Analysis of DNA Motifs Based on a Novel Motif Comparison Method

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

by

Naomi Habib

Supervised by

Prof. Hanah Margalit and Prof. Nir Friedman

March 2007

The School of Computer Science and Engineering
The Hebrew University of Jerusalem, Israel

Abstract

Characterizing the DNA-binding specificities of transcription factors is a key problem in computational biology that has been addressed by multiple algorithms. These usually take as input sequences that are putatively bound by the same factor and output one or more probabilistic DNA motifs. A common practice is to apply several such algorithms simultaneously to improve coverage at the price of redundancy. Two crucial tasks for interpreting such results regard clustering of redundant motifs and attributing the motifs to transcription factors by retrieval of similar motifs from previously characterized motif libraries. Both tasks inherently involve motif comparison. Here we present a novel method for comparing and merging motifs, based on Bayesian probabilistic principles. This method takes into account both the similarity in positional nucleotide distributions of the two motifs and their dissimilarity to the background distribution. We demonstrate the use of the new comparison method as a basis for motif clustering and retrieval procedures, and compare it to several commonly used alternatives. Our results show that the new method outperforms other available methods in accuracy and sensitivity. The resulting motif clustering and retrieval procedures we incorporated in a large-scale automated pipeline for analyzing DNA motifs. This pipeline integrates the results of various DNA motif discovery algorithms and automatically merges redundant motifs from multiple training sets into a coherent annotated library of motifs. Application of this pipeline to recent genome-wide transcription factor location data in *S. cerevisiae* successfully identified DNA motifs in a manner that is as good as semi-automated analysis reported in the literature. Moreover, we show how this analysis elucidates the mechanisms of condition-specific preferences of transcription factors.

Acknowledgements

First and foremost, I would like to thank my advisors Hanah Margalit and Nir Friedman, for guiding my first steps as a researcher and for constantly challenging and assisting me to go further. I was fortunate to have the opportunity to work with them both, since each exposed me to different scientific perspective. I would also like to thank Tommy Kaplan who in many ways served as my third adviser. I will always be grateful for the time he devoted to help me: advising and offering scripts that made my work much easier. I thank all members of both labs, for always finding the time to help, being good listeners, giving good advice and especially for being such good friends. I want to thank Ariel Jaimovich from my lab, who was always supportive and helped me a great deal to submit my thesis on time. Last but not least, I thank my dear friends and family for being with me at all times. Especially my parents, who also provided good professional advise, although they continue to insist they do not understand anything I wrote.

Contents

Chapter 1

Introduction

1.1 From DNA to Function

Living cells transfer information from one living cell to its daughter cells through nucleic acids chains, mainly Deoxyribonucleic Acid - DNA, which contains the genetic instructions for the development and function of living organisms. The information in the DNA encoded by a four letters alphabet, A,C,G,T, which are in fact different nucleotides. The long DNA chain contains short segments called genes. In the transcription process, short ribonucleic acid (RNA) chains are synthesized according to the information encoded in the genes. A RNA molecule encodes the information needed to construct proteins, in a process called translation. The proteins make an essential part of all living organisms and participate in every process in the cells, defining both the cells function and structure.

One of the miraculous phenomena in nature is that cells change their activity significantly, in response to changes in their environment or external signals, while their DNA, which is the blueprint for their function, remains the same. An even more intriguing fact is that different cells in the same multicellular organism have identical copies of the DNA and nevertheless their function and structure vary considerably. For example, an epithelial cell in the skin tissue has a completely different function and shape from a neuron cell in the brain of the same organism. This raises the question how do cells develop different functions and structures when this genetic instructions are identical?

The answer to this question is that in each cell, at a given time, only part of the proteins encoded in the DNA are present. The activity and structure of the cell in a given state is determined by its proteins, implying that exact dosage and content of proteins at a given time is highly important for the correct function of the cell. This highly specific content of functional proteins in each cell is achieved through several layers of regulations. The first layer is transcription regulation, which

controls which genes are expressed and transcribed to RNA. The second layer of regulation regards the control of translation process of RNA to proteins. This post-transcription layer includes regulation of the processing of the RNA molecule in to a mature transcript. The third layer of regulation acts post-translationally, controlling the function, location and degradation of the proteins themselves.

We focus in this work on specific aspects of the first layer of regulation in the cell, transcription regulation. Key players in the transcription regulation are transcription factors which bind to sequence-specific motifs in the DNA and constantly modulate (activate or repress) the expression of nearby genes. These factors recognize specific sequence patterns on the DNA that are called transcription factor binding sites. To understand transcription regulation it is essential to construct a map of transcription factors and their targets, indicating when every transcription factor is active, which genes it regulates, does it lead to activation or repression of genes and how this regulation is carried out. One of the first steps in building such a transcription regulation map is to define the sequence preferences of each transcription factor and the distribution in the genome of its potential binding sites. This initial mapping indicates which factors can bind to the DNA at a given location and consequently are candidates for regulation of proximal genes. In addition, the location of DNA binding motifs can provide evidence of physical interactions between transcription factors. In this work, we address several computational challenges related to identifying the sequence-specific DNA motifs identified by transcription factors. In addition, from the biological aspect, we show here what DNA motifs can teach us about the complex mechanisms of gene expression regulations by transcription factors.

Chapter 2

DNA Motifs

To understand how transcription factors associate with the DNA, one must specify their DNA binding preferences. These preferences are usually characterized by a motif that summarizes the commonalities among the binding sites of a transcription factor.

2.1 DNA Motif Representation

In the literature there is an ongoing discussion regarding the best representation of the DNA binding specificities of transcription factors [Osada et al., 2004, Day and McMorris, 1992, Benos et al., 2002, Stormo, 2000]. A DNA motif is an abstraction that models the sequence preferences of DNA binding proteins. The motif is built on the basis of multiple sequences known to be bound by the transcription factor.

The simplest kind of motif representation is the consensus sequence e.g. [Day and McMorris, 1992], a string of nucleotides that represents the most abundant nucleotides in each positions of the protein's binding site. For example: The consensus sequence TGACTC represents the binding preferences of the transcription factor Gcn4 in *S. cerevisiae*. However, this model is not flexible enough, since proteins often display variations in binding specificities. A common addition to this model is the use of the IUPAC one-letter codes, also known as ambiguity codes. For example, W represents A or T (weak interaction, 2 hydrogen bonds) and S represents G or C (strong interaction, 3 hydrogen bonds). Other commonly used one letter codes are: R = G or A, Y = T or C, M = A or C, K = G or T. For example: The following sequences are all binding sites of the transcription factor Gcn4 in the *S. cerevisiae* genome: TGACTC, TTACTC, TGACTG. Thus a more accurate consensus sequence for GCN4 based on these sequences is: TKACTS.

Another common representation, which has the benefits of being relatively

simple yet flexible, is a matrix of positions in the binding site versus nucleotides. In the matrix each row represents one residue (A, C, G or T), and each column represents a position in a set of aligned binding sites. There are several types of matrix representations which differ in the type of score they hold in the entries. However, all matrix representations assume that the choice of nucleotides in each position of the motif is independent of all other positions. A common matrix representation is a matrix of nucleotide frequencies in each position of the motif (i.e. the frequencies of the nucleotides A, C, G and T in each position). This matrix is called a Position specific Weight Matrix (PWM) (Figure 2.1), often referred to as a Position-Specific Probability Matrix, or a Profile. A profile is a more flexible representation than the consensus sequence described above, for example it allows us to differentiate between binding preferences of 50% A 50% T and preferences of 70% A 30% T. We are often interested in considering the nucleotide counts in each position which are more informative then the frequencies alone. We call such a count matrix an un-normalized PWM (Figure 2.1), and the transformation between such a matrix and a profile is by simple normalization.

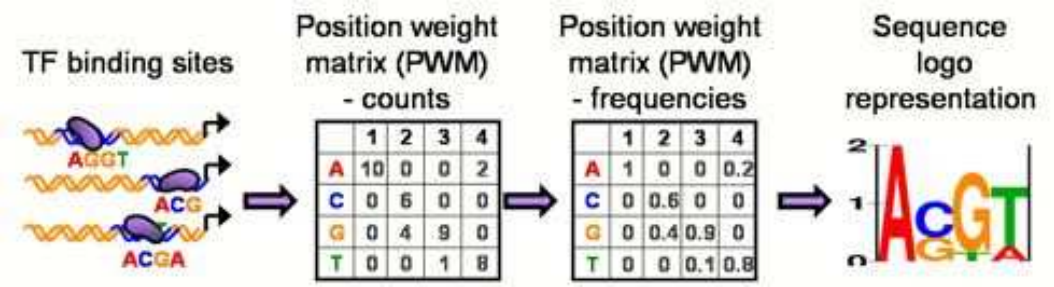

Figure 2.1: **Motif representation.** Constructing a DNA motif for a transcription factor based on given instances of genomic binding sites of the factor. In this illustration a set of binding sites are converted to an unnormalized PWM from the nucleotide counts in each position of the sites. By normalizing each position we get to the alternative PWM (or profile) representation, which contains the nucleotide frequencies in each position in the motif. From the PWM, a graphical representation for the motif can be constructed using a sequence-logo (Schneider and Stephens, 1990). In the logo the y-axis is the information content of a position and the height of each nucleotide is proportional to its frequency in the relevant position of the motif.

Another common representation is a scoring matrix often called a Position specific Scoring Matrix (PSSM)[Staden, 1984], where each position holds the log-likelihood score of each residue to be generated by the motif model. This log likelihood is the log-ratio of two probabilities: the probability to observe the nucleotide given the motif model (the matrix), and the probability to observe the nucleotide given the background model of the frequencies of each nucleotide in

the genomic context. Note that the names of the motif matrices are not always consistent in the literature. For example a PSSM is in some places referred to as a PWM.

One weakness of the above matrix representation for DNA motifs is that they do not take into account higher order dependencies between residues, such as correlations between different positions in the binding site [Man and Stormo, 2001, Bulyk et al., 2002]. For example, if a transcription factor may bind to the sequence: ACGTCC or ACGTGG (CC or GG suffix at the end of the motif), we cannot use a representation that assumes independent positions since this will lead to all possible combinations of suffixes including CG and GC. In this example, the transcription factor has a finite set of binding preferences, which can correlate to different structural configurations of the protein. In this case we can model the binding motif relatively simply by a mixture of PSSMs, in which a transcription factor can bind to any sequence that fits any one of the matrices. Other cases may be more complex and require further modeling of higher order dependencies. Pairwise positional dependencies can be modeled by using a simple correlation matrix with entries for each pair of positions in the motif [Zhang and Marr, 1993]. A more compact and general model is a Bayesian Network, which has been used to model arbitrary dependencies [Barash et al., 2003]. The problem in estimating transcription factor binding preferences as a model with positional dependencies is that it requires a large amount of data. When sufficient data are not available there is a risk of over-fitting.

It has been previously shown that in practice the simpler motif models are often both useful and practical [Benos et al., 2002] and provide a useful approximation to reality. In this work we are using the unnormalized PWM representation (a count matrix) for the description of a DNA motif and we refer to it for simplicity as a PWM.

2.2 Motif Discovery Algorithms

Multiple tools were developed for finding DNA motifs. Most algorithms identify statistically significant overrepresented sequence patterns in a set of related DNA sequences. These groups of related DNA sequences are believed to be control regions in the DNA of a co-regulated group of genes. Thus we expect to find in the control regions the DNA binding motifs of a set of transcription factors that mediate this co-regulation. Deriving the groups of co-regulated sequences can be done from ChIP-on-chip data that characterize a group of DNA sequences bound by the same transcription factor from gene expression data that provide clusters of co-expressed genes, or from functional analysis assays that define genes with a related function, such as genes involved in the same metabolic pathway.

Finding over-represented motifs can be done by enumerative methods, which count exhaustively all words in the dataset. Since this approach is computationally expensive, most algorithms constrain the motif length or the alphabet size. An example of an enumerative algorithm is Weeder [Pavesi et al., 2001].

An alternative approach for finding over-represented motifs is by a probabilistic search which constructs a generative model of the sequence data and searches for a motif that maximizes the likelihood of the observed data. Several probabilistic search algorithms are based on the Expectation Maximization method such as: MEME [Bailey and Elkan, 1995], and EMnEM [Moses et al., 2004], while others are based on the Gibbs sampling method, such as AlignAce [Hughes et al., 2000], MotifSampler [Thijs et al., 2001] and PhyloGibbs [Siddharthan et al., 2005].

One of the problems in these motif discovery algorithms is that the input set of sequences is usually noisy. One property that can be used to decrease the noise is to use the degree of confidence we have a-priori for each DNA sequence that it is co-regulated with the rest of the group. The MDscan [Liu et al., 2002] algorithm is an example for such an algorithm, which receives as an input a ranked group of sequences according to the confidence level of each one. Several algorithms integrate evolutionary conservation information based on the reasoning that important regulatory regions, such as transcription factor binding sites, are under evolutionary pressure and as a consequence are more conserved than other non-coding DNA sequences. The conservation information can be integrated in the algorithms described above by finding conserved and overrepresented motifs in a group of related DNA sequences. This is done in the algorithm PhyloGibbs [Siddharthan et al., 2005] and EMnEM [Moses et al., 2004]. In addition, the conservation information can be used in a genome-wide motif discovery performed on phylogenetically conserved non-coding regions [Kellis et al., 2003]. The drawback of integrating conservation information is that regulatory regions are not always conserved [Tautz, 2000, Moses et al., 2006, Levine and Tjian, 2003], especially in remote species. Several studies show how different regulatory programs in different species lead to similar function e.g. [Tsong et al., 2006], which indicate that the regulation program may be highly flexible. Thus integrating conservation considerations may lead to overlooking of the regulatory signal.

A different approach is to use structural knowledge to infer the binding motifs of transcription factors. This can be done as ab-initio prediction of binding preferences from the structure of the DNA binding domain of proteins [Kaplan et al., 2005, Morozov et al., 2005]. Moreover, structural knowledge can be integrated as a bias to the motif discovery based on prior knowledge of the typical motifs of structural families of transcription factors [Sandelin and Wasserman, 2004, MacIsaac and Fraenkel, 2006]. In addition, the motif discovery can be biased according to positional priors of structural classes in the genome [Narlikar et al., 2006].

Different motif discovery methods were shown to have complementary successes, and no one is clearly superior [Tompa et al., 2005]. It is therefore beneficial to apply multiple methods simultaneously and collate their results [MacIsaac and Fraenkel, 2006].

2.3 Emergent Obstacles

There are several problems in interpreting the output of motif discovery algorithms: (a) Many of these methods output multiple results which require scoring and ranking (b) The outputs of these motif discovery algorithms are frequently redundant and the binding transcription factor is unknown (see example in Figure 2.2). (c) In large-scale experiments the motif output set is very large, and thus the tasks of scoring, merging and identifying motifs need to be done automatically. Since it is beneficial to apply multiple methods simultaneously, the number as well as the redundancy of the discovered motifs is amplified. As similar motifs may represent binding sites of the same protein, eliminating this redundancy is essential for elucidating the true transcriptional regulatory program.

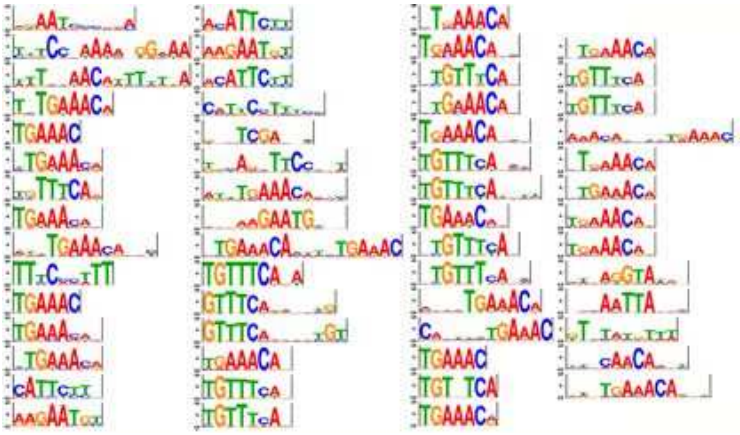

Figure 2.2: **Motif discovery output.** An example for a motif discovery output. six motif discovery tools: MDscan[Liu et al., 2002], AligneAce[Hughes et al., 2000], MEME [Bailey and Elkan, 1995], $MEME_c$ [Harbison et al., 2004], converge[Harbison et al., 2004], SeedSearcher [Barash, 2005] and an additional method [Kellis et al., 2003], were applied on a set of genes found to be bound by the cofactor DIG1 in a ChIP-chip assay [Harbison et al., 2004]. Clearly, deciding which motifs represent the binding site of the same protein is not a trivial task. This task gets more complicated as the number of motifs increases

The general strategy for reducing this redundancy involves clustering similar motifs together and merging motifs within each cluster to create a library of non-redundant motifs [MacIsaac and Fraenkel, 2006] (Figure 2.3.B). An additional

important step in interpreting the output of motif discovery algorithms is to relate the discovered motifs to previously characterized recognition sequences of known transcription factors. This task involves retrieval - given a query motif, find similar motifs in a motif database (Figure 2.3.C). To address both the clustering and the retrieval challenges, we need an accurate and sensitive method for comparing DNA motifs

2.4 DNA Motif Comparison

To compare two PWMs, we can utilize the position-independence assumption to decompose the similarity score of two motifs into the sum of similarities of single aligned positions. Two motifs may be of different length or reverse complement each other (meaning that they are taken from the complementary strand of the DNA), and thus all possible alignments should be considered. The similarity score between two motifs is the highest score of all possible alignments of the motifs. Several similarity scores can be used to compare a pair of aligned positions in a PWM. One possible approach is based on statistical measures, such as the Pearson correlation coefficient (e.g. as used in CompareACE [Hughes et al., 2000],[Xie et al., 2005]). This measure, however, might inappropriately capture similarities between probabilities (Figure 2.4). An alternative approach is to define a similarity between two distributions. This can be a metric distance, such as the Euclidean distance [Harbison et al., 2004] or an information-theoretic measure, such as the Jensen-Shannon divergence [Cover and Thomas, 2001]. The later distances measure distance between vectors, thus they do not have the artifacts of the Pearson correlation. Such measurements, however, equally weight positions with similar nucleotide distributions that are specific (e.g., a strong preference for an A) and similar positions that are non-informative (e.g., identical to the background distribution); (Figure 2.4). It is important to differentiate between the two situations, because the two positions whose similarity is due to a resemblance to the background distribution are less relevant to motif similarity, as they do not contribute to sequence-specific binding of proteins [Yona and Levitt, 2002]. This argument suggests that a proper motif comparison method should reflect the likelihood that the two motifs represent sites bound by the same factor. Hence, the motif comparison method should take into account the sequence similarity between the two DNA motifs and at the same time, also take into account the extent to which they are different from the background distribution. In this work we use this intuition to develop a novel method for comparing and merging DNA motifs, based on Bayesian probabilistic reasoning.

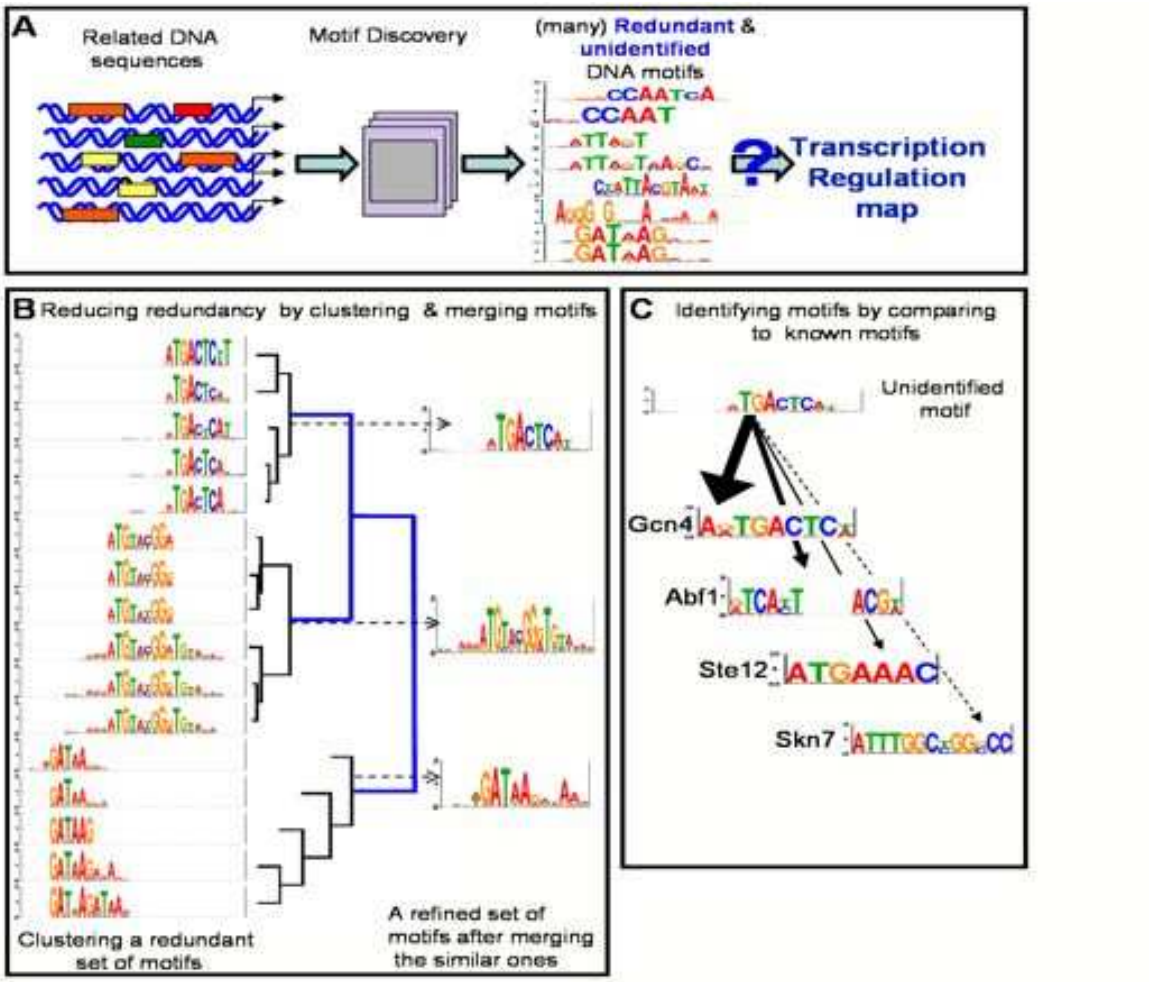

Figure 2.3: **Emergent obstacles and possible solutions .**(A) Identifying DNA binding sites of transcription factors: Applying motif discovery algorithms on a group of related DNA sequences leads to finding putative transcription factor DNA binding sites. These algorithms output a set of unidentified DNA motifs, which are frequently redundant. To infer the correct transcription regulation map from the discovered motif set, it is crucial to reduce this redundancy and identify the newly discovered motifs. (B) Reducing redundancy by clustering and merging motifs: A redundant set of DNA motifs can be reduced by clustering the motifs into groups of related motifs and merging the motifs in each cluster. In this example, a redundant set of 16 DNA motifs (a partial output of several motif search algorithms, as in Figure 2.2) is clustered and merged to a final set consisting of three DNA motifs. For correct clustering an accurate and sensitive DNA motif similarity score is needed. (C) Identifying the binding factors of DNA motifs: The transcription factors that bind unidentified DNA motifs can be revealed based on similarities to previously defined TF binding motifs. In this example, comparison of a newly discovered motif to four known motifs reveals high similarity to the Gcn4 known binding motif. From this comparison the transcription factor that bind the motif is identified with high probability. For the comparison of DNA motifs an accurate and sensitive motif similarity score is needed.

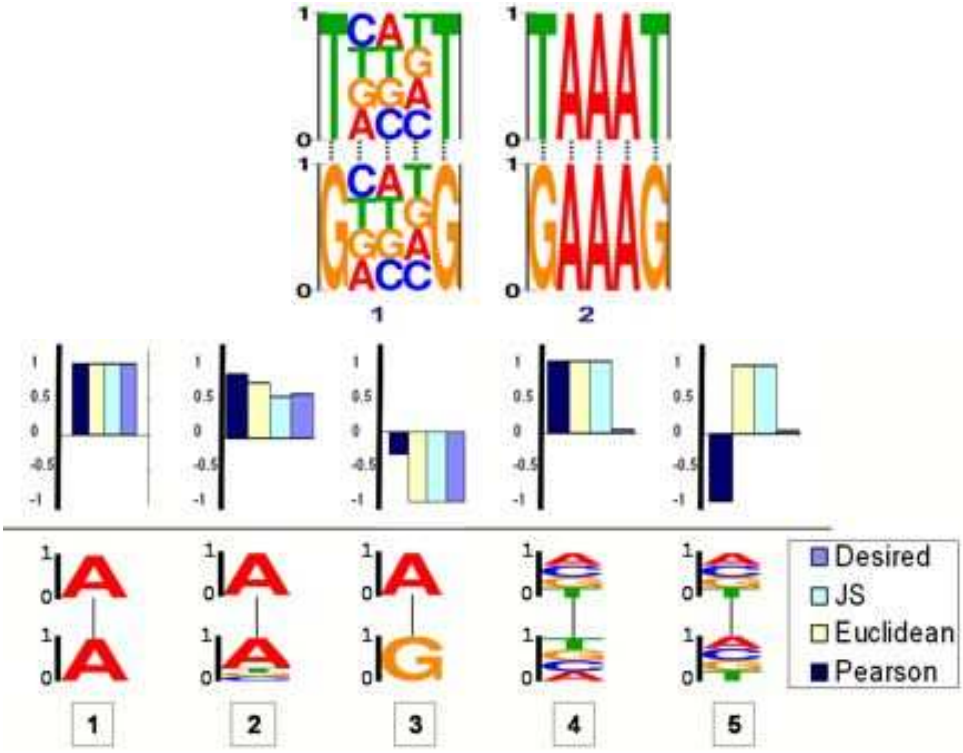

Figure 2.4: **Motif comparison.**(A) Differentiating between informative and non-informative positions. Two pairs of aligned motifs that both have three identical positions and two different ones. However, the identical positions in pair number one are non-informative, while the identical positions in pair number two are informative. Thus we would like our score to differentiate between these two types of similarities and assign a higher similarity score to pair number two. The nucleotide distribution in each motif is represented schematically (with a sequence logo using probabilities instead of information content). (B) Problematic aspects of currently used motif similarity functions. The similarity score of two PWMs decomposes into the sum of similarities of single aligned positions, due to the position-independence assumption in the model. Here we present scores for pairs of positions in DNA motifs by the various similarity functions in addition to a proposed optimal score (all scores are normalized between 1 and -1): The nucleotide distribution in each position is represented schematically (a sequence logo using probabilities). As shown, the Pearson-Correlation does not reflect the true sequence similarity and the Jensen-Shannon divergence (JS) and Euclidean distance do not differ between informative and background uniform positions. Clearly, position 1 should get a higher similarity score than position 2, but the Pearson-Correlation scores for these positions are equal. Position 3 should get the lowest possible score, but Pearson-Correlation does not capture this. Both in positions 1 and 4 identical distributions are compared, but position 1, which should get a higher score, fails to obtain this by all three methods. Positions 4 and 5 should get similar scores, however, Pearson-Correlation grades position 5 significantly lower than position 4.

Chapter 3

A Novel Method for Motif Comparison and Clustering

3.1 A Novel DNA Motif Similarity Score

Our goal is to determine whether two DNA motifs represent the same binding preferences (i.e., they describe binding sites of the same transcription factor). However, when comparing motifs, we have to remember that we wish to differentiate between two motifs with similarity in nucleotide distributions that are specific (e.g. a strong preference for nucleotide A) and two motifs with similar nucleotide distributions that are non-informative (e.g., identical to the background distribution), since the less informative positions in a motif do not contribute to sequence-specific binding of proteins [Yona and Levitt, 2002]. To address this issue, we developed a similarity score that measures the similarity between two DNA motifs, while taking into account their dissimilarity from the background distribution. We now develop the details of the score. Before that we need to clarify how we represent DNA motifs. We can view DNA motifs in two ways. The first, and more common way is as a model that describes the probability of nucleotides in each position of the binding site (see in more detail above). In this work this model is a PWM where the probabilities of nucleotides at different positions are independent of each other. The second view is as the list of sites from which these probabilities were estimated. In this second view we take into account the amount of evidence that we have about the DNA binding preferences. This latter view also allows us to perform statistical evaluation of the motifs. In this view, we assume that each of the binding sites that are presumed to belong to the motif was sampled independently from a common distribution over nucleotides. We assume that this distribution satisfies the position independence properties (in correspondence with the PWM representation). Then, we can evaluate the likelihood

ratio of different source distributions for the sampled binding sites. In practice, we do not keep the actual binding sites, but the sufficient statistics that allow us to evaluate the likelihood of the binding sites. Under the position independence assumption, these statistics are the counts of each nucleotide in each position. Our score is composed of two components: the first measures whether the two motifs were generated from a common distribution, while the second reflects the distance of that common distribution from the background. Statistically, the former component translates to measuring the likelihood-ratio of the two hypotheses:
H_0: The two samples were drawn from a common source distribution.
H_1: The two samples were drawn independently from different source distributions.
The latter component translates to measuring the likelihood-ratio of the two hypotheses:
H_0: The two samples were drawn from a common motif distribution (as above).
H_1: The two samples were drawn from the background distribution.
Our Bayesian Likelihood 2-Components (BLiC) score for motifs m_1 and m_2 is:

$$BLiC_{score} = \log \frac{Pr(m_1, m_2|common \quad source)}{Pr(m_1, m_2|independent \quad source)} + \log \frac{Pr(m_1, m_2|common \quad source)}{Pr(m_1, m_2|background)}$$

An important aspect of this score is that since we assume that all the relevant distributions satisfy position independence, the score decomposes into a sum of local position scores that examine only the distribution of nucleotides at one position in both motifs. More precisely, our likelihood-based score measures the probability of the nucleotide counts in each position of the motif given a source distribution. For two aligned positions in the compared motifs, let n_1 and n_2 be the corresponding count vectors, the similarity score is then:

$$\begin{aligned} BLiC_{score} &= \log \frac{Pr(n_1, n_2|\hat{P_{1,2}})}{Pr(n_1|\hat{P_1})Pr(n_2|\hat{P_2})} + \log \frac{Pr(n_1 + n_2|\hat{P_{1,2}})}{Pr(n_1, n_2|\hat{P_{bg}})} \\ &= \frac{\sum_{y \in A,C,G,T}(n_{1_y} + n_{2_y}) \cdot \log \hat{P_{1,2_y}}}{\sum_{y \in A,C,G,T}(n_{1_y}) \cdot \log \hat{P_{1_y}} + \sum_{y \in A,C,G,T}(n_{2_y}) \cdot \log \hat{P_{2_y}}} + \\ &\qquad \frac{\sum_{y \in A,C,G,T}(n_{1_y} + n_{2_y}) \cdot \log \hat{P_{1,2_y}}}{\sum_{y \in A,C,G,T}(n_{1_y} + n_{2_y}) \cdot \log \hat{P_{1,2_y}}} \end{aligned}$$

Where $\hat{P_1}$, $\hat{P_2}$ and $\hat{P_{1,2}}$ are the estimators for the source distribution of n_1, n_2 and the common source distribution, respectively. And Pbg is the background nucleotide distribution.

3.2 Estimating Distributions

Since the source distribution is unknown, we must estimate it from the nucleotide counts in each position of the PWM. There are alternative approaches towards this goal. The simplest method is to use the maximum likelihood estimator (MLE). For a multinomial distribution, as in our case, estimation using the MLE is very efficient. In addition, this estimator is asymptotically unbiased, i.e., it is assured we will predict the true distribution as the number of samples increases to infinity. However, in the case of DNA motifs, our sample size is far from the required size for calculating the true source distribution. Under these conditions using the MLE is too strict and may lead, for example, to estimations of zero probability of a DNA nucleotide in a certain position of the motif. For this reason it is important to soften our estimation. We use a Bayesian estimation approach, where a priori knowledge, as well as the number of samples, is integrated into the estimation process. We considered two alternative priors. The first is a standard Dirichlet prior [DeGroot, 1970]. The second, more flexible approach, involves a Dirichlet mixture prior [Sjolander et al., 1996], which allows to dynamically choose between several typical distributions. We are using the family of Dirichlet priors because it is conjugate to the multinomial distribution, which enables us to compute the probabilities efficiently. In addition to efficiency considerations, the prior should model the typical distribution of a position in a DNA motif. Using Dirichlet priors is very efficient and has the benefits of Bayesian estimation discussed above. However, using a single component prior does not allow us to model a typical distribution of a position in a DNA motif. In DNA motifs there are several possible typical distributions: informative positions contain positions where the protein has a strong preference for a specific nucleotide: A, c, G or T. In addition, there are non-informative positions in the motif where the protein does not have strong binding preferences. For this reason we chose to use a mixture of priors, which is suitable for representing a complex distribution with more than one typical distribution. More specifically, we used a five-component mixture prior, with four components representing a typical informative distribution, giving high probability for a single DNA nucleotide: A, C, G, or T. The fifth element represents the uniform distribution.

For example: Given the following motif:

	1	2	3
A	5	15	100
C	0	0	0
G	0	0	0
T	0	0	0

Since we are assuming positional independence in the motif, we calculate the source distribution for each position of the motif separately.

Estimator type	5 samples	10 samples	100 samples
Maximum Likelihood Estimator:	(1,0,0,0)	(1,0,0,0)	(1,0,0,0)
Bayesian (Dirichlet prior):	(0.67,0.11,0.11,0.11)	(0.85,0.05,0.05,0.05)	(0.97,0.01,0.01,0.01)
Bayesian (Dirichlet-mix prior):	(0.76,0.08,0.08,0.08)	(0.87,0.04,0.04,0.04)	(0.97,0.01,0.01,0.01)

In this example I used a Dirichlet prior with parameters (1,1,1,1), and Dirichlet mixture prior with uniform weights, where the parameters of the first four components are in the form of : (5,1,1,1) for residue A, etc. and in the form of (2,2,2,2) for the fifth component.

As we can see from this example, the maximum likelihood estimator does not take into consideration the number of samples, and the estimation remains constant and very strict. When the number of samples is small the Bayesian estimation is more flexible since our confidence in the evidence is not high. In addition, The Dirichlet mixture prior is more accurate than the uniform Dirichlet prior, especially when the number of samples is small. This is because we integrated in the Dirichlet mixture prior our prior knowledge on the typical distributions in DNA motifs. As the number of samples grows the differences between the three alternative estimators decreases.

3.2.1 Estimation Details

When using this estimator, the estimated distribution for position n can be calculated as:

$$\hat{p_i} = \frac{n_i}{\sum_{j\in\{A,C,G,T\}} n_j}$$

It is easy to see that when assigning the MLE in the equation of the BLiC score the first component of our score is the known Jensen-Shannon (JS) divergence, which is a similarity measure between two probability vectors based on information theory measure. The JS divergence is the symmetric form of the Kullback-Leibler distance, which is defined for two probabilities P and Q as:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

For probability vectors P, Q and $R = \frac{n_1 P + n_2 Q}{n_1 + n_2}$, the JS divergent is defined as:

$$D_{JS}(P||Q) = \frac{n_1}{n_1+n_2} D_{KL}(P||R) + \frac{n_2}{n_1+n_2} D_{KL}(Q||R)$$

For $\hat{p_{ml}}, \hat{q_{ml}}$ and $\hat{S_{ml}}$ ML estimators for the source distribution of positions n_1, n_2 and the common source distribution, respectively. As we said above the first component of our BLiC score can be represented in the form of a JS divergence:

$$\log \frac{Pr(n_1, n_2|\hat{s})}{Pr(n_1|\hat{p}) Pr(n_2|\hat{q})} = -D_{JS}(P||q)$$

Bayesian estimation using Dirichlet prior: The Dirichlet prior is specified by a set of hyper-parameters $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ and has the form: $Pr(X) = \frac{\Gamma \sum_i \alpha_i}{\prod_i \Gamma \alpha_i} \prod_i x^{\alpha_i - 1}$ For estimating the source distributions with a standard Dirichlet prior, we use uniform hyper-parameters (such as (1,1,1,1)). When using this prior, the estimated distribution for the position n can be calculated as: $\hat{p_i} = \frac{n_i + \alpha_i}{\sum_{j \in \{A,C,G,T} \sum (n_j + \alpha_j)}$ where α is the vector of hyper-parameters. For estimating the source distributions with a five-component mixture of Dirichlet prior [Sjolander et al., 1996], we merge five standard Dirichlet priors using uniform weights. The four components, which represent uni-nucleotide distributions, give high probabilities for a single DNA nucleotide: A, C, G, or T in the hyper-parameters (such as (5,1,1,1) for residue A). The fifth component, which represents the background distribution, is modeled using uniform hyper-parameters (such as (2,2,2,2)). Using this mixture-prior, the estimated source distribution for a count vector n is:

$$\hat{p_i} = \sum_k \left(Pr(\alpha^k | n) \frac{n_i + \alpha_i^k}{\sum_{j \in \{A,C,G,T\}} (n_j + \alpha_j^k)} \right)$$

This estimator is a weighted average of the estimators using each component separately, where the weights are the posterior probability of the component given the data. The posterior reflects our belief that the source distribution in this position of the motif is a certain typical distribution after we are given the vector of counts. The posterior is: $Pr(\alpha^k | n) = \frac{q^k Pr(n | \alpha^k)}{\sum_j q^j Pr(n | \alpha^j)}$

3.2.2 Alignment of Motifs

In the above discussion we assumed that the motifs are aligned. That is, that position 1 in the first motif has the same meaning as position 1 in the second one. In practice, we want to compare two motifs that are not necessarily aligned. Thus, we define the similarity score for two motifs as the score of the best possible alignment between them. Since motifs are short sequences we do not allow gaps in the alignment, and so we only consider the offset of one motif with respect to the other. In addition we consider reverse complement alignment where one motif is complementary to the other (on the opposite DNA strand) Figure 3.1. Since the score decomposes to sum of position scores, we can use a dynamic programming algorithm to find the best scoring alignment between two PWMs (including reverse complement alignments). The unaligned flanks of the motif are scored according to their distance from the background distribution multiplied by a relaxation factor of 0.2.

3.2.3 Assigning P-values to Motif Similarity Scores

To assign the statistical significance of each score, we have devised an empirical p-value estimation, computed for each motif separately. For each motif, we compute the score distribution of alignments with partners from all possible lengths. This is done by comparing the motif to 1000 random motifs of a specified length. These random motifs were

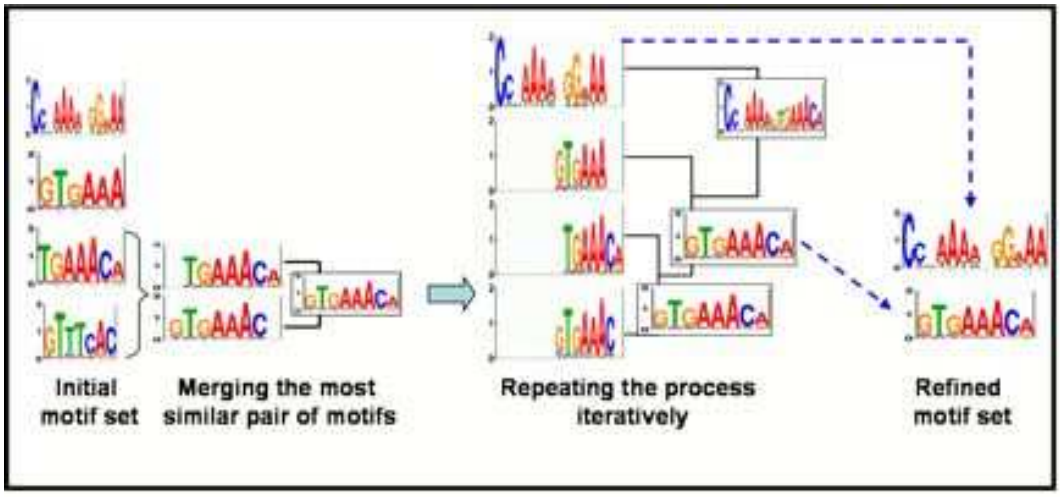

Figure 3.1: **Motif comparison and clustering.** In this short example, the initial set consists of four motifs, which are compared all against all. The score assigned to each pair is the score of the best possible alignment between them (including the reverse complement form). In each step the highest scoring pair is merged into a new motif (by combining the evidence from both motifs). These steps are repeated until we are left with a single motif. The order of merge operations results in a tree, where the leaves are the initial motifs. Each frontier of this tree creates a set of motifs. In this example, a frontier resulting in two motifs is chosen, one is an initial motif and the other is a motif created by merging three initial motifs. These two motifs are the non-redundant set of motifs, derived from the initial set.

generated by sampling positions of known DNA motifs from the TRANSFAC database [Matys et al., 2003]. The TRANSFAC database contains characterized DNA motifs of known transcription factors. By this process we can create random DNA motifs of any length, while using the typical distributions found in transcription factors binding sites. When comparing a given DNA motif to another one, we will use the score distribution of the first motif against the random motifs with the same length as the second motif. We can then assign a p-value to the similarity score by calculating the fraction of random motifs that got the same score or a higher one, which is an approximation for the probability of getting that score or a better one by chance.

3.3 Clustering Motifs

An important application of motif similarity scores is clustering. There are many potential ways of clustering motifs [A. K. Jain, 1988]. Here we consider one of the simplest and straightforward clustering procedures where we combined a similarity score, such as our BLiC score, within a hierarchical agglomerative clustering algorithm. In each iteration of the algorithm we have a set of motifs. The algorithm computes the similarity between

all pairs of motifs and then merges the pair with the highest similarity score into a new motif (Figure 3.1). This merge includes aligning the motifs according to the best scoring alignment between them, and then combining the evidence from both of them. These iterations are repeated until we are left with a single motif. The order of merge operations results in a tree, where the leaves are the initial motifs, and inner nodes correspond to merged motifs that represent all motifs in the relevant sub-tree. Each frontier of this tree stands for a non-redundant clustering of the motifs. We stress that this procedure is different than hierarchical clustering based on the similarities between the initial set of motifs (such as UPGMA(Unweighted Pair Group Method with Arithmetic mean)). Since we merge motifs to create a new one, the similarity of a merged motif to another motif might be different than the average of the similarities of each of the merged motifs to that third motif.

3.3.1 Splitting the Clustering Tree

The clustering tree can be used to distill the set of input motifs into a concise non-redundant group. This is done by splitting the tree into a subset of clusters, each representing a group of redundant motifs. As mentioned above, in this tree, the leaves represent initial motifs, and the inner nodes represent a merging of all motifs in the rooted sub-tree. Thus, to obtain a non-redundant set of motifs, which still covers the initial set, we choose a frontier in the clustering tree. This is done using a bottom-up traversal over the tree. Two adjacent nodes are inserted into the frontier if the ratio between their similarity score and their maximal possible score, is less than a certain threshold (Figure 3.2). After two nodes were inserted to the frontier, usually, additional nodes from their sub-tree should be inserted for consistency. This is demonstrated in Figure 3.2.b, where inserting the top two motifs in the tree to the frontier (separating them to different clusters), derives that the bottom two motif will be separated from the rest of the clustering tree as well.

In this work we use a quite stringent threshold of 60% of the maximal score when splitting the clustering tree. This threshold was chosen as the optimal split threshold compared to hand-curated splits of 10 trees into clusters (with 20 leaves in each).

3.4 Comprehensive Evaluation of Similarity Scores

We set to compare our similarity score to existing ones in the literature. We aim to evaluate scores both in the context of comparing motifs (whether they represent the preferences of the same transcription factor) and clustering motifs. One of the challenges in performing such evaluations is determining the ground truth against which to compare the results. The approach we choose is to generate synthetic datasets where we know the true labeling of motifs. This allows us to benchmark the different procedures, by relating their results with the underlying truth. To make the dataset as realistic as possible, in terms of the properties of binding sites and their preferences, we use predictions of binding sites in real genomic sequences to generate this synthetic dataset. In more detail, we built a library of synthetic motifs where we know the origin of each motif. Each motif is created

by sampling a set of binding sites from the genome-wide catalogue of transcription factor binding locations in S. cerevisiae [Harbison et al., 2004] (see Figure 3.3.a). This dataset simulates the redundant output of motif discovery programs, with similar motifs that arise from different runs over the same transcription factor, as well as partially overlapping motifs that simulate truncated motifs. We compiled a noisy test data of motifs for nine TFs. For each TF, we generated a set of 12 noisy motifs by randomly sampling a subset (of size 5,15 or 35) of its binding site locations, and trimming the original motif by taking only the beginning, end or middle parts. Using these test data we compared different possible similarity scores for DNA motifs. Specifically, we compared the Pearson correlation coefficient; the information-theory based Jensen-Shannon divergence; the Euclidean distance; and our BLiC score.

3.4.1 Motif Comparison Evaluation - Identifying Similar Motifs

We evaluated the sensitivity and specificity of motif similarity scoring methods by comparing pairs of motifs from the test set described above, and testing whether the prediction of close similarity coincide with the true assignment to the pair of motifs, i.e. whether they were generated from the genomic binding locations of the same TF. More precisely, for each pair the significance of the similarity of the first motif to the second was calculated. If the similarity is significant (p-value smaller than a chosen threshold) we label this as a positive pair, and otherwise call it a negative. By comparing this prediction to the true assignment of the motifs (true positive if the two are generated from binding sites of the same transcription factor) we calculated the sensitivity and specificity for each p-value threshold to create an ROC curve for each similarity measure (Figure 3.3.b). Comparing the ROC curves of our score to those of Jensen-Shannon divergence, Euclidean distance and Pearson Correlation coefficient we see that our BLiC score outperformed all other measures throughout the range of possible sensitivity/specificity tradeoffs.

3.4.2 Motif Clustering Evaluation - Reducing the Redundancy

To further evaluate the accuracy of the different similarity scores we clustered the motifs from the test set and examined if motifs originating from the same TF were clustered together. For this, we used the hierarchical agglomerative clustering algorithm described above. The results, based on the 108 noisy motifs for nine different TFs were conclusive. Once again, our two-component score outperformed the other scores, as Figure 3.3.c shows.

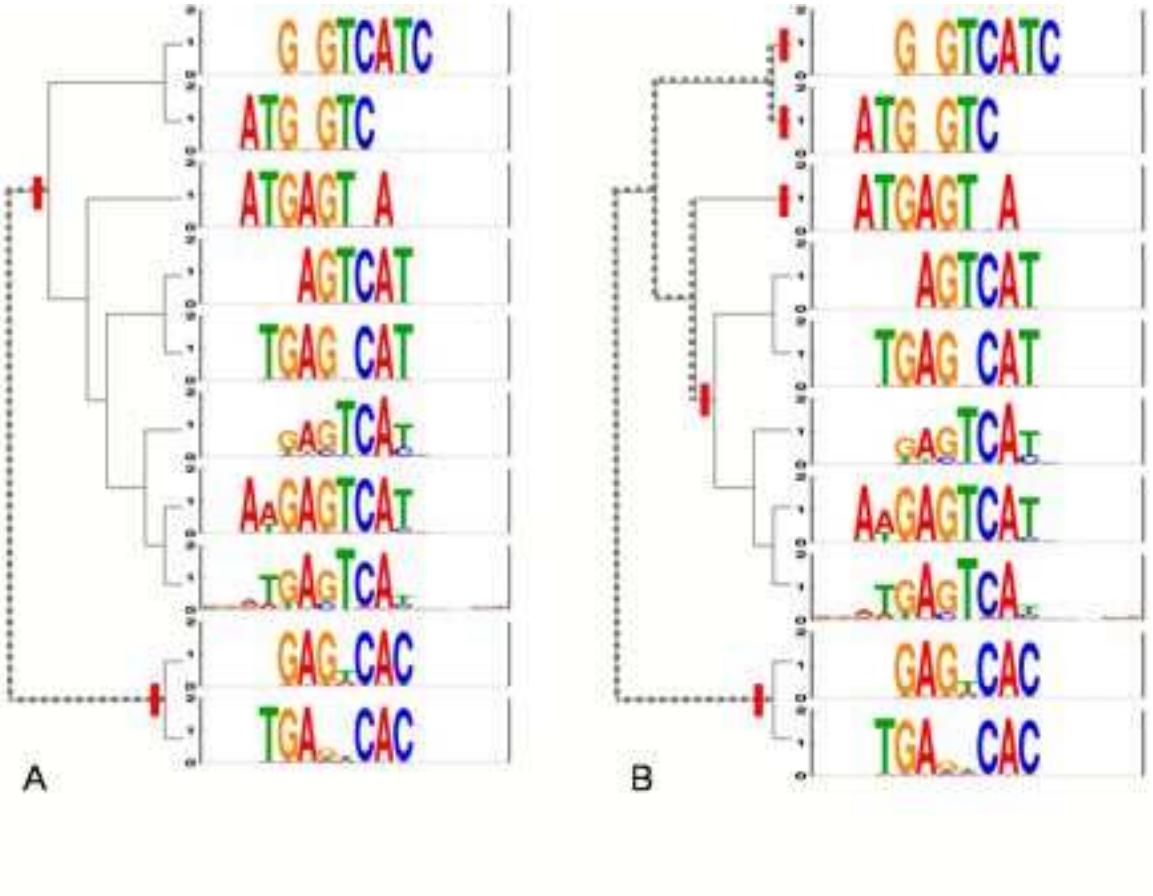

Figure 3.2: **Splitting the clustering tree.** Two alternative splits of the clustering tree. The motifs are discovered in a set of genes found to be bound by the TF Gcn4 under three different environmental conditions in a ChIP-chip assay [Harbison et al., 2004]. We applied six motif discovery algorithms (as in Figure 2.2). The splits are done using a threshold of 50% of the maximal score for these motifs in (a) and 60% of the maximum in (b). Red lines represent the splits of the tree into separate clusters.

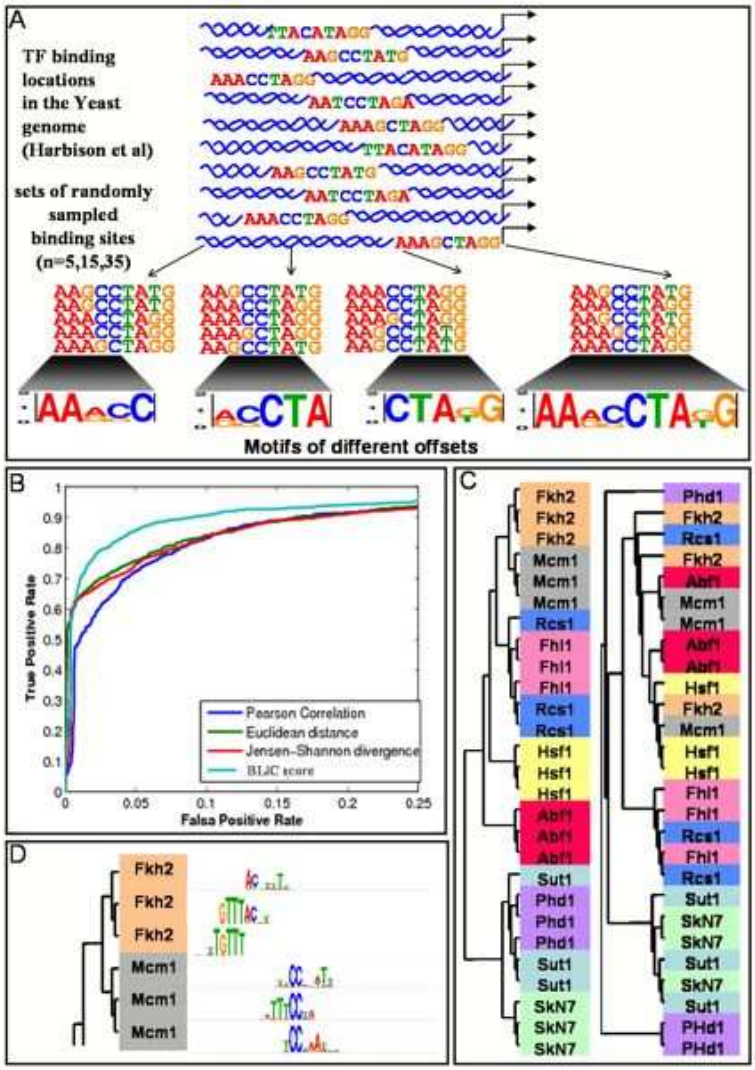

Figure 3.3: **Evaluation of our score.** (A) Generating the test dataset. For a given TF a set of noisy motifs was generated, based on the TF binding locations in the S. cerevisiae genome. First, a subset of the genomic binding site locations was randomly chosen (with varying number of sequences). Second, for several subsets, the length of the original motif was changed by taking only the beginning, end or middle part of the motif. By repeating this procedure a set of noisy motifs was built for each TF. (B) Sensitivity and Specificity of the different scoring methods. Comparing all pairs of motifs from the test set, and assigning an empirical p-value for the score. Motifs generated from the binding sites of the same TF are the true positives. A ROC curve was plotted based on the true positive rate (TPR) and false positive rate (FPR), computed for any choice of p-value threshold. The BLiC score (turquoise) outperformed all other similarity scores: Jensen-Shannon divergence (red), Euclidean distance (green) and Pearson Correlation coefficient (blue). (C) Clustering the test data set using various scores. Here we present the clustering of a partial set of the test data, consisting of all motifs generated from subsets of size 35 with altered lengths. This clustering demonstrates the BLiC score (right) outperforms the Pearson-correlation(left). Similar results show our score outperforms the Euclidean distance and Jensen-Shannon. (D) A more detailed view of the top part of the clustering shown in C, using the BLiC score.

Chapter 4

Large-Scale DNA Motif Analyses

4.1 Analysis Pipeline

Based on our score we developed a three-step method for processing and integrating large-scale data of newly discovered DNA motifs into coherent and reliable sets of non-redundant motifs. The inputs for this procedure are groups of co-regulated DNA sequences. As discussed above, examples for these co-regulated groups are groups of DNA sequences bound by the same transcription factor according to ChIP experiments, or clusters of co-expressed genes from gene expression analysis data. The output for each group of sequences is a set of ranked motifs (Figure 4.1). The advantage of this three-step pipeline is in the accurate and automatic analysis and integration method of DNA motifs. The three steps of the pipeline include:

Step 1: Motif searching and filtering. We begin by applying complementary motif discovery algorithms to each group of sequences. Then, the newly discovered motifs undergo an initial filtration according to their abundance among the group of sequence (see details below).

Step 2: Clustering and merging motifs. The integrated sets of motifs (from all input groups) are clustered and merged to create a non-redundant set. First, the discovered motifs for each group are clustered and merged separately. Then, motifs from all groups are assembled, clustered and merged. After each stage of clustering, a subset of refined motifs is automatically chosen based on the clustering tree (see details below).

Step 3: Ranking and identifying motifs. Finally, the non-redundant set of motifs is ranked and filtered once again, using the abundance of the motifs in the original groups of DNA sequences (see details below). The significant motifs are then coupled with TFs, by comparing them to a known set of DNA motifs from the literature. The output of this analysis is a set of DNA motifs for each TF.

4.2 Analysis Methods

4.2.1 Motif Analysis Pipeline

Motif discovery algorithms. In the analysis pipeline we applied several motif discovery algorithms - Mdscan [Liu et al., 2002], AligneAce [Hughes et al., 2000], MEME [Bailey and Elkan, 1995], MEME$_c$ [Harbison et al., 2004], converge [Harbison et al., 2004], were used through the TAMO package [Gordon et al., 2005], with the default parameters (apart from the MEME algorithm, for which we changed the parameters to output six motifs). We also included conserved and abundant motifs in the yeast genome [Kellis et al., 2003], and the output of the SeedSearcher motif discovery algorithm [Barash, 2005]. All the discovered motifs underwent an initial filtration according to their enrichment among the initial group of sequences, using a hyper-geometric p-value threshold of 1e-5. The hyper-geometric p-value is calculated based on the number of binding sites that match a motif model that can be found within the initial group of genomic sequences as compared to the occurrences in the entire genome (or at least the sequences in all the initial input sets). Since this is only an initial filtration, it is done using an efficient scan for motif matches where a sequence is considered a match to a motif if it had a score of at least 60% of the motif maximum. This is done using the TAMO package [Gordon et al., 2005]. All motifs were converted to a PWM representation, clustered and merged as described in section 3.

Truncating motifs. Uninformative positions at the two edges of motifs were truncated automatically. This was done by testing the null hypothesis that the nucleotides at a motif position distribute according to background distribution. The hypothesis was tested using a chi-square test with a p-value threshold of 0.05.

Ranking and filtering motifs. To score the motifs at the third step of the pipeline, we scanned the entire genome (or set of promoters) using each motif (see below), finding all the set of occurrences of each motif in the genome. Then the enrichment of each motif relative to the input groups of DNA sequences was computed. The statistical significance of the enrichment was evaluated by a hyper-geometric p-value. We filtered the motifs according to a threshold of 1e-3 after applying a Bonferroni correction for multiple hypotheses. We then ranked the motifs by their enrichment in the input groups, assigning enrichment score to each motif as the -log of the hyper-geometric p-value.

Identifying the Motifs. To connect between the discovered motifs and known transcription factor binding specificities, we used our motif comparison method against databases of known motifs (TRANSFAC [Matys et al., 2003], SCPD [Zhu and Zhang, 1999], YPD [Csank et al., 2002]).

4.2.2 Genomic scan

Estimating the motif probabilities from count matrices .To scan the genome with our motifs, we first transferred them from PWMs (count matrices) to profile representation (frequencies). This was done using a Bayesian estimator with the Dirichlet-mixture prior described in section 3.

Identifying binding sites. Finding all the genomic locations of a motif was done using the TestMotif program [Barash et al., 2005] combined with evolutionary conservation data. Particularly, we decide whether a DNA sequence contains a motif if one of three following criteria holds:

- The sequence contains a highly significant binding site (a good sequence match between the motif and the binding site). For this, we used a p-value threshold of 0.01 (after applying a Bonferroni correction for multiple hypotheses according to the average length of the scanned sequences).
- The sequence contains a less significant occurrence of the motif (p-value threshold of 0.1), which is highly conserved among seven species of the genus Saccharomyces. For this, we used the average conservation of the motif, according to the UCSC conservation track, with a conservation threshold of 0.8. (phastCons [Siepel et al., 2005], through the UCSC Genome Browser Database [Karolchik et al., 2003]).
- The sequence may contain a less significant occurrence of the motif but we have high confidence that the sequence contains a motif based on the entire sequence. This criterion is composed of two factors. The first is the Bayesian posterior probability of a motif in any position in the sequence. Here not only strong instances of the motif will indicate this promoter as being a target, but also a few weaker instances of the motif. We require this probability to be at least 0.1. The second factor is the conditional posterior probability of finding a binding site at this specific location, if we know that there is a motif somewhere in the sequence. Here the p-value will be in correlation to the degree of sequence match. We require this probability to be at least 0.5.

For example, scanning all the S.cerevisae promoters with the following variant of the Sko1 motif:

In the output set of targets, we find instances of the motif according to each one of the three criteria. For instance:

1. The sequence TTACGTAATGG has high sequence similarity with a p-value of 1.5e-06.

2. The sequence CTGCGTAAAGG has quite low sequence similarity (p-value 0.07), however the average conservation of the motif is very high and is 8.9.

3. The sequence TCACGTAAAGG has a lower sequence similarity than the first sequence (p-value 0.01), however the probability of the entire promoter to be regulated is 0.14 and the probability of finding a motif in this specific position, assuming the promoter is regulated, is 0.99.

Parameter tuning. The threshold values listed above were chosen according to an extensive search of parameters that maximize the true positive rate, allowing up to 2% false positive calls. This optimization was based on location analysis data of Gcn4 [Harbison et al., 2004], and location and expression data for Sko1 (unpublished data).

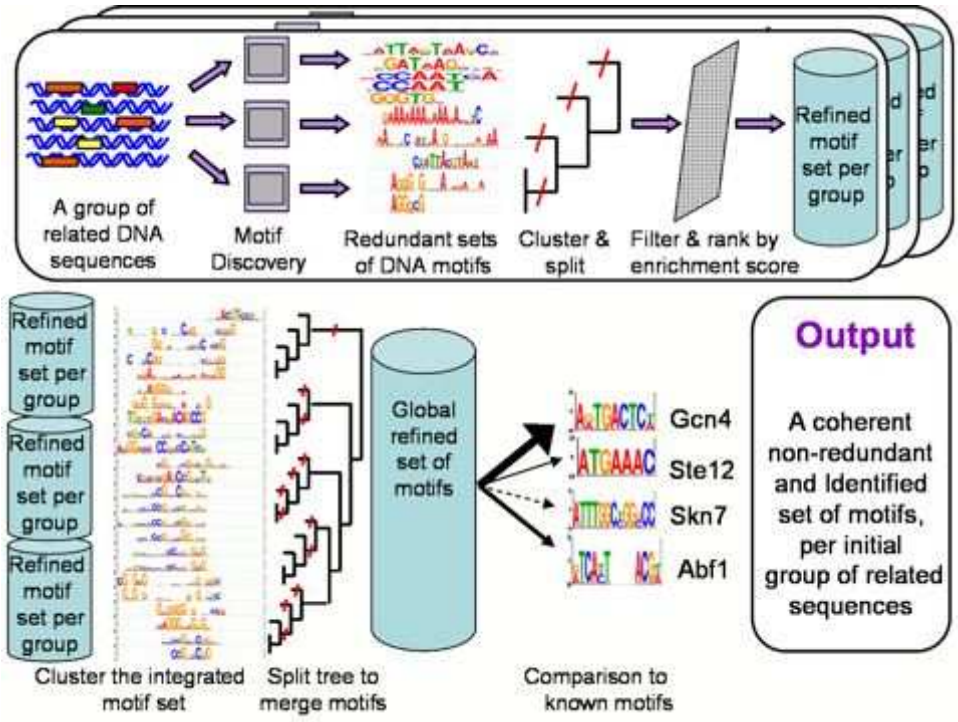

Figure 4.1: **Motif analysis pipeline overview.** Based on our new score we developed a large-scale analysis method of DNA motifs, which gets as an input groups of related sequences and outputs a set of ranked motifs for each group. The first step of the pipeline is searching for motifs in each group of DNA sequences, using complementary motif discovery algorithms. The second step is reducing the redundancy in the newly discovered set of motifs, which is done by clustering and merging the similar motifs. The clustering is done separately within each group and finally the entire set of motifs are clustered and merged. The merging of motifs is done automatically as part of the clustering procedure. The third step of the pipeline is ranking the motifs and identifying their binding factors. The refined motif set is ranked and filtered according to the enrichment score (-log10 of the hyper-geometric p-value in the relevant group of DNA sequences). The motif binding factors are identified by comparing the discovered motifs to the set of known DNA motifs in the literature using the BLiC score. The output of this pipeline can be used in additional analyses such as applying advanced clustering methods and integrating additional sources of information.

Chapter 5

Biological Results

5.1 Yeast Transcription Map

We utilized our pipeline to understand how TFs alter their DNA binding pattern under various environmental conditions. To this end, we applied our DNA motif analysis pipeline to genome-wide ChIP-chip data of 177 TFs under several environmental conditions, a total of 301 experiments for different TFs and conditions [Harbison et al., 2004]. Initially, we used seven motif discovery algorithms to produce a redundant set of motifs for each ChIP experiment (as detailed above). In the second step of the pipeline we clustered the motifs - first the motifs discovered for each TF under each environmental condition were clustered separately, then the (merged) motifs for each TF under all conditions, and finally the entire set of motifs. The motifs were ranked according to their enrichment in the related sets of genes bound by the different TFs. The binding motifs were then compared to the known motifs of the TFs, based on previous information (TRANSFAC [Matys et al., 2003], SCPD [Zhu and Zhang, 1999], YPD [Csank et al., 2002]). This resulted in a concise set of DNA motifs attributed to each TF under each environmental condition (all the motifs sets can be found at www.cs.huji.ac.il/ naomih/conditional_map.html). To further analyze the DNA motifs learned from the entire ChIP data, we used EdgeCluster - a clustering algorithm recently developed in our lab [unpublished results, Hebrew University, 2007]. The novelty in EdgeCluster is in the integration of various sources of information into the clustering process, including pair-wise information. Specifically, we used for each motif data from three different sources. We calculated the motif's enrichment in different groups of genes: the original ChIP data, groups based on functional annotations [Harris et al., 2004], and groups of genes which are up or down regulated according to gene expression data [Segal et al., 2003]. The pair-wise information we used was inter-motif similarity scores (using our BLiC score). In addition, as an input to the algorithm we used an initial partition of the motifs into clusters according to our hierarchical clustering algorithm, which is based only on the sequence similarity of the motifs. The initial partition was done by applying our hierarchical-clustering algorithm with a highly permissive threshold on splitting the tree into clusters (of 0.3). Figure 5.1 demonstrates the

clustering for all the motifs. As we can see groups of similar motifs are grouped together.

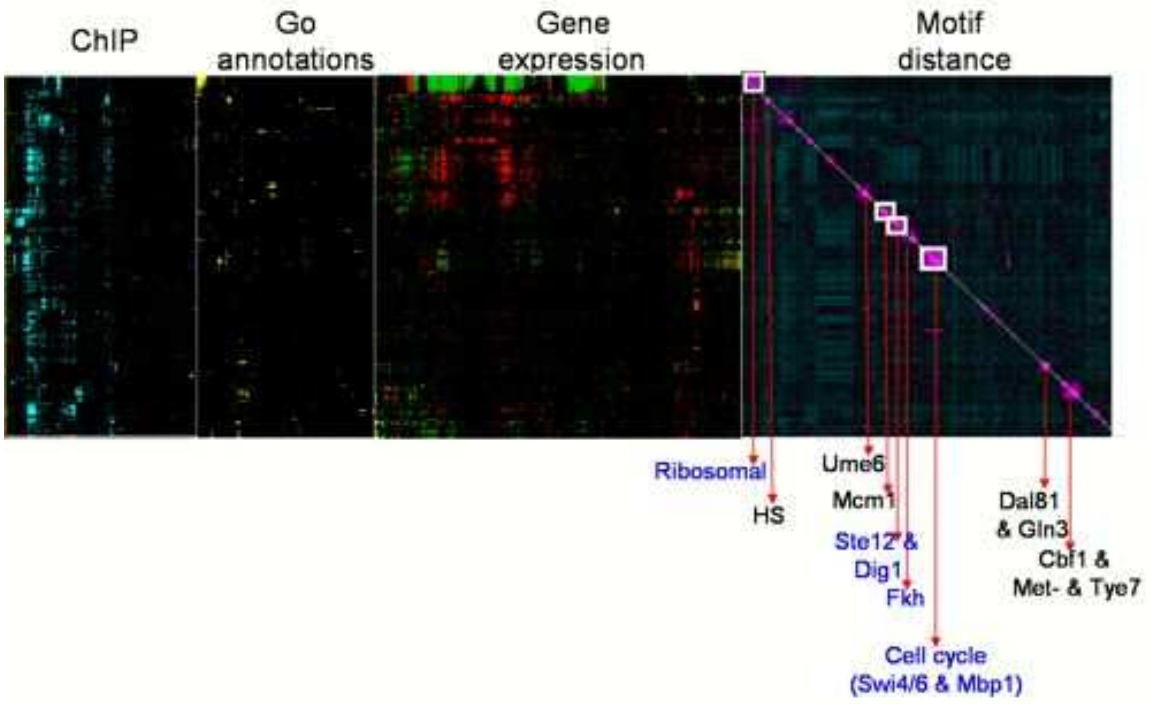

Figure 5.1: **Results overview.**The output of the EdgeCluster algorithm on the set of discovered motifs. The Clustering is based on several types of data: enrichment of each motif in different groups of genes: the original ChIP data, functional annotations [Harris et al., 2004], groups of genes up or down regulated according to gene expression data [Segal et al., 2003], and pair-wise information of inter-motif similarity scores (using our BLiC score).

5.1.1 Comparison to Previous Work

In the work of Harbison et al. (2004) and MacIsaac et al. (2006), the same ChIP data was used to construct a global transcriptional regulatory map in yeast. The motif analyses performed in these two works differ from ours, both in the similarity score used (the Euclidean distance), as well as by applying different motif clustering and merging methods. In addition, the output of these two works was a single motif for each TF, chosen based on the motifs enrichment score and its similarity to the known recognition sequence (when available). We should first note that our motif set might contain several motifs for a single ChIP experiment (TF and condition) - different variants for the binding preferences of that TF, as well as additional motifs for other TFs that interact with it. To be consistent with these previous works in the comparison, we narrow down our set to a single motif. We do that in two different manners. We compared our motifs to the ones learned by these

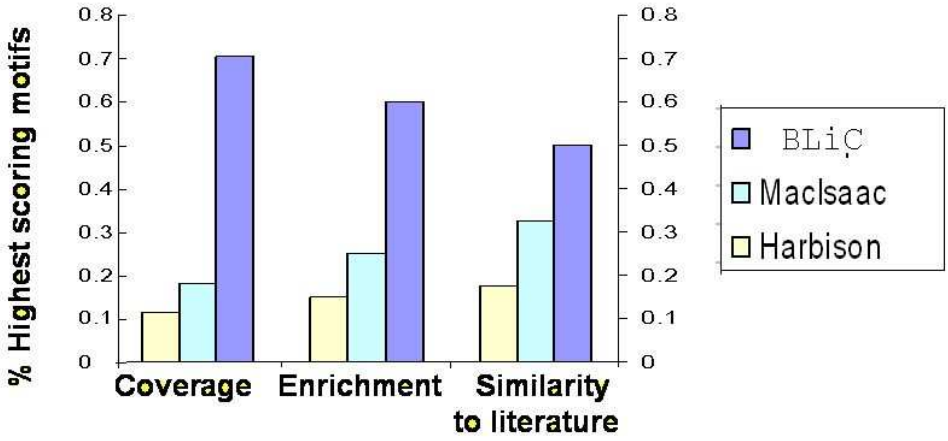

Figure 5.2: **Comparison to previous analysis methods**. Comparison between DNA motifs, of TFs with previously known binding motifs from the literature, discovered by us to the motifs discovered by Harbison et al. and MacIsaac et al. The motifs are compared by three parameters and the fraction of motifs which got the highest score among the three motif sets is presented. The first parameter compared, is the similarity to the known motif from the literature. The similarity was measured using the BLiC score (from our set of motifs, for each TF, the motif most similar to the literature was used). The second and third parameters tested are the enrichment score and the percentage of coverage of the motifs in the relevant chromatin imunoperticipation group (from our set of motifs, for each TF, the top enriched motif which is similar to the known motif from the literature was used for this comparison).

two methods based on similarity to known motifs from the literature (TRANSFAC [Matys et al., 2003], SCPD [Zhu and Zhang, 1999], YPD [Csank et al., 2002]). To choose a single motif for each TF, we considered all motifs in the TF s motif set, and picked the motif that is most similar to the known recognition element (as done in these previous works). We then compared the similarity between the known motifs and the discovered motifs (by all methods). In 50% of the cases, our motifs were found to have the highest similarity to the known motifs. The motifs learned by the algorithms of MacIsaac et al and Harbison et al, had the highest similarity in 32.5% and 17.5% of the motifs, respectively. We further tested the motifs discovered for TFs with known and unknown binding preferences by comparing the motifs based on their enrichments in the ChIP groups of sequences. In this scenario, we have chosen the most significant motif for each TF, similarly to what was done in the previous methods. We scanned for putative target sequences of each motif as described above, and then compared the enrichment (hyper-geometric p-value) of the motif among the bound genes (using ChIP data for the same TF and condition). The same procedure and parameters were applied for motifs from all three methods. Our motifs were found to have higher enrichments in 60% of the cases, see Figure 5.2).

5.2 Elucidating Transcription Factors Conditional Binding

Using the motif sets we have learned, we next turned to examine the change in the binding specificities of the TFs under different conditions, and its effect on the set of targets. Under different conditions a TF may either bind the same targets (condition-independent), or it may change its set of targets from condition to condition (condition-dependent). When changing conditions, such a regulator may expand its targets in addition to the ones it already binds, it may bind to a different set of targets, or it may even not bind any targets at all. Various mechanisms may be involved in monitoring the condition-dependent binding. One possible mechanism regards a change in the dosage of active TF in the nucleus, which may change the number of targets it can bind [Harbison et al., 2004]. Another possible mechanism involves changing the TFs DNA binding specificities. This may be caused by post-translational modifications of the TF or cofactor binding, resulting in variations of the TF recognition site. In addition, when a TF does not bind the DNA on its own, a change in the protein binding partner may be the cause for the altered bound targets, which may be detected through co-occurrence of DNA recognition sites of different TFs. Also, a change of targets may be caused by a change in the accessibility of the binding site due to a modified chromatin state. However, in this case there is no change in the motifs recognition site on the DNA. As stated above, we derived a set of motif variants for each TF at every condition. By analyzing these motif sets, we gain insights into the mechanism through which a TF changes the DNA targets it binds to, either by a change in its DNA binding specificities (different variants of motifs), or by binding of a co-factor (co-occurrence of motifs). Out of the 72 TFs for which ChIP-chip experiments were carried out in more than one condition, 32 TFs alter their target genes between two conditions (in total, 65 pairs of differential conditions). In 27 of these pairs we did not find significant motifs in at least one of the compared conditions and thus could not search for differential motifs. Finding a motif only on one condition could be meaningful on its own, since this may indicate that there is no direct binding of the factor to the DNA. On the other hand it could results from technical reasons, such as noise in the input set of sequences, and thus in this work we do not analyze these cases. Out of the remaining 38 pairs (spanned over 21 different TFs), we found differential motifs in 89% of the pairs (34 cases spanned over 19 TFs) with a p-value of less than 0.05.

5.2.1 Testing for Differential Motifs

We define a TF as altering its target genes between two conditions, if the number of target genes in the intersection is less than half the number in each condition separately. In addition, we consider only TFs with at least 20 target genes in each of the two conditions (a sufficient number for motif discovery purposes). To define a differential motif, we looked for motifs that are enriched among the targets of a TF at one condition, but not in the other (excluding the genes in the intersection). This was calculated using a chi-square test, with a p-value threshold of 0.05

5.2.2 Condition-Dependent Binding of Ste12 Under Conditions of Mating and Filamentous Growth

Ste12 provides a known example of a TF that shows condition-dependent binding. This TF activates genes in two alternative pathways - the mating pathway and filamentous growth pathway [Zeitlinger et al., 2003, Chou et al., 2006] (Figure 5.3.a). Under filamentous growth signaling (induced by Butanol) we find that Ste12 binds to genes whose promoters are enriched with its known recognition sequence [Madhani and Fink, 1997], as well as the known recognition sequence of Tec1 [Madhani and Fink, 1997], a co-factor of Ste12 under filamentous growth [Chou et al., 2004, 2006](Figure 5.3.b). Nevertheless, under mating signaling (induced by Alpha factor) we find that Ste12 binds promoters enriched with another variant of its recognition sequence - a near-perfect tandem repeat of its known site. This motif variant suggests that Ste12 acts as a dimer following Alpha factor induction, as was previously suggested [Schaber et al., 2006, Wang and Dohlman, 2006](Figure 5.3.b). Interestingly, the exact same motifs were learned for Dig1 - a cofactor that apparently does not bind the DNA directly, but is essential for the binding of Ste12 .An additional player found in our analysis is the TF Mcm1. We found its known recognition sequence [Gelli, 2002] enriched among promoters bound by Ste12, both in mating and filamentous growth, consistent with previous knowledge on the role that Mcm1 plays in expression inhibition of mating genes in diploid cells [Gelli, 2002]. We speculate that Mcm1 plays a similar role in the filamentous growth pathway. While haploid cells undergo invasive growth, diploid cells undergo pseudohyphal growth. Thus, using only the motif sets we discovered, we can track a transcription factor altered DNA binding pattern, caused by a change in the DNA binding partner when the environmental conditions is changed.

5.2.3 Condition-Dependent Binding of the Iron-Regulated Factor AFT2

Another interesting example is provided by the iron-regulated transcription factor Aft2, required for iron homeostasis and resistance to oxidative stress [Courel et al., 2005]. This TF exhibits a significant environmental-dependent binding, switching targets between low and high H2O2 conditions (Figure 5.4.a). The role of Aft2 in iron homeostasis and resistance to oxidative stress is poorly understood. In low H2O2, we find that Aft2-bound promoters are highly enriched with a motif similar to the known recognition sequence of Aft2 (GgGTG) [Courel et al., 2005]. However, in high H2O2 we find abundant low-complexity repeats of Poly-GT (Figure 5.4.b). This result indicates that the DNA binding specificity of Aft2 changes over these conditions, a possible explanation for the change in its DNA targets. We can further speculate that the repeated poly-GT motifs under high H2O2 may suggest that Aft2 binds the DNA as a homodimer. However, we could not support this speculation with experimental data. Interestingly, we do find a motif similar to the known recognition sequence of Aft1 (Rcs1)[Courel et al., 2005], a paralog of Aft2, enriched among the Aft2-bound promoters in low H2O2 condition. This implies a possi-

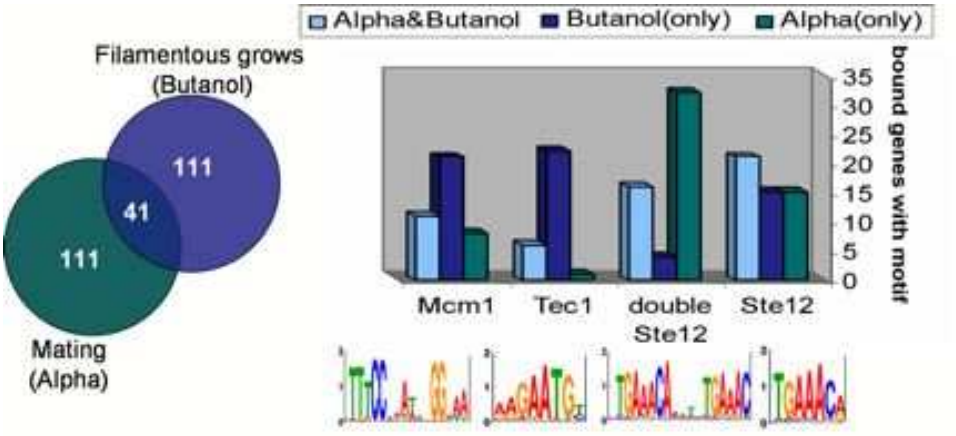

Figure 5.3: **Ste12.**(A) A diagram representing the results of the ChIP-chip experiment [Harbison et al., 2004] for the Ste12 under mating (induced by alpha factor) and filamentous grows (induced by butanol). As it is demonstrated Ste12 alters it's targets significantly between these two conditions. (B) From the motif analysis we see that under filamentous growth signaling we find enrichment for a motif similar to the previously characterized Ste12 motif, as well as the known recognition sequence of Tec1, which is a known co-factor of Ste12 under filamentous growth. Under mating signaling we find a near-perfect tandem repeat of Ste12 known binding site. This motif variant suggests that Ste12 acts as a dimmer in mating. A motif similar to the known Mcm1 motif is found to be enriched under both conditions, especially under filamentous grows. This is consistent with the known role of Mcm1 as an inhibitor of mating in diploid cells.

ble overlap between the targets of Aft2 and Aft1, which is indeed supported by ChIP-chip data of the two TFs (Figure 5.4.b). Based on our analysis, we report two similar (but not identical motifs) for the two paralogs (as suggested by Courel et al. 2005. Rutherford, et al. 2001). Since it is known that Aft2 and Aft1 have independent and partially redundant roles in iron regulation [Rutherford et al., 2001, Courel et al., 2005], we assume that Aft2 DNA binding does not depend on Aft1 and the change in Aft2 targets is due to a change in its specificity to the DNA. The ChIP data and our motif analysis suggest that under high H2O2 conditions Aft2 has a unique role in gene regulation. Here again, using only the motif sets, a transcription factor altered DNA binding pattern was elucidated, caused by a change in its DNA specificity when the environmental conditions have changed.

5.2.4 Condition-Independent Example

As opposed to the cases presented above, the motif sets learned for several TFs remained constant under different environmental conditions. For example, the condition-independent TF Fhl1 is a master regulator of ribosomal genes (Figure 5.5.a). As expected, we find similar sets of motifs enriched in all the conditions tested (Figure 5.3.b), and the most highly enriched motif is similar to the previously known Fhl1 binding motif. This is

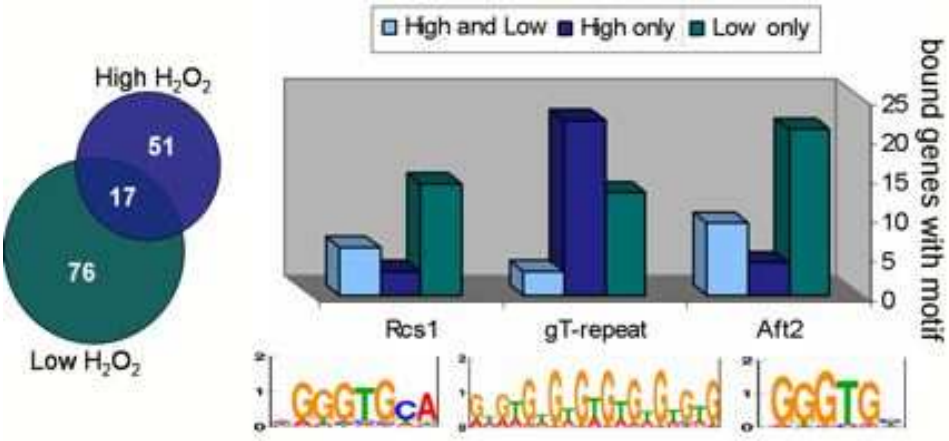

Figure 5.4: **Aft2.**(A) A diagram representing the results of the ChIP-chip experiment [Harbison et al., 2004] for the TF Aft2 under high and low H2O2 stress. As it is demonstrated Aft2 alters it's targets significantly between these two conditions. (B) From the motif analysis we see that under low H2O2 stress we find enrichment for a motif similar to the previously characterized Aft2 motif, as well as the known recognition sequence of Rcs1, a paralog of Aft2 which is known to have an independent and partially redundant role as Aft2 in iron homeostasis. Under high H2O2 stress we only find abundant low-complexity repeats of Poly-GT. This indicates that the DNA binding specificity of Aft2 changes over these conditions, and may suggest that Aft2 binds the DNA as a homodimer

consistent with previous studies that show that Fhl1 remains bound to its target genes, and the environmental dependant regulation of the ribosomal genes is determined by the two cofactors IFH1 (a coactivator) and CRF1 (a corepressor) that bind to Fhl1[Martin et al., 2004].

These results suggest that our two-component score, and the motif analysis pipeline, improve the specificity of the discovered motifs. In addition, our sets of motifs, assigned to each TF in each condition, contain additional information, which is important for understanding the mechanisms of transcriptional regulation. To demonstrate that, we have applied our motif discovery pipeline on transcription factors which alter their behavior under environmental changes.

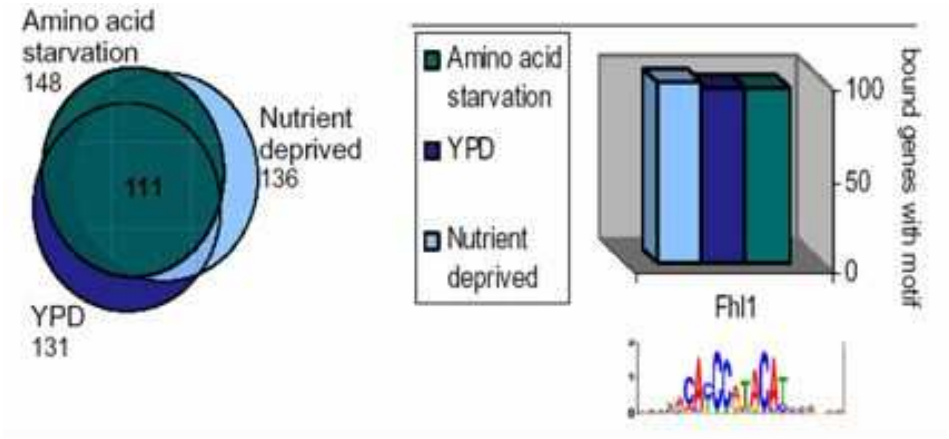

Figure 5.5: **Fhl1.**(A) A diagram representing the results of the ChIP-chip experiment [Harbison et al., 2004] for the TF Fhl1 under YPD conditions, amino-acid starvation and nutrient deproved conditions. As it is demonstrated the targets of Fhl1 remain stable under these changing environments. (C) From the motif analysis we see that under all conditions a motif similar to the known Fhl1 motif is found to be highly enriched under all conditions.

Chapter 6

Discussion

Building maps of transcription regulation requires comparison of DNA motifs. An accurate motif comparison method is important for clustering redundant DNA motifs into coherent groups and for connecting the discovered motifs to previously characterized motifs of known TFs. In this study we present a novel similarity score, the BLiC score, based on Bayesian probabilistic principles. Our score reflects the similarities between transcription factors binding preferences, while taking into account not only the similarity in positional nucleotide distributions of the two motifs but also their dissimilarity to the background distribution. We use the new comparison method as a basis for motif clustering and retrieval procedures, and compare it to several commonly used alternatives. This comparison shows that our BLiC score is more accurate than other possible scores, and improves the specificity and sensitivity of motif comparisons and clustering tasks. The resulting motif clustering and retrieval procedures are incorporated in a large-scale automated pipeline for analyzing DNA motifs, which integrates the output of various DNA motif discovery algorithms and automatically merges redundant motifs from multiple training sets. The output of our pipeline is a coherent annotated library of motifs. Application of this pipeline to genome-wide location data of transcription factors in S. cerevisiae, successfully identified DNA motifs in a manner that is as good as semi-automated analyses reported in the literature. Moreover, we demonstrate how motif analysis can lead to insights into regulatory mechanisms. More specifically we elucidate mechanisms of transcription factor condition-specific binding, by focusing our motif analysis on transcription factors that alter their targets as a response to changes in their environmental conditions, and by searching for differential motifs for these TFs.

Hierarchical agglomerative clustering We used our BLiC score to develop a hierarchical agglomerative clustering algorithm for merging similar motifs. Clustering the motifs hierarchically ensures that the motifs within every sub-tree are properly aligned. Furthermore, such an approach allows us to trim the cluster tree at various heights, thus splitting the motif library into different numbers of non-redundant groups, depending on the requested resolution. In addition, the inner nodes in the tree are created along the run of the algorithm by aligning and merging all motifs in the relevant subtree. Such possibil-

ities are not always available in alternative clustering methods. For instance, the popular k-means clustering algorithm is based on a fixed number of clusters, which is usually not known in advance. Choosing a wrong number of clusters, might lead to either a redundant set of clusters, or to mis-aligned, mixed clusters. Moreover, in k-means clustering there is not necessarily a good alignment for all the motifs within a cluster, thus merging similar motifs cannot be done automatically from the clustering itself.

Motif analysis We developed a motif analysis pipeline based on our BLiC score and the hierarchical agglomerative clustering, designed to process discovered DNA motifs into a set of non-redundant, identified motifs. As we have shown, such an approach improves the sensitivity and specificity of standard motif discovery outputs. By automating all the analyses (including the trimming of cluster trees into discrete sets of motifs), we enable the analysis of hundreds of input groups. In addition, we achieve a wider view on transcription regulation by running several motif discovery algorithms in parallel, and integrating their outputs. By comparing motifs from different input groups we are able to connect between transcription factors that play a role in different processes. In our analysis, we assigned a set of motifs for each input group of genes, and showed that for many input groups, a set of non-redundant motifs captures the regulatory function of the input genes better than a single DNA motif. Many of these cases include TFs that work cooperatively with other TFs (e.g. Ste12). The regulatory mechanism is captured through a set of motifs related to all the involved factors. In addition, some TFs change their binding specificity under different conditions, as we suggest here for the TF Aft2. For these cases, several DNA motifs better capture the DNA binding preferences of the TF than a single motif.

From DNA motifs to regulatory mechanisms Sequence information is a highly accessible resource, and thus it is interesting to ask what can we learn from sequence information alone on transcription regulation? We demonstrated in this study that examining DNA motifs elucidate the regulation mechanisms of transcription factors. We show that motifs can give an indication for the mechanism involved in altered DNA binding, in cases where it involves a change in the TFs specificity to the DNA or its binding partner, as we discussed thoroughly for the TFs Ste12 and Aft2. In addition we examined all the TFs that alter their binding in response to a change in their environment (according to the ChIP-chip data), and found a differential motif for 89% of the TFs (32 TFs). A differential motif is a motif that is over-represented in the set of targets bound by the TF in one condition but not in the other. These differential motifs can point to the cause of the altered DNA binding. An additional important factor affecting DNA binding, which we have not discussed here, is the dosage of the active transcription factors in the nucleus. It has been previously suggested [Harbison et al., 2004] that the dosage of the TF can be inferred as well from DNA motif analysis, by examining the similarity of the bound motif to the consensus sequence. The rational behind this is that when the concentration of the protein is low it will bind sites similar to the consensus since it has a higher affinity for them.

Still, motif analysis obviously does not reveal the whole picture. For instance, we can learn from the motifs if a TF changes its specificity to the DNA, but the cause of that change in specificity still remain unknown. This cause could be, for example, a modification of the protein or binding of a cofactor that does not bind the DNA. In addition, other regulatory mechanisms, such as chromatin remodeling mediated regulation, cannot be inferred from motif analysis. Thus, for a complete understanding of the regulatory mechanisms additional information such as nucleosome positions and dynamics is needed. A significant limitation of motif analysis is that an instance of the DNA motif in the genome is not a sufficient indication for binding of a TF and even less an indication for its activity. There are several methodologies trying to overcome this obstacle, none of which solves the problem completely. A common approach is to consider only conserved instances of motifs eg. [Harbison et al., 2004], since functional motifs are under evolutionary constraints. This reduces the false positives, but may lead to loss of functional sites since the regulatory program undergoes rapid evolution compared to coding sequences [Tautz, 2000, Moses et al., 2006, Levine and Tjian, 2003]. Another possibility is to add information, which may separate between functional and non-functional sites (specific for each TF), such as the distance from the transcription start site, co-occurrence of motifs and more. Our inability to differentiate between functional and nonfunctional motifs raises the question addressed many times before [Barash et al., 2003], if our representation of transcription factor binding preferences is sufficiently accurate? An efficient approach for reducing this noise in motif analysis could be to use additional biological data narrowing down the motif search to certain regions in the genome. We based our work on location data from low resolution arrays which focused mainly on promoter regions. Using genome wide arrays with increased resolution can help point out the genomic bound regions. In work in progress we are using our motif analysis pipeline to analyze data from such arrays.

In this study we overcome a basic obstacle in DNA motif analysis, by developing an accurate motif comparison method. Our motif analysis pipeline, which includes clustering and retrieval procedures based on our novel score, is fully automated and produces accurate results. This is highly important in large-scale analysis, such as that reported here. We showed here the power of motif analyses, which are very useful not only to building regulatory maps, but also for understanding more profoundly regulatory mechanisms.

Bibliography

K. M. Mohiuddin A. K. Jain, J. Mao. Algorithms for clustering data. *Upper Saddle River (New Jersy): Prentice-Hall Callege Division*, (320p), 1988.

T. L. Bailey and C. Elkan. The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol*, 3:21–29, 1995.

Y. Barash. Unified Models for Regulatory Mechanisms. *PhD thesis, Hebrew University, Jerusalem, Israel*, 2005.

Y. Barash, G. Elidan, T. Kaplan, and N. Friedman. Modeling Dependencies in Protein-DNA Binding Sites. *Proc. of the 7th Ann. Int. Conf. in Comp. Mol. Bio. (RECOMB)*, 2003.

Y. Barash, G. Elidan, T. Kaplan, and N. Friedman. CIS: compound importance sampling method for protein-DNA binding site p-value estimation. *Bioinformatics*, 21(5):596–600, Mar 2005.

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, Oct 2002.

M. L. Bulyk, P. L. 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, Mar 2002.

S. Chou, L. Huang, and H. Liu. Fus3-regulated Tec1 degradation through SCFCdc4 determines MAPK signaling specificity during mating in yeast. *Cell*, 119(7):981–990, Dec 2004.

S. Chou, S. Lane, and H. Liu. Regulation of mating and filamentation genes by two distinct Ste12 complexes in Saccharomyces cerevisiae. *Mol Cell Biol*, 26(13):4794–4805, Jul 2006.

M. Courel, S. Lallet, J. M. Camadro, and P. L. Blaiseau. Direct activation of genes involved in intracellular iron use by the yeast iron-responsive transcription factor Aft2 without its paralog Aft1. *Mol Cell Biol*, 25(15):6760–6771, Aug 2005.

T. M. Cover and J. A. Thomas. Elements of Information Theory. *City College of New York*, 2001.

C. Csank, M. C. Costanzo, J. Hirschman, P. Hodges, J. E. Kranz, M. Mangan, K. O'Neill, L. S. Robertson, M. S. Skrzypek, J. Brooks, and J. I. Garrels. Three yeast proteome databases: YPD, PombePD, and CalPD (MycoPathPD). *Methods Enzymol*, 350:347–373, 2002.

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

M. H. DeGroot. Optimal Satistical Decisions. *McGraw-Hill,New York*, 1970.

A. Gelli. Rst1 and Rst2 are required for the a/alpha diploid cell type in yeast. *Mol Microbiol*, 46(3):845–854, Nov 2002.

D. B. Gordon, L. Nekludova, S. McCallum, and E. Fraenkel. TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. *Bioinformatics*, 21(14):3164–3165, Jul 2005.

C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J. B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, Sep 2004.

M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, and R. White. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, 32 (Database issue):D258–D261, Jan 2004.

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, Mar 2000.

T. Kaplan, N. Friedman, and H. Margalit. Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput Biol*, 1(1):e1, Jun 2005.

D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, and W. J. Kent. The UCSC Genome Browser Database. *Nucleic Acids Res*, 31(1):51–54, Jan 2003.

M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937): 241–254, May 2003.

M. Levine and R. Tjian. Transcription regulation and animal diversity. *Nature*, 424(6945): 147–151, Jul 2003.

X. S. Liu, D. L. Brutlag, and J. S. Liu. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol*, 20(8):835–839, Aug 2002.

K. D. MacIsaac and E. Fraenkel. Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Comput Biol*, 2(4):e36, Apr 2006.

H. D. Madhani and G. R. Fink. Combinatorial control required for the specificity of yeast MAPK signaling. *Science*, 275(5304):1314–1317, Feb 1997.

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, Jun 2001.

D. E. Martin, A. Soulard, and M. N. Hall. TOR regulates ribosomal protein gene expression via PKA and the Forkhead transcription factor FHL1. *Cell*, 119(7):969–979, Dec 2004.

V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*, 31(1):374–378, Jan 2003.

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

A. M. Moses, D. Y. Chiang, and M. B. Eisen. Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pac Symp Biocomput*, pages 324–335, 2004.

A. M. Moses, D. A. Pollard, D. A. Nix, V. N. Iyer, X. Y. Li, M. D. Biggin, and M. B. Eisen. Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput Biol*, 2(10):e130, Oct 2006.

L. Narlikar, R. Gordan, U. Ohler, and A. J. Hartemink. Informative priors based on transcription factor structural class improve de novo motif discovery. *Bioinformatics*, 22(14):e384–e392, Jul 2006.

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, Dec 2004.

G. Pavesi, G. Mauri, and G. Pesole. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, 17 Suppl 1:S207–S214, 2001.

J. C. Rutherford, S. Jaron, E. Ray, P. O. Brown, and D. R. Winge. A second iron-regulatory system in yeast independent of Aft1p. *Proc Natl Acad Sci U S A*, 98(25):14322–14327, Dec 2001.

A. Sandelin and W. W. Wasserman. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J Mol Biol*, 338(2): 207–215, Apr 2004.

J. Schaber, B. Kofahl, A. Kowald, and E. Klipp. A modelling approach to quantify dynamic crosstalk between the pheromone and the starvation pathway in baker's yeast. *FEBS J*, 273(15):3520–3533, Aug 2006.

E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, 34(2):166–176, Jun 2003.

R. Siddharthan, E. D. Siggia, and E. van Nimwegen. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol*, 1(7):e67, Dec 2005.

A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–1050, Aug 2005.

K. Sjolander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I. S. Mian, and D. Haussler. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput Appl Biosci*, 12(4):327–345, Aug 1996.

R. Staden. Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res*, 12(1 Pt 2):505–519, Jan 1984.

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

D. Tautz. Evolution of transcriptional regulation. *Curr Opin Genet Dev*, 10(5):575–579, Oct 2000.

G. Thijs, M. Lescot, K. Marchal, S. Rombauts, B. De Moor, P. Rouze, and Y. Moreau. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, 17(12):1113–1122, Dec 2001.

M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23(1):137–144, Jan 2005.

A. E. Tsong, B. B. Tuch, H. Li, and A. D. Johnson. Evolution of alternative transcriptional circuits with identical logic. *Nature*, 443(7110):415–420, Sep 2006.

Y. Wang and H. G. Dohlman. Pheromone-regulated sumoylation of transcription factors that mediate the invasive to mating developmental switch in yeast. *J Biol Chem*, 281 (4):1964–1969, Jan 2006.

X. Xie, J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, and M. Kellis. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–345, Mar 2005.

G. Yona and M. Levitt. Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J Mol Biol*, 315(5):1257–1275, Feb 2002.

J. Zeitlinger, I. Simon, C. T. Harbison, N. M. Hannett, T. L. Volkert, G. R. Fink, and R. A. Young. Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell*, 113(3):395–404, May 2003.

M. Q. Zhang and T. G. Marr. A weight array method for splicing signal analysis. *Comput Appl Biosci*, 9(5):499–509, Oct 1993.

J. Zhu and M. Q. Zhang. SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics*, 15(7-8):607–611, Jul 1999.