Motifs, binding, and expression: computational studies of transcriptional regulation

by

Kenzie Daniel MacIsaac

B.Sc., Biochemistry, Carleton University (1998)
B.Eng., Electrical Engineering, Carleton University (2001)
M.A.Sc., Electrical Engineering, University of Toronto (2003)

Submitted to the Department of Electrical Engineering and Computer Science
In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September, 2009

Signature of Author: ……………………………………………………………………….
Department of Electrical Engineering and Computer Science
July 6, 2009

Certified by: ……………………………………………………………………………..
Ernest Fraenkel
Assistant Professor of Biological Engineering
Thesis Supervisor

Accepted by: ……………………………………………………………………………
Terry P. Orlando
Chairman, Committee on Graduate Students
Department of Electrical Engineering and Computer Science

Motifs, binding, and expression: computational studies of transcriptional regulation

by

Kenzie Daniel MacIsaac

Submitted to the Department of Electrical Engineering and Computer Science
on July 6, 2009, in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

Organisms must control gene expression in response to developmental, nutritional, or other environmental cues. This process is known as transcriptional regulation and occurs through complex networks of proteins interacting with specific regulatory sites in the genome. Recently, high throughput variations of experimental techniques like transcriptional profiling and chromatin immunoprecipitation have emerged and taken on increasing importance in the study of regulatory processes. Mining these experiments for useful biological information requires methods that can handle large quantities of noisy data and integrate information from disparate experimental sources in a principled manner. Not coincidentally, computational and statistical methods for analyzing these data have increasingly become a focal point of research efforts.

In this thesis we address three key challenges in the analysis of genomic sequence, protein localization, and expression data: (1) learning representations of the specific binding interactions that determine connectivity in regulatory networks, (2) developing physically grounded models describing these interactions, and (3) relating binding to its ultimate effect on the expression of regulated genes. To this end, we present several different algorithms and modeling techniques and apply them to real biological data in yeast, mouse, and human.

Our results demonstrate the utility of leveraging multiple sources of information for improving motif analyses of chromatin immunoprecipitation data. Phylogenetic conservation information and knowledge of an immunoprecipitated protein’s DNA binding domain are both shown to have great value in this context. We next present a biophysically motivated framework for modeling protein-DNA interactions and show how it leads to very natural algorithms for analyzing the binding specificity of an immunoprecipitated protein, and jointly analyzing protein localization data for multiple regulators or multiple conditions. Finally, we present an analysis of transcriptional coregulator binding in a variety of mouse tissues and a method for predicting which proteins form complexes with the coregulator based purely on the sequence of the regions it binds. We detail a simple but powerful model relating regulator binding to gene expression, and show how the position of regulatory regions is of crucial importance for predicting the expression level of nearby genes.

Thesis supervisor: Ernest Fraenkel
Title: Assistant professor of Biological Engineering

Acknowledgements

Above all I am indebted to my parents Gordon and Anne MacIsaac for their love, support, and encouragement. They instilled in me the value of education and a love of reading. Perhaps more importantly, they demonstrated by example the importance of hard work and dogged determination. I would also like to thank my dear sisters Tara and Alexis for their support throughout my graduate studies.

My thesis supervisor Professor Ernest Fraenkel has been a wonderful mentor to me over the course of graduate school. He has been supremely patient, allowing me the freedom to pursue ideas (even when those ideas were a bit on the shaky side). At the same time he has provided direction and guidance whenever it was needed or requested. His sound judgment and creativity have made working in the Fraenkel lab a rewarding experience.

I was fortunate to have had two great mentors during my time at MIT. Professor David Gifford was my first research advisor and introduced me to the field of computational genomics. With his insight into computation, statistical learning, and biology, as well as his impressive ability to foster collaboration, I hold him up as a model of what researchers in our highly inter-disciplinary field should aspire to. Both Professors Gifford and Fraenkel share a commitment to innovation, strong scientific ethics, and kindness in their dealings with students and collaborators that I hope to emulate as I leave MIT and embark on my own scientific career.

I also wish to acknowledge the great group of people I've been able to work and collaborate with during my studies. A partial list includes Ben Gordon, Tim Danford, Alex Rolfe, Georg Gerber, Robin Dowell, Duncan Odom, Alan Qi, Lena Nekludova, Rick Young, Alex Marson, Garrett Frampton, Ting Wang, Gary Stormo, Alice Lo, William Gordon, Shmulik Motola, Tali Mazor, Carol Huang, Laura Riva, Esti Yeger-Lotem, Jim Zhang, Katherine Romer, Aparna Kumar, Deepika Dinesh, Tatjana Degenhardt, and Chris Ng. It's difficult to convey just how much talent is represented in the preceding list of names, so I'll just say 'a lot'.

Finally, I'd like to acknowledge and thank the friends I've made here in the Boston and Cambridge area who've made life in Massachusetts such a pleasure: Sonia Timberlake, Des Adler, Annie Kim Adler, Sebastian Stirling, Leia Stirling, Andrew Takahashi, Frank Wei, James Barnthouse, Tina Hinojosa, Elliot Haimes, Justin Buck, Leslie Mebane, Sean Clarke, Dan Buckland, Sarah Miller, Karen Sachs, Jay Gill, Jay Jones, Jon Tyson, Scott Litzelman, Wendy Freedman, Michel-Alexandre Cardin, and Marc Barron among others have all been great friends to me.

Table of Contents

Chapter 1: Introduction

This thesis is about the development and use of computational models and tools to study the biological process known as transcription. Before proceeding any further, it is important to explain what transcription is, why it is an important and interesting subject of study, and why it is helpful to develop computational techniques to study it. To that end, this document will begin with an extremely brief introduction to molecular biology. For a more in depth treatment of this subject matter we refer the reader to a number of excellent text books that cover the field in great depth [1, 2].

1.1 An exceedingly brief overview of molecular biology

Living things, from the simplest microorganisms to the most complex animals, are composed of cells. A cell is, in a sense, just a small bag of organic molecules walled off from its environment by a lipid membrane. Inside the cell there is complex molecular machinery that carries out the chemical and mechanical tasks required to sustain life. These tasks include the catalysis of chemical reactions, physical transport of material, and the orchestration of interactions with the cell's external environment. The machinery that accomplishes this dizzying array of function is composed of diverse interacting organic molecules: carbohydrate, lipid, nucleic acid, and protein. Although all of these molecules play a crucial role in cellular function, it is proteins that have the most diverse range of activities and it is primarily proteins that function as the agents of chemical and physical action in the cell.

Underpinning the biological processes of growth and reproduction is the ability of cells to divide to form two daughter cells. Each daughter cell contains all the information required to continue carrying out the functions necessary for life. This information is encoded in long polymeric strands of nucleotide monomers called deoxyribonucleic acid (DNA). The DNA strand consists of alternating pentose sugar (2-deoxyribose) and phosphate groups linked by ester bonds. Each pentose sugar is covalently linked to one of four bases: adenosine (A), thymine (T), guanine (G), and cytosine (C). Phosphodiester bonds occur between the 5th and 3rd carbons on adjacent sugar molecules, giving the strand an inherent directionality. The terminal with an unesterified phosphate at the 5th carbon is called the 5' end, whereas the end with an unbound hydroxyl group on the 3rd carbon is known as the 3' end. DNA is normally present in the cell as a double-stranded helix, with the two complementary strands wrapped around each other in an anti-parallel fashion. The anti-parallel strands associate through non-covalent base-pairing interactions. Individual bases form specific hydrogen bonding interactions with a complementary base, whereas interactions with non-complementary bases are much less favorable (and generally not present) due to the absence of these hydrogen bonds and other steric constraints. Adenosine and thymine form one complementary pair, while guanine and cytosine form the other.

In higher organisms like mammals, DNA double helices are wrapped around structural proteins, called histones, and packed into a more dense structure termed chromatin. The entire assembly of two anti-parallel DNA strands wrapped and packed around scaffold proteins is termed a chromosome. Chromosomes are located in a membrane-enclosed compartment inside the cell called the nucleus. Different species have variable numbers of chromosomes: humans have 23, mice have 20, and the fruit fly has 4. Ploidy refers to the number of copies of each chromosome an organism has. Nearly all mammals are diploid, meaning they have two copies of each chromosome. One copy is inherited from each parent. Mice, for example, have 2 copies each of 20 distinct chromosomes for a total of 40. All the genetic material making up the various chromosomes in an organism's nucleus is commonly referred to as the genome.

DNA, as has been alluded to previously, is used by the cell for information storage. The central dogma of molecular biology describes how DNA encodes instructions for the cell's machinery: specific regions of DNA act as templates from which a ribonucleic acid (RNA) message is synthesized. This message is then translated by the cell into a protein. RNA is very similar to DNA in that it consists of monomeric nucleotides connected in a linear polymeric strand. The key differences between DNA and RNA are that the pentose sugar making up the sugar-phosphate backbone of RNA is ribose instead of 2-deoxyribose, and the base thymine is replaced with uracil. RNA has diverse roles in the cell: it functions as a crucial component of the cellular machinery required to synthesize protein, it can have catalytic activity, and it has important regulatory roles. However, the particular function of RNA that this thesis is most concerned with is that of an intermediary transcript from which protein is synthesized: messenger RNA (mRNA). The specific regions of DNA that are transcribed to mRNA and which, after processing, are then translated to protein are called genes. Genes consist of protein coding regions called exons, and non-coding regions called introns which are excised by the cell after transcription and prior to translation. The cell interprets the sequence of exonic regions using a genetic code. The genetic code translates three nucleotide chunks of sequence, or codons, into one of twenty amino acid monomers that make up proteins. This process occurs during translation when a molecular machine known as the ribosome reads an RNA transcript one codon at a time and adds amino acids onto the growing polypeptide chain.

1.2 Transcription

Transcription of genes to mRNA, also known as gene expression, is carried out by the enzyme RNA Polymerase. In eukaryotes there are several different RNA Polymerase enzymes; however transcription of most genes and regulatory RNAs is carried out by RNA Polymerase II (Pol II). The process of transcription can conceptually be broken up into five phases: pre-initiation, initiation, promoter clearance, elongation, and termination. Pre-initiation involves the assembly of general transcription factors (GTFs) at the proximal promoter region (approximately 10 to 35bp upstream of the transcription start site (TSS)) to form the pre-initiation complex. The GTFs include TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH, and Mediator. Some of these general transcription factors, such as the TFIID component TATA binding protein (TBP), recognize and bind specific nucleotide sequences at the promoter, although they may also bind non-specifically in certain contexts. Initiation refers to recruitment of RNA polymerase to the promoter to form a productive complex with these GTFs. The resulting initiation complex may repeatedly synthesize short, abortive transcripts that are released before they reach approximately 23 nucleotides in length. The transition between initiation and elongation involves Pol II escape from the interactions tying it to the promoter, a process that is closely tied to phosphorylation of the C-terminal repeat domain of the largest Pol II subunit. Elongation of the transcript proceeds as the Pol II enzyme translocates down the coding strand in the 5’ to 3’ direction, reading the template strand and adding the appropriate nucleotide to the growing mRNA molecule. Finally, transcription is terminated by cleavage of the transcript and the addition of multiple adenosines to the transcript’s 3’ end.

1.3 Transcriptional Regulation

In order for various cell types to perform their specialized functions and respond to environmental cues they must have the capacity to control when and how genes are expressed. By way of example, in mammals liver hepatocytes have a very different role than do the adipocytes that make up white fat tissue. Although many proteins are required by both cell types, others must be synthesized in drastically different quantities, and each tissue must respond differently to external signals. During prolonged periods of starvation, hepatocytes enact a gene expression program leading to higher levels of the glucose production machinery in order to fulfill the energetic requirements of the brain. Meanwhile, since adipose tissue is a major energy source during starvation, it must down-regulate proteins responsible for fat storage and synthesis and upregulate those responsible for lipolysis to produce glycerol and fatty acids [3]. The genome not only encodes instructions for assembling the proteins making up the cell’s molecular machinery, but also contains a regulatory code that is interpreted by the cell and determines when, where, and how these instructions will be used.

Control of gene expression programs across diverse tissues and developmental stages is achieved through complex networks of proteins interacting with specific regulatory sites in the genome. These regulatory proteins have several different mechanisms of action: they may interact directly with components of the basal transcriptional machinery, recruit other important regulators, or affect chromatin structure. The specificity in targeting regulators to particular genomic locations is achieved in a number of ways. Many transcription factors contain a structural domain that recognizes and binds DNA in a sequence-specific manner [4]. A second source of specificity arises from indirect targeting of regulators to regulatory sites via protein-protein interactions with DNA-bound transcription factors [5]. In this way, regulatory complexes of several proteins can assemble on DNA. A third mechanism involves wholesale changes to the chromatin structure of large swaths of the genome. Chromatin modifying complexes, which are themselves likely recruited by DNA-bound proteins, can covalently modify histones affecting nucleosome structure and resulting in the recruitment of other regulators that recognize specific histone modifications [6].

Many important regulatory sites occur in the proximal promoter, where the general transcriptional machinery is recruited to the transcription start site. However equally as important to transcriptional regulation are enhancer regions, which may be located distal to the TSS, sometimes hundreds of kilobases away. Enhancers have been classically defined as regulatory elements that modulate transcription independent of their position or orientation [7]. Enhancers bind transcription factors, which in turn recruit transcriptional coregulators. Coregulators are proteins that do not, themselves, bind DNA but rather are recruited to their targets through protein-protein interactions with DNA bound factors. Once recruited to an enhancer region, these coregulators can have a multitude of different effects on transcription rates. Many are capable of acting as scaffold proteins by interacting with the basal transcriptional machinery or other regulators. In addition, many coregulators have enzymatic activity and can covalently modify histones or transcription factors, affecting chromatin state or regulator activity [5].

The mechanism of enhancer action is an active area of research and several different models have been put forward to explain various aspects of their function. Their ability to affect transcription at long distances is thought to involve some form of communication between enhancer and promoter either through a DNA looping event, translocation of the regulatory complex along the DNA strand from enhancer to promoter, or via effects on chromatin structure that propagate from enhancer to TSS [8]. At least one experimentally characterized enhancer is thought to function through a combination of such mechanisms [9].

Although originally identified as elements that activate transcription, enhancers can also have repressive activity; their precise function is determined by the combinations of regulatory proteins that they bind. The ability of combinations of limited numbers of transcriptional regulators to come together and enact a huge variety of transcriptional programs is thought to be crucial to transcriptional regulation and is referred to as *combinatorial control*.

The nature of combinatorial control at enhancers has been described by two competing models: the enhanceosome and the billboard [10]. The enhanceosome model assumes that enhancer activity relies on the precise and highly cooperative assembly of regulatory proteins on the enhancer. Enhancer function will therefore be disrupted by small changes in binding site position or orientation which may affect any one of the interactions in the complex. The billboard model assumes that enhancer function is an integration over several independent, and possibly opposing, transcriptional signals. The transcriptional effect depends on which signal is ‘observed’ and by whom, hence the name billboard. This model predicts that enhancer function should be much less sensitive to binding site positions and orientations since many of the factors bound at the regulatory region act independently. It may be the case that for many enhancers the truth lies somewhere between these two models with some cooperative interactions having important regulatory roles, but with enhancer function having significant redundancy and being supported by a diversity of regulator binding configurations.

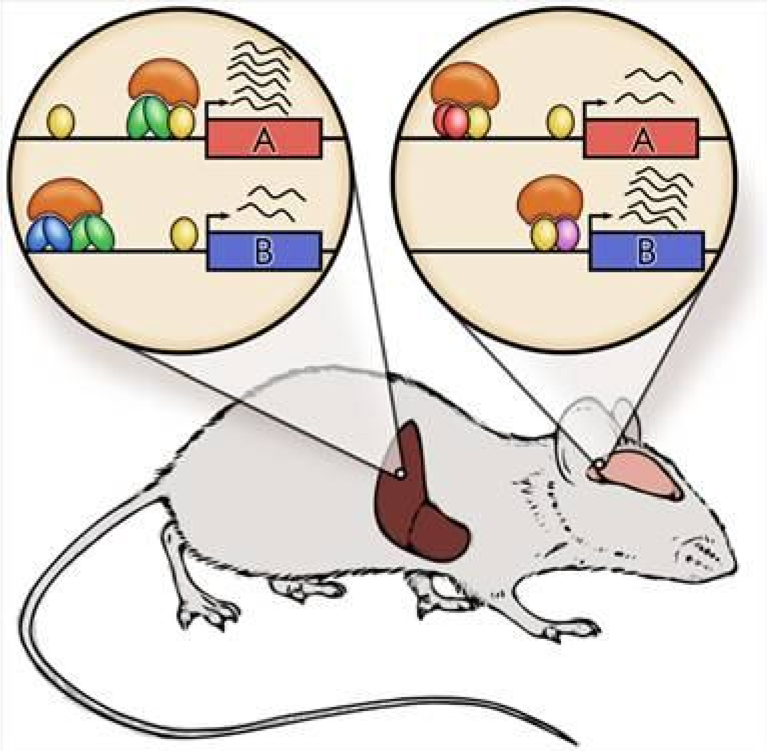


Figure 1.1: Combinatorial control in transcriptional regulation. A variety of regulatory programs in diverse tissues are enacted by a limited number of transcription factors interacting with the genome to achieve transcriptional outcomes. The particular program enacted is determined by the set of active regulators in the tissue, and the set of physical interactions that can occur between the regulators themselves and between regulators and DNA.

At this point, it is likely becoming clear that transcriptional regulation in eukaryotic organisms is very complex and only partially understood. Yet it has important implications in evolution, development, and disease. The genetic mutations that accumulate over evolutionary time and result in the divergence of different species often have a basis in transcriptional regulatory mechanisms, either through changes in *cis* regulatory sequences or sometimes the transcription factors themselves [11]. Promising stem cell based therapies rely on an understanding of how progenitor cells differentiate into the multitude of cell types making up the human body, and importantly, how this process is controlled. Fundamentally this occurs at the level of transcriptional regulation. In fact, it has been famously demonstrated that by activating a simple transcriptional switch consisting of only four transcription factors, fully differentiated mammalian cells can revert to pluripotency [12-14]. Importantly, many diseases have a basis in transcriptional disregulation. These include cancer [15-17], diabetes [18], Rubenstein-Taybi syndrome [19], and many others [20-23].

1.4 Experimental Tools for Studying Transcriptional Regulation

The earliest experimental studies in transcription focused on the important and painstaking work of identifying the components of the transcriptional machinery and characterizing their function [7, 24-27]. From these contributions, general theories and principles of regulation emerged including the notion of repression, activation, modularity, and cooperativity [28]. With the advent of high-throughput microarray and sequencing data, the study of transcriptional regulation has been revolutionized. DNA microarray technology allows us to profile the expression of thousands of genes in a single experiment [29]. Huge experimental efforts have provided us draft sequences of the human genome [30, 31], as well as the genomes of important model organisms like mouse and rat [32, 33], and have afforded us the opportunity to decode the regulatory information present in the sequences of promoters and enhancers on a genome-wide scale. High-throughput chromatin immunoprecipitation experiments paired with microarrays or massively parallel sequencing allow us to map the localization of important regulatory proteins on a genome-wide basis [34, 35]. These rich and varied datasets open up new worlds of scientific opportunity and have led to a greater understanding of transcriptional regulation, while at the same time presenting us with significant analysis challenges that are still being addressed.

DNA microarrays are grids of short DNA oligonucleotide probes, either pre-synthesized and spotted onto a grid square or printed directly onto the array, that have been designed to be complementary to specific mRNA transcripts. The basic idea behind microarray analysis is that mRNAs present in a sample can be reverse-transcribed to cDNA and then detected when they hybridize to a complementary probe on the array [36]. Expression profiling using DNA microarrays takes two forms: single color experiments and two-color experiments. In a single color experiment mRNA is collected and isolated from cells in a particular growth condition. The mRNA is chemically labeled with a fluorescent dye and hybridized to the array. After washing away non-specifically hybridized mRNA, the total quantity of each mRNA can be estimated using the fluorescent intensity measured for each probe on the array [37]. In a two-color experiment, mRNA from cells grown in two different conditions is labeled using two different fluorescent dyes [38]. The samples are then hybridized to the array where they compete for their complementary probe. After normalization, the relative amount of fluorescence observed for each probe measures the relative quantity of the corresponding mRNA in the two profiled samples.

DNA microarrays can also be used to measure the binding of protein to specific genomic regions profiled in a chromatin immunoprecipitation (ChIP) experiment [34].

This technique, known as ChIP-chip, involves first cross-linking chromatin and protein using a chemical agent (e.g. formaldehyde). This forms covalent links between amine groups on nucleotides, which are involved in base-pairing interactions in the minor groove of the DNA helix, and amine groups on proteins bound at those sites. Cross-linking will also covalently link amine groups on nearby proteins, making ChIP suitable for profiling the genomic localization of regulators that may be indirectly targeted to DNA through protein-protein interactions with DNA bound factors. After cross-linking, the DNA is fragmented and isolated. Fragments that are cross-linked to the protein of interest are then immunoprecipitated using an antibody specific to the protein. After-reversing the cross-links and purification, the DNA is amplified using the polymerase chain reaction and fluorescently labeled. This material is then hybridized to a microarray along with differentially labeled, unenriched, whole genome DNA. The microarray intensity measurements allow genomic regions enriched in protein binding to be identified.

Very recently, massively parallel sequencing technologies have emerged that allow for large scale sequencing of individual DNA molecules that are cross-linked to protein immunoprecipitated in a ChIP experiment [35, 39]. These sequence reads may then be aligned to a reference genome to determine binding location. The number and distribution of reads aligning to a specific genomic region serves as a measure of binding enrichment. For chromatin immunoprecipitation experiments, this technique is known colloquially as ChIP-seq.

1.5 Computational Tools for Studying Transcriptional Regulation

The huge size of the datasets produced by microarray and sequencing-based experimental methods, and the significant experimental noise they contain, make them particularly suitable to computational analyses that can model both measurement uncertainty and also deal with large quantities of data. There has been a tremendous amount of work in this area on several fronts. Here I will attempt to summarize prior work in the fields most related to this thesis.

1.5.1 Analysis of chromatin immunoprecipitation data

Computational analysis of ChIP data starts with the basic goal of identifying bound genomic regions. For ChIP-chip data several approaches have been employed. The simplest method is to identify all probes with a raw enrichment ratio greater than some threshold, and to label those probes as bound. More principled approaches employ a statistical model of observed enrichment ratio data to identify bound regions [40, 41]. Current state-of-the-art techniques model the expected peak shape arising from the DNA shear distribution obtained during the sonication or fragmentation step in the ChIP protocol [42, 43]. For ChIP-seq data, the most basic technique for identifying peaks looks at the number of sequence tags that cluster in a particular region and compares this to a background model that assumes a uniform distribution of tags [39]. Bound regions are then identified using some reasonable p-value or FDR cutoff. Recent analyses, however, have demonstrated that the assumption of uniform background tag position is a poor one and may result in a high false positive rate [44]. A better approach is to collect control reads from unenriched whole-genome DNA, and identify binding by assessing enrichment relative to this control [35]. This can help adjust for biases owing to variations in copy number, sequencing efficiency, and cell-type specific chromatin structure [45]. The current state-of-the-art peak-calling methods use control data to estimate local tag distribution backgrounds, thereby achieving significantly improved performance [46]. Once identification of bound region has been accomplished, it is often of interest to identify statistically overrepresented sequence motifs associated with binding. This may provide additional confidence that experimentally identified binding sites are not false positives, and may yield insight into which other proteins cooperate with the immunoprecipitated factor to regulate its identified targets.

1.5.2 Motif Discovery

Many functionally important regions of the genome can be recognized by searching for sequence patterns, or "motifs." There are many biologically meaningful sequence patterns in the genome including CpG islands [47], RNA splice sites [48], and nucleosome positioning motifs [49]. The motifs this thesis is concerned with, however, correspond to the specific sites bound by regulatory proteins. The search for these sites is challenging because a single regulatory protein will often recognize a variety of similar sequences. Computational techniques employed to learn representations of regulatory motifs are termed *motif discovery algorithms*. The motif discovery problem can be formulated as follows: we have a set of genes that are believed, *a priori,* to be co-regulated and thus likely to be bound by one or more common regulatory proteins. We wish to learn motif representations that explain this binding.

There are many ways of representing the sequence specificity of a protein, and the choice of a particular representation is often determined by considerations such as simplicity, interpretability, representational power, or computational convenience.

Perhaps the simplest way of representing a motif is by using a consensus sequence of preferred nucleotides (adenine [A], cytosine [C], guanine [G], or thymine [T]). Degeneracy in the binding specificity of a protein can be incorporated using ambiguity codes (purine [R], pyrimidine [Y], strong [S], weak [W], keto [K], amino [M], and any nucleotide [N]) [50]. A number of methods for generating consensus sequences from data are possible, and several methods have been previously compared [51]. A second widely used motif model is the position weight matrix (PWM). In this formulation, the motif is represented as a matrix of nucleotide scores indexed by letter and position [52]. A closely related approach models a motif as a matrix of probabilities, where each position is represented using a multinomial distribution over observed nucleotides. Under certain assumptions, the nucleotide frequencies observed at different positions in a set of binding sites are related to the theoretical contribution of a particular nucleotide to the free energy of protein binding [53-55]. Motifs represented as frequency matrices can be visualized conveniently using sequence logos. A sequence logo consists of an ordered stack of letters, where a letter's height indicates the information it contains at that position [56].

Consensus sequences and simple matrix models ignore some of the complexity of protein–DNA interaction. Dependencies between nucleotides at different positions in protein binding sites have been observed [57, 58], and several motif models have been proposed that take into account the possibility of positional correlations. Zhou and Liu modeled a motif using a generalized weight matrix that could incorporate pair-wise dependencies [59]. Several other representationally powerful models have been proposed including boosted classifiers [60] and a hidden Markov Dirichlet multinomial model [61]. Of course increasing the model complexity requires more data to estimate the model's parameters. If data are limited, complex models may overfit and yield a poor representation of the factor's true specificity. An important study by Benos, Bulyk, and Stormo suggested that while the consensus sequence and PWM may not fully capture all the subtleties of a protein's binding specificity, these simple and easily interpretable models usually provide a very good approximation to reality [62].

Motif discovery algorithms may be broadly grouped into two categories: enumerative methods and alignment-based methods. Enumerative methods typically involve exhaustive enumeration of words up to some maximum size in a dataset. Once the words are cataloged, they can be scored using an appropriate measure of statistical significance. The computational time complexity of enumerative methods is approximately $O(NmA^eL^e)$, where N is the number of sequences, m is their length, A is the alphabet size, L is the motif length, and e is the number of errors allowed in a match to a catalog entry [63]. Many enumerative methods use trade-offs on the alphabet size and the number of allowable errors to make these searches computationally feasible [63-65].

Alignment methods take on a wide variety of forms, but often involve development of a probabilistic model of the observed sequence data. The MEME program, for example, treats a particular sequence as arising from a mixture model in which the small window of sequence containing the motif is generated from a motif model—represented by a probability matrix—and the rest of the sequence is treated as arising from a Markovian background [66]. The generative model describes a family of parameterized probability distributions, and the motif is described by parameters of this distribution. Any number of optimization techniques may be used to search for the parameter setting that maximizes the likelihood of the observed sequence data. Two frequently used techniques to perform this search are the expectation-maximization (EM) algorithm [67] and Gibbs sampling [68].

1.5.3 Computational methods for probing regulatory mechanism

Although sequence analysis on its own may reveal important information about transcriptional regulation, approaches that integrate sequence and/or binding data with expression data have even greater promise for revealing regulatory mechanism. This has thus far been borne out in simpler model organisms like yeast and bacteria where data in a broad set of growth conditions are plentiful, genome sizes are relatively small, and there are fewer regulators. However, computational forays have also been made with some success into higher eukaryotes as more and more genome-wide sequence and ChIP data becomes available.

One frequently employed approach relates sequence motifs to expression data using regression techniques. This can be particularly valuable when searching for regulatory motifs associated with expression changes in an experimental condition of interest. Bussemaker et al. presented the REDUCE method that enumerates the short DNA sequences present upstream of a set of genes, and then uses multivariate linear regression to associate gene expression level with the presence of these motifs [69]. Similar approaches have been presented by other investigators [70, 71]. These methods can account for combinatorial interactions by including multiple motif features as predictors of expression. Keles et al. modeled cooperativity by including products of motif feature terms in their regression models [71]. Das and coworkers have argued that linear regression-based models cannot accurately represent the significant nonlinearities observed in transcriptional regulation. They presented a different regression technique based on linear splines that captures switch-like behavior in transcriptional regulatory networks [72].

An alternative to the regression-based techniques discussed above is to group genes into co-expressed clusters. The basic notion is simply that groups of functionally related genes will tend to be co-expressed across a wide variety of conditions and therefore their expression is likely to be regulated by a common set of regulators. These regulator combinations may be identified through analysis of the motifs present in the promoter regions of the clustered genes. Beer and Tavazoie used such an approach in yeast and worm, expanding the set of features considered to include specific positional, orientation, and ordering constraints on motifs [73]. They used these features to predict membership in one of 49 pre-defined expression patterns obtained by clustering expression data across a set of 255 conditions and found that promoter sequence alone did a surprisingly good job of predicting a gene's expression program. However, a follow up study by a different group suggested that, in addition to a methodological error that probably resulted in overestimation of their predictive performance on held-out test data, it turns out that a naïve Bayes classifier that ignores position, orientation, and cooperativity performs better on the same data, casting some doubt on the utility of modeling these higher order effects [74].

Another important class of algorithms simultaneously clusters genes into co-regulated groups and attempts to identify regulators responsible for coordinating their expression. The module networks approach of Segal et al. uses gene expression data as evidence of transcription factor activity in order to assign regulators to sets of co-regulated genes and identify functionally coherent modules in yeast [75]. Bar-Joseph and colleagues demonstrated how ChIP binding data could be used to learn regulatory modules with direct physical evidence of regulatory interactions between transcription factors and member genes [76]. More recently, Zhou et al. extended these approaches by introducing 2nd-order expression analysis, where in addition to co-expression analysis, correlations of correlations between groups of co-expressed genes are used to identify functionally-related modules [77]. Several related methods have been presented that incorporate sequence features allowing regulatory modules to be linked to particular motifs and motif combinations [78, 79]. Most module-discovery approaches have been designed and tested using yeast data, where expression and protein localization is plentiful. However, some investigators have developed methods and demonstrated their applicability to mammalian data. Gerber et al. used a model based on a hierarchical Dirichlet process to discover regulatory programs across human tissues [80]. Tissues are automatically clustered into related groups that share similar expression programs. Expression programs can be shared across tissues, and genes can belong to more than one program. This model naturally accounts for heterogeneity in expression between tissues, cell types within a tissue, and between samples, and does not enforce a pre-determined number of expression programs to be specified but rather infers this from the available data. Very recently, the techniques of Bar-Joseph et al. have been applied to the analysis of expression data in human cancer cell lines [81].

A different, but extremely interesting class of algorithms probes regulatory mechanism by focusing directly on the activities of particular regulators or regulator combinations. Segal and colleagues presented a thermodynamic framework that does not rely on ChIP data to identify transcription factor regulatory targets, but rather uses the expression level of those regulators along with a description of their DNA binding specificity to predict when and where they will bind. Each factor is assumed to contribute to expression level independently, and the net expression level of a gene is an integration over the contribution of all factors. They trained their model using spatial expression patterns for eight transcription factors and 44 experimentally characterized gene modules in the developing Drosophila embryo and showed how this framework allowed them to accurately predict segmentation patterns for held out gene modules. Yeang and Jaakkola presented a different probabilistic method that links expression outcomes to the activity of combinations of transcriptional repressors and activators. These regulators are characterized by their effect on target expression in response to changes in their own expression. Their method not only identifies genes sharing a regulatory program, but also learns a description of that program in terms of cooperating regulators which may then plausibly be linked to specific physical mechanisms of regulation [82].

One of the chief remaining computational challenges is developing new techniques, or strategies for applying some of the techniques discussed above, to study mammalian data. Mammalian models are obviously of great interest since what we learn from these models is often directly applicable to human biology and disease processes. However, the development of computational tools that can be applied to mammalian data is complicated by several issues, some of which have been touched on already: huge genome sizes, larger numbers of regulatory proteins, and a great diversity in the number of tissues and cell types that are encountered.

1.6 Thesis roadmap

In this thesis I present a number of novel computational approaches for analyzing sequence, protein localization, and expression data to study transcriptional regulation. Chapter 2 presents a motif discovery algorithm, Converge, which uses phylogenetic conservation information to aid in motif discovery. I then demonstrate how it was used, in combination with a second conservation-based motif discovery algorithm, to expand a previously described map of regulatory sites in yeast. Chapter 3 describes a discriminative motif discovery approach, called THEME, which allows prior knowledge about a protein’s DNA binding specificity to be incorporated into the motif search. In Chapter 4 I present a biophysically motivated framework describing protein-DNA interaction and show how this framework forms the basis of a motif discovery method that can be incorporated into joint analysis of ChIP and sequence data and very naturally extended in a number of interesting directions. Chapter 5 describes a unique and surprisingly accurate probabilistic model that uses ChIP data to predict gene expression level and gain insight into transcriptional enhancer function in mammalian systems. Finally, in Chapter 6 I summarize the work presented here and outline the main contributions of this thesis.

Chapter 2: Converge

In this chapter I will present the Converge algorithm: a motif discovery method that uses phylogenetic conservation information to guide the motif search. I will then discuss how this tool was used in combination with a second motif discovery algorithm, called PhyloCon, to map the set of conserved transcription factor binding events in Saccharomyces cerevisiae. This work was a collaboration with Ting Wang, who developed PhyloCon and applied it to the yeast binding data. Benjamin Gordon's contributions, both in developing software tools for mapping binding events to the yeast genome and for evaluating the statistical significance of sequence motifs, were also instrumental in the success of this study. Discussions with Benjamin, Timothy Danford, David Gifford, and Ernest Fraenkel were very helpful during the development of Converge.

2. 1 Evolution, Phylogeny, and Motif Discovery

Over the course of evolutionary time, species arising from a common ancestor diverge as genetic point mutations, duplications, insertions, deletions, and rearrangements accumulate in their genomes. This process is random, but is also constrained since mutations that disrupt functionally important regions like genes can often have an adverse effect on an organism's fitness. Similarly, since transcription factor binding sites are important for ensuring proper control of gene expression, they tend to be under selective pressure over evolutionary time. A significant fraction of evolutionarily conserved noncoding DNA has been shown to correspond to regions important for regulation [83-86]. One study found that 98% of known binding sites of skeletal muscle–specific transcription factors are confined to the 19% of human sequences most conserved in orthologous rodent sequences [87]. This tendency of transcription factor binding sites to be conserved across species has been exploited in the context of motif discovery by several different research groups.

One approach to leveraging conservation information is to identify blocks of sequence that are conserved across multiple species using phylogenetic footprinting [88, 89]. Phylogenetic footprinting is a general technique for identifying conserved regions based on the evolutionary relationship among species. These conserved blocks can then be used as inputs to standard motif discovery tools and otherwise analyzed [90, 91]. By culling only the conserved sequence from the input data, uninformative background DNA is eliminated, and an effective increase in signal to noise is achieved that facilitates the search for motifs [92].

Other motif discovery tools integrate conservation information directly into the motif search. One approach generates a catalog of motifs with potential regulatory importance by determining, on a genome-wide scale, which consensus sequences are highly conserved across species. Highly conserved motifs are validated by determining their overrepresentation among groups of co-regulated genes [83, 93]. Several algorithms employ a generative probability model of DNA sequence to find conserved motifs using various inference techniques. EMnEM [94] and PhyME [95, 96] both incorporate probabilistic evolutionary models into EM-based motif searches. CompareProspector is a Gibbs sampling algorithm that uses a pre-computed score to measure the conservation level across windows in sequence alignments, and then biases the motif search to regions that are highly conserved [97]. PhyloGibbs is another Gibbs sampling algorithm that leverages conservation by assuming the motif must be present in all species in a conserved region [98]. All these algorithms have been demonstrated, in certain contexts, to outperform similar methods that don't take advantage of conservation information.

2.2 The Converge algorithm

Converge discovers DNA sequence motifs in regions putatively bound by a common regulator in the genome of one species (i.e. a primary genome) by using conservation information in the form of pair-wise sequence alignments from related species. The Converge algorithm makes use of the fact that transcription factor binding sites will tend to be conserved in orthologous regions of related species. Below we describe the algorithm in detail.

2.2.1 Overview

A schematic diagram of the Converge workflow is shown below:

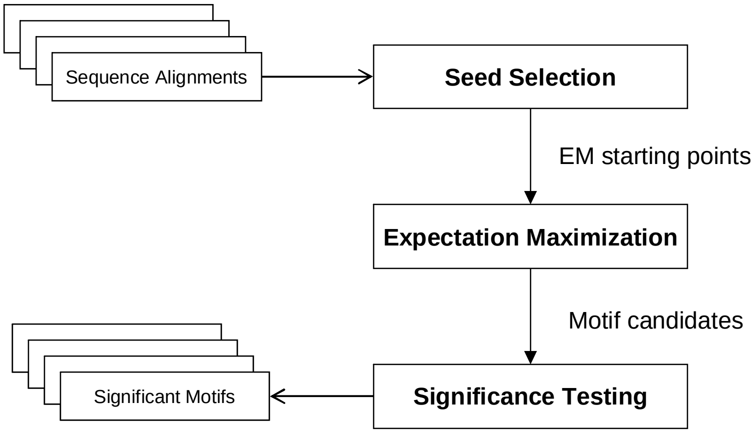


Figure 2.1: Converge Workflow Diagram

Optimization by Expectation Maximization is preceded by a seed selection step where initial starting points are chosen. The sequences in the primary genome are scanned for statistically over-represented k-mers and the top twenty are used to initialize the frequency matrix. Expectation Maximization is then run to convergence for each seed, and the resulting motif candidates are scored using an enrichment statistic to allow their statistical significance to be tested in a principled manner. The enrichment score is fit to a normal distribution estimated using enrichment scores from randomized data runs for a similar number of sequence alignments.

2.2.2 Probabilistic Model

The observed data, **X**, consists of a series of *N* primary genome sequences aligned (pair-wise) to orthologous sequence from supporting genomes. Each set of pair-wise alignments is indexed over *M* possible k-mers and *P* genomes, and is assumed to contain either one or zero motifs in the primary genome as in the zero-or-one-occurrence-per-sequence (ZOOPS) model of Bailey and Elkan [66].

Regions of sequence are treated as arising from either a background distribution or a motif distribution. The motif distribution is modeled using a frequency matrix:

$$\Pi = \left[\pi_1 ... \pi_w\right] \tag{1}$$

Where each π_i is a multinomial distribution representing the expected frequency of each base at position *i* in the motif. The motif has a fixed width, *w*. For sequences flanking the motif region, the distribution is modeled as arising from a 4th order Markov background, which in practice takes the from of a probability table with an entry for each possible 5-mer of sequence, which we denote by Λ_k, where *k* indexes the genome the background was calculated from. Converge assumes that the motif and background probabilities are independent.

Converge attempts to model three important characteristics of the data: regions in the pair-wise alignments that contain gaps should be treated differently than those without gaps, a given alignment may or may not contain the motif we are attempting to learn, and even when the motif is present in the primary genome it may not be present in the aligned supporting genomes. This treatment is made possible by the definition of two additional variables, one observed and the other hidden. The observed gap indicator variables $g_{i,j,k}$ take the value of 1 if for alignment *i*, position *j*, genome *k*, a gap is present in the motif window beginning at position *j*. The hidden variables $z_{i,j,k}$ indicate whether a functional motif is present in alignment *i*, at position *j*, in genome *k*.

We assume that a functional motif is only present in an aligned supporting genome if it also present at the corresponding position in the primary genome. If the primary genome is indexed by $k = 1$, this is equivalent to saying that, for all $k = 2\ldots P$, $z_{i,j,k}$ is equal to zero with probability 1 if $z_{i,j,1}$ is equal to zero. We also assume that given the value of $z_{i,j,k}$, the probability of the sequence for genome *k* is independent of the other aligned sequences and the primary sequence. A graphical model representation of the model is shown below:

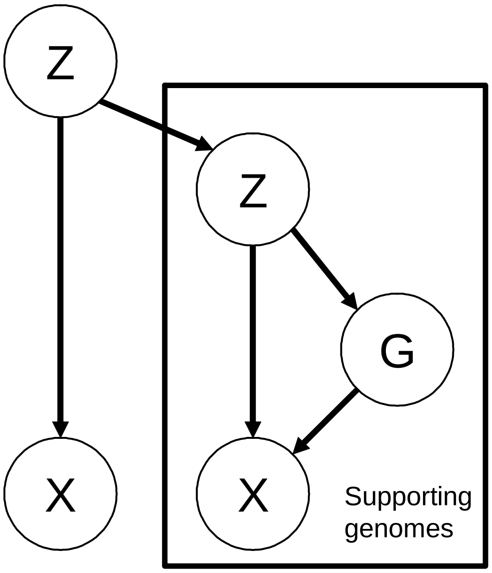


Figure 2.2: Converge probability model. The sequence in the primary genome depends only on the value of the motif indicator variable Z. Sequence in supporting genomes depends on both the motif and gap indicator variables.

Now, the log probability of the data can be factored as follows:

$$\begin{aligned}\log P(\mathbf{X},\mathbf{G},\mathbf{Z}\mid\mathbf{\Psi}) &= \log P(\mathbf{X}\mid\mathbf{Z},\mathbf{G},\mathbf{\Psi})+\log P(\mathbf{G}\mid\mathbf{Z},\mathbf{\Psi})\\ &+\log P(\mathbf{Z}_{k\neq 1}\mid\mathbf{Z}_{k=1},\mathbf{\Psi})+\log P(\mathbf{Z}_{k=1}\mid\mathbf{\Psi})\end{aligned} \qquad (2)$$

Here $Z_{k=1}$ denotes the $z_{i,j,1}$ for all *i* and *j*, $Z_{k\neq 1}$ denotes the $z_{i,j,k}$ for k≠1, and Ψ denotes the parameters associated with the motif and background probability mass functions. We define each term in equation 2 as follows:

$$L(\mathbf{X}\mid\mathbf{Z},\mathbf{G},\mathbf{\Psi})=\sum_{i=1}^{N}Q_i\left(L(X_w\mid\mathbf{Z},\mathbf{G},\Pi)+L(X_f\mid\mathbf{Z},\Lambda_k)\right)+(1-Q_i)L(X\mid\Lambda) \quad (3)$$

Where, $$Q_i=\sum_{j=1}^{M}z_{i,j,1} \qquad (4)$$

Here X_w is the sequence in the motif window, while X_f is the flanking sequence. If there is no functional motif present in the primary sequence, the first term of equation 3 will be equal to zero and the conditional probability of the sequence will simply be the probability it was emitted by the background model. If there is a functional motif present in the primary sequence, one of the $Z_{i,j,1}$'s will be equal to one and the log probability of the observed sequence is given by the sum of the window sequence log probability and the log probability of the flanking sequence. When a motif occurs in alignment *i,* position *j,* and genome *k*, the expressions for the probability of the motif sequence and the flanking sequence are as follows:

$$\begin{aligned}&\log P(X_w\mid z_{i,j,k},g_{i,j,k},\mathbf{\Pi})=\\ &\sum_{c=1}^{W}\left[\begin{aligned}&g_{i,j,k}\left(z_{i,j,k}\Pi_m^1\left[c,x_{i,j+c,k}\right]+(1-z_{i,j,k})\Pi_{bg,k}^1\left[x_{i,j+c,k}\right]\right)+\\ &(1-g_{i,j,k})\left(z_{i,j,k}\Pi_m^0\left[c,x_{i,j+c,k}\right]+(1-z_{i,j,k})\Pi_{bg,k}^0\left[x_{i,j+c,k}\right]\right)\end{aligned}\right]\end{aligned} \qquad (5)$$

$$\log P(X_f\mid\Lambda_k)=\sum_{c\notin\{j\ldots j+W-1\}}\Lambda_k\left[\tilde{x}_{i,c,k}\right] \qquad (6)$$

In equation 5 one of two probability models is selected depending on the value of the gap indicator variable. The $z_{i,j,k}$ selects either a motif model or a background model. When $z_{i,j,k}$ is one, the probability of the sequence in the window is calculated using the appropriate frequency matrix indexed by window position c, and base $x_{i,j+c,k}$ (Π_m^1 for g$_{i,j,k}$=1 or Π_m^0 for g$_{i,j,k}$=0), when its value is zero the probability is calculated using a 1st order background table, indexed by base $x_{i,j+c,k}$, for the appropriate genome k ($\Pi_{bg,k}^1$ for g$_{i,j,k}$=1 or $\Pi_{bg,k}^0$ for g$_{i,j,k}$=0). Equation 6 shows that the probability of the sequence flanking the motif window is calculated using the 4th order Markov background, indexed by $\tilde{x}_{i,c,k}$, the 5-tuple of sequence in the alignment beginning at position c, and genome k. In a similar fashion, the final term in equation 3 is defined as follows:

$$\log P\left(X \mid \mathbf{Z}, \Lambda\right) = \sum_{k=1}^{P} \sum_{j=1}^{M} \Lambda_k \left[\tilde{x}_{i,j,k} \right] \tag{7}$$

The second term of equation 2 models the probability of observing a gap in the motif window, given the value of the z's, and will in general be different for each aligned genome:

$$L\left(\mathbf{G} \mid \mathbf{Z}, \mathbf{\Psi}\right) = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=2}^{P} \left[\begin{array}{l} z_{i,j,k} \left(g_{i,j,k} \log \zeta_{k,1} + \left(1 - g_{i,j,k}\right) \log\left(1 - \zeta_{k,1}\right) \right) \\ + \left(1 - z_{i,j,k}\right) \left(g_{i,j,k} \log \zeta_{k,0} + \left(1 - g_{i,j,k}\right) \log\left(1 - \zeta_{k,0}\right) \right) \end{array} \right] \tag{8}$$

The value of $g_{i,j,k}$ can be generated from two different binomial distributions. The particular distribution that generates $g_{i,j,k}$ is selected by the value of $z_{i,j,1}$. When $z_{i,j,1}$ is equal to one, the value of $g_{i,j,k}$ is generated from the binomial distribution with parameter $\zeta_{k,1}$, otherwise it is generated from a binomial distribution with parameter $\zeta_{k,0}$. This models our belief that the likelihood of observing a gap in an aligned sequence should be different depending on whether a functional motif is present in the primary sequence. While in general it may be useful to model cases where a motif can be present in an aligned genome even though there are gaps in the alignment (e.g. if the binding specificity of a protein has evolved to have a different sized spacer region in one species), we assume that a gap in the alignment means that no motif is present, therefore we fix $\zeta_{k,1}$ to be zero and the frequency matrix Π_m^1 from equation 5 is never used.

The third term in equation 2 describes the probability of the $z_{i,j,k}$'s for k≠1 given the value of the $z_{i,j,1}$'s.

$$\log P\left(\mathbf{Z}_{k\neq 1} \mid \mathbf{Z}_{k=1}, \boldsymbol{\theta}\right) = \sum_{i=1}^{N}\sum_{j=1}^{M} z_{i,j,1} \sum_{k=2}^{P}\left[z_{i,j,k}\log\theta_k + \left(1 - z_{i,j,k}\right)\log\left(1-\theta_k\right)\right] \qquad (9)$$

Since we constrain the $z_{i,j,k}$'s to be zero unless $z_{i,j,1}$ is one, equation 9 models the probability of the $z_{i,j,k}$'s as arising from a binomial distribution with parameter θ_k representing the probability of observing a functional motif in the aligned genome *k*, given the presence of a functional motif in the primary genome.

The final term in the joint log probability distribution models the a priori probability of a functional motif being present at a particular alignment position in the primary genome:

$$\log P\left(\mathbf{Z}_{k=1} \mid \lambda\right) = \sum_{i=1}^{N}\sum_{j=1}^{M} z_{i,j,1}\log\lambda + \sum_{i=1}^{N}\left(1-Q_i\right)\log\left(1-\gamma\right) \qquad (10)$$

In equation 10, the parameter γ , is defined as the *a priori* probability of a motif being present in a given alignment, and the parameter λ is defined as γ / M , or the *a priori* probability of a functional motif being present at any given position in the alignment.

2.2.3 Optimization by Expectation Maximization

Converge learns motifs present in the data set using the EM algorithm; an iterative coordinate ascent on the joint probability function of equation 2, that first calculates the expected value of the hidden variables Z, and then uses that expectation to re-estimate the values of the parameters Π, ζ, θ, and γ. This procedure is repeated iteratively until convergence of the likelihood function.

In the E-step of iteration *t*, Converge calculates the expected log likelihood of the data over the distribution of the hidden variables Z, which takes the form:

$$
\begin{aligned}
&E\left[L(\mathbf{X},\mathbf{Z}\,|\,\mathbf{\Psi})\right] \\
&=\sum_{i=1}^{N}\left(1-\sum_{j=1}^{M}E\left[z_{i,j,1}\right]\right)\sum_{j=1}^{M}\sum_{k=1}^{P}\Lambda_k\left[\tilde{x}_{i,j,k}\right] \\
&\quad+\sum_{i=1}^{N}\sum_{j=1}^{M}E\left[z_{i,j,1}\right]\left\{\sum_{c=1}^{W}\mathbf{\Pi}^0\left[c,x_{i,j+c,1}\right]+\sum_{c\notin\{j\ldots j+W-1\}}\Lambda_k\left[\tilde{x}_{i,c,1}\right]\right\} \\
&\quad+\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=2}^{P}\left\{\begin{array}{l}\displaystyle\sum_{c=1}^{W}\left[\begin{array}{l}g_{i,j,k}\left(E\left[z_{i,j,1}z_{i,j,k}\right]\mathbf{\Pi}^1\left[c,x_{i,j+c,k}\right]+\left(E\left[z_{i,j,1}\right]-E\left[z_{i,j,1}z_{i,j,k}\right]\right)\mathbf{\Pi}^1_{bg,k}\left[x_{i,j+c,k}\right]\right)+\\ \left(1-g_{i,j,k}\right)\left(E\left[z_{i,j,1}z_{i,j,k}\right]\mathbf{\Pi}^0\left[c,x_{i,j+c,k}\right]+\left(E\left[z_{i,j,1}\right]-E\left[z_{i,j,1}z_{i,j,k}\right]\right)\mathbf{\Pi}^0_{bg,k}\left[x_{i,j+c,k}\right]\right)\end{array}\right]\\ \displaystyle+\sum_{c\notin\{j\ldots j+W-1\}}\Lambda_k\left[\tilde{X}_{i,c,k}\right]\end{array}\right\} \\
&\quad+\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=2}^{P}\left[\begin{array}{l}E\left[z_{i,j,1}z_{i,j,k}\right]\left(g_{i,j,k}\log\zeta_{k,1}+\left(1-g_{i,j,k}\right)\log\left(1-\zeta_{k,1}\right)\right)\\ +\left(E\left[z_{i,j,1}\right]-E\left[z_{i,j,1}z_{i,j,k}\right]\right)\left(g_{i,j,k}\log\zeta_{k,0}+\left(1-g_{i,j,k}\right)\log\left(1-\zeta_{k,0}\right)\right)\end{array}\right] \\
&\quad+\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=2}^{P}\left(E\left[z_{i,j,1}z_{i,j,k}\right]\log\theta_k+\left(E\left[z_{i,j,1}\right]-E\left[z_{i,j,1}z_{i,j,k}\right]\right)\log\left(1-\theta_k\right)\right) \qquad (11) \\
&\quad+\sum_{i=1}^{N}\left(1-\sum_{j=1}^{M}E\left[z_{i,j,1}\right]\right)\log(1-\gamma)+\sum_{i=1}^{N}\sum_{j=1}^{M}E\left[z_{i,j,1}\right]\lambda
\end{aligned}
$$

Taking the partial derivative of equation 11 with respect to the parameters Π, ζ, θ, and λ, and setting the result equal to zero, we derive the M-step update equations:

$$
\lambda^{(t+1)}=\frac{\sum_{i=1}^{N}\sum_{j=1}^{M}E\left[z_{i,j,1}\right]}{NM} \qquad (12)
$$

$$\theta_k^{(t+1)} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} E\left[z_{i,j,k}\right]}{\sum_{i=1}^{N}\sum_{j=1}^{M} E\left[z_{i,j,1}\right]} \tag{13}$$

$$\zeta_{k,1}^{(t+1)} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} g_{i,j,k} E\left[z_{i,j,k}\right]}{\sum_{i=1}^{N}\sum_{j=1}^{M} E\left[z_{i,j,k}\right]} \tag{14}$$

$$\zeta_{k,0}^{(t+1)} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} g_{i,j,k}\left(1 - E\left[z_{i,j,k}\right]\right)}{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(1 - E\left[z_{i,j,k}\right]\right)} \tag{15}$$

$$\pi_{c,l}^{(t+1)} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{P}\left[\left(1-g_{i,j,k}\right) I\left(i, j+c, k, \ell\right) E\left[z_{i,j,k}\right]\right]}{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{P}\left[\left(1-g_{i,j,k}\right)\sum_{l}\left(I\left(i, j+c, k, \ell\right) E\left[z_{i,j,k}\right]\right)\right]} \tag{16}$$

Where in equation 16, the indicator variable $I\ (i, j{+}c, k, \ell\,)$ is equal to 1 if $x_{i,j+c,k}$ corresponds to the base indexed by ℓ.

The θ parameter for each genome is initialized to the average number of differences per base position between the aligned genome and the primary genome. The ζ parameters for each genome are simply initialized to 0.5. This simple initialization scheme for the gap indicator prior seems reasonable, since its final value at convergence is very insensitive to the initial guess of its value.

2.3 Algorithm performance

A previously published study reported an initial regulatory map for *Saccharomyces cerevisiae* by analyzing genome-wide chromatin immunoprecipitation (ChIP) data for 203 proteins [99]. Of these 203 proteins, 172 were profiled in a growth condition in which at least four microarray probes were bound with a p-value cutoff of 0.001. Alignments of these probe sequences with three additional yeast species, *S. paradoxus*, *S. mikatae*, and *S. bayanus*, were provided as input to Converge. We then used Converge to re-analyze these data, evaluating its performance by comparing results to experimentally characterized binding specificities for 87 different transcription factors.

2.3.1 Seed Selection and Motif Discovery

We generated initialization points for EM in all data sets at motif widths of 6, 8, 10, 15, and 20 base pairs. For motif widths less than or equal to 10, we selected seeds by first identifying the top 400 k-mers in the data set. We calculated a conservation score for each k-mer by counting the total number of bases where the sequence was conserved across all intergenic regions in at least 50% of the aligned yeast species. We associated a p-value with these scores by fitting the result to a binomial distribution, or when the number of occurrences was sufficiently large, to a normal approximation to the binomial distribution. We discarded all k-mers with a conservation p-value greater than 0.1 from consideration as seeds. The remaining k-mers were scored using the hypergeometric distribution to give an enrichment p-value associated with observing an equal or greater number of occurrences in an equally sized random sample of probe sequences in *S. cerevisiae*. We selected the top 20 enriched surviving k-mers as initialization points.

For motif widths greater than 10, we used gapped k-mers consisting of flanking regions of defined sequence, with an unconstrained center region. This approach was intended to compensate for the paucity of large k-mers with multiple occurrences. Furthermore, many transcription factors are known to bind paired sequences separated by non-specific regions of DNA and it was hoped that this seeding approach would help in the discovery of such motifs. Each flanking region was set to a size equal to one third of the motif width, rounded down. The top 400 gapped k-mers were identified and subjected to the same conservation criterion described above. We scored these gapped k-mers for enrichment and the top 20 were selected as initialization points, with the gapped region initialized to background base frequencies.

For each initialization seed, we ran the Converge algorithm until the mean squared difference between motifs in subsequent iterations was less than 10^{-3} for each position in the matrix, and the value of each θ parameter changed by less than 10^{-3}. We confirmed empirically that this convergence criterion coincided with convergence of the data likelihood, which was computationally expensive to compute. In the M-step, we add 0.01 pseudo counts at each position in the frequency matrix. We used an estimate of the prior probability of motif occurrence in a given probe of 0.2 and set its learning rate to 0.5. The θ parameter was initialized to a simple measure of phylogenetic distance between the aligned species and Saccharomyces cerevisiae: the mean number of matches per position relative to S. cerevisiae in all probe alignments. This gave θ initialization values of 1.00, 0.80, 0.63, and 0.58 for *S. cerevisiae*, *S. paradoxus*, *S. mikatae* and *S. bayanus*, respectively. We estimated background sequence probabilities using a 4th order Markov model calculated separately for each species from its set of intergenic regions.

We used a previously described approach to empirically estimate the significance level of the motif generated by Converge [99]. The number of promoters bound by a regulator in each experiment ranges from 4 to 176, with an average of 55. From all promoters in the yeast genome where an orthologous sequence group could be formed based on sequences of multiple genomes, we randomly created datasets from 4 to 160 orthologous groups in size. For each sample size, 50 to 100 datasets were generated. We applied Converge to these randomized datasets and estimated normal distributions for the hypergeometric enrichment at each sample size. After motif discovery on real datasets, motif scores were compared to the normal distribution of the most closely matching random sequence sample size. P-values were determined using z-scores calculated from the mean and standard deviation of this distribution. The top-ranked motif was accepted as the predicted specificity for the corresponding protein if it had a p-value < 0.001.

2.3.2 Known binding specificities recovered by Converge

We first evaluated how many of the 87 previously described transcription factor binding specificities Converge could recover from the sequence alignment data. When a matrix was available describing the known specificity, a match was defined as an average Euclidean distance between the frequency matrix columns of < 0.18. For the remaining motifs, a match was determined empirically by assessing whether the motif was consistent with reported binding sites. Converge's performance is compared to the six programs used by Harbison et al. in Figure 2.3 below:

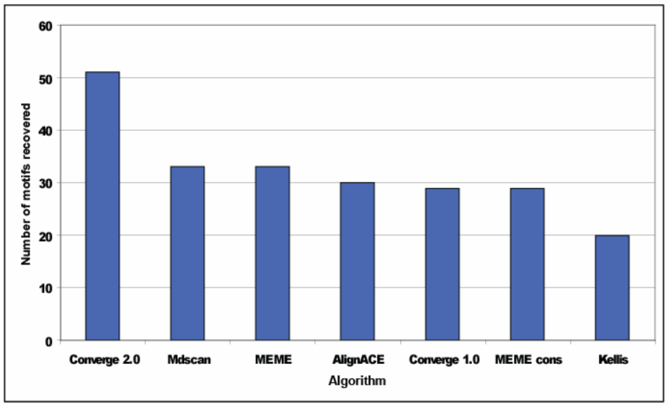


Figure 2.3: Number of known transcription factor binding specificities recovered by Converge and six previously reported motif discovery programs. Converge recovers more motifs than any of the suite of 6 programs employed in Harbison et al.

In total Converge recovered 51 sequence motifs matching the previously described specificity. This was more than any single program used by Harbison et al. and, in fact, was more than the combination of all six programs employed in that study, three of which made use of conservation information to aid in the motif search. In some cases, Converge's motifs differ substantially from the motifs reported in Harbison et al. For example the specificity discovered previously for Pho2 was SGTGCGsygyG. Converge predicts a specificity of AYTAAr. The new motif is more consistent with the results of gel shift and DNAse footprint analysis and with the fact that that Pho2 encodes a homeodomain protein [100], a class of transcription factors that tend to bind to AT-rich sequences. The factor Dal82 is now predicted by Converge to have a specificity of AAaNwTgyG, consistent with previously reported experimental evidence [101]. The motif reported in Harbison et al. (GATAAG) is likely to represent the binding specificities of Gln3, Gat1, and Dal80, which are known to co-regulate allophanate/oxalurate-dependent genes along with Dal82 [102].

One of the programs employed by Harbison et al. was a previous version of Converge that assumed a motif was always present in the aligned species when it was present in the primary genome. The improved ability of the newer version of Converge to recover correct motifs in these data (51 recovered vs. 29 for the older version) underscores the value of learning phylogenetic relationships through the θ parameters and making use of information in alignment gaps. A particularly striking example of this emerges from the analysis of Rds1 binding data. Converge determines that there is a very low probability that a match to the Rds1 motif will occur in *S. bayanus* in positions that contain the motif in *S. cerevisiae*. The θ parameter, which measures the genome-wide probability of observing a motif in *bayanus* when it is present in the primary genome, falls to 0.058. As a result, the *S. bayanus* sequences have almost no influence on the discovered motif. Interestingly, the Rds1 protein from *S. bayanus* is only 32% identical to its *S. cerevisiae* ortholog, compared to approximately 72% for other transcription factors in these two species. These data suggest that in *S. bayanus* Rds1 does not regulate the orthologs of the genes that are bound by Rds1 in *S. cerevisiae*, and that both the protein and its former binding sites have evolved.

We also compared Converge's performance on these data to results reported for a conservation-based approach that directly estimates mutation rates using a set of substitution matrices for motif and background in each species [103]. Li and Wong tested their algorithm on 53 data sets from Harbison, finding the correct motif in 39 of those cases, whereas Converge found the correct motif in 43 of these data sets demonstrating that our simple approach is, at worst, competitive with that of Li and Wong.

2.3.3 Comparison with PhyloCon and merging of motif results

We next wished to compile an expanded motif catalog by merging the results of Converge with a complementary conservation-based motif discovery algorithm, PhyloCon [104]. This complementarity arises from differences in the evolutionary assumptions made by each algorithm. PhyloCon dynamically realigns orthologous sequences, making no assumptions regarding the relative location of binding sites. However, it assumes that the sequences from each species should contribute equally to motif discovery. Converge, by contrast, assumes that the position of binding sites will be aligned in the orthologous sequences, but it makes no assumptions about the importance of the sequences from each species. We assessed the performance of each program, and the combination of programs, using empirical estimates of false positive, true positive, false negative, and true negative rates. True positives were defined as top-ranked statistically significant motifs that matched the known specificity. A false positive occurred when the top-ranked motif did not match the known specificity. A false negative was defined as the case when the program produced no statistically significant motif, but the correct specificity was discovered by another program (PhyloCon, Converge, or one of the six programs from Harbison). A true negative was defined as the case when the program produced no significant motif, and no other program was able to discover the known specificity.

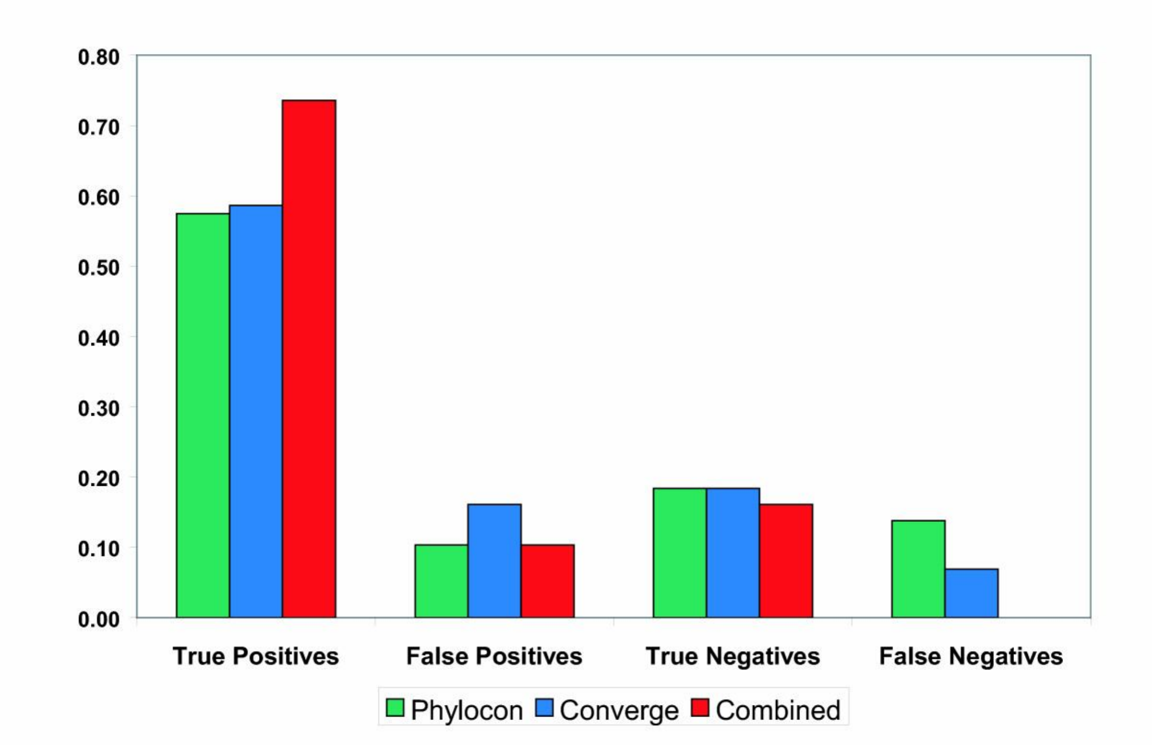


Figure 2.4 Performance of PhyloCon, Converge, and the combined motif set on data for factors of known specificity. Combining the results of PhyloCon and Converge increases the number of true positives recovered, and eliminates false negatives, without an adverse effect on the false positive rate.

Converge and PhyloCon have very similar performance with Converge showing somewhat greater sensitivity but less specificity than PhyloCon. Both Converge and Phylocon show significantly better performance than the combined results from the six programs used in Harbison. In Harbison *et al*., the predicted specificities derived from a combination of six programs matched the known specificities for 44 of the 87 proteins (51%). In contrast, Converge found 51 true positives (59%) and 14 false positives (16%). Converge was unable to find statistically significant motifs for 22 (25%) of these factors.

Combining the Converge and PhyloCon results allowed us to increase the number of transcription factors for which we could predict binding specificities with high-confidence. Our strategy for combining motifs was as follows: We first identified all motifs with a p-value < 0.001 for each program. We then identified the subset of motifs common to both programs and reported the motif with the best p-value (using the minimum p-value over both programs). If there were no significant motifs common to both programs, the most statistically significant motif from either program (p < 0.001) was reported. We discovered significant motifs for 98 of 172 factors. This is 33 more than were found by Harbison and co-workers, who used the same strict selection criteria. Of the 98 motifs, 43 were discovered by both programs, 22 were found only by PhyloCon, and 33 were discovered only by Converge. The discovered motifs were augmented with 26 factor specificities from the literature, to produce a final catalogue of 124 motifs.

2.4 An updated yeast regulatory map

Using the new catalogue of yeast specificities we built a more complete and comprehensive regulatory map for *Saccharomyces cerevisiae*. We scanned the *S.*

cerevisiae genome for putative regulatory interactions using the updated motif catalogue and the same criteria used by Harbison et al. As in that study, we restricted our analysis to the highest confidence sites, defined as those containing conserved motif matches that were bound by the corresponding factor at a p-value < 0.001. The new map contains a total of 4,229 conserved and bound motif sites across 2,022 genes, compared to the 3,353 sites across 1,883 genes in Harbison et al. The new and the old sets of motifs have similar information content (mean information content of 11.77 bits and information content per base of 1.24 bits in the new code, compared to 11.10 bits and 1.25 bits in the old code), suggesting that this increase is not due to an overall loosening of the specificity estimates.

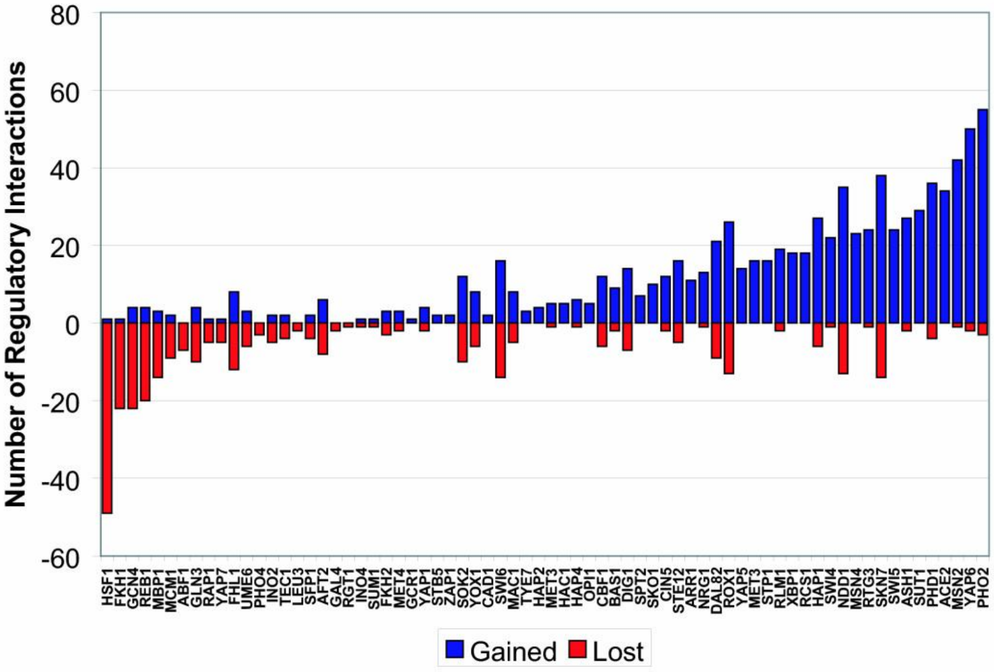


Figure 2.5: Changes in the number of putative regulatory interactions in the new yeast regulatory map. For each modified motif, the number of regulatory interactions added and lost relative to the previously reported map is shown. Our analysis produced modified factor binding specificities for 85 factors, resulting in a net gain of 398 putatively regulated genes.

Figure 2.5 and Figure 2.6 show the change in the number of bound genes by factor between the new and old maps. The net gain in the number of putative regulatory interactions between transcription factors and proteins is 636, with 133 of these accounted for by new binding specificity estimates for 18 factors that had no previously reported motif.

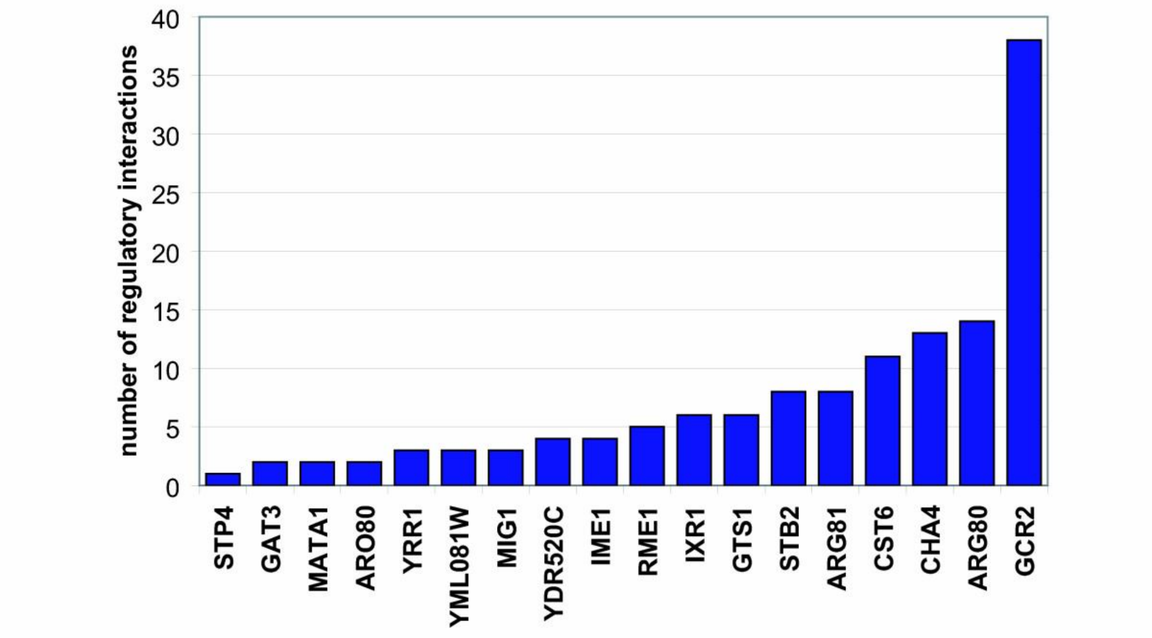


Figure 2.6: Regulatory interactions added through the addition of new factor specificity estimates. A total of 200 genes were identified as being putatively regulated by factors with newly reported motifs.

The new map reveals regulatory interactions for a number of transcription factors that are consistent with their known functions. For example, the refined motif for Msn2 detects regulatory sites in 39 genes that were not detected in the previous study. Msn2 is known to function in the transcriptional response to stress [105]. Of the newly identified targets, there is a significant ($p < 0.01$) over-representation of genes with the GO annotation "stress-response". Similarly, the refined Xbp1 motif results in a gain of 18 regulatory interactions. The new targets are enriched at a p-value < 0.02 for genes with the GO annotation "morphogenesis", consistent with a previously reported regulatory role for this transcription factor [106].

The revised map also provides new insights into the regulatory roles of several transcription factors. For example, the revised motif for Hap1 reveals that this transcription factor has an extensive role in regulating synthesis of ergosterol, a fungal-specific pathway that is a target for anti-fungal drugs. The previous map revealed regulatory interactions of Hap1 with genes for the ergosterol biosynthetic enzymes Erg5, Erg9 and Erg11. In the new map, we find interactions with genes for six additional enzymes in this pathway: Erg2, Erg8, Erg10, Erg25, Faa1, and Hmg1. In addition, the new map details an expanded role for Hap1 in regulating expression of components of the electron transport chain. Regulatory interactions with genes for two components of the cytochrome c oxidase complex, Cox7 and Cox8, were added to the three already present (Cox4, Cox6, and Cox13). A regulatory interaction with the gene for Qcr6, a component of ubiquinol cytochrome c reductase, was added to the previously reported interaction with the gene for Cor2, also a member of this complex. Finally, a Hap1 regulatory interaction with cytochrome c isoform 2, Cyc7, was added to previously discovered interactions with three other cytochromes, Cyc1, Cyb2, and Cyt1.

We examined the network of regulatory interactions between transcription factors in order to understand the system-level implications of the improved map. The previously reported regulatory code and the revised code were used to generate interaction networks for all the yeast transcription factors. This network is shown in Figure 2.7. Thirty-nine new interactions are reported, with six interactions lost. We searched the network for occurrences of six regulatory network motifs: autoregulation, feed-forward regulation, multi-component loops, single-input, multi-input, and regulatory chains [107]. The new network reveals several cases of feedback regulation that were not present in the previous version. The regulators Arg81, Rox1, Sut1, and Zap1 are all found to have an autoregulatory interaction in the new map. Of these, Rox1 [108] and Zap1 [109] have been previously shown to regulate their own expression.

The map also contains evidence of enhanced roles for a number of factors in the yeast transcriptional regulatory network. Yap6 acts as a regulatory hub, displaying five new interactions with transcription factors, three of which (Phd1, Sok2, and Hms2) are involved in pseudohyphal differentiation [110-112]. The stress-induced factor Xbp1, previously implicated in cell-cycle function [113], now has interactions with the pseudohyphal determinant Phd1, and Smp1, a factor required for cell viability in the stationary phase [114]. Table 2.1 details the regulatory motifs present in the new and old networks.

Table 2.1: Transcription factor network motifs in the old and new regulatory codes

Regulatory motif type	This study	Harbison et al.
Autoregulation	16	12
Multi-component loop	15	5
Feed-forward loop	71	55
Single-input motif	91	72
Multi-input motif	481	392
Regulatory chain	1452	168

There is an increase in the number of all six regulatory motif types, with a particularly striking increase in the number of regulatory chain motifs, owing to the motif's combinatorial dependence on the total number of interactions in the network. The overall picture that emerges from this analysis is of a more complex interplay of transcription factor influences in yeast regulatory networks than could be deduced from the previously reported regulatory code.

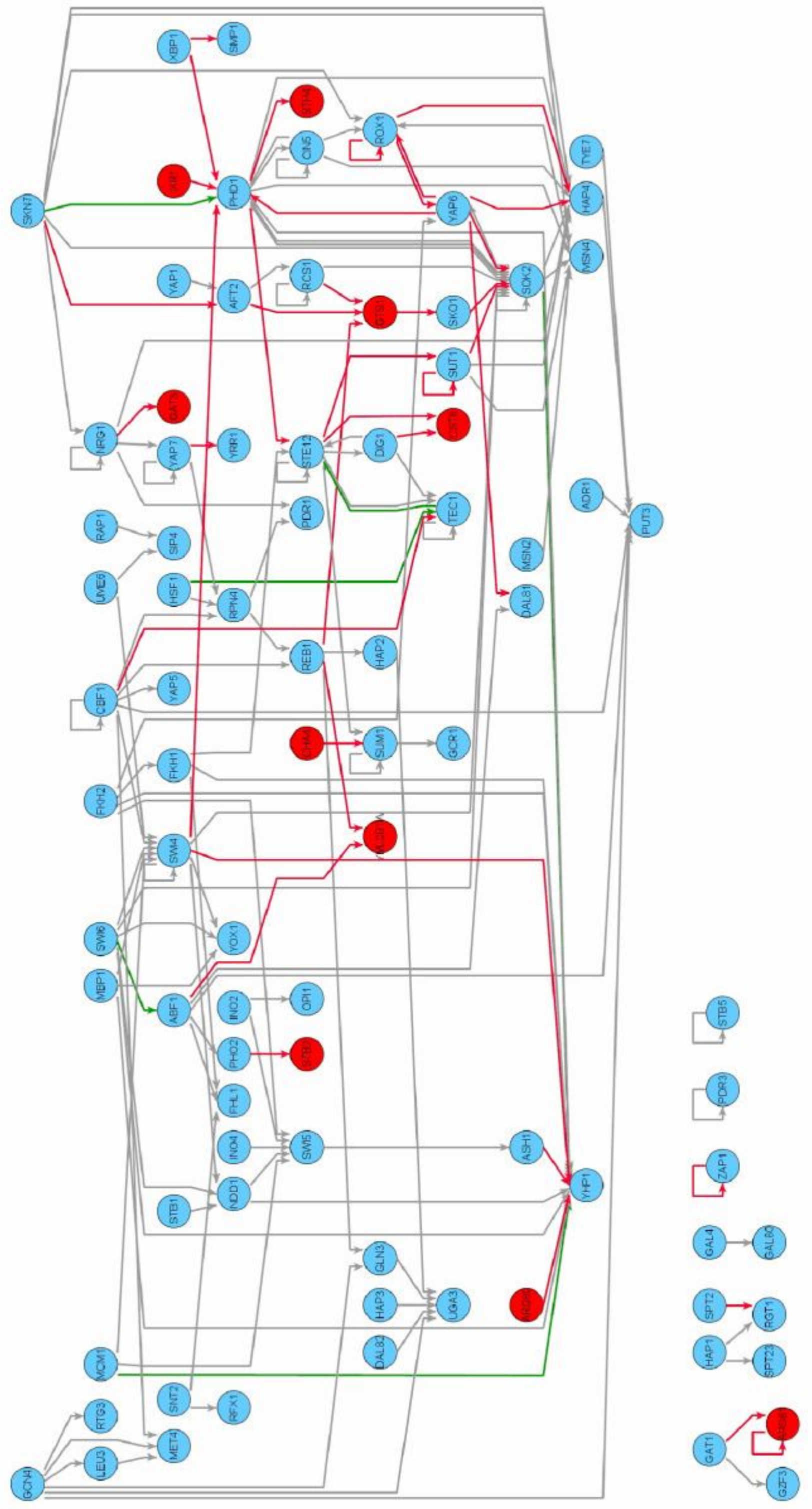


Figure 2.7: Yeast transcriptional regulatory network. Nodes correspond to transcription factors and an edge from one node to another indicates a putative regulatory interaction. Red nodes correspond to factors without a previously reported specificity. Edges are colored red for interactions unique to the new map, grey for interactions common to the old and new maps, and green for interactions unique to the old map. There are 39 new interactions and 6 interactions lost relative to the previously reported map.

2.5 Conclusion

In this chapter I have presented a framework for performing motif discovery using phylogenetic conservation information. This method is distinguished from similar approaches on several fronts. First, unlike many methods, the Converge algorithm incorporates conservation information directly into its probability model rather than as a pre-processing or post-processing step. Second, Converge is unique in explicitly make use of gaps in the pre-computed sequence alignments by weighting these regions differently during motif discovery, and third we learn a simple but meaningful measure of evolutionary distance between species that allows conservation information to be weighted differently across those species.

I have also demonstrated the application of the algorithm to real ChIP-chip data and shown how Converge's use of conservation information leads to an improvement in motif discovery performance, as measured by recovery of correct motifs for proteins with an experimentally characterized binding specificity. The results from analyzing RDS1 also demonstrate that the simple measure of phylogenetic distance we employ has real biological meaning and can provide insight into the evolution of regulatory networks across related species. Finally I have shown that merging the analyses of Converge and the PhyloCon program allows us to significantly expand the yeast regulatory map. This provided a different view of the regulatory role of several transcription factors, and showed that the regulatory network of transcription factors in yeast is more highly connected than previously thought.

Chapter 3: THEME

In this chapter I will present the THEME algorithm: a discriminative motif discovery method that tests specific, biologically informed hypotheses regarding the binding specificity of a protein and selects the best hypothesis using a principled cross-validation procedure. I will demonstrate that this technique performs exceptionally well on ChIP data from mammals and present several applications. Benjamin Gordon performed the groundwork necessary for the development of THEME by showing that informative, biological priors could dramatically improve the performance of a motif discovery tool. Benjamin and Lena Nekludova derived motif priors from binding site data for various binding domain families. Duncan Odom and Joerg Schreiber provided ChIP-chip experiments used to test the algorithm.

3.1 Hypothesis testing for motif analysis of ChIP data

Identification of functionally relevant motifs in genomes of higher eukaryotes is more challenging than in yeast. Regulatory regions are substantially larger and more complex, and sequence features common in mammalian genomes, such as CpG islands, further confound motif discovery methods. An evaluation of 13 motif discovery tools demonstrated the limitations of these techniques for analyzing mammalian promoter sequences [115]. At the same time, there is a need for robust motif analysis methods as an explosion in the quantity of mammalian ChIP-chip and ChIP-seq data is imminent, if not already upon us.

THEME is a hypothesis-driven method that is effective in identifying biologically meaningful sequence motifs from ChIP-chip data in human and mouse tissues. THEME uses principled statistical methods to test hypotheses about the binding specificity of an immunoprecipitated protein. It evaluates hypotheses based on their ability to accurately predict which sequences from a held-out test set were bound by the protein and which were not. The most predictive hypothesis is either accepted or rejected by comparing its predictive value to those of motifs derived by applying the same algorithm to randomly selected input sequences.

The hypothesis driven approach is particularly appealing since it allows us to merge information from the sequence of bound regions with prior biological knowledge when searching for motifs. By deriving initial hypotheses from the binding sites of related proteins in the TRANSFAC database [116] we can determine whether or not there is a motif that both explains the binding data and is consistent with the domain structure of the transcriptional regulator. Most DNA-binding domains show a limited repertoire of sequence specificity, and family members usually recognize variants of the same core sequences. For example, many bZIP proteins bind to variations of the AP-1 site (TGANTCA), the ATF-CREB (TGANNTCA) or the C/EBP site (ATTKC). Similarly, HLH proteins often bind to E-boxes (CANNTG), and differ largely in their specificity for the two middle base pairs and the flanking regions. THEME provides a method for determining if the specificity of the immunoprecipitated protein is similar but not necessarily identical to the prototypes for its family.

3.2 The THEME algorithm

An overview of the THEME workflow is shown in Figure 3.1. The initial hypothesis to be tested consists of a position weight matrix (PWM) model of the binding specificity, describing the probability distribution for bases at each position of a binding site.

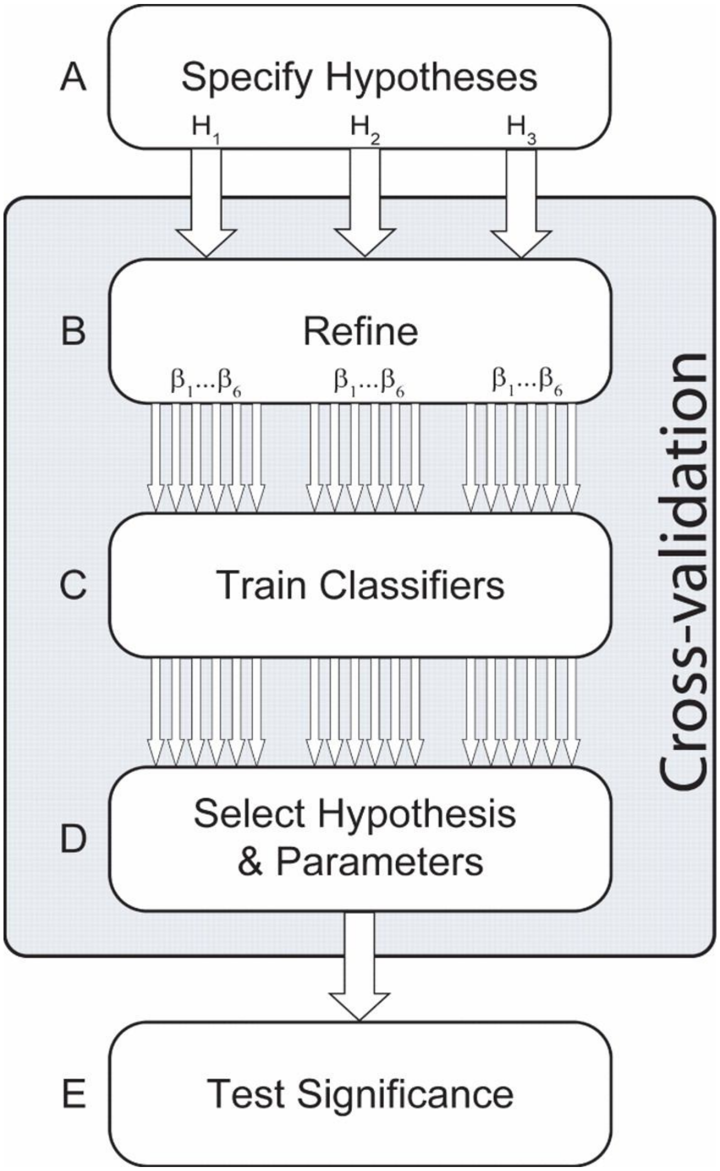


Figure 3.1: Overview of THEME. (A) THEME requires that one or more binding hypotheses be specified in the form of a frequency matrix. (B) The data are partitioned for cross-validation. Using only the training data, the hypotheses are refined using the EM algorithm. (C) The refined hypotheses are used to train a classifier and the classification error on the held-out test data is evaluated. (D) The hypothesis that yields the best mean cross-validation error is identified. (E) The statistical significance of the observed cross-validation error is estimated by comparing it with a distribution obtained by applying the hypothesis to randomly chosen promoter sequences.

Hypotheses can be derived from a variety of sources. Input consists of a set of sequences bound by the protein of interest (the positive data), as well as sequences that are not bound (the negative data). Using cross-validation, hypotheses are refined with training data and evaluated on held-out test data to identify the most predictive motif. The statistical significance of the best motif is then determined.

3.2.1 Hypothesis Generation

While hypotheses from any source can be tested, a particularly effective approach is to derive hypotheses using known binding sites of proteins that belong to the same DNA-binding domain family as the immunoprecipitated protein. Individual members of protein families generally bind related DNA sequences due to structural constraints. These preferences can be represented as PWMs and have been designated Family Binding Profiles [117]. Family Binding Profiles capture sequence features common to the binding sites of many members of the family, but are consequently poor representations of the specificity of individual family members.

We used profiles derived from unaligned binding sites in the TRANSFAC v7.2 database [116]. Pfam hidden-Markov models (Bateman et al., 2004) identify 37 families of DNA-binding domains in TRANSFAC with at least 4 proteins and 30 sites. Sites for a family were pooled and family binding profiles were generated using two motif discovery programs: AlignACE (Roth et al., 1998) and DimerFinder [118]. On average, a family is represented by three profiles. To demonstrate the utility of this approach even when there are no close homologs of a protein of interest, we used profiles derived using only the binding site data for proteins with <70% sequence identity to the DNA-binding domain of the protein of interest. For example, the Family Binding Profiles that we use to discover the specificity of HNF4α exclude binding data for all HNF4α, HNF4β and HNF4γ proteins from any species. RXRβ2 is the most similar protein to HNF4α that is included in the profiles.

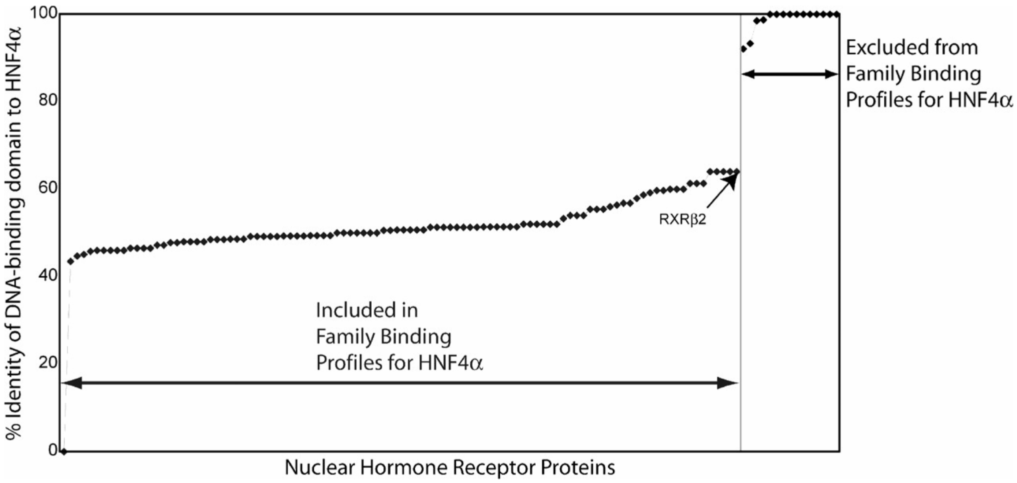


Figure 3.2: Similarity of the nuclear hormone receptor DNA-binding domains to HNF4α. The graph shows the percent identity between the DNA-binding domain of HNF4α and each nuclear hormone receptor protein in TRANSFAC. Proteins with >70% sequence identity were excluded when the Family Binding Profiles were derived.

3.2.2 Hypothesis testing by cross-validation

The Family Binding Profiles for each protein are refined and tested using cross-validation to find the hypothesis that best explains the binding data. We define the set of bound probe sequences in a ChIP experiment as the positive set. We produce a negative set by randomly undersampling the set of unbound probes until it is 10 times larger than the positive set. We partition the sequences into test and training sets and perform THEME hypothesis testing using the following five-step procedure:

1. Refine the hypothesis on the positive training set
2. Score each sequence in the training and test data using the refined model
3. Oversample the positive training and test data
4. Train a classifier on the training examples
5. Classify the test examples and report the classification error.

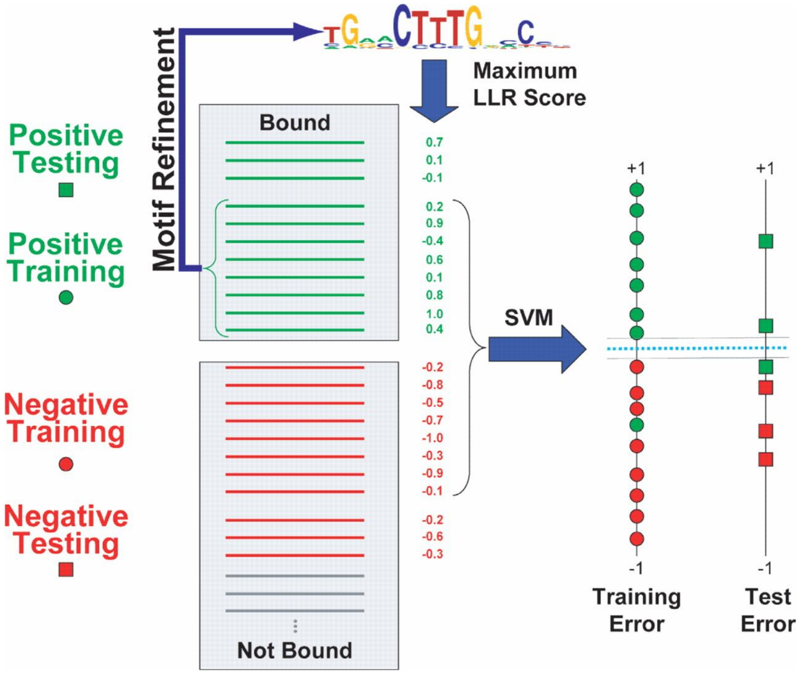


Figure 3.3: Hypothesis refinement and cross-validation by THEME. Positive sequences were bound in the ChIP experiment. The remaining sequences on the array are the negative set. Data is divided into training and test sets. The positive training data are used to refine the hypothesis. All training and test examples are then mapped to a one-dimensional feature space by evaluating the LLR score of their best match to the refined hypothesis. A classifier is trained using both the positive and negative training examples and used to evaluate the classification error on the positive and negative test sets.

Each hypothesis is refined on the positive data using a standard motif discovery algorithm to ensure it represents the motif signal present in the data as much as possible. We used the ZOOPS probability model and optimize the motifs using the EM algorithm [66]. Each hypothesis acts as the initialization point for EM. The E and M steps are alternated until the Euclidean distance between the motif models in subsequent M steps was less than 10–3. In the M step, motif is updated using the expected counts in each position of the matrix. The change in the motif during the M step is restrained using pseudo counts added to the matrix in proportions determined by the original hypothesis and the value of the β parameter. β is defined as the fraction of the total counts added to the matrix during the M step that are pseudo counts used to restrain the model. A β of 0.0

indicates that EM refinement proceeds without restraint. When β = 1.0, no refinement is carried out. Refinements occur in parallel with β values of 0.05, 0.1, 0.33, 0.5, 0.67 and 1.0. The Markov background model used in EM was estimated from the set of all sequences represented on the microarray for a given experiment.

In order to train a classifier and perform cross-validation, the refined hypotheses must be used to define one or more features used to score the sequences. THEME uses the log-likelihood ratio (LLR) score of the best match to the refined hypothesis. This score is an intuitive feature that measures our belief that the best match is an instance of the motif, described by the PWM model, after taking into account the single-nucleotide base distribution of the background sequences. Typically, an arbitrary threshold is used to determine when the LLR is high enough to constitute a match to a motif [99]. THEME uses a more principled approach, by training a very simple linear-kernel support vector machine (SVM) to determine the threshold that best separates the bound and unbound sequences in the training data. The scores of the training data are scaled so that they fall between –1.0 and 1.0, and used to train the SVM at a particular setting of the parameter, *C*, which is used in the regularization term. The test data are then scaled in an identical manner and classified using the SVM. The classification error of the SVM on the test data is evaluated using the optimal value of *C* determined from the training data.

When building classifiers from datasets with a significant imbalance in the proportion of positive and negative examples, it is important to ensure that the classifier has sufficient sensitivity to the minority class. One solution is to resample the dataset to achieve greater balance between the two classes. We combine undersampling of the negative dataset with SMOTE oversampling of the positive training and test sets so that the number of positive and negative examples is equal. This technique has been shown to improve classification performance on datasets with large disparities in the sizes of the minority and majority classes [119].

For each hypothesis we perform a grid search over the two parameters β and C. We repeat the five-step procedure at each parameter setting to find the setting yielding the lowest 3-fold cross-validation error. Due to the non-deterministic nature of the sampling procedure, cross-validation results could, in principle, vary among trials with the same input and parameters. To test this, we compared hypotheses using three separate THEME trials with different randomly selected negative datasets. The refined motifs did not vary significantly across these trials. Here we report the average cross-validation errors. The best refined motif model is the one that has the lowest mean error on the test sets after 3-fold cross-validation.

3.2.3 Determining statistical significance

For the best candidate family binding profile we determine the empirical distribution of mean cross-validation errors, under the null hypothesis that the input sequences are unrelated to the profile, by running THEME multiple times using sets of randomly selected sequences. These sets are equal in size to the original dataset, and the calculations are conducted using the same parameter settings. We assume the observed cross-validation errors are normally distributed and perform randomization runs until the standard error on our estimate of the standard deviation is ~10%. We then compare the observed cross-validation errors to the computed distribution and perform a Z-test to assess the statistical significance of the refined hypothesis.

3.3 Performance of the THEME algorithm

We tested THEME by applying it to published ChIP-chip experiments for 14 human transcriptional regulators, which are members of 9 different DNA-binding domain families. These data are quite diverse and thus constitute a good set of experiments with which to evaluate THEME. Initial hypotheses were generated using Family Binding Profiles derived from the TRANSFAC binding sites, excluding data for close homologs as described.

Each profile was refined using the positive training data for the most strongly bound genes (binding P-value < 0.001). The mean test errors for these hypotheses after 3-fold cross-validation are shown in Table 3.1 below. In each case, the refined hypothesis with the best cross-validation error is statistically significant and agrees with previously reported motifs or binding sites for the protein.

NeuroD1 illustrates the power of THEME when there is little prior knowledge about the DNA-binding specificity of a protein or that of its close homologs. The most similar protein that has known binding sites in TRANSFAC (v7.2) is the T-cell acute lymphocytic leukemia-1 protein, SCL/TAL1, which is only 48% identical to NeuroD1 in its DNA-binding domain. Nevertheless, we find a motif, sCAgcTGs, which is statistically significant, present in 97% of the bound probes on the mouse array and consistent with known sites for NeuroD1 in the promoter of Pax6 [120].

Table 3.1. Family Binding Profiles and Associated Refined Motifs

Protein	Profile	Refined Motif	Mean Cross-Validation Error
c-Rel	gGGr.tTyC	gGGr.tTyC	0.35
c-Rel	krGAAAa.y	.gGrAAwcc	0.42
c-Rel	GGaawttCC	GGaawttCC	0.34
c-Rel	GGawwtCC	GgrwwycC	0.38
c-Rel	GGGgAwTcCCC	gGGrawtyCCc	0.35
E2F4	GCGssaaa	GCGssaaa	0.35
HNF3b	arTAAACA	.GYaAACA	0.39
HNF3b	kTTGTT	gkyGTt	0.46
HNF3b	TGTTTrTT	TGTTtrY.	0.44
HNF4α	..RGGTCA	marGGyCA	0.40
HNF4α	rGwaCA...tGTwC	rg.rCw..rkGkmC	0.48
HNF4α	aGaACA...TGTtCt	aGaACa...tGTtCt	0.46
HNF4α	AGGTCAc.gTGACCT	.gG.cwc.gwg.Cc.	0.42
HNF4α	AGGTCATGACCT	rGkyC..GrmCy	0.42
HNF4α	tcAAGkTCAag	tcaaGgtCaag	0.44
HNF4α	TGACCT...kTGACCT	tkaCCyymw.tkmyCy	0.43
HNF4α	TGACCTTTGACCyy	tGgmCytTGmCcy.	0.30
HNF6	ATCGAT.s	ATCGAT.s	0.32[1]
HNF6	CAcm.Ata..TaTkG	CAcm.Ata..TaTkG	0.47
HNF6	CgATcG	cgATcg	0.43
HNF6	cgATCGAT	cgATCGAT	0.32[1]
Nanog	TAATTrsy	tAAtkrsy	0.42
Nanog	AAgyrcTT	AAgyrcTT	0.43
Nanog	AaT.AtT	Aak.mtT	0.44
Nanog	TAATt.aATTA	taat...atta	0.44
Nanog	TAATTAat	tAAtkr.t	0.44
NeuroD1	cCACGTGg	cCamktGg	0.42
NeuroD1	CgCaCGC	CgCaCGC	0.46
NeuroD1	rCAgcTGy	rCAgcTGy	0.35
NeuroD1	tCACGTGa	tCACGTGa	0.44
Oct4	ATGCAAAT	ATGCAAAt	0.40
Oct4	TAAwTTA	kaAwTtm	0.44
p50	GraAw.cCCm	GGraAwyCCC	0.30
p52	GGrAw.yCCc	GGrAw.yCCc	0.28
p52	GGaawttCC	GGaawttCC	0.30
p52	GGawwtCC	GGawwtCC	0.33
p52	GGGgAwTcCCC	GGGrawtyCCC	0.21
p65	GGrAw.mCCc	ssRrAwycCc	0.40[1]
p65	GGGGAwTCCCC	sggrawtyccs	0.40[1]
P-CREB	rTGACgyr	rTGaCGy.	0.44
P-CREB	ttrtGYAA	tkrcGtMA	0.44
P-CREB	caCGTGGc	caCGTGGc	0.47
P-CREB	mCACGTGk	w.aCGt.w	0.45
P-CREB	aTGACGTCAt	aTgACGTcAt	0.40
P-CREB	aTGAsTCAt	.w.msk.w.	0.49
P-CREB	aTTg..cAAt	.wwscgsww.	0.46
P-CREB	gcCACGTGgc	.ysaCGtsr.	0.41
P-CREB	GtG.CaC	skkwmms	0.50
P-CREB	gTGacGTG	rTGaCGt.	0.43
P-CREB	TtACGTaA	TkaCGtmA	0.41
P-CREB	tTGCAa	tyGCra	0.48
RelB	GGrAw.yCCc	GGrAw.yCCc	0.30
RelB	GGaawttCC	GGrawtyCC	0.32
RelB	GGawwtCC	GGawwtCC	0.39
RelB	GGGGAwTCCCC	gGGrawtyCCc	0.33
Sox2	AACAAWRr	AACAAwrr	0.39

[1] For two factors, HNF6 and p65, the two best profiles tested gave very similar mean cross-validation errors. We note that in both cases the refined motifs are also quite similar.

3.3.1 The importance of hypothesis testing

Leveraging prior biological knowledge is crucial for successfully identifying the correct motif in mammalian datasets. To demonstrate this, we ran the THEME algorithm using an uninformative hypothesis, equal in length to the correct motif, but consisting of background nucleotide frequencies. The uninformative hypotheses produced the correct motif in only one case (HNF6). The cross-validation error for motifs derived from uninformative priors was always higher than when Family Binding Profiles were used. Of the motif discovery programs that we tested on these data, AlignACE performed the best, discovering motifs consistent with the known specificities in six cases (Table 3.2). The cross-validation errors for AlignACE motifs were always higher than those discovered by THEME. To obtain the AlignACE results, we needed to run the program multiple times using different random number seeds. A typical AlignACE calculation required 21 h to complete, compared with 18 min for THEME.

3.3.2 Deriving hypotheses with limited prior data

In the absence of Family Binding Profiles THEME can take advantage of other available data, such as known binding sites. To demonstrate this, we derived hypotheses from each of the three known and distinct NeuroD1 binding sites [120] by assigning 99% of the probability mass to the nucleotide represented in the sequence and distributing the remaining mass among the other 3 nt at each position and tested them with THEME. The refined motifs, shown below in Table 3.3, match the NeuroD1 motif reported in Table 3.1 and display similar cross-validation errors.

Table 3.2: Importance of hypothesis testing

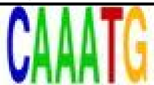
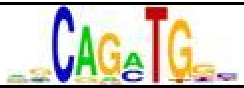
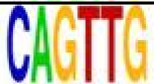
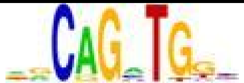
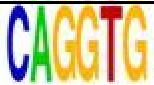
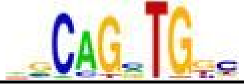
Factor	THEME: Uninformative Hypothesis		AlignACE		THEME
	Motif	Mean 3-fold CV error	Rank[1]	Mean test error[2]	Mean 3-fold CV error[3]
c-Rel	Not Found	0.46	Not Found	0.40	0.34
E2F4	Not Found	0.36	Not Found	0.39	0.34
HNF3b	Not Found	0.47	Not Found	0.47	0.39
HNF4	Not Found	0.40	Not Found	0.48	0.30
HNF6	Found	0.34	Not Found	0.50	0.32
Nanog	Not Found	0.45	Not Found	0.47	0.42
NeuroD1	Not Found	0.49	1	0.44	0.35
Oct4	Not Found	0.43	1	0.45	0.41
p50	Not Found	0.40	1	0.32	0.30
p52	Not Found	0.42	1	0.26	0.21
p65	Not Found	0.45	Not Found	0.46	0.40
P-CREB	Not Found	0.43	Not Found	0.47	0.40
RelB	Not Found	0.46	1	0.33	0.30
Sox2	Not Found	0.44	3	0.44	0.39

[1] Rank of motif matching known specificity

[2] AlignACE motifs were ranked by hypergeometric enrichment score. THEME was used without refinement to evaluate the classification error of the top-ranked AlignACE motif. In the case of Sox2, the motif that matched the known specificity was used in place of the top-ranked motif.

[3] Cross-validation error for best THEME results shown in Table 3.1.

Table 3.3: NeuroD1 results obtained using hypotheses derived from single binding sites

Binding Site	Initial Hypothesis	Refined Hypothesis	Optimal β	Mean 3-fold CV Error
CAAATG			0.05	0.34
CAGTTG			0.05	0.32
CAGGTG			0.05	0.36

In many cases, THEME is able to identify the correct motif, even if the DNA-binding domain or binding sites of the factor are not specified. To demonstrate this, we ran THEME for each factor in Table 3.1, using every profile across all families as initial hypotheses. We ranked the resulting refined motifs by their cross-validation errors. In 10 out of 14 cases, we observe that the correct motif, derived from a hypothesis corresponding to the factor's DNA-binding domain family, has the lowest cross-validation error (Table 3.4). Furthermore, in 13 out of 14 cases, the correct motif and the correct family were ranked in the top 5 families (the correct family for Nanog was ranked 8th out of 36 families).

Table 3.4: Top-ranked Family Determined by THEME after Testing with Profiles from All Families[1]

Factor	PFAM Family	Hypothesis	Refined Motif	Mean 3-fold CV error	Rank
c-Rel	PF00554 (RHD)	GGrAw.yCCc	GGrAw.yCCc	0.34	1
E2F4	PF02319 (Winged helix)	GCGSsAAa	GCGssAAa	0.30	1
HNF3b	PF00250 (Forkhead)	rYAAACAa	ryAAACA.	0.41	1
HNF4	PF00105 (NHR)	TGACCTTTGACCyy	tGgmCytTGsCcy.	0.28	1
HNF6	PF02376 (CUT)	cgATCGAT	srATCgAT	0.31	1
Nanog	PF00172 (Zn clus)	CGGm.ga.	CgG.....	0.41	1
	PF00046 (Homeobox)	TAATTrsy	yAAtkrsy	0.43	8
NeuroD1	PF00170 (bZIP)	gcCACGTGgc	rsCAgcTGsy.	0.38	1
	PF00010 (HLH)	cCACGTGg	sCAgcTGs	0.41	4
Oct4	PF02257 (RFX)	GTTGCya.G..am	.ttgw.atg..aa	0.40	1
	PF00157 (POU)	ATGCAAAT	ATGcaaAt	0.41	4
p50	PF00554 (RHD)	GGGGAwTCCCC	GGGrawtyCCC	0.22	1
p52	PF00554 (RHD)	GGGGAwTCCCC	GGGGAwTCCCC	0.23	1
p65	PF00554 (RHD)	GGGGAwTCCCC	sggrawtyccs	0.35	1
P-CREB	PF00170 (bZIP)	aTGACGTCAt	.TgACGTcA.	0.40	1
RelB	PF00554 (RHD)	GGrAw.yCCc	GGrAw.yCCc	0.29	1
Sox2	PF02376 (CUT)	cgATCGAT	racAAw.g	0.37	1
	PF00505 (HMG)	AACAAWRr	AACAAwrr	0.41	5

[1] The top-ranked motif is always shown. In those cases where this motif is derived from a family other than that of the immunoprecipitated protein, the results for the expected family are also shown. Similarities between these motifs and the top-ranked motif are indicated by the underlined letters.

THEME does not require highly accurate initial hypotheses. To demonstrate this we used THEME to refine noisy versions of the hypotheses that yielded the lowest cross-validation error for each factor. We obtained these hypotheses by combining, in various ratios, 1000 sequences derived from the uncorrupted PWM and from the background base frequencies. Noise levels of up to 40% have little effect on the cross-validation errors and in 13 of the 14 datasets the motifs obtained with 40% noise are consistent with the known specificities.

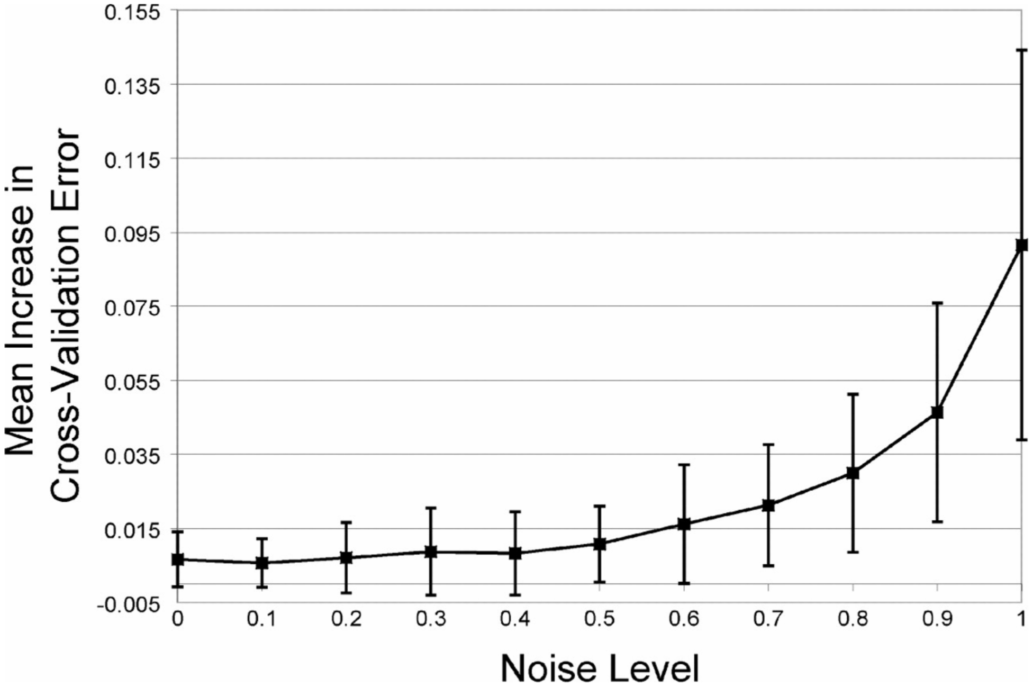


Figure 3.4: Effect of noise on cross-validation error. The Family Binding Profiles yielding the lowest cross-validation error for each dataset were corrupted with varying amounts of noise to produce matrices of gradually decreasing quality. These were used as hypotheses in THEME. The mean cross-validation error for the refined motif from each hypothesis is compared with the best hypothesis for the same dataset.

3.4 Conclusions

In this chapter I have presented THEME: a hypothesis-driven approach to analyzing ChIP-chip data that differs from standard motif discovery programs in that it begin by specifying biologically-informed hypotheses, and then establishes whether these hypotheses are supported by the data. THEME is able to determine whether to accept or reject a hypothesis because it seeks to solve a classification problem. A good motif distinguishes between bound and unbound sequences in the test set. An incorrect hypothesis may produce a motif that appears significant on the training data, but it will be poorly represented in the test data. THEME was the first discriminative motif analysis method to employ cross-validation to rank candidate motifs and to protect against overfitting.

THEME's unique hypothesis-testing framework is particularly valuable because it addresses the issue of interpreting motifs. THEME not only assesses whether there is a motif that can distinguish bound and unbound sequences, but also whether that motif is consistent with prior biological knowledge. When prior biological knowledge is available, either in the form of a known DNA-binding domain or known binding sites (as for NeuroD1), the accuracy of THEME is dramatic. THEME identifies a statistically significant motif consistent with the expected specificity for all 14 datasets we analyzed. In contrast, using the cross-validated approach without an informative prior fails to identify the correct motif in all but one of these mammalian datasets. In the absence of information about the DNA-binding domain of the protein, THEME is often able to identify the correct motif by exhaustively testing all available Family Binding Profiles. These results suggest that THEME may be a valuable tool in the analysis of diverse data.

Chapter 4: A biophysically motivated framework for motif analysis

In this chapter I will present an extension of some of the ideas introduced in Chapter 3, reformulated into a probabilistic framework based on biophysical principles. I will demonstrate how this approach leads to a particularly straightforward and interpretable motif discovery algorithm and then present results of its application to ChIP data in a variety of mouse and human tissues. I also present extensions to the model for comparing binding specificity and concentration across growth conditions for the same protein and for analyzing binding data for proteins that compete for the same binding site. Previously unpublished ChIP-chip and ChIP-seq data presented in this chapter were collected by William Gordon, Alice Lo, and Shmulik Motola.

4.1 From frequency matrices to affinity matrices

Standard frequency matrix or consensus sequence motif models ignore the role of protein concentration in DNA-protein interaction. While on the surface this may appear to be a minor drawback, it has important implications. Once a motif has been identified, it is often of interest to identify potential regulatory regions by scanning genomic regions for matches to the binding specificity. However, the probability that a particular binding site will be occupied by a regulator strongly depends on the nuclear concentration of that regulator, especially for weaker binding sites. It is also unclear how to handle biological phenomena like competition between regulators for a common binding site using standard models. An alternative approach treating protein-DNA interactions thermodynamically would allow these concerns to be addressed in a natural way.

The feasibility of modeling protein-DNA interactions thermodynamically has been demonstrated in several different contexts. Very detailed structure-based methods that predict energetic interactions between protein and binding sites allow prediction of a protein's sequence-specific binding affinity without the need for a training set of binding site sequences [121, 122]. Notably, these studies continue to indicate that common simplifying assumptions regarding the dependence of binding energy on DNA sequence (e.g. positional independence) are reasonable for most DNA-binding proteins. In fact, starting from this assumption it can be shown that under conditions of binding site saturation the information theoretic log-odds position weight matrix is a matrix of scaled binding free energy contributions [123, 124]. Adapting a biophysically-based approach to motif discovery would thus appear to be, if not a straightforward extension of previous approaches, at least a natural one. Indeed, Tsang and coworkers presented an algorithm that moved toward this goal by estimating a matrix of binding energy contributions (position-specific affinity matrix, or PSAM) from a set of bound sequences in a ChIP-chip experiment [54]. However in Tsang et al.'s method energy contributions are not directly interpretable as thermodynamic parameters. Furthermore they ignore protein concentrations and assume only one binding event is possible per sequence when calculating binding probabilities. Nevertheless, they were able to show that their approach was superior to the frequency matrix based MEME algorithm and the alignment-based AlignACE algorithm for discovering motifs with very weak sequence signals.

Djordjevic and colleagues introduced the QPMEME algorithm which estimates a PSAM and a chemical potential for a transcription factor based on a set of pre-defined bound examples [53]. They showed that their binding model was superior to the information matrix approach in the context of binding site identification in *E. coli*. Their method considers the effect of concentration in the form of a chemical potential which is closely related to the best threshold for classifying bound and unbound sites. However, they assume that known regulatory sites are bound with probability very near 1 under physiological conditions. Furthermore, they do not make explicit use of unbound examples but rather assume that random DNA sequences have a Gaussian distribution of binding energies. They then perform a constrained optimization, minimizing the probability that random sites are bound subject to the constraint that observed sites are bound. Their all or none approach is not appropriate for analyzing binding of regulators for which modulation of the binding probability (e.g. by altering the concentration) plays a physiological role.

The MatrixREDUCE algorithm is used to analyze protein binding data and estimate a PSAM that is directly interpretable as binding free energy contributions [125]. MatrixREDUCE assumes that ChIP ratios are linearly proportional to the occupancy of a sequence and estimates binding energies by performing a least-squares fit to the intensity ratios. Foat *et al.* showed that these estimates are in good agreement with *in vitro* experimental binding affinity measurements and the predictions of structure-based models. However, the MatrixREDUCE algorithm does not directly model the effect of protein concentration on binding and assumes that this concentration is very small relative to the dissociation constant of the protein-DNA complex. This may not always be a valid assumption for low affinity sites.

Biophysically-based models have also been used in the context of expression pattern prediction in the developing fruit fly [126]. Segal and coworkers used transcription factor concentration data and estimates of their binding specificities in a model that predicted both transcription factor binding occupancy in promoter regions, as well as downstream expression effects. They validated their model's performance by testing its ability to recapitulate observed spatial gene expression patterns for held-out test modules. In a similar vein, Gertz et al. used biophysical modeling to explain the ability of a library of synthetic yeast promoters to drive expression in yeast [127]. They found that such models could explain a large fraction of the variance in observed expression levels and revealed the importance of weak and/or cooperative binding events. These results suggest that biophysical approaches have significant promise in capturing the behavior of transcriptional regulatory systems; however neither the work of Segal et al. nor that of Gertz and coworkers uses direct evidence of protein binding to train their models. Methods for directly learning and testing biophysical models of protein-DNA interaction from high-throughput ChIP data would thus represent a useful contribution to the field.

4.2 A biophysical model of DNA-protein interaction

In this section we present a very simple model of protein binding to DNA that makes no unwarranted assumptions about protein concentration level in the nucleus. In later sections we will show how this framework may be adapted for use in motif discovery and the analysis of ChIP data. We assume that protein, *A*, binding to a site, *B*, to form a complex, *C*, can be modeled as a bimolecular reaction at equilibrium. We further assume that the nucleus acts as a constant pressure and volume "reaction vessel". The equilibrium constant of the reaction is:

$$K_a = \frac{[C]}{[A][B]} \tag{4.1}$$

This leads to a simple expression for the probability, *p*, that protein and DNA form a complex:

$$p = \frac{[C]}{[C]+[B]} \tag{4.2}$$

$$\log \frac{p}{1-p} = \log[A] - \log K_a \tag{4.3}$$

and since,

$$\log K_a = {\Delta G}/{RT} \tag{4.4}$$

then

$$p = \frac{1}{1+\exp\left(-\log[A] + {\Delta G}/{RT}\right)} \tag{4.5}$$

In equation 4.4, *ΔG* is the free energy of the binding reaction relative to the unbound state. We now make the simplifying assumption that the free energy (scaled by the temperature and gas constant) can be expressed as the sum of contributions of individual nucleotides, x_i. Replacing the protein concentration term with β_0 yields the logistic function:

$$p = g(X) = \frac{1}{1+\exp\left(-\beta_0 - \sum_i \beta_i x_i\right)} \tag{4.6}$$

The logistic form of equation 4.6 immediately suggests a straightforward method for estimating the position-specific binding energy contributions of each nucleotide as well as the nuclear protein concentration. If we were provided with a representative distribution of bound and unbound genomic sites, we could find maximum likelihood estimates of the *β* parameters by simply training a logistic regression classifier to distinguish them.

4.3 The THEME+ algorithm

THEME+ is an extension of THEME that adapts the biophysical framework presented above to perform motif analysis of ChIP data. In THEME+, the separate steps of motif optimization and classifier training are replaced with a joint motif optimization and classification procedure. THEME+ is similar to the QPMEME algorithm in that it learns a motif and concentration that discriminates between bound sites and unbound background. However unlike QPMEME, THEME+ makes explicit use of unbound regions and does not require that the precise binding site of the factor be known. It is thus suitable for analyzing ChIP data, where there is uncertainty about the exact location(s) of protein binding within an immunoprecipitated region.

4.3.1 THEME+ Probability Model

We have a set of sequences with associated labels, *y*, taking on the value of 1 if a sequence was bound in the ChIP experiment and 0 otherwise. For a motif with width *w*, each sequence of length *L* contains 2(*L*-*w*+1) potential binding sites (on both the forward and reverse complement strands). We ignore any steric constraints that may exclude overlapping binding sites. The probability that such a sequence will not be bound anywhere is given by the product of the probabilities that each individual site, with nucleotide content X_i, is not bound:

$$P\left(y=0\right)=\prod_{i}^{2(L-w+1)}\left(1-g\left(X_i\right)\right) \qquad (4.7)$$

The log probability of the label data given the motif model is thus:

$$\log P=\sum_i\left[\left(1-y_i\right)\sum_j\log\left(1-g\left(X_j\right)\right)+y_i\log\left(1-\prod_j\left(1-g\left(X_j\right)\right)\right)\right] \qquad (4.8)$$

Maximization of this expression in terms of the motif parameters is complicated by the log of ‘1 minus the product of probabilities’ term. We circumvent this difficulty by augmenting the likelihood function 4.8 with the hidden variables, *Z*, which indicate which positions in each sequence are bound by the protein of interest:

$$\begin{aligned}\log P(Y,Z) &= \log P(Y\mid Z)+\log P(Z)\\ &= \sum_i y_i \log q_i + (1-y_i)\log(1-q_i)\\ &+\sum_i\sum_j z_j \log g(X_j)+(1-z_j)\log\left(1-g(X_j)\right) \qquad (4.10)\end{aligned}$$

$$q_i = \begin{cases} 1 & \sum_j z_{i,j} \geq 1 \\ 0 & o.w. \end{cases}$$

The first term in equation (4.10) expresses constraints on the label for a given binding configuration (i.e. the probability of observing y = 1 is 0 unless the sequence is bound in at least one position, and the probability of observing y = 0 is 1 if the sequence is unbound, and 0 otherwise). The second term is simply the probability of the binding configuration given the energy and concentration parameters. Our strategy is to use Expectation Maximization to obtain estimates of the parameters that maximize the probability of the observed data labels.

4.3.2 Expectation Maximization procedure

E step

In the E-step we need to calculate the expected likelihood function given the label data and the current motif parameters. For sequences with y = 0, this is simple since all hidden variables are 0 with probability 1. For sequences with y = 1, we must calculate:

$$
\begin{aligned}
E\left[\log P\left(Z \mid y\right)\right] &= \sum_i \sum_j E\left[z_j\right] \log g\left(X_j\right) + \left(1 - E\left[z_j\right]\right) \log\left(1 - g\left(X_j\right)\right) \\
&= \sum_i \sum_j E\left[z_j\right]\left(\beta^T x_{i,j} - \log\left(1 + e^{\beta^T x_{i,j}}\right)\right) - \left(1 - E\left[z_j\right]\right) \log\left(1 + e^{\beta^T x_{i,j}}\right) \qquad (4.11) \\
&= \sum_i \sum_j E\left[z_j\right] \beta^T x_{i,j} - \log\left(1 + e^{\beta^T x_{i,j}}\right)
\end{aligned}
$$

For a particular binding site, the expected value of z is:

$$
E\left[z_{i,j} \mid y_i\right] = \frac{P\left(y_i \mid z_j = 1\right) P\left(z_j = 1\right)}{\sum_{z_j} P\left(y_i \mid z_j\right) P\left(z_j\right)} = \frac{P\left(z_j = 1\right)}{1 - \prod_j \left(1 - g\left(X_{i,j}\right)\right)} \qquad (4.12)
$$

M step

In the M-step we maximize the expected likelihood (4.11) in terms of the motif model parameters. The convenient logistic form of *g(X)* means that this is equivalent to training a logistic regression classifier to distinguish bound sites from unbound sites using the nucleotides at each position as the predictive features. The difference being that, instead of hard labels, the classifier uses the soft binding labels calculated in the E-step.

As in the THEME algorithm, we restrain the model parameters during optimization. This is accomplished by placing a Gaussian prior on each regression parameter and estimating their values using Bayesian logistic regression. For the energy terms, the prior mean is set to the log-odds of each nucleotide in the original motif PWM hypothesis. The concentration is initialized to the value that yields optimal separation of bound and unbound examples given the initial motif. The variance of the prior is treated as a regularization parameter that is selected using an internal round of cross-validation during the M-step. We use the previously reported algorithm of Genkin et al. to perform Bayesian logistic regression [128].

4.3.3 Performance and results

We examined the performance of the THEME+ algorithm on several ChIP-chip and ChIP-seq datasets from a variety of tissues in both human and mouse. For each data set we first identified all bound regions from the ChIP-chip or ChIP-seq experiment. For previously published ChIP-chip data we used the regions reported as bound from each respective study. For unpublished ChIP-seq data we used the MACS algorithm with a p-value cutoff of 1e-6 to identify bound regions. Unbound negative examples were obtained by randomly selecting unbound genomic regions, taking care to match the sequence length distribution as well as the distribution of distances from nearby transcription start sites. The negative and positive datasets were of equal size. For each protein, we ran the THEME+ algorithm, testing family binding profiles from the appropriate DNA-binding domain family of the protein. Each motif hypothesis was optimized with 5 iterations of EM and evaluated by its mean 3-fold cross-validation error. In order to evaluate the performance of the algorithm when the DNA-binding domain of the protein is not known *a priori*, we evaluated the entire library of 105 family binding profiles on each dataset and determined the rank of the motif matching each protein's true DNA binding specificity. The results are summarized and compared to the original THEME algorithm in Table 4.1. The THEME+ algorithm performs well on these data, identifying a motif consistent with the known binding specificity of the protein in all cases when we restrict the set of starting hypotheses to family binding profiles derived from the protein's binding domain. When we test all profiles, THEME+ still performs well, with the correct motif ranking first for 17 of 22 datasets. Its performance is comparable, and perhaps slightly better, than the standard THEME algorithm.

Table 4.1 Performance of THEME+

Protein	Tissue	Top THEME+ motif	Mean cv error	Rank	Top THEME motif	Mean cv error	Rank
C/EBPα	mouse liver	TkrCGymA	27%	1	aTTg..cAAt	29%	1
E2F4	mouse liver	GCGssAAa	22%	1	GCGssAAa	24%	3
FOXA2	mouse liver	ryAAACAa	39%	1	ryAAACA.	41%	1
FOXP3	mouse CD4+ T-cells	rtAAACAn	35%	2	rYAAACAa	37%	2
HNF4α	mouse liver	nnrGgtca	38%	10	tgacCTytGacCy.	40%	6
pCREB	mouse liver	grTGACGy	27%	1	.tgaCGtca.	27%	1
PPARγ	mouse 3T3-L1	TGACCTTTGACCyy	31%	1	tgaCCTtTgaCCy.	27%	1
RXR	mouse 3T3-L1	anrGGtCA	36%	4	tgaCCTyTgaCCy.	32%	1
c-Rel	human U937	GgAwwTCC	36%	1	GGrAw.yCCc	34%	1
E2F4	human HepG2	GCGcsAAA	35%	2	GCGssAAa	35%	1
FOXA2	human liver	ryAAACAa	37%	1	ryAAACA.	39%	1
HNF4α	human liver	TGACCTTTGACCyy	32%	1	tGgmCytTGsCcy.	30%	1
HNF6	human liver	saATCGAT	30%	1	srATCgAT	32%	1
p50	human U937	GGGgAwTcCCC	28%	1	GGGrawtyCCC	30%	1
p52	human U937	GGGgAwTcCCC	25%	1	GGGGAwTCCCC	21%	1
Nanog	human ES cells	tAATTrat	42%	2	yAAtkrsy	42%	8
NeuroD1	mouse MIN6	CAgcTG	29%	1	sCAgcTGs	38%	4
p65	human U937	GGGGAwTcCCC	39%	1	sggrawtyccs	40%	1
RelB	human U937	GGAawtTCC	34%	1	GGrAw.yCCc	30%	1
Sox2	human ES cells	aaCAAwgn	35%	1	AACAAwrr	39%	5
Oct4	human ES cells	AtGCaaak	40%	1	ATGcaaAt	40%	4
pCREB	human islets	snTGaCkt	37%	1	.TgACGTcA.	40%	1

For eleven transcription factors THEME+ finds a motif with better cross-validation error than THEME, whereas the reverse is true for six datasets. When the entire set of 105 family binding profiles are tested, THEME+ ranks the correct motif higher than THEME for 5 datasets, whereas THEME gives a better rank to the correct motif for 3 datasets. These results suggest that the biophysical model of protein-DNA binding is well-suited for motif analysis of ChIP data. The slight differences in performance between THEME+ and THEME may be attributable to the differences in the assumptions made by the two approaches. In THEME, the possibility of false positives in the bound data is built into the generative model of sequence used to learn the motif model. In contrast, THEME+ does not account for this possibility and assumes that every sequence in the positive set is bound by the protein. The original THEME algorithm might therefore be expected to perform better on noisier datasets. A second difference is that THEME assumes a single binding site for the protein in each bound sequence. Classification is based on the single best match to the motif in each sequence. THEME+ makes no such assumption and includes the contributions of every potential binding site in the sequence. Thus for proteins with multiple weak binding events in immunoprecipitated regions THEME+ might be expected to demonstrate improved performance.

4.4 Incorporating ChIP-seq count data

A key limitation of the THEME+ algorithm is that it uses hard labels corresponding to bound or unbound regions and ignores the fact that, due to the dynamic nature of protein-DNA association, binding level is actually a continuum of fractional values. When we perform a ChIP experiment we are measuring a signal arising from this fraction of bound sites. It is reasonable to assume that there is information about transcription factor concentration and binding site strength in the ChIP signal. In fact, several groups have observed that ChIP-chip ratio signals and ChIP-seq count data appear to be related to *in* *vivo* binding occupancy [129, 130]. In this section we extend the probabilistic framework presented in section 4.3 to better take advantage of the information present in raw ChIP-seq count data. We begin by reviewing, in Figure 4.1, the basic experimental workflow of a ChIP-seq experiment:

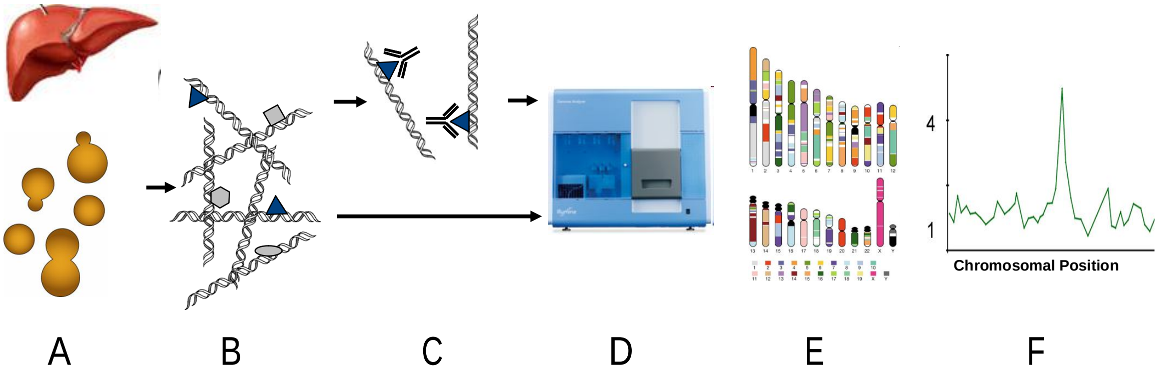


Figure 4.1: ChIP-sequencing workflow. In a ChIP-seq experiment cells from a tissue of interest are isolated (A) and a chemical cross-linking agent is used to covalently link amino groups resulting in persistent association of protein to genomic DNA or other protein with which it associates *in vivo* (B). Genomic DNA is isolated, fragmented, and enriched for bound sites by immunoprecipitation using an antibody specific to the protein of interest (C). Unenriched whole-genome DNA, or DNA obtained by mock-IP, is reserved as a control. The cross-linking procedure is reversed, the DNA is purified and is then sequenced (D). The resulting sequence reads are aligned to a reference genome (E). Finally, genomic regions enriched for reads in the IP channel relative to the control are identified and reported.

To extend our model to use raw ChIP-seq counts we treat the genome as having been divided up into a set of sequence regions. Rather than a set of binding labels, in this setting the observed data is a set of counts: the number of individual sequence tags that align to a given genomic region. As shown in Fig 4.1, in order for an individual binding event to be detected as a count in a ChIP-seq experiment the protein must be bound to the site in the cell, cross-linked to the DNA, immunoprecipitated by an antibody, sequenced, and aligned to the reference genome. At each step in this process, there is a certain loss. Starting from a total population of identical binding sites in *N* cells, only a small fraction will be detected as a tag count in the ChIP-seq experiment. In addition, some counts will arise from unbound or non-specifically bound background DNA. The probability that a particular genomic location in the population will be detected as a tag is given by:

$$\begin{aligned} f\left(detect\right) &= \sum_b P\left(detect \mid bound\right) P\left(bound\right) \\ &= \theta p + \phi\left(1-p\right) \end{aligned} \tag{4.13}$$

In equation (4.13) the probability the site is bound *in vivo* is given by *p*. The probability a sequence tag is detected given that the site is bound is denoted by θ, and ϕ is the probability of detecting a tag given that there is no binding.

The probability of detecting a sequence tag that aligns to a genomic region given that there is no binding could, in theory, be estimated on a region-by-region basis from the control experiment where DNA obtained from a mock-IP, or unenriched whole genome DNA, is sequenced. This estimation is difficult since it depends on knowing the total number of cells in the starting sample. Although this could be measured experimentally, in general this information is not available to us. We therefore use an alternative strategy to account for background binding. We take the control experiment and linearly scale the number of reads so it is the same as in the IP, and then subtract these scaled control reads from the IP reads to obtain an estimate of the reads arising from immunoprecipitated DNA. We now introduce a modified version of equation (4.13) that gives the probability of observing an aligned sequence tag in the vicinity of a binding site, after background subtraction:

$$\begin{aligned} f\left(detect\right) &= \theta p + \phi\left(1-p\right) \\ &= \left(\theta - \phi\right) p + \phi \\ &\approx \theta p + \phi \end{aligned} \tag{4.14}$$

In equation 4.14 we make the assumption that the probability of observing a tag count (after background subtraction) given the site is bound is much greater than the probability of observing a count given that the site is not bound. The ϕ parameter now accounts for the presence of any residual background sequence tags that are not eliminated by background subtraction. For a particular region in the genome, we now express the probability of observing *k* aligned sequence tags from *N* total cells using the binomial distribution:

$$P(k)=\binom{N}{k}(\theta p+\phi)^{k}(1-\theta p-\phi)^{N-k} \tag{4.15}$$

Employing the Poisson approximation to the binomial distribution allows us to express this probability as:

$$\begin{aligned} P(k) &= \frac{\left(N(\theta p+\phi)\right)^{k} e^{-N(\theta p+\phi)}}{k!} \\ &= \frac{(\gamma p+a)^{k} e^{-\gamma p-a}}{k!} \end{aligned} \tag{4.16}$$

The observed tag count can thus be viewed as the sum of two independent Poisson random variables: the first with firing rate γp arising from bound instances in the population of cells, and the second with a firing rate of *a* that arises from unbound background.

4.4.1 Unified Probability Model

The probability of observing a given tag count number in a sequence region depends on the fractional *in vivo* occupancy of the protein of interest in that region (*p* in equation 4.16). This fractional occupancy is assumed to be a function of the motif and protein concentration. Specifically we assume that the probability a sequence is not bound anywhere is given by (4.7), that the probability it is bound is 1 minus this value, and further that in equilibrium the binding probability is equivalent to the fractional occupancy of that region across the entire population of cells. Rather than directly estimating the parameters of our model from the data, we again employ an expectation maximization approach to decouple the parameter estimation step. This approach will allow us to easily extend the model to analyze binding data for multiple proteins as discussed below. We again introduce hidden variables indicating where the protein of interest binds in each region. Here we wish to estimate fractional occupancy, so unlike in section 4.3 where a single set of hidden variables was used, we introduce *M* virtual copies of each sequence region; each copy has its own set of indicator variables that specify its binding configuration:

$$\begin{aligned} \log P(Y,Z) &= \log P(Y \mid Z) + \log P(Z) \\ &= \sum_i k_i \log(\gamma p_i + a) - \gamma p_i - a - \log(k_i!) \\ &\quad + \sum_i^N \sum_j^{2(L-w+1)} \sum_k^M z_{i,j,k} \beta^T x_{i,j} - \log\left(1 + e^{\beta^T x_{i,j}}\right) \end{aligned} \tag{4.17}$$

$$p_i = \frac{1}{M} \sum_k^M q_{i,k}$$

$$q_{i,k} = \begin{cases} 1 & \sum_j z_{i,j,k} \geq 1 \\ 0 & o.w. \end{cases}$$

As before, parameters are learned using an iterative EM-like procedure. Given the expected value of the hidden variables, estimation of the motif and concentration parameters in the M step proceeds as before, by training a logistic regression classifier to distinguish bound sites from unbound sites. Estimation of the parameters γ and a can be accomplished by numerically solving the equations:

$$\begin{aligned}&\sum_i p_i\left(\frac{k_i}{\gamma p_i + a} - 1\right) = 0\\&\sum_i \frac{k_i}{\gamma p_i + a} = N\end{aligned} \tag{4.18}$$

In the E step we employ a sampling strategy to obtain estimates of the hidden variables given the observed ChIP tag count data. In the section that follows we describe this strategy in detail.

4.4.2 Modified Stochastic Simulation algorithm

To obtain expected binding occupancies given the tag count data in each genomic region we employ a sampling method, inspired by the Gillespie stochastic simulation algorithm [131], which can be derived by viewing each genomic region as a chemical system at equilibrium. Imagine we have a well-mixed reaction vessel where i=1…N different reactions can occur. Gillespie showed that the time of the next reaction of type i is distributed exponentially:

$$\begin{aligned}P(\tau) &= h_i c_i \exp(-h_i c_i \tau)\\&= a_i \exp(-a_i \tau)\end{aligned} \tag{4.19}$$

Where h_i is the number of distinct combinations of reactants that can combine according to reaction i, and c_i is the rate constant. Assuming constant chemical potentials, each reaction is an independent Poisson arrival process and therefore the time of the next reaction of any type is also a Poisson process with inter-arrival times distributed exponentially:

$$P(\tau) = \sum_{i=1}^{N} a_i \exp\left(-\sum_{i=1}^{N} a_i \tau\right) \tag{4.20}$$

We can therefore view the unfolding of a reaction time course as a Poisson splitting process with inter-arrival times distributed according to (4.20). Each arrival is sent to an individual reaction channel, *i*, with probability:

$$p_i = \frac{a_i}{\sum_{j=1}^{N} a_j} \tag{4.21}$$

The execution of a reaction changes the chemical potential for other reactions so this Poisson process is non-homogeneous. However, by sequentially drawing samples from these two distributions and updating the a_i's we arrive at the well-known Gillespie stochastic simulation algorithm (SSA) for exactly sampling the trajectory of a reacting system.

Here we modify the SSA to obtain samples from an approximation to the posterior equilibrium distribution of binding events. The reactions that must be modeled are simply protein association and dissociation from DNA, however in general this scheme could be expanded to include other reactions like association of proteins to form a bound complex. The original Gillespie algorithm assumes that all relevant rate parameters are known and then proceeds by iteratively selecting the next reaction based on these rate constants and the current number of reactant molecules (i.e. the chemical potential). To obtain the rate constant for binding and dissociation reaction we make two observations: the relative rates of the forward and reverse reactions at equilibrium are given by the equilibrium constant which we can calculate for any binding site using the motif and concentration parameters of equation (4.6), and secondly, since we are concerned only with the equilibrium behavior of the system rather than its time-course kinetics, these relative rates are all that is required in our simulation as long as sampling is run for a time course long enough to achieve equilibrium.

We incorporate posterior binding evidence by calculating an effective chemical potential for the reaction. Consider the case where we have evidence, *X*, regarding the equilibrium state of the system. We wish to bias the reaction time course so that reactions more consistent with the evidence are favored. We accomplish this by altering the Poisson splitting process. Rather than splitting the arrivals according to (4.21), we split according to the posterior probability of a reaction given the evidence:

$$\begin{aligned} P\left(r_i \mid X\right) &= \frac{P\left(X \mid r_i\right)P\left(r_i\right)}{\sum_{j=1}^{N} P\left(X \mid r_j\right)P\left(r_j\right)} \\ &= \frac{P\left(X \mid r_i\right)a_i}{\sum_{j=1}^{N} P\left(X \mid r_j\right)a_j} \end{aligned} \tag{4.22}$$

Near the posterior equilibrium point the probability of the evidence given that the next reaction is r_i will be approximately equal for each reaction, (4.22) will be approximately equivalent to (4.21), and we will approach the standard SSA. Intuitively, this strategy achieves a balance between the likelihood of the observed count data and prior information in the form of the current binding free energy and concentration parameters. Reactions consistent with *both* the binding specificity of the protein and the posterior evidence will be favored.

The argument that follows suggests that this scheme allows us to obtain samples from a reasonable approximation to the posterior distribution over binding configurations: The stochastic simulation algorithm allows for exact simulation of the time course of any chemical system.

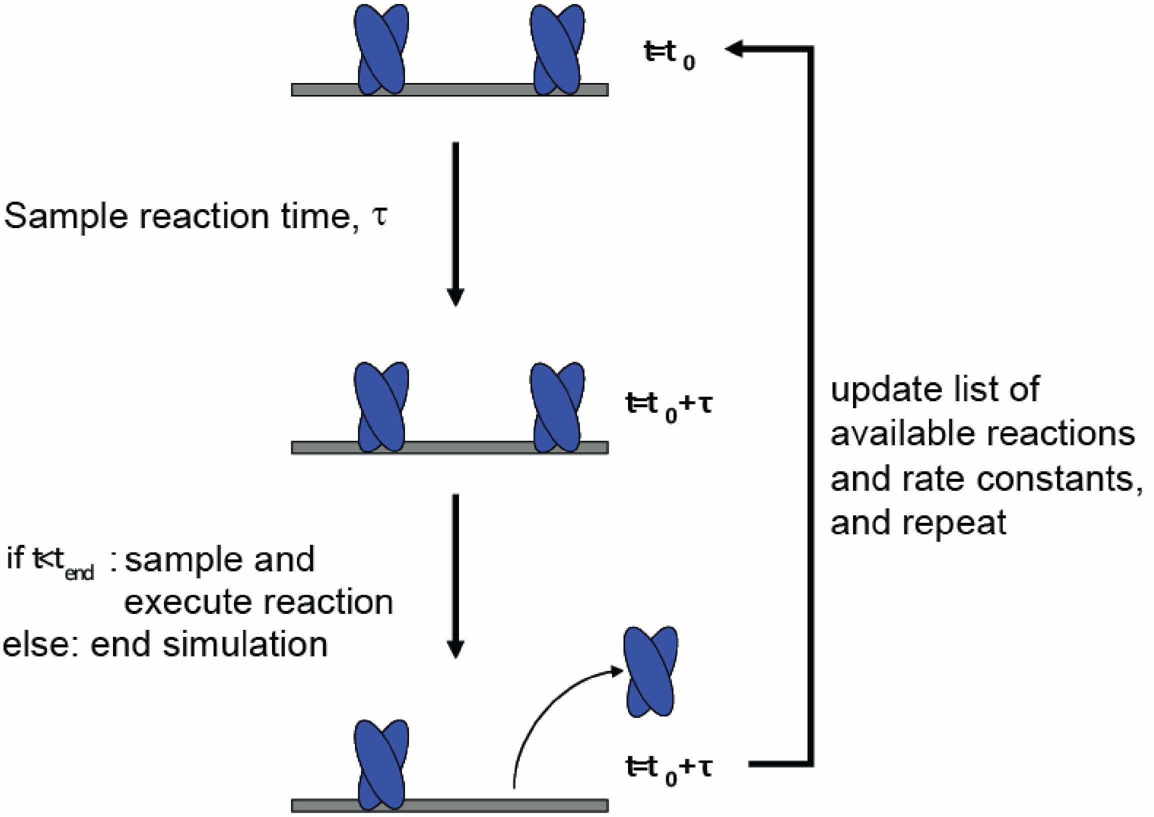

t=t 0
Sample reaction time, τ
t=t 0+τ
update list of available reactions and rate constants, and repeat
if t<t_end : sample and execute reaction
else: end simulation
t=t 0+τ

Figure 4.2: Stochastic simulation algorithm schematic. The stochastic simulation algorithm is used to obtain samples from an approximation to the posterior distribution of binding configurations. The algorithm first selects a reaction time according to equation (4.20) followed by random selection of a binding or dissociation reaction according to equation (4.22). After reaction selection, the reactant numbers and rate constants are updated and the procedure is repeated until the end time, t_{end}, is reached.

Thus if an equilibrium point exists, running the SSA for a time course long enough to achieve equilibrium allows us to obtain samples from the Boltzmann distribution over the states of the system, *i*:

$$\frac{N_i}{N} = \frac{e^{-E_i/(k_B T)}}{Z} \tag{4.23}$$

In equation (4.23) the total number of states (binding configurations) is *N*, E_i refers to the free energy of state *i*, k_B is the Boltzmann constant, and *Z* is the normalization constant obtained by summing over all configurations. Equation (4.23) is equivalent to the prior probability distribution over the hidden binding variables shown in equation (4.17):

$$\frac{N_i}{N} = \frac{\prod_i z_{i,k} e^{\log[A]-\Delta G/RT} \prod_i 1/\left(1+e^{\log[A]-\Delta G/RT}\right)}{\sum_j \prod_i z_{i,j} e^{\log[A]-\Delta G/RT} \prod_i 1/\left(1+e^{\log[A]-\Delta G/RT}\right)}$$
$$= \frac{[A]^p \prod z_{i,k} e^{-\Delta G/RT}}{[A]^p \sum_j \prod z_{i,j} e^{-\Delta G/RT}} = \frac{e^{-E_i/(k_B T)}}{Z} \tag{4.24}$$

Now, if we introduce another energy term that corresponds to the “likelihood energy” of each configuration, we arrive at an expression equivalent to the joint likelihood of the observed count data and hidden binding variables:

$$\frac{N_i}{N} = \frac{e^{-\log P(X|i)} e^{-E_i/(k_B T)}}{Z} \tag{4.25}$$

Simulating a chemical system with an equilibrium partition function equal to equation (4.25) should therefore allow us to obtain samples from the posterior distribution of configurations given the observed count data. Our sampling strategy is, in fact, equivalent to simulating such a system. Imagine grouping binding and dissociation reactions into sets that all lead to identical values for the data likelihood. We now treat each binding and dissociation event as a two-step process where one of these reaction sets is first selected, and then a particular reaction from the set is executed:

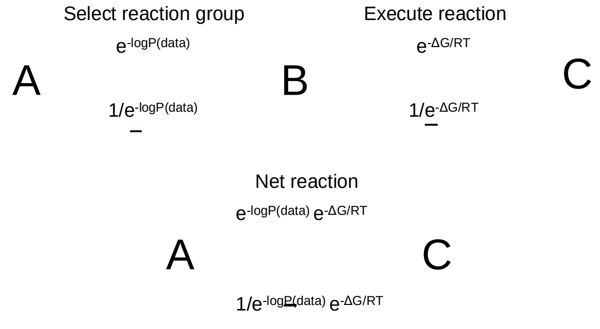

Select reaction group
$e^{-\log P(\text{data})}$
$1/e^{-\log P(\text{data})}$
Execute reaction
$e^{-\Delta G/RT}$
$1/e^{-\Delta G/RT}$
A
B
C
Net reaction
$e^{-\log P(\text{data})}\, e^{-\Delta G/RT}$
$1/e^{-\log P(\text{data})}\, e^{-\Delta G/RT}$
A
C

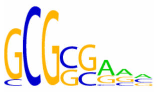
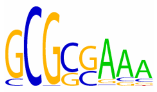
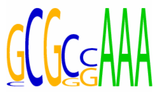
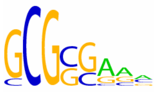
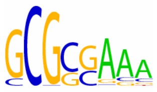
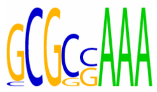
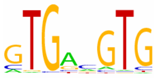
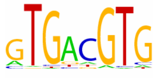
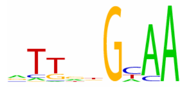
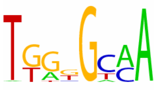
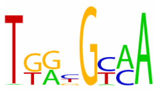
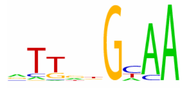
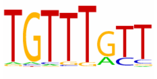
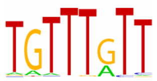
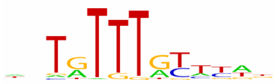
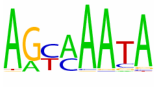
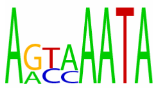
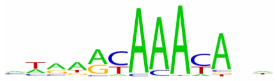
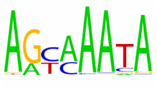
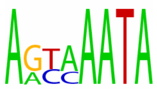
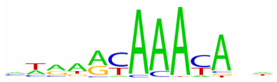
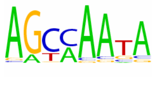
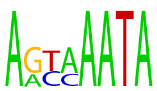
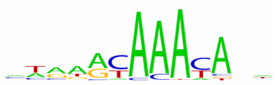
Figure 4.3: The reaction selection strategy can be represented as a reaction with two free energy components: one related to the contribution of the data likelihood and one related to the *a priori* favorability of the reaction given the binding site sequence.

If the “group selection” step has free energy equal to the data log-likelihood for that group, then the net free energy for any group selection followed by binding/dissociation reaction is given by $\Delta G_{net} = \Delta G_{rxn} + \log P\left(Data \mid rxn\right)$, and the relative rates of the forward and reverse reactions are given by $e^{\Delta G_{rxn} + \log P\left(Data \mid rxn\right)}$ as shown in Figure 4.3 above. This leads to selecting reactions according to equation (4.22) during the SSA and suggests that, provided simulations are run long enough to achieve equilibrium, our method should do a reasonable job of sampling from the posterior distribution of binding configurations.

4.4.3 Results

We tested the algorithm on ChIP-seq datasets for several different transcriptional regulators in two different tissues. For each factor, a set of ChIP-enriched regions was first identified by running the MACS algorithm [46] with a low stringency p-value cutoff threshold of 1e-5. We then determined the number of counts that aligned to each region. We subtracted out *kN* reads based on the control data, where *N* is the number of control reads aligning to the region of interest and *k* is the ratio of total IP reads to total control reads that passed the sequencer manufacturer’s quality filters. The background subtracted count data and each region’s DNA sequence was provided as the input to our algorithm. Our strategy for analyzing each dataset is similar to the strategy employed by THEME+. We test a set of starting motif hypotheses, optimizing the motif by taking advantage of the count data, and evaluating the optimized motif models using their cross-validated likelihood scores. The motif yielding the best mean likelihood score after cross-validation is reported. The results are summarized below:

Table 4.2: Top-ranked motifs obtained from analysis of raw ChIP-seq count data

Experiment	Top-ranked motif	Starting hypothesis	Previously-reported binding specificity
E2F4 3T3-L1 cells			
E2F4 liver			
C/EBPα 3T3-L1 cells			
C/EBPα liver			
FOXA1 liver			
FOXA1 liver (high fat diet)			
FOXA2 liver			
FOXA2 liver (high fat diet)			

For all experiments we detect a statistically significant relationship between the predicted binding probability (calculated from the optimized motif parameters) and the observed count data. Figure 4.4 below shows 100-point moving average plots of the mean number of counts for held out test regions, sorted by predicted binding probability. These plots validate the basic assumption behind our approach: namely that the quantity of *in vivo* binding is related to the affinity a protein has for a particular genomic region.

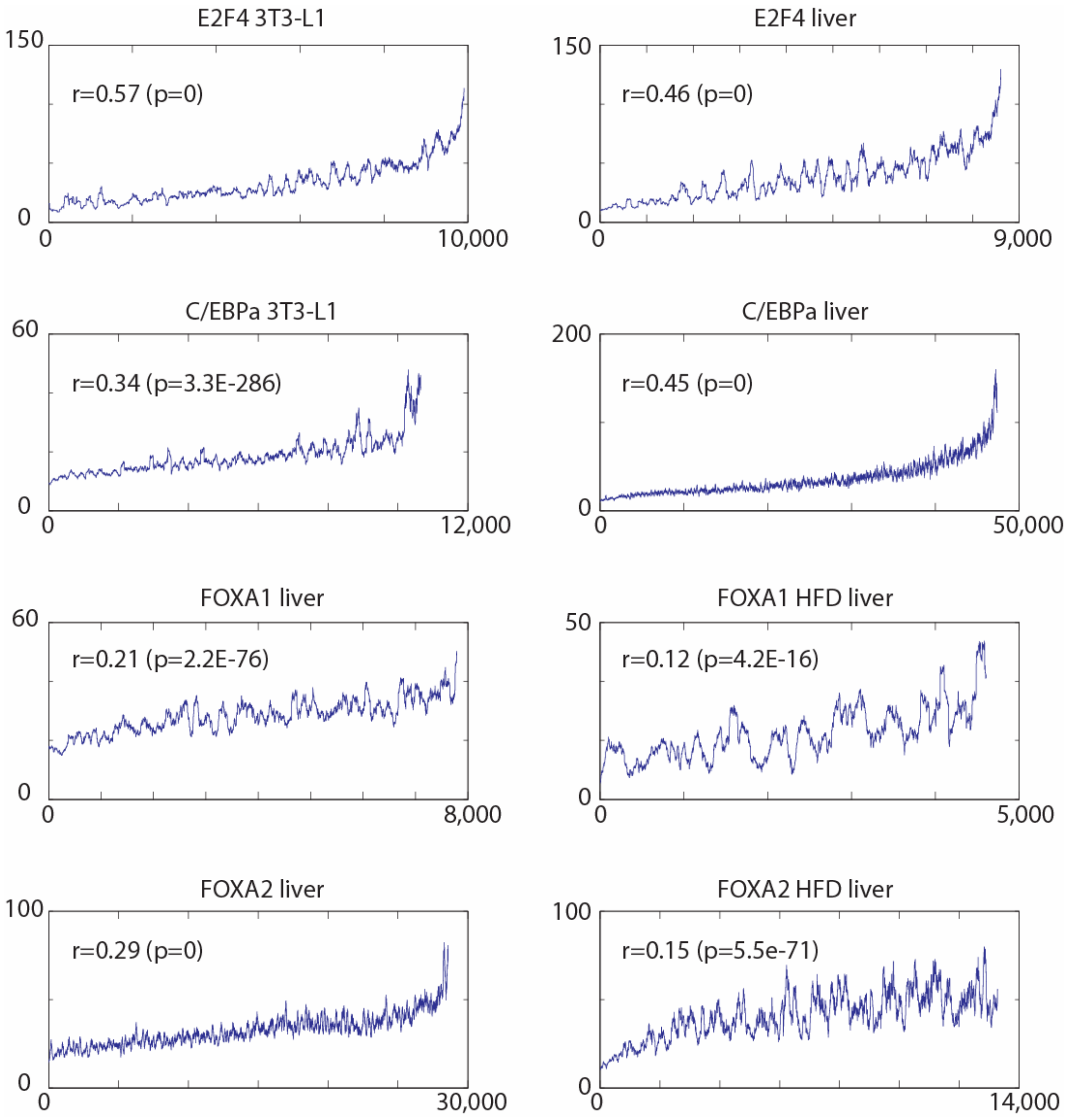


Figure 4.4: Motifs reported for each experiment are significantly correlated with ChIP-seq count data. Held out test regions were scored using the final motif parameters obtained from training. Test regions were then sorted by predicted binding probability. We then calculated a 100-point moving average of ChIP-seq counts for the sorted regions, which is shown above for each experiment. Also shown is the Spearman rank correlation and associate p-value from a two-tailed t-test between binding score and ChIP-seq tag counts.

After model training, we can perform a final round of sampling on all the sequence regions in order to estimate posterior binding occupancies. This leveraging of motif information could, in principle, help weed out false positives or assign more confidence to weakly bound regions with good motif matches. However, this type of strategy should be approached with caution since, in addition to interacting with the genome through their DNA-binding domain, many transcription factors are recruited to their genomic targets via protein-protein interactions with other regulators. Another difficulty in using posterior binding estimates to improve the identification of bound regions in a ChIP experiment is that, in general, we lack of a ‘gold standard’ set of bound and unbound regions with which to evaluate the change, if any, in accuracy over standard approaches. In principle, binding predictions could be validated by performing chromatin immunoprecipitation followed by a set of gene specific PCR experiments. In the absence of such data, one can attempt to roughly estimate performance by evaluating the predictions accordance with other biological data sources. One imperfect method is to identify genes located near the identified binding events and then to evaluate their functional coherence using the Gene Ontology [132] or biological pathway information [133]. This assumes that each protein will regulate a functionally coherent set of targets (which may or may not be true). This method is also sensitive to the quality of the annotation data used to assess functional coherence. An alternate strategy is to evaluate the expression of target genes under the assumption that regulator binding is associated with a consistent effect on expression across all targets. Again, this is a strong assumption since many regulators are known to have both activating and repressive effects of varying magnitude in different contexts. In the figure below, we show moving average plots of absolute expression intensity vs. predicted binding occupancy for C/EBPα and E2F4 in 3T3-L1 cells, and for C/EBPα, E2F4, FOXA1, and FOXA2 in liver hepatocytes. To generate these plots, for each region we identified the nearest transcript with detectable expression from Affymetrix Mouse 2.0 arrays. We then removed from the analysis any binding event located further than 10kb from a transcription start site to avoid biasing the analysis by including many potentially nonfunctional events. We sorted the remaining regions by both total tag counts and mean posterior binding occupancy, and then plotted 500-point moving averages of absolute expression.

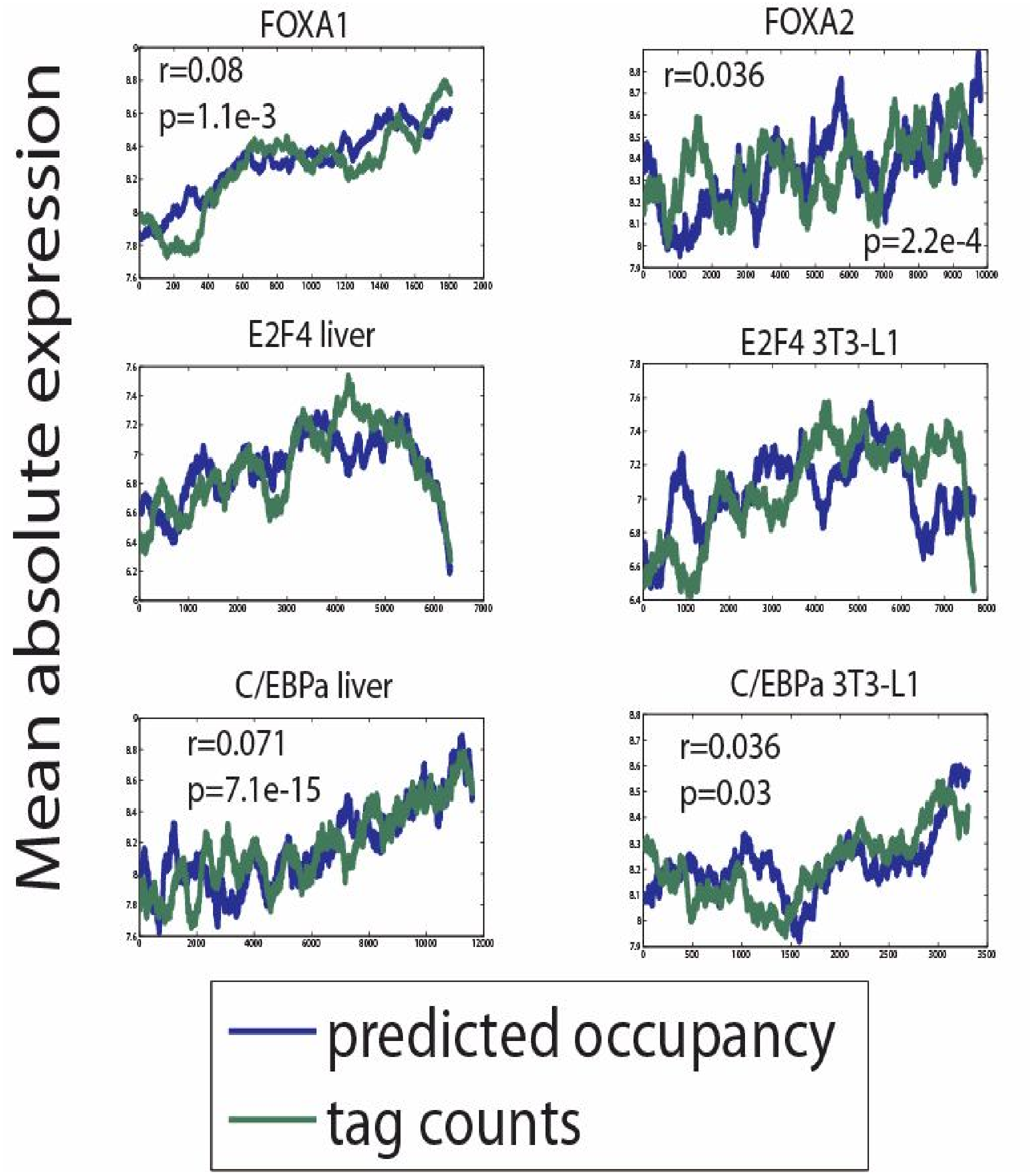


Figure 4.5: Expression vs. predicted occupancy and ChIP-seq tag counts. 500-point moving average plots of mean absolute expression are show for genomic regions sorted by predicted binding occupancy and by raw ChIP-seq tag count number. Also shown are rank correlations between expression and predicted occupancy for factors with significant correlation between expression and binding.

Figure 4.5 shows there is a weak, but statistically significant correlation between expression and predicted occupancy for several factors. We interpret this as further confirmation that the ChIP-seq count data contains biologically meaningful information about binding occupancy that our method successfully leverages. Several additional points are worth noting here: First, C/EBPα, FOXA1, and FOXA2 are all known to have important transcriptional activation function in liver (as well as 3T3-L1 cells in the case of C/EBPα) [134-136] and all show a positive correlation between expression and binding. Second, consistent with E2F4's known role as a transcriptional repressor [137], its binding seems to be negatively correlated with expression, although this relationship only becomes apparent at high levels of occupancy. Third, and unfortunately, we find no strong evidence that posterior binding estimates are better correlated with expression outcome than is the raw count data. The utility of employing motif information to assess the confidence of binding events from a ChIP-seq experiment remains an open question which will likely only be resolved through careful experimental validation of binding predictions made with and without the use of motif data.

4.5 Joint analysis of ChIP-seq data from two conditions

A frequently encountered problem when analyzing ChIP data for a protein in different tissues or growth conditions is determining which portion of the observed differences in binding arise from biological sources (e.g. changes in the quantity of protein in the nucleus, the binding specificity of the protein, chromatin accessibility, or the activity of binding partners) and which portion arises from experimental sources (e.g. differences in IP efficiency or other experimental noise). In theory, there should be information in the relative occupancy of binding sites with different affinities that allows us to address this question. Specifically, the protein concentration determines the relative occupancy of sites with different affinities. Consider a strong and a weak binding site exposed to the same concentration of a transcription factor. At low concentrations, only the high affinity site will be highly occupied, whereas at high concentrations even the weak site will have high occupancy:

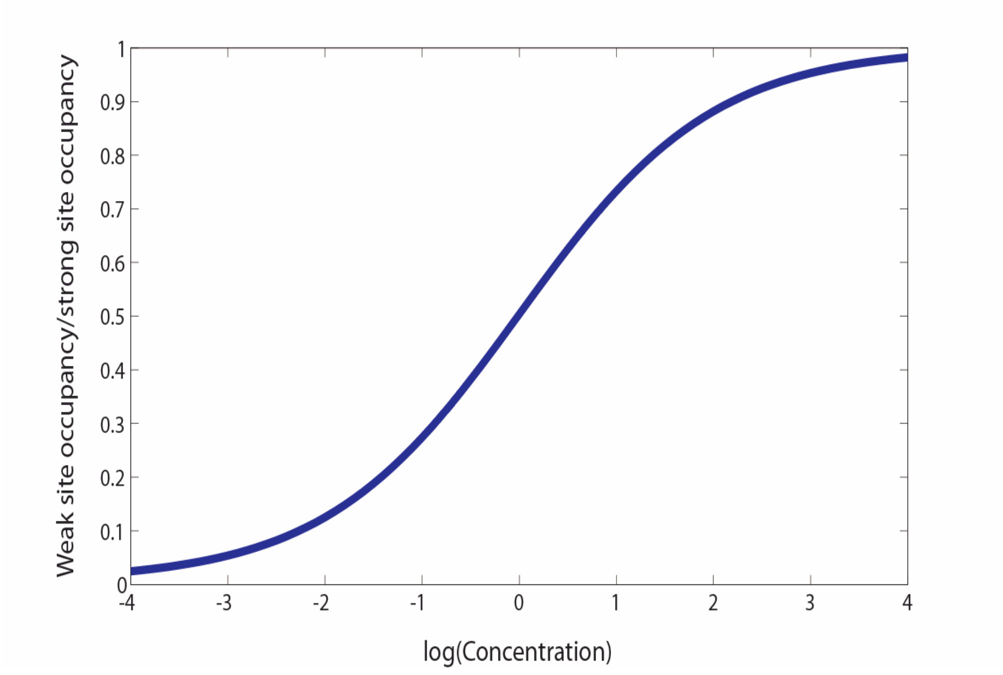


Figure 4.6: Weak site to strong site occupancy ratio as a function of concentration. Here we plot the fractional occupancy ratio of a weak binding site with a 0kcal/mol association free energy and a strong site with a -4kcal/mol binding free energy at various concentration levels.

Joint analysis of binding data for a factor in two conditions should in principle allow us to compare the concentration of the factor in these conditions. Assuming that the binding specificity of the protein does not change, the relative occupancy of low and high affinity sites within each condition is determined by the concentration. In this setting, we share statistical strength across conditions to estimate the motif binding energies, while the relative occupancy of weak and strong affinity sites within a condition is used to predict concentrations. Of course, it is also possible that any occupancy changes between conditions may be rooted in changes in site accessibility (via chromatin structure changes), specific protein-protein interactions, competition with other factors, or changes in binding affinity rooted, perhaps, in post-translational modifications of the protein. Such changes might be poorly modeled allowing only the concentration parameter to change across conditions. Therefore, for each experiment pair we also test an alternative hypothesis that both the binding specificity and the concentration may change between conditions. The results of this analysis for C/EBPα, E2F4, FOXA1, and FOXA2 profiled in two separate tissues/growth conditions are summarized below:

Table 4.3: Concentration and specificity comparisons across conditions

Protein	**Condition 1**	**Condition 2**	**Predicted concentration ratio (condition 1 vs. 2)**	**Fold expression change (condition 1 vs. 2)**	**predicted specificity difference**
E2F4	liver	3T3-L1	1.0	0.95	No
C/EBPα	liver	3T3-L1	2.6	3.0	Yes
FOXA1	normal diet liver	high fat diet liver	1.3	1.09	No
FOXA2	normal diet liver	high fat diet liver	0.6	1.25	No

For three of four factors our results show that the binding data is more consistent with a model that does not allow binding specificity to change. It is only for C/EBPα that the cross-validated likelihood score was improved by allowing the specificity to change. This is consistent with the results of section 4.4.3 which suggested that C/EBPα binding in liver and 3T3-L1 was best explained by different bZIP-like motifs. The predicted concentrations in each condition are largely consistent with expectations based on expression data. E2F4 and FOXA1 are predicted to have very little difference in their nuclear concentrations in each condition, and show no significant difference in their expression levels in these conditions. In contrast, C/EBPα is upregulated approximately 3–fold in liver relative to 3T3-L1 cells and our method predicts a 2.6 fold difference in their concentrations. It must be stated, however, that this predicted concentration difference should be interpreted with extra caution since the binding specificity was also predicted to change in liver and 3T3-L1. The concentration parameters are therefore not directly comparable. FOXA2 is a very interesting case. It shows somewhat higher expression in normal diet, however our method predicts an almost 2-fold decrease in concentration in normal diet relative to high fat diet. This observation may be rooted in the biology of FOXA2 regulation. During starvation, FOXA2 is localized to its targets in liver; however during feeding it is phosphorylated and sequestered outside of the nucleus in response to insulin signaling [138]. In high fat diet induced diabetes, insulin signaling by the protein IRS2 is compromised. This leads to activation of forkhead proteins like FOXO1, who are no longer phosphorylated. However FOXA2, since it is phosphorylated by both IRS1 and IRS2 (whose function is not affected in insulin resistance), was thought to remain largely inactive [139]. Our results here offer up the interesting hypothesis that, although FOXA2 may remain partially inactive in high fat diet induced insulin resistance due to the action of IRS2, there may in fact be a detectable increase in its nuclear localization, and hence activity, that has been overlooked in previous studies.

4.6 Modeling competition for binding sites

One of the advantages of the biophysical framework presented above is that extending the model to incorporate other regulatory interactions can often be accomplished in a very natural and straightforward manner. An example is competition between different regulatory proteins for the same binding site. When two proteins, A and D, bind similar sequence motifs to form complexes C and E respectively, the probability that a particular protein binds a site can no longer be expressed using (4.5). Instead, one must consider the fact that the site can now undergo two separate reactions with equilibrium constants:

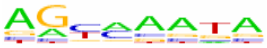
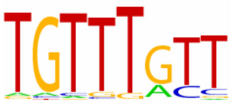
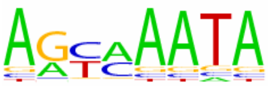
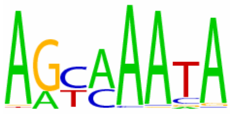
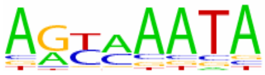
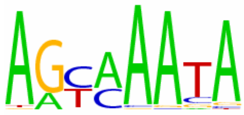
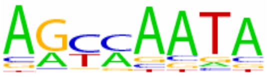
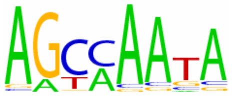
$$K_{a,1} = \frac{[C]}{[A][B]}, \quad K_{a,2} = \frac{[E]}{[D][B]} \tag{4.26}$$

Now the probability of binding by each individual protein is given by:

$$\begin{aligned} p_C &= \frac{C}{C+E+B} & p_E &= \frac{E}{C+E+B} \\ &= \frac{AK_{a,1}}{1+DK_{a,2}+AK_{a,1}} & &= \frac{DK_{a,2}}{1+DK_{a,2}+AK_{a,1}} \\ &= \frac{\mathrm{e}^{\Delta G_1 + \log A}}{1+\mathrm{e}^{\Delta G_1 + \log A}+\mathrm{e}^{\Delta G_2 + \log D}} & &= \frac{\mathrm{e}^{\Delta G_2 + \log D}}{1+\mathrm{e}^{\Delta G_1 + \log A}+\mathrm{e}^{\Delta G_2 + \log D}} \end{aligned} \tag{4.27}$$

Again, binding probability assumes a convenient logistic form. Given a representative set of binding sites and their *in vivo* occupancies, in the 1-protein case the parameter estimation problem was analogous to logistic regression. Here motif and concentration parameter estimation involves multinomial logistic regression. This extends in a straightforward manner to any number of competing proteins; binding by each protein is treated as an additional class in the regression. The sampling procedure is also extended in a straightforward manner by considering binding reactions for each protein. We tested this framework by jointly analyzing FOXA1 and FOXA2 ChIP-seq data in normal diet and high fat diet liver. The results of the analysis produce motifs that are largely consistent with the motifs reported when the datasets were analyzed individually. We compared the likelihood of held out ChIP-seq count data according to the competition model, with likelihoods calculated when the same data were analyzed individually without considering competition. Curiously, in high fat diet a model that did not consider competition performed better, whereas in normal diet the opposite was true: accounting for competition improved the test likelihood score.

Table 4.4: Binding specificities of FOXA1 and FOXA2 learned from a competitive binding model

Condition	Protein	Specificity with Competition	Specificity from Independent Analysis
Normal diet	FOXA1		
	FOXA2		
High fat diet	FOXA1		
	FOXA2		

This result seems plausible since the binding overlap between these factors is significantly higher in normal diet than high fat diet: 81.7% of FOXA1 sites overlap a FOXA2 site in normal diet, whereas only 43.1% of FOXA1 sites overlap a FOXA2 site in high fat diet.

4.7 Conclusions and Future Work

In this chapter we have presented a biophysically motivated framework for modeling protein-DNA interaction. This framework is different from classical consensus sequence or generative sequence models in that it attempts to realistically model the physical interaction between protein and DNA. This turns out to be directly analogous to a conditional motif model, and in fact the expression we derive for binding probability takes on a convenient logistic form. We have demonstrated how this framework can be adapted to the hypothesis-testing motif analysis method presented in Chapter 3 without any loss in performance. We then presented a probabilistic model relating binding occupancy to the tag count measurements made in a ChIP-seq experiment that allows us to perform a hypothesis testing analysis without imposing strict binding cutoffs.

A clear attraction of this approach is how it naturally extends to physically realistic and interesting regulatory scenarios. We have demonstrated its application for estimating the extent of protein concentration or binding specificity changes across different tissues or conditions. We have also demonstrated that the framework can easily be extended to model competition between regulators for common binding sites. There are several other interesting and straightforward extensions that have occurred to us that we outline below:

Mixture models of binding specificity

It is possible that a single affinity matrix does a poor job of representing the binding specificity of a transcriptional regulator. Indeed, there are examples of proteins with binding specificities that would be more naturally represented as a multi-component mixture of matrices such as SREBP, which may bind the consensus sequence TCACCCyA as well as the E-box CACGTG [140]. There are also proteins that may be recruited to DNA both by binding their consensus sequence and by interacting with another DNA-bound factor. Such mixture models are easily handled in this framework. Consider a two-component mixture model: here our approach is to model binding as occurring through two different reactions:

$$K_{a,1} = \frac{[C]}{[A][B]}, \quad K_{a,2} = \frac{[D]}{[A][B]}$$

This leads to an expression for binding probability that is very similar to the binding competition scenario presented above. After sampling, estimation of motif parameters may proceed via multinomial logistic regression as before, with the additional constraint that the protein concentration parameter is shared across classes:

$$
\begin{aligned}
p_C &= \frac{C}{C+B+D} & p_D &= \frac{D}{C+B+D} \\
&= \frac{AK_{a,1}}{1+AK_{a,1}+AK_{a,2}} & &= \frac{AK_{a,2}}{1+AK_{a,1}+AK_{a,2}} \\
&= \frac{\mathrm{e}^{\Delta G_1+\log A}}{1+\mathrm{e}^{\Delta G_1+\log A}+\mathrm{e}^{\Delta G_2+\log A}} & &= \frac{\mathrm{e}^{\Delta G_2+\log A}}{1+\mathrm{e}^{\Delta G_1+\log A}+\mathrm{e}^{\Delta G_2+\log A}}
\end{aligned}
$$

Coregulator recruitment

Coregulators are recruited to their targets through interactions with DNA-bound proteins. Many of these coregulators can interact with multiple proteins, and competition for limiting concentrations of coregulator among promoters is thought to be an important regulatory mechanism in certain contexts [141, 142]. Our biophysical framework can be adapted to jointly analyze the binding of a DNA-binding transcription factor or factors and a coregulator that is recruited by that factor. A DNA-binding protein is assumed to exist in two forms in the nucleus: free protein and protein complexed with a coregulator. Free protein A, can bind a site, B, to form a complex C. In addition, a protein-coregulator complex, D, can also bind the site to from a separate complex E. The complex, C, can be bound by free coregulator to form E. The entire set of reactions that need to be considered is shown below:

$$
\begin{aligned}
A+B &\xrightarrow{k_1} C, & C &\xrightarrow{k_{-1}} A+B \\
D+B &\xrightarrow{k_1} E, & E &\xrightarrow{k_{-1}} D+B \\
C+F &\xrightarrow{k_2} E, & E &\xrightarrow{k_{-2}} C+F \\
A+F &\xrightarrow{k_2} D, & D &\xrightarrow{k_{-2}} A+F
\end{aligned}
$$

This leads to the following expression for the probability a binding site will be found bound by the DNA-binding protein only (species C), and the probability it will be bound by the complex E:

$$
\begin{aligned}
p_C &= \frac{C}{C+E+B} & p_E &= \frac{E}{C+E+B} \\
&= \frac{AK_1}{1+AK_1+AK_1FK_2} & &= \frac{AK_1FK_2}{1+AK_1+AK_1FK_2} \\
&= \frac{\mathrm{e}^{\Delta G_1+\log A}}{1+\mathrm{e}^{\Delta G_1+\log A}+\mathrm{e}^{\Delta G_1+\log A+\Delta G_2+\log F}} & &= \frac{\mathrm{e}^{\Delta G_1+\log A+\Delta G_2+\log F}}{1+\mathrm{e}^{\Delta G_1+\log A}+\mathrm{e}^{\Delta G_1+\log A+\Delta G_2+\log F}}
\end{aligned}
$$

Physically realistic models have a very important advantage over many other approaches: interpretability. Interpretability is particularly crucial in the highly collaborative settings where computational biology techniques are most valued since results must often be communicated to experts from a range of disciplines. This chapter has hopefully demonstrated that developing such models does not necessarily require sacrificing computational convenience or principled statistical methodology.

Chapter 5: A model of transcriptional enhancer function

In this chapter I present the results of a study examining transcriptional enhancer structure and function in three mouse tissues. This chapter applies many of the ideas presented in the previous chapters, while at the same time introducing a novel predictive model of gene expression. Of all the projects I have been involved in during my thesis work, I feel that this project best exemplifies the power that joint computational and experimental studies can bring to bear on furthering our understanding of biology. The success of this study depended on very significant experimental efforts on the part of William Gordon, Alice Lo, Shmulik Motola, and Tali Mazor who performed either ChIP or gene expression microarray experiments.

5.1 Introduction

Control of gene expression programs across diverse tissues and developmental stages is achieved through complex networks of proteins interacting with specific regulatory sites in the genome. Chromatin immunoprecipitation (ChIP) coupled with high throughput microarray (ChIP-chip) or sequencing (ChIP-seq) technology has allowed the structure of some of these networks to be mapped on a genome-wide scale [143-146]. These draft networks must be interpreted with caution since there is evidence that only a subset of regulator binding sites identified in a ChIP experiment are functional, while many binding events play no direct role at all in determining transcription levels [147]. Even if all functional regulatory regions in a tissue could be identified, there is currently no simple and accurate quantitative framework describing how the resulting regulatory architecture relates to transcription levels of regulated genes, or indeed even how to associate binding events with the genes they may regulate in a principled manner.

Finally, the effect of a regulatory site on expression levels will depend on complex combinatorial interactions among the multiple transcriptional activators and repressors they bind, and it is currently unknown how large a role such interactions play in determining tissue-specific expression levels.

In this chapter we present a model of transcriptional regulation that successfully addresses these key challenges. We identified enhancers and proximal promoter regulatory sites by performing high throughput ChIP experiments on CREB-binding protein (CBP), the deacetylase SIRT1 and on multiple DNA-binding transcription factors in three different tissues. Sequence analysis of the immunoprecipitated DNA reveals important tissue-specific DNA-binding proteins that recruit CBP to distinct regions in different cell types. We analyze binding and expression data to reveal the quantitative effect that each regulatory complex has on a gene's expression. Remarkably, we find that the function of a regulatory site is, to a large extent, dependent on its proximity to the transcription start site of a gene. Our approach also reveals the relative contributions of each protein to combinatorial control of transcription.

5.2 Experimental identification of enhancer regions

A recently described strategy used ChIP to profile the genomic localization of the p300 enhancer-binding protein [148, 149]. We employed a similar strategy by performing ChIP on mouse liver and cerebellum samples using an antibody specific to CBP, a transcriptional coregulator closely related to p300. Immunoprecipitated DNA from liver was sequenced, the 35bp reads were aligned to the reference mouse genome, and regions with significant levels of CBP binding relative to a set of control reads were identified. We also performed ChIP-chip experiments in liver and cerebellum using mouse promoter microarrays. The ChIP-seq analysis identified 18,264 CBP-bound regions in liver, while the ChIP-chip experiments revealed 2,608 and 2,452 CBP-bound regions near proximal promoters in liver and cerebellum respectively. Most CBP binding occurs outside of the proximal promoter: 79% of sites in liver ChIP-seq, 70% in liver ChIP-chip, and 51% in cerebellum ChIP-chip occur outside a 500bp window centered on a transcript's TSS. Several of these more distal sites, hereafter referred to as enhancers for simplicity, directly overlap previously characterized transcriptional enhancers [150-156].

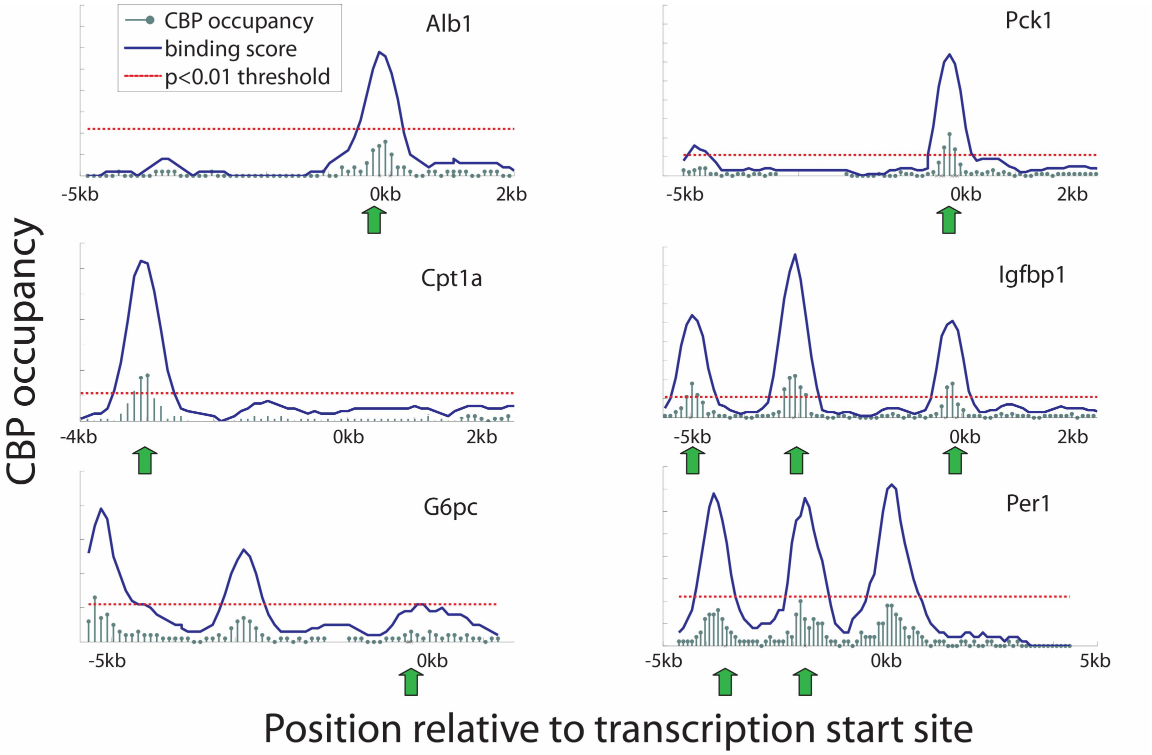


Figure 5.1: CBP binding overlaps previously characterized enhancers. CBP binding occupancy predicted by Redwing is shown in green, and smoothed CBP occupancy scores (obtained by convolving the Redwing predictions with a 400bp sliding window) are shown in blue. The red dashed line in each figure corresponds to a high confidence binding threshold (FDR <= 0.01). This threshold was estimated by running Redwing on randomly permuted ratio data. Positions of previously characterized enhancer regions are denoted by the green arrows. These include a proximal region between -170bp and the TATA box in the Alb1 promoter (Maire et al. 1989), a region at -3kb in the Cpt1 promoter (Louet et al. 2002), regulatory regions A and B between -231bp and -158bp in the glucose-6-phosphatase promoter (Onuma et al. 2009), a CRE and several other regulatory sites in the region -300bp to -100bp in the Pck1 promoter (Hanson and Reshef 1997), three distinct DNaseI hypersensitive regions at approximately -5kb, -3kb, and -300bp in the Igfbp1 promoter (Crissey et al. 1999), and characterized cAMP-response and glucocorticoid-response elements at -1728bp (Travnickova-Bendova et al. 2002) and -3566bp (Yamamoto et al. 2005) respectively.

These regions typically span 400-800bp in length, and are generally located within 100kb of annotated transcription start sites (TSS).

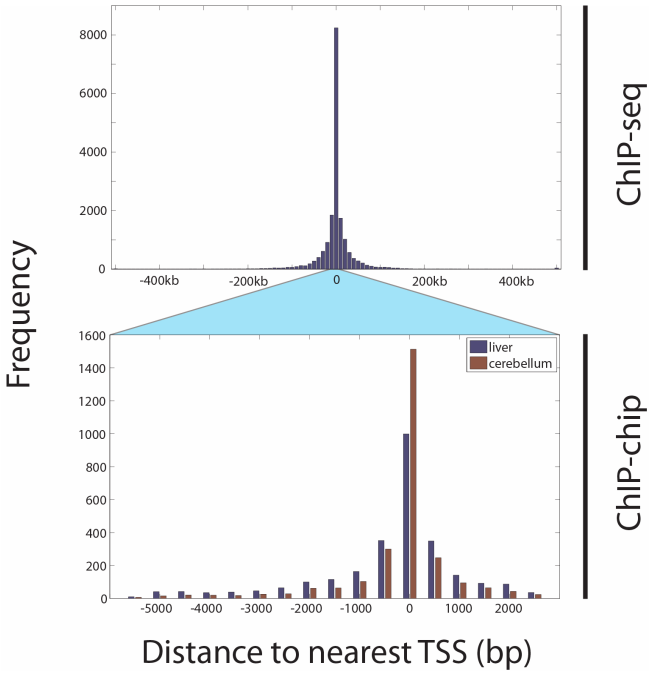


Figure 5.2: Distribution of position relative to the nearest transcription start site for CBP-bound regions from ChIP-seq in liver (upper plot) and ChIP-chip in liver and cerebellum (lower plot).

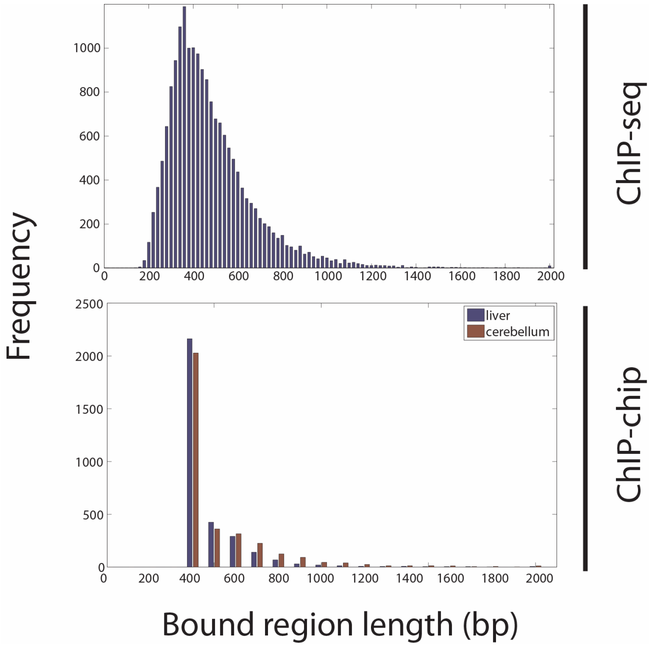


Figure 5.3: Distribution of lengths for CBP-bound regions identified in ChIP-seq experiments in liver (upper plot) and ChIP-chip experiments in liver and cerebellum (lower plot).

5.3 Sequence analysis of enhancer regions

CBP and p300 are recruited to their genomic targets via protein-protein interactions with DNA-bound transcription factors [157, 158]. We wished to identify the transcription factors that these coregulators form complexes with, to identify any higher order motif combinations or positional constraints associated with enhancer regions, and to determine whether these predictive features were different in different tissues. To that end, we analyzed the DNA sequence CBP-bound sites in liver and cerebellum, and at regions previously identified as bound by p300 in embryonic mouse limb and forebrain [148].

5.3.1 Identification of overrepresented motifs

A compendium of 530 motif position weight matrices (PWMs) was assembled by combining motifs from various databases [116, 118, 159] and removing those originating from non-mammalian sources. PWMs were clustered to eliminate redundancy using Affinity Propagation (AP) [160] yielding a set of 233 distinct motifs. The motif distance used for clustering was calculated by assessing the mean KL-divergence between columns of the motif frequency matrices for all possible alignments of the motif pair (both forward and reverse complement, subject to a minimum overlap of 6bp) and taking the minimum value. Motifs were evaluated for statistical enrichment using the THEME algorithm. Motifs with a cross-validation error less than 0.50 and with an FDR-corrected p-value<=1e-3 were then clustered by AP to produce the set of motifs associated with CBP recruitment in each tissue. These motifs were then used to generate motif-based feature sets used to predict coregulator recruitment.

5.3.2 Predictive models of coregulator binding

We evaluated four motif-based feature sets for predicting CBP/p300 binding in a series of logistic regression classifiers. Two models, called *Enhanceosome A* and *B*, assume that some motif combinations may be more important than others. The remaining two models, called *Billboard A* and *B* do not distinguish among various motif combinations.

Generation of motif features

In each experiment the N statistically significant PWMs were used to score training set sequences as previously described [161]. THEME provides a threshold, t, for each PWM that optimally distinguishes bound from unbound sequences. We convert a PWM score, s, to a normalized score, m, using the transformation:

$$m = {}^{1}\!\big/_{\left(1+\exp\left(t-s\right)\right)} \qquad (5.1)$$

For each sequence, with length L, we calculate a 1 x N vector, X, of maximum scores for each PWM. We also define an N x L indicator matrix, Y, encoding the location of any motif matches in the sequence. Matches are defined as sites where the normalized score is >= 0.5. We slide a 100bp rectangular window along the matrix Y, and identify the location with the maximum number of distinct matches to the N PWMs. This maximum, z, can range in value from 0 to N. Finally, for each sequence we define a vector, V, of indicator variables, $v_{j,k}$, which encode whether there is a 100bp window in Y with a match to both PWMs j and k.

Description of predictive models

The coregulator binding model we employ is identical to the model described for protein binding in chapter 5. We assume that coregulator binding to an enhancer can be modeled as a bimolecular reaction at equilibrium. The free energy of this reaction is a simple sum of contributions, β_i, from motif features, x_i, present at the enhancer leading to the now familiar expression for binding probability:

$$p = \frac{1}{1 + \exp\left(-\beta_0 - \sum_i \beta_i x_i\right)} \tag{5.2}$$

The four models tested differ in the sets of features that are used as statistical predictors in equation 5.2. The *Billboard A* feature set consists of only the vector of maximum motif scores for each sequence, *X*. The *Billboard B* feature set augments the feature vector *X* with the clustering score *z*. The *Enhanceosome A* feature set augments *X* with pair-wise products of motif scores $x_j x_k$ (for $j,k = 1\ldots N$ and $j \neq k$) to capture the effect of specific combinatorial interactions between TFs. Finally, the *Enhanceosome B* feature set concatenates the feature vectors *X* and *V* to capture combinatorial interactions among closely spaced motifs.

5.3.3 Model performance

Classifiers were trained on the same sequences used to screen motifs with THEME. Logistic regression parameters were estimated as previously described [128]. Features were sequentially removed from the model by backward elimination, using 3-fold cross-validated classification error for evaluation and the one standard error rule as a stopping criterion. The surviving features were used to train a classifier using the entire training set, whose performance was evaluated on the held-out data. This procedure was repeated ten times for each model/tissue pair. After feature selection and performance evaluation on held out test data we found that in four out of five datasets, a simple feature set that used individual motif match scores, ignoring specific motif combinations, performed as well or better than more complex models.

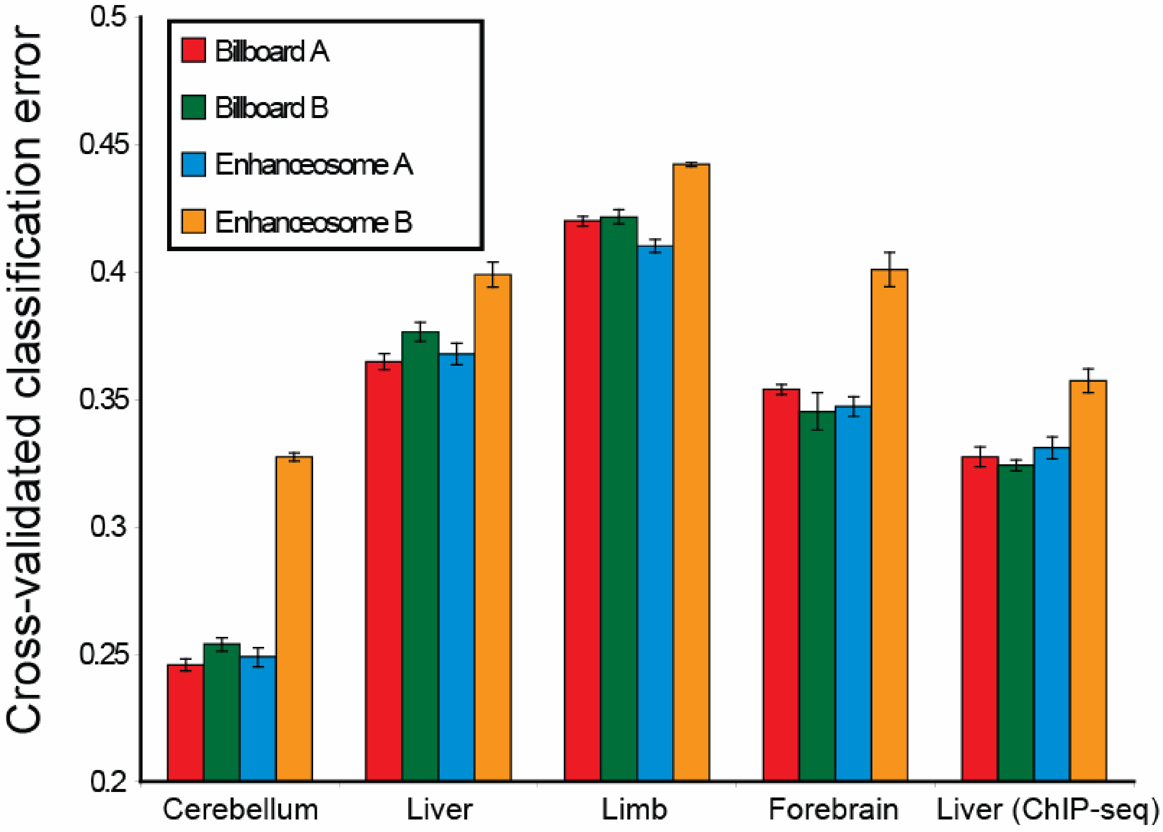


Figure 5.4: Classification performance of CBP/p300 recruitment models. Mean 3-fold cross-validated prediction error is shown for the four feature sets used to distinguish CBP/p300 bound regions identified in ChIP experiments from unbound regions. Error bars correspond to the standard error of the mean calculated from ten separate trials for each model/tissue pair.

Interestingly, the most important motifs for predicting CBP/p300 recruitment are different in each tissue. Tables 5.1-5.5 show the motifs that survived feature selection in at least 5 of 10 trials in each data set. In tables 5.1-5.5, mean weight is the mean logistic weight assigned to the motif when it was included in the final binding model. The survival rate refers to the number of training runs in which the motif was included in the model.

Table 5.1 – Core motifs at liver enhancers (ChIP-chip)

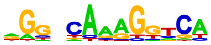
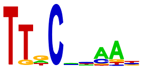
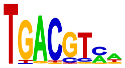
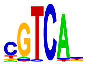
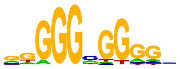
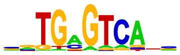
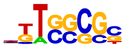
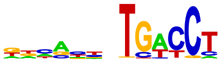
Logo	Factor	Mean weight	Survival rate
	PPAR/HNF4	1.22	100%
	C/EBPα	0.57	100%
	CREB	0.73	80%
	ATF/CREB	0.39	80%
	SP	0.50	70%
	Bach2	0.23	60%
	E2F	0.14	50%
	ER/AR	0.00	50%

Table 5.2 – Core motifs at liver enhancers (ChIP-seq)

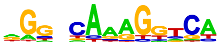
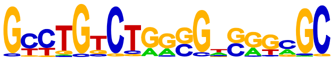
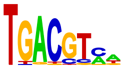
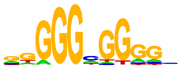
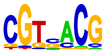
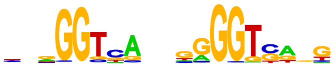
Logo	Factor	Mean weight	Survival rate
	NHR	0.75	100%
	STAT	0.95	90%
	CREB	0.18	60%
	SP	0.59	50%
	PAX	0.15	50%
	RXR/PPAR	0.14	50%

Table 5.3 – Core motifs at cerebellum enhancers

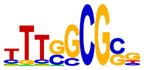
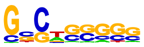
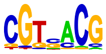
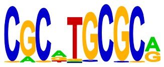
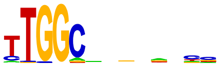
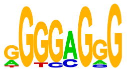
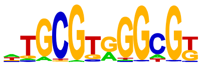
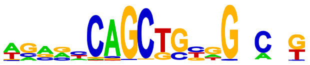
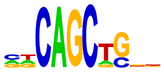
Logo	Factor	Mean weight	Survival rate
	E2F	2.05	100%
	AP-2	1.49	90%
	PAX	0.79	60%
	NRF-1	0.57	60%
	NF-I	0.48	60%
	MAZ	0.58	50%
	EGR-1	0.43	50%
	AP-4	0.10	50%
	AP-4	0.00	50%

Table 5.4 – Core motifs at embryonic forebrain enhancers

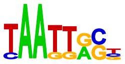
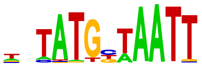
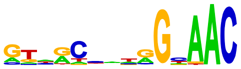
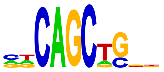
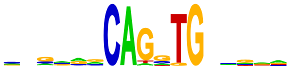
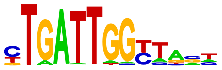
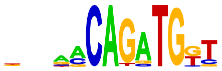
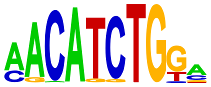
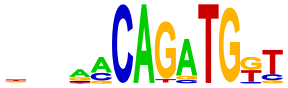
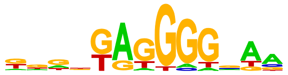
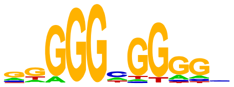
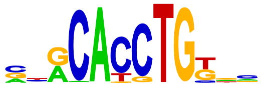
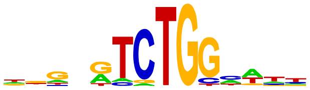
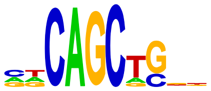
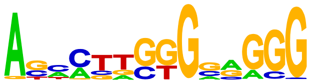
Logo	Factor	Mean weight	Survival rate
	Homeobox	1.13	100%
	Pou2f1	0.69	100%
	Rfx1	0.75	80%
	AP-4	0.33	80%
	E12/E47/MYOD	0.00	80%
	NF-Y/CBF	0.55	70%
	Tal-1/E47	0.00	60%

Table 5.5 – Core motifs at embryonic limb enhancers

Logo	Factor	Mean weight	Survival rate
	RP58	0.53	100%
	Tal-1/E47	0.17	70%
	MZF-1	0.10	60%
	SP	0.01	60%
	Areb6	0.00	60%
	Hand1	0.04	50%
	AP-4	0.00	50%
	Sp3	0.00	50%

In all tissues putative enhancers are enriched for the binding sites of a wide variety of transcription factors, many of which have previously described regulatory roles in the tissues examined. Cerebellum and liver enhancers share enrichment for the motifs of many ubiquitously expressed TFs known to interact with CBP/p300 including E2F [162], SP [163], and AP-2 [164] factors. Liver enhancers are distinguished by enrichment for a nuclear hormone receptor motif consistent with binding sites for liver-enriched PPAR [165] and HNF4α [166] transcription factors. Embryonic forebrain enhancers are uniquely enriched in an RFX and POU motif, TF families that have been implicated in brain development [167, 168], while limb enhancers show unique enrichment for an HLH motif consistent with the binding specificity of the Hand TFs, key regulators of limb bud development [169].

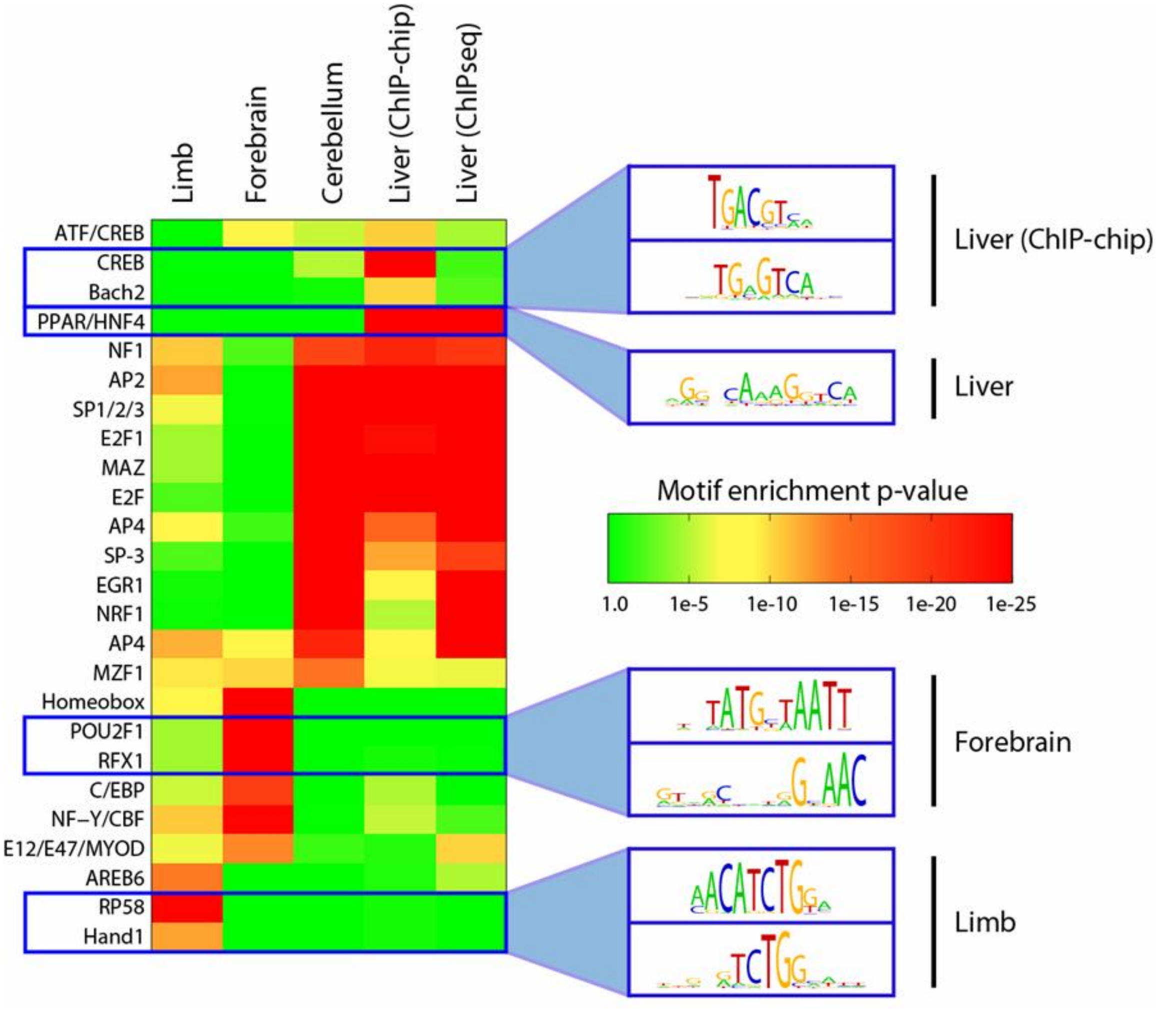


Figure 5.5: DNA sequence motif enrichment at enhancers. Tissue specific enrichment of DNA sequence motifs associated with CBP/p300 recruitment in mouse liver, cerebellum, embryonic forebrain, and embryonic limb is indicated by color in the heat map. Sequence motifs were assessed for statistical enrichment in held out validation data sets using a hypergeometric test. Motifs with unique enrichment in a single tissue are shown in boxes.

5.4 Enhancers bind clusters of regulators

Our sequence analysis of liver enhancers identified motifs corresponding to the binding specificities of several regulators that drive liver-specific gene expression, including C/EBPα, CREB, and HNF4α [166, 170, 171]. To validate the prediction that CBP binds the same regions as these transcription factors, we analyzed ChIP-chip experiments in liver using antibodies specific to C/EBPα and the phosphorylated form of CREB (pCREB), and re-analyzed previously published ChIP-chip data for HNF4α [172]. We assessed the observed binding overlap between these DNA-bound transcription factors and CBP relative to a null model of random binding in the genome. Expected overlap and z-scores were calculated based on 100 randomized trials where TF binding positions were permuted by first randomly selecting a transcription start site, sampling a position relative to this TSS from the empirical distribution of CBP binding positions, and adding a random integer between -200 and 200. All three factors associate with putative enhancers much more than predicted by chance. We performed separate genome-wide ChIP-seq experiments in liver on C/EBPα, E2F4, whose motif is also highly enriched at liver enhancers, and the known CBP-interacting transcription factor FOXA2 [173], and examined the overlap in binding location between these factors and CBP from genome-wide ChIP-seq. Again, the overlap in binding sites is significant for all factors.

Table 5.6 – Binding site overlap of DNA-bound factors with CBP

Factor	Observed overlap	Expected overlap	z-score
pCREB (ChIP-chip)	915	139	66.6
C/EBPα (ChIP-chip)	702	111	52.3
HNFa (ChIP-chip)	578	160	36.7
E2F4 (ChIP-seq)	4,831	1,194	104.8
FOXA2 (ChIP-seq)	8,162	3,917	67.5
C/EBPα (ChIP-seq)	9,509	6,436	38.1

Interestingly, clusters of overlapping transcription factor binding sites are surprisingly accurate predictors of putative enhancer location, with over 80% of regions bound by all three factors directly overlapping CBP-bound regions from ChIP-chip experiments. This observation is not restricted to regions proximal to transcription start sites. For the factors profiled with ChIP-seq we also observed that binding by two or more factors was significantly more predictive of CBP-binding than binding by any single factor. Of the regions bound by all three factors, 70.2% were also bound by CBP.

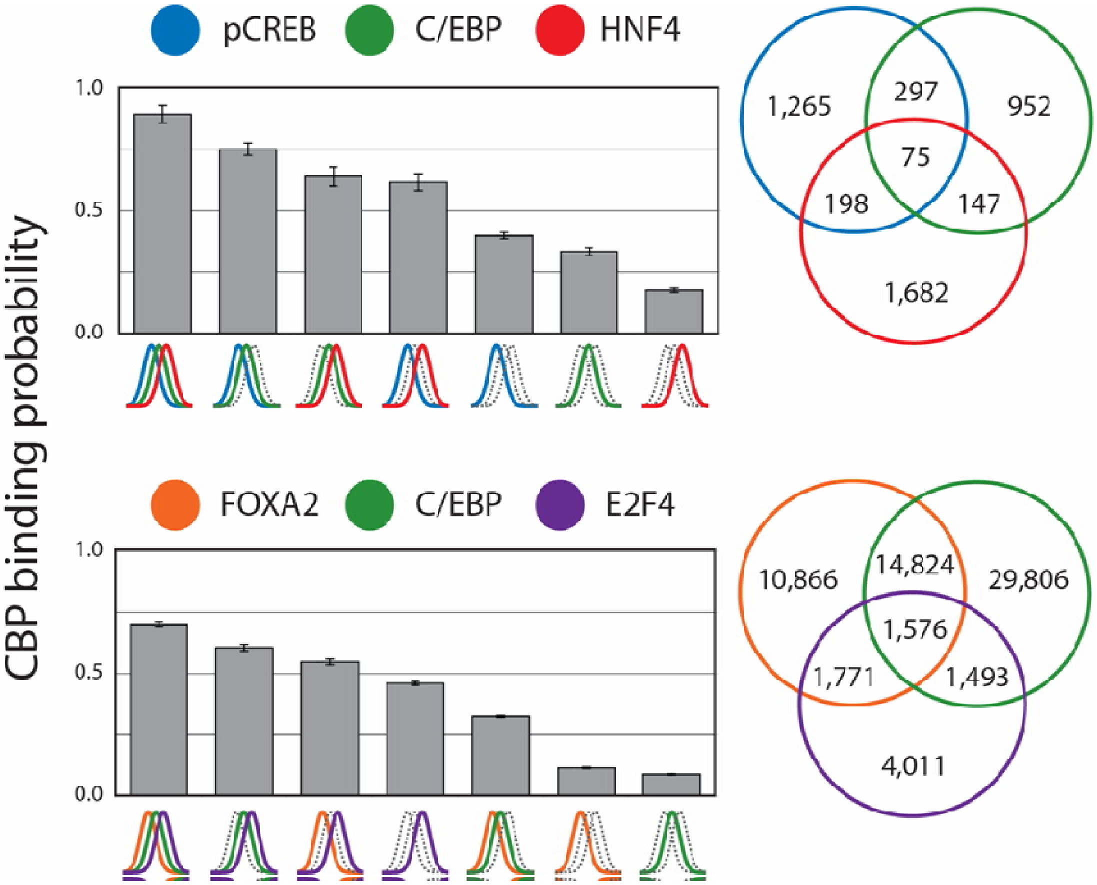


Figure 5.6: CBP binding is associated with overlapping binding sites for multiple regulators. The probability that a region is bound by CBP is denoted by bar height for different transcription factor binding configurations. Colored peaks indicate DNA-binding transcription factors with overlapping binding at a site. CBP binding probability at proximal promoter regions from ChIP-chip experiments is shown for pCREB, C/EBPα, and HNF4α in the upper graph. CBP binding probability at genome wide locations for FOXA2, C/EBPα, and E2F4 sites from ChIP-seq experiments is shown in the lower graph. Error bars indicate +/- s.e.m.

We examined the sites bound by E2F4, C/EBPα, and FOXA2 in ChIP-seq experiments for evidence of spatial constraints on binding position and orientation. For each pair of factors we identified all sites bound by both proteins and CBP. Then, using sequence motifs representing each factor's known binding specificity, we enumerated all binding site spacings and orientations in these short regions and searched for statistical overrepresentation relative to a set of 5,000 equally sized data sets with randomly permuted binding site positions. Interestingly, we found no evidence of binding site constraints occurring at a statistically significant frequency. Our results suggest that enhancer function in liver encompasses the action of a variety of different regulators and accommodates a diversity of TF binding site configurations. These data further suggest that a viable alternative strategy for identifying enhancers would be to search for regions bound by a cluster of factors in a series of ChIP experiments.

5.5 Enhancer proximity is correlated with transcript levels

To understand the relationship between regulator binding and transcription we identified sites of combinatorial control, in a fashion similar to that described in section 5.2, by performing ChIP on samples from mouse liver and 3T3-L1 cells using an antibody specific to p300, which has been used similarly in previous studies[148, 149], as well as antibodies for several proteins with transcriptional activation function in these tissues (Table 5.7) and by analyzing previously published data for PPARγ and RXR in 3T3-L1 cells[174].

The ChIP-seq analysis identified 22,191 and 7,821 putative regulatory regions in liver and 3T3-L1 cells respectively. The vast majority of these sites occur within 100kb of known genes but most are located outside of the proximal promoter (Figures 5.7 and 5.8): 92% of regulatory sites liver and 93% in 3T3-L1 cells occur outside the 500bp window centered on each transcript's transcription start site (TSS). The ChIP-chip promoter array experiments revealed 3,326 and 3,187 CBP-bound regions in liver and cerebellum; 70% of these sites in liver and 51% in cerebellum occur outside the proximal promoter.

Table 5.7 – Anti-sera used in ChIP experiments

Protein	Antibody	Source	Tissues
CBP	sc-369X	Santa Cruz	liver, cerebellum
C/EBPα	sc-9314X	Santa Cruz	liver, 3T3-L1
E2F4	sc-1082X	Santa Cruz	liver, 3T3-L1
FOXA1	ab5089	Abcam	liver
FOXA2	sc-6554	Santa Cruz	liver
p300	sc-585	Santa Cruz	liver, 3T3-L1
pCREB	sc-7978X	Santa Cruz	liver
Sirt1	sc-19857	Santa Cruz	cerebellum

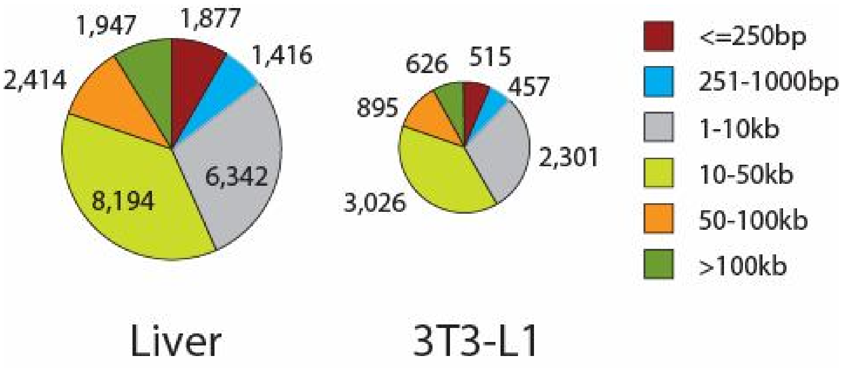


Figure 5.7: Distribution of proximities to the nearest transcription start site for regulatory regions identified in ChIP-seq experiments in liver and 3T3-L1 cells.

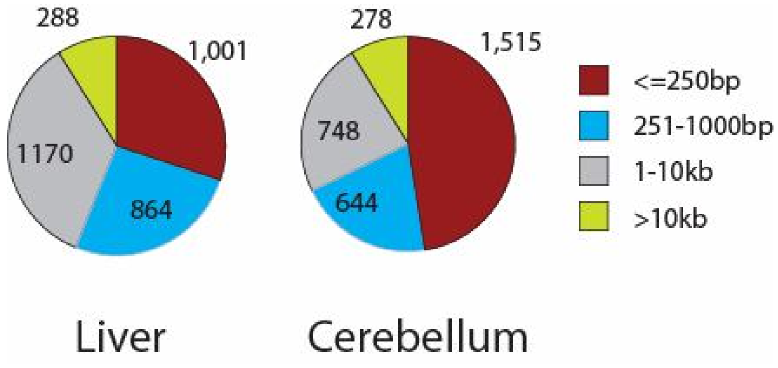


Figure 5.8: Distribution of proximities to the nearest transcription start site for regulatory regions identified in CBP ChIP-chip experiments in liver and cerebellum.

Understanding the relationship regulator binding and transcription is a complicated task. Although binding within 5 kilobases (kb) of a gene’s transcription start site (TSS) is associated with higher expression in each tissue (Figure 5.9A), it provides limited information about tissue-specific transcription levels as these genes display a wide range of expression values (Figure 5.9B).

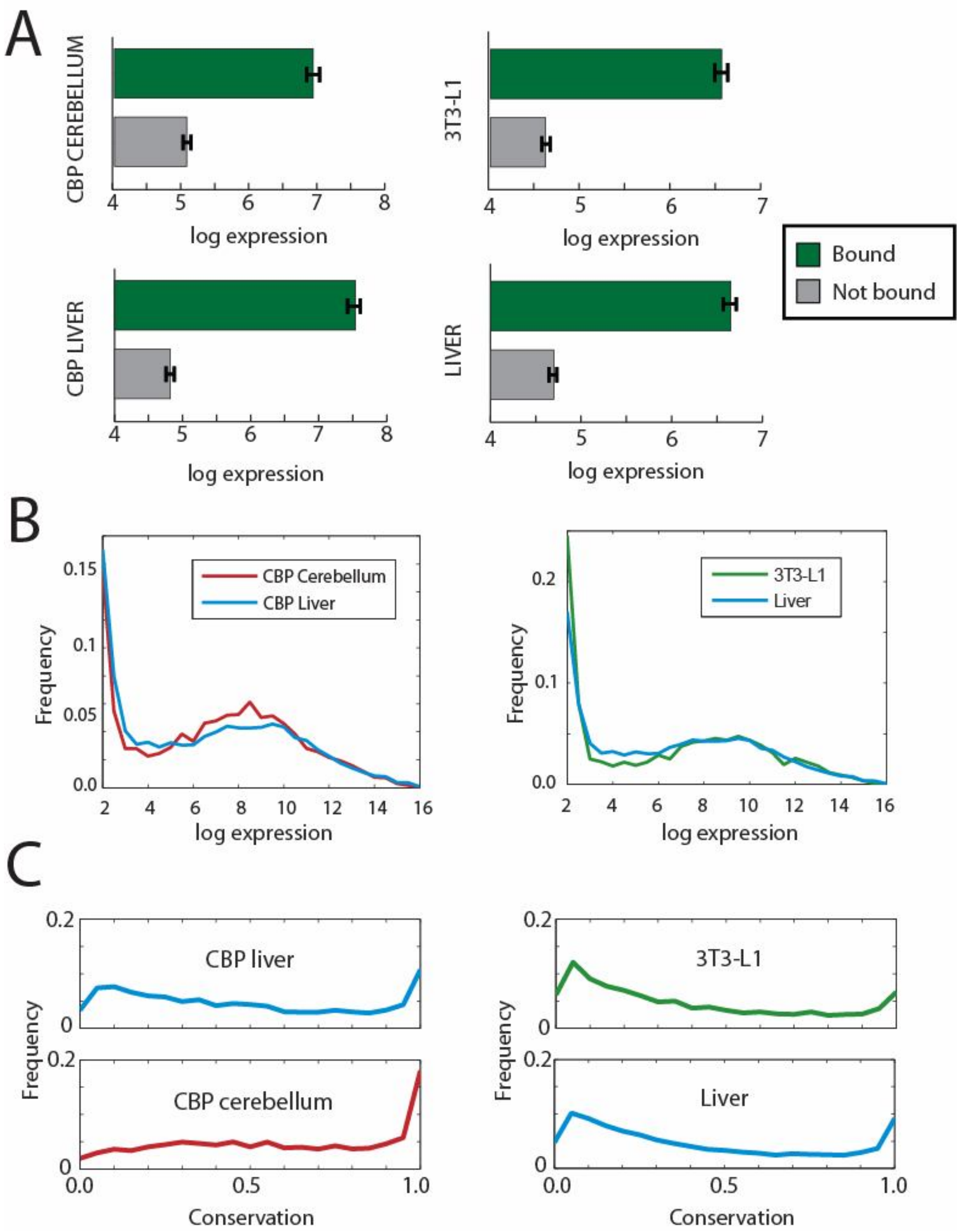


Figure 5.9: Characteristics of bound genes and bound regions. (A) Genes with a regulatory region within 5kb of their transcription start site have a higher mean expression level than genes with no binding event. Error bars indicate +/- s.e.m. (B) Bound genes display large variation in levels of absolute gene expression. (C) Putative regulatory regions show great variation in their sequence conservation levels. Conservation level was calculated as the maximum 100bp moving average of Phastcons scores from alignments of placental mammal genomes.

This variation may be explained, in part, by the action of distal regulatory sites located further than 5kb from the gene. However, as we begin to consider binding events further from the TSS the situation becomes increasingly complicated as more, potentially non-functional, binding sites become associated with each gene. It is also difficult to associate binding events with the genes they regulate. For example, approximately 41% of regulatory sites identified in liver and 45% in 3T3-L1 cells are located within 50 kb of the TSS of two or more genes.

The problem of identifying functional regulatory regions has been addressed using sequence conservation [175, 176]. We found that bound regions vary significantly in their degree of sequence conservation (Figure 5.9C) and wished to explore whether more highly conserved sites were more likely to be functional. When we examined the mean expression level of genes in each tissue as a function of the conservation level of their binding events, we found a weak or non-existent relationship (Figure 5.10A). Surprisingly, transcription levels *are* highly correlated with the proximity between a gene's TSS and the closest bound region (Figure 5.10A). This statistical relationship persists over tens of kilobases and is highly statistically significant, even at a distance resolution of hundreds of nucleotides within the proximal promoter (Figure 5.11). Although we are aware of an *in vitro* study where a linear falloff in transcription rate was observed as an enhancer's location was moved further from the TSS in a series of reporter constructs [177], to our knowledge this intriguing effect has not been previously reported as a general feature of transcriptional regulation in an *in vivo* system.

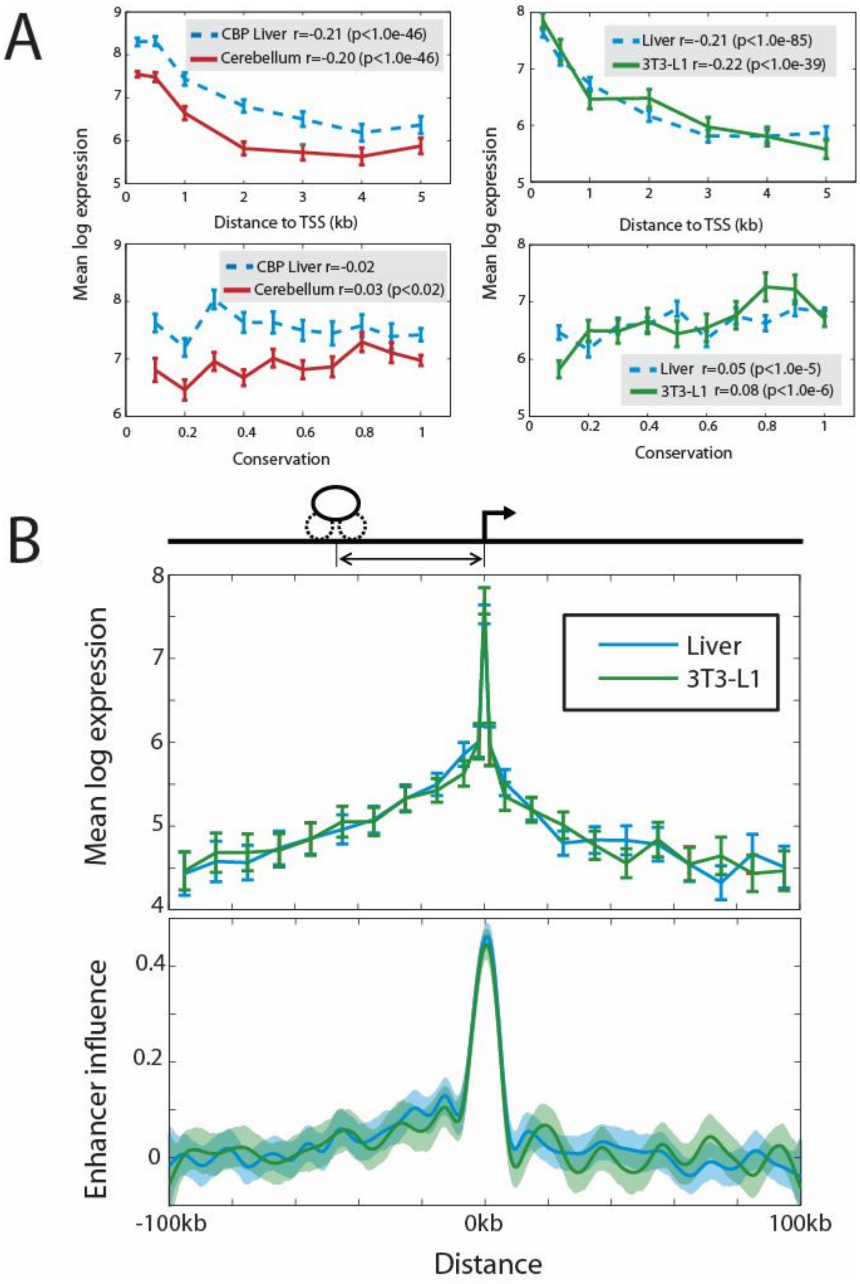


Figure 5.10: Binding site position, but not sequence conservation, is strongly associated with gene expression level. (A) The mean log expression of bound genes is shown in each tissue as a function of the distance between the transcription start site and the nearest regulatory region identified by ChIP, and the maximum conservation score of any regulatory region within 5kb of that gene's TSS. Error bars indicate +/- s.e.m. Also shown is the Spearman correlation, and associated p-value from a right-tailed t-test, between log expression and the distance and conservation measures. (B) In the upper plot the mean log expression of genes in liver and 3T3-L1 cells is shown as a function of the location of the nearest binding site over a 200kb window. Error bars indicate +/- s.e.m. In the lower plot we show the influence function, which measures a binding event's predicted effect on expression as a function of position, obtained by fitting our predictive model to 1,000 bootstrapped samples of ChIP and expression data in each tissue. Shaded regions show the empirical 99% confidence intervals obtained from the bootstrap iterations.

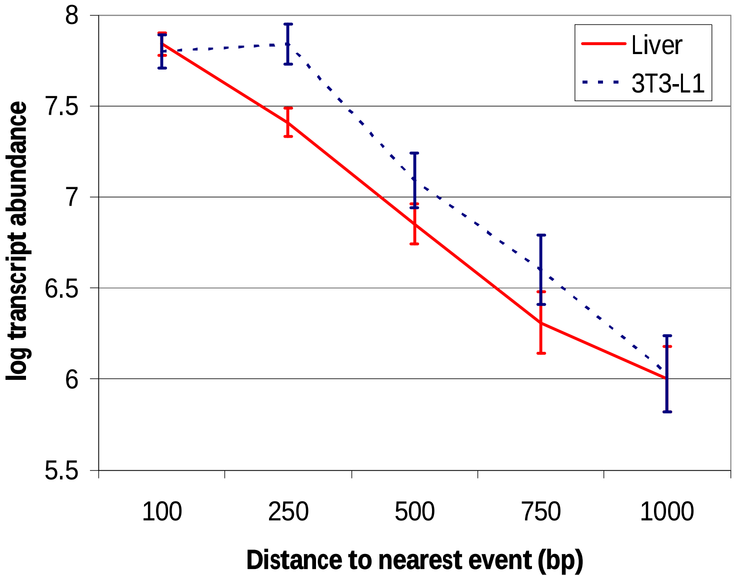


Figure 5.11: Expression level is related to regulatory site proximity in the proximal promoter. Mean log expression intensity of transcripts from Affymetrix microarrays is plotted vs. the distance between their TSS and the nearest putative regulatory site in liver and 3T3-L1 cells. Only genes with binding events in their proximal promoter are considered.

To further understand the relationship between expression and regulator binding location we developed a simple quantitative model that predicts transcription level as a function of transcription factor binding position. We assume that the mean expression level of a gene is determined by contributions from all individual regulatory sites in the vicinity of that gene, and that each regulatory site may regulate the expression of multiple genes. The functional relevance of a region depends on its position relative to the TSS; this relationship takes the form of an influence function that is fit to the data during model training. This approach allows proximal sites to be treated differently than distal sites, or upstream and downstream sites to be treated differently. The details of this model are presented in section 5.5.1

5.5.1 Predictive model of expression from enhancer location

Our goal is to predict log absolute expression level, as measured by a microarray experiment, using predicted enhancer locations. The rate of expression of a transcript, k_1, is assumed to be a function of its basal expression rate, k_0, and the action of nearby enhancers:

$$\begin{aligned} k_1 &= e^{\lambda} k_0 \\ \lambda &= \sum_{enhancers} \alpha_i f\left(d_i\right) \end{aligned} \tag{5.3}$$

Each enhancer is assumed to contribute additively to the expression rate modifier, λ. The effect that enhancer i has on this modifier is a function of its distance to the TSS, d_i. It may also depend on other considerations, for example the particular regulators bound at the enhancer. Such effects are subsumed into the parameter α_i, which unless otherwise specified, is taken to be 1.

We assume 0th order kinetics of mRNA production with rate constant k_1, and 1st order mRNA degradation kinetics with rate constant k_2. These processes, measured across the population of cells, are assumed to be at equilibrium. The log transcript abundance is then given by:

$$k_2\left[A\right] = e^{\lambda} k_0, \quad \log\left[A\right] = \lambda + \log\left({k_0}/{k_2} \right) \tag{5.4}$$

The log intensity levels, y, from the Affymetrix arrays are noisy measurements of these transcript abundances. The mean squared error between the N observations and model predictions is given by:

$$MSE = \sum_i \left(y_i - \lambda - \log\left({k_b}/{k_2} \right) \right)^2 / N \tag{5.5}$$

We now express the enhancer influence function *f(d)* using a basis set of *P* 3rd order B-splines:

$$f\left(d\right)=\sum_{k=1}^{P}c_kB_k\left(d\right) \tag{5.6}$$

Assuming that the term incorporating a transcript's basal expression rate and degradation rate, log(k_0/k_2), can be ignored leads to the following expression for MSE:

$$MSE=\sum_i\left(y_i-\sum_k c_k\left(\sum_j B_k\left(d_{i,j}\right)\right)\right)^2/N \tag{5.7}$$

The innermost sum over values of the B-spline basis functions for each enhancer position can be pre-computed. We introduce a penalty on an approximation to the integrated square of the 2nd derivative of the fitted function to control complexity. The objective function we wish to minimize, *F*, then becomes:

$$F=\sum_i\left(y_i-\sum_k c_kb_{i,k}\right)^2+\sigma\Lambda \tag{5.8}$$

Here $b_{i,k}$ are the pre-computed B-spline value sums over enhancers for basis function *k* and transcript *i*, σ is a regularization parameter that controls complexity, and Λ is the penalty term. The parameters defining the shape of the influence function, $c_{k,}$ can now be estimated by solving the system of equations:

$$B^Ty=\left(B^TB+\sigma D^TD\right)c \tag{5.9}$$

where *D* is a matrix representation of the penalty term [178].

5.5.2 Modeling relative expression levels

To predict relative expression levels between tissues *a* and *b*, we assume that basal expression rate and degradation rate for each transcript is identical in both tissues. The log fold change in expression, *y*, is then given by:

$$\begin{gathered} k_2 \left[A\right]_a = e^{\lambda_a} k_0, \quad k_2 \left[A\right]_b = e^{\lambda_b} k_0 \\ \log \frac{\left[A\right]_a}{\left[A\right]_b} = \lambda_a - \lambda_b \end{gathered} \tag{5.10}$$

and the mean-squared error is given by:

$$\begin{aligned} MSE &= \sum_i \left(y_i - \left(\lambda_a - \lambda_b \right) \right)^2 / N \\ &= \sum_i \left(y_i - \sum_k c_k \left(\sum_j B_k \left(d_{i,j}^a \right) - \sum_n B_k \left(d_{i,n}^b \right) \right) \right)^2 / N \end{aligned} \tag{5.11}$$

Here enhancers present in tissue *a* are indexed by *j*, while those in tissue *b* are indexed by *n*. The influence function parameters are then solved as described above.

5.5.3 Modeling the effect of regulators bound at the enhancer

When data for the binding of several regulators at enhancer regions is available, we can model their individual effects on enhancer function by introducing a parameter θ_m for each regulator that modifies an enhancer's effect on transcription as follows:

$$\begin{gathered} \lambda = \sum_{enhancers} \alpha_i f \left(d_i \right) \\ \alpha_i = \prod_m \theta_m^{I_{i,m}} \end{gathered} \tag{5.12}$$

Here $I_{i,m}$ is an indicator variable taking the value of 1 if regulator *m* is present at enhancer *i*, and 0 otherwise. The objective function we wish to minimize is then given by:

$$F = \sum_i \left(y_i - \sum_k c_k \sum_j \left[\left(\prod_m \theta_m^{I_{j,m}} \right) B_k\left(d_{i,j}\right) \right] \right)^2 + \sigma\Lambda \tag{5.13}$$

Taking the derivative with respect to c_k yields:

$$\partial F \big/ \partial c_p = -2 \sum_i \left\{ \begin{array}{l} \left(y_i - \sum_k c_k \sum_j \left[\left(\prod_m \theta_m^{I_{j,m}} \right) B_k\left(d_{i,j}\right) \right] \right) \\ \times \left(\sum_j \left[\left(\prod_m \theta_m^{I_{j,m}} \right) B_p\left(d_{i,j}\right) \right] \right) \end{array} \right\} \tag{5.14}$$

Taking the derivative with respect to θ_m yields:

$$\partial F \big/ \partial \theta_r = -2 \sum_i \left\{ \begin{array}{l} \left(y_i - \sum_k c_k \sum_j \left[\left(\prod_m \theta_m^{I_{j,m}} \right) B_k\left(d_{i,j}\right) \right] \right) \\ \times \left(\sum_k c_k \sum_j \left[I_{j,r} \left(\prod_{m \neq r} \theta_m^{I_{j,m}} \right) B_k\left(d_{i,j}\right) \right] \right) \end{array} \right\} \tag{5.15}$$

We then set these derivatives to zero and solve the resulting system of equations using the nonlinear equation solver fsolve in Matlab to obtain parameter estimates for the c's and θ's.

5.6 Predicting absolute expression from enhancer location

We first used our model to predict the absolute expression levels of genes in liver and 3T3-L1 cells from the location of p300 and clustered transcription factor binding sites. We considered all binding events located within 100kb of each gene's TSS. The correlation between predicted and observed transcript abundance in held-out test data is highly statistically significant (Table 5.8). Notably, the predicted relationship between position and expression influence is nearly identical in both tissues (Figure 5.10B). There is an approximately linear fall-off in influence as enhancer position moves further away from the TSS. Sites located within approximately 10kb of the TSS are statistically associated with the highest transcription levels, and regulatory regions located upstream of the TSS are predicted to have a somewhat greater effect on transcription than downstream events. Although proximal sites have the greatest influence, binding sites located up to 50kb away from the TSS are predicted to have a significant effect on transcription, consistent with previous observations that enhancers may act at very long distances to affect expression [179, 180].

Table 5.8 – Prediction of absolute expression level from enhancer position

Experiment	MSE (random)	MSE	Correlation
CBP liver (ChIP-chip)	1.007+/-.003	0.915+/-.003	0.30 (p=1.6E-32)
CBP Cerebellum (ChIP-chip)	1.000+/-.003	0.939+/-.003	0.26 (p=3.3E-27)
CBP liver (ChIP-seq)	0.992+/-.002	0.911+/-.002	0.30 (p=1.0E-239)
3T3-L1 (ChIP-seq)	1.000+/-.002	0.939+/-.002	0.239 (p=1.2E-127)

Pearson correlation between absolute expression intensity measurements and model predictions on held out test data. The mean squared error and s.e.m. obtained by randomly guessing the training sample mean is shown in the first column. The mean squared error and s.e.m. of our model predictions are given in the 2nd column, and their correlation with observed intensities (and associated p-value from a 2-tailed t test) is given in column 3.

Notably, the shape of the enhancer influence function is nearly identical in both tissues. There is a sharp fall-off in influence as enhancer position moves further away from the TSS, and putative enhancers within approximately 10kb of the TSS are statistically associated with the highest transcript levels. Regulatory regions located upstream of the TSS are predicted to have a somewhat greater effect on transcription than downstream events. Interestingly, although some enhancers are known to act at very long distances, there is little predicted effect on transcript levels for enhancers located greater than 50kb away from the TSS.

5.7 Enhancer position predicts tissue-specific expression level

Regulatory sites identified in our ChIP experiments are statistically associated with tissue-specific expression of nearby genes (Figure 5.11). We therefore sought to examine whether observed expression differences between tissues could be explained by differences in binding of regulatory proteins.

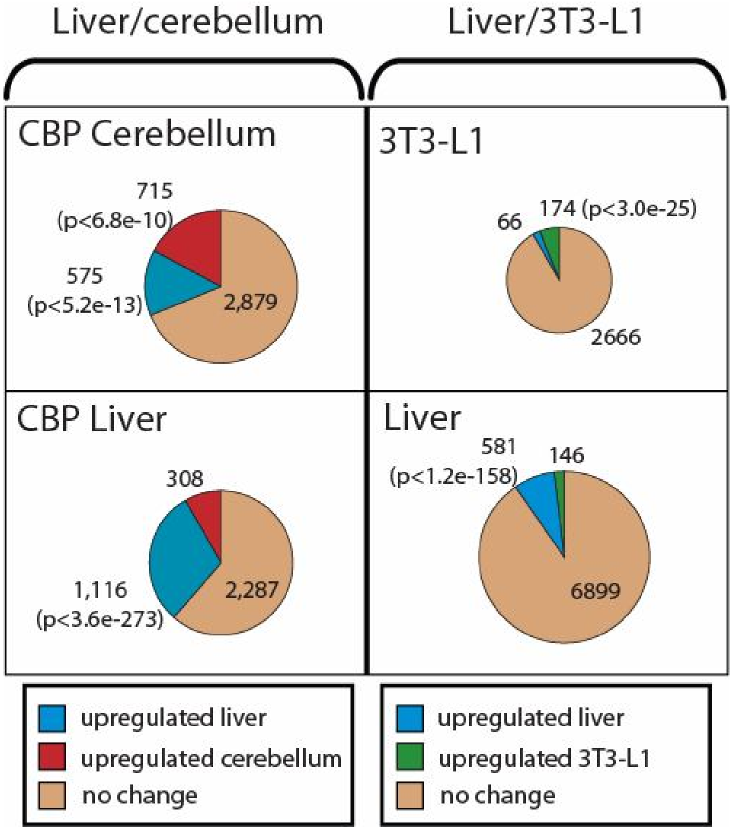


Figure 5.11: Bound genes are statistically associated with differential expression. Refseq genes with a CBP binding event within 5kb of their transcription start site were evaluated for differential expression in liver vs. cerebellum and hypergeometric p-values were calculated for the number of observed genes in each category. An identical analysis was performed for sites in liver and 3T3-L1 cells that were bound by p300 or at least two other transcriptional activators in ChIP-seq experiments.

We used all the liver and 3T3-L1 binding events identified in ChIP-seq experiments to predict relative expression of differentially expressed genes in these tissues. In order to evaluate the importance of binding site position in predicting the functional relevance, we compared our model's performance to two competing models: one that weighted binding events equally regardless of position, and a second that weighted the contributions of bound regions by sequence conservation, allowing highly conserved regulatory regions to be weighted differently than regions with low conservation. We fit each model using two-thirds of the bound, differentially expressed genes, and evaluated their ability to predict the magnitude of expression differences for the remaining third of the genes, repeating this process 100 times using randomly sampled test and training data.

The position-based model of transcription produces significantly more accurate predictions than the uniform weighting and the conservation-based approaches (Figure 5.12).

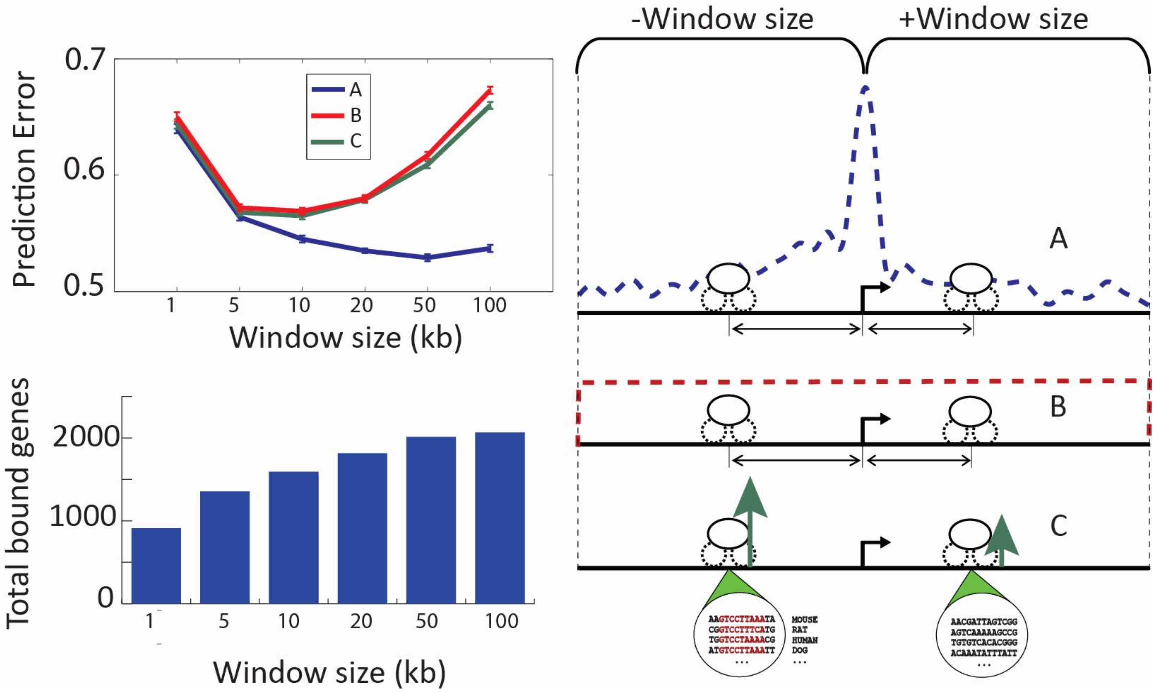


Figure 5.12: Binding site position, but not sequence conservation, is strongly associated with gene expression level. (A) The mean log expression of bound genes is shown in each tissue as a function of both the distance between the transcription start site and the nearest regulatory region identified by ChIP, and the maximum conservation score of any regulatory region within 5kb of that gene's TSS. Error bars indicate +/- s.e.m. Also shown is the Spearman correlation, and associated p-value from a right-tailed t-test, between log expression and the distance and conservation measures. (B) In the upper plot the mean log expression of genes in liver and 3T3-L1 cells is shown as a function of the location of the nearest binding site over a 200kb window. Error bars indicate +/- s.e.m. In the lower plot we show the influence function, which measures a binding event's predicted effect on expression as a function of position, obtained by fitting our predictive model to 1,000 bootstrapped samples of ChIP and expression data in each tissue. Shaded regions show the empirical 99% confidence intervals obtained from the bootstrap iterations.

To evaluate the importance of distal binding events in predicting expression, we identified bound genes using several distance cutoffs, ranging from the 1kb proximal promoter to a distance of 100kb from the gene's TSS. The position-based model outperforms the other models across a wide range of distance windows. At the 100kb cutoff, 2,205 of the 2,309 differentially expressed genes identified are bound in at least one tissue (Figure 5.12). Even when including these very distal sites in the analysis, many of which are presumably non-functional, our predictions have an extraordinary median correlation of 0.69 with observed expression levels of held-out test genes compared to 0.58 for the conservation-based model and 0.57 for the model that weights binding events uniformly. This value approaches the correlation level observed for relative expression measurements made using different experimental platforms [181, 182] and indicates that regulatory site position has a substantial effect on transcription levels in these tissues. Including binding events up to 50kb away from the TSS improves expression predictions, demonstrating the importance of these distal sites. However, weighting the influence of each regulatory region appropriately is crucial; the models that do not consider position both show a drastic deterioration in prediction accuracy as the distance cutoff increases. Interestingly, the simple uniform weighting model performs about as well as the model that weights sites by sequence conservation, indicating that conservation is of limited use in identifying functional binding events from ChIP data.

To address whether these data support the hypothesis that individual regulatory sites regulate multiple genes, we compared the prediction accuracy of our model to one where regulatory sites are assumed to regulate expression of only the closest transcript. We first associated binding events in liver and 3T3-L1 cells to transcripts, assuming they regulate only the nearest gene. We then trained our position-based transcriptional model and predicted the expression of held-out genes. These predictions were compared to those obtained, for the same set of genes, without the constraint that a site regulates a single gene. The difference in prediction accuracy is dramatic. The mean-squared prediction error over 100 bootstrapped trials was 0.73+/0.03 s.d. when we assume that binding events regulate only the closest gene. This improved by approximately 8 standard deviations to 0.48+/-0.02 s.d. when binding were allowed to regulate many genes.

5.8 Non-conserved binding is functional

To further explore the role of non-conserved regulatory sites we identified bound regions in each tissue that showed low sequence conservation levels, using the threshold that best distinguished bound regions from randomly selected DNA sequences (Figure 5.13).

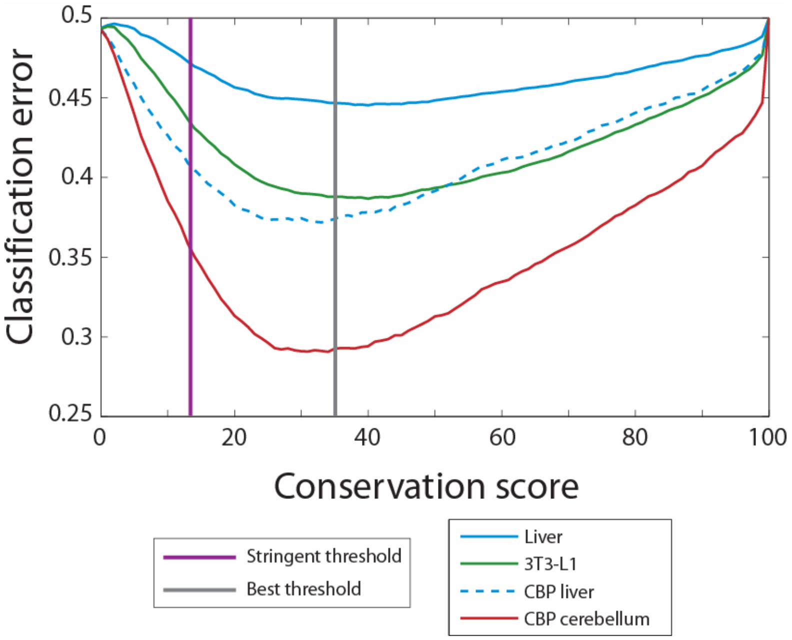


Figure 5.13: Conservation thresholds. We used conservation scores to distinguish bound regions in each tissue from sequences randomly selected from the mouse genome, evaluating classification error at 100 different thresholds. A conservation score of less than 35 was used to identify non-conserved sites since it yielded the lowest error rate across the four datasets. We identified a second, more stringent, threshold of 13 which yielded approximately 50% fewer conserved random sequences than the best threshold of 35.

At this threshold approximately 59% of sites from ChIP-seq experiments in liver and 47% in 3T3-L1 cells are non-conserved. Similarly, 44% of CBP sites in liver and 28% of sites in cerebellum are non-conserved. Genes located within 5kb of these sites in our experiments were associated with high levels of gene expression (Figure 5.14).

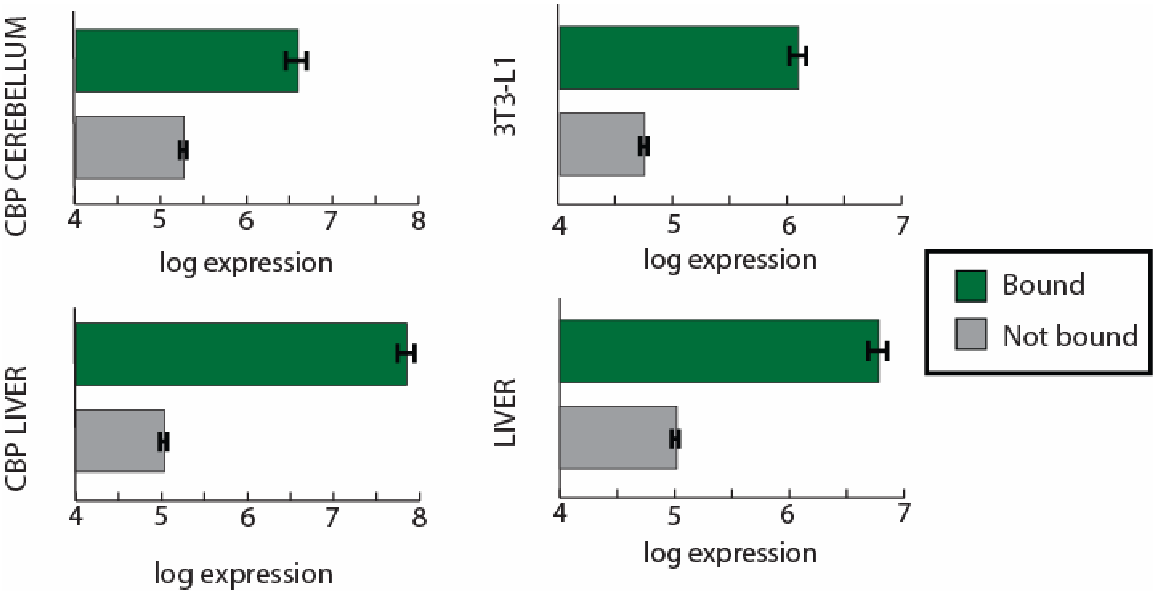


Figure 5.14: Expression of genes bound at nonconserved sites. Genes with a nonconserved binding event (identified using the most stringent threshold) located within 5kb of their transcription start site have a higher mean expression level than genes with no binding events. Error bars indicate +/- s.e.m.

Next we identified 261 differentially expressed genes in liver and 3T3-L1 cells bound (within 50kb) at only non-conserved regions. In a similar fashion we identified 884 differentially expressed genes bound only at non-conserved regions by CBP in liver and cerebellum. We performed the training and test procedure described above and determined whether the locations of these non-conserved sites predicted gene expression (Figure 5.15A). In both liver/3T3-L1 cells and in liver/cerebellum the position of non-conserved binding is a strong predictor of relative expression level. Our predictions have a mean correlation of 0.56 with observed expression values in liver/3T3-L1, significant at p<2.6e-9 by a right-tailed t-test. In liver/cerebellum the mean correlation is 0.57, significant at p<5.4e-26.

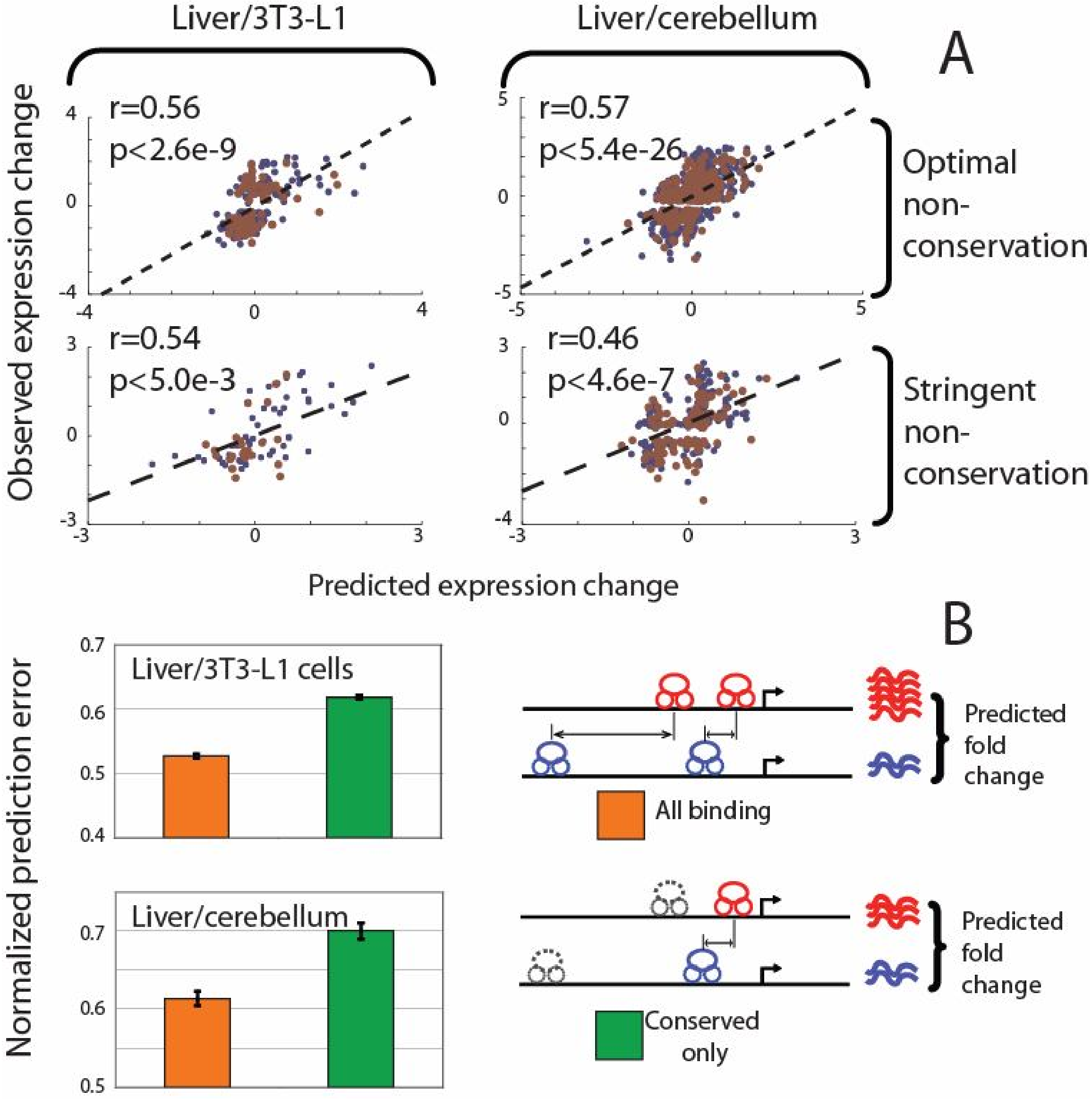


Figure 5.15: Non-conserved binding events predict expression. (A) Scatter plots of observed and predicted expression difference are presented for differentially expressed genes bound only at non-conserved regions at stringent and moderate conservation thresholds. Training data points are shown in blue and test data is shown in red. Each non-conserved binding event's effect on transcription was modulated by its distance to the TSS. In both tissue pairs, and at both conservation thresholds, the model's predictions are strongly correlated with observed expression differences. (B) The expression difference of genes bound at both conserved and non-conserved sites was predicted using only conserved sites, and the prediction error was compared to that obtained when both conserved and non-conserved binding sites were used. Including non-conserved regions significantly improved performance in both tissue pairs. Error bars indicate +/- s.e.m.

We then repeated the analysis using the stringent conservation threshold and found that non-conserved sites were still highly predictive of expression (Figure 5.15B). We also examined genes bound at both conserved and non-conserved sites within 100kb of their TSS and asked whether the conserved sites alone were adequate to predict expression. We first predicted expression using only conserved sites and then repeated the analysis using all bound regions. Underlining the importance of non-conserved regulatory regions, we find that considering both the conserved and non-conserved sites results in significantly more accurate predictions (Figure 5.15B).

5.9 Revealing the role of specific regulators

Although binding site position is very important in determining expression influence, the function of a regulatory region is also determined by the particular transcription factors that bind to it. We therefore extended our transcriptional model so that the relevance of any particular regulatory site was determined by both its location and the particular regulators bound. Each protein’s effect on transcription was estimated by including a protein-specific weight that modulated the expression influence of the site. We tested this approach on ChIP-seq and expression data in liver and 3T3-L1 cells, including binding data for an additional regulator, E2F4, in each tissue. We estimated the influence of p300, C/EBPα, FOXA1/A2, and E2F4 in liver, and p300, C/EBPα, PPARγ/RXR, and E2F4 in 3T3-L1 cells. In total, 2,038 differentially expressed genes were analyzed. Our predictions have an extraordinary median correlation of 0.74 with observed expression differences on held out test data, ranging between 0.72 and 0.76 in 11 separate trials (Figure 5.16). Our simple predictive framework remarkably accounts for over 50% of the variance in observed relative expression levels, and gives better predictions than a model that considers only binding site position.

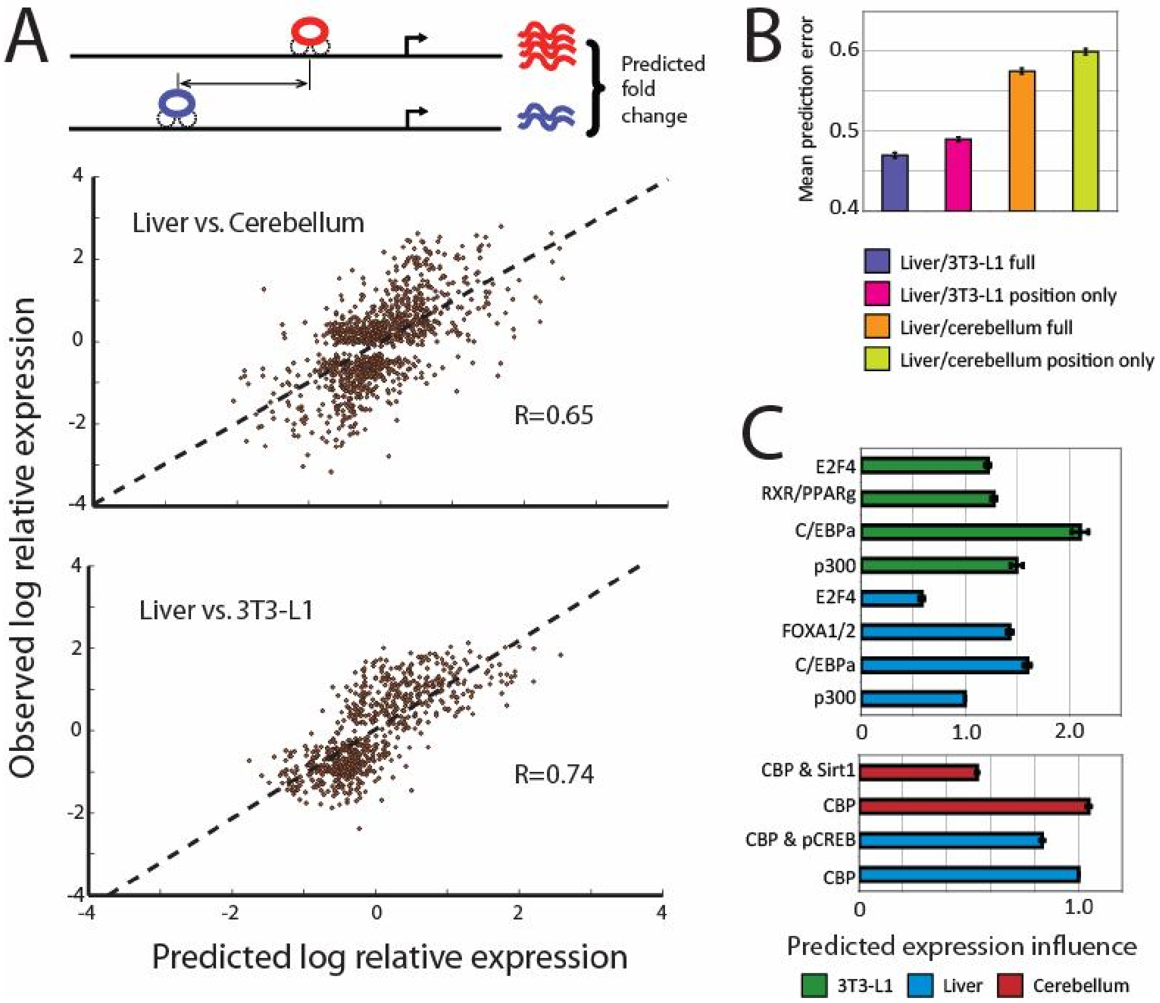


Figure 5.16: Transcriptional regulators have distinct influences on expression. (A) Shown are representative scatter plots of predicted vs. observed expression differences for held out test genes in liver/cerebellum and liver/3T3-L1 cells. Predictions were made using a transcriptional model that takes into account the influence of both the genomic position and the particular proteins bound by a site. The median correlation from 11 separate trials was 0.65 and 0.74 for liver/cerebellum and liver/3T3-L1 respectively. (B) The prediction error of the full model that includes individual transcription factor influence weights is compared to a model that uses only position to predict influence. Modeling the influence of bound regulators improves predictive performance. Error bars indicate +/- s.e.m. (C) The expression influence for each protein is learned in our transcriptional model. Sites bound by proteins with known repressive activity (E2F4 and Sirt1) are predicted to have the smallest influence.

The influence learned for each protein provides evidence of its function in these tissues. For example, C/EBPα is associated with the strongest activation in both cell types, in agreement with its well-characterized role in these tissues [135]. In contrast E2F4 is associated with the lowest levels of activation in both tissues; its influence weight of 0.52 in liver indicates that it actually attenuates an enhancer's effect on expression in this tissue, consistent with its previously described transcriptional repressor activity [137]. We performed a similar analysis in liver and cerebellum by collecting ChIP-seq data for the histone deacetylase Sirt1 in cerebellum, and ChIP-chip data for the transcription factor pCREB in liver. Modeling the different transcriptional influences of CBP sites that are also bound by pCREB or Sirt1 resulted in more accurate expression predictions. The median correlation between observed and predicted expression difference in liver and cerebellum was 0.65, ranging between 0.62 and 0.68 over 11 separate trials. Sirt1 has the opposite enzymatic activity to CBP/p300, and is known to repress p300 activation of transcription in certain contexts [183]. As expected, sites in cerebellum that are bound by Sirt1 have only about half as much influence on expression levels as CBP sites that do not recruit Sirt1.

5.10 Conclusions

In this chapter, we address a central problem in the study of transcriptional regulation by developing a model that reveals the function of transcription factor binding sites. Experimental approaches combining ChIP with microarray and sequencing technologies have led to tremendous progress in mapping transcriptional regulatory sites across the genome. However, progress in determining the function of these sites has been slower. In part this is because static maps of regulator binding give an incomplete picture of the complexity that arises from dynamic signaling and binding events, but progress has also been slowed by the absence of a simple framework that links regulatory network architecture (as defined by the location of regulatory regions in the genome) to transcription.

To understand the functional role of these regulatory sites, we developed a simple model that accurately predicts the expression difference between tissues based only on binding site positions. The correlation of the predictions with measured values approaches the correlation observed between different *experimental* platforms and can remarkably explain over half the variance in the relative transcription levels of differentially expressed genes.

Our findings suggest the need for a re-evaluation of how we understand and describe transcriptional controls. Regulatory sites are typically divided into promoter-proximal elements, which are within approximately 200 base pairs of the start site, and enhancer elements [2]. Surprisingly, we find an almost linear decrease in the effect of a regulatory site over a region of many kilobases, encompassing both proximal promoters and distal enhancers. Our results suggest that a more critical distinction may be between those binding events within or beyond 50 kilobases. We suggest that the latter be thought of as “remote enhancers,” the function of which remains to be elucidated.

Overall, our results suggest that regulatory events should not be thought of as belonging to enhancer/promoter categories or of being associated with a particular gene, but rather in quantitative terms. The net transcription level of a gene is the result of integrating a potentially large number of binding events. Although binding events that are very close to the transcription start site have a disproportionately large effect on expression, many genes show large differences in tissue-specific expression that are apparently driven by much more remote events. Our transcriptional model can accurately predict expression even when no binding event is detected within 1kb of the TSS (Figure 5.17).

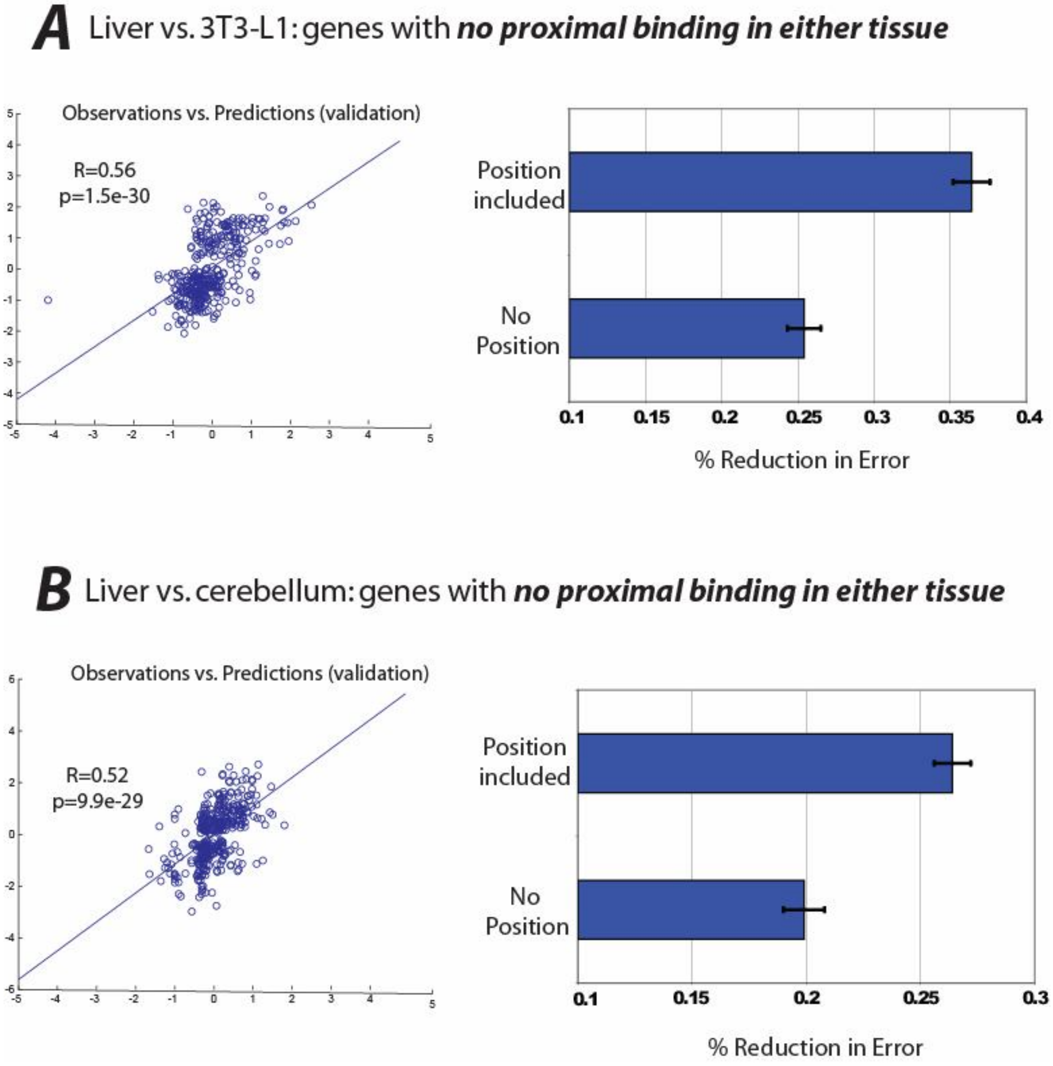


Figure 5.17: Enhancer position predicts expression for genes with no proximal binding events. Representative scatter plots of observed and predicted normalized expression difference for held-out validation data are presented for differentially expressed genes in liver and cerebellum (A) and genes in liver and 3T3-L1 cells (B) with no binding events within 1kb of the TSS. To demonstrate the importance of position even for binding events within 10kb of regulated genes we excluded any binding event located more than 10kb from the gene's TSS. Median correlation between observed and predicted expression difference is greater than 0.5 and highly statistically significant for both analyses. The reduction in mean test error relative to random guessing is shown for the full predictive model and for a model that ignores binding position. In both cases, modeling the effect of position significantly improves performance.

Interestingly, our analysis supports a model of transcription where binding events frequently regulate the expression of multiple genes. Based on our observation that binding sites located within 50kb of a gene significantly influence its expression level, we estimate that approximately 40-45% of regulatory sites affect the expression of more than one transcript.

In contrast to the strong relationship between the location of binding and transcription, there is little relationship between sequence conservation and expression. Including binding to non-conserved sequences in our models improves their accuracy significantly over models built using only binding to conserved sequences. Previously, we have shown that the sites targeted by individual DNA-binding proteins can vary across species even when tissue-specific gene expression is conserved [172]. Taken together, these findings suggest that organisms can achieve similar gene expression patterns through diverse mechanisms. Because transcription integrates binding events that are distributed over great distances, there is a reasonable probability that the evolutionary gain or loss of regulatory regions at one locus can be compensated for by mutations at other sites. More work is needed to whether the quantitative relationship between binding and expression is similar across mammals.

The results presented here represent a significant step towards a quantitative framework for understanding gene expression. The statistical relationship between enhancer position and transcription level is clear, and this observation should lead to more accurate models of transcriptional regulation. However, many other factors have a profound effect on enhancer function including which coregulators are recruited, the nuclear concentrations of transcription factors, binding of small molecules that modulate enzymatic activities and interaction surfaces, and any signaling events leading to post-translational modification of regulators. Enriching the modeling framework presented here by incorporating data describing such events may lead to a greater understanding of regulatory networks and their relationship to developmental and disease processes.

Chapter 6: Conclusions and Thesis Contributions

I will conclude this thesis by summarizing the work presented in previous chapters and outlining what I feel are its main contributions. The focus of the research presented here has been on understanding an important subset of the interactions that control transcriptional regulation: namely the binding interactions that result in recruitment of transcriptional regulators to their genomic targets. We have developed tools to probe three key aspects of these interactions: descriptions of their specificity, physical models of their behavior, and their ultimate effect on transcription. Describing the specificity of binding interactions involves either learning representations of a protein's binding specificity from experimental data, or predicting which proteins recruit it to its targets when it has no DNA binding activity. Accurate descriptions of binding interactions are invaluable in obtaining a reasonable representation of transcriptional regulatory architecture. However, understanding the behavior of networks in response to perturbations relies on a reasonable physical model of how the components of the system interact and behave in different settings. Finally, any useful description of a regulatory network must relate how binding events ultimately affect transcription. Each chapter of this thesis has focused on one or more of these issues.

In chapter 2 we presented a motif discovery method, called Converge, which uses phylogenetic conservation information to improve motif discovery performance. Here our goal was to develop a tool that could be used to obtain accurate descriptions of a protein's DNA binding specificity from a set of sequences that are likely to be bound by the protein *in vivo* as well as pair-wise alignments of orthologous sequence from related species. The chief contributions of this work are:

1. a novel generative model of sequence that leverages conservation information allowing both a motif and a simple measure of evolutionary relatedness to be learned from alignment data.
2. a re-analysis of ChIP-chip data for 172 yeast transcription factors resulting in new and corrected binding specificities for several factors, and a significantly expanded set of high-confidence regulatory interactions.

Chapter 3, like chapter 2, focused on the problem of determining a protein's binding specificity. Here we presented a hypothesis-testing approach, called THEME, which formulates motif discovery as a model selection problem. The motif most likely to correspond to a protein's true binding specificity is assumed to be the motif that best discriminates bound and unbound sequences in a ChIP experiment. The chief contributions of this chapter are:

1. a novel computational framework for integrating prior information about a protein's DNA binding specificity into the motif search
2. the first discriminative motif analysis method to employ a cross-validated approach for ranking candidate motifs and protecting against overfitting.

In chapter 4, we presented a biophysically-motivated modeling framework for DNA-protein interactions. Here we were concerned not only with learning accurate representations of binding specificities, but also with developing interpretable and physically realistic models of protein recruitment to the genome *in vivo*. We first demonstrated that treating protein binding as a simple bimolecular reaction at equilibrium with a reaction free energy that is a simple sum of contributions from each position in the binding site allows us to derive a convenient logistic function expression for the binding probability. We then demonstrated how this framework could be adapted for use in THEME's hypothesis-testing approach to motif discovery. Next, we extend the approach to use raw ChIP-seq count data and demonstrate how this allows us to test hypotheses about binding specificity and concentration changes across conditions, and jointly analyze ChIP-seq data for factors that compete for common binding sites. The main contributions of this chapter of the thesis are:

1. an intuitive model of sequence-specific protein binding grounded in biophysical principles.
2. integration of this model into the THEME hypothesis-testing framework for motif analysis, with performance validation on a varied group of mammalian datasets.
3. extension of the biophysical framework to include experimental evidence of relative binding levels in the form of ChIP-seq count data.
4. a general stochastic simulation method for obtaining samples from an approximation to the posterior distribution of binding configurations in our biophysical modeling setting.
5. presentation of several natural applications of the proposed modeling framework including: joint analysis of binding for a factor in two conditions, analysis of competitive binding, mixture models of binding specificity, and coregulator recruitment.

Finally in chapter 5 we present an analysis of ChIP-seq data for several transcriptional regulators in a number of mouse tissues. We first performed a sequence analysis of bound regions from ChIP-seq experiments for the coregulator CBP/p300 in order to predict which transcription factors it associates with in each tissue, and then validated the predictions of this analysis with follow-up ChIP experiments. We then presented a simple model relating gene expression to the location and proteins bound at the regulatory regions identified by ChIP. The chief contributions of this chapter are:

1. the key insight that the location of regulatory regions relative to the TSS is strongly associated with expression level of the gene. This has important implications for understanding regulatory network function and evolution.
2. a novel and accurate model of expression as a function of regulator binding location that explains a large fraction of variance in expression level between tissues and can help infer the regulatory roles of transcription factors.
3. a method for performing sequence analysis of genomic regions bound by coregulators. Several predictions were validated experimentally, and further analysis suggested that a model of independently contributing proteins is better supported by the data than more complex models that include cooperative recruitment by pairs of regulators.

In this thesis I have, hopefully, helped lay the foundation for further work in understanding the relationship between genomic binding events and downstream transcriptional effects.

Bibliography

1. Alberts, B., Johnson, Lewis, Raff, Roberts, Walter, *Molecular Biology of the Cell*. 2002: Garland.
2. Lodish, H., Berk, Krieger, Kaiser, Scott, Bretscher, Ploegh, Matsudaira, *Molecular Biology of the Cell*. 2007: W. H. Freeman.
3. Cahill, G.F., Jr., *Fuel metabolism in starvation.* Annu Rev Nutr, 2006. **26**: p. 1-22.
4. Kummerfeld, S.K. and S.A. Teichmann, *DBD: a transcription factor prediction database.* Nucleic Acids Res, 2006. **34**(Database issue): p. D74-81.
5. Naar, A.M., B.D. Lemon, and R. Tjian, *Transcriptional coactivator complexes.* Annu Rev Biochem, 2001. **70**: p. 475-501.
6. Reid, G., R. Gallais, and R. Metivier, *Marking time: the dynamic role of chromatin and covalent modification in transcription.* Int J Biochem Cell Biol, 2009. **41**(1): p. 155-63.
7. Banerji, J., S. Rusconi, and W. Schaffner, *Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences.* Cell, 1981. **27**(2 Pt 1): p. 299-308.
8. Bondarenko, V.A., et al., *Communication over a large distance: enhancers and insulators.* Biochem Cell Biol, 2003. **81**(3): p. 241-51.
9. Hatzis, P. and I. Talianidis, *Dynamics of enhancer-promoter communication during differentiation-induced gene activation.* Mol Cell, 2002. **10**(6): p. 1467-77.
10. Arnosti, D.N. and M.M. Kulkarni, *Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards?* J Cell Biochem, 2005. **94**(5): p. 890-8.
11. Lynch, V.J. and G.P. Wagner, *Resurrecting the role of transcription factor change in developmental evolution.* Evolution, 2008. **62**(9): p. 2131-54.
12. Takahashi, K., et al., *Induction of pluripotent stem cells from adult human fibroblasts by defined factors.* Cell, 2007. **131**(5): p. 861-72.
13. Takahashi, K. and S. Yamanaka, *Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.* Cell, 2006. **126**(4): p. 663-76.
14. Yu, J., et al., *Induced pluripotent stem cell lines derived from human somatic cells.* Science, 2007. **318**(5858): p. 1917-20.
15. Bai, Z. and R. Gust, *Breast cancer, estrogen receptor and ligands.* Arch Pharm (Weinheim), 2009. **342**(3): p. 133-49.
16. Bassett, E.A., et al., *Structural and functional basis for therapeutic modulation of p53 signaling.* Clin Cancer Res, 2008. **14**(20): p. 6376-86.
17. Christensen, K.L., et al., *The six family of homeobox genes in development and cancer.* Adv Cancer Res, 2008. **101**: p. 93-126.
18. Sugden, M.C. and M.J. Holness, *Role of nuclear receptors in the modulation of insulin secretion in lipid-induced insulin resistance.* Biochem Soc Trans, 2008. **36**(Pt 5): p. 891-900.
19. Roelfsema, J.H. and D.J. Peters, *Rubinstein-Taybi syndrome: clinical and molecular overview.* Expert Rev Mol Med, 2007. **9**(23): p. 1-16.
20. Cousins, D.J., J. McDonald, and T.H. Lee, *Therapeutic approaches for control of transcription factors in allergic disease.* J Allergy Clin Immunol, 2008. **121**(4): p. 803-9; quiz 810-1.

21. Villard, J., *Transcription regulation and human diseases.* Swiss Med Wkly, 2004. **134**(39-40): p. 571-9.
22. Seidman, J.G. and C. Seidman, *Transcription factor haploinsufficiency: when half a loaf is not enough.* J Clin Invest, 2002. **109**(4): p. 451-5.
23. Firestein, G.S. and A.M. Manning, *Signal transduction and transcription factors in rheumatic disease.* Arthritis Rheum, 1999. **42**(4): p. 609-21.
24. Englesberg, E., et al., *Positive control of enzyme synthesis by gene C in the L-arabinose system.* J Bacteriol, 1965. **90**(4): p. 946-57.
25. Jacob, F. and J. Monod, *Genetic regulatory mechanisms in the synthesis of proteins.* J Mol Biol, 1961. **3**: p. 318-56.
26. Ippen, K., et al., *New controlling element in the Lac operon of E. coli.* Nature, 1968. **217**(5131): p. 825-7.
27. Tjian, R., *The binding site on SV40 DNA for a T antigen-related protein.* Cell, 1978. **13**(1): p. 165-79.
28. Ptashne, M., *Regulation of transcription: from lambda to eukaryotes.* Trends Biochem Sci, 2005. **30**(6): p. 275-9.
29. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.
30. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
31. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
32. Gibbs, R.A., et al., *Genome sequence of the Brown Norway rat yields insights into mammalian evolution.* Nature, 2004. **428**(6982): p. 493-521.
33. Waterston, R.H., et al., *Initial sequencing and comparative analysis of the mouse genome.* Nature, 2002. **420**(6915): p. 520-62.
34. Ren, B., et al., *Genome-wide location and function of DNA binding proteins.* Science, 2000. **290**(5500): p. 2306-9.
35. Johnson, D.S., et al., *Genome-wide mapping of in vivo protein-DNA interactions.* Science, 2007. **316**(5830): p. 1497-502.
36. Bilitewski, U., *DNA microarrays: an introduction to the technology.* Methods Mol Biol, 2009. **509**: p. 1-14.
37. Lockhart, D.J., et al., *Expression monitoring by hybridization to high-density oligonucleotide arrays.* Nat Biotechnol, 1996. **14**(13): p. 1675-80.
38. Shalon, D., S.J. Smith, and P.O. Brown, *A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization.* Genome Res, 1996. **6**(7): p. 639-45.
39. Robertson, G., et al., *Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing.* Nat Methods, 2007. **4**(8): p. 651-7.
40. Buck, M.J., A.B. Nobel, and J.D. Lieb, *ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data.* Genome Biol, 2005. **6**(11): p. R97.
41. Keles, S., et al., *Multiple testing methods for ChIP-Chip high density oligonucleotide array data.* J Comput Biol, 2006. **13**(3): p. 579-613.
42. Qi, Y., et al., *High-resolution computational models of genome binding events.* Nature Biotechnology, 2006. **24**(8): p. 963-970.

43. Zheng, M., et al., *ChIP-chip: data, model, and analysis.* Biometrics, 2007. **63**(3): p. 787-96.
44. Zhang, Z.D., et al., *Modeling ChIP sequencing in silico with applications.* PLoS Comput Biol, 2008. **4**(8): p. e1000158.
45. Vega, V.B., et al., *Inherent signals in sequencing-based Chromatin-ImmunoPrecipitation control libraries.* PLoS ONE, 2009. **4**(4): p. e5241.
46. Zhang, Y., et al., *Model-based analysis of ChIP-Seq (MACS).* Genome Biol, 2008. **9**(9): p. R137.
47. Gardiner-Garden, M. and M. Frommer, *CpG islands in vertebrate genomes.* J Mol Biol, 1987. **196**(2): p. 261-82.
48. Murray, J.I., et al., *Identification of motifs that function in the splicing of non-canonical introns.* Genome Biol, 2008. **9**(6): p. R97.
49. Segal, E., et al., *A genomic code for nucleosome positioning.* Nature, 2006. **442**(7104): p. 772-8.
50. Cornish-Bowden, A., *Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984.* Nucleic Acids Res, 1985. **13**(9): p. 3021-30.
51. Day, W.H. and F.R. McMorris, *Critical comparison of consensus methods for molecular sequences.* Nucleic Acids Res, 1992. **20**(5): p. 1093-9.
52. Stormo, G.D., et al., *Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli.* Nucleic Acids Res, 1982. **10**(9): p. 2997-3011.
53. Djordjevic, M., A.M. Sengupta, and B.I. Shraiman, *A biophysical approach to transcription factor binding site discovery.* Genome Res, 2003. **13**(11): p. 2381-90.
54. Leung, H.C., et al., *Finding motifs with insufficient number of strong binding sites.* J Comput Biol, 2005. **12**(6): p. 686-701.
55. Stormo, G.D., *DNA binding sites: representation and discovery.* Bioinformatics, 2000. **16**(1): p. 16-23.
56. Schneider, T.D. and R.M. Stephens, *Sequence logos: a new way to display consensus sequences.* Nucleic Acids Res, 1990. **18**(20): p. 6097-100.
57. Bulyk, M.L., 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, 2002. **30**(5): p. 1255-61.
58. Man, T.K. 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, 2001. **29**(12): p. 2471-8.
59. Zhou, Q. and J.S. Liu, *Modeling within-motif dependence for transcription factor binding site predictions.* Bioinformatics, 2004. **20**(6): p. 909-16.
60. Hong, P., et al., *A boosting approach for motif modeling using ChIP-chip data.* Bioinformatics, 2005. **21**(11): p. 2636-43.
61. Xing, E.P., et al., *Logos: a modular bayesian model for de novo motif detection.* J Bioinform Comput Biol, 2004. **2**(1): p. 127-54.
62. Benos, P.V., M.L. Bulyk, and G.D. Stormo, *Additivity in protein-DNA interactions: how good an approximation is it?* Nucleic Acids Res, 2002. **30**(20): p. 4442-51.

63. Pavesi, G., G. Mauri, and G. Pesole, *An algorithm for finding signals of unknown length in DNA sequences.* Bioinformatics, 2001. **17 Suppl 1**: p. S207-14.
64. Blanchette, M. and S. Sinha, *Separating real motifs from their artifacts.* Bioinformatics, 2001. **17 Suppl 1**: p. S30-8.
65. Eskin, E. and P.A. Pevzner, *Finding composite regulatory patterns in DNA sequences.* Bioinformatics, 2002. **18 Suppl 1**: p. S354-63.
66. Bailey, T.L. and C. Elkan, *Fitting a mixture model by expectation maximization to discover motifs in biopolymers.* Proc Int Conf Intell Syst Mol Biol, 1994. **2**: p. 28-36.
67. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum Likelihood from Incomplete Data Via Em Algorithm.* Journal of the Royal Statistical Society Series B-Methodological, 1977. **39**(1): p. 1-38.
68. Geman, S. and D. Geman, *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images.* Ieee Transactions on Pattern Analysis and Machine Intelligence, 1984. **6**(6): p. 721-741.
69. Bussemaker, H.J., H. Li, and E.D. Siggia, *Regulatory element detection using correlation with expression.* Nat Genet, 2001. **27**(2): p. 167-71.
70. Conlon, E.M., et al., *Integrating regulatory motif discovery and genome-wide expression analysis.* Proc Natl Acad Sci U S A, 2003. **100**(6): p. 3339-44.
71. Keles, S., M. van der Laan, and M.B. Eisen, *Identification of regulatory elements using a feature selection method.* Bioinformatics, 2002. **18**(9): p. 1167-75.
72. Das, D., N. Banerjee, and M.Q. Zhang, *Interacting models of cooperative gene regulation.* Proc Natl Acad Sci U S A, 2004. **101**(46): p. 16234-9.
73. Beer, M.A. and S. Tavazoie, *Predicting gene expression from sequence.* Cell, 2004. **117**(2): p. 185-98.
74. Yuan, Y., et al., *Predicting gene expression from sequence: a reexamination.* PLoS Comput Biol, 2007. **3**(11): p. e243.
75. Segal, E., et al., *Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.* Nat Genet, 2003. **34**(2): p. 166-76.
76. Bar-Joseph, Z., et al., *Computational discovery of gene modules and regulatory networks.* Nat Biotechnol, 2003. **21**(11): p. 1337-42.
77. Zhou, X.J., et al., *Functional annotation and network reconstruction through cross-platform integration of microarray data.* Nat Biotechnol, 2005. **23**(2): p. 238-43.
78. Kato, M., et al., *Identifying combinatorial regulation of transcription factors and binding motifs.* Genome Biol, 2004. **5**(8): p. R56.
79. Lemmens, K., et al., *Inferring transcriptional modules from ChIP-chip, motif and microarray data.* Genome Biol, 2006. **7**(5): p. R37.
80. Gerber, G.K., et al., *Automated discovery of functional generality of human gene expression programs.* PLoS Comput Biol, 2007. **3**(8): p. e148.
81. Niida, A., et al., *Gene set-based module discovery in the breast cancer transcriptome.* BMC Bioinformatics, 2009. **10**(1): p. 71.
82. Yeang, C.H. and T. Jaakkola, *Modeling the combinatorial functions of multiple transcription factors.* J Comput Biol, 2006. **13**(2): p. 463-80.

83. Kellis, M., et al., *Sequencing and comparison of yeast species to identify genes and regulatory elements.* Nature, 2003. **423**(6937): p. 241-54.

84. Moses, A.M., et al., *Position specific variation in the rate of evolution in transcription factor binding sites.* BMC Evol Biol, 2003. **3**: p. 19.

85. Chin, C.S., J.H. Chuang, and H. Li, *Genome-wide regulatory complexity in yeast promoters: separation of functionally conserved and neutral sequence.* Genome Res, 2005. **15**(2): p. 205-13.

86. Lenhard, B., et al., *Identification of conserved regulatory elements by comparative genome analysis.* J Biol, 2003. **2**(2): p. 13.

87. Wasserman, W.W., et al., *Human-mouse genome comparisons to locate regulatory sites.* Nat Genet, 2000. **26**(2): p. 225-8.

88. Duret, L. and P. Bucher, *Searching for regulatory elements in human noncoding sequences.* Curr Opin Struct Biol, 1997. **7**(3): p. 399-406.

89. Zhang, Z. and M. Gerstein, *Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements.* J Biol, 2003. **2**(2): p. 11.

90. Qin, Z.S., et al., *Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites.* Nat Biotechnol, 2003. **21**(4): p. 435-9.

91. Jensen, S.T., L. Shen, and J.S. Liu, *Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes.* Bioinformatics, 2005. **21**(20): p. 3832-9.

92. Cliften, P., et al., *Finding functional features in Saccharomyces genomes by phylogenetic footprinting.* Science, 2003. **301**(5629): p. 71-6.

93. Xie, X., et al., *Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.* Nature, 2005. **434**(7031): p. 338-45.

94. Moses, A.M., D.Y. Chiang, and M.B. Eisen, *Phylogenetic motif detection by expectation-maximization on evolutionary mixtures.* Pac Symp Biocomput, 2004: p. 324-35.

95. Sinha, S., M. Blanchette, and M. Tompa, *PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences.* BMC Bioinformatics, 2004. **5**: p. 170.

96. Sinha, S., *PhyME: a software tool for finding motifs in sets of orthologous sequences.* Methods Mol Biol, 2007. **395**: p. 309-18.

97. Liu, Y., et al., *Eukaryotic regulatory element conservation analysis and identification using comparative genomics.* Genome Res, 2004. **14**(3): p. 451-8.

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

99. Harbison, C.T., et al., *Transcriptional regulatory code of a eukaryotic genome.* Nature, 2004. **431**(7004): p. 99-104.

100. Burglin, T.R., *The yeast regulatory gene PHO2 encodes a homeo box.* Cell, 1988. **53**(3): p. 339-40.

101. Dorrington, R.A. and T.G. Cooper, *The DAL82 protein of Saccharomyces cerevisiae binds to the DAL upstream induction sequence (UIS).* Nucleic Acids Res, 1993. **21**(16): p. 3777-84.

102. Scott, S., A.T. Abul-Hamd, and T.G. Cooper, *Roles of the Dal82p domains in allophanate/oxalurate-dependent gene expression in Saccharomyces cerevisiae.* J Biol Chem, 2000. **275**(40): p. 30886-93.
103. Li, X. and W.H. Wong, *Sampling motifs on phylogenetic trees.* Proc Natl Acad Sci U S A, 2005. **102**(27): p. 9481-6.
104. Wang, T. and G.D. Stormo, *Combining phylogenetic data with co-regulated genes to identify regulatory motifs.* Bioinformatics, 2003. **19**(18): p. 2369-80.
105. Gasch, A.P., et al., *Genomic expression programs in the response of yeast cells to environmental changes.* Mol Biol Cell, 2000. **11**(12): p. 4241-57.
106. Miled, C., C. Mann, and G. Faye, *Xbp1-mediated repression of CLB gene expression contributes to the modifications of yeast cell morphology and cell cycle seen during nitrogen-limited growth.* Mol Cell Biol, 2001. **21**(11): p. 3714-24.
107. Lee, T.I., et al., *Transcriptional regulatory networks in Saccharomyces cerevisiae.* Science, 2002. **298**(5594): p. 799-804.
108. Deckert, J., et al., *Multiple elements and auto-repression regulate Rox1, a repressor of hypoxic genes in Saccharomyces cerevisiae.* Genetics, 1995. **139**(3): p. 1149-58.
109. Zhao, H. and D.J. Eide, *Zap1p, a metalloregulatory protein involved in zinc-responsive transcriptional regulation in Saccharomyces cerevisiae.* Mol Cell Biol, 1997. **17**(9): p. 5044-52.
110. Gimeno, C.J. and G.R. Fink, *Induction of pseudohyphal growth by overexpression of PHD1, a Saccharomyces cerevisiae gene related to transcriptional regulators of fungal development.* Mol Cell Biol, 1994. **14**(3): p. 2100-12.
111. Lorenz, M.C. and J. Heitman, *Regulators of pseudohyphal differentiation in Saccharomyces cerevisiae identified through multicopy suppressor analysis in ammonium permease mutant strains.* Genetics, 1998. **150**(4): p. 1443-57.
112. Ward, M.P., et al., *SOK2 may regulate cyclic AMP-dependent protein kinase-stimulated growth and pseudohyphal development by repressing transcription.* Mol Cell Biol, 1995. **15**(12): p. 6854-63.
113. Mai, B. and L. Breeden, *Xbp1, a stress-induced transcriptional repressor of the Saccharomyces cerevisiae Swi4/Mbp1 family.* Mol Cell Biol, 1997. **17**(11): p. 6491-501.
114. de Nadal, E., L. Casadome, and F. Posas, *Targeting the MEF2-like transcription factor Smp1 by the stress-activated Hog1 mitogen-activated protein kinase.* Mol Cell Biol, 2003. **23**(1): p. 229-37.
115. Tompa, M., et al., *Assessing computational tools for the discovery of transcription factor binding sites.* Nat Biotechnol, 2005. **23**(1): p. 137-44.
116. Wingender, E., et al., *TRANSFAC: an integrated system for gene expression regulation.* Nucleic Acids Research, 2000. **28**(1): p. 316-319.
117. Sandelin, A. and W.W. Wasserman, *Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics.* J Mol Biol, 2004. **338**(2): p. 207-15.
118. MacIsaac, K.D., et al., *A hypothesis-based approach for identifying the binding specificity of regulatory proteins from chromatin immunoprecipitation data.* Bioinformatics, 2006. **22**(4): p. 423-9.

119. Chawla, N.V., et al., *SMOTE: Synthetic minority over-sampling technique.* Journal of Artificial Intelligence Research, 2002. **16**: p. 321-357.

120. Marsich, E., et al., *The PAX6 gene is activated by the basic helix-loop-helix transcription factor NeuroD/BETA2.* Biochem J, 2003. **376**(Pt 3): p. 707-15.

121. Morozov, A.V., et al., *Protein-DNA binding specificity predictions with structural models.* Nucleic Acids Res, 2005. **33**(18): p. 5781-98.

122. Liu, L.A. and J.S. Bader, *Decoding transcriptional regulatory interactions.* Physica D, 2006. **224**(1-2): p. 174-181.

123. Stormo, G.D. and D.S. Fields, *Specificity, free energy and information content in protein-DNA interactions.* Trends Biochem Sci, 1998. **23**(3): p. 109-13.

124. Berg, O.G. and P.H. von Hippel, *Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters.* J Mol Biol, 1987. **193**(4): p. 723-50.

125. Foat, B.C., A.V. Morozov, and H.J. Bussemaker, *Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE.* Bioinformatics, 2006. **22**(14): p. e141-9.

126. Segal, E., et al., *Predicting expression patterns from regulatory sequence in Drosophila segmentation.* Nature, 2008. **451**(7178): p. 535-40.

127. Gertz, J., E.D. Siggia, and B.A. Cohen, *Analysis of combinatorial cis-regulation in synthetic and genomic promoters.* Nature, 2009. **457**(7226): p. 215-8.

128. Genkin, A., D.D. Lewis, and D. Madigan, *Large-scale Bayesian logistic regression for text categorization.* Technometrics, 2007. **49**(3): p. 291-304.

129. Tanay, A., *Extensive low-affinity transcriptional interactions in the yeast genome.* Genome Res, 2006. **16**(8): p. 962-72.

130. Jothi, R., et al., *Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data.* Nucleic Acids Res, 2008. **36**(16): p. 5221-31.

131. Gillespie, D.T., *Exact Stochastic Simulation of Coupled Chemical-Reactions.* Journal of Physical Chemistry, 1977. **81**(25): p. 2340-2361.

132. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.

133. Kanehisa, M., et al., *The KEGG resource for deciphering the genome.* Nucleic Acids Res, 2004. **32**(Database issue): p. D277-80.

134. Pedersen, T.A., et al., *Distinct C/EBPalpha motifs regulate lipogenic and gluconeogenic gene expression in vivo.* EMBO J, 2007. **26**(4): p. 1081-93.

135. Roesler, W.J., *The role of C/EBP in nutrient and hormonal regulation of gene expression.* Annu Rev Nutr, 2001. **21**: p. 141-65.

136. Friedman, J.R. and K.H. Kaestner, *The Foxa family of transcription factors in development and metabolism.* Cell Mol Life Sci, 2006. **63**(19-20): p. 2317-28.

137. Trimarchi, J.M. and J.A. Lees, *Sibling rivalry in the E2F family.* Nat Rev Mol Cell Biol, 2002. **3**(1): p. 11-20.

138. Koo, S.H. and M. Montminy, *Fatty acids and insulin resistance: a perfect storm.* Mol Cell, 2006. **21**(4): p. 449-50.

139. Wolfrum, C., et al., *Foxa2 regulates lipid metabolism and ketogenesis in the liver during fasting and in diabetes.* Nature, 2004. **432**(7020): p. 1027-32.

140. Brown, M.S. and J.L. Goldstein, *The SREBP pathway: regulation of cholesterol metabolism by proteolysis of a membrane-bound transcription factor.* Cell, 1997. **89**(3): p. 331-40.

141. Ting, H.J., et al., *Androgen-receptor coregulators mediate the suppressive effect of androgen signals on vitamin D receptor activity.* Endocrine, 2005. **26**(1): p. 1-9.

142. Wei, L.N., *Retinoid receptors and their coregulators.* Annu Rev Pharmacol Toxicol, 2003. **43**: p. 47-72.

143. An, W., J. Kim, and R.G. Roeder, *Ordered cooperative functions of PRMT1, p300, and CARM1 in transcriptional activation by p53.* Cell, 2004. **117**(6): p. 735-48.

144. Chen, X., et al., *Integration of external signaling pathways with the core transcriptional network in embryonic stem cells.* Cell, 2008. **133**(6): p. 1106-17.

145. Ferrari, R., et al., *Epigenetic reprogramming by adenovirus e1a.* Science, 2008. **321**(5892): p. 1086-8.

146. Odom, D.T., et al., *Core transcriptional regulatory circuitry in human hepatocytes.* Mol Syst Biol, 2006. **2**: p. 2006 0017.

147. Li, X.Y., et al., *Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm.* PLoS Biol, 2008. **6**(2): p. e27.

148. Visel, A., et al., *ChIP-seq accurately predicts tissue-specific activity of enhancers.* Nature, 2009. **457**(7231): p. 854-8.

149. Heintzman, N.D., et al., *Histone modifications at human enhancers reflect global cell-type-specific gene expression.* Nature, 2009.

150. Crissey, M.A., et al., *Liver-specific and proliferation-induced deoxyribonuclease I hypersensitive sites in the mouse insulin-like growth factor binding protein-1 gene.* Hepatology, 1999. **30**(5): p. 1187-97.

151. Hanson, R.W. and L. Reshef, *Regulation of phosphoenolpyruvate carboxykinase (GTP) gene expression.* Annu Rev Biochem, 1997. **66**: p. 581-611.

152. Louet, J.F., et al., *The coactivator PGC-1 is involved in the regulation of the liver carnitine palmitoyltransferase I gene expression by cAMP in combination with HNF4 alpha and cAMP-response element-binding protein (CREB).* J Biol Chem, 2002. **277**(41): p. 37991-8000.

153. Maire, P., J. Wuarin, and U. Schibler, *The role of cis-acting promoter elements in tissue-specific albumin gene expression.* Science, 1989. **244**(4902): p. 343-6.

154. Onuma, H., et al., *Insulin and epidermal growth factor suppress basal glucose-6-phosphatase catalytic subunit gene transcription through overlapping but distinct mechanisms.* Biochem J, 2009. **417**(2): p. 611-20.

155. Travnickova-Bendova, Z., et al., *Bimodal regulation of mPeriod promoters by CREB-dependent signaling and CLOCK/BMAL1 activity.* Proc Natl Acad Sci U S A, 2002. **99**(11): p. 7728-33.

156. Yamamoto, T., et al., *Acute physical stress elevates mouse period1 mRNA expression in mouse peripheral tissues via a glucocorticoid-responsive element.* J Biol Chem, 2005. **280**(51): p. 42036-43.

157. Arany, Z., et al., *A family of transcriptional adaptor proteins targeted by the E1A oncoprotein.* Nature, 1995. **374**(6517): p. 81-4.

158. Kasper, L.H., et al., *Conditional knockout mice reveal distinct functions for the global transcriptional coactivators CBP and p300 in T-cell development.* Mol Cell Biol, 2006. **26**(3): p. 789-809.

159. Sandelin, A., et al., *JASPAR: an open-access database for eukaryotic transcription factor binding profiles.* Nucleic Acids Research, 2004. **32**: p. D91-D94.

160. Frey, B.J. and D. Dueck, *Clustering by passing messages between data points.* Science, 2007. **315**(5814): p. 972-6.

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

162. Morris, L., K.E. Allen, and N.B. La Thangue, *Regulation of E2F transcription by cyclin E-Cdk2 kinase mediated through p300/CBP co-activators.* Nat Cell Biol, 2000. **2**(4): p. 232-9.

163. Billon, N., et al., *Cooperation of Sp1 and p300 in the induction of the CDK inhibitor p21WAF1/CIP1 during NGF-mediated neuronal differentiation.* Oncogene, 1999. **18**(18): p. 2872-82.

164. Braganca, J., et al., *Physical and functional interactions among AP-2 transcription factors, p300/CREB-binding protein, and CITED2.* J Biol Chem, 2003. **278**(18): p. 16021-9.

165. Jump, D.B., et al., *Fatty acid regulation of hepatic gene transcription.* J Nutr, 2005. **135**(11): p. 2503-6.

166. Watt, A.J., W.D. Garrison, and S.A. Duncan, *HNF4: a central regulator of hepatocyte differentiation and function.* Hepatology, 2003. **37**(6): p. 1249-53.

167. Sugitani, Y., et al., *Brn-1 and Brn-2 share crucial roles in the production and positioning of mouse neocortical neurons.* Genes Dev, 2002. **16**(14): p. 1760-5.

168. Zhang, D., et al., *Identification of potential target genes for RFX4_v3, a transcription factor critical for brain development.* J Neurochem, 2006. **98**(3): p. 860-75.

169. McFadden, D.G., et al., *Misexpression of dHAND induces ectopic digits in the developing limb bud in the absence of direct DNA binding.* Development, 2002. **129**(13): p. 3077-88.

170. Schrem, H., J. Klempnauer, and J. Borlak, *Liver-enriched transcription factors in liver function and development. Part II: the C/EBPs and D site-binding protein in cell cycle control, carcinogenesis, circadian gene regulation, liver regeneration, apoptosis, and liver-specific gene regulation.* Pharmacol Rev, 2004. **56**(2): p. 291-330.

171. Mayr, B. and M. Montminy, *Transcriptional regulation by the phosphorylation-dependent factor CREB.* Nat Rev Mol Cell Biol, 2001. **2**(8): p. 599-609.

172. Odom, D.T., et al., *Tissue-specific transcriptional regulation has diverged significantly between human and mouse.* Nat Genet, 2007. **39**(6): p. 730-2.

173. Rausa, F.M., Y. Tan, and R.H. Costa, *Association between hepatocyte nuclear factor 6 (HNF-6) and FoxA2 DNA binding domains stimulates FoxA2 transcriptional activity but inhibits HNF-6 DNA binding.* Mol Cell Biol, 2003. **23**(2): p. 437-49.

174. Nielsen, R., et al., *Genome-wide profiling of PPARgamma:RXR and RNA polymerase II occupancy reveals temporal activation of distinct metabolic*

pathways and changes in RXR dimer composition during adipogenesis. Genes Dev, 2008. **22**(21): p. 2953-67.

175. Nobrega, M.A., et al., *Scanning human gene deserts for long-range enhancers.* Science, 2003. **302**(5644): p. 413.

176. Loots, G.G., et al., *Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons.* Science, 2000. **288**(5463): p. 136-40.

177. Ross, E.D., A.M. Keating, and L.J. Maher, 3rd, *DNA constraints on transcription activation in vitro.* J Mol Biol, 2000. **297**(2): p. 321-34.

178. Eilers, P.H.C. and B.D. Marx, *Flexible smoothing with B-splines and penalties.* Statistical Science, 1996. **11**(2): p. 89-102.

179. Nerenz, R.D., M.L. Martowicz, and J.W. Pike, *An enhancer 20 kilobases upstream of the human receptor activator of nuclear factor-kappaB ligand gene mediates dominant activation by 1,25-dihydroxyvitamin D3.* Mol Endocrinol, 2008. **22**(5): p. 1044-56.

180. Yeamans, C., et al., *C/EBPalpha binds and activates the PU.1 distal enhancer to induce monocyte lineage commitment.* Blood, 2007. **110**(9): p. 3136-42.

181. Bammler, T., et al., *Standardizing global gene expression analysis between laboratories and across platforms.* Nat Methods, 2005. **2**(5): p. 351-6.

182. Petersen, D., et al., *Three microarray platforms: an analysis of their concordance in profiling gene expression.* BMC Genomics, 2005. **6**(1): p. 63.

183. Motta, M.C., et al., *Mammalian SIRT1 represses forkhead transcription factors.* Cell, 2004. **116**(4): p. 551-63.