

國立臺灣大學電機資訊學院資訊工程學研究所


碩士論文

Graduate Institute of Computer Science Engineering
College of Electrical Engineering and Computer Science
National Taiwan University

Master Thesis

從一級結構預測 DNA 結合蛋白之標的序列

Predicting target sequences of DNA-binding proteins based on
primary structure



林志瑋

Chih-Wei Lin

指導教授：歐陽彥正 博士

陳倩瑜 博士

Advisor: Yen-Jen Oyang, Ph.D.

Chien-Yu Chen, Ph.D.

中華民國 100 年 7 月

July, 2011

誌謝

兩年前憑著喜歡生物的一股傻勁，就決定進入生物資訊這個領域的研究，因此就決定加入了歐陽彥正老師的實驗室。在念碩士班的這兩年之間，得到了許多人的幫助。最重要的就是歐陽彥正老師、陳倩瑜老師以及張天豪老師的教誨，在參與跨領域計畫的研究過程中，我學到了許多東西。另外還要感謝廷因學長、鈺峰學長、翊鍾學長在我摸索生物資訊的過程中不時的給我許多指教，令我成長許多。還要謝謝元鴻以及又仁，雖然我們研究的主題不同，但是有兩位夥伴在這兩年中一起打拼是很棒的事情，非常開心可以認識兩位。最後要感謝我的前女友湯穎婷小姐，雖然在去年的時候離開了我，讓我低迷了一陣子。但是在這之後我在各方面的確成長了許多，感謝你。



中文摘要

結合特定 DNA 序列的蛋白質在基因調控中扮演重要的角色，利用計算方法預測或設計生物實驗尋找這些 DNA 結合蛋白質的標的序列可以幫助我們了解基因調控如何進行，並解釋基因組中序列的變異如何擾亂正常的基因表現。位置頻率矩陣 (position frequency matrix) 是最常被拿來描述這些標的序列的模型，對大部分的物種而言，截至目前為止，只有一小部分的轉錄因子已經從相關生物實驗中取得這樣的模型。由於生物實驗往往需要高資金與人力成本，因此，如何利用計算方法準確預測位置頻率矩陣，加速這個研究領域的進展，一直以來是生物資訊學家非常關心的研究議題之一。這篇論文針對這個問題，提出一個利用蛋白質 DNA 複合物結構與紀錄不同胺基酸和核酸之間結合偏好的知識庫去預測 DNA 結合蛋白之標的序列的新方法。當我們拿到一條蛋白質序列，會先挑選一個適當的樣板複合物結構，接著利用該樣板與所得之知識庫進行位置頻率矩陣的預測。

這篇論文使用了兩組資料去評估新方法的表現，和其他利用三級結構的方法比較起來，這篇論文提出的新方法可以達到和它們一樣的預測效果；但若與另一個同樣以序列資訊為基礎且利用已知位置頻率矩陣訓練所得之預測模型相比，本論文所提之方法表現略差。由於現存這些以序列資訊為基礎的預測方法仍各有其侷限處，本論文所提之方法，仍可幫助一些相關的研究，針對其同源序列已有蛋白質 DNA 複合物結構之蛋白質序列預測其標的序列，所得之預測結果將有助於相關研究之進行。

ABSTRACT

Proteins that bind specific DNA sequences play important roles in regulating gene expression. Identifying target sequences of a DNA-binding protein helps to understand how genes are regulated in cells and explain how genetic variations cause disruption of normal gene expression. Position frequency matrices (PFMs) are one of the most widely used models to represent such target sequences. However, up to now, for most species, only a small fraction of the transcription factors (TFs) have experimentally determined PFMs. Since biological experiments usually require much time and cost, it is strongly desired to develop computational methods with satisfied accuracies to speedup the progress. Here, a new method based on existing protein-DNA complex structures and the knowledgebase containing the preference of contacts between amino acids and nucleotides is proposed to predict quantitative specificities of protein-DNA interactions. When given a query protein sequence, a protein-DNA complex structure of homologues proteins is selected and the PFM prediction is made based on the selected template incorporated with the built knowledgebase.

The proposed method is evaluated by two datasets and compared with existing computational methods. It turns out that the proposed method can predict as well as the compared structure-based methods. On the other hand, when a sequence-based method that is trained by collected experimentally determined PFMs is compared, the proposed method performs slightly worse. Even though, the proposed method still has its value since different predictors usually have their own advantages and limitations. In summary, it is concluded that a DNA-binding protein's binding preference can be predicted based on its primary structure using the complexes of its homologues. This facilitates related studies in the future because target sequences of proteins without a

solved structure could be predicted now.



CONTENTS

口試委員會審定書	#
誌謝	i
中文摘要	ii
ABSTRACT	iii
CONTENTS	v
Tables	vii
Figures	viii
Chapter 1 Introduction.....	1
Chapter 2 Literature Review	4
2.1 Algorithms for predicting protein-DNA binding specificities.....	4
2.1.1 Predicting the binding preference of DNA-binding proteins.....	4
2.1.2 Predicting PFM by homologues' annotated PFMs.....	5
2.1.3 Predicting PFM based on structural model and potential functions	6
2.2 Algorithms of sequence alignment	10
2.2.1 BLAST	10
2.3 Algorithms of structure alignment.....	11
2.3.1 TM-align	11
Chapter 3 Methods.....	13
3.1 Materials	13
3.1.1 Collection of protein-DNA complex structures	13
3.1.2 Collection of PFMs	14
3.1.3 Relating PFMs to protein-DNA complex structures	14

3.2	Building the knowledgebase.....	14
3.3	Prediction framework	16
3.3.1	Template selection and contact residue substitution	18
3.3.2	Building the predicted PFM by DNA sequence in the template	18
3.3.3	Refining the PFM by knowledgebase	18
Chapter 4	Results	21
4.1	Measuring performance	21
4.2	Validation sets	21
4.2.1	Training data of SABINE.....	21
4.2.2	Protein-DNA complexes with annotated PFMs	22
4.3	Performance	22
4.3.1	Training data of SABINE.....	22
4.3.2	Protein-DNA complexes with annotated PFMs	25
4.4	Evaluating SABINE	30
4.5	Discussion.....	30
4.5.1	Differences between DNA sequences in protein-DNA complex structures and their annotated PFMs	30
4.5.2	The effect of different contact distance cut-off	32
4.5.3	How to select a template	33
4.5.4	Similar protein sequences bind similar DNA sequences.....	33
4.5.5	Using the number of contact atoms of contact residues.....	34
4.5.6	The frequency of amino acids and nucleotides	35
Chapter 5	Conclusions.....	38
	REFERENCE	40

Tables

Table 4-1 PFM become more similar with annotation after refinement.	23
Table 4-2 Difference before and after PFM refinement by the knowledgebase.....	26
Table 4-3 Similarities of 24 proteins of different algorithms	29
Table 4-4 Similarities of 12 proteins with similarity > 0.7.....	32
Table 4-5 Average similarity under different distance cut-off.....	33



Figures

Figure 2-1 Prediction framework of Schroder, A., et al's work	6
Figure 2-2 Workflow of the study of <i>Alamanova, et al.</i>	10
Figure 3-1 Preference between amino acids and nucleotides.....	16
Figure 3-2 Prediction framework of this study.....	17
Figure 4-1 Similarities under training data set of SABINE.....	24
Figure 4-2 Distribution of similarities between DNA sequences and annotated PFMs..	31
Figure 4-4 Correlation between protein sequence similarities and PFM similarities.....	34
Figure 4-5 Contact counts between amino acids and nucleotides	36
Figure 4-6 Scores between amino acids and nucleotides under the older scheme	37



Chapter 1 Introduction

Proteins that bind to specific DNA sequences are extremely important for the proper regulation of gene expression. Identifying the target sequences of DNA-binding proteins binds can help to understand how gene regulation proceeds and how genetic variations cause disruption of normal gene expression within cells. In recent years, proteins can be assigned to a certain molecular function (e.g., transcription factor) by biologists or computational methods efficiently. However, quantitative functional information (e.g., DNA-binding specificities) remains insufficient for the requirement. Despite recent progress in the development of high-throughput technologies for the measurement of protein-DNA interaction parameters, the determination of highly resolved quantitative binding specificity information is still laborious.

Position frequency matrix (PFM) is a simple probabilistic model to represent the consensus of target DNA sequences that can be recognized by a DNA-binding domain of transcription factors (TFs) [1-3]. PFMs indicate for a certain TF how frequently the nucleotides A, C, G and T occur at each position within the binding site.

There are experimental methods for determining the binding specificity of a protein. Surface plasmon resonance (SPR) is one of the methods for measuring the binding affinity of a protein-DNA interaction directly. SPR is often used to study protein-ligand interactions, but it can also be used to measure protein-DNA interactions. SPR is based on the fact that the angle of light reflection from a surface depends on the mass of molecular attached to the other side of the surface. DNA can be attached to the surface,

and then proteins are added to change the reflection angles of the light. The on-rate for the formation of protein-DNA complex and the off-rate for its dissociation can thus be measured, and then the binding affinities can be measured. Protein binding microarrays (PBMs) are another technology for measuring binding specificity of proteins. PBM uses arrays of over 44,000 spots which contain all possible combinations of DNA 10-mer. A protein is added into the array, and is then washed to remove nonspecific binding. The amount of proteins that bind to a specific DNA spot is determined with a fluorescent antibody to the protein. Despite the significant progress of experimental methods, proteins still need to be prepared for these methods, either purified from cells or synthesized in vitro. So it still spends some time using experimental methods to determine the binding specificity of protein-DNA interactions.

Using computational methods to model binding specificity can spend much less time than using experimental methods does. One of the most widely used computational methods for PFM inference of a transcription factor is to collect a set of promoter sequences to which the transcription factor can bind, and then to conduct motif finding and determine the frequency of each position among the detected over-represented subsequences. Such methods require sufficient sequences for pattern discovery, which are currently only available for a small amount of DNA-binding proteins. Previously, some structure-based approaches were also presented to predict PFMs. Several approaches are based on analyses of protein-DNA complex structures. These methods are shown to perform well in telling which positions in a PFM should be more conserved and do not allow degeneration [4-7]. On the other hand, *Schroder, A., et al.* applied the method support vector regression (SVR) to predict a quantitative measure for the PFM similarity of between proteins based on their primary structure, and further

predict the PFM of a protein [8].

However, in [8], the preference between amino acids in proteins and bases in nucleotides was not considered. In addition, it requires homologues of the query protein to have an annotated PFM, which is not usually available. In this regard, this thesis aims at providing an alternative to predict target DNA sequences of a DNA-binding protein from primary structure, provided that any homologue of the query protein has a protein-DNA complex structure. In this thesis, a novel approach to predict PFM of a protein based on its primary structure and a collection of protein-DNA complex structures is proposed. The DNA sequence in a protein-DNA complex can be regarded as a PFM that every column contains only one of the nucleotides (A, T, C, or G) with probability equals to one. Based on the idea that similar contact residues of proteins might bind to similar DNA bases, the sequences in protein-DNA complex structures were used to infer the PFM of a given protein sequence. Furthermore, a knowledgebase that describes the contact frequency between amino acids and bases is built to refine the PFMs built by the DNA sequence in a protein-DNA complex structure.

In Chapter 2, several methods proposed for predicting target sequence of DNA-binding proteins and previous studies of protein-DNA interactions are introduced. The proposed method is introduced in Chapter 3, and in Chapter 4, the performance of the proposed method is shown and some discussion is made. Finally, conclusions of this thesis are included in Chapter 5.

Chapter 2 Literature Review

In section 2.1, different methods for predicting protein-DNA binding specificities are introduced. Given a protein with a solved protein-DNA complex structure, its binding specificity can be inferred by structure-based PFM predictions. For a protein without solved structure (i.e. protein with only primary structure information), its binding specificity can be inferred by protein-DNA binding information of its homologues, for example, protein binding microarray (PBM) data, annotated PFMs, or protein-DNA complex structures.

The technique of sequence alignment was used in this thesis, thus the basic idea of sequence alignment is introduced in section 2.2. In addition, structure alignment is used while comparing the proposed method in this thesis with the existing structure-based PFM prediction, so an algorithm of structure alignment, TM-align is introduced in section 2.3.

2.1 Algorithms for predicting protein-DNA binding specificities

2.1.1 Predicting the binding preference of DNA-binding proteins

In the study of *Berger, et al.* [9] the Z-score transformed relative signal intensities for 168 homeodomains against all the probabilities of 32,896 8-mer DNA sequences were obtained using protein binding microarrays (PBMs). After that, *Alleyne, et al.* applied different machine learning algorithms to predict the Z-scores between mouse

homoedoms and individual DNA 8-mers [10].

In the study of [10], the contact region of these proteins were aligned and transferred into feature sets for the training of prediction models. The sequence alignment was converted into numerical encodings representing amino acid sequences of length l as binary vectors of length $l \times 20$ digits, i.e. the 20 different amino acids were encoded as 20 orthogonal vectors and an amino acid sequence was represented by concatenating binary vectors corresponding to residues at each position. Gaps were encoded as a vector of 20 zeros. With derived features, classifiers including nearest neighbor, random forest regression, support vector regression with linear, polynomial, and radial basis function kernel, and principal components regression were applied to predict the Z-score. It turns out that nearest neighbor performance best comparing to other algorithms. The root mean square error (RMSE) of nearest neighbor is 0.76. This reveals the fact that, the amino acids within contact regions do have influences on binding preference of proteins.

2.1.2 Predicting PFM by homologues' annotated PFMs

Schroder, et al. applied the regression method support vector regression (SVR) to predict PFMs of proteins based on their primary structure [8]. Pairwise alignment scores, structural and physicochemical properties, and phylogenetic distances, were used as features to train the SVR models in order to predict the PFM similarity of two proteins. The prediction frame work is described as followed (Figure 2-1):

- i) Feeding the sequence, organism, and structural superclass of the protein to the trained SVR models;
- ii) (the SVR model) reporting the PFMs with similarity higher than a threshold;

- iii) Filtering outlier PFMs;
- iv) Merging the remaining PFMs to get the final prediction;

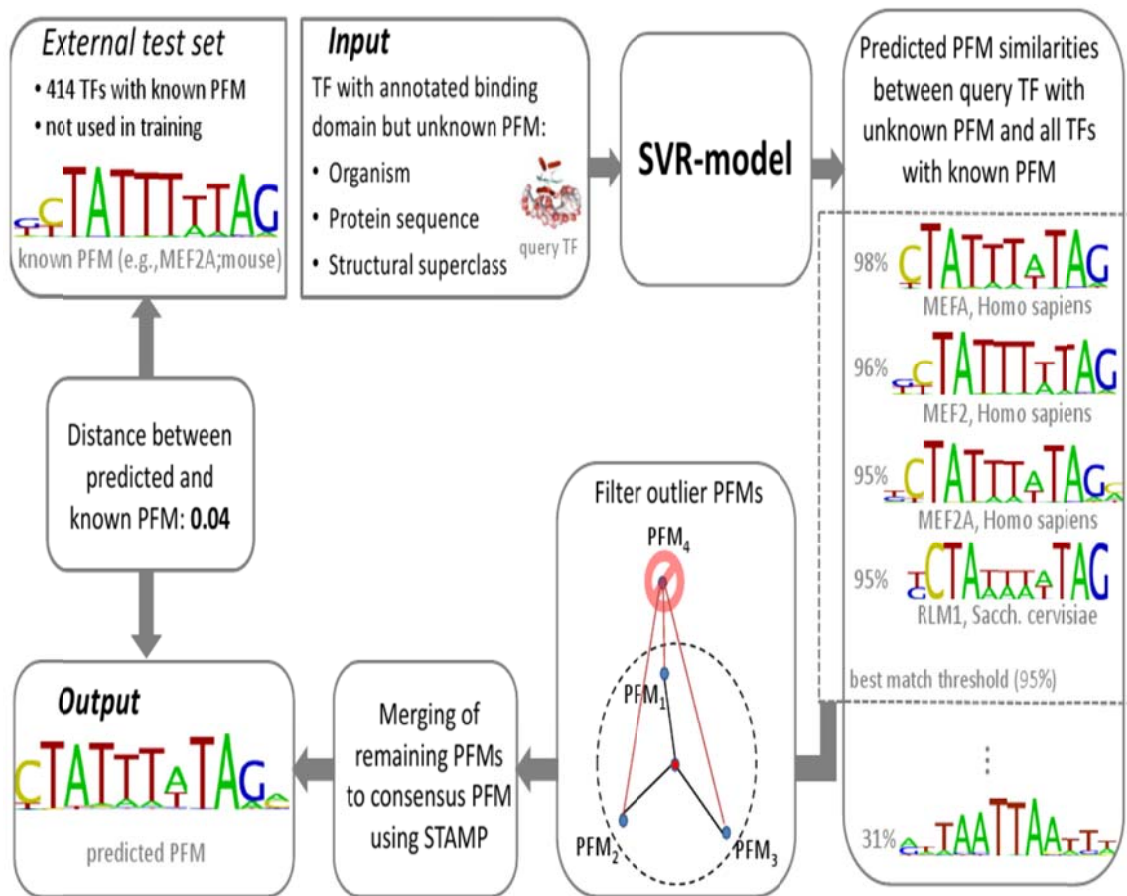


Figure 2-1 Prediction framework of Schroder, A., et al's work

The average absolute error (AAE) of this framework with default parameter is 0.12 on a scale from 0 to 2. In comparison, the average similarity between two PFMs that are randomly sampled is 0.64, indicating that the predicted performance of SVR model is significantly higher than the performance expected by random guesses.

2.1.3 Predicting PFM based on structural model and potential functions

Morozov, et al. developed a simple null model called 'contact' model for predicting PFMs [5, 6]. This model only cares about the number of atomic contacts between

protein side chains and DNA bases. The probability of the base pair type α in the PFM column i is calculated as follows:

$$p_{\alpha}^i = \begin{cases} \frac{1}{4}(1 - N/N_{max}) \text{ if } N \leq N_{max}, 0 \text{ if } N > N_{max} & (i \neq wt) \\ \frac{1}{4}(1 + 3N/N_{max}) \text{ if } N \leq N_{max}, 1 \text{ if } N > N_{max} & (i = wt) \end{cases}$$

, where N is the observed number of atomic contacts (a heavy atom pair from the protein and DNA respectively is defined to be in contact if they are closer than 4.5\AA), wt denotes the base type observed in the position i in the complex, N_{max} and is a free parameter. This model has been shown to be generally as good as the static model using physics-based potential function when a native protein-DNA complex is considered [5], and has the advantage of consuming much less computation cost.

Different to the contact model, which considers only the coordinates of residue atoms, the all-atom knowledge-based potential function presented by *Zhou, et al.* takes the amino acid types into consideration. The FIRE potential function [7] is a succinct knowledge-based potential function that considers interactions between all atom types. Among the series of all-atom scoring functions presented in [7], FIRE has the advantage of easy implementation and is shown to be generally as good as two of its extended functions, cFIRE and vcFIRE, in predicting PFMs.

To construct the knowledgebase, the number of pairs of atom types i and j with the distance falling within a specified range $(r - \Delta r, r]$ were denoted as $N_{obs}(i, j, r)$, where $r = 3, 4, 5, 6, 7, 8, 9, \text{ and } 10 (\text{\AA})$, and Δr is set as 3 for $r = 3$ and 1 for the rest of the values of r . In this study, the number of pairs of atom types i and j with the distance falling within a specified range, $N_{obs}(i, j, r)$, are calculated based on the 990 protein-DNA complex structures collected from PDB [11]. With $N_{obs}(i, j, r)$ of all the combinations, the potential between atom types i and j is represented as follows:

$$u^{FIRE}(i, j, r) = \begin{cases} -RT \ln \frac{P(i, j, r)}{P_{ref}(r)}, & \text{if } r < r_{cut} \\ 0 & \text{if } r > r_{cut} \end{cases}$$

where $P(i, j, r) = N_{obs}(i, j, r) / \sum_r N_{obs}(i, j, r)$, $P_{ref}(r) = r^\alpha \Delta r / \sum_r r^\alpha \Delta r$, $r_{cut} = 10 \text{ \AA}$. The value of α is set as 1.61 because it best fits of r^α to the actual distance-dependent number of ideal-gas points in finite protein-size spheres [7]. For a given complex, the binding free energy, ΔG , is defined as the sum of all the potentials of the observed atom pairs [5]:

$$\Delta G = \sum_{i, j} u^{FIRE}(i, j, r)$$

Assume that influences on binding free energy from different positions are independent, and thus ΔG can be represented as follows:

$$\Delta G = \sum_i \Delta G_\alpha^i$$

where ΔG_α^i is the binding free energy of a base α (A, T, C, or G) at position i . By combining two equations above, we can estimate the probabilities in each column of PFMs as follows:

$$p_\alpha^i = \frac{\exp(-\beta \Delta G_\alpha^i)}{\sum_{b \in \{A, T, C, G\}} \exp(-\beta \Delta G_b^i)}$$

where β is a free parameter.

Alamanova, et al. used another potential function proposed in [12] to predict PFM [4].

$$G \approx -\ln P(C|D) = -\sum_i^{N_P} \sum_j^{N_D} \ln P(C|d_{ij}, t_i, t_j)$$

where D is the set of atomic distances d_{ij} between interface atoms, t_i and t_j are the atom types. N_P and N_D are the number of atoms in the protein and DNA. The

probability of atomic contacts was modeled as the likelihood:

$$P(C|d_{ij}, t_i, t_j) = P(C) \frac{P(d_{ij}, t_i, t_j|C)}{P(d_{ij}, t_i, t_j)}$$

where $P(C|d_{ij}, t_i, t_j)$ is the likelihood function, $P(d_{ij}, t_i, t_j)$ is the marginal probability, $P(C)$ is the Bayesian prior and was set to one. The following formula expresses the likelihood of observation a native-like interatomic distance:

$$P(d_{ij}, t_i, t_j|C) \approx f(d_{ij}, t_i, t_j) = \frac{N_{obs}(d_{ij}, t_i, t_j)}{\sum_{d_{ij}} N_{obs}(d_{ij}, t_i, t_j)}$$

Where $N_{obs}(d_{ij}, t_i, t_j)$ is the number of contacts observed between t_i and t_j separated by distance d_{ij} .

For each DNA sequence of length N in the protein-DNA complex, 4N+X random sequences were generated, and their potentials when bound by the query protein are calculated by the formulas above. The weights of each position in the PFM were estimated by solving the linear equation:

$$AX = b$$

, where X is a vector of 4N dimensions of the estimated weights, and A is a binary matrix of dimensions (4N, 4N+X), with each row of the matrix A corresponding to one DNA sequence. In addition, b is a vector consists of 4N+X potential values. The linear equation is solved by least squares optimization. Finally, the probabilities in each column of PFMs were estimated as follows:

$$p_{\alpha}^i = \frac{\exp(-\beta\Delta G_{\alpha}^i)}{\sum_{\gamma \in \{A,T,C,G\}} \exp(-\beta\Delta G_{\gamma}^i)}$$

The workflow of this study is described as Figure 2-2

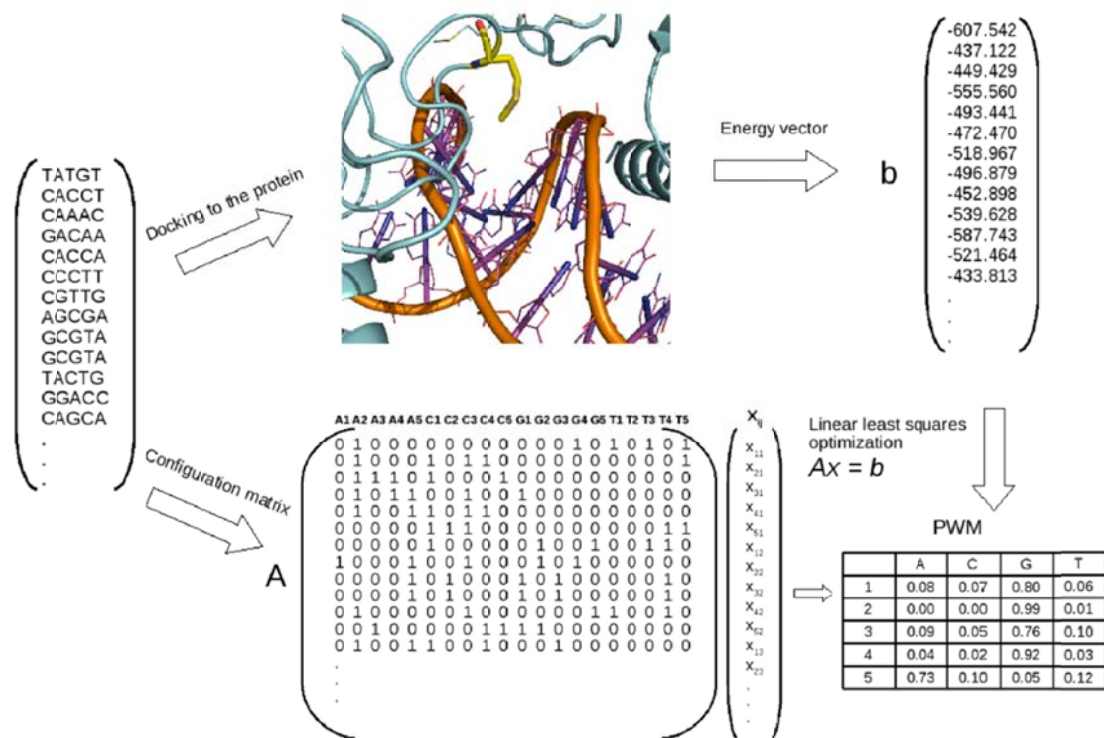


Figure 2-2 Workflow of the study of *Alamanova, et al.*

2.2 Algorithms of sequence alignment

Sequence alignment is a fundamental problem of bioinformatics. Smith-Waterman and BLAST are the most famous algorithms for local sequence alignment. Smith-Waterman uses dynamic programming to solve the problem, and its time complexity is $O(n \times m)$, where n and m are the length of the sequences. So it takes lots of time to align sequences. In this regard, it is not useful while searching a huge database. On the other hand, BLAST uses a heuristic approach to search the solution, and it takes much less time than Smith-Waterman while aligning long sequences.

2.2.1 BLAST

BLAST (Basic Local Alignment Search Tool) is a local sequence alignment algorithm

for biological sequences, and we can use it for protein sequence or DNA sequence alignment [13]. We can align a query sequence with a sequence database by BLAST, and further find its similar sequences in the database.

BLAST is one of the most commonly used programs in the field of bioinformatics due to its high speed for searching a solution. Although it does not guarantee an optimal solution as Smith-Waterman does, its high speed still makes it very useful, and thus being widely used.

The major steps of BLAST are listed as followed:

- Making a list of neighborhood words, for example, length 4 for amino acid sequences;
- Constructing a dictionary for all words in the query;
- Matching query with score higher than a threshold;
- Scanning the database for words;
- Extending the match in both directions when a match is found;

2.3 Algorithms of structure alignment

Protein structural comparisons are employed in many branches of structure biology, ranging from protein fold classification, protein structure modeling to structure-based protein function annotation. TM-align is one of the most famous algorithms for structure alignment [14]. It is ~4 times faster than CE [15] and 20 times faster than DALI [16] and SAL [17].

2.3.1 TM-align

TM-align only employs the information of backbone C_{α} coordinates of the given

protein structure. The algorithm executes the following steps:

1. Initial structure alignment

There are three initial alignments being exploited. The first type of initial alignment is obtained by assigning secondary structure to each residue of two proteins using dynamic programming. The second type of initial alignment is based on the gapless matching of two proteins. The smaller protein is gapless threaded against the larger protein. The third initial alignment is also by dynamic programming, but the scoring matrix is a combination of the first and the second initial alignment.

2. Heuristic iteration

In this procedure, structures are rotated by TM-score rotation matrix based on the initial alignment [18]. The score similarity is defined as:

$$S(i, j) = \frac{1}{1 + d_{ij}^2/d_0(L_{min})}$$

where d_{ij} is the distance between the i th residue of the first structure and the j th residue of the second structure under the TM-score superposition. $d_0(L_{min}) = 1.24\sqrt[3]{L_{min} - 15} - 1.8$, L_{min} is the length of the smaller structure. A new alignment can be obtained by implementing dynamic programming on the matrix $S(i, j)$. Then a new TM-score rotation matrix is obtained by the new alignment, and a new score matrix is obtained. The procedure is repeated until the alignment converges. Finally, the alignment with the highest TM-score is returned.

Chapter 3 Methods

Although the structure-based PFM predictors have been shown to perform well [5-7], especially in telling which positions in a PFM should be more conserved and do not allow degeneration, it is strongly desired to design a new algorithm because structure-based PFM prediction can be applied only when the query protein has a structure solved. However, there are still many proteins do not have a solved structure. In this regard, this thesis aims to develop a method to predict the target sequence of these proteins.

As introduced in Chapter 2, there are algorithms for predicting protein-DNA binding specificities based on their homologues' protein-DNA binding information, such as PBM or PFM. However, protein-DNA complex structures provide some kind of protein-DNA binding information. Here a novel method for inferring PFMs of a protein by using protein-DNA complex structures of its homologues is proposed.

3.1 Materials

3.1.1 Collection of protein-DNA complex structures

Protein-DNA complex structures from the 27 February 2009 release of PDB were collected [11]. Protein-DNA complexes were collected for finding a proper template of given query protein sequences. Since similar proteins binds similar DNA sequences, a protein-DNA complex structure of the query's homologue helps predicting target sequences of the query. The template structures were required to satisfy the following criteria: a) it is an X-ray structure with resolution better than 3.0Å, b) the DNA

molecule has ≥ 6 paired bases and has less than 30% non-paired bases, c) the protein chain has ≥ 5 contact residues (residues within 4.5Å to the DNA molecule) and d) the protein chain has ≥ 40 residues. There are 990 protein chains satisfying the criteria above.

3.1.2 Collection of PFMs

Annotated PFMs of human, and mouse were collected from TRANSFAC [19], and annotated PFMs of yeast were collected from MYBS [20]. Totally, 592 annotated PFMs of human, 802 of mouse, and 117 of yeast were collected.

3.1.3 Relating PFMs to protein-DNA complex structures

After collecting the data above, the next step is to know with which protein-DNA structures the collected PFMs are associated. PFMs were related to protein-DNA complex structures by their entry names in UniProt database [21]. If a PFM and a protein have the same entry name, they are associated with each other, i.e. this protein is supposed to bind to the DNA sequences similar with those of the same entry name. After mapping PFMs to the protein-DNA complex structures, 119 protein chains with an annotated PFM are remained, corresponding to 26 different proteins.

3.2 Building the knowledgebase

Before building the knowledgebase, we clustered the collected protein chains by a hierarchical clustering algorithm HomoClust [22] based on pair-wise sequence identity reported by BLAST [13]. In each cluster, the protein chain with the largest average similarity to the other protein chains is assigned as the representative of the cluster. The

sequences of collected 990 protein-DNA complexes were used as input for HomoClust. In the end, 260 representatives were used to build the knowledgebase.

The contact residues of protein chains (residues with 4.5Å to any nucleotide present) in 260 templates were defined first. With this information, contact counts between twenty types of amino acids and four types of nucleotides were reported respectively. Note that for amino acids, only the atoms of the side chains are considered, because they discriminate different amino acids. Similarly, for nucleotides, only the atoms of bases are considered. After counting the contacts the scores between amino acids and nucleotides are calculated by the following equation.

$$S(a, n) = \log \frac{c(a, n)}{e(a, n)}$$

,where a means a type of amino acid , n means a type of nucleotide. $c(a, n)$ is the observed contact counts of a and n , and $e(a, n)$ is the expected contact frequency between a and n . The expected contact number is calculated by

$$e(a, n) = f(a) \cdot f(n) \cdot C$$

where $f(a)$ is the frequency that amino acid a appears in the Swiss-Prot database [21], and $f(n)$ is the frequency of the appearance of the nucleotide n , which was set to 0.25 in this study and C is the observed total contact counts between all amino acids and nucleotides.

The scores between amino acids and nucleotides are shown as Figure 3-1. It appears that nucleotides prefer amino acids with positive charges (Arginine, Histidine, and Lysine) and amino acids with polar (Asparagine and Glutamine) to other amino acids. On the other hand, it is interesting that Serine prefers Thymine to other nucleotides.

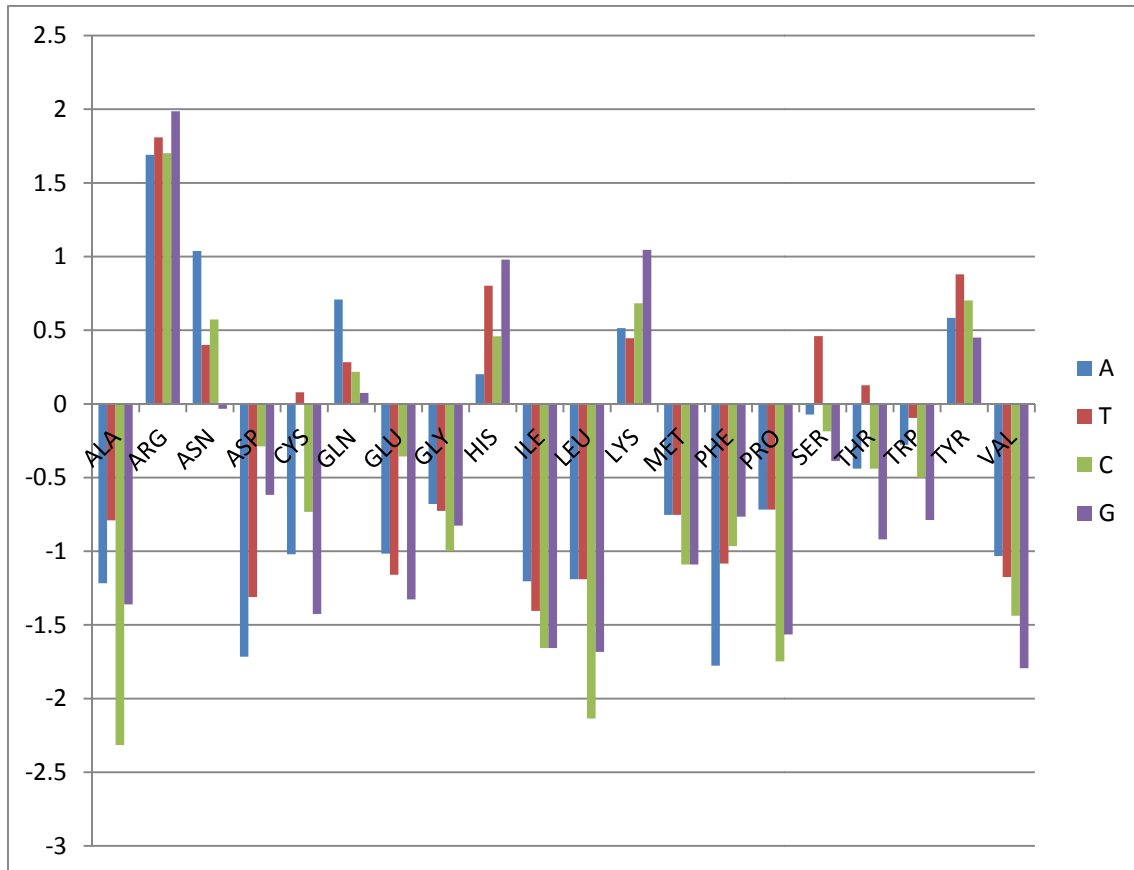


Figure 3-1 Preference between amino acids and nucleotides

3.3 Prediction framework

When a query protein sequence is given, the PFM prediction will be performed by the following three steps.

- i) Template selection and contact residue substitution
- ii) Building a predicted PFM by DNA sequence in the template
- iii) Refining the PFM by the knowledgebase

.

The prediction framework of this study is shown as Figure 3-2

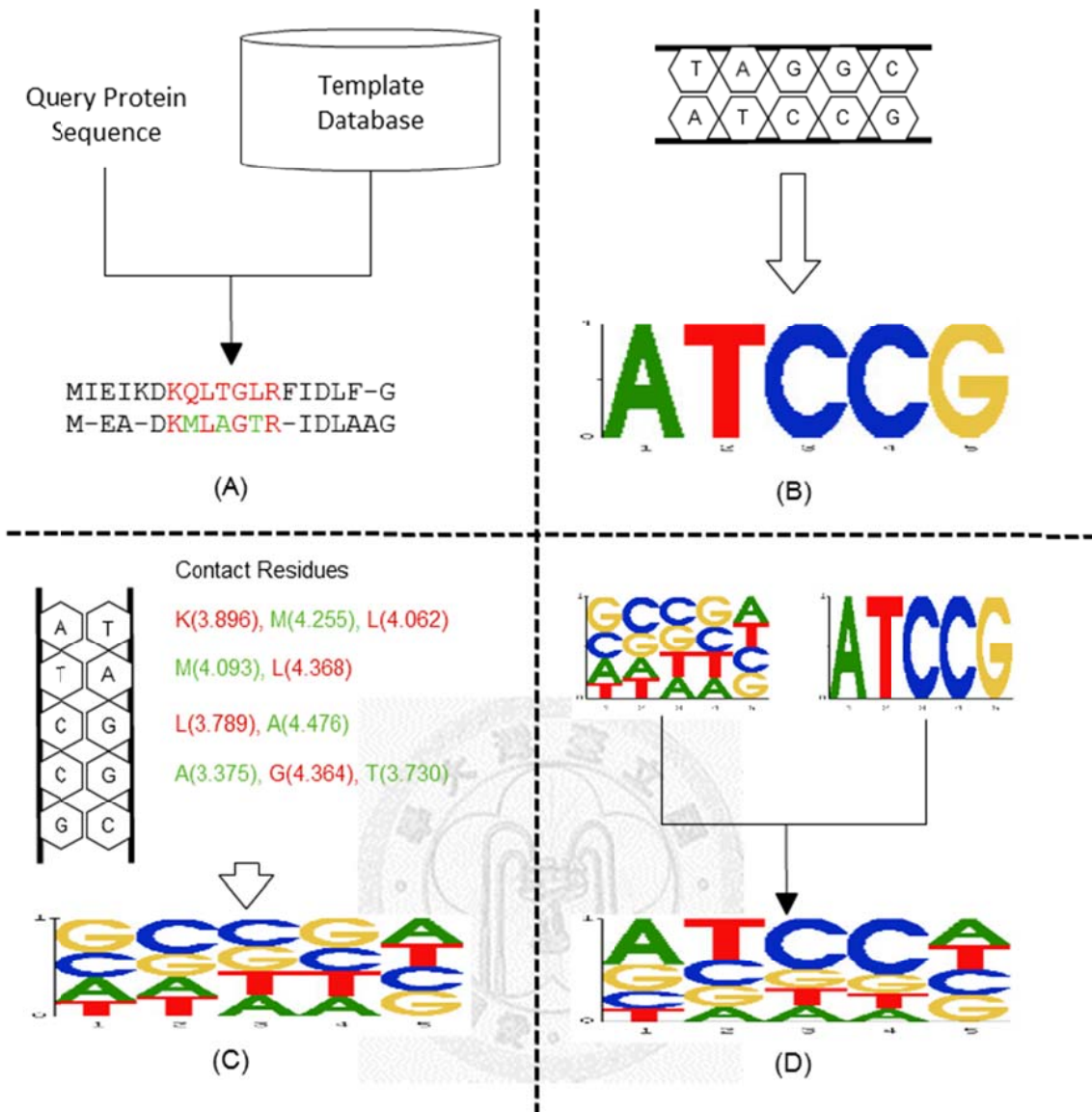


Figure 3-2 Prediction framework of this study. (A) Selecting a template from template database by BLAST and find out the contact residues that are substituted. (B) Generating a predicted PFM from the DNA sequence. (C) Calculating the PFM by the knowledge base and the distance between nucleotides and the contact residues. (D) PFM refinement by DNA sequence in the template structure.

3.3.1 Template selection and contact residue substitution

Given a query protein sequence, a proper homologue is selected from the 990 protein chains in template database using BLAST. The protein with the lowest e-value is selected as template of the query. According to the sequence alignment reported by BLAST, there will be mismatches on the contact residues (residues within 4.5Å to the nucleotide) of the template protein. The contact residues of the template protein are then replaced by. Note that not for all the given query protein sequences, the PFM prediction will be performed. If there's no template with e-value < 0.001, no PFM prediction is performed.

3.3.2 Building the predicted PFM by DNA sequence in the template

Since similar protein sequences bind similar DNA sequences, the DNA sequence in the template structure gives us important information to infer PFMs. After a template was selected, a PFM with probability of either one or zero based on the DNA sequence can be obtained. If the position i of the DNA sequence is base n , then $PFM_s(i, n) = 1$, otherwise, $PFM_s(i, n) = 0$

3.3.3 Refining the PFM by knowledgebase

Before refining the PFM built by the DNA sequence in the template structure, another PFM is built by the knowledgebase first. For each DNA base pair in the template structure, we can calculate a column of the PFM. If the DNA sequence in the template structure is L , then a PFM of length L is constructed. To construct the PFM, a scoring matrix M is calculated by the contact residues of each base pair and the distances between amino acids and nucleotides.

$$M(i, n) = \frac{\sum_{a \in \Gamma_i} S(a, n) \cdot d_a^{-w}}{\sum_{a \in \Gamma_i} d_a^{-w}} \quad (i = 1 \sim L, n \in \{A, T, C, G\})$$

$M(i, n)$ is the score of nucleotide n at position i and it equals to the weighted sum of the score between contact residue a and the nucleotide n defined by the knowledge base, i.e. $S(a, n)$. For each position i in the template structure, we have the set of contact residues Γ_i of this position. The weight of each contact residue a is given in an inverse ratio of the distance d_a to the DNA molecule in the template, and d_a is defined as the shortest atom pair distance between amino acid and DNA molecule in the template. Here w is a free parameter that can be adjusted by the user. The contact residues closer to the nucleotide are favored as w gets larger.

After calculating M , we can obtain a PFM by

$$PFM_k(i, n) = \frac{e^{\beta \cdot M(i, n)}}{\sum_{k \in \{A, T, C, G\}} e^{\beta \cdot M(i, k)}} \quad (i = 1 \sim L, n \in \{A, T, C, G\})$$

where β is a free parameter. Now we have a second PFM, which is constructed by the knowledgebase.

For refining the PFM built by the DNA sequence by the PFM built with the knowledgebase, the two PFMs built (PFM_s and PFM_k) are going to be merged together. If the contact residues of a base pair are conserved, then the information given by the DNA sequence in the template structure should be kept, so the refinement can be done by the following equation.

$$PFM_r(i, n) = PFM_k(i, n) \cdot (1 - \delta) + PFM_s(i, n) \cdot \delta \quad (i = 1 \sim L, n \in \{A, T, C, G\})$$

, δ is the ratio of the number of substituted contact residues over the total number of contact residues of the position i and $\delta = 0$ if there is no contact residue at the

position.

For the positions that are close to the starting or ending positions and without contact residues, they are regarded as unimportant positions in this protein-DNA interaction. In these positions, four types of nucleotides would be assigned with the same probability : 0.25. These positions are trimmed before reporting the prediction because they do not provide any useful information.



Chapter 4 Results

In this chapter, a measure to compare PFMs is first introduced in section 4.1. Introduction of validation sets that used to test performance is followed in section 4.2. In section 4.3 and 4.4, the performance of the proposed method and comparison with other methods is introduced. Finally, discussions are made in section 4.5.

4.1 Measuring performance

To compare the predicted PFM and the annotated PFM, a measure called ‘Similarity’ was employed, referring to the similarity function used in [23] for comparing PFMs. To calculate the similarity between two PFMs, say a and b , they are aligned to maximize

$$1 - \frac{1}{w} \sum_{i=1}^l \frac{1}{\sqrt{2}} \sqrt{\sum_{n \in \{A,T,C,G\}} (a_{i,n} - b_{i,n})^2}$$

where w is the number of positions that two PFMs aligned together, and $a_{i,n}$ and $b_{i,n}$ are the probabilities of nucleotides n at position i in PFM a and b , respectively.

4.2 Validation sets

4.2.1 Training data of SABINE

In [8], 1239 protein sequences with their annotated PFMs with were collected from different databases. 453 of them were public and attached in the source code of SABINE. These protein sequences were used to evaluate the proposed method.

4.2.2 Protein-DNA complexes with annotated PFMs

As mentioned in section 3.1.3, we have 26 different protein-DNA complex structures with known PFMs, and they were also used to evaluate the proposed method. The protein sequences of the protein-DNA complex structures were used as an input query for prediction. When evaluating the proposed method, the query protein sequence runs BLAST with other 989 template sequences and finds a template. However, there are two proteins cannot find a template with e-value < 0.001 , so we finally use 24 proteins for evaluation.

4.3 Performance

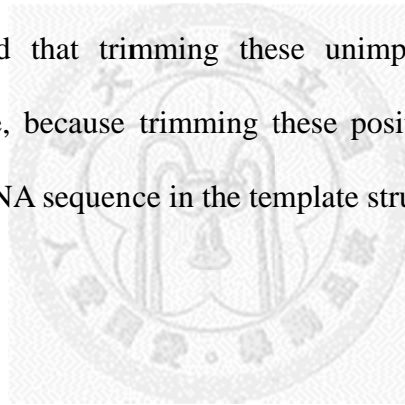
Two validation sets described in 4.2 were used to evaluate the proposed method. In 4.3.1, we can see the improvement after the refining the PFMs built by DNA sequences in the template structures by the training data of SABINE. Structure-based methods require a protein-DNA complex structure to perform prediction, so in 4.3.2, the proposed method is compared with structure-based methods using protein-DNA complexes with annotated PFMs, and it turns out that the proposed method can predict as well as structure-based methods predict.

4.3.1 Training data of SABINE

453 protein sequences contained in the training data of SABINE were used as query for the proposed method. 96 of these protein sequences cannot find a template structure with e-value < 0.001 , so 357 protein sequences were tested. As shown in Figure 4-1, PFMs built by employing the DNA sequence in the template (PFM_s) have an average similarity of 0.632, after refining PFM_s by the knowledgebase, an average similarity

of 0.682 is achieved. This shows that refinement by the knowledgebase do help inferring the PFM of a query protein sequence.

In Table 4-1, it is shown that PFM becomes more similar to the annotated PFM after refinement. For example, at the 11th position of annotated PFM, PFM_s gives probability = 1 to Thymine, after refinement, the knowledgebase divides some probability to Cytosine, so the PFM after refinement becomes more similar to the annotated PFM. Similar situations can be observed at 7th, 8th, 9th, 10th and 13th position of annotated PFM. On the other hand, PFM after refinement is shorter than PFM_s , because there is no any contact residue at 1st ~ 4th position of DNA sequence in the template structure, so predictions at these positions were trimmed before reporting the result. It can be observed that trimming these unimportant positions also helps improving the performance, because trimming these positions filters the unimportant information given by the DNA sequence in the template structure.



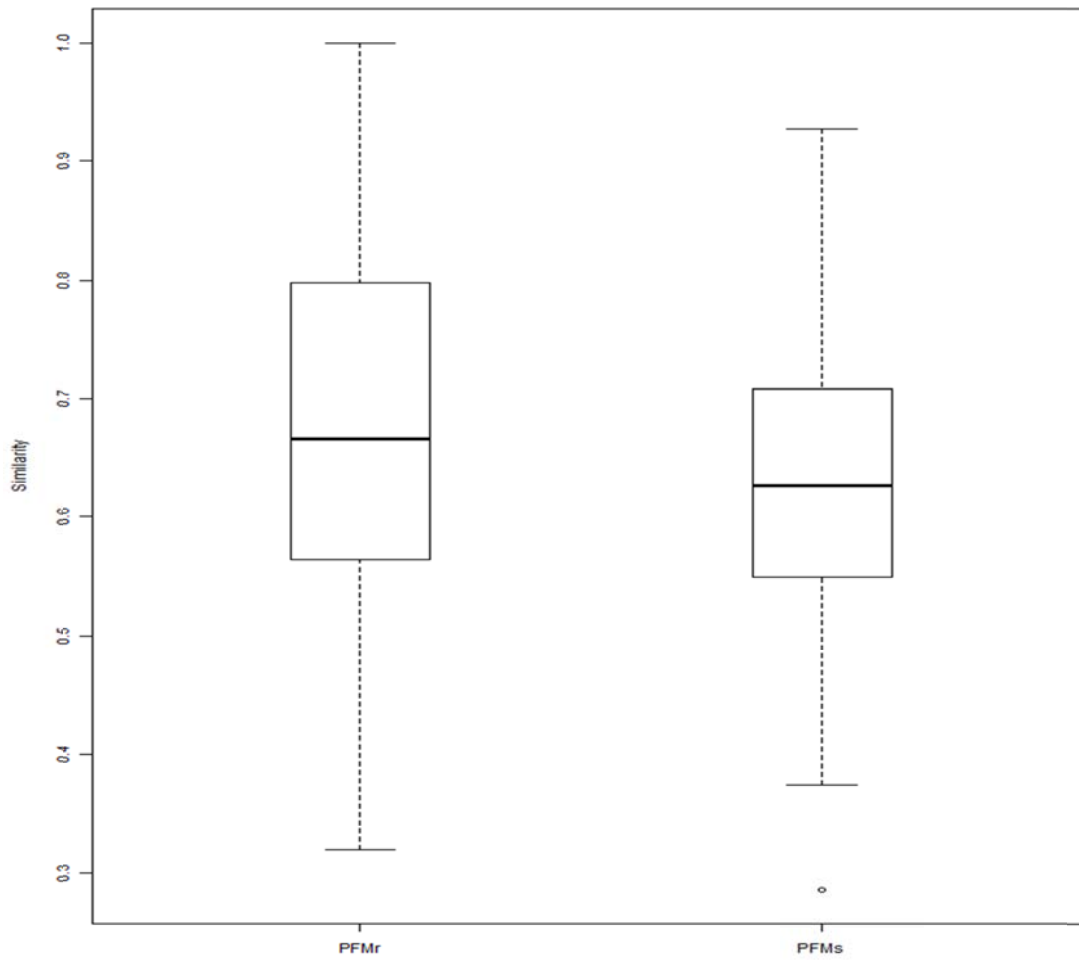
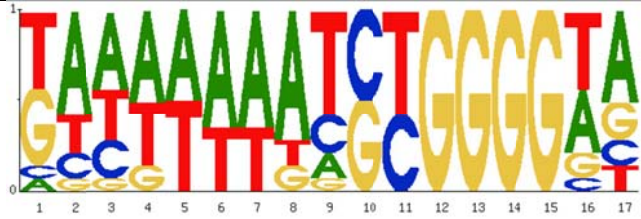
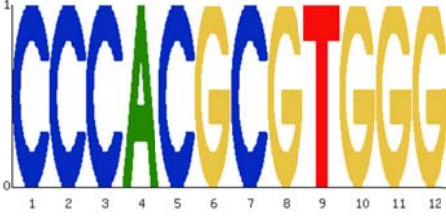
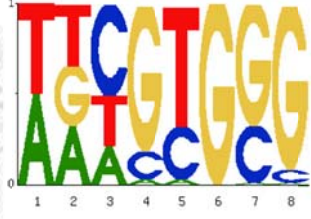


Figure 4-1 Similarities under training data set of SABINE

Table 4-2 PFM become more similar with annotation after refinement

	Binding Profile	Similarity
Annotated PFM		
DNA sequence of template		0.498
PFM after refinement by knowledge		0.796

4.3.2 Protein-DNA complexes with annotated PFMs

Next, 24 protein sequences in protein-DNA complex structures were used for validation. For the PFMs built by employing the DNA sequence in the template (PFM_s), they have an average similarity of 0.642 over the 24 proteins. After incorporating the PFMs built by knowledgebase, an average similarity of 0.711 is achieved. It shows again that the knowledgebase we built do help refining the prediction of PFM. In Table 4-2, it can be observed that most of all the cases have improvement after refinement. (PFM_s denotes PFMs built by employing the DNA sequences of templates, and PFM_r denotes PFMs after refinement)

Table 4-3 Difference before and after PFM refinement by the knowledgebase

PDB ID	PFM_s	PFM_r	Difference
1MDMA	0.498	0.535	0.037
1GTWA	0.618	0.660	0.042
1MNNA	0.650	0.689	0.039
3CO7C	0.706	0.904	0.198
2NNYA	0.507	0.639	0.132
3BPYA	0.723	0.853	0.130
1HLOA	0.662	0.751	0.089
1A0AA	0.876	0.862	-0.014
1NKPA	0.879	0.957	0.078
2UZKA	0.671	0.835	0.164
3G73A	0.473	0.556	0.083
2PI0A	0.678	0.659	-0.019
6PAXA	0.440	0.547	0.107
1YSAC	0.405	0.375	-0.03
1B72B	0.600	0.669	0.069
1MHDA	0.509	0.695	0.186
1TSRB	0.850	0.960	0.110
1D66A	0.596	0.994	0.398
1ZMEC	0.454	0.611	0.157
1KB2A	0.544	0.587	0.043
1MNMA	0.751	0.433	-0.318
1AKHA	0.962	0.962	0
2HAPC	0.657	0.622	-0.035
1CDWA	0.716	0.716	0
Average	0.642	0.716	0.074

The contact model proposed in [5, 6] and the all-atom knowledgebase potential (called ‘all-atom model’ in the following context) proposed in [7] were implemented in order to make comparison with the proposed method. Superimposed structures and native structures of these 24 proteins were used for PFM prediction respectively.

The superimposed structures were constructed by applying the rotation matrix reported by TM-align [14]. The original protein chains in the template were removed and the transformed coordinates of the query structure was appended into the template structure to generate a superimposed protein-DNA complex structure for structure-based PFM prediction. The templates selected here are the same as the templates selected by the proposed method. As a result, contact model has an average similarity of 0.692 and all-atom model has an average similarity of 0.679. On the other hand, the native structures of these 24 proteins were also used as input of the two methods for PFM prediction. As a result, contact model has an average similarity of 0.716, and all-atom model has an average similarity of 0.689.

In Table 4-4, similarities of different algorithms are included. Contact model and all-atom model are the algorithms described in this section. Native and superimposed denote using native structure and superimposed structure as query for structure-based prediction, respectively.

Compared with structure-based methods with superimposed structures, the proposed method achieves a better performance. For users have unbound query protein structures (protein structures without DNA), they could have better predicted PFMs if they apply the proposed method with the protein sequence of the unbound structure. As for

structure-based method with native structures, sequence-based method can achieve the same (or better, compared with all-atom model) performance as it does. This is a great accomplishment that the proposed, sequence-based method can do as well as structure-based method.



Table 4-4 Similarities of 24 proteins of different algorithms

PDB ID	Contact Model	All-atom Model	Contact Model	All-atom Model	Proposed
	native	native	superimposed	superimposed	method
1MDMA	0.724	0.623	0.728	0.654	0.535
1GTWA	0.735	0.691	0.708	0.661	0.660
1MNNA	0.794	0.745	0.713	0.733	0.689
3CO7C	0.721	0.860	0.738	0.677	0.904
2NNYA	0.851	0.754	0.751	0.713	0.639
3BPYA	0.742	0.702	0.735	0.821	0.853
1HLOA	0.854	0.716	0.835	0.737	0.751
1A0AA	0.538	0.974	0.490	0.705	0.862
1NKPA	0.488	0.618	0.525	0.701	0.957
2UZKA	0.809	0.638	0.738	0.725	0.835
3G73A	0.626	0.620	0.649	0.552	0.556
2PI0A	0.651	0.678	0.616	0.621	0.659
6PAXA	0.763	0.677	0.619	0.685	0.547
1YSAC	0.545	0.643	0.589	0.467	0.375
1B72B	0.723	0.806	0.597	0.700	0.669
1MHDA	0.829	0.659	0.697	0.659	0.695
1TSRB	0.681	0.592	0.733	0.616	0.960
1D66A	0.838	0.817	0.767	0.756	0.994
1ZMEC	0.759	0.647	0.752	0.676	0.611
1KB2A	0.737	0.730	0.686	0.682	0.587
1MNMA	0.632	0.603	0.550	0.734	0.433
1AKHA	0.709	0.813	0.765	0.749	0.962
2HAPC	0.730	0.588	0.830	0.621	0.622
1CDWA	0.720	0.644	0.808	0.625	0.716
Average	0.716	0.689	0.693	0.680	0.711

4.4 Evaluating SABINE

Sequences of these 24 proteins were also used as queries for SABINE [8]. When using default parameter, there is no prediction performed by SABINE. That is, SABINE cannot find a PFM with similarity higher than default threshold (0.95) from its training data. The similarity threshold was thus set to 0, forcing SABINE report as many results as possible, but SABINE still only report 21 results under this parameter. It appears that there are queries that SABINE cannot perform a prediction.

SABINE achieves an average similarity of 0.794 for the 21 queries, and the proposed method in this study has an average similarity of 0.710. However, after scanning the public parts of SABINE's training data set, 18 of 21 queries' annotated PFMs have similarity > 0.8 with SABINE's training data set. It is quite strict for us under this circumstance because SABINE already knows the annotated PFMs of queries. As for the rest four queries, SABINE has an average similarity of 0.704, and the proposed method has an average of 0.697. Given the query that is not contained in the training set of SABINE, the proposed method can do as well as SABINE can.

4.5 Discussion

This section discusses several interesting points observed in this study and suggests different parameters and methods to achieve a better performance for future studies.

4.5.1 Differences between DNA sequences in protein-DNA complex structures and their annotated PFMs

In the process of this study, it is observed that some DNA sequences in protein-DNA

complex structures have low similarities with their annotated PFMs. DNA sequences in the protein-DNA complex structures were transferred into PFMs that contains probabilities of 0 and 1, and these PFMs were used to calculate the similarity between them and annotated PFMs of proteins in the structure. It turns out that only 4 of the 24 protein chains have similarity > 0.8 , and 12 of them have similarity > 0.7 (). This is bad for the proposed method, which predicts target sequences by the DNA sequences in protein-DNA complex structures. Somehow it appears some protein-DNA complex structures and their annotated PFMs give different information to us. Figuring out the reason why this is happening might help improving the performance of the proposed method.

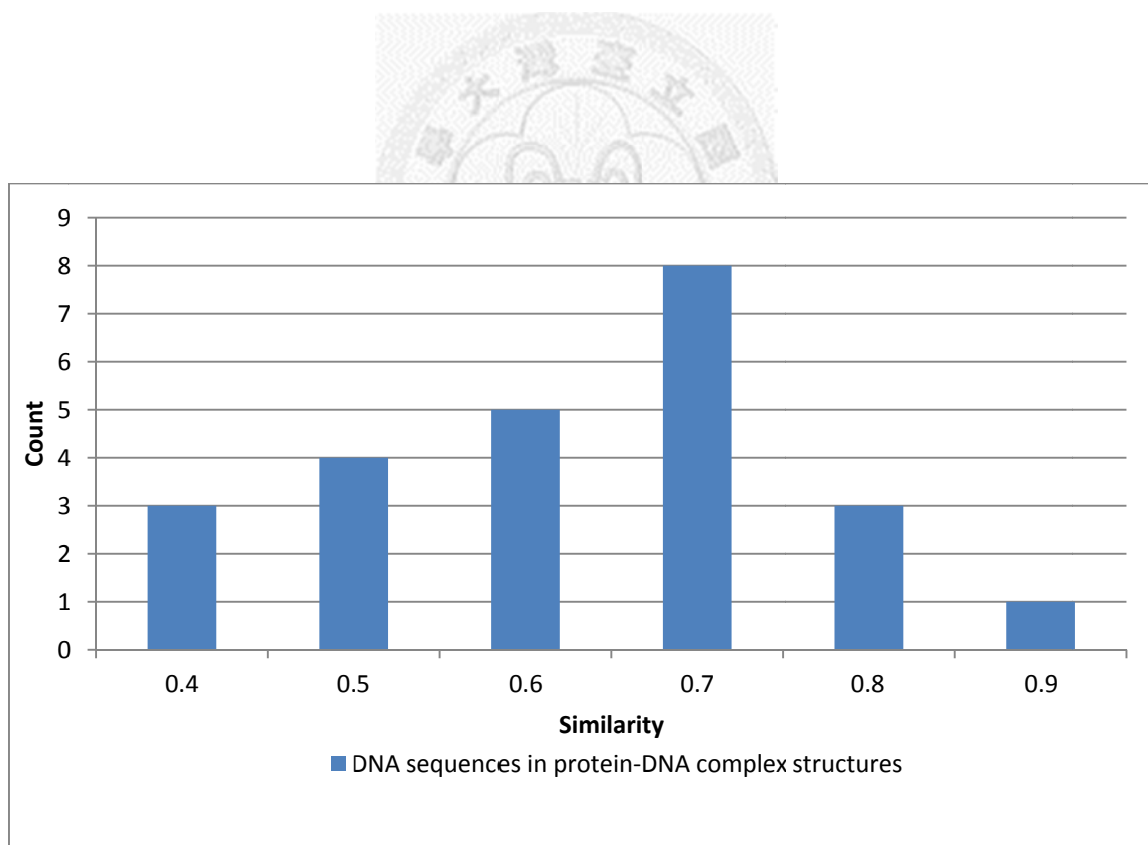


Figure 4-2 Distribution of similarities between DNA sequences and annotated PFMs

Among these 12 proteins chains which have similarity > 0.7 , the proposed method achieved an average similarity of 0.75. Six of them have higher similarity than their own DNA sequences have. This reveals that given cases that their DNA sequence in the protein-DNA complex structures and annotated PFMs give similar information, the proposed method can predict better.

Table 4-5 Similarities of 12 proteins with similarity > 0.7

QUERY	Similarity of DNA sequence in the structure and	Similarity of prediction by the proposed method
	annotated PFM	and annotated PFM
1MNNA	0.703	0.688
3CO7C*	0.748	0.904
3BPYA*	0.754	0.852
1HLOA	0.767	0.750
1A0AA	0.876	0.861
1NKPA*	0.879	0.956
2UZKA*	0.832	0.971
1YSAC	0.794	0.375
1B72B	0.722	0.669
1MNMA	0.750	0.432
1AKHA*	0.962	0.962
1CDWA*	0.715	0.715

*: prediction has higher similarity than query's own DNA sequence

4.5.2 The effect of different contact distance cut-off

While building the knowledgebase and calculating the PFM by the knowledgebase, different contact distance cut-offs were applied in order to build a better knowledgebase. 4.5, 6, and 7.5 Å were applied as the distance cut-offs. As a result, using 4.5 Å as distance cut-off achieves the best average similarity (Table 4-6), so finally 4.5 Å is used. It appears that only atoms close enough to DNA molecules can provide correct

information of protein-DNA interaction.

Table 4-7 Average similarity under different distance cut-off

<i>Distance cut-off</i>	4.5	6	7.5
Average Similarity	0.711	0.670	0.641

4.5.3 How to select a template

While selecting a template for a query protein sequence, it can be selected by e-value or sequence identity between query and template. As a result, using e-value can select a better template and thus have better performance (0.711 vs. 0.660). This might result from many evolutionary related proteins share low sequence homology.

4.5.4 Similar protein sequences bind similar DNA sequences

The basic idea of the proposed method is that, similar protein sequences bind similar DNA sequences, so the DNA sequences in templates are used for inferring target sequences of proteins. The sequence similarities and PFM similarities between 119 protein chains that have an annotated PFM were calculated. The result is showed as Figure 4-3, and it appears that the protein sequences with high similarity usually have high similarity between their annotated PFMs.

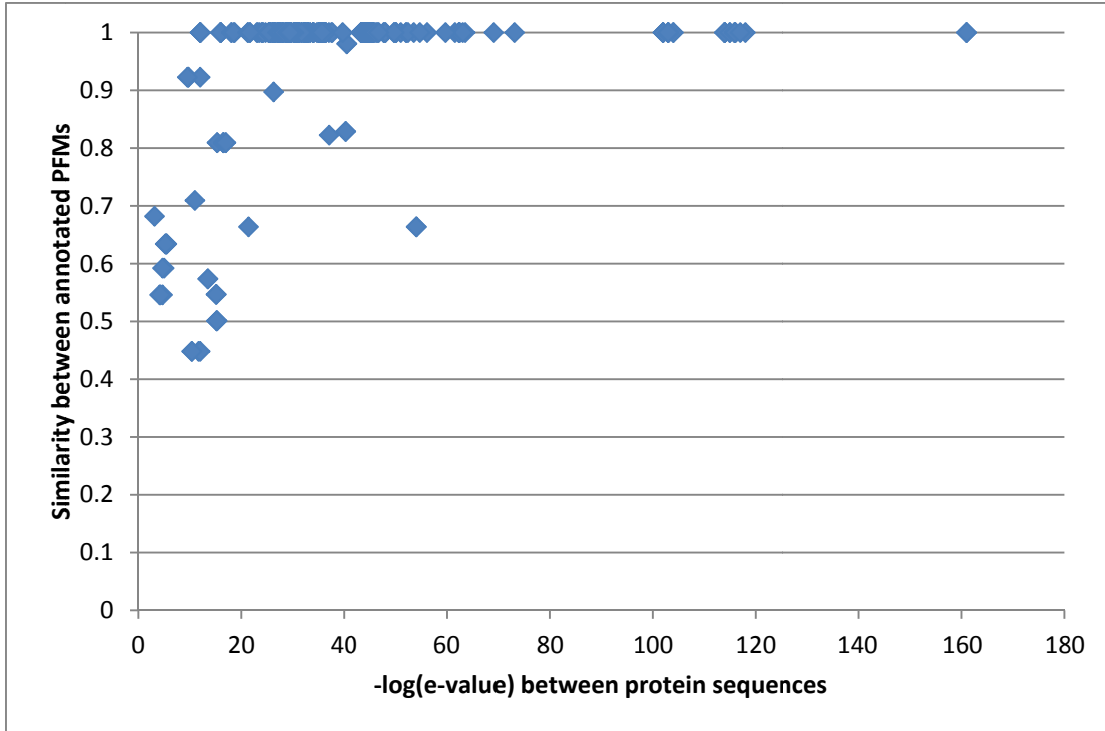


Figure 4-4 Correlation between protein sequence similarities and PFM similarities

4.5.5 Using the number of contact atoms of contact residues

It is expected that the contact residues with more contact atoms are more important. In this regard, this information, the number of contact atoms in a residue is used when building PFMs. There are two ways to exploit this information. One is assigning weights proportional to the number of contact atoms to the contact residues when calculating the scoring matrix M . Another is defining δ when merging the PFM built by DNA sequence in the template and the PFM built by the knowledgebase.

While giving weights by the number of contact atoms of contact residues, the formula in 3.3.2 for calculating the scoring matrix M would become:

$$M(i, n) = \frac{\sum_{a \in \Gamma} S(a, n) \cdot (d_a * c_a)^{-w}}{\sum_{a \in \Gamma} (d_a * c_a)^{-w}} \quad (i = 1 \sim L, n \in \{A, T, C, G\})$$

where c_a is the number of contact atoms of contact residue a , and the other variables

are defined as described in section 3.3.2.

When applying the number of contact atoms of contact residues to merging the PFM built by DNA sequence in the template and PFM built by the knowledgebase together, the variable δ in the formula described in section 3.3.3 is redefined.

$$PFM_r(i, n) = PFM_k(i, n) \cdot (1 - \delta) + PFM_s(i, n) \cdot \delta \quad (i = 1 \sim L, n \in \{A, T, C, G\})$$

where δ is redefined as the number of contact atoms which is substituted over the total number of contact atoms of position i .

Although this is a nice idea, however the two methods described above did not improve the performance, either using them respectively or using them together. The average similarity is still 0.71 under the validation set of protein-DNA complex structures when using the methods described above. This might be caused by the fact that most of the contact residues of the template were not substituted, so the information of DNA sequence is usually kept. Thus slightly adjusting how to use the knowledgebase cannot provide significant improvement.

4.5.6 The frequency of amino acids and nucleotides

While building the knowledgebase which describes the preferences between amino acids of proteins and bases of nucleotides, the frequency of them was applied to calculate the expected contact frequency. Finally the frequencies of amino acids were obtained from Swiss-Prot, and the frequencies of nucleotides were set to 0.25. However, these frequencies were obtained by the contact counts in the 260 non-redundant protein-DNA complex structures at first. The frequency of a type of amino acid or nucleotide amounts to its frequency obtained by the contact counts. The scores between

amino acids and nucleotides are obtained by the same formula described in section 3.2.

The contact counts and scores under this scheme are shown in Figure 4-6 and Figure 4-5 respectively. It is observed that Arginine has lots more contacts with nucleotides than other amino acids. However, under this scoring scheme, Arginine has lower scores. This is caused by the fact that Arginine has lots of contacts so that its expected contact count becomes very large, so Arginine gets lower scores. As a result, this scoring scheme is discarded because it is believed that amino acids with more contacts should have higher score.

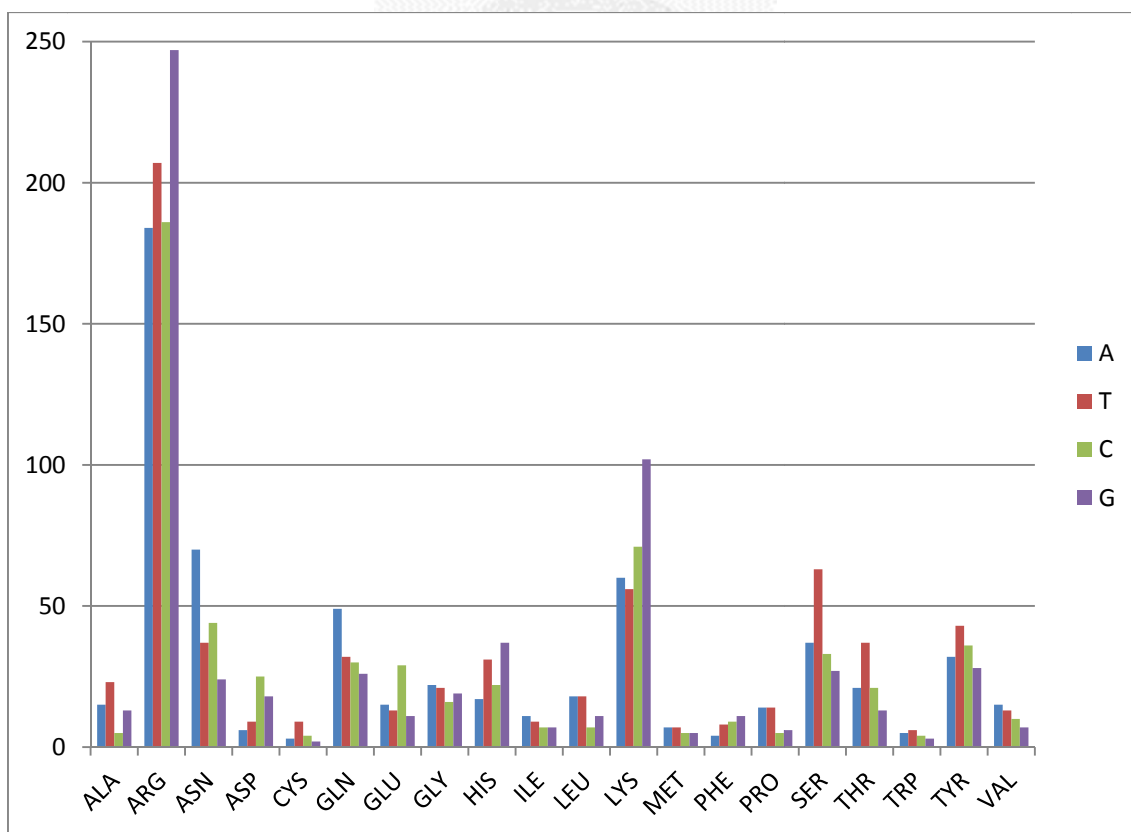


Figure 4-6 Contact counts between amino acids and nucleotides

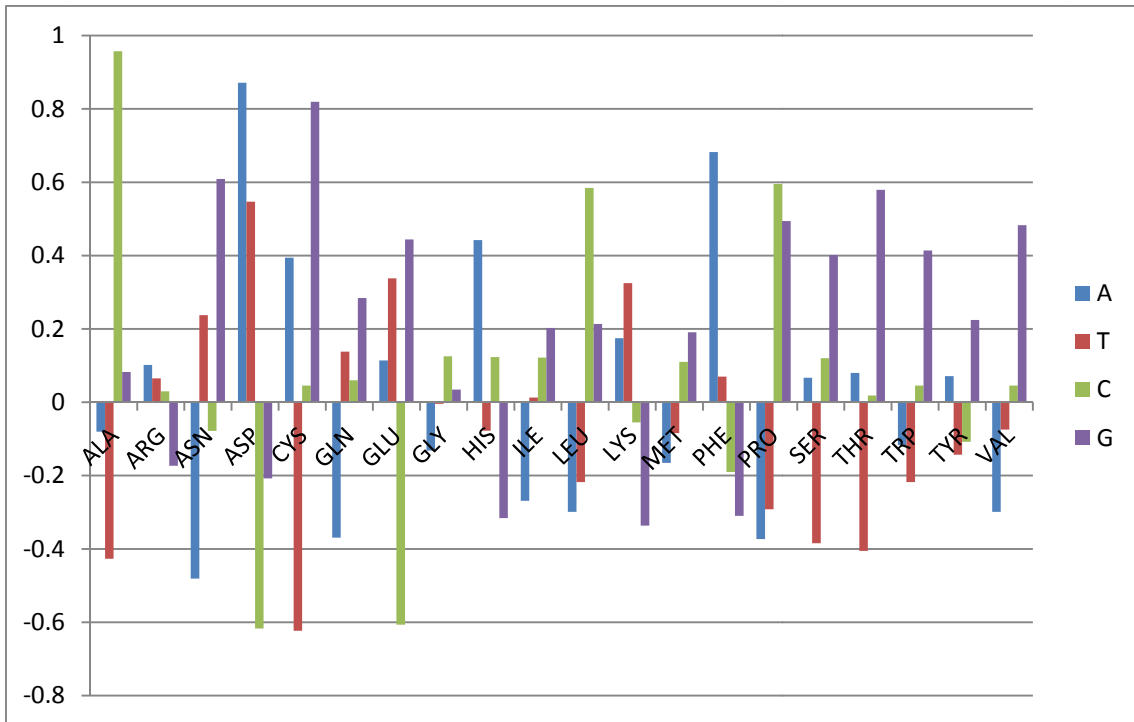
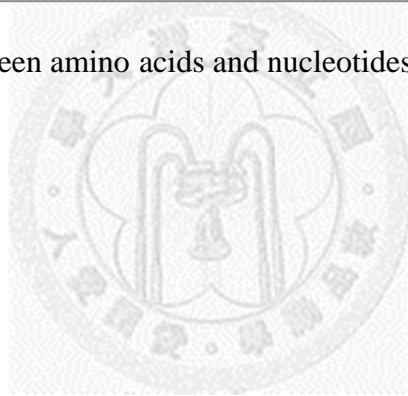


Figure 4-7 Scores between amino acids and nucleotides under the older scheme



Chapter 5 Conclusions

A novel method for predicting target sequences of DNA-binding proteins based on primary structure is proposed in this thesis. Given a query protein primary structure, the proposed method predicts its target sequences based on a knowledgebase that describes the preference between amino acids of proteins and nucleotides of DNA sequences in protein-DNA complex structures.

In the process of this study, different methods and parameters were implemented to achieve a better performance. Different distance cut-offs were implemented to build the knowledgebase of preference between amino acids and nucleotides. As a result, using 4.5Å as distance cut-off achieves the best performance, telling us that only residues close enough to DNA molecules can provide useful information. On the other hand, number of atoms of a contact residue was considered as an important issue. However, it does not have much influence on the performance of the proposed method, because most of the contact residues of the template were not substituted. Because contact residues are seldom substituted, it becomes important to select a proper template when the proposed method is applied. Template can be selected by sequence identity or e-value between two protein sequences, and it turns out that using e-value to select a template of query can achieve a better performance.

As a result, the proposed method is shown to perform well when compared to the structure-based methods [5-7]. However, users need a structure of query protein to predict target sequences by structure-based method. This thesis provides an easier way

for users want to know target sequences of proteins. The proposed method was also compared with SABINE [8], which predicts target sequences of protein by support vector regression (SVR). Although the proposed method cannot predict as well as SABINE does for the validation set, however, there are protein sequences SABINE cannot predict but the proposed method can. This thesis provides another method for when SABINE cannot predict a given protein sequence.



REFERENCE

1. Wrzodek, C., et al., *ModuleMaster: A new tool to decipher transcriptional regulatory networks*. Biosystems, 2010. **99**(1): p. 79-81.
2. Rodionov, D.A., *Comparative genomic reconstruction of transcriptional regulatory networks in bacteria*. Chemical Reviews, 2007. **107**(8): p. 3467-3497.
3. Bonneau, R., et al., *The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo*. Genome Biology, 2006. **7**(5).
4. Alamanova, D., P. Stegmaier, and A. Kel, *Creating PWMs of transcription factors using 3D structure-based computation of protein-DNA free binding energies*. BMC Bioinformatics, 2010. **11**: p. -.
5. Morozov, A.V., et al., *Protein-DNA binding specificity predictions with structural models*. Nucleic acids research, 2005. **33**(18): p. 5781-98.
6. Morozov, A.V. and E.D. Siggia, *Connecting protein structure with predictions of regulatory sites*. Proceedings of the National Academy of Sciences of the United States of America, 2007. **104**(17): p. 7068-73.
7. Zhou, Y.Q., et al., *An all-atom knowledge-based energy function for protein-DNA threading, docking decoy discrimination, and prediction of transcription-factor binding profiles*. Proteins-Structure Function and Bioinformatics, 2009. **76**(3): p. 718-730.
8. Schroder, A., et al., *Predicting DNA-binding specificities of eukaryotic transcription factors*. PloS one, 2010. **5**(11): p. e13876.
9. Berger, M.F., et al., *Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences*. Cell, 2008. **133**(7): p. 1266-1276.
10. Alleyne, T.M., et al., *Predicting the binding preference of transcription factors to individual DNA k-mers*. Bioinformatics, 2009. **25**(8): p. 1012-1018.
11. Wolber, G., et al., *The Protein Data Bank (PDB), Its Related Services and Software Tools as Key Components for In Silico Guided Drug Discovery*. Journal of Medicinal Chemistry, 2008. **51**(22): p. 7021-7040.
12. Robertson, T.A. and G. Varani, *An all-atom, distance-dependent scoring function for the prediction of protein-DNA interactions from structure*. Proteins-Structure Function and Bioinformatics, 2007. **66**(2): p. 359-374.
13. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. Nucleic acids research, 1997. **25**(17): p. 3389-402.
14. Zhang, Y. and J. Skolnick, *TM-align: a protein structure alignment algorithm based on the TM-score*. Nucleic acids research, 2005. **33**(7): p. 2302-2309.
15. Shindyalov, I.N. and P.E. Bourne, *Protein structure alignment by incremental combinatorial extension (CE) of the optimal path*. Protein Engineering, 1998. **11**(9): p. 739-747.
16. Holm, L. and C. Sander, *Protein-Structure Comparison by Alignment of Distance Matrices*. Journal of Molecular Biology, 1993. **233**(1): p. 123-138.
17. Kihara, D. and J. Skolnick, *The PDB is a covering set of small protein structures*.

- Journal of Molecular Biology, 2003. **334**(4): p. 793-802.
18. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality*. Proteins-Structure Function and Bioinformatics, 2004. **57**(4): p. 702-710.
 19. Matys, V., et al., *TRANSFAC (R) and its module TRANSCompel (R): transcriptional gene regulation in eukaryotes*. Nucleic acids research, 2006. **34**: p. D108-D110.
 20. Tsai, H.K., et al., *MYBS: a comprehensive web server for mining transcription factor binding sites in yeast*. Nucleic acids research, 2007. **35**: p. W221-W226.
 21. Chan, W.M. and U. Consortium, *The UniProt Knowledgebase (UniProtKB): a freely accessible, comprehensive and expertly curated protein sequence database*. Genetics Research, 2010. **92**(1): p. 78-79.
 22. Chen, C.Y., W.C. Chung, and C.T. Su, *Exploiting homogeneity in protein sequence clusters for construction of protein family hierarchies*. Pattern Recognition, 2006. **39**(12): p. 2356-2369.
 23. Tsai, H.K., et al., *Method for identifying transcription factor binding sites in yeast*. Bioinformatics, 2006. **22**(14): p. 1675-81.

