each model.

Model1	Model2	t-value	p-value
NBM without ep	NBM with ep	2.41	0.018
NBM without ep	SVM without ep	-8.95	9.74e-15
NBM without ep	NN without ep	-3.29	0.0013
SVM without ep	NN without ep	4.48	1.81e-05
NBM with ep	SVM with ep	-11.26	5.06e-20
NBM with ep	NN with ep	-5.30	6.14e-07
SVM with ep	NN with ep	4.48	1.81e-05

Table 3: The result of t-test between different models. NBM, SVM, and NN are Naive Bayes Model, Support Vector Machine, and Neural Network respectively. ep represents electrostatic potential.

3 Result

3.1 Overview

During the implementation, all the models are tested with leave-one-out cross validation and the result is below (Table 2):

From Table 2, it seems that:

1. The support vector machine has the best performance when predicting the binding sites on the protein molecules, while the performance of the naive bayes model is relatively worst comparing to other two models.

2. It doesn't affect the performance of the model too much when adding the electrostatic potential as the features into the model.

3.2 t-test

Though some conclusions seems obvious in Table 2, it is better to run the t-test to comparing different models. The t-test result is below Table 3: From the table Table 3, for a confidence of 99%, it is concluded that:

1. The p-value between NBM with and without electrostatic potential is larger than 0.01, thus the average accuracy between these two models have no significant difference. What's more, the average accuracy of SVM or NN with and without electrostatic potential are exactly the same.

2. The p-value between different models when both taking or not taking electrostatic potential are smaller than 0.01, thus they have significant difference.

3.3 Discussion

From the two tables above (Table 2 and Table 3), there are at least two conclusions: 1. Electrostatic potential didn't affect the performance of the model significantly. What's more, it doesn't affect the result at all in SVM and NN. 2. SVM has the best performance comparing to NN and NBM.

For the first conclusion, it is reasonable. When converting the neighboring residue sequence of the instance into the float number sequence for SVM and NN, the corresponding values obtained from PSSM indicated the characteristic of the sequence of the neighboring residues. On the other hand, according to the previous description of obtaining the value of the electrostatic potential, the value of the electrostatic potential is determined by the type of residue completely. Thus, different order of the residues in the sequence will cause the different order of the electrostatic potentials of the a single instance. In other words, the residue sequence and the electrostatic potential sequence are nearly equivalent. Thus, adding the sequence of the electrostatic potential will not improve the performance of the model based on the residue sequence at all. For NBM, since the model is based on the probability of the residues instead of the sequence of the residues, thus adding the electrostatic potential will affect the performance. To be more specific, there are twenty types of residues, while the number of types of the electrostatic potential is only three (-1, 0, 1), so the distribution of the probability of these two classes of parameters are not equivalent. Thus, when implementing the model for predicting the binding sites based on the residue sequence, it is not necessary to add the electrostatic potential as the feature at all.

For the second conclusions, it is also reasonable. NBM is based on the probability of the residues, so it doesn't considering the order of the residues in the neighboring residue sequence. Thus, it is reasonable that the performance of NBM is worse than SVM. For NN, since the model implemented here is the simplest one and the iteration is not very large. However, the performance of NN is affected by these two properties significantly, thus the performance of NN here is not better than that of the SVM. But with the improvement of the NN and increase of the iteration, the test accuracy of the NN should be higher than SVM finally.

4 Conclusion and Future work

4.1 Conclusion

From the previous discussion, it indicated two conclusions from the experiment and the analysis:

1. Since the electrostatic potential is determined by the type of the residue, adding the electrostatic potential as the features when implementing the model based on the residue sequence is unnecessary.

2. Since the Support Vector Machine(SVM) and the Neural Network(NN) predict the tag of the instance based on the sequence of the residues, the performance of these two models are better than that of the Naive Bayes Model(NBM) which is implemented based on the probability of the residues in the residue sequence.

4.2 Future work

Based on the previous conclusions, it is possible to improve the model from several aspects below:

1. Instead of using the electrostatic potential as the additional features in the sequence-based model, the structural-based features like the Solvent Accessible Surface Area (SASA), Structural Neighbors might be more proper additional features in the sequence-based model.

2. Try different kernal functions in the Support Vector Machine (SVM), because some kernal functions might generate much more accurate hyperlane in the SVM.

3. Improve the Neural Network (NN) and increase the time of iterations, which might improve the performance of the NN very well.

References

[1] Wen Yi Chu, Yu Feng Huang, Chun Chin Huang, Yi Sheng Cheng, Chien Kang Huang, and Yen Jen Oyang. ProteDNA: A sequence-based predictor of sequence-specific DNA-binding residues in transcription factors. *Nucleic Acids Research*, 37(SUPPL. 2):396–401, 2009.

[2] Gerald Z Hertz and Gary D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics (Oxford, England)*, 15(7-8):563–77, 1999.

[3] Susan Jones, Hugh P. Shanahan, Helen M. Berman, and Janet M. Thornton. Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucleic Acids Research*, 31(24):7189–7198, 2003.

[4] Chris Kauffman and George Karypis. An analysis of information content present in protein-DNA interactions. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 488:477–88, 2008.

[5] Matthias Keil, Thomas E. Exnep, and Jürgen Brickmann. Pattern recognition strategies for molecular surfaces: III. Binding site prediction with a neural network. *Journal of Computational Chemistry*, 25(6):779–789, 2004.

[6] D. D. Kirsanov, O. N. Zanegina, E. A. Aksianov, S. A. Spirin, A. S. Karyagina, and A. V. Alexeevski. NPIDB: nucleic acid–protein interaction database. *Nucleic Acids Research*, 41(D1):D517–D523, 2013.

[7] Yanay Ofran, Venkatesh Mysore, and Burkhard Rost. Prediction of DNA-binding residues from sequence. *Bioinformatics*, 23(13):347–353, 2007.

[8] Jingna Si, Rui Zhao, and Rongling Wu. An overview of the prediction of protein DNA-binding sites. *International Journal of Molecular Sciences*, 16(3):5194–5215, 2015.

[9] A L Tarca, V J Carey, X W Chen, R Romero, and S Draghici. Machine Learning and Its Applications to Biology. *PLoS Comput Biol*, 3(6):e116, 2007.

[10] Harianto Tjong and Huan-Xiang Zhou. DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. *Nucleic Acids Research*, 35(5):1465–1477, 2007.

[11] Jean-philippe Vert. Machine Learning in Computational Biology (the frequentist approach). (September), 2013.

[12] Changhui Yan. Identi cation of interface residues involved in protein-protein and protein-DNA interactions from sequence using machine learning approaches. 2005.