Computational Methods for Predicting Transcription Factor Binding Sites

Robert Radoslaw Zygmunt Osada

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
By the Department of
Computer Science

November, 2006

Abstract

A major challenge in computational biology is to understand the mechanisms that control gene expression. Transcription factor proteins mediate this process by interacting with a cell's DNA. Here the problem of identifying sequence-specific DNA binding sites of transcription factors is studied, taking two complementary approaches, one based primarily on identifying sequence features and the other exploiting a transcription factor's structure.

The first approach considers the problem of developing a representation for DNA binding sites known to be bound by a particular transcription factor, in order to recognize its other binding sites. The effectiveness of several commonly used approaches is compared, including position-specific scoring matrices, consensus sequences and match-mismatch based methods, showing that there are statistically significant differences in their performances. Furthermore, the use of per-position information content improves all basic approaches, and including local pairwise nucleotide dependencies within binding site models results in statistically significant improvements for approaches based on nucleotide matches. Based on the analysis, the best results when searching for DNA binding sites of a transcription factor are obtained by methods that use both information content and local pairwise correlations.

The second approach focuses on a particular structural class of transcription factors, the C_2H_2 zinc fingers, that comprise the largest family of eukaryotic transcription factors. Zinc finger protein-DNA interactions are modeled by their pairwise residue-base interactions that make up their structural interface using a modified support vector machine framework to find the favorability of each residue-base interaction. Unlike previous approaches, this framework includes not only examples of known interactions but also quantitative information about the relative binding affinities between different protein-DNA configurations. The resulting classifier performs well in a variety of cross-validation testing.

Acknowledgments

Chapter 2 is joint work with Elena Zaslavsky and Mona Singh, and initially appeared in the journal Bioinformatics in 2004 [1]. Chapter 3 is joint work with Mona Singh.

I would like to thank Mona Singh for being my advisor, and acknowledge Professor Chazelle, Professor Schapire, Professor Troyanskaya, and Professor Wingreen for being members of my graduate committee.

Funding for this research was provided by a Department of Defense Natural Science and Engineering Graduate Fellowship, a Wu fellowship through the Princeton University School of Engineering and Applied Sciences, research associate support from Mona Singh via grants NSF MCB-0093399, DARPA MDA972-00-1-0031 and NIH GM076275, and other support from Princeton University.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Biological Background

The entire genetic complement of hundreds–soon to be thousands–of organisms has been determined. A major task is that of uncovering the relationship between an organism's genome and how it functions. In this thesis, computational methods to help understand the process of gene expression are developed, in particular to identify the sequence-specific elements that transcription factor proteins bind in order to control gene expression.

Each cell of an organism contains a copy of its genome, which itself contains the encoding for every protein the cell can produce. At any given moment, however, only a few of these potential genes are actually expressed, as it is both uneconomical and potentially deadly to produce proteins unless they are needed. Similarly, it is also dangerous to underproduce necessary proteins. A cell has several mechanisms for controlling gene expression and these mechanisms allow the cell to respond to changes in its environment (for example, when a bacteria finds a source of food), to signals sent from other cells in a multi-cellular organism (e.g., hormones), or to changes in the cell's internal state (e.g., progressing from cell growth to cell division). In multi-cellular organisms, tissues express different genes depending on their location and role within the organism.

There are several mechanisms for controlling which proteins are present in a cell, corresponding to distinct stages in gene production. Gene production proceeds from DNA to mRNA to protein, according to the central dogma of molecular biology. The production of mRNA from DNA is called transcription, while producing protein from mRNA is called translation. Of these, the most often used and most economical for controlling protein levels is transcriptional regulation: turning genes off before a mRNA transcript is made. If a protein is produced, it may be later modified, or it may also need to combine with other proteins to form larger complexes.

In order to start the production of genes, transcriptional machinery attaches to a segment of DNA nearby the target gene (called the promoter region), assembles itself, and proceeds to read the DNA sequences while producing the corresponding mRNA chain. The transcriptional machinery itself is composed of several proteins and the entire complex is referred to as RNA polymerase. In order for RNA polymerase to attach and begin, it often needs the help of additional proteins called (positive) transcription factors. Positive factors bind near the start of the gene and help RNA polymerase find its place. Performing exactly the opposite role, negative factors prevent the production of genes, for example, by competing directly with RNA polymerase for the same piece of DNA, or by binding to a piece of DNA between the promoter region and the gene (physically preventing RNA polymerase from transcribing the gene). Several transcription factors often coordinate the production of a single gene. For example, in *E. coli*, lactose metabolizing proteins are only produced in the presence of lactose and absence of glucose (glucose is the preferred energy source), and each condition corresponds to an individual regulating protein.

Transcription factors need to accomplish two goals: bind to a specific sequence of DNA, and interact with RNA polymerase or another protein. They perform these tasks using structurally distinct parts, called domains. Domains designed to bind DNA are called DNA binding domains (DBDs), or sometimes called a binding motif. If present, another domains may interact with RNA polymerase (either helping or hindering the

production of the nearby gene), other transcription factors, or other proteins. In many examples, these two domains functions independently, so it is possible to swap the DBDs of two transcription factors in order to change which genes they effect. The specific mechanisms of binding are rather intricate and will be explained for a single DBD in a subsequent chapter, but all that is needed currently is that transcription factors bind to specific regions of DNA near the beginning of a coding sequence, thereby influencing the production of nearby genes.

A transcription factor typically influences the production of many genes, so its presence or absence will most likely have a significant effect on the life of the cell. As factors are themselves proteins, their expression is also controlled, often by other transcription factors. In addition, transcription factors commonly negatively regulate themselves in order to stop production when their concentration is high enough.

Mutations in transcription factors can also have serious consequences. One example of this is found in *Drosophila* (fruit fly), where the change in a single transcription factor can cause the development of an extra leg or an extra pair of wings.

Besides transcription factors, the cell also has other techniques for regulating gene expression. For example, methyl groups can be added to individual bases in gene, which usually results in decreased transcription. Another technique for regulating gene expression involves the way in which DNA is packaged. Inside the cell, DNA is wound twice around nucleosomes, much in in the same way that thread is wound around a spool. Nucleosomes are connected together by linking DNA sequences, and when they are packaged together tightly outside molecules are unable to access the DNA sequence.

Gene expression is a complicated and rather marvelous process. Controlling gene expression involves the interactions of multiple proteins in long signaling pathways with several transcription factors often regulating one another. If that was not enough, all these molecules collaborate to produce the right proteins only where, and when, they are needed [2].

1.2 Contributions

This thesis focuses on one piece of the complex gene expression process: the relationship between a transcription factor and the specific DNA sites that it binds. This is an important component of unraveling the transcriptional circuitry for any genome, as the DNA binding sites of a transcription factor also reveal the corresponding regulated proteins. Since a single transcription factor can bind sites of considerable variability, it is difficult to find rules for identifying novel binding sites for a given protein, and much research in computational biology has focused on this problem.

Here, two complementary approaches to identifying the binding sites for a given transcription factor are considered. The first approach exploits sequence features of known binding sites for a transcription factor in order to recognize its other binding sites. The second approach builds a model for a particular single structural class of transcription factors in order to recognize binding sites for any protein within that class. While the first approach can predict binding sites for any transcription factor for which some binding sites are known, the second permits prediction of binding sites for proteins for which there are no binding sites, as long as the protein has the studied structure.

The first approach (described in chapter 2) considers the problem of developing a representation for a group of DNA binding sites known to be bound by a given transcription factor, in order to recognize its other binding sites. Commonly used methods for this problem include consensus sequences (e.g., [3]) and probabilistic approaches [4, 5]. A consensus sequence of a group of aligned binding sites is one that contains the most frequently occurring residue (or pair of residues) in each column, and nucleotide matches to a consensus sequence are used to evaluate the suitability of other putative binding sites. Probabilistic approaches, commonly referred to as position-specific scoring matrices (PSSMs) or weight matrices, assess the likelihood of observing a base in a particular position of the binding site.

The following basic methods for representing and searching for transcription factor binding sites are evaluated: consensus sequences, PSSMs, and a novel method that computes the average number of nucleotide matches between a putative site and all known sites. Whereas each of these basic methods assume that each base contributes independently to binding, it has been demonstrated that there are interdependent effects between bases [6, 7]. Similarly, the use of information content has been shown to be useful in representing binding sites [8] and in motif discovery [9]. Therefore, each basic method is extended to include either interdependent bases and/or information content. Cross-validation testing on a dataset of known *E. coli* transcription factor binding sites [10] shows that there are statistically significant differences between how well these methods identify binding sites. The use of per-position information content improves the performance of all basic approaches. Furthermore, including local pairwise dependencies within binding site models result in statistically significant performance improvements for approaches based on nucleotide matches. Based on this analysis, the best results when searching for DNA binding sites are obtained by methods that include both information content and local pairwise correlations.

The second approach (described in chapter 3) considers the problem of identifying binding sites of a regulatory protein when the overall structural interface is known. In this case, solved crystal structures of protein-DNA complexes for a structural family of transcription factors are used to determine conserved interactions between specific amino acids and nucleotides. Proteins within the same structural family exhibit different binding sites by varying residues in key DNA-binding positions.

The C_2H_2 zinc finger family of transcription factors were used as a model transcription family with a well conserved binding interface. This family is the largest known DNA-binding family in eukaryotes, and has been studied extensively (review, see [11]). C_2H_2 zinc finger proteins typically bind DNA target according in a well-known and conserved model, with specific residue and base combinations mediating the protein-DNA interac-

tion. There have been several previous approaches to uncover the favorability of different residue and base combinations in C_2H_2 zinc finger contacts, including a statistical mechanics based formulation [12] and one that uses an expectation-maximization approach on C_2H_2 zinc finger binding data [13]. The approach presented here differs from these in that it includes not only known C_2H_2 zinc finger protein-DNA interactions but also quantitative information about the relative binding affinities between different protein-DNA configurations. As high-throughput datasets with quantitative information about protein-DNA binding become more widely available (e.g., [7]), methods that can use such information will become increasingly important.

A modified support vector machine (SVM) framework was used to find the favorability of each residue-base interaction. Information about relative binding affinities of C_2H_2 zinc finger protein-DNA interactions are included as pairwise constraints which influence the model. This method was tested using stringent per-protein cross-validation, and shown to outperform or perform comparably with previously published methods. The results show that the SVM-based method holds great potential, especially as more high-throughput experiments give quantitative information about protein-DNA binding.

Chapter 2

Comparative Analysis of Methods for Representing and Searching for Transcription Factor Binding Sites

2.1 Introduction

This chapter analyzes how to represent binding sites for a particular transcription factor with the goal of searching for additional binding sites. A single transcription factor can bind sites of considerable variability; as a result, a number of different methods have been proposed (e.g., [3–5, 14–16]). Traditionally, a transcription factor's preference for binding has been represented by a consensus sequence (e.g., [3]), and more recently as a sequence logo [8]. Novel sites are typically found by either matching to a consensus sequence, or using position-specific scoring matrices (PSSMs) [4].

While many methods for identifying regulatory binding sites have been proposed, the availability of online datasets of transcription factors and their aligned binding domains (e.g., [10, 17]) allows us to quantify the effectiveness of different approaches. In particular, cross-validation testing is used to quantify how well each method performs in

distinguishing between the DNA binding sites for a one transcription factor and those of other proteins. While there may be some overlap between the binding domains for different transcription factors, the known DNA binding sites for the transcription factor under consideration should be among the top-ranked sites. Such an empirical evaluation is important and timely, as whole-genome scans in search of the binding sites are increasingly used to make functional annotations of uncharacterized proteins, and to infer properties of transcriptional regulatory networks (e.g., [18]). Additionally, the previously mentioned methods are the basis for other more sophisticated approaches for predicting transcription factor binding sites, including motif discovery and cross-genomic approaches (e.g., [9, 19–25]).

This chapter evaluates four basic methods for representing and searching for transcription factor binding sites: consensus sequences [3], two variants of position specific scoring matrices (log-odds matrices, and the statistical mechanics based Berg and von Hippel method [5]), as well as a method based on nucleotide matches, called *Centroid*, that computes the average number of nucleotide matches between a putative site and all known binding sites.

Each basic method is considered with two natural extensions: pairwise nucleotide dependencies and per-position information content. Whereas the basic methods assume that each base contributes independently to binding, it has been demonstrated that there are interdependent effects between bases [6, 7]. Though the independence assumption has clearly been useful in practice and seems to provide a good approximation to the energetics of DNA-protein binding [26], here it is assessed whether improvement is possible by using pairwise dependencies. Similarly, the use of per-position information content was shown to be useful in representing binding sites [8] and in motif discovery [9]; here, it is applied directly to the problem of searching for binding sites by using using the information content of a position to weigh its contribution towards the overall score.

These methods and their extensions are compared in how well they perform in identifying the binding sites for a particular transcription factor without additionally identifying binding sites for other proteins. Improvement in performance are assessed using the matched-pairs signed-ranks test as well as receiver operating characteristic (ROC) curves. The rank test evaluates whether the frequency with which one method outperforms another is statistically significant, and a ROC curve compares the performance of two or more methods over a range of possible false positive rates.

Testing on a dataset of *E. coli* transcription factor binding sites [10], there are statistically significant differences between these methods. The main findings are:

1. Using per-position information content to weigh positional scores improves the performance of all methods, sometimes dramatically. For example, consensus sequences have by far the poorest performance of all basic methods in discriminating between binding sites for the transcription factor of interest and binding sites of other transcription factors; however, weighing each match to a consensus base by the appropriate per-position information content makes consensus sequences much more competitive with other methods.

2. Methods based on nucleotide matches, such as consensus sequences and Centroid, show statistically significant improvements when including pairwise nucleotide dependencies. Furthermore, in these cases, the choice of which pairs to include in the model is important; in particular, considering all possible pairs of nucleotides in a binding site is not as effective as using just neighboring pairs. Somewhat unexpectedly, probabilistic methods, such as log-odds PSSMs, do not show statistically significant improvements when including pairwise dependencies.

3. The difference in performance between methods decreases substantially when both information content and pairs are used.

In general, when searching for DNA binding sites, methods using information content and pairwise dependencies were found to be most effective. For organisms like *E. coli* with many well-characterized transcription factors and binding sites, analysis similar to the one performed here should aid in choosing a specific method and suitable threshold.

Software implementing all the methods discussed in this chapter was included with the original publication [1]. Appendix 2.A briefly describes some of the issues that were addressed during implementation. Finally, appendix 2.B describes a statistical procedure used throughout the chapter to assess the significance of multiple hypotheses.

2.2 Methods

2.2.1 Dataset

[10, 22] contains 68 regulatory proteins and their aligned DNA binding sites; the dataset used in this chapter was constructed from it as follows. First, only proteins with at least four binding sites were considered. Second, in the original database, occasionally the binding sites for a single regulatory protein were split into multiple groups based on the number of tandem duplications; individual sites for ArgR, MetJ, and PhoB were included rather than their tandem-repeated counterparts. Third, binding sites from sigma factors were removed, as were binding sites from NarP, since all the latter are also binding sites for NarL. Fourth, duplicate binding sites were removed in order to preserve leave-one-out cross-validation. Finally, each binding site was located within the *E. coli* K-12 genome (version M54 of strain MG1655 [27]), and was extracted along with flanking regions on each side. Binding sites that could not be located unambiguously within the genome were excluded from the study. This process left 35 transcription factors and 410 binding sites, with an average of 11.7 ± 8.5 (standard deviation) sites per transcription factor. Figure 2.1 shows the number of sites and site length for each transcription factor in the final dataset.

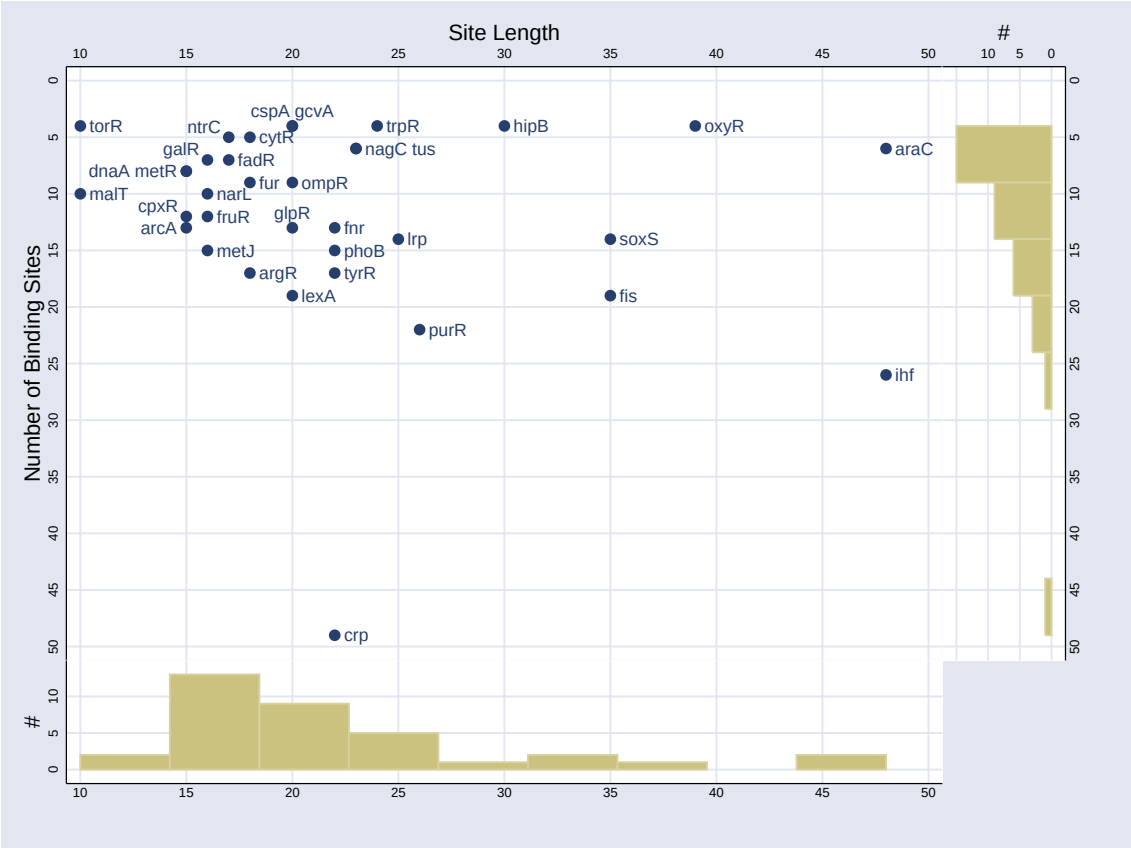


Figure 2.1: Number of sites and site length for each transcription factor in the final dataset. Marginal distributions are shown as histograms near the bottom and right hand side. This and subsequent charts were designed according to the principles of graphical excellence and data integrity, as demonstrated and explained in [28–30].

2.2.2 Approaches for Predicting Binding Sites

Four basic approaches for searching for transcription factor binding sites were evaluated. Specific implementation details were found to affect performance, and so each of the methods is described briefly below.

First, a word about notation. Let S be the N binding sites for a particular transcription factor. Each binding site has length l and that these binding sites are aligned. Define $n_i(b)$ to be the number of times base b appears in the i-th position of any sequence in S, and $f_i(b)$ to be the corresponding frequency. Similarly, define $n(b)$ to be the number of times base b appears overall in the N binding sites, and $f(b)$ to be the overall frequency for base b. Each method is used to score a new DNA subsequence t (also of length l) in an attempt to predict whether t is a binding site of the given protein. Let t_i denote the i-th base of the sequence t to be scored.

Extending the above notation to pairs of positions, let $n_{ij}(b, d)$ be the number of times the ordered pair of bases (b, d) occurs in positions i and j of any sequence belonging to S, and $f_{ij}(b, d)$ be the corresponding frequency. Ideally, pairwise interdependencies should only be included for those pairs that are known, perhaps through structural studies, to act together in determining DNA-protein binding specificity. Since such precise information is not always readily available, as a first approximation consider only pairwise dependencies between nearby positions. We introduce the notion of *scope* to delimit which pairs are considered important when determining specificity. For instance, a scope of one restricts dependent positions to adjacent pairs while a scope of two considers both adjacent pairs and pairs separated by an intermediate base.

Next, define the information content (**IC**) of a position in a binding site. Information content is based on the information theoretic notion of entropy introduced in a seminal paper by Claude Shannon [31]. In the current application, the entropy of a position expresses the average number of bits necessary to describe the position in a binding site, and the information content of a position is calculated by subtracting its entropy from

the value of the maximum possible entropy. That is, the higher the information content, the more conserved (and presumably more important) the position. More specifically, the information content IC_i of position i in S is defined as $2 + \sum_{b \in DNA} f_i(b) \log f_i(b)$. The total information content of a transcription factor's binding sites is computed by summing IC_i over all positions, and varies from 12 to 42 bits. The information content IC_{ij} of a pair of positions is $4 + \sum_{b,d \in DNA} f_{ij}(b,d) \log f_{ij}(b,d)$ [14]. Information content is used in Sequence Logos by [8] mainly as a visualization tool to identify important positions in a binding site. A different, more direct usage in a scoring scheme is proposed here, namely by including the IC of a position as a multiplicative factor in scoring a target binding site sequence.

The following basic methods were used:

Consensus: These methods vary considerably [3]; a version of consensus sequences described by [32] was used. For each position i, let b be the most frequent base and d be the second most frequent base. If $f_i(b) > 0.5$, then b is the consensus base for position i (denoted by $\mathrm{consensus}_i$); otherwise if $f_i(b) + f_i(d) > 0.75$ then both b and d are the consensus bases. If neither is true, there is no consensus base for this position. The score of a new sequence t is obtained by counting the number of times t_i agrees with the consensus base for the i-th position.

PSSM: Typically, this method assumes independence between positions, and computes a log-odds score for a potential binding site. A commonly used Bayesian estimate to handle the zero frequency case was used, replacing $f_i(b)$ by $\hat{f}_i(b) = \frac{n_i(b) + \hat{f}(b)}{N+1}$ [33], where $\hat{f}(b)$ is the estimate of overall background frequency of base b, computed as $\frac{n(b)+.25}{Nl+1}$.

Berg and von Hippel: The full analysis was conducted using a statistical mechanics-based method that makes the connection between base-pair statistics of sites and its binding free energy. Denoting the number of occurrences of the most common

base in position i of binding sites by $n_i(0)$, the method scores a new sequence t by computing a per-positional corrected log-odds score of observing a base of t versus the most frequent base in the corresponding position of the sequences [5, 34].

Centroid: This novel method scores a sequence t by computing the average shared identity between t and every sequence in S.

Next, extensions of the above methods that include pairwise dependencies were considered:

Consensus-P: For a sequence t, this method counts both the number of nucleotides matching the consensus sequence and the number of nucleotide pairs within a given scope matching the corresponding bases in the consensus sequence.

Centroid-P: This method considers the number of shared bases as well as the number of shared pairs of bases within a particular scope between the sequence t and each sequence in S.

PSSM-P: This method is an extension of the PSSM log-odds method that also accounts for pairwise dependencies. Although rigorously generalizing PSSM-P beyond adjacent pairs is not difficult in principle, in practice the small number of known sites per transcription factor limits the rigorous probabilistic derivation of the method to only adjacent pairs [35]. For example, a derivation of scope two requires calculating triplet frequencies. Instead, the analysis evaluates an intuitive definition of the method that considers only pairwise dependencies regardless of scope.[1] A standard Bayesian approach was used to handle the zero frequency case by replacing $f_{ij}(b,d)$ by $\hat{f}_{ij}(b,d) = \frac{n_{ij}(b,d)+\hat{f}(b)\hat{f}(d)}{N+1}$. Several different ways of computing the reference "background" pair frequencies were evaluated; modeling this as the product of single column frequencies had the best overall performance.

[1] For scope value of one, the rigorous derivation that assumes that position i depends on position $i+1$ subtracts single columns log-odds scores; the method described here tests better without subtracting these singlet scores.

Method	Column Score	Pair Score ($j = i + s$)
Consensus	$[t_i \in \text{consensus}_i]$	$[(t_i, t_j) \in \text{consensus}_{ij}]$
PSSM	$\log \frac{\hat{f}_i(t_j)}{\hat{f}(t_i)}$	$\log \frac{\hat{f}_{ij}(t_i,t_j)}{\hat{f}(t_i)\hat{f}(t_j)}$
Berg and von Hippel	$\log \frac{n_i(t_i)+0.5}{n_i(0)+0.5}$	$\log \frac{n_{ij}(t_i,t_j)+0.5}{n_{ij}(0,0)+0.5}$
Centroid	$f_i(t_i)$	$f_{ij}(t_i, t_j)$

Table 2.1: Column and pair scores computed by various methods. The final score for a basic method is $\sum_{i=1}^{l} \bullet$, where $\bullet$ is the column score listed above. The final score for a pair method is the basic score plus $\sum_{s=1}^{\text{scope}} \sum_{i=1}^{l-s} \bullet\bullet$, where $\bullet\bullet$ is the pair score listed above.

Berg and von Hippel-P: The Berg and von Hippel method was extended to include pairs of bases in a similar manner, with $n_{ij}(0, 0)$ giving the most frequent pair of bases in positions i and j.

Finally, for every method considered, its variation in which per-position information content is used to weigh the contribution of each position (or pair of positions) towards the overall score is examined.[2] For instance, the score computed by Centroid IC is $\sum_{i=1}^{l} \text{IC}_i \, f_i(t_i)$, and the score computed by its pair counterpart Centroid-P IC with scope parameter *scope* is the sum of the Centroid IC score and $\sum_{s=1}^{\text{scope}} \sum_{i=1}^{l-s} \text{IC}_{ij} \, f_{ij}(t_i, t_j)$, where $j = i + s$. All of these methods and variations are summarized concisely in table 2.1.

2.2.3 Cross-validation Testing and Analysis

The most common use of any of the methods described above would be to scan non-coding regions in a genome in order to find possible binding sites. This entails scoring consecutive windows of appropriate length and considering windows that score above a chosen threshold to be predicted binding sites. However, such a framework is not easily applicable when evaluating and comparing different methods; for example, the *E. coli* genome contains many yet uncharacterized binding sites, and predicted windows may

[2] [14] suggest using a sampling error correction based on the expected information content of n random samples. This correction did not improve performance during testing and so only uncorrected information content is reported.

correspond to true binding sites even if they are not annotated in the dataset. Instead, leave-one-out cross-validation studies are conducted to evaluate each method, considering each binding site s in turn. Suppose s belongs to known binding sites S, each of length l, for transcription factor TF. The method under consideration then uses all the sites except s, i.e. $S - \{s\}$ to build the binding site representation for TF, and scores s as well as the negative examples. Negative examples consist of all binding sites except those known to be bound by TF. A negative binding site t is scored by examining all possible alignment positions of this binding site against the binding site representation of TF such that either the representation of TF is completely contained within t, or t is completely contained within the representation of TF. In the latter case, genomic flanking regions around t are used for scoring. Six pairs of binding sites were found to reside completely inside one another in the genome. In these cases, when scoring the negative binding site, a true binding site for the transcription factor of interest is present; thus these corresponding binding sites were removed from the pool of negative examples during cross-validation testing. The final score for a target sequence is taken to be the higher score when considering both the original sequence and its reverse complement. It is still possible that transcription factor TF can bind some of the negative examples, but nevertheless s should be among the top scoring sites.

The discriminatory power of each method, exhibited in the relative score of the actual binding site among all scored sites, is analyzed using two data-mining tests: averaged ranks and receiver operating characteristic (ROC) curves (e.g., [36]). In particular, for each site s of a transcription factor under consideration, its rank in cross-validation testing is computed by counting how many negative examples score as well or better than s, with lower rank indicating better performance. Then, to compare how well two methods perform, a matched-pairs signed-ranks test is used. Briefly, the number of times one method outperforms the other is compared with how many times such an event would happen merely by chance under the assumption that both methods perform equally well.

P-values of less than .05 are considered significant. For ROC analysis, a ROC curve was first created for each individual leave-one-out test (i.e., keeping track of whether the binding site was found as a function of the number of false positives allowed) and then averaged over all sites for that transcription factor. Curves are then further averaged across the various transcription factors to arrive at a final curve for each method.

2.3 Experimental Results

2.3.1 Comparison of Basic Methods

This section established baseline performance of each of the four basic methods. Figure 2.2 compares the performance of the Consensus, PSSM, Berg and von Hippel and Centroid methods using ROC analysis. Each curve plots the fraction of correctly classified positive examples (TP rate) as a function of the incorrectly classified negative examples (FP rate).

As expected, Consensus performs markedly the poorest, consistently lying to the lower-right of the other curves. The remaining methods are comparable, as their curves lie very close to one another and cross at various FP rates.

As seen in this test, and in all of the following testing scenarios, PSSM and its variants perform virtually identically to Berg and von Hippel's method and its variants. Therefore, results omit the latter method in order to simplify the analysis.

2.3.2 Influence of Pairwise Dependencies

Next, the performance of the basic methods described above is evaluated considering the effect of adding pairwise correlations. Ideally, a method for including pairwise correlations should only take into account those pairs that are known, perhaps through structural studies, to act together in determining DNA-protein binding specificity. Such precise information is not readily available, and so, a first approximation focuses on considering pairwise correlations between bases that are nearby in sequence. A comparison of results

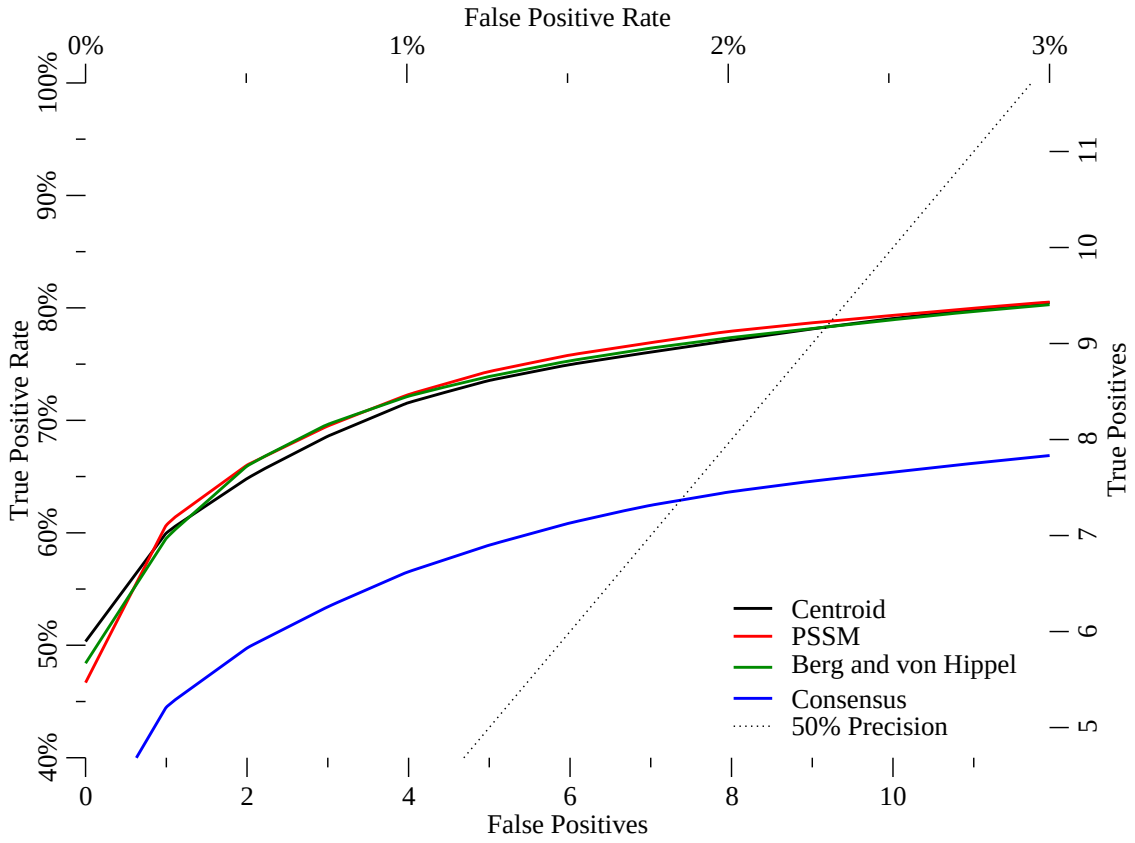
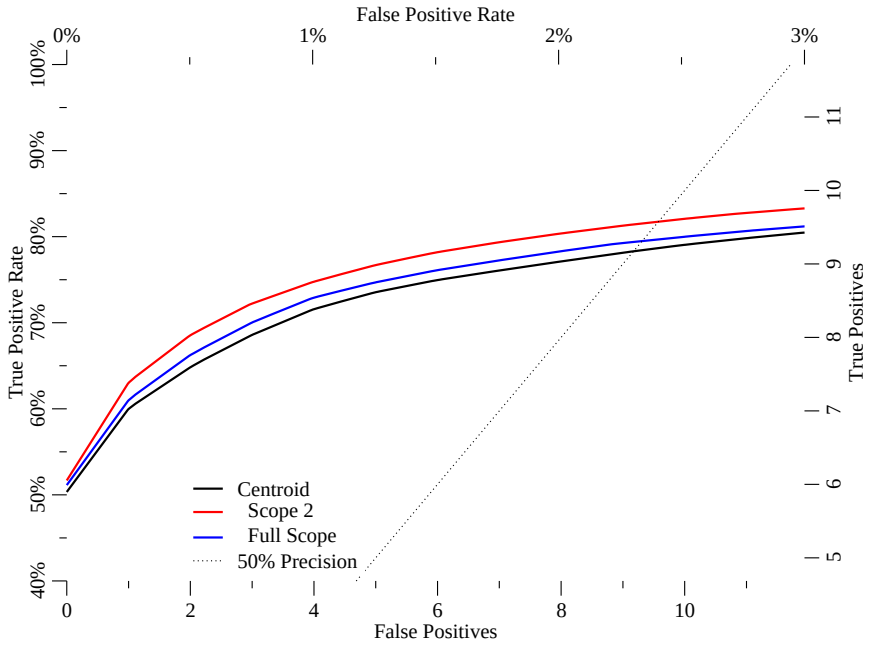
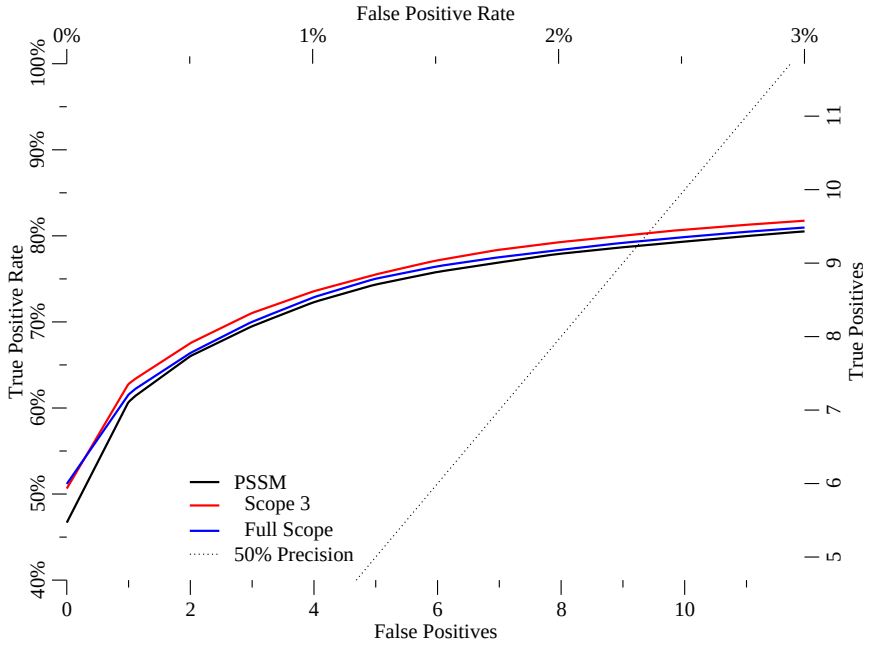


Figure 2.2: ROC curves comparing performance of the four basic methods: Centroid, PSSM, Berg and von Hippel, and Consensus. The top and left axes indicate average false positive rate and true positive rate, respectively (for a given false positive rate, true positive rates were averaged over all transcription factors to give the shown curves). The bottom and right axes shows the average number of binding sites corresponding to each rate (there are a total of 410 binding sites and 35 transcription factors in the current dataset). The 50% precision line indicates the boundary at which the methods would predict as binding sites as many incorrect sites as correct ones. Consensus is clearly outperformed by the other three methods.

is done using pairwise correlations as the positional distance allowed between pairs of bases (i.e., the scope) is varied. Figure 2.3 summarizes the effect of considering various pairwise correlations for centroid and PSSM. The effect of nearby pairwise correlations are further quantified by scope parameters between zero (where no pairwise dependencies are assumed) and four are considered, as well as full scope. For Centroid-P, neither zero scope nor full scope performs best, whereas curves with scopes in the range of two to four consistently achieve higher TP rates across the relevant range of FP rates (results are only shown for scope two). The improvement for scope two is modest (3% improvement when allowing no false positives, and approximately 5% when allowing one false positive) yet significant with a p-value of less than 10^{-4}, as judged by the matched-pairs signed-ranks test (section 2.3.4 on page 22). Thus, including pairwise correlations improves the discriminatory ability of the centroid method; however, it is important to consider only certain pairs of positions. In the remainder of this chapter, Centroid-P is used with scope two as it is less computationally intensive than those with scopes three and four and yet performance is comparable. A similar trend is observed for Consensus (not shown), where a dramatic improvement in performance occurs with the addition of pairwise correlations. At scope two, used for all subsequent analysis, performance increases over scope zero by 25% when allowing no false positives, and 26% when allowing one false positive. In contrast, for PSSM-P, including pairs at small scopes results in a performance decline (section 2.4 on page 26). At larger scopes PSSM-P performs very similarly to PSSM. For the remaining analysis, PSSM-P is used with scope three; performance increases over scope zero by 8% with no false positives and by 3% with one false positive. However, this improvement is not statistically significant (see below). To summarize, for the methods based on tallying up nucleotide matches, such as Centroid and Consensus, considering pairwise correlations clearly helps. However, the same claim cannot be made for probabilistic methods such as PSSM; perhaps because these methods are less stable when very few sites are known.



(a)



(b)

Figure 2.3: ROC curves comparing performance when pairs are considered for Centroid (a) and PSSM (b). For each, the basic method (scope zero) is shown, along with the method using all possible pairs (full scope) and the method using the best performing scope (scope two for Centroid-P and scope three for PSSM-P). In all further testing Centroid-P is shown with scope two while PSSM-P is shown with scope three.

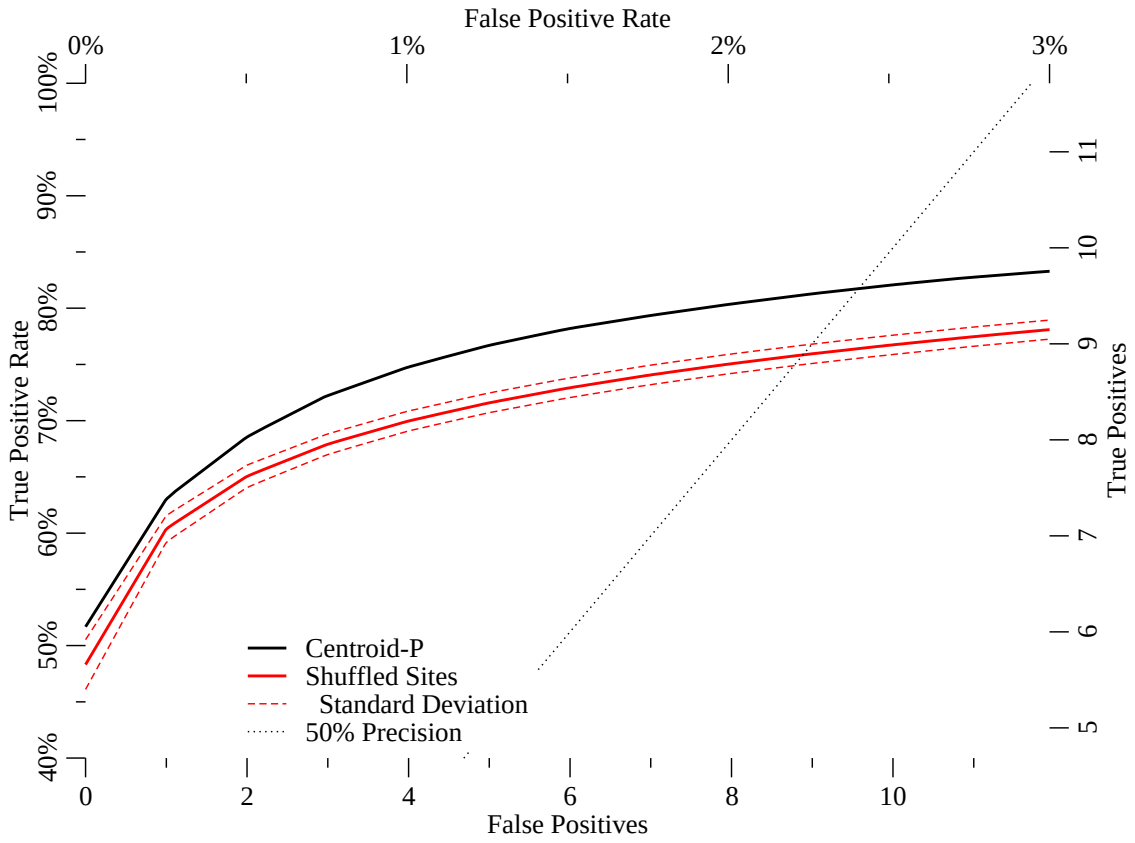


Figure 2.4: ROC curves comparing Centroid-P with scope two using regular sites and sites with columns shuffled. The solid curve is an average over 1,024 different shuffles, while dashed curves show performance out to one standard deviation with random shuffling.

The improvement gained by using pairwise correlations is further quantified by considering the performance of the Centroid-P method in the same cross-validation scenario but on a perturbed dataset, produced by randomly shuffling the columns of the binding sites used as positive examples. While shuffling the columns for binding sites preserves per-column nucleotide composition, it also, on average, destroys any local pairwise correlations found in the original alignment. The results are shown in figure 2.4 where a ROC curve for Centroid-P tested on the original dataset is plotted against the same method tested on the shuffled dataset. Shuffling and cross-validation are averaged over 1,024 trials producing the solid shuffling curve, while the dashed curves show performance out to one standard deviation (due to the effects of randomness). The benefit of including

nearby inter-column correlations is clearly observed, as performance on shuffled sites is consistently worse than performance on the original sites.

2.3.3 Influence of Per-position Information Content

Next, the performance of the basic methods described above is evaluated considering the effect of adding per-position information content to each method. A rank chart compares the performance of the Consensus, PSSM and Centroid methods, along with their pair counterparts with and without information content.

Figure 2.5 shows average ranks (as computed over the binding sites for each transcription factor) for both versions of each method and its pair extension. Comparing median performance, it is clear that adding per-position information content results in improved performance in both the original and pairwise versions of the basic methods. Noticeably, the addition of information content to the Consensus method dramatically improves its performance, and in fact makes it much more competitive with the other methods.

When considering performance differences of both pairwise dependencies and per-position information content at particular values of false positives, basic Consensus shows a 36% improvement when allowing no false positives and a 37% improvement when allowing one false positive. These values for Centroid are 2% and 9%, and for PSSM are 10.5% and 8%.

2.3.4 Statistical Significance of Methods Comparison

A matched-pairs signed-ranks test is used to compare methods and assess whether the differences in performance (partially described above) of various methods are statistically significant. The change in the rank of the left-out-example is calculated for every comparison and each cross-validation test. These rank differences are converted into p-values under the assumption that both methods perform equally well. Calculated p-values represent the probability that the observed differences in performance could have occurred

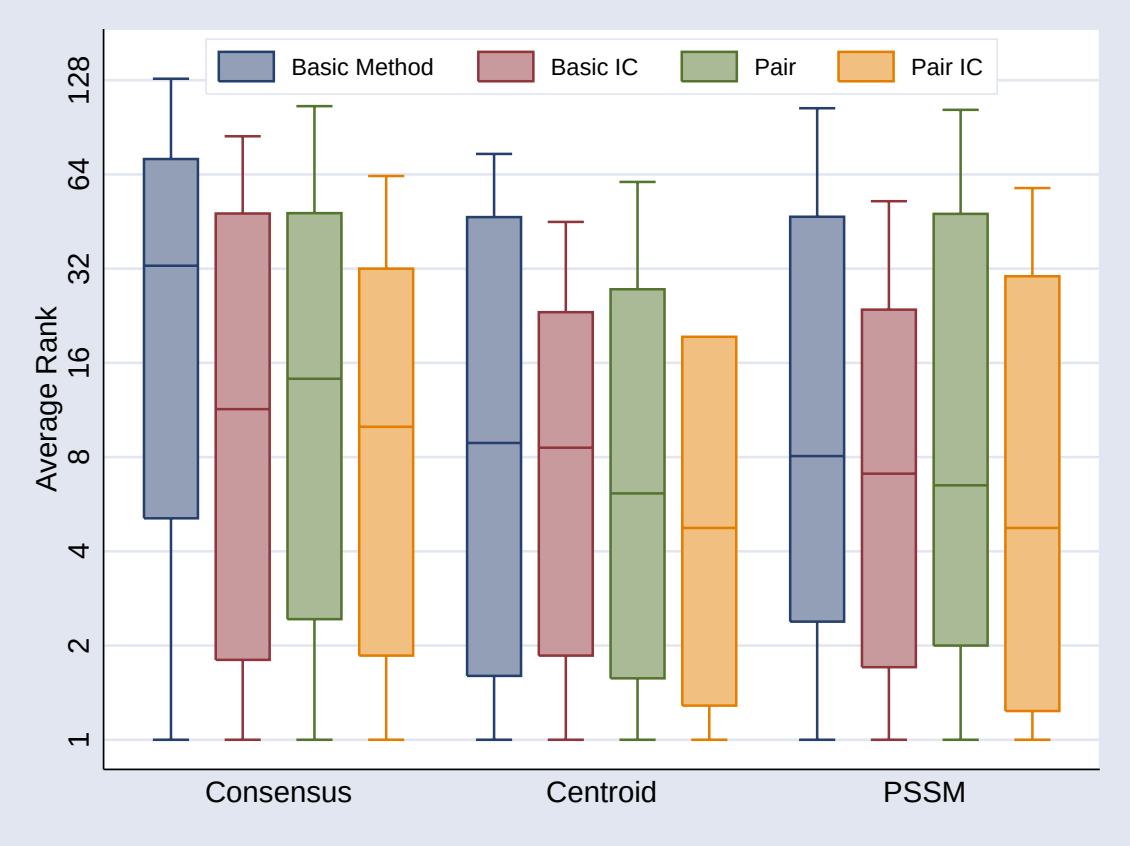


Figure 2.5: Performance of methods based on averaged rank. For each transcription factor, an average rank is computed from the rank of each of its binding sites in cross-validation testing. The horizontal line in each box is the median transcription factor's average rank while each box shows the 25th–75th percentiles for average ranks.

by chance alone, and an individual test is considered significant if the p-value associated with it is small enough. The criteria for significance is taken from a multiple hypothesis test, which limits the overall probability of judging any combination of individual tests as statistically significant due to chance alone. For example, the classical Bonferroni multiple hypothesis test requires that $p \leq \alpha/n$ for each individual test, where n is the number of tests being performed and $1 - \alpha$ is the desired overall significance level for the entire procedure. An improved multiple hypothesis test called sequentially rejective Bonferroni [37] was used. This procedure is described fully in appendix 2.B. Sequentially rejective Bonferroni is statistically more powerful than the classical Bonferroni test while still guaranteeing that the overall probability of accepting any combinations of comparisons as statistically significant by chance alone is less than α [37]. An alpha value of 5% was used during testing, making the overall procedure significant at the 95% level.

Several possible pairs of methods were chosen to be tested with the goal of identifying the best performing methods and quantifying improvement (if any) resulting from including information content and pairwise dependencies. For each basic method, all its variations are compared to each other; additionally, the versions of every method that includes both pairs and information content are compared to each other. The results are shown in figure 2.6, producing a graph in which a directed edge connects a pair of methods, one with a significant performance improvement over the other. The overall conclusion is that including pairs and information content for each basic method outperforms the other methods in its group (with the exception of Consensus-P IC which did not perform significantly better than Consensus-IC). As for the overall best method, Centroid-P IC has the best average rank; and both Centroid-P IC and PSSM-P IC statistically outperform the highest number of other methods. All information content weighted methods perform significantly better than their non-weighted counterparts, with a qualification that although Centroid-P IC and Centroid IC perform better than Centroid-P and Centroid at individual p-values of .02, these differences are not statistically significant with the

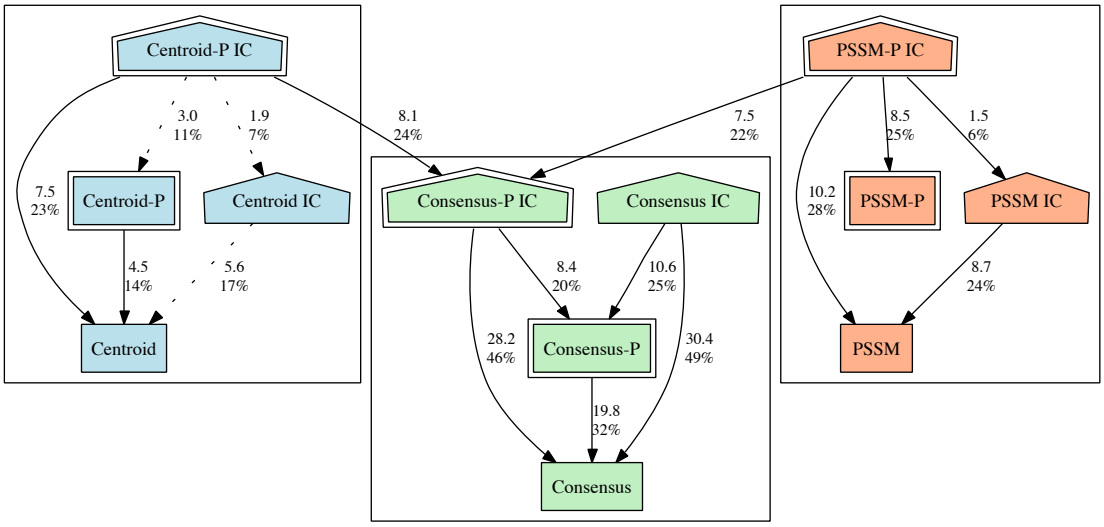


Figure 2.6: Partial ordering (at the 95% significance level) of methods based on a signed ranks test. Arrows point towards worse performing methods, with labels indicating the difference in average rank between the methods, both as an absolute number and as percentage improvement. Pairwise comparisons were performed between all four variations of each method, as well as between the three P IC methods, for a total of 21 tested hypotheses. P-values were adjusted for multiple hypothesis testing using a sequentially rejective Bonferroni test [37]. Dotted edges show differences that were not found to be significant with the Bonferroni correction, but have individual p-values $< .05$.

Bonferroni correction. The improvement resulting from adding pairwise dependencies is more modest, observing only some of the possible arrows relating a method and its pair counterpart.

2.4 Discussion

Based on the results presented in section 2.3, both pairwise dependencies and especially information content can be used to improve the discriminatory power of computational methods for binding site recognition and prediction.

The influence of pairwise dependencies within binding sites has been a topic of debate within the computational biology research community, and numerous papers have been published supporting both viewpoints (e.g., [7, 38]). In part, the findings are consistent with both opinions, as considering pairs improves performance for all three methods, though the improvements are not significant for PSSM-P (Figure 5). Moreover, while the resulting improvement is dramatic for Consensus-P, it is more modest for Centroid-P. Nevertheless, testing suggests that inter-positional information can provide additional binding domain specificity, especially when the appropriate pairs are considered. This is demonstrated by the fact that considering some pairs of positions (i.e., those within close sequential proximity) results in improved performance for the nucleotide match methods such as Centroid-P; it may be possible that more careful selection of pairwise dependencies, perhaps from crystal structures, would result in further improvements.

The performance of the statistical methods, such as PSSM-P, do not show statistically significant improvements when using pairs most likely because of the non-occurrence of many base-pair combinations in the small dataset; such a case is more severely penalized in PSSM-P scoring than in Centroid-P or Consensus-P scoring.

It is not clear that the addition of information content should improve performance of methods that already include frequency information. Nevertheless, information content benefits all methods tested, and a clear trend is observed when considering the methods

and their IC counterparts (figure 2.5 on page 23). Additionally, the performance of most methods is comparable once per-position IC weights have been included. This suggests that information content is the measure that allows us to rigorously identify the highly conserved positions in the binding site; these are presumably the functionally important positions in the interaction between a transcription factor and its DNA binding domain. Given those key positions, the precise way of making use of that information appears less critical; even Consensus, a clearly inferior method utilizing the simplest matching criterion, receives noteworthy improvement from including information content.

Finally, some variation in method performance per transcription factor was observed. This suggests that no single method is optimal for all situations. This is not surprising given the high degree of variation observed in protein-DNA interactions. Whereas in general methods using information content and pairwise nucleotide information are expected to be most effective when searching for DNA binding sites, for a specific transcription factor and its binding sites, an alternate method may perform better. Additionally, in some scenarios it is desirable to allow a higher number of predicted binding sites that can be later eliminated using other approaches (e.g., using cross-genomic information). Analysis similar to the one performed here is likely to prove useful in choosing, for different contexts, a specific method and suitable threshold for finding binding sites.

2.A Appendix: Software Implementation

This chapter compared different methods for identifying previously known binding sites. However, in more natural circumstances the task is to find probable binding sites in a new sequence. A collection of known binding sites for a given protein is still assumed to be available, but now the task is to give the user the ability to gain confidence that a newly discovered subsequence is an actual binding site.

A software utility was created containing all the methods described in section 2.2.2. All implementational issues are hidden to the user and he or she simply specifies which method (basic method, scope, and whether or not to use information content) and criteria are used to identify potential new sites. These criteria are described briefly below.

Top: Only show the top scoring subsequence within a given sequence. This could be useful, for example, when scanning the upstream region of a gene to determine whether a transcription factor has any direct influence on the expression of that gene.

Cutoff: Only show subsequences with a score greater than a previously determined cutoff. This could be useful when the user has already performed analysis of scores on positive and negative sites and has determined a threshold suitable for his or her needs. For example, choosing a cutoff equal to the lowest score for known sites allows the user to find sites which score as well or better than all known sites.

P-value: Choose a cutoff so that the probability of a randomly generated site scoring higher than the cutoff is equal to a given p-value. This is further explained below.

False-Positive Rate: In this case, negative example sequences are also provided by the user. Each example sequence is scanned, recording its highest scoring subsequence. A threshold is chosen so that only a given percentage of these sites would contain at least one potential new site. This could be useful when the user has a collection

of upstream regions (for example, all upstream regions in a genome), and would like to control the number of false positive sites returned.

Additionally, criteria can be combined as required (for example, only showing the top site in a sequence, and only if it has a low p-value). The p-value criteria requires some additional clarification. The user specifies the overall expected GC content of scanned sequences and base frequencies for random sequences are chosen accordingly, with complementary bases assumed to occur with the same probability. Random subsequences are generated by choosing a base independently for each position over the entire length of the site. Given this random model and assuming a non-pairs method, the score for a randomly generated subsite equals,

$$S = \sum_i S_i$$

where S_i are random variables for position scores, $S_i = S_i(b)$ with probability $\mathrm{P}[\, b \,]$, and i ranges over the length of the site. The expectation and variance of S can be calculated using linearity of expectation and linearity of variance for independent variables (e.g., [39]),

$$\mathrm{E}[\, S \,] = \sum_i \mathrm{E}[\, S_i \,]$$
$$\mathrm{Var}(S) = \sum_i \mathrm{Var}(S_i)$$

For pair-based methods the score includes pair terms,

$$S = \sum_i S_i + \sum_{ij} S_{ij}$$

where S_{ij} are random variables for pair scores, $S_{ij} = S_i(b, d)$ with probability $\mathrm{P}[\, (b, d) \,] = \mathrm{P}[\, b \,]\mathrm{P}[\, d \,]$, and ij iterates over all pairs within a given scope.

Expected score and variance are calculated as follows,

$$\mathrm{E}[\,S\,] = \sum_i \mathrm{E}[\,S_i\,] + \sum_{ij} \mathrm{E}[\,S_{ij}\,]$$

$$\mathrm{Var}(S) = \sum_i \mathrm{Var}(S_i) + \sum_{ij} \mathrm{Var}(S_{ij}) + 2\sum_{i,\,jk} \mathrm{Cov}(S_i, S_{jk}) + 2\sum_{ij<kl} \mathrm{Cov}(S_{ij}, S_{kl})$$

where $\mathrm{Cov}(S_i, S_{jk}) = 0$ and $\mathrm{Cov}(S_{ij}, S_{kl}) = 0$ unless the column/pair(s) overlap. Covariances are needed because the random variables are correlated. The last summation computes the covariance between pair-scores, considering each pair of pairs only once.

Having calculated the mean (μ) and standard deviation ($\sigma = \sqrt{\mathrm{Var}(S)}$) of scores for random subsites, the distribution of scores is approximated by a normal distribution with first two moments matching, $S \approx N(\mu, \sigma)$. Finally, a cutoff is chosen that gives the desired p-value.

$$\mathrm{P}[\,N(\mu, \sigma) \geq \text{cutoff}\,] = \text{p-value}$$

2.B Appendix: Bonferroni Multiple Hypothesis Testing

Calculated p-values represent the probability that the observed differences could have occurred by chance alone. An individual test is considered significant if its p-value is small enough.

A multiple hypothesis test limits the overall probability of judging any combination of individual tests as statistically significant due to chance alone. The classical Bonferroni multiple hypothesis test requires that $p_i \leq \alpha/n$ for each individual test, where n is the number of tests being performed. Let $i \in I$ be the hypotheses that are false (that is, their null-hypotheses are true), and let $m = |I|$. The probability of rejecting all false hypotheses can be bounded using the union bound.

$$\begin{aligned}
&\mathrm{P}[\, p_i > \alpha/n \text{ for all } i \in I \,] \\
&= 1 - \mathrm{P}[\, p_i \leq \alpha/n \text{ for any } i \in I \,] \\
&\geq 1 - \sum_{i \in I} \mathrm{P}[\, p_i \leq \alpha/n \,] \\
&= 1 - m(\alpha/n) \\
&\geq 1 - \alpha
\end{aligned}$$

Therefore, $1 - \alpha$ is the overall significance level for the entire procedure. A less strict criteria α/m could have been used instead of α/n, except that m is not known.

Sequentially rejective Bonferroni [37] realizes a less strict bound without knowing m, while still ensuring the same overall level of significance. First, p-values are ordered so that $p^{(1)} \leq \ldots \leq p^{(n)}$. The first $p^{(i)}$ is found such that $p^{(i)} > \alpha/(n+1-i)$. Hypotheses $p^{(1)}$ through $p^{(i-1)}$ are accepted as statistically significant, while $p^{(i)}$ through $p^{(n)}$ are rejected as statistically inconclusive (tied p-values are either all accepted or all rejected). The criteria for the smallest p-value is the same as in the classical test, while the criteria for the largest p-value (if it wasn't rejected earlier) is the same as in a single hypothesis test.

To show that this procedure is valid, consider the event $A = \{p_i > \alpha/m \text{ for all } i \in I\}$. $\mathrm{P}[\,A\,] \geq 1 - \alpha$, as shown above. Consider the smallest p_i for $i \in I$. We don't know the position of the p_i among the other p-values, but it is the smallest of m values, so $i \leq n + 1 - m$. This is equivalent to $m \leq n + 1 - i$. Given that event A has occurred,

$$p^{(i)} > \alpha/m \geq \alpha/(n + 1 - i)$$

Therefore, the first false hypothesis will be rejected, and the same for all false hypotheses following it.

The criteria for individual tests is not as strict as in classical Bonferroni, so the overall procedure is statistically more powerful (accepts more hypotheses as true). This advantage can be most clearly seen in the following scenario. Suppose hypotheses p_1 through p_a are true, but are nevertheless included in the multiple hypothesis testing. Their p-value will be approximately zero, so they will be the smallest p-values tested and accepted as true. The remaining hypotheses are tested as if the true cases were never present, while, in the classical multiple hypothesis test, all hypotheses are tested using a more stringent criteria.

Chapter 3

Protein-DNA Interactions: Including Relative Binding Affinity Using SVMs

This chapter presents a framework for predicting the binding sites of a transcription factor using knowledge about its 3D structure. The method presented may be seen as a natural extension of the previous chapter—whereas there binding sites were represented with no reference to the actual protein, in this chapter the overall structural interface between the transcription factor and DNA is known and modeled explicitly. This allows a more expressive model of binding to be built, representing both DNA bases and the amino acids contacting them, and provides a means for predicting the binding sites of other structurally similar proteins, even those for which no binding sites are known.

The framework presented in this chapter for structure-based prediction of transcription factor binding sites has been developed for the C_2H_2 zinc finger protein family. C_2H_2 zinc fingers comprise the largest family of eukaryotic transcription factors, with several hundred C_2H_2 zinc finger proteins known in the human genome [40].

Zinc finger proteins[1] have been extensively studied, with crystal structures and experimental studies having explained many of the determinants of binding specificity (reviews, [11, 41]). Zinc finger proteins bind DNA in a well-characterized manner which specifies the exact interactions between specific residues in the DNA binding regions of the protein with nucleotides at the DNA site (see section 3.1.2 and figure 3.4 on page 44). Knowledge of this "canonical" structural interface can be used in predicting zinc finger specificity.

This chapter models zinc finger protein-DNA interactions by the pairwise residue-base interactions. A modified support vector machine (SVM) framework was used to find the favorability of each residue-base interaction. This framework includes not only examples of known zinc finger-DNA interactions but also quantitative information about the relative binding affinities between different protein-DNA configurations. Previous bioinformatics methods for predicting zinc finger protein-DNA interactions utilize only known examples of protein-DNA interactions (e.g., [12, 13, 42]); they are not able to use information about relative binding affinities. As high-throughput datasets with quantitative information about protein-DNA binding become more widely available (e.g., [7]), methods that can use such information will become increasingly important.

The SVM method developed in this chapter was tested using stringent per-protein cross-validation[2] and shown to be overall compatible with previous experimental data. Additionally, this method is shown to be competitive with previously published methods in a wide range of cross-validational testing. Overall, the SVM method holds great potential, especially as more quantitative information about binding is made available in high-throughput experiments.

[1] Though there are many types of zinc fingers, throughout this chapter, the term zinc fingers or zinc finger proteins to refer exclusively to C_2H_2 zinc finger proteins.

[2] Proteins with identical amino acids in DNA binding positions are considered the same for the purposes of creating a training dataset. See section 3.2.4 on page 48 for more details.

The rest of the chapter is organized as follows. Section 3.1 gives introductory structural information. First, DNA will be briefly described, concentrating on the way in which proteins are able to differentiate among bases. Then, the conserved sequence and structure of C_2H_2 zinc fingers will be shown. Finally, the pattern of binding between individual fingers and DNA sequence is outlined—this pattern is the central building block used to create the model of binding for the entire protein. Section 3.2 describes the framework used to train the model on collected zinc finger data, and gives implementation details. Section 3.3 describes the process of gathering examples of binding and non-binding configurations from the literature. References and brief descriptions of the experimental techniques used in these papers are provided, including a description of the quantitative way in which binding affinity is measured. Section 3.4 briefly outlines previous methods used to predict zinc finger binding. Section 3.5 describes extensive testing used to verify the integrity of the model, and includes the final derived weight vector. Finally, section 3.6 has some concluding remarks, while appendix 3.A describes binding affinity.

3.1 Background

This section provides some general structural background to DNA, protein-DNA binding, and zinc finger proteins.

3.1.1 DNA

DNA is composed of nucleic acids—each nucleic acid has a sugar, phosphate, and base. Sugars and phosphates are water soluble, and the structure of DNA attempts to hide the hydrophobic bases inside while exposing the hydrophilic sugars and phosphates outside [2].

Nucleic acids attach to one another forming a strand of DNA, while two complementary strands combine in opposite directions to form the famous double helix. The overall shape of the double helix can be thought of as a twisted ladder or a helical staircase with the bases acting as rungs or steps. Phosphates and sugars are located on the outside edges, holding

the bases in place. The double helix is constantly in motion, twisting, turning, winding and unwinding due to forces exerted on it from its changing environment (e.g., transcription factors, RNA polymerase, nucleosomes, and water) [2].

DNA can take on several different forms. The preferred confirmation in the cell is called the B-form, containing 10 bases per complete turn of DNA. The backbone of the B-form has two openings—the major and minor groove—through which proteins can bind with the bases. It is easier for amino acids to enter the larger major groove than the smaller minor groove. Each of the four nucleic bases presents a unique pattern of chemical groups (donors and acceptors) in the major groove of DNA, allowing proteins the ability to bind to specific sequences of DNA (called the base pair recognition code). The minor groove also presents an opportunity for proteins to enter, although in this case complementary base pairs present the same chemical groups making it impossible to distinguish between them chemically [43, DNA Structures].

Another form of DNA, the A-form, contains an extra twist per turn, making its major and minor grooves more similar to one another. There are other forms of DNA (e.g., the Z-form twists in the other direction as the A- and B-forms), although *in vivo* DNA is usually found between the A- and B-form [2].

A protein attaches itself to a sequence of DNA when it can make enough stable contacts with bases and backbone. Transcription factors bind DNA in a number of common structural conformations; common classes of transcription factors include leucine zippers, helix-turn-helix proteins, and zinc finger proteins [44]. Transcription factors are generally very specific, binding only to a sequence or range of sequences but not binding to the vast majority of sequences found within a genome. This level of specificity is required for meaningful regulation and many other biological molecules also show such a high level of specificity (e.g., hemoglobin combines only with oxygen, enzymes digest only certain foods, and antibodies attach only to specific antigens) [45]. The internal workings of the cell also include non-specific molecules, such as nucleosomes and histones.

3.1.2 C_2H_2 Zinc Finger Proteins

An authoritative description of the C_2H_2 binding domain may be found in [46]. More recent reviews can be found in [11, 41].[3] This section begins by describing sequence features of C_2H_2 zinc finger proteins, following by a description of their structure.

Sequence Features of the Zinc Finger Domain

Zinc finger domains are readily identifiable via sequence-based methods. While reported consensus sequences corresponding to the zinc finger domain vary [11,13,41,46–48], there is general agreement that zinc fingers share a sequence pattern of C-X_a-C-X_{12}-H-X_b-H, where X represents any residue and X_a and X_b is an arbitrary sequence of amino acids of length 2 to 5. Of the 7,005 zinc finger domains annotated via PROSITE [48], the vast majority of these (97%) match this consensus pattern, while 94% match the consensus pattern with parameters $a = 2$ and $b = 3$.

Figure 3.1 shows a sequence logo for the most frequent occurring zinc finger pattern. The height of each letter corresponds to the level of conservation for that amino acid, while the total height of each column of letters represents the overall conservation for that position. The observed residues agree well with published consensus sequences. There is a high level of conservation in the linker region (TGEKP), which connects adjacent finger domains [11,41]. The positions along the domain that are marked with stars make frequent contact with DNA bases (see next section). The stability of the domains allows a high level of variability in these positions, which in turn, allows different zinc finger proteins to bind a range of DNA, and allows the design of novel zinc finger proteins [41].

[3] Another source of information is the US Patent and Trademark Office, which (at the time of writing) listed 36 patents with the words 'zinc finger' in the title. The earliest patent was given in 1998 for designing zinc fingers for DNA binding.

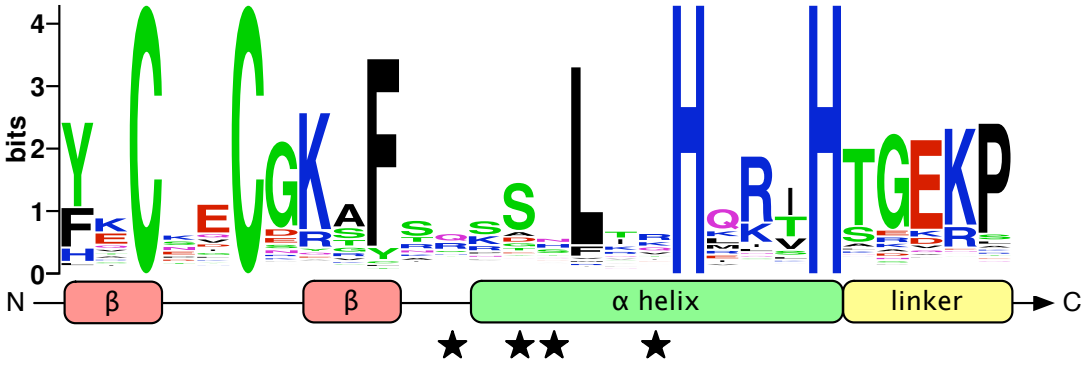


Figure 3.1: Sequence logo of zinc finger domains matching the pattern C-X_2-C-X_{12}-H-X_3-H. The height of each position corresponds to the level of conservation for that position, measured in bits (a perfectly conserved position corresponds to $\log_2(20)$ bits). Below the sequence logo is the secondary structure, including two beta strands and an alpha helix. A well conserved linker region, which connects tandem zinc fingers, is shown near the end. Positions predominantly responsible for recognizing DNA sequences in canonical zinc fingers are marked with stars, and are referred to as the -1, 2, 3 and 6 (numbered relative to the start of the alpha helix). Sequence logo was created using [49].

Organism	Genes with Domain		Domains
	Number	Percent	Per Gene
A. mellifera	92	1.5%	5.0
G. gallus	301	1.7%	4.9
D. melanogaster	347	2.1%	4.6
R. norvegicus	469	2.1%	6.9
C. familiaris	556	3.3%	8.5
P. troglodytes	629	2.9%	8.1
M. musculus	715	2.7%	7.9
B. taurus	759	2.0%	6.4
H. sapiens	858	3.1%	8.9
D. rerio	936	3.2%	10.1

Table 3.1: Number of genes in an organism that contain a C_2H_2 zinc finger domain, as judged by PROSITE [48]. Columns: model organism, number of genes with at least one C_2H_2 zinc finger binding domain and percentage of the organism's genes that this number represents, and average number of zinc fingers in genes with at least one. Organisms represented (from top to bottom): honeybee, chicken, fruit fly, rat, dog, chimpanzee, mouse, cattle, human, and zebra fish.

Table 3.1 shows the result of scanning for C_2H_2 zinc finger domains in available genomes [50] using a profile-based representation [48].[4] As seen in the table, zinc fingers are prevalent in multicellular organisms. The number of genes with at least one zinc finger motif varies, from less than a hundred in the western honeybee to over 800 in human, and more than 900 in zebra fish. Roughly 1 to 3% of the genes in the genomes for these organisms code for proteins containing at least a single zinc finger. Of the organisms shown, dog has the highest concentration of zinc finger containing genes, while the honeybee has the lowest concentration. The number of finger motifs per gene varies from a low of five in fruit fly to over ten in zebra fish.

[4] [48] reports 99.9% and 98.94% precision and recall rates when scanning SwissProt.

Structure of Zinc Finger Proteins

The C_2H_2 zinc finger binding domain has a strongly conserved secondary structure, consisting of two beta strands followed by an alpha helix [43, Structural Motifs of Eucaryotic TFs]. All three secondary structures are connected to one another through an intermediate zinc ion. Specifically, two cysteines (located in the beta strands) and two histidines (located in the alpha helix) bind with zinc. These cross-links between distant parts of the domain make zinc fingers highly stable, resulting in a smaller, more compact structure than would be possible without such cross-links [47].

The structure of Zif268 [51], a mouse transcription factor (also known as *Egr-1* and Krox-24) with three zinc finger domains has served as a model system for studying the specificity of zinc finger protein-DNA interactions. Figure 3.2 shows a rendering of crystal structure `1AAY` [52], which is wild type Zif268 binding to its native binding site. The protein is seen to wrap around the DNA, with the alpha helix in each finger fitting into the major groove of DNA. The three binding domains are spaced so that each finger binds different bases.

Studies of C_2H_2 zinc finger proteins (e.g., [52, 54]) have found that a large portion of zinc fingers, called canonical zinc fingers [41], bind DNA in a manner similar to Zif268. This pattern of side chain-base interactions, which is referred to as the canonical binding model, is shown schematically in figure 3.3. The amino acid sequence proceeds from right to left while the mainly contacted DNA strand is shown 5' to 3'. Only the alpha helix is situated close enough to the DNA to make contact with bases, and four amino acid side chains in or near it are responsible for the majority of contacts. These positions are 6, 3, -1 and 2, numbered relative to the start of the alpha helix. Canonical C_2H_2 zinc fingers provide a conserved, modular domain which serves as an essential starting point for predicting the protein's DNA binding.

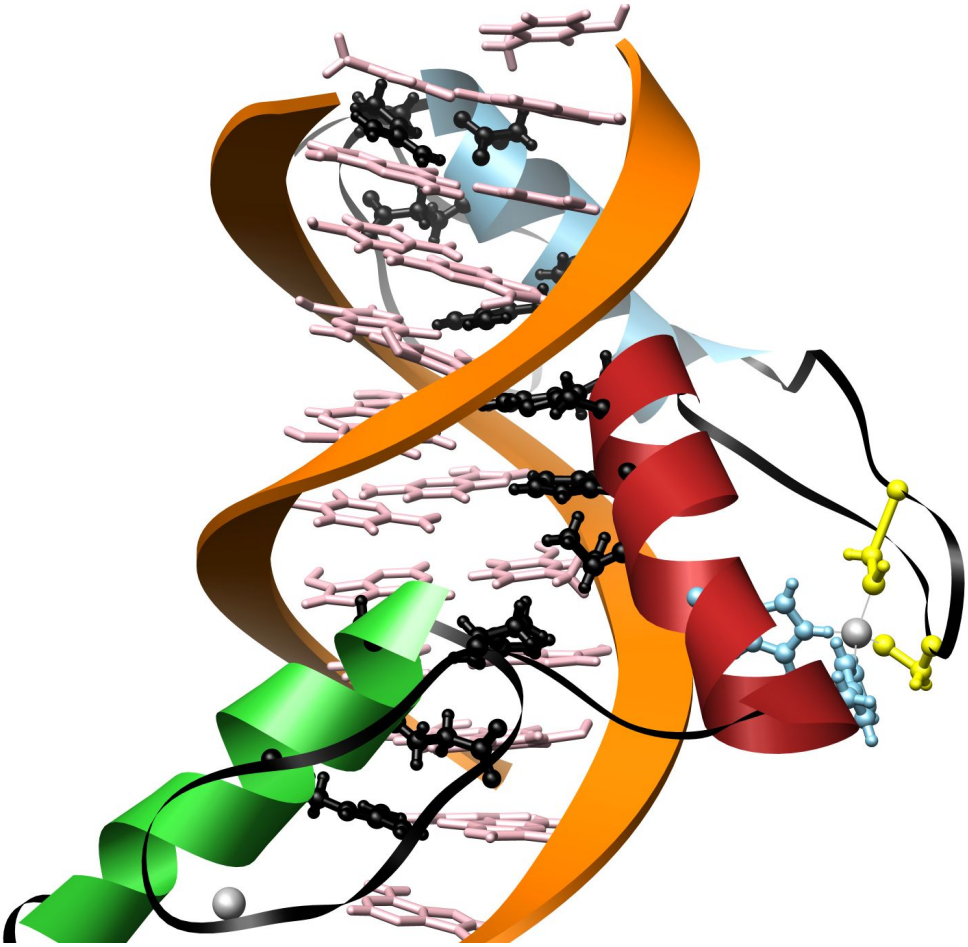


Figure 3.2: Wild type Zif268 protein binding to its native binding site as captured in the PDB crystal structure `1AAY` [52]. Residues in the alpha helix of each finger make contact through the major groove of DNA. Alpha helical regions are depicted as shaded ribbons. Contacting amino acids are shown in black. Linker regions and beta strands are shown as a black string while zinc atoms are shown as gray spheres. The four centrally coordinated cysteine (yellow) and histidine amino acids (blue) are shown for the second zinc finger (red ribbon; far right). Crystal structure was rendered using Chimera [53].

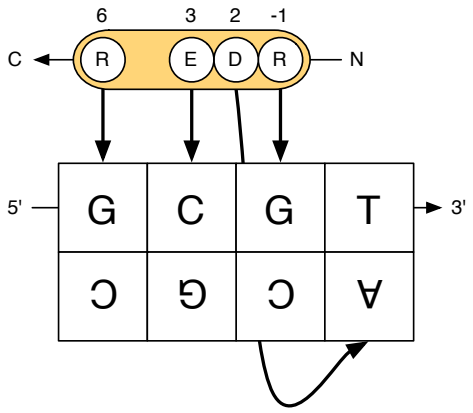


Figure 3.3: Canonical binding pattern frequently seen in C_2H_2 zinc fingers [41]. Relative to the start of the alpha helix, positions 6, 3, and -1 contact three consecutive bases on the main strand, whereas position 2 contacts the following base, but on the complementary strand. Zinc fingers exhibiting this interaction pattern are referred to as canonical fingers. Amino acids and bases shown correspond to fingers 1 and 3 of wildtype Zif268 binding to its natural site.

3.1.3 Analysis of Zinc-Finger Structural Interface

With the availability of many crystal structures, it is possible to estimate how well the canonical binding model approximates observed bonding in solved structures. Nineteen crystal structures were gathered from the PDB: `1A1F`, `1A1G`, `1A1H`, `1A1I`, `1A1J`, `1A1K`, `1A1L`, `1AAY`, `1F2I`, `1G2D`, `1G2F`, `1JK1`, `1JK2`, `1P47`, `1TF6`, `1UBD`, `1ZAA`, `2DRP`, and `2GLI`. For each structure, Chimera was used to remove solvent atoms, add hydrogen atoms, and calculate potential hydrogen bonds between proteins and DNA [55].[5] A total of 197 distinct bonded amino-acid base pairs were identified. Considering only the amino acid involved, table 3.2 shows which amino acid positions (with respect to the start of the alpha helix) are responsible for the binding between protein and base. Positions -1, 2, 3, and 6 (canonical amino acid positions) account for the vast majority of found contacts (98%), supporting the use of only these positions in predicting future interactions.

Next, the binding pattern of amino acids for an individual finger were considered. Chemical bonds were grouped by their zinc finger, and binding with DNA was compared

[5]Similar results were obtained using HBPLUS [56].

Position	Base Contacts		Backbone Contacts	
	Number	Percent	Number	Percent
-1	62	31%	7	6%
1			2	2%
2	35	18%	2	2%
3	45	23%	1	1%
5	1	1%	11	10%
6	51	26%		
7			51	46%
9			4	4%
10	2	1%		
Other	1	1%	33	30%
Total	197	100%	111	100%

Table 3.2: Hydrogen bonds found in nineteen crystal structures containing C_2H_2 zinc finger domains. Position refers to the position of the amino acid side chain contacting DNA. Positions -1, 2, 3, and 6 correspond to canonical position and include 98% of all base contacts.

against the canonical binding model (several fingers could not be registered successfully and were removed from the sample). Table 3.3 shows which contacts are most prevalent in the remaining 76 zinc fingers. Contacts following the canonical binding pattern account for 98% of all contacts found, although the average finger was found to contain only 2.5 out of the 4 contacts specified in the model.

Position	Contacts	
	Number	Percent
6	46	61%
3	41	54%
-1	58	76%
2	35	46%
Other	7	9%

Table 3.3: Contacts following the canonical binding pattern account for 96% (180 out of 187) of all contacts found in 76 fingers from nineteen crystal structures. The canonical binding pattern states that positions 6, 3, -1 contact three consecutive bases in one strand whereas position 2 contacts the next base, but on the complementary strand.

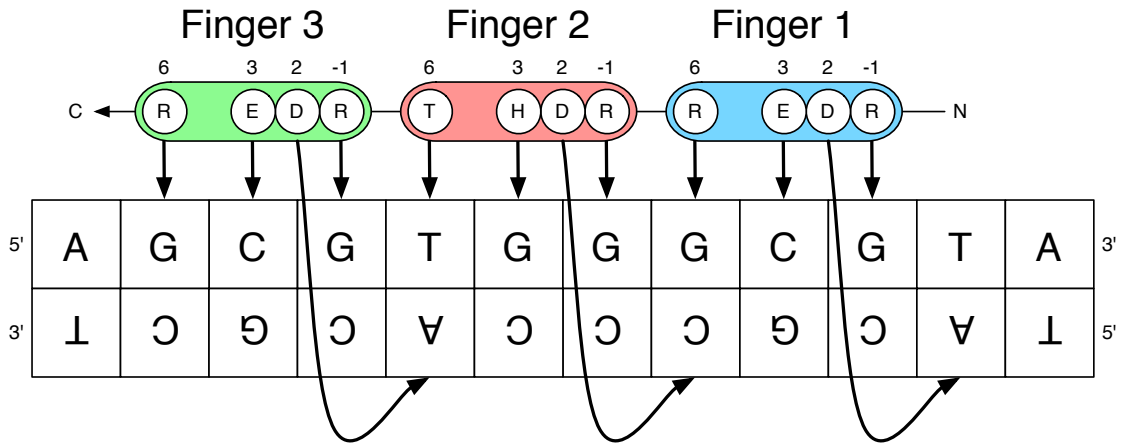


Figure 3.4: Canonical binding model applied to a three fingered proteins, such as *Egr-1* (shown here). Actual crystal structures do not exhibit all the depicted contacts, although a majority of contacts were found in nineteen crystal structures in the PDB.

Of the nineteen crystal structures considered, seventeen exhibited a simple binding pattern, where consecutive fingers bind consecutive patches of DNA, with a single base-pair overlap. Figure 3.4 shows this type of binding for a three fingered protein, which constitutes most of the examples gathered from the literature (as described in section 3.3). Exceptions to this binding pattern were found in `1tf6` and `2gli`. `1tf6` (TFIIIA) is a 6-fingered protein: fingers 1-3 bind in a manner similar to Zif268, fingers 4 and 6 bind through the minor groove of the DNA, and finger 5 binds again in the major groove [57]. The spacing between binding sites is not regular. `2gli` is a 5-fingered protein: finger 1 does not bind to DNA, fingers 2-4 bind in a manner similar to Zif268, and a gap lies before the bases contacted by finger 5 [58].

3.2 Methods

3.2.1 Representing Zinc Fingers

Zinc fingers will be assumed to bind according to the canonical binding model (as shown in figure 3.4 on the previous page), which will be the basis for representing zinc finger protein-DNA interactions. The canonical binding model is an approximation often made when modeling these type of interactions [12,13,42]. A protein-DNA configuration (either binding or non-binding) is represented mathematically by a feature vector x, where the coordinate $x_{p,a,b}$ is the number of times amino acid a is in position p in one of the fingers of the protein and base b would be the contact in the canonical binding model.

The goal is to find a weight vector w that represents the "favorability" of each possible amino-acid base-pair pairings in each of the four canonical positions.

3.2.2 Standard Support Vector Machines

Given a dataset of binding and non-binding examples, support vector machines (SVMs) are one means for learning a way to classify the two [59]. SVMs try to find a weight vector w that best separates binding and non-binding examples, as follows:

$$\begin{aligned} &\text{minimize } \|w\|^2 \\ \text{subject to } &\begin{cases} w \cdot x_i + b \geq 1 & \text{for binding examples} \\ w \cdot x_i + b \leq -1 & \text{for non-binding examples} \end{cases} \end{aligned} \tag{3.1}$$

This optimization searches for the feature vector of minimum length, which can be interpreted as the 'least complicated' weight vector classifying all examples correctly. Model 3.1 can be used when the dataset is separable (when a weight vector exists that is consistent with all observed examples). When this is not the case, the following formulation is used.

$$\text{minimize } \frac{1}{2}\|w\|^2 + C \sum\nolimits_i \epsilon_i$$

$$\text{subject to } \begin{cases} w \cdot x_i + b \geq 1 - \epsilon_i & \text{for binding examples} \\ w \cdot x_i + b \leq -1 + \epsilon_i & \text{for non-binding examples} \end{cases} \tag{3.2}$$

$$\text{where } \epsilon_i \geq 0$$

The optimization finds a compromise between a 'least complicated' weight vector and fitting the training data. C is called the cost factor, and is the tradeoff between these two alternatives.

The trained weight vector can be used to make predictions for unknown configuration by calculating $p = w \cdot x + b$ for the feature vector x corresponding to a novel configuration of protein and DNA. A more positive score predicts stronger binding.

3.2.3 Modified SVM

The SVM model shown in model 3.2 has been used successfully to solve a wide range of machine learning problems [59, 60]. However, it poses two problems when applied to the current situation. First, in many cases, quantitative information about the binding affinity of zinc finger-DNA pairs is known. Affinity information is lost in the binary positive and negative classification, and could be very valuable. Second, experimental protocols vary in the sources used for gathering experimental data. Specifically, the sources do not agree on what is considered 'non-binding' (for example, 'non-binding' in one experiment may be considered 'weakly-binding' in another). Therefore, it is unclear how to extract clean and comparable negative examples from combined sources.

Both concerns are addressed by substituting 'comparative examples' in place of 'non-binding' examples in the optimization. As the name suggests, comparative examples capture the binding preference between two configurations. Suppose x and y are feature corresponding to two protein-DNA configurations and it is known that configuration x binds more strongly than configuration y. In terms of the weight vector w, this corresponds

to $w \cdot x + b > w \cdot y + b$, so that $z = y - x$ can be added as a negative example, capturing the desired relation. The modified SVM used during training is shown in model 3.3.

$$\begin{gathered} \text{minimize } \frac{1}{2}\|w\|^2 + C \sum\nolimits_i \epsilon_i \\ \text{subject to } \begin{cases} w \cdot x_i + b \geq 1 - \epsilon_i & \text{for binding examples} \\ w \cdot z_i \leq -1 + \epsilon_i & \text{for comparative examples} \end{cases} \\ \text{where } \epsilon_i \geq 0, \text{ and} \\ z_i = y_i - x_i \text{ when } x_i \text{ binds more strongly than } y_i \end{gathered} \tag{3.3}$$

Including comparative examples ($w.x > w.y$) require a linear classifier, which precludes the use of non-linear kernels in the optimization. Kernel functions were tried using a standard dataset of positive and negative examples. However, the resulting classifier did not perform as well as the simpler linear model, possibly due to the difficulty in trying to establish comparable positive and negative examples from various sources.

Implementation

SVM-light [61] version 6.01 was used to solve model 3.3, and found an optimal weight vector in a time-efficient manner even with over a hundred thousand examples used during training. Alternative solvers were also tried [62, 63], giving similar results.

A cost factor of $C = 50$ was used, based on trying several different C values for cross-validation testing (section 3.5.4).

A heuristic that improved performance during testing was to scale error terms for positive examples so that they are not dominated by other examples [36]. Specifically, ϵ_i was scaled by O/P for error terms associated with positive examples, where P is the number of positive examples and O be the number of other examples. A second heuristic was to eliminate the constant term b, thus reducing the number of variables in the model.

3.2.4 Cross-Validation

It is tempting to calculate the number of misclassified comparative examples using the learned SVM model. However, this error rate (called the resubstitution error rate) does not give an accurate indication of the anticipated error for novel examples, as the classifier may *overfit* on the learned data [36]. With a potential of 320 variables in the final model, overfitting is a concern in the current scenario.

There are several standard methods to compensate for overfitting or similar machine learning difficulties. When the amount of available data is large, a viable option is to divide the dataset into two parts [36]. Training is done on one part, while the other is used for the final evaluation. If there are several stages in training, then the data can be divided into several portions, always reserving one for final testing. In this case, however, there is a limited supply of data so a more conservative approach is needed.

Cross-validation, used throughout this chapter, removes from training the portion of the dataset used during testing. Because examples are often related to one another (e.g., comparative example $y - x$ is related to examples x and y), training examples are removed on a per-protein basis. As a specific example, human SP1 contains three binding zinc fingers with contacting amino acids `KSHA`, `RDER`, and `RDHK` (positions -1, 2, 3, 6). When testing on human SP1, any examples with the same contacting amino-acids is removed from training. Examples with the same contacting amino acids, but in an alternate order, are also removed from training. Comparative examples are treated conservatively, removing them if either x or y match a testing example. Using per-protein cross-validation, performance tests are less likely to be influenced by overfitting or chance.

3.3 Gathering Experimental Data

3.3.1 Sources of Experimental Data

An extensive literature search was performed to gather examples of binding and non-binding configurations (as stated by the authors) of C_2H_2 zinc fingers and DNA. Some of the experiments quantified the binding affinity binding between protein and DNA, usually in the form of an association or dissociation value (higher association and lower dissociation values correspond to stronger binding). Examples with binding affinity information will be referred to as quantitative examples. Binding affinity, association and dissociation values are described in more detail in appendix 3.A. The following sources were used to gather experimental examples, most recent source first.

[64] changed eight three-zinc-finger proteins into repressors by fusing their DNA binding domain (DBD) with a known repressor domain. The proteins were designed to bind upstream of HIV-1 genes. One was found to inhibit HIV-1 replication by 75%, demonstrating the potential of designing novel zinc finger proteins for antiviral therapy.

[65] combined a single zinc finger domain from the human genome with two domains in Zif268, creating a novel three fingered DBD. These DBDs were later fused with activator or repressor domains, creating novel activator and repressor transcription factors.

[66] designed several three-fingered and six-fingered proteins to bind upstream of a human skin-specific gene, allowing the expression of that gene in non-skin cells.

[7] analyzed the binding affinity of wild-type Zif268 and four variants against a variety of binding sites using a high-throughput microarray based approach. Briefly, a microarray glass slide was filled with different DNA strands, corresponding to different binding sites. This array was exposed to a solution containing a known concentration of a zinc finger protein and a non-specific control. The protein and the control were tagged using fluorescent dyes of differing color. The solution was rinsed, retaining only bound proteins. The intensities of colors seen on the microarray were measured, indicating the binding

affinity of the tested protein against all binding sites on the array. This high-throughput approach allowed the measurement of the binding affinity of 64 binding sites for five proteins, although some combinations of protein-DNA were below experimental detection.

[67] constructed a library of zinc fingers that are capable of binding to 3-base subsites beginning with adenine. This library was based on a protein that binds more strongly to its binding site than wild-type Zif268 (the majority of protein variants bound less strongly than the original).

[68] constructed zinc finger proteins for a nine base-pair site by first creating two three-fingered proteins, each designed to bind to overlapping five base pair sequences. The three-fingered proteins were created in such a way so that it is possible to combine them into a single protein which will bind to the desired site. The resulting proteins were tested to bind upstream of a HIV-1 promoter.

[69, 70] constructed and analyzed zinc fingers capable of binding to 3-base subsites beginning with guanine.

[71, 72] used a unique selection protocol to create novel three-fingered zinc finger proteins. Briefly, the desired binding site was joined to a known two-finger site. The known site was used as an 'anchor' to optimize a single finger extending into the desired site. Once one finger was known, binding domains where shifted and the process was repeated, each time adding more of the desired site.

[73] investigated the interactions of the overlapping base between adjacent zinc finger binding domains. [74] changed individual fingers of Zif268 one at a time, giving information about the modularity of the C_2H_2 binding domains.

[75] used phage display to investigate zinc finger DNA binding. Briefly, randomized zinc finger genes are injected into bacterial viruses (phages) which express the proteins on their surface. A solution containing a pool of phages with different expressed zinc fingers is mixed with immobilized strands of DNA. Bacteria with binding zinc fingers attach themselves to DNA and are not washed away during rinsing. Examples of proteins

that survived several rounds of selection were given. [76] investigated zinc finger binding using a SELEX protocol. Unlike phage display, during SELEX the protein is fixed while a pool of randomized DNA segments compete to bind with it. Binding and non-binding examples often differed in only a single base, demonstrating the specificity of zinc finger proteins.

[77, 78] mutated the first finger of Zif268. [79] investigated the binding of three zinc-finger proteins. Identical and non-identical binding domains were considered, and the change in binding affinity was measured after exchanging binding domains. [80] changed key positions in the second finger and reported dissociation values for several binding sites. [81] reported the relative binding affinity of ten zinc finger proteins (Zif268 and nine mutations) on a total of fourteen binding sites each.

Definite information about zinc finger interactions was extracted from crystal structures found in the PDB [82]. Structures that follow the canonical binding model were taken into consideration: `1A1F`, `1A1G`, `1A1H`, `1A1I`, `1A1J`, `1A1K`, `1A1L`, `1AAY`, `2DRP`, `1F2I`, `1G2D`, `1G2F`, `1JK1`, `1JK2`, `1ZAA`, `1P47`.

Binding sites for variants of *Egr-1* were taken for a previously published study of zinc finger binding [12, 83]. Over a thousand positive examples of binding zinc finger proteins were included from fifty-four journals. Some of this data overlaps data gathered above. The vast majority of this data were for proteins such as *Egr-1*, although some examples were for two fingered proteins. No direct information about the binding affinity for protein-DNA complexes is given. Many of the experiments were the result of selection experiments where key amino acids or key nucleotides were randomized and the given protein-DNA were among the highest binding combinations found.

Finally, binding sites detailed by [13] were added to the dataset.

Type of Example	Number of Examples
Binding	494
Quantitative	350
Qualitative	144
Non-binding	310
Quantitative	300
Qualitative	10
Comparative	1,302

Table 3.4: Number of examples in the gathered dataset (before processing), not including data from [12, 83]. Some examples contained quantitative information about binding affinity (either an association or dissociation value), while qualitative examples were stated as either binding or non-binding. Comparative examples give information about the relative binding affinity of two configurations.

3.3.2 Data Processing

Data for [13, 64–81] was entered manually; data for [7] was downloaded from the supplementary information web site; crystal structures were taken from the PDB [82]; data for [12,83] was taken from a public file. In order to lessen the chance of error, a file format was created which resembles the way in which experimental results are presented in the literature. As an example, figure 3.5 shows the input for one source. A small parser was written to convert such datafiles into a format suitable for further analysis. Table 3.4 shows the number of examples of the initial dataset, before any processing.

Experiments differ on what is considered 'binding' and 'non-binding'. It is not clear how to compare association and dissociation values for sources with differing experimental protocols. Thus, it is advantageous to convert the original dataset into a form that is independent of such differences. While the original dataset contains positive, negative, and comparative examples, the final dataset contains only positive and comparative examples. Although sources may disagree on the what is considered binding and non-binding, the relation between weak and strong binding should be more easily conserved.

```
/**
@articleICK97,
 author ="M. Isalan and Y. Choo and A. Klug",
 title  ="Synergy between adjacent zinc fingers in sequence-specific DNA recognition",
 journal=PNAS,
 year   =1997, volume=94, number=11, pages="5617--5621", month=may
**/
source=ICK97

dna=tatatagcg___gcgtatata
#   xxx===---...---===xxx

# amino acid positions: -1 1 2 3 4 5 6 7 8 9 10
zf=3
f1=RSDELTRHIRI
f2=__________T
f3=RS_ERKRHTKI

# figure 4
ex=Kd
{
  f2=RSDHLTTHIR dna=TGG               # Zif
  { f3=D Kd=2.8  KdSd=0.6; } # wt
  { f3=A Kd=10.0 KdSd=3.3; } # mut
//
  f2=REDVLIRHGK dna=GTG               # F2-Arg
  { f3=D Kd=1.3 KdSd=0.1; } # wt
  { f3=A Kd=5.6 KdSd=1.3; } # mut
}
```

Figure 3.5: Example of file in the gathered dataset [73]. Essential features of the file format include: arbitrary tags declared using `tag=value`, semicolons end examples, curly brackets designate scope (in the sense of computer languages), and underscores are considered placeholders. In this data format, placeholders must be filled-in before an example is printed and tags can only be filled-in, never redefined. These safety mechanisms help catch errors. For example, the `dna=` declaration near the beginning ensures that all examples from this source will have a DNA sequence of equal length. The parser that processes this file format also checks whether every example that could have been printed was printed (for example, in case of a missing semicolon).

Type of Example	Number of Examples
Binding	937
Comparative	19,353

Table 3.5: Number of examples in the final dataset (after processing). The large number of comparative examples were derived from positive and negative examples in the gathered dataset (table 3.4 on page 52). The number of binding examples has changed, as data from [12, 83] was added, and duplicate examples in the combined dataset were removed.

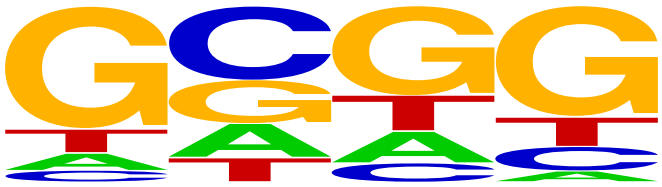
The dataset used for training was created as follows. Positive and comparative examples are passed to the final dataset unaltered. Additional comparative examples are created considering each source individually. First, every pair of quantitative examples which association or dissociation values differ by a ratio of two or more are used to create a comparative example. Second, all positive examples are paired with negative examples, representing negative examples indirectly in the form of many comparative examples. Table 3.5 shows the number of examples in the final dataset, while figure 3.6 shows the frequency of amino acids and binding subsites for individual fingers.

3.3.3 Alternative Sources of Experimental Data

There are several general databases of transcription factors and their binding sites, including Transfac [84] and JASPAR [85]. Neither is ideal for a structural-based approach as the binding pattern is not given explicitly. However, there are computational methods for predicting the most likely binding site within a larger sequence. For example, [13] used an iterative approach (expectation maximization) to learn both binding preferences and the actual binding sites within sequences given in Transfac.

However, there are advantages for using the constructed dataset instead of the databases described above. In particular, most of the experiments provide quantitative information about binding affinity, and this is used to construct a large number of comparative examples. Given high enough concentration, even weakly binding proteins will bind at some

(a)



(b)

Figure 3.6: Frequency of amino acids and bases for individual fingers and binding subsites in the final dataset. (a) shows the frequency of amino acids in positions -1 through 6, while (b) shows the frequency of binding subsites contacted by an individual C_2H_2 zinc finger. DNA is shown 5' to 3' on the primary strand. Only binding examples are included in the figure, and only zinc fingers and bases contacting or within varied subsites (many experiments varied bases contacted by a single zinc finger, keeping the remaining binding site fixed).

level, so a measurement of binding affinity may be helpful in building a computational model. Moreover, as mentioned above, in the sources gathered for the dataset experiments were designed in such a way that the alignment of protein to DNA is explicitly known.

3.4 Previous Methods

3.4.1 Sequence Based Methods

The constructed SVM model was compared against several alternate sequence based methods, briefly described below.

Kaplan *et al.* [13] A probabilistic method based on the binding sites in the Transfac database. Transfac does not provide exact sites, but provides a longer sequence in which the binding site resides, so in order to use this database expectation maximization was used to learn both the probabilities associated with different amino acids and bases and the locations of binding sites in the database. Potential binding sites were scored using a log-odds score, assuming a uniform background probability.

SAMIE [12] A probabilistic computational model of Zif268 binding trained on SELEX and phage display experimental data gathered from the literature. The model was fit in order to maximize the *specificity* of the binding zinc finger.

Mandel-Gutfreund *et al.* [42] A computational method based on the hydrogen bonding patterns extracted from crystal structures from crystal structures of various proteins in the PDB and NDB. The weights used in this method represent more general trends found in protein-DNA interactions, although testing was done on C_2H_2 zinc finger proteins.

Suzuki *et al.* [86] A method based on expert knowledge of biochemical principles. Two distinct components of protein-DNA binding are considered: chemical and stereochemical. Chemical rules are general and are based on the inherent chemical com-

patibility of amino acids and bases; a table of numerical compatibilities is given. Stereochemical rules are specific to individual classes of transcription factors and correspond to amino acid-base contacts. As in the previous method, the weights given in this method represent more general principles. C_2H_2 zinc fingers were among several transcription factors considered.

3.4.2 Physics-Based Methods

Some methods predict protein-DNA interactions based on detailed crystal structures. These physics-based methods (e.g., [87,88]) extract the binding pattern between the DNA and protein (or a close homolog), and use detailed structural information and energy functions to evaluate protein-DNA configurations. SVM was not compared against these two, because of computational concerns and because detailed structures are not available for use for all test proteins.

3.5 Results

3.5.1 Evaluating Adding Comparative Examples

A heuristic which improved overall performance is to expand comparative examples to include shifted binding sites. Specifically, if a protein-DNA configuration x binds more strongly than protein-DNA configuration y, then it also binds more strongly than configuration y', where y' has the DNA strand shifted left or right, or reverse complemented.

First, the classifier was trained on comparative examples originally in the dataset. Then, reverse-complemented examples were added. Then, the DNA strand of the less-strongly-binding examples were shifted one, two, or three bases from the original position. Table 3.6 shows the number of comparative examples in each dataset. Shifted datasets were compared using binding data for Human SP1, described next.

Dataset	Comparative Examples
Original examples	19,353
Reverse-complemented	38,674
Shifted-by-one base	114,799
Shifted-by-two bases	191,942
Shifted-by-three bases	265,046

Table 3.6: Number of comparative examples after adding shifted examples.

Human SP1

[89] have made available binding sites for transcription factors along human chromosomes 21 and 22. Of the proteins tested, SP1 is a zinc finger protein containing three fingers binding in tandem as in the canonical binding model described in this chapter. Binding preferences are given in the form of a p-value comparing the binding of SP1 versus two controls. From this, binding and non-binding regions were extracted, merging overlapping regions and removing overlapping regions between binding and non-binding regions. A p-value of 10^{-5} or less was used for positive binding sites and 99% or more for negative sites, resulting in 1,992 binding regions (positives) and 5,646 non-binding regions (negatives).

This dataset gives valuable information for a real protein and gives *in vivo* information about binding on a very large sequence. Therefore, it represents a natural testing scenario that is what a zinc finger classifier may be asked to perform in practice. The majority of data is also not found in the training dataset, so it represents an external way of testing the built classifier and was used to optimize model parameters and decide which examples should be included during training.

Figure 3.7 shows the results of classifying Human SP1 binding sites using different comparative examples. For each curve, several cost factors were tried and only the optimal performing value is shown. ROC curves were generated using leave-protein-out cross-validation, so that Human SP1 protein was not used during training. The classification power of the resulting classifier increases with the addition of shifted sites, as seen on

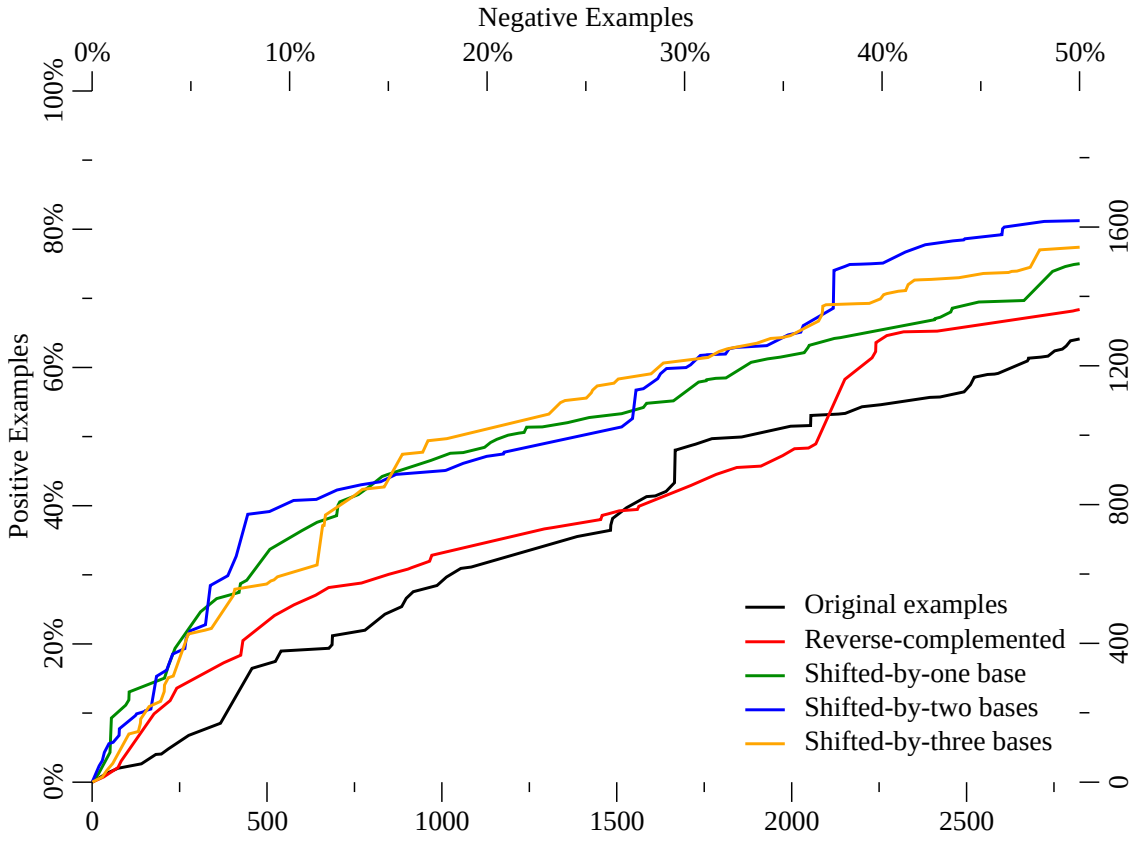


Figure 3.7: Improvement classifying Human SP1 binding sites by adding shifted examples. The first curve uses only comparative examples explicitly in the dataset. The following curve adds reverse-complemented examples. Then, the DNA strand of the less-strongly-binding examples are shifted by one, two, or three bases from the original position. ROC curves were generated using leave-protein-out cross-validation.

the ROC curve. Quantitatively, the AUC[6] is 17% (original dataset), 22%, 32%, 33%, and 31% (shifted-by-three bases) as more comparative examples are added. As a result, all figures and calculations remaining in this chapter have been done using the expanded shifted-by-two dataset.

3.5.2 Evaluating Data Sources

The learning model introduced in section 3.2 depends on the examples gathered in section 3.3. Therefore, it is important that the data used during learning is as accurate and as complete as possible. Here, this question is addressed by using cross-validation testing to test whether including each source of data improves upon the overall results. Specifically, examples from source (a) were treated as testing examples and a training dataset was created using leave-protein-out cross-validation. For each remaining source (b), the classifier was trained and evaluated using the full training dataset and again without any examples from (b). The percentage of comparative examples classified correctly from source (a) was used as an indication of performance, and a matched-pairs signed-rank test was used to test whether adding (b) improves overall performance. Table 3.7 shows the results of the signed rank test.

All p-values fall within expected values, meaning that all sources either improve performance during cross-validation or do not hurt performance enough to be deemed statistically significant. This supports using all of the data sources during training.

3.5.3 Predicting Binding Affinity

[7] analyzed the binding affinity of wild-type Zif268 and four variants against 64 binding sites using a high-throughput microarray based approach. Binding affinity measurements were given for a total of 124 protein-DNA combinations.

[6] Area underneath the curve, calculated by integrated the roc curve from 0 to 20% false-positive rate (x-axis).

Source	Better	Worse	Same	Z-score	P-value
[66]	14	1	6	2.96	0%
[69]	12	1	8	2.75	1%
[70]	13	4	4	2.41	2%
[74]	9	4	8	1.60	11%
[64]	8	4	9	1.44	15%
[73]	7	3	11	1.40	16%
[68]	7	3	11	1.30	19%
[77]	8	4	9	1.22	22%
[79]	9	4	8	1.21	23%
[82]	3	1	17	1.05	29%
[65]	9	7	5	0.93	35%
[75]	4	2	15	0.86	39%
[67]	6	5	10	0.50	62%
[13]	3	2	16	0.49	62%
[76]	7	6	8	0.45	65%
[7]	8	6	7	0.37	71%
[78]	6	6	9	0.26	80%
[72]	2	2	17	0.10	92%
[80]	4	4	13	-0.08	94%
[12, 83]	6	6	9	-0.16	87%
[71]	5	7	9	-0.44	66%
[81]	6	6	9	-0.46	65%

Table 3.7: Testing whether including a source improves overall performance during cross-validation. Sources are listed on the left. The following three columns show the number of times including a source helped/worsened/or did not effect performance when calculating the percentage of comparative examples classified correctly during cross-validation on the remaining sources. The final two columns give the z-score and p-value associated with a matched-pair signed-rank test [90], accessing the statistical significance of the previous three columns.

Zif268 Variant	#	SVM	Kaplan *et al.*	SAMIE	Mandel-Gut-freund *et al.*	Suzuki *et al.*
RSDH	15	.51	.30	.38	.40	.39
RGPD	17	.47	.50	.47	.37	.43
REDV	15	.51	.35	.48	.23	.41
LRHN	13	.50	.30	.46	.49	.44

Table 3.8: Correlation coefficients between predicted binding site scores and experimental binding affinities for the built SVM classifier and previously published methods. First two columns list the Zif268 protein and number of observations used to calculate the correlation coefficient. RSDH is the wildtype Zif268 protein. The observations used correspond to those with binding affinity above experimental detection.

Table 3.8 shows the correlation coefficient between predicted scores and actual binding affinities for four variants (the non-specific KASN protein was omitted from consideration; each method performed poorly trying to predicting its binding preference). SVM performance was measured using cross-validation, leaving out the target protein (and any example made using this protein, or any other protein having the same DNA binding amino acids) from the dataset during training.

SVM had the highest correlation coefficient for RSDH (wildtype Zif268), REDV and LRHN (although very close with the second ranking method). It had second best performance for RGPD (tying with another method). Overall, SVM scored very well correlating with published binding affinities, showing that a strong model can be built using a combination of support-vector-machines, comparative examples, and high-throughput quantitative data.

3.5.4 Human SP1

As described in section 3.5.1, [89] contains binding data for a single human transcription factor, SP1, along two human chromosomes. Positive (binding) and negative (non-binding) regions were extracted from this dataset, giving information about *in vivo* binding of a naturally occurring protein.

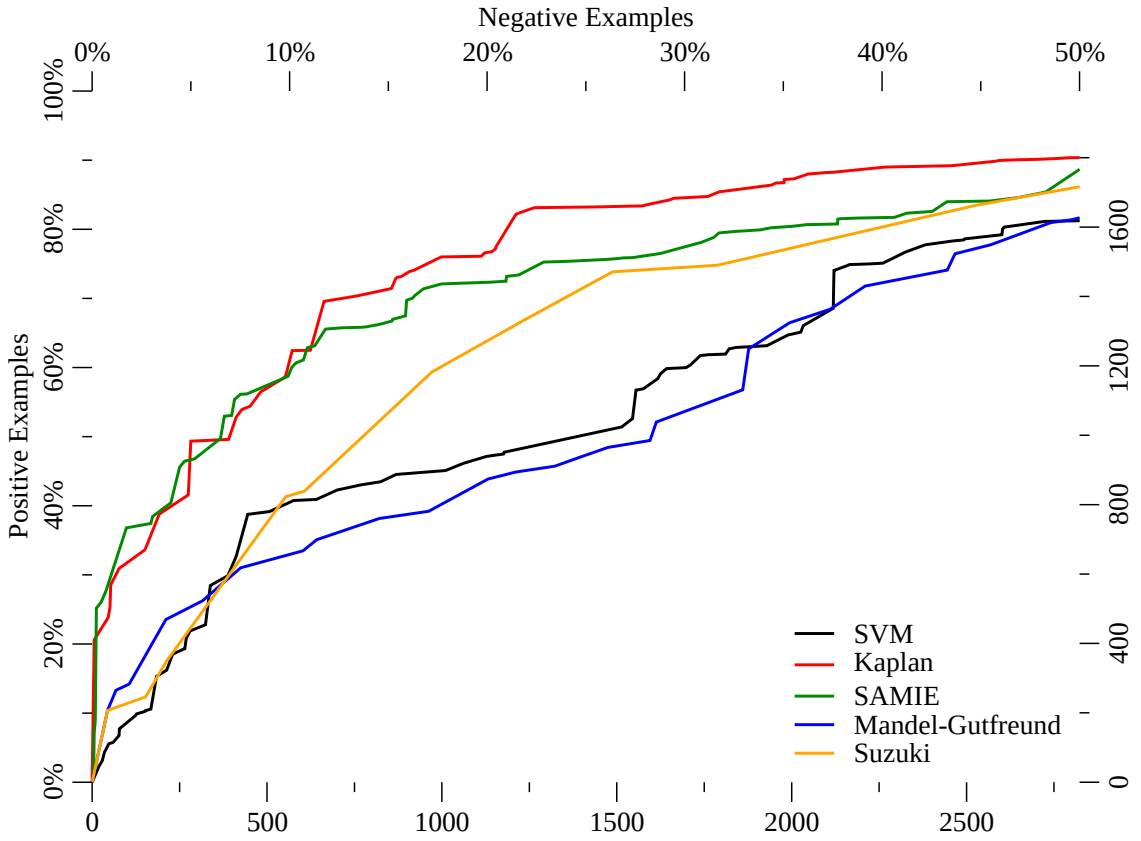


Figure 3.8: Comparison of SVM and four previously published methods in finding Human SP1 binding sites. ROC curve for SVM was generated using leave-protein-out cross-validation. The AUC (area underneath the curve, integrated from 0 to 20% false-positive rate) for each method is: 57% (Kaplan), 56% (SAMIE), 38% (Suzuki), 33% (SVM), 31% (Mandel-Gutfreund). See section 3.4 on page 56 for details on previous methods.

Figure 3.8 compares the SVM built classifier with four previously published methods. ROC curves for the SVM were generated using cross-validation, while other methods were taken from their respective sources. As can be seen, when predicting the binding sites of human SP1, SVM classifies comparably with [42], yet is outmatched by more recent methods. It should be stressed, however, that other methods used different datasets for training. For example, Kaplan *et al.* [13] extracted data from Transfac [84] using an expectation-maximization approach, and the authors themselves note they have a large number of SP1 targets in their dataset. In general, SP1 binds G-rich binding sites, and it

ZF	SVM	Kaplan *et al.*	SAMIE	Mandel-Gut-freund *et al.*	Suzuki *et al.*
A	21	5	1	11	1
B	5	1	65	39	46
D	58	22	2	11	28
E	110	201	59	43	345
F	138	306	520	2,040	1,016
G	606	65	1,385	92	140
Avg.	157	100	339	373	263

Table 3.9: Rankings of target sites within the HIV-1 genome. For each protein, the method with the lowest ranking site is shown underlined (lower numbers in the table are better). Protein C was omitted from the table because it was predicted as one of the weakest binding sites by all computational methods.

is possible that the current dataset, comprised of primarily of mutational data binding a large variety of sites, does not adequately reflect this bias.

3.5.5 HIV-1

[68] listed several proteins designed to bind to sequences of DNA found within the Human Immunodeficiency Virus type 1 genome. Seven proteins were engineered to have high binding affinity to its target site and low binding affinity to the target sites of other proteins. This section extends (computationally) this type of analysis to the entire HIV-1 genome (Genbank accession number K03455).

Previous scoring methods and the built SVM classifier were used to build a weight matrix capturing predicted binding preferences for each of the engineered proteins in the study. This weight matrix was used to scan HIV-1, noting the ranking of the target site among the protein's top scoring sites. Because each protein was engineered to bind to its intended target with high specificity, the target site is expected to be among the highest scoring sites. SVM classifier was built using cross-validation, while other methods were taken as-is. Table 3.9 shows the result of these scans (lower numbers in the table are better). Overall, the SVM classifier had the second best average of all methods,

most accurately predicting the target site among all possible sites for a single engineered protein, making SVM competitive with previously published methods. There was no clear leading approach in this test, as each method had the lowest ranking target site for at least one protein.

3.5.6 Learned Weights

Table 3.10 show the final SVM weight vector trained using all examples in the dataset. Learned weights were rounded to two decimal places to increase the numerical stability of the method. Rounded weights were always used during testing. Empty entries indicate features no found in the dataset.

3.6 Conclusions

This chapter investigated using structural information and relative binding affinities for modeling protein-DNA interactions, dealing with a prevalent class of transcription factors known as C_2H_2 zinc fingers. Structural analysis verified that these types of proteins bind DNA in a conserved binding pattern, and that this binding pattern can be used build a sequence-based model of the protein-DNA interface. Then, an extensive database of C_2H_2 Zinc Finger interactions was gathered from the literature, including (when available) information about the relative binding affinity of various protein-DNA configuration. Finally, the collected dataset was used to train a modified linear SVM, including a large number of comparative examples encoding relative binding affinities as well as non-binding (negative) examples.

Overall results are very promising—the SVM approach performed competitively with previously published methods in a wide variety of scenarios. SVM was tested using stringent per-protein cross-validation while other methods were run as is. As a result, testing was done conservatively with respect to the SVM method–so, if it were possible to perform cross-validation for the other methods, SVM's relative performance may improve.

Position 6

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
a	4.78		0.00	1.45		0.00	0.00	0.00	-4.99	0.00		0.00		0.00	-1.77	-0.05	0.89	1.26		0.00
c	0.00		0.00	0.00	0.00	0.00		-4.22	-2.19	0.00		0.00	0.00	3.14	1.73	2.39	2.38	-2.29		
g	2.24		0.00	0.00		0.00	0.00	0.00	-1.29	4.56		0.00	0.00	0.00	2.14	-1.38	-0.17	0.00		
t	5.19		0.00	4.56		0.00	0.00	0.00	-5.03	0.00		0.00	0.00	1.04	-0.31	0.41	2.59	0.00	0.00	

Position 3

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
a	2.09		0.06	0.79		-0.33	1.86		-2.12	0.00	0.00	3.35		1.88		1.44	0.90	-6.72		
c	1.03	0.04	3.31	1.18		-0.33	-0.91	0.00	-3.12	2.45		2.09		1.18	0.00	1.88	0.34	-2.96		0.00
g	2.09	0.00	-0.70	0.88		0.00	2.55		1.44	-0.80	0.00	2.61	0.00	0.09	0.00	0.95	0.90	-6.78		0.00
t	3.09	0.00	1.31	1.08		0.67	0.08	0.00	-4.07	0.00	0.00	1.62	0.00	0.09	0.00	2.44	1.90	-3.75	0.00	

Position -1

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
a	-0.34	0.00	-1.59	-1.29		-0.27	-1.88	0.00	-0.73	-4.74	0.00	2.38	0.00	2.77	1.17	0.84	-0.47	0.00		
c	-0.48	0.00	2.10	-1.29		0.73	-3.67	0.00	-0.21	-3.22	0.00	-2.44	0.00	1.55	1.91	-1.15	-1.41	0.00	0.00	
g	-2.19	0.00	-1.89	-0.29	0.00	-0.27	-0.83	0.00	-0.21	-3.22	0.00	0.00	2.93	2.33	4.64	0.84	0.53	1.23	0.00	0.00
t	1.04	0.00	-0.51	-1.72		-0.18	2.24	5.64	1.28	0.04	3.87	-0.37	3.50	1.86	0.67	1.84	3.10	0.42	2.46	

Position 2

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
a	1.91	0.00	-1.04	0.00	-2.90	1.26	-5.68		-3.11	0.00	0.00	0.00	0.00	0.69	0.00	3.02	2.77	0.00	0.00	
c	1.71	0.00	0.51	0.00	0.00	0.94	2.89		0.00	0.00	5.38	0.00	-2.77	0.00	-0.35	0.82	-1.32	6.04	0.00	0.00
g	1.13	6.22	-2.78	0.00	0.00	1.11	0.00		0.00	0.00	0.00	0.00	-4.29	1.65	2.37	4.11	-3.93	0.00	0.00	0.00
t	-2.61	0.00	-0.34		0.00	1.63	-0.94	0.00	3.11	0.00	0.00	0.00	0.00	0.00	-1.45	1.28	0.00	0.00		

Table 3.10: Learned weights for the trained SVM model. Each weight is represented by an entry in the table corresponding to a one of the twenty amino acids contacting one of four nucleic bases. Empty entries indicate features not found in the dataset.

In the future, one possibility would be to expand the database further, perhaps by including data from Transfac [84] or JASPAR [85]. Methodological improvements are possible by changing the optimization to include prior knowledge, perhaps in the form of previously estimated residue and base recognition preferences. Additional improvements may be possible by changing the basic binding model–for example, by considering contact 2 of one finger together with contact 6 of the next finger, or by using structural information further to eliminate from consideration residue-base combinations whose small size would prevent contact.

The current line of research could also benefit by including testing in other settings. For example, there have been several recent high-throughput *in vitro* studies mapping the DNA-binding specificities of transcription factors using DNA microarrays [91] and protein microarrays [92]. These studies include some zinc finger proteins, and while the exact binding sites are not known, and so the data cannot be used for training (at least not as binding examples), they provide a good external source of testing. Similarly, zinc finger binding preferences have also been explored in large-scale *in vivo* data (e.g., [93,94]), and may also provide a useful, though less direct, means for testing.

In conclusion, judging from the current findings, structural models of protein-DNA binding benefit from including binding affinity information. These types of approaches are expected to play an increasingly important role as more high-throughput data of protein-DNA binding is made available.

3.A Appendix: Binding Affinity

From a chemical point of view, protein-DNA binding are described succinctly with the following chemical reaction,

$$\text{Prot} + \text{DNA} \underset{\text{K}_{\text{off}}}{\overset{\text{K}_{\text{on}}}{\rightleftharpoons}} \text{Complex} \tag{3.4}$$

Equation 3.4 states that protein (in this case, zinc fingers) binds with DNA, producing a protein-DNA complex. According to the law of mass action, the rate of a reaction is proportional to the concentration of the reactants: the rate of the forward reaction is $\text{K}_{\text{on}}[\text{Prot}][\text{DNA}]$, while the rate of the reverse reaction is $\text{K}_{\text{off}}[\text{Complex}]$. At equilibrium these two rates are equal, which implies that the following quantity is a constant.

$$\frac{\text{K}_{\text{off}}}{\text{K}_{\text{on}}} = \frac{[\text{Prot}][\text{DNA}]}{[\text{Complex}]} \tag{3.5}$$

$\text{K}_{\text{d}} = \text{K}_{\text{off}}/\text{K}_{\text{on}}$ is known as the dissociation constant, while $\text{K}_{\text{a}} = 1/\text{K}_{\text{d}}$ is known as the association constant. K_{d} is expressed in units of concentration and is equal to the concentration of DNA required so that exactly half of the protein is bound (i.e.: when $[\text{Prot}] = [\text{Complex}]$). Higher K_{a} values correspond to stronger binding. Examples with either a K_{a} or K_{d} value are referred to as quantitative examples in the text.

While binding affinity measures the attraction between two molecules, the ability of a molecule to selectively bind to its intended target is called its binding specificity. Both binding affinity and specificity are important in controlling transcription. An ideal protein would have both high affinity and high specificity for its intended target (which implies low affinity for other targets).

Chapter 4

Conclusions

This thesis has made made contributions towards solving an important problem that arises when studying gene expression and regulation: that of identifying transcription factor binding sites. Approached from a computational viewpoint, identifying binding sites is a first step in uncovering the transcriptional circuitry of a cell, an understanding of which ultimately will inform how a cell functions by responding to its changing circumstances.

Two distinct approaches were taken. The first approach, described in chapter 2, is based on conserved statistical patterns in binding sites. A comprehensive study of various binding site representation methods was performed, evaluating how well they could identify additional binding sites of a transcription factor, when given a group sites it is already known to bind. Cross-validation testing and a rank sum test have shown that including information content and including local pairwise information results in statistically significant improvements in classifying binding sites.

The second approach, described in chapter 3, concentrates on a single family of transcription factors with a known structure and well-conserved binding pattern. This binding pattern was verified using crystal structures from the PDB and used to construct a binding model. A modified support vector machine (SVM) was used to learn from data gathered from the literature. The structural approach allows the addition of protein information,

while the modified SVM allowed including relative binding information through comparative examples. Overall, current performance was comparable to previous methods, as demonstrated on numerous results, and performance is expected to improve as further data of relative binding affinity is made available.

While these two approaches towards recognizing transcription factor binding sites are very different, future research combining the two may be beneficial. For example, in the sequence-based approach, including pairwise nucleotide correlations resulted in improved performance in recognizing binding sites, but only if the nucleotides considered were close together in sequence. It may be possible that more careful selection of pairwise dependencies, perhaps using some structural information, would result in further improvements. Similarly, the use of several known binding sites for certain zinc finger proteins may be helpful in further evaluating the effect of residue-base combinations.

Finally, the research presented in this thesis on predicting zinc finger interactions motivates studying protein-DNA interactions based on other structural motifs. Transcription factors in the PDB fall into eight structural groups, which can be further classified into approximately 50 structural families [44]. As quantitative affinity data and examples of non-binding DNA segments are gathered via high-throughput technologies for structurally diverse transcription factors, an optimization strategy similar to what was done with the zinc fingers may be useful for predicting the binding of other structural families.

Bibliography

[1] R. Osada, E. Zaslavsky, and M. Singh. Comparative analysis of methods for representing and searching for transcription factor binding sites. *Bioinformatics*, 20(18):3516–3525, December 2004.

[2] C.R. Calladine and H.R. Drew. *Understanding DNA: The Molecule and How It Works*. Academic Press, San Diego, California, second edition, 1997.

[3] W.H.E. Day and F.R. McMorris. Critical comparison of consensus methods for molecular sequences. *Nucleic Acids Res.*, 20(5):1093–1099, March 1992.

[4] R. Staden. Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.*, 12(1):505–519, January 1984.

[5] O.G. Berg and P.H. von Hippel. Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, 193(4):723–750, February 1987.

[6] T.-K. Man and G.D. Stormo. Non-independence of Mnt repressor-operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.*, 29(12):2471–2478, June 2001.

[7] M.L. Bulyk, P.L.F. Johnson, and G.M. Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.*, 30(5):1255–1261, March 2002.

[8] T.D. Schneider and R.M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, 18(20):6097–6100, October 1990.

[9] G.Z. Hertz and G.D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7):563–577, July 1999.

[10] K. Robison, A.M. McGuire, and G.M. Church. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol.*, 284(2):241–254, November 1998.

[11] S.A. Wolfe, L. Nekludova, and C.O. Pabo. DNA recognition by Cys_2His_2 zinc finger proteins. *Annu. Rev. Biophys. Biomol. Struct.*, 29:183–212, June 2000.

[12] P.V. Benos, A.S. Lapedes, and G.D. Stormo. Probabilistic code for DNA recognition by proteins of the EGR family. *J. Mol. Biol.*, 323(4):701–727, November 2002.

[13] T. Kaplan, N. Friedman, and H. Margalit. Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Computational Biology*, 1(1):5–13, June 2005.

[14] T.D. Schneider, G.D. Stormo, L. Gold, and A. Ehrenfeucht. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 188(3):415–431, April 1986.

[15] M.S. Gelfand. Prediction of function in DNA sequence analysis. *J. Comput. Biol.*, 2(1):87–115, February 1995.

[16] G.D. Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, January 2000.

[17] H. Salgado, S. Gama-Castro, A. Martínez-Antonio, E. Díaz-Peredo, et al. RegulonDB (version 4.0): Transcriptional regulation, operon organization and growth conditions in *Escherichia Coli* K-12. *Nucleic Acids Res.*, 32(Database):D303–D306, 2004.

[18] D. Thieffry, H. Salgado, A.M. Huerta, and J. Collado-Vides. Prediction of transcriptional regulatory sites in the complete genome sequence of *Escherichia coli* K-12. *Bioinformatics*, 14(5):391–400, June 1998.

[19] J.D. Hughes, P.W. Estep, S. Tavazoie, and G.M. Church. Computational identification of *Cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, 296(5):1205–1214, March 2000.

[20] S. Sinha and M. Tompa. A statistical method for finding transcription factor binding sites. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 344–355, San Diego, CA, August 2000. AAAI.

[21] M.S. Gelfand, E.V. Koonin, and A.A. Mironov. Prediction of transcription regulatory sites in archaea by a comparative genomic approach. *Nucleic Acids Res.*, 28(3):695–705, February 2000.

[22] A.M. McGuire, J.D. Hughes, and G.M. Church. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.*, 10(6):744–757, June 2000.

[23] L. McCue, W. Thompson, C. Carmack, M.P. Ryan, J.S. Liu, V. Derbyshire, and C.E. Lawrence. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.*, 29(3):774–782, February 2001.

[24] K. Tan, G. Moreno-Hagelsieb, J. Collado-Vides, and G.D. Stormo. A comparative genomics approach to prediction of new members of regulons. *Genome Res.*, 11(4):566–584, April 2001.

[25] M. Blanchette and M. Tompa. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.*, 12(5):739–748, May 2002.

[26] P.V. Benos, M.L. Bulyk, and G.D. Stormo. Additivity in protein–DNA interactions: how good an approximation is it? *Nucleic Acids Res.*, 30(20):4442–4451, October 2002.

[27] F.R. Blattner, G. Plunkett III, C.A. Bloch, N.T. Perna, V. Burland, et al. The complete genome sequence of *Escherichia coli* K-12. *Science*, 277(5331):1453–1462, September 1997.

[28] E.R. Tufte. *The Visual Display of Quantitative Information.* Graphics Press, Cheshire, Connecticut, 2nd edition, 2001.

[29] E.R. Tufte. *Envisioning Information.* Graphics Press, Cheshire, Connecticut, 1990.

[30] E.R. Tufte. *Visual Explanations.* Graphics Press, Cheshire, Connecticut, 1997.

[31] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, July and October 1948.

[32] K. Yamauchi. The sequence flanking translational initiation site in protozoa. *Nucleic Acids Res.*, 19(10):2715–2720, May 1991.

[33] C.E. Lawrence, S.F. Altschul, M.S. Boguski, J.S. Liu, A.F. Neuwald, and J.C. Wootton. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–214, October 1993.

[34] O.G. Berg and P.H. von Hippel. Selection of DNA binding sites by regulatory proteins: Ii. the binding specificity of cyclic AMP receptor protein to recognition sites. *J. Mol. Biol.*, 200(4):709–723, April 1988.

[35] M.O. Zhang and T.G. Marr. A weight array method for splicing signal analysis. *CABIOS*, 9(5):499–509, October 1993.

[36] I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco, California, 2000.

[37] S. Holm. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, 6:65–70, 1979.

[38] Y. Barash, G. Elidan, N. Friedman, and T. Kaplan. Modeling dependencies in protein-DNA binding sites. In *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology, Berlin, Germany, April 2003*, pages 28–37. ACM, April 2003.

[39] S.M. Ross. *Introudction To Probability And Statistics For Engineers And Scientists*. John Wiley & Sons, New York, New York, 1987.

[40] J.C. Venter, M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, et al. The Sequence of the Human Genome. *Science*, 291(5507):1304–1351, February 2001.

[41] C.O. Pabo, E. Peisach, and R.A. Grant. Design and selection of novel Cys_2His_2 zinc finger proteins. *Annu. Rev. Biochem.*, 70:313–340, 2001.

[42] Y. Mandel-Gutfreund and H. Margalit. Quantitative parameters for amino-acid base interaction: implication for prediction of protein-DNA binding sites. *Nucleic Acids Res.*, 26(10):2306–2312, May 1998.

[43] C. Branden and J. Tooze. *Introduction to Protein Structure*. Garland Publishing, New York, New York, 1991.

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

[45] L. Pauling. Nature of forces between large molecules of biological interest. *Nature*, 161(4097):707–709, May 1948.

[46] A. Klug and J.W.R. Schwabe. Protein motifs 5. zinc fingers. *FASEB Journal*, 9:597–604, May 1995.

[47] J.M. Berg. Zinc fingers and other metal-binding domains. *J. Biol. Chem.*, 265(12):6513–6516, April 1990.

[48] N. Hulo, C.J.A. Sigrist, V.L. Saux, P.S. Langendijk-Genevaux, et al. Recent improvements to the PROSITE database. *Nucleic Acids Res.*, 32(Database):D134–D137, January 2004.

[49] G.E. Crooks, G. Hon, J.M. Chandonia, and S.E. Brenner. WebLogo: A sequence logo generator. *Genome Research*, 14(6):1188–1190, June 2004.

[50] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and D.L. Wheeler. GenBank: update. *Nucleic Acids Res.*, 32(Database):D23–D26, January 2004.

[51] N.P. Pavletich and C.O. Pabo. Zinc finger-DNA recognition: Crystal structure of a Zif268-DNA complex at 2.1 Angstrom. *Science*, 252(5007):809–817, May 1991.

[52] M. Elrod-Erickson, M.A. Rould, L. Nekludova, and C.O. Pabo. Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. *Structure*, 4(10):1171–1180, October 1996.

[53] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, and T.E. Ferrin. UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612, July 2004.

[54] M. Elrod-Erickson, T.E. Benson, and C.O. Pabo. High-resolution structures of variant Zif268-DNA complexes: implications for understanding zinc finger-DNA recognition. *Structure*, 6(4):451–464, April 1998.

[55] J.E. Mills and P.M. Dean. Three-dimensional hydrogen-bond geometry and probability information from a crystal survey. *J. Comput-Aided Mol. Des.*, 10(6):607–622, December 1996.

[56] I.K. McDonald and J.M. Thornton. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, 238(5):777–793, May 1994.

[57] R.T. Nolte, R.M. Conlin, S.C. Harrison, and R.S. Brown. Differing roles for zinc fingers in DNA recognition: Structure of a six-finger transcription factor IIIA complex. *Proc. Natl. Acad. Sci. USA*, 95(6):2938–2943, March 1998.

[58] N.P. Pavletich and C.O. Pabo. Crystal structure of a five-finger GLI-DNA complex: New perspectives on zinc fingers. *Science*, 261(5129):1701–1707, September 1994.

[59] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, New York, New York, 2000.

[60] M.A. Hearst. Support vector machines. *IEEE Intelligent Systems*, 13(4):18–28, July/August 1998.

[61] T. Joachims. Making large-scale SVM learning practical. In B. Scholkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods–Support Vector Learning*. MIT-Press, 1999.

[62] R. Fourer, D.M. Gay, and B.W. Kernighan. *AMPL: A Modeling Language for Mathematical Programming*. Brooks/Cole Publishing Company, Pacific Grove, CA, 2002.

[63] R. Vanderbei. LOQO: An interior point code for quadratic programming. *Optimization Methods and Software*, 12:451–484, 1999.

[64] L. Reynolds, C. Ullman, M. Moore, M. Isalan, M.J. West, P. Clapham, A. Klug, and Y. Choo. Repression of the HIV-1 5' LTR promoter and inhibition of HIV-1

replication by using engineered zinc-finger transcription factors. *Proc. Natl. Acad. Sci. USA*, 100(4):1615–1620, February 2003.

[65] K.-H. Bae, Y.D. Kwon, H.-C. Shin, M.-S. Hwang, E.-H. Ryu, K.-S. Park, et al. Human zinc fingers as building blocks in the construction of artificial transcription factors. *Nat. Biotechnol.*, 21(3):275–280, March 2003.

[66] P. Blancafort, L. Magnenat, and C.F. Barbas III. Scanning the human genome with combinatorial transcription factor libraries. *Nat. Biotechnol.*, 21(3):269–274, March 2003.

[67] B. Dreier, R.R. Beerli, D.J. Segal, J.D. Flippin, and C.F. Barbas III. Development of zinc finger domains for recognition of the 5'-ANN-3' family of DNA sequences and their use in the construction of artificial transcription factors. *J. Biol. Chem.*, 276(31):29466–29478, August 2001.

[68] M. Isalan, A. Klug, and Y. Choo. A rapid, generally applicable method to engineer zinc fingers illustrated by targeting the HIV-1 promoter. *Nat. Biotechnol.*, 19(7):656–660, July 2001.

[69] B. Dreier, D.J. Segal, and C.F. Barbas III. Insights into the molecular recognition of the 5'-GNN-3' family of DNA sequences by zinc finger domains. *J. Mol. Biol.*, 303(4):489–502, November 2000.

[70] D.J. Segal, B. Dreier, R.R. Beerli, and C.F. Barbas III. Toward controlling gene expression at will: Selection and design of zinc finger domains recognizing each of the 5'-GNN-3' DNA target sequences. *Proc. Natl. Acad. Sci. USA*, 96(6):2758–2763, March 1999.

[71] S.A. Wolfe, H.A. Greisman, E.I. Ramm, and C.O. Pabo. Analysis of zinc fingers optimized *via* phage display: Evaluating the utility of a recognition code. *J. Mol. Biol.*, 285(5):1917–1934, February 1999.

[72] H.A. Greisman and C.O. Pabo. A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science*, 275(5300):657–661, January 1997.

[73] M. Isalan, Y. Choo, and A. Klug. Synergy between adjacent zinc fingers in sequence-specific DNA recognition. *Proc. Natl. Acad. Sci. USA*, 94(11):5617–5621, May 1997.

[74] H. Wu, W.-P. Yang, and C.F. Barbas III. Building zinc fingers by selection: Toward a therapeutic application. *Proc. Natl. Acad. Sci. USA*, 92(2):344–348, January 1995.

[75] Y. Choo and A. Klug. Toward a code for the interactions of zinc fingers with DNA: Selection of randomized fingers displayed on phage. *Proc. Natl. Acad. Sci. USA*, 91(23):11163–11167, November 1994.

[76] Y. Choo and A. Klug. Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc. Natl. Acad. Sci. USA*, 91(23):11168–11172, November 1994.

[77] A.C. Jamieson, S.-H. Kim, and J.A. Wells. *In Vitro* selection of zinc fingers with altered DNA-binding specificity. *Biochemistry*, 33(19):5689–5695, May 1994.

[78] E.J. Rebar and C.O. Pabo. Zinc finger phage: Affinity selection of fingers with new DNA-binding specificities. *Science*, 263(5147):671–673, February 1994.

[79] J.R. Desjarlais and J.M. Berg. Use of a zinc-finger consensus sequence framework and specificity rules to design specific DNA binding proteins. *Proc. Natl. Acad. Sci. USA*, 90(6):2256–2260, March 1993.

[80] J.R. Desjarlais and J.M. Berg. Toward rules relating zinc finger protein sequences and DNA binding site preferences. *Proc. Natl. Acad. Sci. USA*, 89(16):7345–7349, August 1992.

[81] J. Nardelli, T. Gibson, and P. Charnay. Zinc finger-DNA recognition: analysis of base specificity by site-directed mutagenesis. *Nucleic Acids Res.*, 20(16):4137–4144, August 1992.

[82] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Res.*, 28(1):235–242, January 2000.

[83] P.V. Benos, A.S. Lapedes, D.S. Fields, and G.D. Stormo. SAMIE: Statistical algorithm for modeling interaction energies. *Pac. Symp. Biocomput.*, pages 115–26, 2001.

[84] E. Wingender, X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, et al. The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, 29(1):281–283, January 2001.

[85] A. Sandelin, W. Alkema, P. Engstrvm, W. Wasserman, and B. Lenhard. JASPAR: an open access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, 32(Database):D91–D94, January 2004.

[86] M. Suzuki, S.E. Brenner, M. Gerstein, and N. Yagi. DNA recognition code of transcription factors. *Protein Eng.*, 8(4):319–328, 1995.

[87] R.G. Endres, T.C. Schulthess, and N.S. Wingreen. Toward an atomistic model for predicting transcription-factor binding sites. *Proteins: Structure, Function, and Bioinformatics*, 57(2):262–268, November 2004.

[88] A.V. Morozov, J.J. Havranek, D. Baker, and E.D. Siggia. Protein-DNA binding specificity with structural models. *Nucleic Acids Res.*, 33(18):5781–5798, October 2005.

[89] S. Cawley, S. Bekiranov, et al. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, 116(4):499–509, February 2004.

[90] StataCorp. *Stata Statistical Software: Release 8*. StataCorp, College Station, Texas, 2003.

[91] S. Mukherjee, M.F. Berger, G. Jona, X.S. Wang, D. Muzzey, M. Snyder, R.A. Young, and M.L. Bulyk. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.*, 36(12):1331–1339, December 2004.

[92] S.-W. Ho, G. Jona, C.T.L. Chen, M. Johnston, and M. Snyder. Linking DNA-binding proteins to their recognition sequences by using protein microarrays. *Proc. Natl. Acad. Sci. USA*, 103(26):9940–9945, June 2006.

[93] T.I. Lee, N.J. Rinaldi, F. Robert, et al. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298(5594):799–804, October 2002.

[94] C.T. Harbison, D.B. Gordon, T.I. Lee, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, September 2004.