

國立臺灣大學電機資訊學院資訊工程學系

碩士論文

Department of Computer Science and Information Engineering

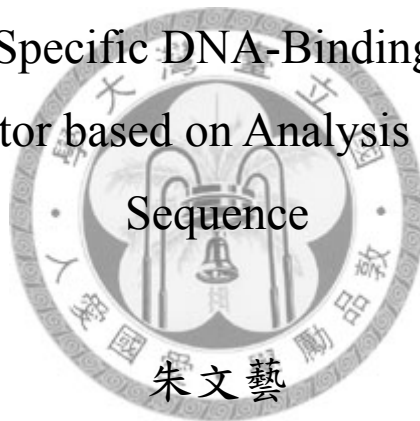
College of Electrical Engineering & Computer Science

National Taiwan University

master thesis

預測轉錄因子與去氧核糖核酸作用之專一性結合殘基

Prediction of Specific DNA-Binding Residues in a  
Transcription Factor based on Analysis of the Polypeptide  
Sequence



Chu Wen-Yi

指導教授：歐陽彥正 博士

Advisor: Oyang Yen-Jen, Ph.D.

中華民國 97 年 6 月

June, 2008

## 誌謝

兩年的碩士生涯很快就過去了，然而相信在這段時間裡所遇見的人與事，已發生的各式各樣奇妙的見聞或者反省，都令自己成長了不少。而這些片段也一定會成爲回憶裡的養分，在未來偶爾駐足歇息時偷偷取出。

由於歐陽教授的指導，自己可以很大膽地嘗試許多的想法，雖然過程中會不斷地遇到障礙和挫折，但是大家的激勵和創意使得這項研究終於成型，有了可見的成果。我想自己這一輩子都不會忘記會議室裡教授爽朗的笑聲。



很感謝從開始到最後都給了我許多幫助的中才學長，許多事後自己都覺得好笑、不成熟的靈感，他都能夠耐心地傾聽甚至給予讚美，學長真的是相當友善慷慨的人。

還要感謝黃乾綱教授、陳倩瑜教授、張天豪教授不時的關心以及建議。倩瑜老師還大方地借了我一把很棒的小提琴，可惜我無法說「就讓我來拉一曲以表達我的感謝吧」，因爲最近都沒有練琴，這樣做會讓我良心不安。最後，得感謝實驗室裡的成員，鈺峰、諭承、翊鍾、右昇、智棚、志宏、振傑、凱維、孟翰、廷因、凱勳、榮元、玫如。

## 摘要

本篇論文旨在設計一個能從多肽序列擷取資訊的自動分類器，以預測轉錄因子上會與 DNA 之鹼基產生鍵結的殘基。正如最近一些研究所揭露的，有大量轉錄因子之三級結構是不穩定序(disordered)，因此若能只純粹利用序列的資訊，進而預測轉錄因子與 DNA 之關鍵殘基，將非常有助於下一步的實驗。

有鑑於此，設計、發展一個預測器並使之能夠分辨與 DNA 進行專一性結合的殘基更形迫切需要。此外，專一性結合不僅能反應基因上的特異序列辨識，在正確的基因調控中也扮演極重要的角色。



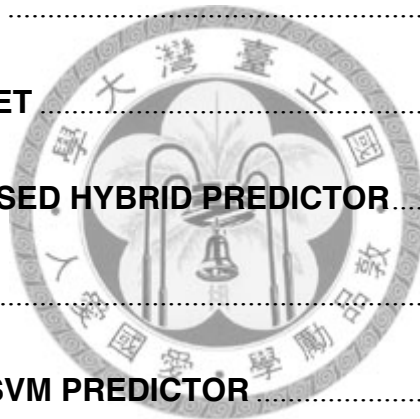
本論文呈現的設計混合了兩種不同方法，分別是以 SVM 為主的機器學習方式以及原先應用於預測蛋白質域(protein domain)的演算法。觀察後發現兩個方法的預測表現在不同的蛋白質二級結構上各有優劣，於是我們嘗試設計一套機制以混合兩種方法的輸出結果以取得最佳的成績。在本文最終的實驗結果，呈現的新預測器能提供 59.5%的涵蓋度、77.4%的精確度，以及 98.8%的專一度(specificity)。

## ABSTRACT

This thesis presents the design of a polypeptide sequence based predictor aiming to identify the residues in a transcription factor that are involved in specific binding with the DNA. As a recent study has revealed that the tertiary structures of a large number of transcription factors are mostly disordered, the capability to identify the residues in a transcription factor that play key roles in interaction with the DNA based purely on analysis of the polypeptide sequence is highly desirable. In this respect, it is further desirable to have a predictor capable of distinguishing those residues involved in specific binding with the DNA, since specific binding corresponds to sequence-specific recognition of a gene, which is essential for correct gene regulation. The design of the proposed predictor is distinctive by employing a hybrid approach. That is, two prediction mechanisms specialized for making predictions in different types of protein secondary structures have been incorporated. In the experiments reported in this thesis, the proposed hybrid predictor delivered precision of 77.4%, sensitivity of 59.5%, and specificity of 98.8%

# Contents

誌謝.....	I
摘要.....	II
ABSTRACT.....	III
Chapter 1 INTRODUCTION.....	1
Chapter 2 RELATED WORK.....	7
2.1 OVERVIEW.....	7
2.2 CLASSIFIER.....	10
2.3 FEATURE SET.....	15
Chapter 3 THE PROPOSED HYBRID PREDICTOR.....	17
3.1 OVERVIEW.....	18
3.2 PRIMARY SVM PREDICTOR.....	20
3.3 AUXILIARY SSEP PREDICTOR.....	21
Chapter 4 EXPERIMENTAL RESULTS.....	25
4.1 DESIGN OF EXPERIMENTS.....	25
4.2 RESULTS AND DISCUSSIONS.....	26
Chapter 5 CONCLUSIONS AND FUTURE WORKS.....	36
REFERENCES.....	38
APPENDIX.....	43



## List of Figures

Fig. 1 Different mechanisms of specific and non-specific binding .....	6
Fig.2 Basic idea of SVM working principals .....	11
Fig.3 The overview of the proposed hybrid method .....	17
Fig. 4 Ideas of the merging process .....	24
Fig. 5 An example of TF-DNA interaction with PDB ID 1YSA.....	30
Fig .6 A prediction result of query protein chain with PDB ID 1YSA.....	31
Fig. 7 An example of TF-DNA interaction with PDB ID 1BDV .....	32
Fig.8 An example of TF-DNA interaction with PDB ID 1RPE .....	33
Fig.9 An example of TF-DNA interaction with PDB ID 1LAT.....	35
Fig. 1 Different mechanisms of specific and non-specific binding .....	6
Fig.2 Basic idea of SVM working principals .....	11
Fig.3 The overview of the proposed hybrid method .....	17
Fig. 4 Ideas of the merging process .....	24
Fig. 5 An example of TF-DNA interaction with PDB ID 1YSA.....	30
Fig .6 A prediction result of query protein chain with PDB ID 1YSA.....	31
Fig. 7 An example of TF-DNA interaction with PDB ID 1BDV .....	32
Fig.8 An example of TF-DNA interaction with PDB ID 1RPE .....	33

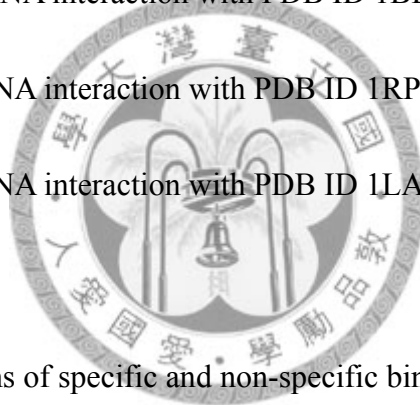


Fig.9 An example of TF-DNA interaction with PDB ID 1LAT..... 35

## List of Tables

Table 1. Prediction results with the SVM based primary predictor..... 26

Table 2. Prediction results with the SSEA based auxiliary predictor ..... 27

Table 3. Prediction results with the hybrid predictor..... 28

Table 4. Breakdown of the prediction results with the hybrid predictor in respect of  
different types of TF-DNA interactions..... 28



## Chapter 1 INTRODUCTION

In the field of molecular biology, a transcription factor (sometimes called a sequence-specific DNA binding factor) is a protein that binds to specific parts of DNA using DNA binding domains and is part of the system that controls the transfer (or transcription) of genetic information from DNA to RNA [1, 2]. The importance of transcription factors lies in that, without transcription factors, the creation of new RNA from DNA cannot occur. Transcription factors perform this function alone, or by using other proteins in a complex, by increasing (as an activator), or preventing (as a repressor) the presence of RNA polymerase, the enzyme which activates the transcription of genetic information from DNA to RNA [3-5]. Transcription factors are one of the groups of proteins that read and interpret the genetic "blueprint" in the DNA. They are the key to determining where the DNA chain becomes "unzipped," creating a single strand to which RNA can be bound while it's being built. They bind DNA and help initiate a program that decreases or increases gene transcription. As such, they are vital for many important cellular processes. Transcription factors at least involve in these important functions and biological roles below:

**Basal transcription regulation [6, 7]:** In eukaryotes, an important class of transcription factors called general transcription factors (GTFs) which are necessary for transcription to occur. Many of these GTFs don't actually bind DNA but are part of the large



transcription preinitiation complex that interacts with RNA polymerase directly. These GTFs are, for example, TFIIA, TFIIB, and TFIID.

**Development:** Many TFs in multicellular organisms are involved in development.

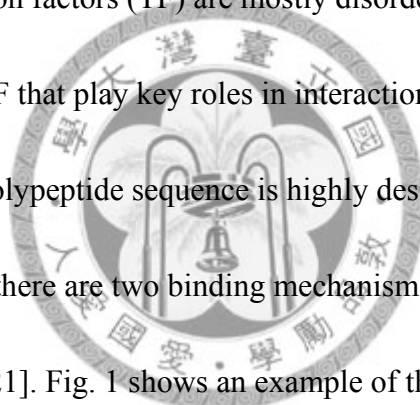
Responding to cues (stimuli), these TFs turn on or turn off the transcription of the appropriate genes which in turn allows for changes in cell morphology or activities needed for cell fate determination and cellular differentiation. For example, the Hox transcription factor family (which is important for proper body pattern formation) and the TF encoded by the Sex-determining Region Y (SRY) gene [8] (which plays a major role in determining gender in humans) can be regarded in this category.

**Response to environment:** Not only do transcription factors act downstream of signaling cascades related to biological stimuli, but they can also be downstream of signaling cascades involved in environmental stimuli. Examples include heat shock factor (HSF) [9] which upregulates genes necessary for survival at higher temperatures, hypoxia inducible factor (HIF) [10, 11] which upregulates genes necessary for cell survival in low oxygen environments, and sterol regulatory element binding protein (SREBP) which helps maintain proper lipid levels in the cell.

**Cell cycle control:** Many transcription factors, especially some that are oncogenes or tumor suppressors, help regulate the cell cycle and as such determine how large a cell will get and when it can divide into two daughter cells. One example is the Myc [12]

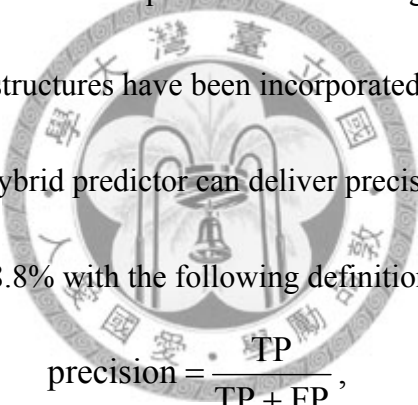
oncogene, which has important roles in cell growth and apoptosis.

In recent years, prediction of residues in a protein chain that may be involved in interaction with the DNA has been a research topic that attracts a high level of interest [13-15]. Some of the studies were purely based on analysis of the polypeptide sequence [13, 15-18], while the others took the structural information into account [17, 19]. In this respect, as it has been reported in a recent article that the tertiary structures of a large number of transcription factors (TF) are mostly disordered [20], the capability to identify the residues in a TF that play key roles in interaction with the DNA based purely on analysis of the polypeptide sequence is highly desirable. Concerning protein-DNA interactions, there are two binding mechanisms involved: specific binding and non-specific binding [21]. Fig. 1 shows an example of these two binding mechanisms. Specific binding occurs between protein side chains and nucleotide bases, while non-specific binding occurs between protein side chains and the DNA sugar-phosphate backbone. In this thesis, we define the residues with heavy atoms which are within 4.5 Å from the bases of the DNA as the “specific-binding residues”, or “base-specific binding residues”. According to previous literature, many had set the distance with the range from 4 Å to 6 Å. With the point not to include too much noise and retain fidelity, we set the distance to 4.5 Å in determining a base-specific binding



residue.

In molecular biology, specific binding corresponds to sequence-specific recognition of a gene and therefore is essential for correct gene regulation. Therefore, in this thesis, we have aimed to design a polypeptide sequence based predictor capable of identifying those residues in a TF that are involved in specific binding with the DNA. The design of the proposed predictor is distinctive by employing a hybrid approach. In the hybrid predictor, two prediction mechanisms specialized for making predictions in different types of protein secondary structures have been incorporated. Based on the experiments reported in this thesis, the hybrid predictor can deliver precision of 77.4%, sensitivity of 59.5%, and specificity of 98.8% with the following definitions:


$$\text{precision} = \frac{TP}{TP + FP},$$

$$\text{sensitivity} = \frac{TP}{TP + FN},$$

$$\text{specificity} = \frac{TN}{TN + FP},$$

where

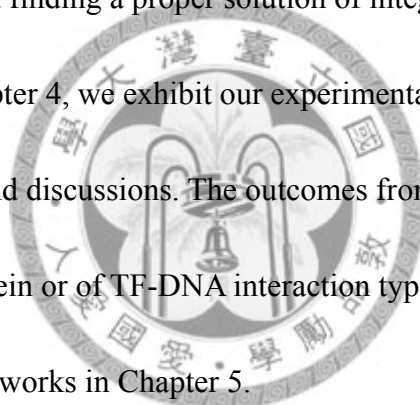
TP = the number of correctly classified specific binding residues;

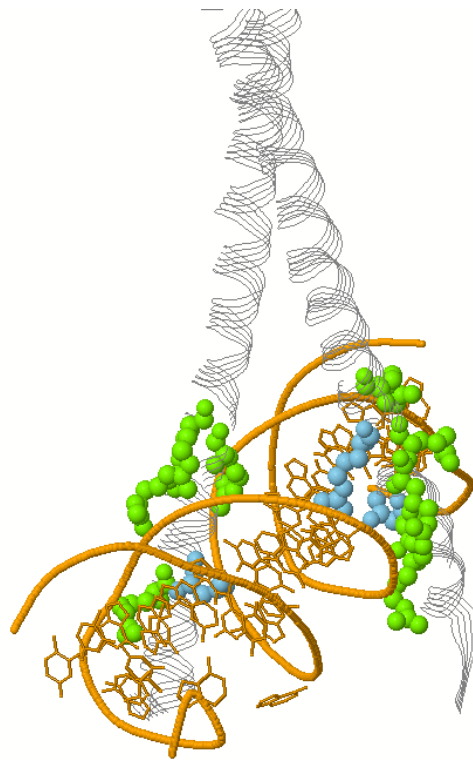
TN = the number of correctly classified non-specific binding residues;

FP = the number of non-specific binding residues incorrectly classified as specific;

FN = the number of specific binding residues incorrectly classified as non-specific.

This thesis is organized as follows: in Chapter 2 we introduce the related works in which many researchers had studied the characteristics of DNA-binding proteins and the properties of DNA-binding residues. In wake of this, many predictors have been developed in order to provide clues for the correct binding sites on proteins. In Chapter 3, we first exhibit the overall architecture of the proposed hybrid predictor with the purpose that the operational flow can help one easily understand the process much more clearly. Then we describe the primary SVM predictor and the auxiliary SSEP predictor. The way to utilize both and finding a proper solution of integration is elaborated in the end of this chapter. In Chapter 4, we exhibit our experimental results. It contains numerous tables, figures and discussions. The outcomes from the view point of secondary structure of protein or of TF-DNA interaction type are also included. We give our conclusions and future works in Chapter 5.





Jmol

Fig. 1 Different mechanisms of specific and non-specific binding. This picture shows the specific binding and non-specific binding residues. The residues have specific binding occur between protein side chains and nucleotide bases are colored in blue. The residues have non-specific binding occur between protein side chains and the DNA sugar-phosphate backbone are colored in green.

## Chapter 2 RELATED WORK

In this chapter, we first introduce the related studies and the history concerning DNA-binding site prediction. We introduce two different systems, one is based on protein sequence and the other is based on three-dimensional structure of protein. Then we illustrate the ideas of the SVM classifier and the SSEP-Domain algorithm, both are critical components involved in our solution. Finally, we describe the feature sets including PSSMs and predicted secondary structure, which are the learning materials for our approach to capture the characteristics of base-specific binding residues.

### 2.1 OVERVIEW

Many studies have recently attempted to use structural and even sequence properties of unbound proteins to predict protein-protein interface [13-17, 19, 22]. Ahmad *et al.* [16] made first attempt to adopt sequence and evolutionary features for predicting DNA-binding sites in DNA-binding proteins. Since then, some methods using three-dimensional structure of the protein were also proposed [15, 19, 22]. Though those articles deal with not exactly the same problem as we are addressing (base-specific binding residues of proteins), there are some standard procedures and useful mechanisms which can't be ignored. In the following paragraphs, we will go through those DNA-binding site predictors and have basic knowledge of their working

principles.

For DNA-binding site predictors which are based on structural information of proteins, there are some studies: in 2003, Jones *et al.* [17] used electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. They collected a dataset composed of 56 non-homologous DNA-binding proteins and achieved 68% correction prediction of the dataset. The result reveals that the electrostatic potentials are strong and significant features for prediction. Later in 2004, Tsuchiya *et al.* [19] proposed a structure-based predictor for DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. They obtained the electrostatic potentials by solving the Poisson-Boltzmann equation numerically, which may consume a lot of time for calculating. They focused on 63 protein-DNA complexes and in the end developed prediction schemes with 86% and 96% accuracy for DNA-binding and non-DNA-binding proteins, respectively. It is interesting that the result still supports the importance of electrostatic potentials on protein surfaces.

Recently in 2007, Tjong , H. and Zhou, H.-X [15] analyzed 264 protein-DNA complexes, rather than electrostatic potentials, they featured position-specific sequence profiles, solvent accessibilities of each residue and its spatial neighbors to the neural network. Overall, they claimed the predictor achieved accuracy over 80% and coverage

over 60% of actual DNA-contacting residues.

For those which are based on sequence and evolutionary information of the proteins, in other words, the three-dimensional structure of the query protein is not required, take some for representatives: DISIS [14], BindN [23], and DP-Bind [24]. We will give a brief description of these tools below.

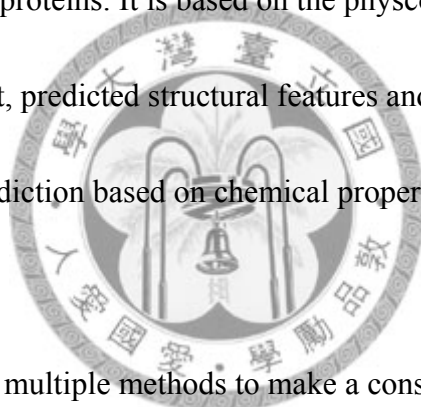
1) **DISIS** [14] predicts DNA-binding sites directly from amino acid sequence and hence is applicable for all known proteins. It is based on the physicochemical properties of the residue and its environment, predicted structural features and evolutionary data.

2) **BindN** [23] makes a prediction based on chemical properties of the input protein sequence.

3) **DP-Bind** [24] combines multiple methods to make a consensus prediction based on the profile of evolutionary conservation and properties of the input protein sequence.

Profile of evolutionary conservation is automatically generated by this web-server.

As mentioned in the introduction, a large number of transcription factors (TF) may be mostly disordered, which may make it difficult for researchers in the laboratories to have those TFs crystallized or structured. This is the reason why it's necessary and crucial to develop an accurate specific-binding site predictor from sequence only.



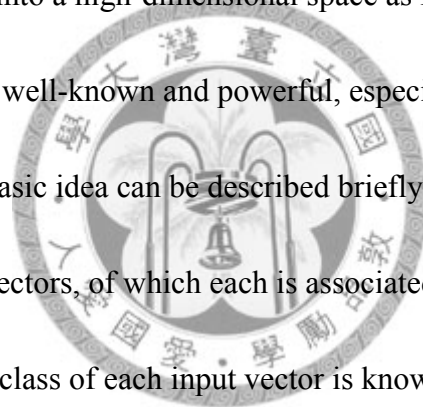


## 2.2 CLASSIFIER

In this thesis, we hybridize two learning mechanisms to obtain optimal performance.

One is LIBSVM [25] and the other one is SSEP-Domain predictor. In this section we will focus on introducing the core concepts of these two algorithms, leaving their detailed operations during methods later in chapter 3.

LIBSVM is a learning machine based on statistical learning theory. The advantages of translating the training set into a high-dimensional space as Fig. 2 and avoiding overfitting make LIBSVM well-known and powerful, especially in addressing biological problems. The basic idea can be described briefly as follows. First, the inputs are formulated as feature vectors, of which each is associated with one of two classes. In the training procedure, the class of each input vector is known in advance. In prediction, the class is the output of SVM. Secondly, the feature vectors are mapped into a feature space (possibly with high dimensionality) by a kernel function, either linearly or nonlinearly. Thirdly, a division is computed in the feature space to optimally separate the two classes of training vectors. SVM training always automatically seeks global optimum and avoids over-fitting. These characteristics make it particularly suitable to deal with large numbers of features.



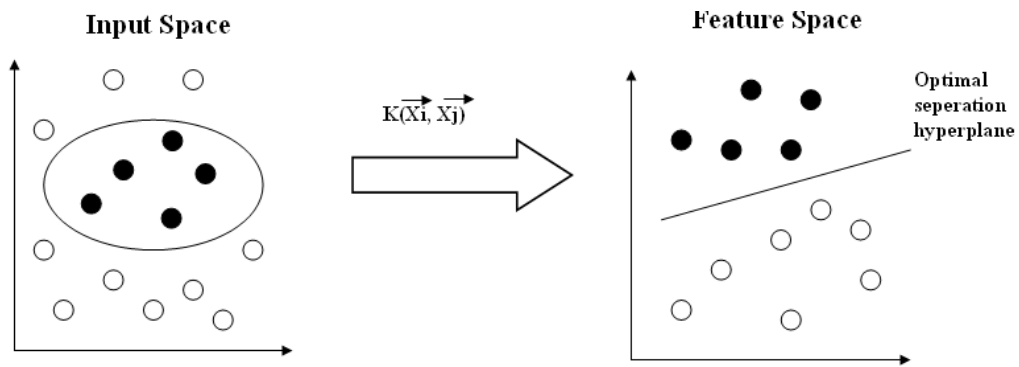
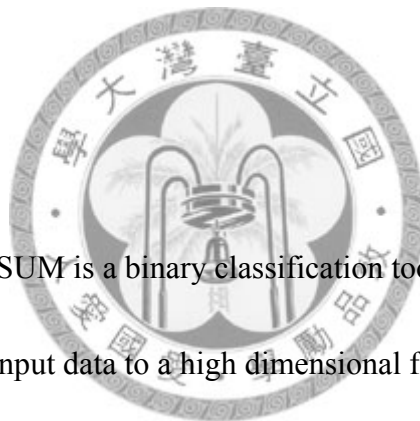


Fig.2 Basic idea of SVM working principals. The basic idea of SVM is to employ a mapping function to transform data from the input space to a feature space where the border can be represented by a linear optimal separation hyperplane.



As mentioned in pervious, SVM is a binary classification tool that uses a non-linear transformation to map the input data to a high dimensional feature space where linear classification is performed. It is equivalent to solving the quadratic optimization

problem:

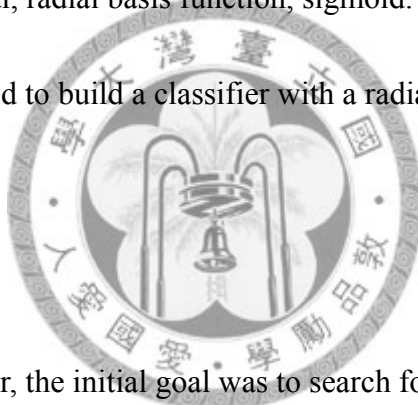
$$\min_{w, b, \zeta_i} \frac{1}{2} w \cdot w + C \sum_i \zeta_i \quad (1)$$

Such that,

$$y_i (\Phi(x_i) \cdot w + b) \geq 1 - \zeta_i \quad i = 1, \dots, m, \quad (2)$$

$$\zeta_i \geq 0 \quad i = 1, \dots, m$$

Where  $x_i$  is a feature vector labeled by  $y_i \in \{+1, -1\}$ ,  $\{x_i, y_i\}$ ,  $i=1, \dots, m$ , and the penalty parameter of the error term  $C$ , also called the cost. It does the classification by generating a separating hyper-plane using the equation  $f(x) = \Phi(x) \cdot w + b = 0$ . Use  $\sum_j \alpha_j \Phi(x_j)$  to represent  $w$ , we obtain  $\Phi(x_i) \cdot w = \sum_j \alpha_j \Phi(x_j) \cdot \Phi(x_i)$ . This provides an efficient approach to solve SVM without the explicit use of the non-linear transformation. Further  $K(x_i, x_j) \equiv \Phi(x_i)^T \Phi(x_j)$  is called the kernel function and it is this function that maps the data to a higher dimension. There are several kernel types including linear, polynomial, radial basis function, sigmoid. In this thesis, public available LIBSVM was used to build a classifier with a radial basis function kernel and a set of parameters.



For SSEP-Domain predictor, the initial goal was to search for protein domain boundary and to conduct domain prediction. Because of its high sensitivity and precision of single domain, we adopted its idea and made some modification to suite our purpose of finding specific-binding sites. The fundamentals of the whole process include two major steps. The first step is searching for domain boundary, picking up suitable and significant domains in the template library created beforehand; the other one is scoring of domain regions, which introduces a technique called profile-profile alignment (PPA). The details of the implementation are elaborated by psudeo-code of this algorithm:

## Step 1: Domain Boundary Search

```
1: // initialization

2: Centers ← centers of coil regions predicted on target t

3: Regions ← {rij = t[ci..cj]} | ci, cj ∈ Centers ∧ ci < cj}

4: Images ← {}

5: PFAM_DNAbindings ← Domains annotated as DNAbinding by PFAM

6: Domains ← PFAM_DNAbindings

7: // generation of domain images

8: For all template domains d ∈ Domains do

9:     // get highest scoring region of similar length

10:    smax(d) ← maxrij ∈ Regions, |rij| ≈ |d| SSEA(d, rij)

11:    //significance filtering: score high enough?

12:    if smax(d) > sthresh(d) then

13:        add corresponding region rij to Images

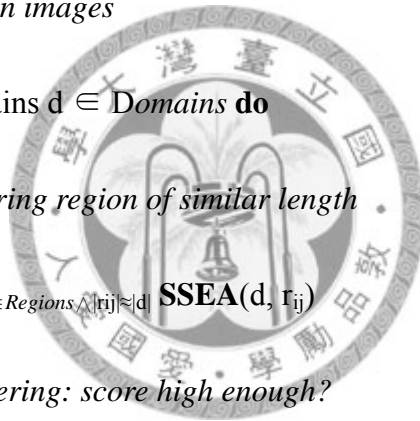
14:        with score(rij) ← smax(d)

15:    end if

16: end for

17: // accumulative scoring of coil centers

18: ∀ c ∈ Centers : score(c) ← 0
```



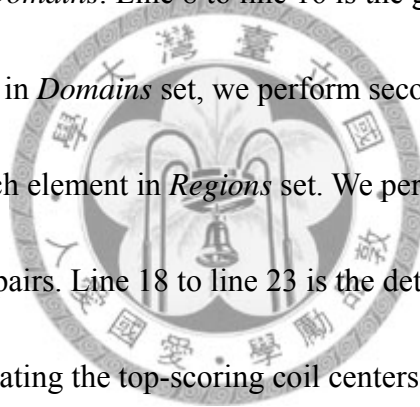
```

19: for the top-scoring  $r_{ij} \in Images$  do
20:     score( $c_i$ )  $\leftarrow$  score( $c_i$ ) + score( $r_{ij}$ )
21 :     score( $c_j$ )  $\leftarrow$  score( $c_j$ ) + score( $r_{ij}$ )
22: end for
23: select the top-scoring coil centers

```

Line 2 to line 6 is initialization. In this we create several sets: *Centers*, *Regions*, *Images*, *PFAM\_DNAbindings* and *Domains*. Line 8 to line 16 is the generation of domain images.

For every template domain in *Domains* set, we perform secondary structure element alignment (SSEA) with each element in *Regions* set. We perform significance filtering by discarding low-scoring pairs. Line 18 to line 23 is the determination of proper domain boundary by calculating the top-scoring coil centers.



## Step 2: Scoring of Domain Regions

```

1: Regions  $\leftarrow$  potential domain regions
2: For all  $r \in Regions$  do
3:     // score fold classes by highest-scoring members
4:     for all fold classes  $Fold \subset Domains$  do
5:         score( $Fold$ )  $\leftarrow$   $\max_{d \in Fold \wedge |d| \approx |r|} SSEA(r, d)$ 

```

```

6:      end for

7:      // select members of potential fold classes

8:       $D_{\text{top}} \leftarrow$  members of top-scoring fold classes

9:      // score normalization for multiplicative scoring

10:      $\text{score}_{\text{raw}}(r) \leftarrow \max_{d \in \text{Fold} \wedge |d| \approx |r|} \text{PPA}(r, d)$ 

11:      $\text{score}_{\text{final}}(r) \leftarrow \text{score}_{\text{raw}}(r) / (10 \log |r|)$ 

12: end for

```

Line 3 to line 6 we perform SSEA between each element in Fold set and each element in Regions set. Line 8 to line 12 the potential fold classes were selected by performing profile-profile-alignment (PPA).

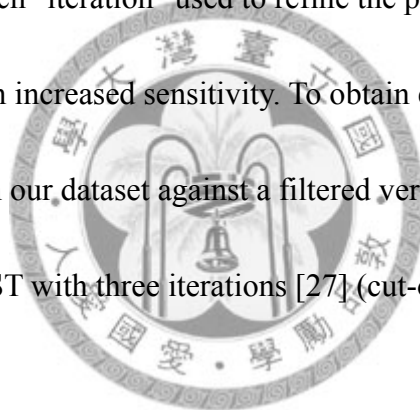


## 2.3 FEATURE SET

In our study, we use two strong features to capture the characteristics of base-specific interaction residues. They are position specific scoring matrix (PSSM) and secondary structures of the protein. The power of the first feature may be due to its enriched conservation information of a protein chain. And the power of the second feature may contribute to the stability of the structures and the unique composition of secondary structure elements which may reveal the preference of base-specific interaction.

Position specific iterative BLAST (PSI-BLAST) [26] was executed and a profile (or position specific scoring matrix, PSSM) is constructed (automatically) from a multiple alignment of the highest scoring hits in an initial BLAST search. The PSSM is generated by calculating position-specific scores for each position in the alignment.

Highly conserved positions receive high positive scores and weakly conserved positions receive scores near zero or negative. The profile is used to perform a second BLAST search and the results of each "iteration" used to refine the profile. This iterative searching strategy results in increased sensitivity. To obtain evolutionary profiles, we first aligned each protein in our dataset against a filtered version of all currently known sequences using PSI-BLAST with three iterations [27] (cut-off at 10<sup>-3</sup>).



HYPROSPII [28] is a knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence. PSIPRED [26] is another protein secondary structure predictor which is based on position-specific scoring matrices. We used both outputs to the evaluation. Outcomes showed HYPROSPII was more accurate in predicting  $\beta$ -sheet segments which is crucial for some specific DNA-binding sites. In this respect, we submitted each protein chain to HYPROSPII and then its generated secondary structure profile was stored.

### Chapter 3 THE PROPOSED HYBRID PREDICTOR

In this chapter, we first exhibit the overall structure of the proposed method. The observation during the experiments that evolves and results in the hybrid method will also be elaborated. Then we introduce the primary SVM predictor and auxiliary SSEP predictor, their performance will also be discussed. Concerning the problem of integration, the solution to combine and utilize two different mechanisms will be described in the end of this chapter.

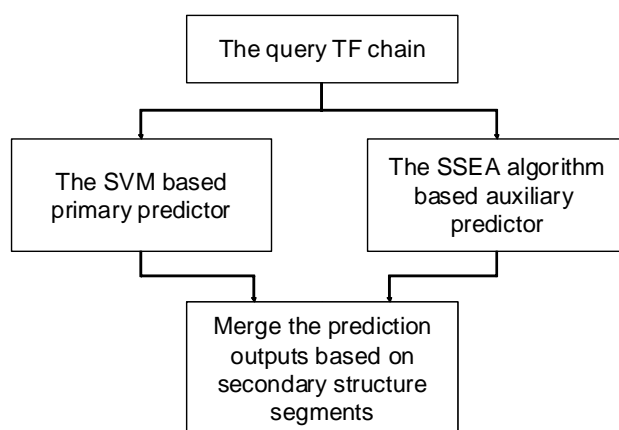


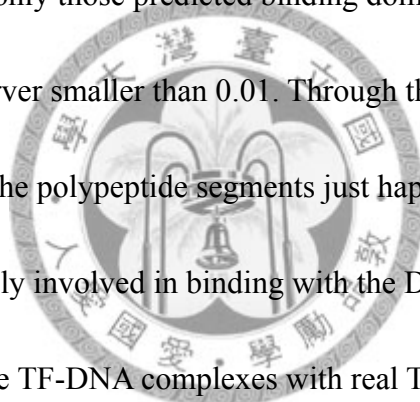
Fig.3 The overview of the proposed hybrid method. Each testing case was delivered to two different predictors, SVM and SSEP-Domain, simultaneously. It is their innate design that SVM and SSEP-Domain may learn the characteristics of specific-binding sites in different way.



### 3.1 OVERVIEW

Fig. 3 presents an overview of the hybrid predictor proposed in this thesis. The entire hybrid predictor consists of the primary predictor and the auxiliary predictor. The primary predictor is a support vector machine (SVM) with its parameter settings optimized for delivering high precision. As a result, one can expect that sensitivity of the SVM-based primary predictor is traded, since one common phenomenon in tuning the parameters of a predictor is that raising precision typically means that sensitivity is traded and vice versa. In fact, it has been observed in our experiments that the SVM with the parameter settings employed in this thesis is capable of delivering reasonably well precision with respect to identifying those residues in  $\alpha$ -helix and coil types of secondary structures that are involved in specific binding with the DNA. On the other hand, it has also been observed that the SVM hardly identifies the residues in a  $\beta$ -sheet segment that are involved in specific binding with the DNA. Therefore, one straightforward way to improve the overall sensitivity of prediction is to incorporate a mechanism that can accurately identify those binding residues in a  $\beta$ -sheet segment. As shown in Fig. 3, in the proposed hybrid predictor, we have incorporated a mechanism based on secondary structure element alignment (SSEA) to complement the prediction power of the SVM. The hybrid predictor then take the union of the predicted binding residues output by the primary and auxiliary predictors as its output.

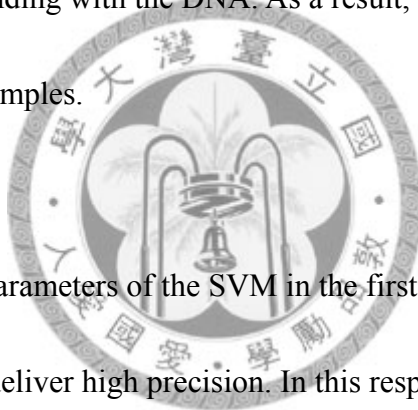
For training the hybrid predictor presented in Fig. 3, we have created a data set containing 228 TF-DNA complexes extracted from the 691 protein-DNA complexes that Yanay Ofran *et al.* [14] collected from the protein data bank (PDB) [29]. In this process, we included only those complexes in the Ofran collection that contain a TF. We then queried the PFAM server [30] to exclude those complexes in which no polypeptide segment is within the DNA-binding domains predicted by the PFAM server. In this respect, we submitted the full sequences of the proteins in the complexes to the PFAM server and adopted only those predicted binding domains with the p-value computed by the PFAM server smaller than 0.01. Through this process, we excluded those complexes in which the polypeptide segments just happen to be in the proximity of the DNA but are not really involved in binding with the DNA. It might happen that we accidentally exclude some TF-DNA complexes with real TF-DNA interactions. Nevertheless, it is our intention to be conservative. In the end, 228 out of the 691 complexes initially in Ofran collection remained. This collection of 228 TF-DNA complexes is then adopted to generate the training data set and testing data set in the experiments reported in this thesis.



### 3.2 PRIMARY SVM PREDICTOR

For the design of the primary predictor, we have employed the LIBSVM [25] package with the Gaussian kernel. The model of the SVM has been generated based on a training data set derived from the data set containing 228 TF-DNA complexes described above.

The training data set was generated by associating each residue in the 228 protein chains with a position specific scoring matrix (PSSM) computed by the PSI-BLAST package with window size set to 11 [15]. In addition, each residue was labeled based on whether it is involved in specific binding with the DNA. As a result, the training data set contains a total of 22097 samples.



As mentioned earlier, the parameters of the SVM in the first stage of the proposed predictor have been set to deliver high precision. In this respect, we have set parameters  $C$  and  $g$  with the Gaussian kernel to 32 and 0.03125, respectively.

### 3.3 AUXILIARY SSEP PREDICTOR

As mentioned earlier, the auxiliary predictor has been designed with a mechanism based on secondary structure element alignment (SSEA) and profile-profile alignment (PPA), which was firstly proposed in CASP 6 and CASP 4 [31]. The kernel of the SSEA-based mechanism refers to a template library containing  $\beta$ -sheet segments involved in specific binding with DNAs. The template library has been created with the following steps.

(1) Each protein chain in the data set containing 228 TF-DNA complexes was submitted to the HYPROSP II server, which is a predictor of protein secondary structures.

Then, each residue in the predicted  $\beta$ -sheet segments was examined to determine whether it is involved in specific binding with the DNA.

(2) Each DNA-binding domain with one or more  $\beta$ -sheet segments involved in specific binding with DNA was deposited into the template library and each residue in the domain was labeled by the HYPROSP II as one of the following three types of residues:  $\alpha$ -helix,  $\beta$ -sheet, and turn.

With the template library, we then can invoke the following procedure to predict the specific binding residues in  $\beta$ -sheet segments of the query transcription factor.

(1) Invoke the HYPROSP II server to label each residue in the query transcription factor with one of following three types:  $\alpha$ -helix,  $\beta$ -sheet, and turn.

(2) Invoke the BLAST package [26] to align the sequence of the labels of the query transcription factor with the sequence of labels of each template in the library. The similarity score between the query TF and a template is then computed as follows.

$$score = \log \sum_{i=1}^{20} \sum_{j=1}^{20} \alpha_i^T \beta_j \frac{P_{rel}(i,j)}{P_i P_j} \quad (3)$$

The principle is to compare two PSSM matrices of two aligned protein chain sequences.

The higher the score is, the more similarity these two PSSM matrices have. In this respect, we obtained the score by calculating equation (1), where  $\alpha_i$  is a row vector of PSSM representing the characteristics of an amino acid, and  $\beta_j$  is a row vector of PSSM representing the characteristics of an amino acid which is aligned to  $\alpha_i$ .  $P_{rel}$  is a function which can be derived from BLOSUM62 to indicate the preference to substitute the type of amino acid  $i$  to the type of amino acid  $j$ . The denominators,  $P_i$  and  $P_j$  stand for background probability of the occurrence of amino acid  $i$  and amino acid  $j$ , which could also be derived from BLOSUM62. It is important to understand that the score was calculated by two aligned, corresponding residues, and to make the work finished, we have to sum up all individual scores to make it meaningful. At this moment we are spending time elaborating on the idea of scoring function but the fraction  $P_{rel}(i,j)/P_i P_j$ , which has important statistical and biological meanings, is still missing. Thus we tried

to find practical ways to solve the problem as follows.

$$S(a,b) = \frac{1}{\lambda} \log \frac{P_{ab}}{P_a P_b} \quad (4)$$

$$\frac{P_{rel}(i,j)}{P_i P_j} = e^{\lambda S(i,j)} \quad (5)$$

In BLOSUM62 matrix, the value of each cell was determined by equation (4), where  $P_{ab}$  is the related probability as described previously.  $P_a$  and  $P_b$  stands for background probability of the occurrence of different amino acids. It is mentioned in nature biotechnology website that  $\lambda$  is set as 0.347 for creating BLOSUM62 matrix. We adopted the value and because now we have the exact value of the constant  $\lambda$ , we can immediately modify equation (4) and transform it into equation (5).  $P_{rel}(i,j)/P_i P_j$  is then substituted into equation (3) to obtain the final score.

(3) The positions of the specific binding residues in the 5 templates that give the highest similarity scores are then superimposed to predict the position of the specific binding residues in the query TF.

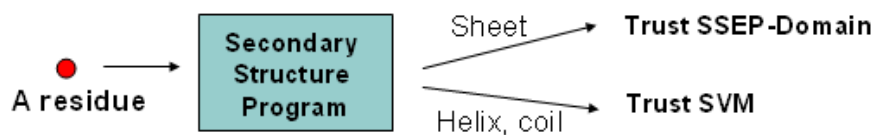


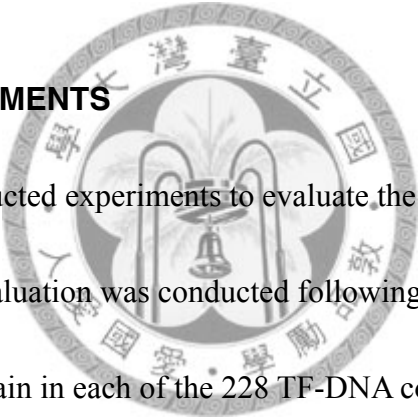
Fig. 4 Ideas of the merging process. In the merging process, the type of secondary structure of each residue in the protein chain was first predicted by a secondary structure program where each residue could come out as helix, sheet, or coil. As mentioned previously, SSEP-Domain had better accuracy in predicting  $\beta$ -sheet residue, SVM rendered good performance in predicting  $\alpha$ -helix residues. In this respect we take union of the outputs from these two predictors according to which type of secondary structure this residue belongs to.

In the phase of “merging the output”, the process could be better depicted as in Fig. 4. For a residue which is of helix or coil type in the query protein chain, we refer to SVM for prediction. On the other hand, for a residue which is of sheet type in the query protein chain, we use the other one, SSEP predictor to obtain the result.

## Chapter 4 EXPERIMENTAL RESULTS

In this chapter, we first depict how the performance of prediction is evaluated. Then we arrange residues into different groups according to their secondary structure type:  $\alpha$ -helix, coil, and  $\beta$ -sheet. We also have breakdowns of residues in respect of their different TF-DNA interaction. Then we show the results of each individual classifier and the hybrid predictor. In the end of this chapter, we exhibit a real scenario of TF-DNA interaction and discuss the corresponding predicted result.

### 4.1 DESIGN OF EXPERIMENTS



In our study, we have conducted experiments to evaluate the performance of the proposed approach. The evaluation was conducted following the leave-one-out practice. Accordingly, the protein chain in each of the 228 TF-DNA complexes was used as the testing case once. In order to avoid bias caused by homologous protein chains, the training data set for the SVM and the template library for the SSEP algorithm were re-generated for each testing protein chain with the protein chains in the remaining 227 TF-DNA complexes that has a sequence identity higher than 20% when aligned with the testing protein chain removed. In our experiment, the bl2seq package was invoked to obtain a score of sequence identity between two protein chains.



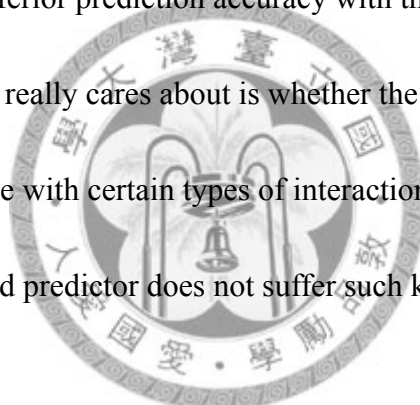
**Table 1. Prediction results with the SVM based primary predictor.**

Type of the secondary structure element	# in residues tested	Prediction results						
		TP	TN	FP	FN	Prec.	Sens.	Spec.
Helix	12781	573	11670	156	382	0.786	0.6	0.987
Sheet	1465	0	1358	3	104	0.0	0.0	0.998
Coil	7921	186	7506	58	171	0.762	0.521	0.992

## 4.2 RESULTS AND DISCUSSIONS

Table 1 shows how the SVM based predictor in Fig. 3 performed in the leave-one-out process. As mentioned earlier, the parameters of the SVM based predictor has been tuned to deliver high precision. As a result, sensitivity was traded. The results in Table 1 reveal that the SVM based predictor, to a certain extent, is capable of identifying the specific DNA-binding residues in  $\alpha$ -helix and coil elements. On the other hand, the SVM based predictor can hardly identify the specific DNA-binding residues in  $\beta$ -sheet elements. Therefore, in order to raise sensitivity of prediction, we have resorted to the SSEA based mechanism to complement the prediction power of the SVM. Table 2 shows how the SSEA based predictor performed in identifying the specific DNA-binding residues in  $\beta$ -sheet elements. Combining the results in Tables 1 and 2, one

can easily conclude that the prediction power of the SSEA based mechanism complements that of the SVM. With the SVM based predictor and the SSEA based predictor integrated as shown in Fig. 3, the hybrid predictor has been able to deliver the performance shown in Table 3. Table 4 shows a breakdown of the experimental results with the hybrid predictor based on the classification of TF-DNA interactions proposed by J.M. Thornton *et al.* [32]. It should not be a surprise to observe that the hybrid predictor can deliver superior prediction accuracy when dealing with certain types of interactions and delivers inferior prediction accuracy with the other types. In this respect, what a biologist or chemist really cares about is whether the predictor could deliver extremely poor performance with certain types of interactions. The results reported in Table 4 show that the hybrid predictor does not suffer such kind of deficiency.



**Table 2. Prediction results with the SSEA based auxiliary predictor.**

Type of the secondary structure element	# in residues tested	Prediction results						
		TP	TN	FP	FN	Prec.	Sens.	Spec.
Sheet	1465	83	1329	32	21	0.722	0.798	0.984

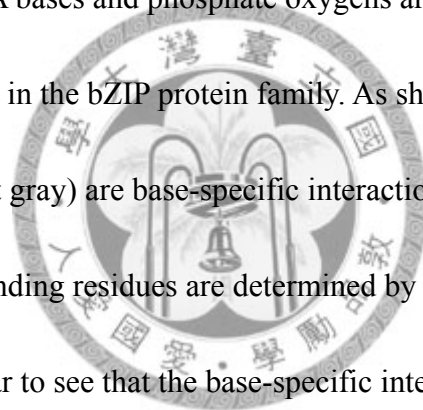
**Table 3. Prediction results with the hybrid predictor.**

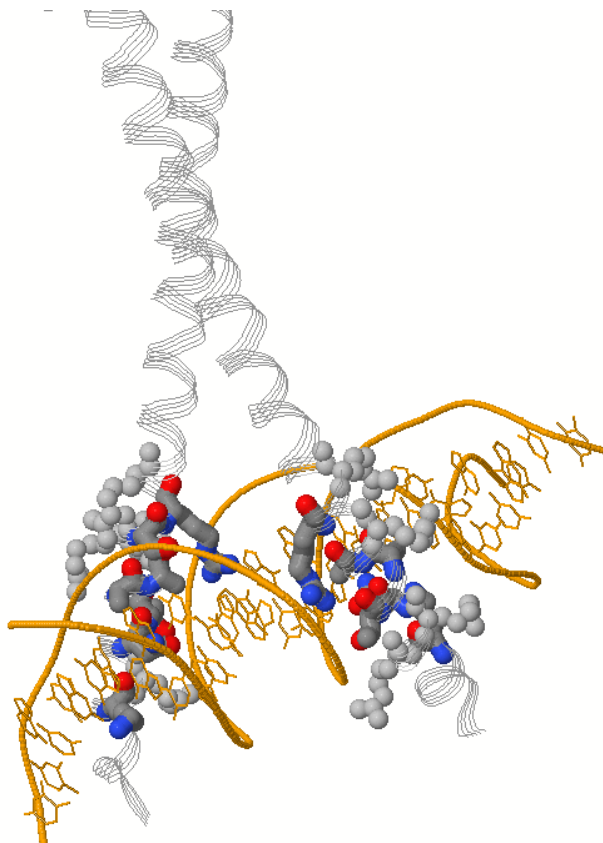
Type of the secondary structure element	# in residues tested	Prediction results						
		TP	TN	FP	FN	Prec.	Sens.	Spec.
Helix	12781	573	11670	156	382	0.786	0.6	0.987
Sheet	1465	83	1329	32	21	0.722	0.798	0.976
Coil	7921	186	7506	58	171	0.762	0.521	0.992
Overall	22167	842	20505	246	574	0.773	0.594	0.988

**Table 4. Breakdown of the prediction results with the hybrid predictor in respect of different types of TF-DNA interactions**

Type of the DNA-binding group	# of chains involved	# in residues tested	Prediction results						
			TP	TN	FP	FN	Precision	Sensitivity	Specificity
Zipper-type	44	3109	213	2821	30	45	0.876	0.826	0.989
Helix-turn-helix	97	12480	316	11712	123	329	0.72	0.49	0.99
Zinc-coordinating	57	4792	230	4332	74	156	0.757	0.596	0.983
$\beta$ -hairpin/ribbon	30	1786	83	1640	19	44	0.814	0.654	0.989
Overall	228	22167	842	20505	246	574	0.774	0.595	0.988

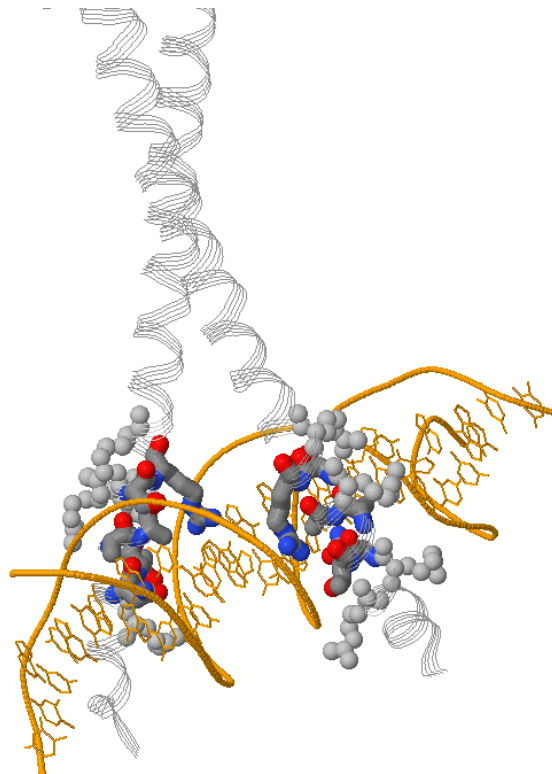
The yeast transcriptional activator GCN4 is 1 of over 30 identified eukaryotic proteins containing the basic region leucine zipper (bZIP) DNA-binding motif. The bZIP dimer is a pair of continuous alpha helices that form a parallel coiled coil over their carboxy-terminal 30 residues and gradually diverge toward their amino termini to pass through the major groove of the DNA-binding site. The coiled-coil dimerization interface is oriented almost perpendicular to the DNA axis, giving the complex the appearance of the letter T. There are no kinks or sharp bends in either bZIP monomer. Numerous contacts to DNA bases and phosphate oxygens are made by basic region residues that are conserved in the bZIP protein family. As shown in Fig.5, the atoms with colors but white (light gray) are base-specific interaction residues according to our definition. Non-specific binding residues are determined by literature and then colored white (light gray). It is clear to see that the base-specific interaction residues are much closer to DNA bases than others. The same TF-DNA complex was tested by using our approach, and the prediction result is presented in Fig.6. As in this case of Zipper-type domain, our predictor gives 85% (6/7) sensitivity and 100% precision.





Jmol

Fig. 5 An example of TF-DNA interaction with PDB ID 1YSA. The atoms with colored but white (light gray) are the heavy atoms in the sidechains which are within 4.5 Å from the bases of the DNA. The atoms colored by white (light gray) are the heavy atoms in the sidechains of the non-specific DNA-binding residues.



Jmol

Fig .6 A prediction result of query protein chain with PDB ID 1YSA. The atoms with colored but white (light gray) are the heavy atoms in the sidechains which are “predicted” as the specific DNA-binding residues. The atoms colored by white (light gray) are the heavy atoms “predicted” in the sidechains of the non-specific DNA-binding residues.

Arc repressor [33] is one member of  $\beta$ -hairpin/ribbon family, and it acts by the cooperative binding of two Arc repressor dimers to a 21-base-pair operator site. Each Arc dimer uses an antiparallel beta-sheet to recognize bases in the major groove. As depicted in Fig. 7, two antiparallel beta-sheet are the binding interface stretched to the major groove [33]. In this case, our predictor gives 50% (3/6) sensitivity and 100% precision.

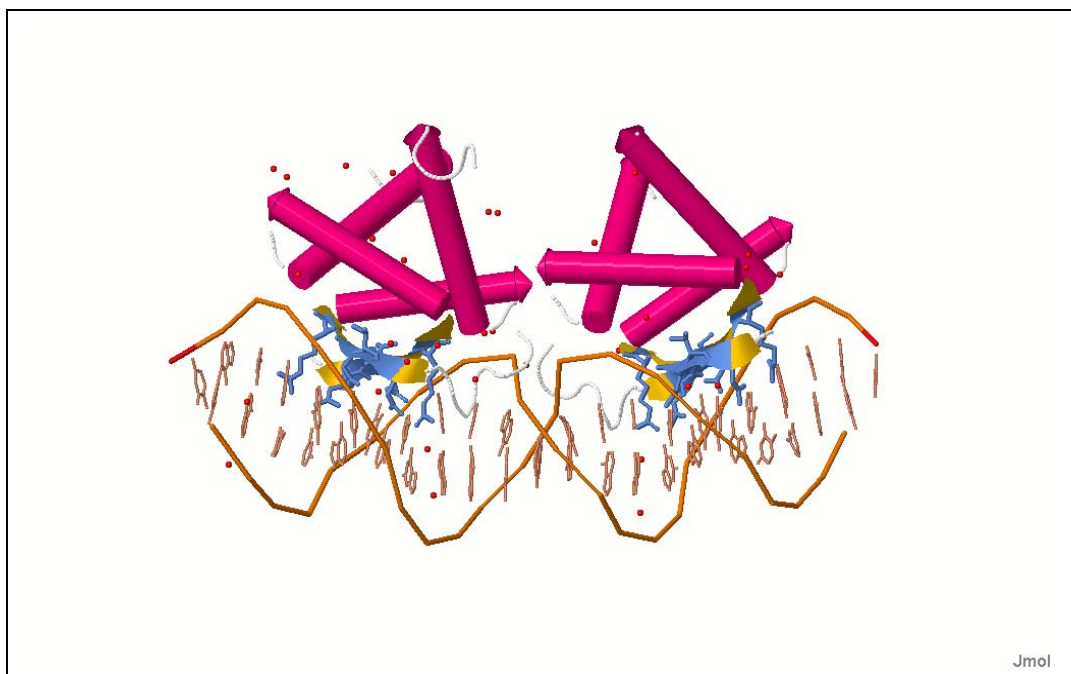


Fig. 7 An example of TF-DNA interaction with PDB ID 1BDV. There is an Arc domain, containing beta-sheet to recognize bases in the major groove, which is a DNA-binding domain in each protein chain. The atoms colored in blue are heavy atoms which are within 4.5 Å from the bases of the DNA.

The helix-turn-helix clan contains many members; HTH\_3 is one of these members, and it is a large family of DNA binding helix-turn helix proteins that include a bacterial plasmid copy control protein, bacterial methylases, various bacteriophage transcription control proteins and a vegetative specific protein from Dictyostelium discoideum (Slime mould). Fig. 8 depicted one example of protein with HTH\_3 domain interacts with DNA. In this case, our predictor gives 50% (4/8) sensitivity and 100% precision.

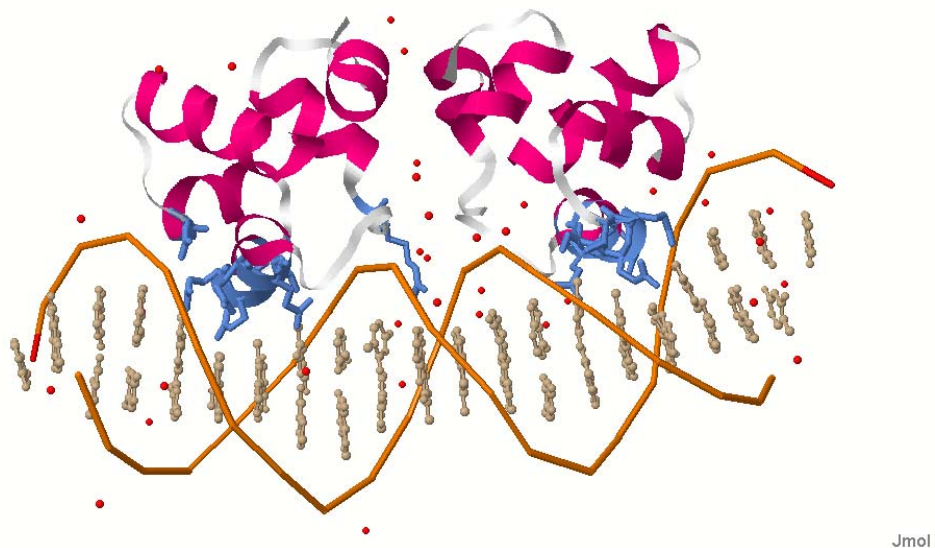


Fig.8 An example of TF-DNA interaction with PDB ID 1RPE. There is a HTH\_3 which is a DNA-binding domain in the protein chain. The atoms colored in blue are heavy atoms which are within 4.5 Å from the bases of the DNA. There are many helix-turn-helix motifs in the protein chain, while only some of these can be the DNA-binding domain.



Zf-C4 is a member of the Zinc-coordinating family, also a zinc finger. In nearly all cases, this is the DNA binding domain of a nuclear hormone receptor. The alignment contains two Zinc finger domains that are too dissimilar to be aligned with each other.

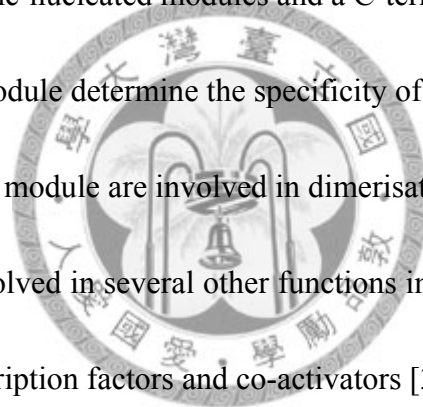
The DNA-binding domain can elicit either an activating or repressing effect by binding to specific regions of the DNA known as hormone-response elements [34, 35]. These response elements position the receptors, and the complexes recruited by them, close to the genes of which transcription is affected. The DNA-binding domains of nuclear

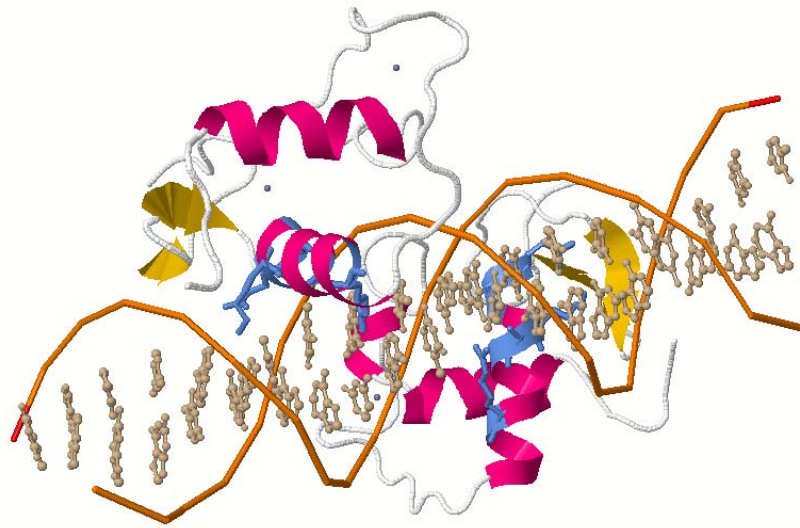
receptors consist of two zinc-nucleated modules and a C-terminal extension, where residues in the first zinc module determine the specificity of the DNA recognition and residues in the second zinc module are involved in dimerisation. The DNA-binding domain is furthermore involved in several other functions including nuclear localization, and interaction with transcription factors and co-activators [34]. This is a rather

sophisticated DNA-binding domain which involved in many functions, as depicted in

Fig. 9. Our predictor failed in locating the correct specific-binding residues, rendering

0% sensitivity (0/5) and 0% precision (0/2).





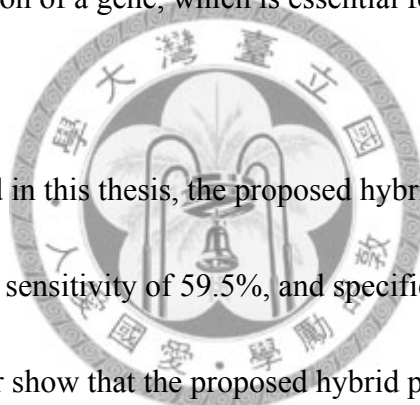
Jmol

Fig.9 An example of TF-DNA interaction with PDB ID 1LAT.



## Chapter 5 CONCLUSIONS AND FUTURE WORKS

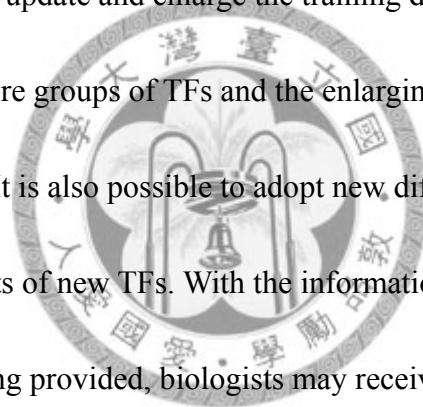
This thesis presents the design of a sequence-based predictor aiming to identify the specific DNA-binding residues in a TF. As a recent study has revealed that the tertiary structures of a large number of transcription factors are mostly disordered, a sequence based predictor is essential for analyzing how a TF interacts with the DNA. Furthermore, it is highly desirable to have a predictor capable of identifying those residues involved in specific binding with the DNA, since specific binding corresponds to sequence-specific recognition of a gene, which is essential for correct gene regulation. .



In the experiments reported in this thesis, the proposed hybrid predictor delivered overall precision of 77.4%, sensitivity of 59.5%, and specificity of 98.8%. The experimental results further show that the proposed hybrid predictor is capable of delivering the same level of prediction accuracy when dealing with different types of TF-DNA interactions. It is anticipated the prediction accuracy delivered by the hybrid predictor will continue to improve as the number of TF-DNA complexes deposited in the PDB continues to grow and therefore the number of training samples that can be exploited continues to increase. Nevertheless, it is our primary objective to continue to develop more advanced prediction mechanisms. In this respect, we believe that, as the number of TF-DNA complexes deposited in the PDB increases, we can obtain more

insights about the key physiochemical properties that play essential roles in TF-DNA interactions and then we will be able to develop more advanced prediction mechanisms accordingly.

Besides those four types of DNA-binding domains in TFs mentioned in the study, there are other DNA-binding domains such as P53, GATA..etc, which also play important role in regulatory network. An obvious way to support new forthcoming DNA-binding domains is to continuously update and enlarge the training data set, therefore the hybrid predictor could support more groups of TFs and the enlarging would also possibly enhance the performance. It is also possible to adopt new different learning methods or features for the unique traits of new TFs. With the information of binding sites and candidate domain type being provided, biologists may receive more information for conjecturing and understating the functionality of given protein chain. Because of the high precision and reliability, the proposed method in the thesis would be deserving of extension or application in the future, making contribution to connecting some DNA-binding domains with specific functions.

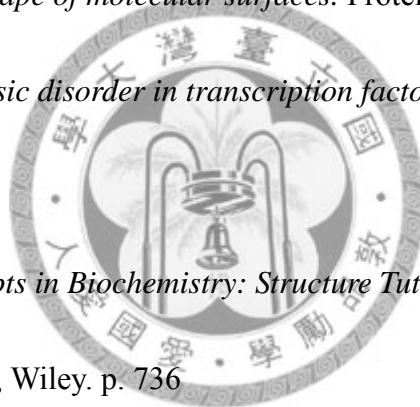


## REFERENCES

1. Latchman, D.S., *Transcription factors: an overview*. Int J Biochem Cell Biol, 1997. **29**(12): p. 1305-12.
2. Karin, M., *Too many transcription factors: positive and negative interactions*. New Biol, 1990. **2**(2): p. 126-31.
3. Roeder, R.G., *The role of general initiation factors in transcription by RNA polymerase II*. Trends Biochem Sci, 1996. **21**(9): p. 327-35.
4. Nikolov, D.B. and S.K. Burley, *RNA polymerase II transcription initiation: a structural view*. Proc Natl Acad Sci U S A, 1997. **94**(1): p. 15-22.
5. Lee, T.I. and R.A. Young, *Transcription of eukaryotic protein-coding genes*. Annu Rev Genet, 2000. **34**: p. 77-137.
6. Goodrich, J.A., et al., *Drosophila TAFII40 interacts with both a VP16 activation domain and the basal transcription factor TFIIB*. Cell, 1993. **75**(3): p. 519-30.
7. Xiao, H., et al., *Binding of basal transcription factor TFIIF to the acidic activation domains of VP16 and p53*. Mol Cell Biol, 1994. **14**(10): p. 7013-24.
8. Poulat, F., et al., *The human testis determining factor SRY binds a nuclear factor containing PDZ protein interaction domains*. J Biol Chem, 1997. **272**(11): p. 7167-72.
9. Sorger, P.K. and H.R. Pelham, *Yeast heat shock factor is an essential*

- DNA-binding protein that exhibits temperature-dependent phosphorylation. Cell*, 1988. **54**(6): p. 855-64.
10. Wang, G.L., et al., *Hypoxia-inducible factor 1 is a basic-helix-loop-helix-PAS heterodimer regulated by cellular O<sub>2</sub> tension. Proc Natl Acad Sci U S A*, 1995. **92**(12): p. 5510-4.
11. Maxwell, P.H., et al., *The tumour suppressor protein VHL targets hypoxia-inducible factors for oxygen-dependent proteolysis. Nature*, 1999. **399**(6733): p. 271-5.
12. Cole, M.D., *The myc oncogene: its role in transformation and differentiation. Annu Rev Genet*, 1986. **20**: p. 361-84.
13. Yan, C., et al., *Predicting DNA-binding sites of proteins from amino acid sequence. BMC Bioinformatics*, 2006. **7**: p. 262.
14. Ofran, Y., V. Mysore, and B. Rost, *Prediction of DNA-binding residues from sequence. Bioinformatics*, 2007. **23**(13): p. i347-53.
15. Tjong, H. and H.X. Zhou, *DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. Nucleic Acids Res*, 2007. **35**(5): p. 1465-77.
16. Ahmad, S. and A. Sarai, *PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatics*, 2005. **6**: p. 33.

17. Jones, S., et al., *Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins*. *Nucleic Acids Res*, 2003. **31**(24): p. 7189-98.
18. Ferrer-Costa, C., et al., *HTHquery: a method for detecting DNA-binding proteins with a helix-turn-helix structural motif*. *Bioinformatics*, 2005. **21**(18): p. 3679-80.
19. Tsuchiya, Y., K. Kinoshita, and H. Nakamura, *Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces*. *Proteins*, 2004. **55**(4): p. 885-94.
20. Liu, J., et al., *Intrinsic disorder in transcription factors*. *Biochemistry*, 2006. **45**(22): p. 6873-88.
21. Boyer, R.F., *Concepts in Biochemistry: Structure Tutorials*, in *Concepts in Biochemistry*. 2005, Wiley. p. 736
22. Tsuchiya, Y., K. Kinoshita, and H. Nakamura, *PreDs: a server for predicting dsDNA-binding site on protein molecular surfaces*. *Bioinformatics*, 2005. **21**(8): p. 1721-3.
23. Wang, L. and S.J. Brown, *BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences*. *Nucleic Acids Res*, 2006. **34**(Web Server issue): p. W243-8.
24. Hwang, S., Z. Gou, and I.B. Kuznetsov, *DP-Bind: a web server for*



- sequence-based prediction of DNA-binding residues in DNA-binding proteins.*
- Bioinformatics, 2007. **23**(5): p. 634-6.
25. Chang, C. and C. Lin, *{LIBSVM}: a library for support vector machines.* 2001.
26. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
27. Przybylski, D. and B. Rost, *Alignments grow, secondary structure prediction improves.* Proteins, 2002. **46**(2): p. 197-205.
28. Lin, H.N., et al., *HYPROSP II--a knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence.* Bioinformatics, 2005. **21**(15): p. 3227-33.
29. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
30. Finn, R.D., et al., *Pfam: clans, web tools and services.* Nucleic Acids Res, 2006. **34**(Database issue): p. D247-51.
31. Gewehr, J.E. and R. Zimmer, *SSEP-Domain: protein domain prediction by alignment of secondary structure elements and profiles.* Bioinformatics, 2006. **22**(2): p. 181-7.
32. Luscombe, N.M., et al., *An overview of the structures of protein-DNA complexes.*



Genome Biol, 2000. **1**(1): p. REVIEWS001.

33. Raumann, B.E., et al., *DNA recognition by beta-sheets in the Arc repressor-operator crystal structure*. Nature, 1994. **367**(6465): p. 754-7.
34. Claessens, F. and D.T. Gewirth, *DNA recognition by nuclear receptors*. Essays Biochem, 2004. **40**: p. 59-72.
35. Moehren, U., M. Eckey, and A. Baniahmad, *Gene repression by nuclear hormone receptors*. Essays Biochem, 2004. **40**: p. 89-104.



## APPENDIX

Chain ID	start	End	Bits score	E-value	Pfam-A	Chain Length	Number of specific-binding residues
1A0A_A	2	58	55.6	1.70E-13	HLH	63	7
1A0A_B	2	58	55.6	1.70E-13	HLH	63	6
1AM9_A	6	56	71.4	3.10E-18	HLH	80	6
1AM9_B	6	56	71.4	3.10E-18	HLH	76	7
1AM9_C	6	56	71.4	3.10E-18	HLH	82	6
1AM9_D	6	56	71.4	3.10E-18	HLH	76	7
1AN2_A	3	54	81.9	2.00E-21	HLH	86	4
1NKP_A	6	58	73.5	7.20E-19	HLH	88	6
1NKP_B	2	53	81.9	2.00E-21	HLH	83	5
1NKP_D	6	58	73.5	7.20E-19	HLH	85	5
1NKP_E	2	53	81.9	2.00E-21	HLH	83	5
1NLW_B	1	52	81.9	2.00E-21	HLH	76	6
1NLW_E	1	52	81.9	2.00E-21	HLH	76	5
1HLO_A	12	63	81.9	2.00E-21	HLH	80	5

1HLO_B	12	63	81.9	2.00E-21	HLH	80	6
1AN4_A	5	60	66.7	7.60E-17	HLH	65	6
1AN4_B	5	60	66.7	7.60E-17	HLH	65	5
1MDY_A	12	63	69.6	1.00E-17	HLH	68	5
1MDY_C	6	57	69.6	1.00E-17	HLH	62	5
1MDY_D	6	57	69.6	1.00E-17	HLH	62	5
1NLW_A	1	53	56.4	9.80E-14	HLH	79	5
1NLW_D	1	53	56.4	9.80E-14	HLH	77	6
1A02_F	2	52	28.9	1.80E-05	bZIP_2	55	5
1FOS_E	1	51	28.9	1.80E-05	bZIP_2	60	4
1FOS_G	1	51	28.9	1.80E-05	bZIP_2	60	6
1A02_J	1	55	53.1	9.90E-13	bZIP_1	56	5
1FOS_F	1	60	72.4	1.50E-18	bZIP_1	60	5
1FOS_H	1	60	72.4	1.50E-18	bZIP_1	60	6
1JNM_A	1	60	72.4	1.50E-18	bZIP_1	56	6
1JNM_B	1	60	72.4	1.50E-18	bZIP_1	57	6
1IO4_A	12	65	82	2.00E-21	bZIP_2	73	7
1IO4_B	12	65	82	2.00E-21	bZIP_2	76	7
1H88_A	12	65	82	2.00E-21	bZIP_2	78	6

1H88_B	12	65	82	2.00E-21	bZIP_2	78	7
1H8A_A	12	65	82	2.00E-21	bZIP_2	76	6
1H8A_B	12	65	82	2.00E-21	bZIP_2	76	7
1H89_A	1	51	67	6.40E-17	bZIP_2	64	6
1H89_B	1	51	67	6.40E-17	bZIP_2	64	7
1HJB_A	12	65	82	2.00E-21	bZIP_2	75	6
1HJB_B	12	65	82	2.00E-21	bZIP_2	76	7
1HJB_D	12	65	82	2.00E-21	bZIP_2	77	6
1HJB_E	12	65	82	2.00E-21	bZIP_2	77	7
1YSA_C	2	56	66.2	1.10E-16	bZIP_1	58	6
1YSA_D	2	56	66.2	1.10E-16	bZIP_1	57	7
1CGP_A	20	111	104.7	2.90E-28	cNMP_binding	205	5
1CGP_B	20	111	104.7	2.90E-28	cNMP_binding	205	5
1LB2_A	20	111	104.7	2.90E-28	cNMP_binding	209	6
1CF7_A	7	73	137.1	5.00E-38	E2F_TDP	72	4
1CF7_B	7	88	155.4	1.50E-43	E2F_TDP	90	4
1C0W_A	2	62	106.4	8.60E-29	Fe_dep_repress	222	7
1C0W_B	2	62	106.4	8.60E-29	Fe_dep_repress	225	7
1C0W_C	2	62	106.4	8.60E-29	Fe_dep_repress	221	7

1C0W_D	2	62	106.4	8.60E-29	Fe_dep_repress	221	7
1DDN_A	3	63	106.4	8.60E-29	Fe_dep_repress	120	6
1DDN_B	3	63	106.4	8.60E-29	Fe_dep_repress	120	7
1DDN_C	3	63	106.4	8.60E-29	Fe_dep_repress	120	5
1DDN_D	3	63	106.4	8.60E-29	Fe_dep_repress	120	7
1F5T_A	3	63	106.4	8.60E-29	Fe_dep_repress	121	6
1F5T_B	3	63	106.4	8.60E-29	Fe_dep_repress	121	7
1F5T_C	3	63	106.4	8.60E-29	Fe_dep_repress	121	7
1F5T_D	3	63	106.4	8.60E-29	Fe_dep_repress	121	8
1L3L_A	17	161	164.5	2.80E-46	Autoind_bind	234	6
1L3L_B	17	161	164.5	2.80E-46	Autoind_bind	234	6
1L3L_C	17	161	164.5	2.80E-46	Autoind_bind	234	5
1L3L_D	17	161	164.5	2.80E-46	Autoind_bind	234	5
1H9T_A	12	75	90.2	6.80E-24	GntR	243	8
1H9T_B	12	75	90.2	6.80E-24	GntR	234	8
1HW2_A	8	71	90.2	6.80E-24	GntR	228	8
1HW2_B	8	71	90.2	6.80E-24	GntR	228	8
1AKH_A	6	61	41.5	3.00E-09	Homeobox	60	6
1YRN_A	6	61	41	4.10E-09	Homeobox	60	6

1LE8_A	2	53	30.7	5.40E-06	Homeobox	53	6
1MNM_C	30	87	29.9	9.50E-06	Homeobox	87	7
1MNM_D	30	87	29.9	9.50E-06	Homeobox	87	5
1AKH_B	5	62	29.9	9.50E-06	Homeobox	78	7
1APL_C	5	62	29.9	9.50E-06	Homeobox	62	8
1APL_D	5	62	29.9	9.50E-06	Homeobox	62	7
1LE8_B	5	62	19.8	0.00013	Homeobox	78	6
1YRN_B	5	62	29.9	9.50E-06	Homeobox	78	7
1K61_A	1	58	29.9	9.50E-06	Homeobox	60	6
1K61_B	1	58	29.9	9.50E-06	Homeobox	59	6
1K61_C	1	58	29.9	9.50E-06	Homeobox	58	4
1K61_D	1	58	29.9	9.50E-06	Homeobox	58	6
1CQT_A	4	78	190.5	4.20E-54	Pou	163	13
1AU7_A	1	71	167.6	3.40E-47	Pou	146	14
1AU7_B	1	71	167.6	3.40E-47	Pou	146	15
1HF0_A	1	75	185.4	1.40E-52	Pou	158	13
1HF0_B	1	75	185.4	1.40E-52	Pou	158	12
1B72_B	2	61	71.4	2.90E-18	Homeobox	75	5
1PUF_B	2	61	71.4	2.90E-18	Homeobox	73	5

1B8I_B	2	61	70	7.80E-18	Homeobox	62	4
1FJL_A	19	75	115.3	1.80E-31	Homeobox	81	6
1FJL_B	19	75	115.3	1.80E-31	Homeobox	75	5
1FJL_C	19	75	115.3	1.80E-31	Homeobox	75	5
1PUF_A	14	70	110.6	4.80E-30	Homeobox	77	6
9ANT_A	4	60	113.7	5.40E-31	Homeobox	62	6
1HDD_C	4	60	105.5	1.60E-28	Homeobox	61	6
1HDD_D	4	60	105.5	1.60E-28	Homeobox	61	3
2HDD_A	4	60	101.5	2.70E-27	Homeobox	61	4
2HDD_B	4	60	101.5	2.70E-27	Homeobox	59	3
1JGG_A	2	58	109.2	1.30E-29	Homeobox	59	7
1JGG_B	2	58	109.2	1.30E-29	Homeobox	59	9
3HDD_A	2	58	105.5	1.60E-28	Homeobox	59	4
3HDD_B	2	58	105.5	1.60E-28	Homeobox	58	4
1DU0_A	1	56	89.6	1.00E-23	Homeobox	57	4
1DU0_B	1	56	89.6	1.00E-23	Homeobox	56	4
1LLI_A	21	76	53.5	7.10E-13	HTH_3	92	5
1LLI_B	21	76	53.5	7.10E-13	HTH_3	92	11
1LMB_3	21	76	50.4	6.10E-12	HTH_3	92	5

1LMB_4	21	76	50.4	6.10E-12	HTH_3	92	10
1PER_L	6	59	56.1	1.20E-13	HTH_3	63	9
1PER_R	6	59	56.1	1.20E-13	HTH_3	63	6
1RPE_L	6	59	56.1	1.20E-13	HTH_3	63	8
1RPE_R	6	59	56.1	1.20E-13	HTH_3	63	6
2OR1_L	6	59	56.1	1.20E-13	HTH_3	63	9
2OR1_R	6	59	56.1	1.20E-13	HTH_3	63	8
1GDT_A	3	139	199.4	8.80E-57	Resolvase	183	13
1GDT_B	3	139	199.4	8.80E-57	Resolvase	183	10
1D5Y_A	7	53	48.3	2.80E-11	HTH_AraC	292	5
1D5Y_B	7	53	48.3	2.80E-11	HTH_AraC	292	2
1D5Y_C	7	53	48.3	2.80E-11	HTH_AraC	292	5
1D5Y_D	7	53	48.3	2.80E-11	HTH_AraC	292	0
1JWL_A	4	29	48.2	2.80E-11	LacI	330	6
1JWL_B	4	29	48.2	2.80E-11	LacI	330	6
1KU7_A	7	60	90.9	3.90E-24	Sigma70_r4	73	7
1JT0_A	7	53	69.6	1.10E-17	TetR_N	189	7
1JT0_B	7	53	69.6	1.10E-17	TetR_N	189	7
1JT0_C	7	53	69.6	1.10E-17	TetR_N	189	7



1JT0_D	7	53	69.6	1.10E-17	TetR_N	186	7
1TRO_A	17	104	146.5	7.60E-41	Trp_repressor	108	7
1TRO_C	17	104	146.5	7.60E-41	Trp_repressor	108	7
1TRO_E	17	104	146.5	7.60E-41	Trp_repressor	108	7
1TRO_G	17	104	146.5	7.60E-41	Trp_repressor	105	6
1TRR_A	16	103	145.6	1.30E-40	Trp_repressor	105	9
1TRR_B	16	103	145.6	1.30E-40	Trp_repressor	105	3
1TRR_D	16	103	145.6	1.30E-40	Trp_repressor	105	9
1TRR_E	16	103	145.6	1.30E-40	Trp_repressor	105	8
1TRR_G	16	103	145.6	1.30E-40	Trp_repressor	105	9
1TRR_H	16	103	145.6	1.30E-40	Trp_repressor	105	6
1TRR_J	16	103	145.6	1.30E-40	Trp_repressor	105	9
1TRR_K	16	103	145.6	1.30E-40	Trp_repressor	105	3
1BDT_A	4	53	123.2	7.40E-34	Arc	52	5
1BDT_B	4	53	123.2	7.40E-34	Arc	53	5
1BDT_C	4	53	123.2	7.40E-34	Arc	50	6
1BDT_D	4	53	123.2	7.40E-34	Arc	50	6
1BDV_A	4	53	116.2	9.80E-32	Arc	52	4
1BDV_B	4	53	116.2	9.80E-32	Arc	53	6

1BDV_C	4	53	116.2	9.80E-32	Arc	49	4
1BDV_D	4	53	116.2	9.80E-32	Arc	50	6
1PAR_A	4	53	123.2	7.40E-34	Arc	52	5
1PAR_B	4	53	123.2	7.40E-34	Arc	53	5
1PAR_C	4	53	123.2	7.40E-34	Arc	50	6
1PAR_D	4	53	123.2	7.40E-34	Arc	53	6
1B01_A	4	42	53.5	7.20E-13	RHH_1	43	3
1B01_B	4	42	53.5	7.20E-13	RHH_1	43	4
1EA4_A	4	42	53.5	7.20E-13	RHH_1	43	3
1EA4_B	4	42	53.5	7.20E-13	RHH_1	41	6
1EA4_D	4	42	53.5	7.20E-13	RHH_1	43	3
1EA4_E	4	42	53.5	7.20E-13	RHH_1	44	5
1EA4_F	4	42	53.5	7.20E-13	RHH_1	45	3
1EA4_G	4	42	53.5	7.20E-13	RHH_1	42	5
1EA4_H	4	42	53.5	7.20E-13	RHH_1	45	3
1EA4_J	4	42	53.5	7.20E-13	RHH_1	44	7
1EA4_K	4	42	53.5	7.20E-13	RHH_1	43	3
1EA4_L	4	42	53.5	7.20E-13	RHH_1	44	4
1CMA_A	1	104	305.2	1.20E-88	MetJ	104	2

1CMA_B	1	104	305.2	1.20E-88	MetJ	104	2
1MJM_A	1	104	301.7	1.40E-87	MetJ	104	2
1MJM_B	1	104	301.7	1.40E-87	MetJ	104	2
1MJP_A	1	104	301.7	1.40E-87	MetJ	104	3
1MJP_B	1	104	301.7	1.40E-87	MetJ	104	3
1KB2_A	7	82	166.6	6.50E-47	zf-C4	95	5
1KB2_B	7	82	166.6	6.50E-47	zf-C4	91	4
1KB4_A	7	82	166.6	6.50E-47	zf-C4	99	5
1KB4_B	7	82	166.6	6.50E-47	zf-C4	105	4
1KB6_A	7	82	166.6	6.50E-47	zf-C4	99	5
1KB6_B	7	82	166.6	6.50E-47	zf-C4	106	4
2NLL_B	2	79	176.9	5.00E-50	zf-C4	103	5
1A6Y_A	8	84	184.2	3.20E-52	zf-C4	85	8
1A6Y_B	8	84	184.2	3.20E-52	zf-C4	88	7
1DSZ_A	5	80	195	1.90E-55	zf-C4	80	4
1DSZ_B	6	81	190.2	5.20E-54	zf-C4	84	5
1HCQ_A	5	80	180.2	5.20E-51	zf-C4	74	5
1HCQ_B	5	80	180.2	5.20E-51	zf-C4	74	5
1BY4_A	6	81	190.2	5.20E-54	zf-C4	82	4

1BY4_B	6	81	190.2	5.20E-54	zf-C4	82	5
1BY4_C	6	81	190.2	5.20E-54	zf-C4	82	4
1BY4_D	6	81	190.2	5.20E-54	zf-C4	81	4
1LAT_A	5	80	168.6	1.60E-47	zf-C4	75	5
1LAT_B	5	80	168.6	1.60E-47	zf-C4	77	6
1R0N_A	4	79	185.8	1.10E-52	zf-C4	81	4
2NLL_A	1	66	145.9	1.20E-40	zf-C4	66	4
1F2I_G	20	44	35.9	1.40E-07	zf-C2H2	73	7
1F2I_H	20	44	35.9	1.40E-07	zf-C2H2	73	7
1F2I_I	20	44	35.9	1.40E-07	zf-C2H2	73	7
1F2I_J	20	44	35.9	1.40E-07	zf-C2H2	73	7
1F2I_K	20	44	35.9	1.40E-07	zf-C2H2	73	7
1F2I_L	20	44	35.9	1.40E-07	zf-C2H2	73	7
1G2D_C	5	29	32.5	1.50E-06	zf-C2H2	89	14
1G2D_F	5	29	32.5	1.50E-06	zf-C2H2	88	14
1G2F_C	5	29	32.5	1.50E-06	zf-C2H2	89	14
1G2F_F	5	29	32.5	1.50E-06	zf-C2H2	88	15
1MEY_C	5	27	45.4	2.00E-10	zf-C2H2	84	13
1MEY_F	5	27	45.4	2.00E-10	zf-C2H2	84	13

1MEY_G	5	27	45.4	2.00E-10	zf-C2H2	83	0
1P47_A	4	28	35.9	1.40E-07	zf-C2H2	87	12
1P47_B	4	28	35.9	1.40E-07	zf-C2H2	85	13
1LLM_C	4	26	27.8	4.00E-05	zf-C2H2	87	8
1TF6_A	13	37	23.7	0.0007	zf-C2H2	188	15
1TF6_D	13	37	23.7	0.0007	zf-C2H2	188	15
2DRP_A	11	34	21.3	0.0037	zf-C2H2	65	11
2DRP_D	11	34	21.3	0.0037	zf-C2H2	66	9
1D66_A	9	47	59.1	1.50E-14	Zn_clus	64	4
1D66_B	9	47	59.1	1.50E-14	Zn_clus	64	3
1HWT_C	8	48	37.1	6.30E-08	Zn_clus	74	8
1HWT_D	8	48	37.1	6.30E-08	Zn_clus	74	4
1HWT_G	8	48	37.1	6.30E-08	Zn_clus	74	7
1HWT_H	8	48	37.1	6.30E-08	Zn_clus	74	4
2HAP_C	8	48	28.7	2.20E-05	Zn_clus	76	4
2HAP_D	8	48	28.7	2.20E-05	Zn_clus	76	6
1QP9_A	8	48	37	6.70E-08	Zn_clus	76	6
1QP9_B	8	48	37	6.70E-08	Zn_clus	75	5
1QP9_C	8	48	37	6.70E-08	Zn_clus	74	4

1QP9_D	8	48	37	6.70E-08	Zn_clus	75	5
1PYI_A	5	44	56.6	8.80E-14	Zn_clus	90	3
1PYI_B	5	44	56.6	8.80E-14	Zn_clus	72	3
1ZME_C	2	39	59	1.60E-14	Zn_clus	70	3
1ZME_D	2	39	59	1.60E-14	Zn_clus	70	7

