Modeling Multi-factor Binding of the Genome

by

Todd Steven Wasson

Program in Computational Biology and Bioinformatics
Duke University

Date: _____

Approved:

_____
Alexander J. Hartemink, Supervisor

_____
Sayan Mukherjee

_____
Daniel Lew

_____
Mike West

_____
Jeanette McCarthy

Dissertation submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
in the Program in Computational Biology and Bioinformatics
in the Graduate School of Duke University
2010

Abstract
(Bioinformatics)

Modeling Multi-factor Binding of the Genome

by

Todd Steven Wasson

Program in Computational Biology and Bioinformatics
Duke University

Date: _____

Approved:

_____
Alexander J. Hartemink, Supervisor

_____
Sayan Mukherjee

_____
Daniel Lew

_____
Mike West

_____
Jeanette McCarthy

An abstract of a dissertation submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
in the Program in Computational Biology and Bioinformatics
in the Graduate School of Duke University
2010

Abstract

Hundreds of different factors adorn the eukaryotic genome, binding to it in large number. These DNA binding factors (DBFs) include nucleosomes, transcription factors (TFs), and other proteins and protein complexes, such as the origin recognition complex (ORC). DBFs compete with one another for binding along the genome, yet many current models of genome binding do not consider different types of DBFs together simultaneously. Additionally, binding is a stochastic process that results in a continuum of binding probabilities at any position along the genome, but many current models tend to consider positions as being either binding sites or not.

Here, we present a model that allows a multitude of DBFs, each at different concentrations, to compete with one another for binding sites along the genome. The result is an 'occupancy profile', a probabilistic description of the DNA occupancy of each factor at each position. We implement our model efficiently as the software package COMPETE. We demonstrate genome-wide and at specific loci how modeling nucleosome binding alters TF binding, and vice versa, and illustrate how factor concentration influences binding occupancy. Binding cooperativity between nearby TFs arises implicitly via mutual competition with nucleosomes. Our method applies not only to TFs, but also recapitulates known occupancy profiles of a well-studied replication origin with and without ORC binding.

We then develop a statistical framework for tuning our model concentrations to further improve its predictions. Importantly, this tuning optimizes with respect

to actual biological data. We take steps to ensure that our tuned parameters are biologically plausible.

Finally, we discuss novel extensions and applications of our model, suggesting next steps in its development and deployment.

To Ana Paula, to whom I was married a mere 15 days at the time of my defense.

To our journey together.

To my family, without whom I would never have come as far.

Contents

List of Tables

List of Figures

List of Abbreviations and Definitions

ChIP-chip	A microarray specifically used for determining the location along DNA of bound DBFs
Decoding	A probabilistic deconstruction of the likely paths through the state space of a hidden Markov model given an input sequence
DBF	See *DNA binding factor*
DNA	Deoxyribonucleic acid
DNA binding factor	A protein or protein complex that physically binds to DNA
Expression	See *Gene expression*
Gene expression	The transcription of a gene into its resulting mRNA; alternatively, the level of the abundance of that mRNA
Gene regulatory function	A function that maps transcription factor concentration to gene expression output
GRF	See *Gene regulatory function*
Hidden Markov model	A probabilistic framework for representing spatio-temporal series data as the output of a Markovian path through a hidden set of unobservable states
HMM	See *Hidden Markov model*
M-H	See *Metropolis-Hastings algorithm*
Markov chain Monte Carlo	A technique for sampling from a complex probability distribution

MCMC	See *Markov chain Monte Carlo*
Metropolis-Hastings algorithm	A specific type of MCMC in which the desired probability distribution is explored via another probability distribution from which sampling is easier
Microarray	A tool for measuring the abundance of many different mRNA or DNA sequences simultaneously
mRNA	Messenger RNA
Occupancy profile	A probabilistic representation of the landscape of binding probabilities across a given DNA sequence
ORC	The Origin Recognition Complex, a key component in initiation DNA replication
Position specific scoring matrix	A description of (usually) TF binding preferences in terms of the probability of binding each possible nucleotide at each possible position in the binding pocket
Posterior decoding	A specific type of decoding depicting the posterior-probability-averaged paths through the state space of an HMM, giving a view of distributions of paths (as opposed to just the best path, for example)
PSSM	See *Position specific scoring matrix*
RNA	Ribonucleic acid
TF	See *Transcription factor*
Transcription factor	A DBF that usually binds in gene promoters to play a role, either directly or indirectly, in transcriptional initiation

Acknowledgments

To a first approximation, my experience at Duke can be divided between the direct pursuit of my scholastic education and my life in the university and Durham community apart from my scholarly endeavors. It must be stated out of the gate that I make this distinction with more than a little hesitation, however, because among the great pleasures of my six years at Duke has been the delightful extent to which these parts of my life blended into one another, to form one cohesive growing, learning, and bonding experience.

This said, regarding my formal pursuit of my degree, I must first and foremost thank my advisor Alex Hartemink, without whom my graduate education would have been far less fulfilling, comprehensive, and enjoyable. Indeed, without Alex I may not have ever attended Duke: while exploring potential schools I first experienced his generosity via a lengthy telephone call during which he helped me see that Duke was a place I could pursue my wide-ranging interests with few restraints. From that time when I was an unknown voice on the telephone until my successful defense of my Ph.D. and beyond, he has been patient and thoughtful. It is a commonly passed-around axiom among graduate students that for whom you work is as or more important than on what you work, and I have been exceedingly fortunate on both accounts.

Additionally, thanks to my committee members Steve Haase, Danny Lew, Sayan Mukherjee, Mike West, and Jeanette McCarthy, who stepped in at the last minute to

replace Steve for my defense, allowing me to successfully satisfy a number of difficult constraints. In particular, thanks to Mike for his excellent course in Probability and Statistical Models, which had much to do with my pursuit of advanced techniques in statistics, and his course notes from which I have borrowed for my brief overview of MCMC and Metropolis-Hastings in this dissertation. And of course, I would be remiss if I did not especially thank Sayan for doing his utmost to blur the boundaries between mentorship and friendship, on the kickball field, at parties, or over a beer.

Although my graduate experience has been at various times trying and rewarding, it has above all been richly fulfilling, and the greatest portion of this fulfillment has come from my life within the community of people that I have come to know at Duke. This community has helped me grow as a person and a scientist, and I have had the good fortune to have numerous brilliant researchers become my close friends. I am no longer certain of what my expectations of my fellow students were upon first arriving in Durham, but it is safe to say that they were surpassed in every way. It has been my great pleasure to spend my time with these dynamic individuals in circumstances as diverse as camping out for basketball tickets and attending countless games in Cameron Indoor Stadium, playing numerous sports, going on road trips, barbecuing in the park, camping in the mountains, and relaxing at the beach. I will resist the urge to give a list of these people here because it would consist of dozens of names. Suffice it to say that you all know who you are, and you have all left indelible marks on my life that I will always cherish.

More than anything else, I will remember Durham as the place that I met my wife, and Duke as the institution that brought us together and in whose gardens we had our unforgettable wedding in the heart of our beloved community. Duke and Durham have given me more than I ever could have hoped for, and more than I can ever repay. I have fallen in love with Durham and the Duke community, and my life has forever been changed for the better by both. This community has given

me lifelong friends, wonderful memories, and my beautiful wife, and I am endlessly grateful.

To everyone that I have had the pleasure of encountering during my life-changing tenure at Duke, I sincerely thank you.

1

Introduction

Those of us living now have been fortunate enough to have been born in what may be the most exciting time to be an intellectual being in this planet's history. The amount of information that has been (and continues to be) gathered and generated about our world and ourselves is unprecedented, and inconceivable to our ancestors. Perhaps even more staggering is the almost embarrassing facility with which we can access it, often at no charge or penalty to ourselves. This situation is both the cause and effect of our curiosity, our dedication to the scientific method, and our ability and desire to pose ever more complex questions.

However, as our ability to ask questions has expanded, we have come to better recognize the extent to which our understanding of our universe is defined as much by our ignorance as our knowledge. This is as true in biology as in any discipline, where our views of fundamental biological processes are forever being challenged, revised, or overturned altogether.

Our inherent curiosity and capability to ask nuanced questions has long ago gained a purchase in the study of our physical selves, and of the other forms of life that inhabit our world, but in the modern age the extent to which these questions

can be answered has exploded. With the genomic revolution has come a view of life's inner workings that was unimaginable mere decades ago.

From philosophy to religion to science, we have throughout our history looked to far-flung areas of thought to learn more about what makes us human; what differentiates us from non-sentient life. As more species' genomes have been sequenced, we had hoped to find answers to the fundamental questions about the source of our pretended uniqueness and differences amongst organisms. To our surprise, we have been confronted with genomes of remarkable similarity of DNA content; far from unique, the various forms of life we have investigated are stunningly similar.

Setting bombastic rhetoric aside, we face a fundamental and practical question: if so many organisms are so similar at the most basic levels, where, then, does differentiation come from? And does it arise amongst different organs and tissues in a multicellular organism in the same way that it arises between organisms of different species? To begin to address questions such as these, we turn to studying the interconnections within the web of genes' control over one another, and specifically how that control is exerted.

The work presented in the body of this dissertation is concerned with developing an understanding of the physical binding of DNA by various actors in the cell, which is necessary to further address the questions discussed above. This focus, and our particular way of thinking about it, arises as a consequence of a collection of factors. As such, this introduction tells the story from the ground up, arriving at the body of our work as a next logical step in the series of events and thoughts beginning with a discussion of gene regulation. This order of presentation continues throughout this dissertation: our work is the end of a logical train of thought, and our discussion flows in that direction.

1.1 Gene regulation in eukaryotes

Although it has been known for some time that DNA encodes all of the genetic information of an organism, large portions of the nature of that encoding and how that information is acted upon are still shrouded in mystery. We know that DNA is a massive molecule constructed of two paired strands of *nucleotides*, denoted as A, C, G, or T for adenine, cytosine, guanine, and thymine respectively, and that certain regions of DNA, known as *genes*, encode information that is subsequently used for the production of functional units within the cell (such as *proteins*, which may combine to form *protein complexes*). However, DNA is more than just the tape of some microscopic Turing machine; its information content is layered not just in the identity of the nucleotides that make it up (or, by extension, the genes within it), but in how it moves and folds in three-dimensional space, how it interacts with the slurry of molecules and complexes in which it is suspended, and how these interactions allow it to exchange information with itself indirectly through these third parties.

As more organisms' genomes have been sequenced, it has been increasingly observed that the DNA sequence of these genes often remains markedly similar even between distantly-related species. What accounts for much of the inter-species differences, it has been revealed, is network of genes' interactions with one another; the *gene regulatory network*.

The mechanisms by which DNA is interpreted and instructs the creation of the resultant functional biological components display a remarkable complexity, with *regulation*, or modulation of rate of completion of a particular facet of the processes involved, happening at virtually all steps along the way. In order to appreciate the various forms that regulation can take, it is instructive to briefly introduce the various steps in the process of a gene's activation and interpretation into a functional unit. For the sake of simplicity, this discussion will focus on the production of proteins.

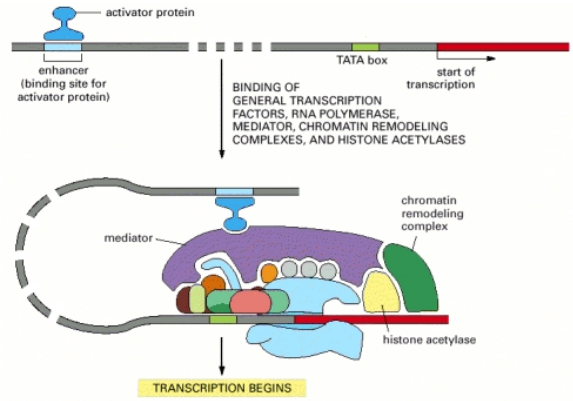

FIGURE 1.1: Transcription initiation in eukaryotes. Figure taken from Alberts et al. (2002).

1.1.1 The central dogma, then and now

"The central dogma" of molecular biology, which is in hindsight considered only a rough approximation to the process by which genes are used to produce proteins, was first proposed by Francis Crick in 1958 (Crick, 1958), five years after his publication of the structure of DNA with James Watson (Watson and Crick, 1953). This theory stated that genes in DNA encode RNA, which in turn encode proteins. While not necessarily incorrect at the broadest interpretation, it has been since revealed that this process involves many more steps including a vast amount of minutia not present in the original theory. A more modern view of the process by which DNA is used to produce the variety of gene products present in the cell, as well as some of the means by which it is regulated, is summarized below, with a particular focus on transcription, transcriptional regulation, and transcription factors.

Transcription

Transcription is the process by which genes are interpreted and used to generate RNA. It is a nuanced process, but involves at its coarsest resolution three primary steps, known as initiation, elongation, and termination.

Initiation is the step during transcription in which the majority of regulation takes place. Like transcription as a whole, it can be described in terms of three fundamental stages. First is the binding of an *RNA polymerase*, the molecule that interprets the DNA of the coding region of the gene and constructs the resultant RNA (or *transcript*), to the DNA in the *promoter* region of the gene (specifically, at the *transcription start site*), a region of DNA found upstream (toward the *5'* end of the DNA molecule; as opposed to downstream, or toward the *3'* end) of the beginning of the *coding region* (which encodes the gene product, i.e. protein) of the gene itself. To facilitate this binding, the *core promoter* of the gene, located within the promoter, must be first bound by a set of proteins called *general transcription factors*. The general transcription factors are required to recruit an RNA polymerase, forming (along with some other proteins and complexes) the *preinitiation complex*. Along with the general transcription factors, gene-specific *transcription factors* (TFs) can play a key role in initiation, via binding of the DNA in the promoter at *transcription factor binding sites* (Jacob and Monod, 1961). The relationship between TFs and rate of transcription has been heavily investigated, and understanding of this relationship is a fundamental driving force for this dissertation, as discussed later.

The remainder of initiation is composed of the opening, or "melting", of the DNA double helix (as a gene is encoded on a single strand, and an RNA polymerase traverses only one strand at a time), and the escape of the polymerase molecule from the preinitiation complex as it begins transcribing the gene and producing the resultant mRNA.

Upon a successful initiation, the polymerase escapes the promoter region, separating from the general transcription factors, and proceeds into the next step of transcription, known as elongation, during which the RNA molecule produced by the RNA polymerase is extended one nucleotide at a time until a complementary copy of the entire gene has been constructed. Lastly, the transcript is cleaved and a

sequence of adenines placed at the end in the final step of transcription, known as termination.

Post-transcriptional regulation

After an RNA transcript is produced, the transcript itself is subject to regulation. Among other steps, the transcript may potentially be *spliced*, and must be *exported* from the nucleus into the cytoplasm. During splicing, portions of the RNA transcript are selectively removed. The method by which splicing occurs is somewhat similar to transcriptional regulation by transcription factors: it involves physical binding of the RNA by RNA-binding proteins, differing permutations of which can result in differing sites for the 5' (*donor*) site or the 3' (*acceptor*) site, and hence, in differing splicing behaviors. During nuclear export, differing protein chaperones guide the RNAs into the cytoplasm, and can localize them to specific physical locations in the cell. The RNA is then subject to various other phenomena, including sequestration for future use or even destruction. Additionally, throughout its existence an mRNA may be targeted by *microRNAs*, which physically bind mRNA and can lead to various events including destruction of the mRNA, or *gene silencing* via the prevention of translation (Bartel, 2004, 2009). See Prasanth and Spector (2007) for a review of the known roles of various non-coding RNA.

Translation, translational regulation, and post-translational regulation

After transport to the cytoplasm, RNA can finally be *translated*, or used to guide the construction of a protein. Translation is carried out by *ribosomes*, whose recruitment can be a point of further regulation. Once the protein product has been constructed, it is now subject to direct regulation, by means such as *phosphorylation* or *dephosphorylation* (addition or subtraction of phosphate groups, leading to conformational changes and hence alteration of function) or other reversible activities, or targeted

destruction of the protein via *proteolysis*.

1.1.2 A focus on transcriptional regulation and transcription factors

Transcriptional regulation is the most-studied area of gene regulation, for a variety of reasons. Standing large among them is the allure of the *transcriptional regulatory code*, the set of rules by which the combinatorial permutations of transcription factor binding elicit a given effect on the rate of transcription of a gene.

As mentioned briefly above, transcriptional regulation happens primarily during initiation, and largely via the presence (or absence) of DNA binding proteins called *transcription factors* (TFs). However, unlike the general transcription factors, which are required for all initiation, different permutations of TFs are necessary at different promoters and can have differing effects on the resultant rates of transcription. Transcription factors have energetic preferences for (and against) certain sequences of nucleotides, and these preferences help guide them to their binding sites along the genome. TFs can occur as either single proteins or as part of protein complexes, and generally bind DNA mutually exclusively of other DNA binding entities, although, as seems to forever be the case in biology, there are exceptions.

To a first approximation, TFs can act as either *activators*, which increase the rate of transcription, and can be necessary for any significant levels of transcription at all in some genes, or *repressors*, which reduce the rate of transcription, including virtually eliminating it in some genes. Activators increase transcription via mechanisms such as direct binding to DNA and subsequent recruitment and/or activation of RNA polymerase, recruitment of other proteins to activate polymerase, protein modification, or DNA conformational changes (Ptashne and Gann, 2002). Repressors decrease transcription via mechanisms such as thermodynamic competition with activators, steric hindrance of activator and/or polymerase binding, activator destruction, and gene silencing via chromatin modification. Transcription factors, being the product

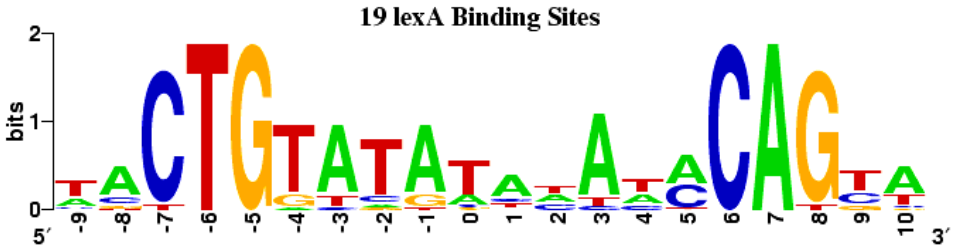

FIGURE 1.2: DNA sequence logo (Robison et al., 1998; Crooks et al., 2004).

Table 1.1: Example position specific scoring matrix. To score an example sequence "GACCTTCA", the product is calculated for each cell in the corresponding column of the PSSM. In this case, $0.70 \cdot 0.80 \cdot 0.25 \cdot 0.25 \cdot 0.40 \cdot 0.80 \cdot 0.10 \cdot 0.40 = 0.000448$.

	1	2	3	4	5	6	7	8
A	0.10	0.80	0.25	0.25	0.40	0.03	0.40	0.40
C	0.10	0.05	0.25	0.25	0.10	0.07	0.10	0.10
G	0.70	0.01	0.25	0.25	0.10	0.10	0.10	0.10
T	0.10	0.04	0.25	0.25	0.40	0.80	0.40	0.40

of genes, are themselves also regulated, creating feedback loops and vastly increasing the complexity of the gene regulatory network.

Transcription factor binding representations

Transcription factor binding is a probabilistic process, suggesting sequence preferences that maybe be more or less strong but are rarely binary. That is, TFs will generally bind virtually any DNA sequence with some probability; that probability may just be quite small. TF binding preferences are routinely called binding *motifs*, and are commonly given via *position-specific scoring matrices*, or *PSSMs*, which are capable of representing the probabilistic behavior of TF binding.

A PSSM is simply a matrix with four rows, corresponding to A, C, G, or T, and n columns, where n is the number of nucleotides commonly bound by the TF (Table 1.1). This number of nucleotides is an expression of the size of the binding

pocket of the TF protein, but is usually determined indirectly in the PSSM construction process (Bulyk, 2004). The cells of the PSSM contain real numbers between 0 and 1 that reflect the frequency with which each nucleotide is observed in each position of experimentally observed DNA binding events. In order to "score" a sequence S of N nucleotides as a prospective binding site for a given TF described by PSSM p, the probability of S being a binding site is calculated via the product of the terms in the cells in each column 1 through N corresponding to the nucleotide given in the sequence at that position. Explicitly,

$$P(S|p) = \prod_{i=1}^{N} p^i_{s_i} \tag{1.1}$$

where s_i is the ith nucleotide of S and $p^i_{s_i}$ is the probability of nucleotide s_i in column i of the PSSM p. This implies a (generally unrealistic) assumption of independence between binding positions, but this inaccuracy has proven empirically in most cases to be a worthwhile trade-off for the simplification it provides.

In lieu of a complete PSSM, or occasionally alongside it, shorter TF binding preference representations are often used. A visual representation of both the PSSM relative nucleotide per-position frequencies is often depicted via a *sequence logo* (Figure 1.2). Each column of a sequence logo consists of a stack of the letters A, C, G, and T, scaled by their respective coefficients from the corresponding PSSM cell. A sequence logo conveys an additional piece of information, in that the information content of each column of the PSSM is used to scale the height of the stack of letters (often with columns below some information threshold left empty). A shorthand alternative to the sequence logo and PSSM is the *consensus sequence*, or the sequence of the highest-probability nucleotides per-position in the binding motif. This is frequently augmented at positions of degeneracy using IUPAC codes such as V, indicating a position is usually occupied by *not* T (i.e., the position is A, C, or G).

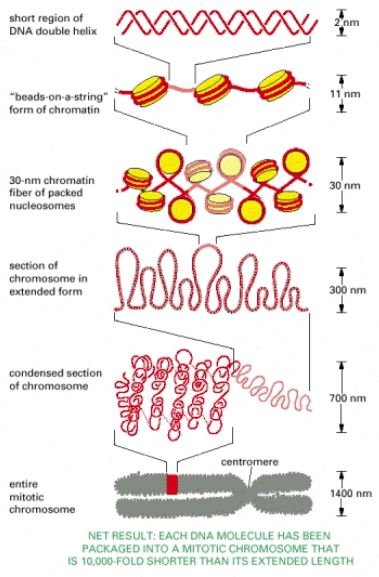

FIGURE 1.3: Chromatin packing. Figure taken from Alberts et al. (2002).

Among the most common ways to determine TF sequence preference are using experimental techniques known as ChIP-chip microarrays and Protein Binding Microarrays, detailed in sections 1.5.2 and 1.5.3 respectively.

1.2 Chromatin and its effects on transcriptional regulation

Gene regulation, and transcriptional regulation in particular, is subject to practical implications of spatial considerations within the cell. In order to fit into the nucleus, DNA must necessarily be densely compacted. This compaction takes place at a hierarchy of levels, as illustrated in Figure 1.3. The initial level of compaction is called a *nucleosome*, and involves about 150 base pairs of DNA coiling approximately 1.7 times around a histone core, a pair of protein tetramers, like thread wrapping around a spool. The compacted combination of DNA and nucleosomes is known as *chromatin*, and it plays a central role in transcriptional regulation via restriction of the accessibility of DNA to transcription factors and other DNA binding proteins and protein complexes, as indicated by the fact that nucleosomes occupy

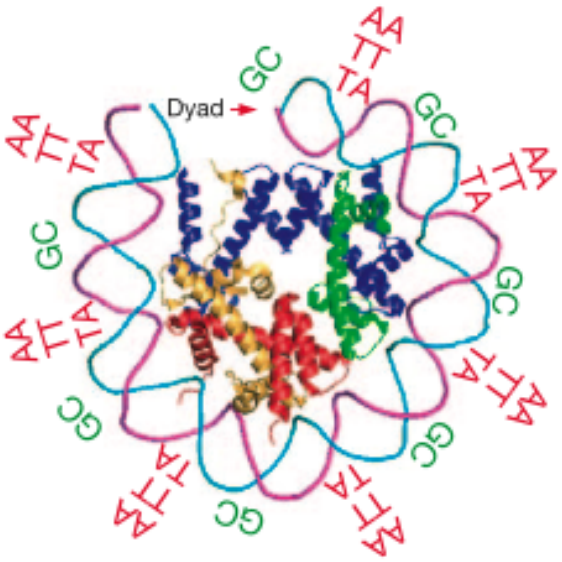

FIGURE 1.4: DNA dinucleotide positions around a nucleosome histone core. Figure taken from Segal et al. (2006).

75–90% of the genome at any given time in various eukaryotes (Van Holde, 1989). However, nucleosome occupancy shows a wide range of behaviors suggestive of the link between nucleosome and transcriptional regulation. For example, nucleosomes are depleted genome-wide at active regulatory regions (Lee et al., 2004), nucleosome occupancy is known to change over the cell cycle (Hogan et al., 2006) and due to conditional perturbations (Lee et al., 2004), and nucleosome positioning is correlated with transcription rate (Shivaswamy et al., 2008).

Like transcription factors, nucleosomes are driven by sequence preferences to bind stretches of nucleotides with particular composition (Figure 1.4). However, nucleosomes' sequence preferences are of a fundamentally different nature than the majority of TFs'. Where transcription factors can have strong preferences for or against particular nucleotides at particular positions in a prospective binding site, the sequence that a nucleosome tends to occupy is best characterized by its physical flexibility; particularly, by its ability to bend in the tight coil necessary to wrap

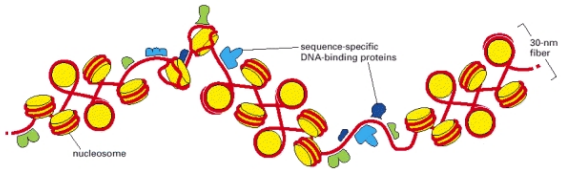

FIGURE 1.5: Nucleosomes and TFs coexist on DNA, but generally mutually exclusively at any single location. Figure taken from Alberts et al. (2002).

around the histone core.

Additionally, DNA bound by nucleosomes is generally prevented from simultaneously being bound by TFs, although there are certain TFs for which this is not the case (Alberts et al., 2002). For the majority of TFs, however, nucleosomes and TFs bind exclusively of one another, and as such are effectively in competition for binding of DNA, as illustrated in Figure 1.5.

The role of chromatin in transcriptional regulation has drawn a high level of interest of late, with high resolution genome-wide nucleosome maps being produced for a variety of organisms (Yuan et al., 2005; Shivaswamy et al., 2008; Mavrich et al., 2008) as well as focused studies of nucleosome behavior in promoters of particular genes (Lam et al., 2008) or gene families (Kharchenko et al., 2008).

1.3 Transcriptional regulation is driven largely by stochastic, competitive DNA binding

The physical binding of DNA by transcription factors, nucleosomes, and other DNA-binding proteins and protein complexes, collectively henceforth referred to as DNA binding factors (DBFs), is a common theme across the varied stages and mechanisms involved in enacting gene regulation. The number of types of DBFs bound to DNA at any given time reaches into the hundreds, and they demonstrate a remarkable diversity. While the genome is broadly bound by nucleosomes and is being traversed by DNA and RNA polymerases, transcription factors and other specialized proteins

and protein complexes bind very specific regions of the genome, often only at specific times or under specific conditions. The spatio-temporal interactions between these various DBFs and DNA, and between one another, regulate key cellular processes involving the genome, including replication, chromatin packaging, and transcription. Consequently, it is critical that we understand how these factors interact as they bind to the genome.

In particular, DBFs bind to the genome in competition with one another, jockeying for position along the DNA. To take one example, as discussed above, since transcription factors bind primarily exclusively of nucleosomes, TF binding sites depend not only on TF sequence specificity, but also on competition with nucleosomes (and other DBFs, including other TFs). In considering how to best represent the underlying biological phenomena that result in the DNA occupancy we observe in the cell, we need approaches that can model the interactions along the entire genome of large numbers of DBFs of varying kinds, including both nucleosomes and arbitrary numbers of TFs.

Additionally, the energetics of binding clearly indicate that DBF binding is a continuous phenomenon. That is, a DBF may bind anywhere in the genome, though more energetically favorable locations will be bound more frequently. A continuum view of DBF binding is consistent with a wealth of recent experimental data. Large-scale transcript sequencing studies reveal that many genes do not have a single well-defined transcription start site (TSS) but rather a distribution over TSSs (Miura et al., 2006), that many genes are not spliced using well-defined acceptor/donor sites but rather using a distribution over acceptor/donor sites (Chern et al., 2006), and that transcription is initiated not just at the TSSs of genes but in many places in the genome (Suzuki et al., 2001), leading to most of the genome being transcribed (Nagalakshmi et al., 2008; Johnson et al., 2005; Cheng, 2005; Carninci, 2005). In addition, large-scale ChIP studies reveal that RNA polymerases are located not only

upstream of genes, but in many other places in the genome (Steinmetz et al., 2006), and that TFs are bound not only to some promoters and not others, but rather in a continuum of detectable binding (Harbison et al., 2004). Computational studies have demonstrated that a continuum view of TF binding is useful in determining the probability of a promoter being bound somewhere by a single TF (Granek and Clarke, 2005), in building models of TF sequence specificity (Foat et al., 2006; Berger et al., 2006), and in understanding the contributions of weak binding to transcriptional regulation (Tanay, 2006; Segal et al., 2008).

This dichotomy of continuous evidence versus discrete explanation hints at a larger issue within current biological thought. For ease of explanation and instruction, all sorts of biological phenomena have been repeatedly described as if they were discretized, like steps in an algorithm. For example, pathways are depicted with the output of one step necessarily becoming the input of the next, and transcription is described with the stepwise assembly of the initiator complex leading necessarily to initiation. However it is important to note that, in plain terms, we know better. Just because these processes may require a particular ordering of events to lead to completion does not in turn mean that they are guided by some order-imposing force through the various steps. In reality, down to the smallest movements and interactions of individual molecules in the cell, the actions of biological systems are driven by randomness. This is not to say that this randomness is uniform; clearly there are peaks and valleys in the distributions of behaviors, but although simplifications may be of use to us in our modeling choices (as George Box famously said, "essentially, all models are wrong, but some are useful") and for pedagogical purposes, embracing and accounting for randomness and discounting unnecessary oversimplifications (like binary views of TF binding, or the very notion of a TF binding site itself) will yield an understanding of the underlying system that is more in line with biological reality.

1.4 The role of DNA binding factors in replication

Along with the above-mentioned gene regulatory processes, numerous other cellular processes are enacted by DNA binding. A prominent example of this is DNA replication, a critical component of a cell's reproductive process. During the duplication of a cell's genome, DNA replication occurs across all chromosomes at many different sites called *origins of replication* (Nasheuer et al., 2002). The first step of initiation replication at an origin is the binding of DNA by a protein complex called the *origin recognition complex*, or ORC. The regulatory mechanisms that control the timing of DNA replication are still not entirely understood, but a clearer view of the DNA binding behavior of ORC may lead to elucidating the nature of this regulation.

Like transcription factors and nucleosomes, ORC has a sequence preference and binds mutually exclusively of other DBFs. We can therefore consider it along with other DBFs, allowing investigation of the dynamics of the role of DNA binding in replication as well as transcription.

1.5 Experimental measurement techniques

A large factor in driving the exponential increases in the power of scientific inquiry in modern times has been our ability to actually address the nuanced questions that in the past would not have been answerable, leading invariably to new questions that would not have been askable. This is certainly true in molecular biology, where advances in measurement technology have allowed experiments to gain broad views of a multitude of processes in a cell simultaneously, to increase in speed and resolution, and to decrease in cost.

The primary experimental sources of data used in this dissertation are gene expression microarrays, ChIP-chip microarrays, and protein binding microarrays, and hence the following discussion of measurement techniques will focus on these tech-

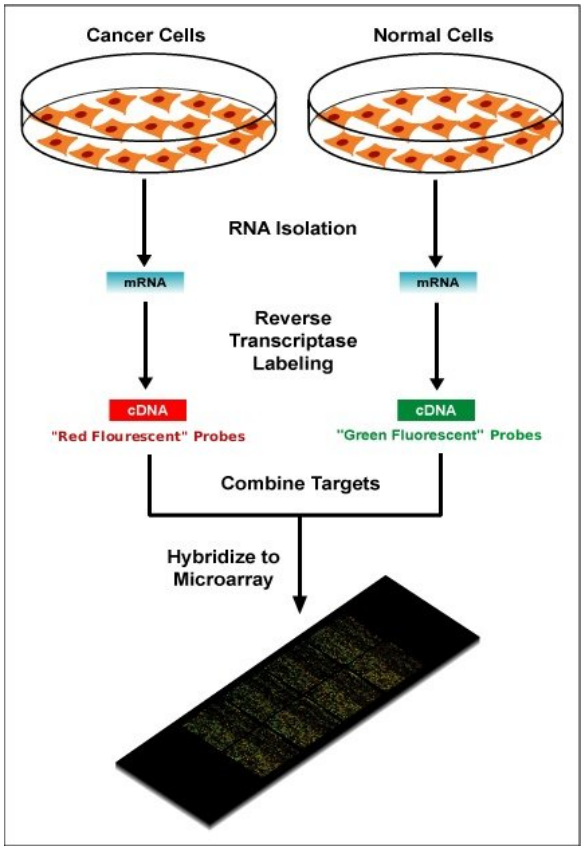

FIGURE 1.6: Gene expression microarray schema depicting competitive hybridization experiment. Although the two samples in this example are cancerous and normal cells, they may represent any two conditions.

nologies. While sequencing technology clearly has played a vital role in this work as well—without a genome to operate on our work would be limited to toy sequence data—for the sake of brevity it will not be discussed as a deep understanding of it does not inform our modeling choices, discussed at length in chapters 2 and 3.

1.5.1 Gene expression microarrays

Microarrays allow widespread measurement of the presence in a solution of thousands of different fragments of DNA simultaneously. Their most common use is to measure transcriptional expression of thousands of genes and gene variants by detecting the presence of different mRNA transcripts, and this is the specific use that will be

discussed here.

Microarrays are glass or silicone slides covered with grids of DNA fragments referred to as *probes*. The protocol to carry out a gene expression microarray experiment consists, at a high level, of the purification of mRNA from a solution whose mRNA content one would like to determine, generation of fluorescently-labeled cDNA from the mRNA via reverse transcription, and washing the cDNA-rich solution over the microarray itself.

In gene expression microarrays, the probes are specifically chosen to match mRNAs of interest. Ideally, this ensures that copies of the matching mRNA will *hybridize*, or bind, to the appropriate probe, and not to others. After hybridization, the spots on the microarray are excited with a laser, which causes the dyes on the cDNAs to fluoresce, and this fluorescence is measured as an indication of the quantity of cDNA hybridized at each probe.

There are two primary types of microarrays, differentiated by the construction of their probes. Spotted arrays are constructed by placing small portions of cDNA pools in a grid, where the cDNAs chosen are complementary to the mRNAs of interest. Spotted array facilities are commonly available at universities and research institutes, and, because the cDNA pools may be easily provided by an on-site researcher, they are quite customizable for specific tasks or experiments. This customizability, combined with relatively low costs, can make them an attractive option for experiments with relatively low requirements for sensitivity (Bammler et al., 2005).

Oligonucleotide arrays are higher sensitivity than spotted arrays, and are synthesized such that the probes have sequences pre-specified to a single-nucleotide resolution, and can match unique DNA sequences within the mRNAs (as opposed to entire mRNA strands) (Lockhart et al., 1996; Wodicka et al., 1997). This higher fidelity of control allows creation of probes of more or less specificity to a variety of mRNAs. Among the more commonly-used oligonucleotide arrays are those manufactured by

Affymetrix, which provide a variety of beneficial features including out-of-the-box chips available to measure expression of known genes for a variety of organisms.

Microarrays are typically used in either one-channel or two-channel experimental configurations. In one-channel experimental design, typically employed with oligonucleotide arrays like Affymetrix arrays, the cDNA sample washed across the chip contains one experimental condition dyed with a single fluorescent dye. In this way, the fluorescence value corresponds to some absolute value of gene expression. However, to do a comparative study of expression profile differences, more than one experiment must be performed using more than one chip (as each chip is single-use) in order to have a point of reference.

Two-channel experimental design involves a cDNA sample containing two experimental conditions (Figure 1.6), differentially tagged with fluorescent dyes (usually a red Cy5 and a green Cy3 dye) (DeRisi et al., 1997). The leads to *competitive hybridization* between the samples, in which the samples hybridize proportionally at each probe according to the quantity of mRNA transcripts present in each sample. The fluorescence of each dye is measured independently, and the two fluorescence levels are then combined after the fact for visualization purposes. Thus, probes will appear to fluoresce somewhere on a scale between red and green indicating which sample had more mRNAs for that probe, and yielding ratios of mRNA content between the samples. The actual level of mRNA content is lost in this approach, as only ratios are measured, but considering that many experiments serve to discern changing expression patterns between samples, measurement of ratios is often enough to suit the research purpose.

The inherent noisiness and experimental biases of microarray measurements have led to a thriving area of statistical research attempting to efficiently extract the information content present in the fluorescence measurements from the microarrays (Li and Wong, 2001; Irizarry et al., 2003), to the extent that papers attempting to

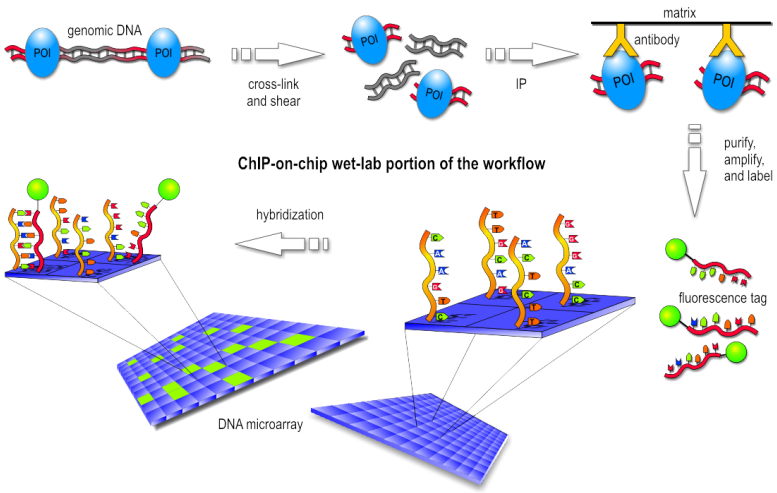

FIGURE 1.7: ChIP-chip experimental workflow.

disambiguate which method is "best" have received as much attention as the original papers themselves (Bolstad et al., 2003). It is commonplace presently to not use the original measurements from microarray experiments at all, but to instead use normalized, corrected statistics that attempt to account for biases within a single chip (or across chips), such as experimental error or effects from nonspecific or "background" hybridization.

1.5.2 ChIP-chip

Although microarrays are most commonly used for gene expression studies, their ability to detect the presence of any type of DNA has been applied to a variety of other uses. Prominent among these is the chromatin immunoprecipitation microarray, or *ChIP-chip* (Ren et al., 2000; Iyer et al., 2001; Lieb et al., 2001).

A ChIP-chip experiment allows measurement of the location of binding of DNA binding proteins or protein complexes, such as transcription factors, along the genome *in vivo* (Figure 1.7). To perform the experiment, a DBF of interest is first selected. A population of cells are exposed to a cross-linking agent, which binds any DBF on the genomic DNA at the time of introduction to the DNA molecule. The cells are then

lysed, releasing their DNA, which is bound by some collection of DBFs. The genomic DNA is sheared, and using antibodies specific to the target DBF, DNA-bound copies of the DBF of interest are extracted. The cross-linking is reversed, yielding a pool of DNA fragments that were bound by the DBF, and these fragments are amplified and washed over the microarray. Usually, the microarray has been prepared with probes that correspond to gene promoter regions genome-wide, such that the array may be used to discern genes in whose promoters the DBF of interest binds, and by extension for which genes the DBF plays a regulatory role.

Similarly to gene expression microarray studies, ChIP-chip experiments introduce statistical issues regarding interpretation of fluorescence as an indication of binding of the target DBF to the region corresponding to a given probe. For these experiments, the derived statistic is generally a p-value. Although a Bayes factor indicating the probability of binding would be preferable, p-values are used for historical reasons, though mapping from p-values to bounds on Bayes factors can be performed (Sellke et al., 2001).

As mentioned briefly in section 1.1.2, ChIP-chip experiments are often used to determine the sequence preferences for various DNA binding proteins and protein complexes, like transcription factors. These experiments are often carried out in multiple conditions, to help understand changes in TF binding behavior and the roles TFs play in the differentiation of regulatory behavior as a response to various states of the cell. A particularly widely-used and comprehensive example of such a study for the yeast *Saccharomyces cerevisiae* is detailed in Harbison et al. (2004) and MacIsaac et al. (2006).

1.5.3 Protein binding microarrays

Like ChIP-chip experiments, protein binding microarrays (PBMs) (Mukherjee et al., 2004; Berger et al., 2006) are often used to gain an understanding of the sequence

specificities of transcription factors. However, unlike the *in vivo* sequence preferences yielded by experiments involving TFs binding DNA in the cell, the preferences from PBM experiments give a view of *in vitro* specificities.

PBM experiments are carried out fundamentally differently from ChIP-chip in a variety of ways. A DNA binding protein of interest (or at least, a protein with the DNA binding domain of interest) is tagged, purified, and washed over a microarray that contains doubled-stranded DNA (as opposed to the single-stranded DNA on typical gene expression or ChIP-chip microarrays). Current PBM techniques generally employ microarrays with probes representing not the genome of a particular organism, but a "universal" set of DNA probes, with every possible DNA sequence of a given length represented. Using this library of disparate sequences, the inherent sequence specificity for the protein can be determined by its hybridization pattern on the microarray.

The *in vitro* specificities gained from PBM data, which are gathered, by definition, without the influence of the busy natural environment of the cell and hence free from influence of factors other than sequence preference, reveal much about the way that the cell functions. Within the cell, the location along the DNA of transcription factors, for example, is the result of a complex interplay between a collection of driving forces. While sequence preference certainly plays a role, *in vivo* locations are effectively the *output* of this system of forces, albeit an output indicative of sequence preferences; *in vitro* preferences are one of the *inputs*. This idea dovetails intellectually with the driving ideas behind this dissertation: understanding complex binding phenomena as a function of a set of simpler rules and inputs.

1.6 Outline of this dissertation

This dissertation has two primary goals: first, to elucidate the nature of DNA binding and its role in transcriptional regulation, and second, to promote a continuous,

stochastic view of biological phenomena as a whole, and to do this through the development and introduction of ways of thinking about DNA binding in this light. To crystallize these thoughts into a tangible entity, a model of DNA is presented that takes DBF sequence specificities and concentrations, pits them in competition with one another for occupancy of the genome of *Saccharomyces cerevisiae*, and evaluates the resultant binding predictions in a probabilistic sense.

Chapters 2 and 3 present and expound upon ideas appearing in Wasson and Hartemink (2009). In chapter 2, we present the theoretical underpinnings of our approach as a consequence of the nature of DNA binding as described above, which we refer to as the extended 'musical chairs' contest view of genome binding. We show that our approach has characteristics inherent to it that lends itself well to being modeled by particular approaches from statistical mechanics. Finally, we give an introduction to Boltzmann chains via hidden Markov models, show their relationship, and give the algorithmic and mathematical depiction of our model.

Chapter 3 takes the ideas presented in chapter 2 and presents an implementation in efficient software. Our model is then applied to a variety of scenarios in yeast, showing desirable traits with respect to observed biological behaviors. Furthermore our model is shown to be an improvement on simpler techniques, both on a localized small scale, and on a whole-genome large scale.

We describe a Markov chain Monte Carlo method to tune the concentration parameters in chapter 4. Our technique increases the agreement between transcription factor binding predictions by our model and binding data from ChIP-chip experiments. We then extend the technique to constrain predictions to biologically plausible values by restricting extremely high or low occupancy of the genome by nucleosomes in conjunction with the tuned parameters.

In chapter 5 we propose several avenues for improving upon our model, as well as potential next applications of the model to other problems. Regarding improvements,

we discuss alterations of the theory underlying our model including alternative TF binding representations, introducing cooperativity among factors, and accounting for higher-order chromatin compaction. We also discuss practical issues such as the benefits of restructuring the model in terms of Hierarchical HMMs and other related frameworks. Finally, we briefly discuss the need for developing a link between the binding predictions produced by our model and rates of transcription initiation as a function of bound TFs.

2

Theoretical motivation for, and construction of, our model

2.1 Desired traits in a DNA occupancy model

Given our understanding of how biological systems function, and how DNA binding in particular arises as a consequence of a multitude of forces, a reasonable model of DNA binding must demonstrate, at a minimum, the following characteristics:

- Incorporation of multiple factors together
- Accounting for effects of DBFs on one another, including competition and cooperativity
- Consideration of all possible binding configurations
- Continuous description of the likelihood of all possible configurations, even if extremely unlikely

Yet, few current models of TF binding take competition with nucleosomes and other TFs into account, and few current models of nucleosome binding take competition with TFs into account. Current models that do consider nucleosomes and TFs

together suffer from various drawbacks, such as being restricted to small genomic regions or coarse resolution (Teif, 2007), or a lack of genome-wide improvement of positioning as a result of this incorporation (Morozov et al., 2008). Although a few models of TF binding consider a small number of TFs at once at single-nucleotide resolution (Sinha, 2006; Segal et al., 2008), these do not consider nucleosomes. We need approaches that can model the interactions along the entire genome of large numbers of DBFs of varying kinds, including both nucleosomes and arbitrary numbers of TFs.

A separate problem of current models is that most consider DBF binding to be a discrete, or binary, phenomenon. This leads to a view in which certain genomic positions are annotated as DBF binding sites, while the rest are assumed to be unbound. But the energetics of binding clearly indicates that DBF binding is a continuous phenomenon. This leads to a contrasting view in which genomic positions are not annotated as binding sites using the set $\{0, 1\}$ but rather the interval $[0, 1]$. Under such a view, a DBF may bind anywhere in the genome, but more energetically favorable locations will be bound more frequently. Furthermore, the frequency of binding at any particular location will also depend on DBF concentration, a point ignored by nearly every model (some notable exceptions being Djordjevic et al. (2003), Bintu et al. (2005a), and Segal et al. (2008)).

In summary, we need models of continuous genome binding that take into account competition by many DBFs of various kinds and at varying concentrations, and are capable of doing so at genome-wide scales and single nucleotide resolutions. In the following sections, we present a first such model. Our model takes as input the sequence specificities and concentrations of multiple kinds of DBFs, along with a genome sequence; it then computes as output the probability that each position in the genome will be occupied by each kind of DBF when these factors are in competition with one another for binding locations. This 'occupancy profile' for each

DBF across the genome is computed as the thermodynamic ensemble average over all valid binding configurations, each occurring with a different frequency in accordance with a Boltzmann distribution. Later, we presented our model implemented in a software package called COMPETE (Competitive Occupancy: Multi-factor Profile Evaluation of a Thermodynamic Ensemble). COMPETE is computationally efficient: entire yeast chromosomes with hundreds of different kinds of DBFs can be processed in minutes. It is flexible and extensible, allowing incorporation of any DBF with sequence-specific binding preferences. We demonstrate this by considering simultaneously the binding of TFs, nucleosomes, and the origin recognition complex across the actual sequence of the yeast genome. Our model represents a significant early step toward the goal of understanding, at a mechanistic level, the interactions between factors regulating different genomic processes like replication, transcription, and chromatin packaging. As a first step toward layout out our model framework, it is necessary to embark on an introduction of statistical mechanics, and how it presents a language from which the formulation of our model flows naturally.

2.2 Statistical mechanics, and its application to DNA binding

Statistical mechanics is a reductionist field of study involving attempting to understand a thermodynamic system as a function of its constituent parts, specifically via statistical methods. The roots of statistical mechanics lie in the kinetic theory of gases (Bloch and Walecka, 2001), but the area has quickly grown to applications in a number of fields, including biology. Generally, problems that can be phrased such that the goal is understanding the mechanics of macroscopic systems with many degrees of freedom, where the components of the systems are individual actors with known behaviors and the systems may adopt known configurations (even if these configurations are innumerable) lend themselves to techniques from statistical mechanics.

Statistical mechanics is a broad field capable of addressing many types of problems, each with its own intricacies and twists. In the interest of maintaining focus, the discussion here will focus on the facets of thermodynamics either directly relevant to an application to our DNA binding problem, or on those topics necessary to frame the relevant issues.

2.2.1 Statistical ensembles

In statistical mechanics, a fundamental tenet is that the possible states of a given system can completely described, and the energy of the system can be evaluated at each of the possible states. The states (or configurations) of the system are called *microstates*, and the system adopts each of the microstates with a probability proportional to the energy of the state. The complete set of microstates for a given system is called the *statistical ensemble*, or just *ensemble* (Stowe, 2007; Halley, 2007). Although the term "ensemble" is heavily overloaded in the literature of various disciplines, in this context it refers to the complete set of possible microstates of a thermodynamic system. The connotation here, therefore, is similar to a *probability space*.

Ensembles are usually described with a number of different variables, including a system temperature T, total number of microstates N, volume of the system V, system energy E_i at each microstate i, and internal energy U. Given these values and the selection of a statistical ensemble, thermodynamic values of note can be calculated such as entropy and system free energy, as well as the *partition function* for the ensemble, necessary for transforming E_i into a probability.

There exist a large number of ensembles, describing various system configurations, but most are framed in terms of the three "canonical" ensembles: the *microcanonical ensemble*, the *canonical ensemble*, and the *grand canonical ensemble*. These three ensembles are differentiated from one another largely by which of the above values

are fixed and which are not. Importantly, all three of the canonical ensembles describe systems *at equilibrium*; this has consequences for our system, discussed in section 2.2.1.

The *microcanonical ensemble* describes a system in which N, V, and E are fixed. The system is therefore thermally isolated by definition, and each microstate is equiprobable. The *grand canonical ensemble* describes a system in which V and T are fixed, along with chemical potential. The system is used to model systems in which particles and energy can be freely exchanged with a given reservoir, allowing each microstate to differ in both of these quantities. The ensemble that we use in our model of DNA binding is the *canonical ensemble*, in which N, V, and T are all fixed, but the system is allowed to exchange energy with a reservoir, allowing each microstate a different energy E_i.

The canonical ensemble

In the canonical ensemble, the probability of the system adopting microstate i is expressed by:

$$p_i = \frac{1}{Z} e^{-E_i/(k_B T)} \tag{2.1}$$

where k_B is Boltzmann's constant and Z is the partition function,

$$Z = \sum e^{-E_i/(k_B T)} \tag{2.2}$$

For simplicity, let $\beta = \frac{1}{k_B T}$. Then Equation (2.1) and Equation (2.2) become:

$$p_i = \frac{1}{Z} e^{-\beta E_i} \tag{2.3}$$

and

$$Z = \sum e^{-\beta E_i} \tag{2.4}$$

The term $e^{-\beta E_i}$ is called the *Boltzmann factor*, and the resulting probability distribution described in Equation (2.3) and Equation (2.4) is called the *Boltzmann distribution.*

DNA occupancy couched in terms of the canonical ensemble

Consider a closed system at a fixed temperature, consisting of a molecule of DNA and various DNA binding factors, each at a given concentration but present in infinite supply, such that "usage" of a DBF does not alter its concentration. Assume that every DBF d_j (with a DNA binding site width of n_j) can bind every n_j-long stretch of sequence along the DNA molecule. Furthermore, assume that each position along the DNA can be either occupied by a DBF or unoccupied and open, but cannot be occupied by more than one DBF at a time. Therefore the DNA can exist in permutations ranging from completely naked and unoccupied to completely occupied by any combination of DBFs; we call these *configurations.* Finally, assume that given a system temperature T, an energy function exists to calculate E_i for each configuration. The system, described in this way, can be represented by the canonical ensemble, with configurations representing microstates. The critical link necessary to complete this ensemble representation is the energy function that evaluates E_i for each configuration i. This function is the method by which concentrations are introduced into the model, and is described in great detail later.

Although this view will demonstrate great utility going forward, along with the assumptions listed above, there are other assumptions implicit in the canonical ensemble that are relevant to the biological coherence of our model.

- *The system is always at equilibrium.* This implies at least two biologically implausible conditions: equilibrium exists such that the concentration of DBFs is always equal at all places along the DNA molecule, and this equilibrium is achieved instantaneously. Clearly the concentrations throughout the nucleus

are not equal, but they are, to a reasonable first approximation, at *local* equilibrium.

- *DBF-DNA binding is mutually exclusive of other DBFs.* While there are known cases of transcription factors binding opposite strands of DNA, as well as binding DNA that is already bound by nucleosomes, these cases are rare in comparison with the majority of binding, which is mutually exclusive.

- *DBFs are available at an infinite supply.* Clearly this is not the case in the cell.

- *All DBFs associate and disassociate with DNA instantaneously.* Although this activity is extremely rapid for transcription factors, it is not instantaneous. Moreover, association/disassociation likely does not even occur at the same rate for all TFs, and nucleosomes are certainly orders of magnitude slower.

Clearly it is important (and scientifically ethical) to recognize modeling shortcomings, and these assumptions are each more or less incorrect. However, this framework leads to a view of DNA binding very much in line with what we know of the biology as described in chapter 1, and the model demonstrates great utility, as demonstrated in chapter 3. Continuing forward, we will see how this system view meshes with our biological understanding, and how its modeling advantages quickly overwhelm its drawbacks; specifically, we will see that it satisfies the criteria put forth in section 2.1.

2.3 The extended 'musical chairs' contest view of genome binding

Couched in terms of the canonical ensemble, our view of the binding of DBFs along the genome can be likened to an extended 'musical chairs' contest. Imagine the eponymous children's game in which children circle a finite set of chairs while music plays. In the traditional game, when the music stops, the children hurry to sit in a

chair. Because there are one fewer chairs than children, one child is unseated and hence out of the game. This completes one round; at the end of the round a chair is removed, the next round starts, and the entire process is repeated until there is only one chair, and only one child seated in it, being declared the winner.

With a few adjustments and extensions, we envision DNA binding in a similar fashion. Consider the children to be different DBFs and the chairs to be positions along the genome, such that at the end of each round the genome is occupied by some valid configuration of DBFs. Further imagine that the children are grouped into teams (multiple copies of the same DBF) of different sizes (concentration levels); as a consequence, a team with more children (higher concentration) will occupy more chairs than a team with fewer children (lower concentration), all other things being equal. Finally imagine that no chair is removed after each round, allowing an extended contest to proceed indefinitely.

In this extended contest, we record how often each chair is occupied by each team over time, resulting in an *occupancy distribution* for each chair. By considering these occupancy distributions across all chairs, we obtain an aggregate summary for each team, which we call an *occupancy profile*. Our model is almost perfectly analogous, but is more complicated in that we assume each DBF occupies some number of consecutive nucleotides along the genome, and exhibits varying preferences for different locations in the genome according to its sequence specificity. Because of this last point, the various valid binding configurations will be sampled with different frequency, with more energetically favorable configurations being sampled more frequently.

This view of DNA binding possesses all of the traits described in section 2.1. Multiple DBFs are considered simultaneously in this approach for each genomic location, and DBF binding precludes binding of overlapping DBFs, resulting in direct competition between factors. Cooperativity is also realized through this model; the

mechanism through which this takes place is described later.

Because of the framing of this view within the canonical ensemble, all legal configurations are explicitly considered and weighted with Boltzmann factors. Given these weights, the probability of each configuration is calculable, to a single nucleotide resolution. That is, the probability of each position on DNA being occupied by any DBF (or none at all) can be calculated under the Boltzmann distribution. To realize this, we leverage a statistical framework called Boltzmann chains.

2.4 Boltzmann chains

As discussed previously, the key component in realizing our 'musical chairs' view of genome binding via a statistical mechanics canonical ensemble is the energy function E_i used to evaluate a given microstate, or configuration, i. We use a statistical mechanical framework called a *Boltzmann chain* for this purpose (Saul and Jordan, 1995). Boltzmann chains were originally intended to model discrete time series, but in our usage we model the genome sequence as a discrete series, spatially instead of temporally.

Boltzmann chains contain within them *hidden Markov models* (HMMs) as a special case. Because HMMs are more widely known, and because the mathematics for evaluating HMMs and Boltzmann chains is identical, we shift focus to discuss HMMs in detail. Our specific usage of HMMs requires alterations of the 'vanilla' HMM algorithm; this will be discussed in detail, followed by a return to Boltzmann chains, special considerations present with them as compared to HMMs, and a demonstration of their containing of HMMs as a special case.

2.4.1 Hidden Markov models

A brief introduction to HMMs

A process in which a sequential output is observed that depends on some sequentially evolving internal state that is unobserved may often be modeled by a statistical framework known as a *hidden Markov model*. An illustrative example of this kind of process, given in a standard textbook on biological sequence analysis Durbin et al. (1998), is entitled "The Occasionally Dishonest Casino." Imagine a casino in which a dice game is played. Two dice are cycled into use; one is fair, one is loaded. Each time a die is cast, the result produced depends on which die is in play. The fair die will produce uniformly distributed results, and the loaded die will produce skewed results. After the roll, the die in play is either kept in play or exchanged with the other die according to some given probability. The problem becomes: if one knew the probabilities of exchanging the dice and the probabilities of each die landing on each face, and were given only the list of results of the casting of the dice, how could the identity of the die in use for each roll be determined?

Techniques to calculate the probability of the sequence of rolls being generated by every possible combination of switching of the dice exist and have been in use since the late 1980s, though they were first developed decades earlier Rabiner (1989). Of course, this work applies to the general case where the concern is the state of the system at each position in the output sequence, corresponding to which die is in play for each roll in our example. Given this exhaustive set of probabilities of the putative "paths" of states through the sequence, that the highest probability path can be determined (called the *Viterbi decoding*, where a *decoding* is a probabilistic deconstruction of the likely paths through the sequence), and the probability of being in each state at each position can also be calculated (called the *posterior decoding*). As these probabilities can be calculated, the parameters of the model in

question can be manipulated in order to maximize these probabilities. A closed form of Expectation-Maximization (Dempster et al., 1977), known as the Baum-Welch algorithm, accomplishes this task (Baum, 1972).

The occasionally dishonest casino problem is similar in essence to one of placing DNA binding elements along a sequence. If each position on a sequence is considered to be either unbound or bound by an element, and both the unbound sequence and the binding element have a given nucleotide distribution or *motif*, the question of where the element is binding along the sequence is effectively the same as when the casino switched to the loaded die. Though the example given is simplistic and restricted to only two states, these models can be complex and may contain an arbitrarily large number of states. Because of this, our model can account for complexities like the incorporation of multiple binding motifs of different lengths and the fact that the energetics of nucleosome binding depends on dinucleotides instead of mononucleotides.

2.4.2 Mathematical definition of an HMM, and the Forward-Backward algorithm

In this section we define an HMM, using the notation from Durbin et al. (1998). We begin with this basic definition because, as will be demonstrated in later sections, the nature of our problem requires several alterations to the algorithm described here. By demonstrating these alterations as they impose necessary adjustments along the way, it will be clear why our final algorithm arises as a consequence. Additionally, the algorithms given here are not generally available in the wider literature; they were derived for our purposes, but hold true for any HMM (or Boltzmann chain).

An HMM is a type of *graphical model*, defined by a set of N states, probabilities of transitioning among states, an alphabet of symbols, and per-state probabilities of emitting each symbol. For a given sequence through the states, called a *path* π, call the ith state in the path π_i. Additionally, let x be the given input sequence, x_i be

the ith symbol (or character) of x, and L be the length of x. Then the set of state transition probabilities is:

$$a_{kl} = P(\pi_i = l | \pi_{i-1} = k) \tag{2.5}$$

and the set of symbol emission probabilities is:

$$e_k(b) = P(x_i = b | \pi_i = k) \tag{2.6}$$

The joint probability of a given output sequence and path through the states is:

$$P(x, \pi) = a_{0\pi_1} \prod_{i=1}^{L} e_{\pi_i}(x_i) a_{\pi_i \pi_{i+1}} \tag{2.7}$$

where $\pi_{L+1} = 0$.

We have a particular interest in calculating the posterior decoding, $P(\pi_i = k|x)$. To calculate either the Viterbi or posterior decoding, a dynamic programming algorithm called the Forward-Backward algorithm may be employed. Going forward, we describe this algorithm and focus only on the posterior decoding, and the Viterbi decoding is uninteresting in our domain.

The Forward-Backward algorithm involves calculating two tables: the forward table f and the backward table b. Both tables have N rows and L columns.

Forward algorithm:

$$\begin{gathered} \text{Initialization } (i = 0)\text{: } f_0(0) = 1, f_k(0) = 0 \text{ for } k > 0 \\ \text{Recursion } (i = 1 \ldots L)\text{: } f_l(i) = e_l(x_i) \sum_{k=1}^{N} f_k(i-1) a_{kl} \\ \text{Termination: } P(x) = \sum_{k}^{N} f_k(L) a_{k0} \end{gathered} \tag{2.8}$$

Backward algorithm:

$$\begin{aligned} &\text{Initialization } (i = L)\text{: } b_k(L) = a_{k0} \text{ for all } k \\ &\text{Recursion } (i = L-1, \ldots, 1)\text{: } b_k(i) = \sum_{l=1}^{N} a_{kl} e_l(x_{i+1}) b_l(i+1) \\ &\text{Termination: } P(x) = \sum_{l}^{N} a_{0l} e_l(x_1) b_l(1) \end{aligned} \tag{2.9}$$

Given the completed forward and backward tables, the posterior decoding becomes:

$$P(\pi_i = k|x) = \frac{f_k(i) b_k(i)}{P(x)} \tag{2.10}$$

2.4.3 Our algorithm for calculating the posterior decoding

The algorithm as described in Equation (2.8) through Equation (2.10) is effective for most applications, but we have several requirements that necessitate slight alterations to the method.

Silent states

Often times the structure of a highly interconnected graphical model can be simplified via the introduction of *silent states*. Silent states emit no symbols and thus allow state transitions without progressing along the input sequence. The benefits of simplification extend beyond the ease of understanding of a model's structure. Because the size of the forward and backward tables grows linearly in the number of states, reducing the number of states reduces memory requirements and increases speed of the algorithm's execution. The importance of silent states in our model will be made clear in section 3.1.

Numerical stability

When actually computing the values in the forward and backward tables, they quickly become vanishingly small. A common workaround to counteract this problem is to rewrite the forward and backward algorithms using logarithms and exponentiation. This technique is effective and simple to implement, yet computationally intensive. Another workaround is to scale each column of the forward table by its respective sum, normalizing the column entries to [0, 1]. The backward table is then normalized with the scaling constants determined from the forward table, placing its entries in the general neighborhood of [0, 1] as well. This is much higher performance and provides some mathematically convenient bonuses regarding calculation of the posterior decoding, but unfortunately using the same scaling factors for both the forward and backward tables mean that they must be computed in order. Therefore, we derive updating equations that break this dependence with separate scaling factors.

The resulting complete algorithm

Combining all of these alterations together yields the following version of the algorithm:

Enumerate the states such that all 'normal' states are numbered before all silent states. Furthermore, there can be no loops consisting of only silent states, and if there connections between silent states, they must go from lower-numbered states to higher-numbered states. Finally, silent states must be explicitly handled in the initialization, recursion, and termination steps. Define functions $U_f(i)$ and $U_b(i)$ to update silent states for column i, as described below.

Several new variables are necessary. The complete list is now:

$$\begin{aligned} x &: \text{input sequence vector} \\ L &: \text{length of } x \\ N_n &: \text{number of normal states} \\ N_s &: \text{number of silent states} \\ s_f &: \text{scaling vector for forward table} \\ s_b &: \text{scaling vector for backward table} \\ s_r &: \text{scaling ratio vector for posterior decoding} \end{aligned} \tag{2.11}$$

Forward algorithm:

$$\begin{aligned} &\text{Initialization } (i = 0): \ f_0(0) = 1, f_k(0) = 0 \text{ for } k > 0 \\ &\text{Recursion } (i = 1 \dots L): \ f_l(i) = e_l(x_i) \sum_{k=1}^{N_n} f_k(i-1) a_{kl} \\ &\text{Termination: } \ P(x) = \sum_{k}^{N_n} f_k(L) a_{k0} \end{aligned} \tag{2.12}$$

After updating each column i:

1. Update silent states via:
   (a) For all silent states l, set $f_l(i) = \sum_{k=1}^{N_n} f_k(i) a_{kl}$
   (b) Beginning with lowest numbered silent state $l = N_n + 1 \dots N_s$, set $f_l(i) = f_l(i) + \sum_{k=N_n+1}^{l-1} f_k(i) a_{kl}$
2. Calculate scaling factor: $s_{f_i} = \sum_{k=1}^{N_n} f_k(i)$
3. Apply scaling factor: For all states l, set $f_l(i) = f_l(i)/s_{f_i}$

Backward algorithm:

$$\begin{aligned} &\text{Initialization } (i = L)\text{: } b_k(L) = a_{k0} \text{ for all } k \\ &\text{Recursion } (i = L-1, \ldots, 1)\text{: } b_k(i) = \sum_{l=1}^{N_n} a_{kl} e_l(x_{i+1}) b_l(i+1) \\ &\text{Termination: } P(x) = \sum_{l}^{N_n} a_{0l} e_l(x_1) b_l(1) \end{aligned} \tag{2.13}$$

After updating each column i:

1. Update silent states via:
   (a) For all silent states k, set $b_k(i) = \sum_{l=1}^{N_n} a_{kl} e_l(x_i) b_l(i)$
   (b) Counting backwards beginning with highest numbered silent state $l = N_s \ldots N_n + 1$,
   set $b_l(i) = b_l(i) + \sum_{k=l}^{N_s} b_k(i) a_{lk}$
2. Calculate scaling factor: $s_{b_i} = \sum_{k=1}^{N_n} b_k(i)$
3. Apply scaling factor: For all states l, set $b_l(i) = b_l(i)/s_{b_i}$

Posterior decoding: The posterior decoding will now depend on s_f and s_b. For computational efficiency, we calculate s_r:

1. $s_{r_L} = 1$
2. For $i = L \ldots 2$, $s_{r_{i-1}} = s_{r_i} \frac{s_{b_i}}{s_{r_i}}$

The final posterior decoding now becomes:

$$P(\pi_i = k|x) = s_{b_i} s_{r_i} f_k(i) b_k(i) \tag{2.14}$$

2.4.4 Hidden Markov models as a subset of Boltzmann chains

Boltzmann chains relax the restriction that transition and emission weights must be probabilities. The weights, labeled A_{kl} for the transition weight from state k to state l and $B_k(b)$ for the emission weight of emitting symbol b in state k, can be any nonnegative real number.

Letting a microstate i of the ensemble modeled by a Boltzmann chain be defined by a path π and an input sequence x, the energy of the system in microstate i is:

$$E_i = -A_{0\pi_1} - \sum_{i=1}^{L} \left(A_{\pi_i \pi_{i+1}} + B_{\pi_i}(x_i) \right) \tag{2.15}$$

where $\pi_{L+1} = 0$.

The probability of each possible microstate may now be calculated via Equation (2.3), and the partition function may be calculated via Equation (2.4).

For a given HMM with transition probabilities a_{kl} and emission probabilities $e_k(b)$, we can now represent this HMM via a Boltzmann chain by setting $A_{kl} = T \log(a_{kl})$ and $B_k(b) = T \log(e_k(b))$.

2.4.5 Why Boltzmann chains instead of HMMs?

Given that HMMs are a much more thoroughly studied framework and our introduction to Boltzmann chains is given largely through the prism of HMMs, it is fair to ask why it is that we do not use an HMM for our modeling needs. Clearly any model is chosen for its applicability to the domain of interest, and this is true in our case. The specific reasons are twofold: First, because of the relaxed nature of transition weights with Boltzmann chains, we can more accurately represent the effects of DBF concentrations than if we normalized the concentrations against their sum total. That is, concentrations of 10 and 100, for example, have markedly different effects on binding predictions than concentrations of 0.1 and 1; the magnitudes of

the concentrations matter, not just the ratios. If the entire set of transition probabilities were scaled by the same multiple, the results would be identical, but given the structure of our model (see section 3.1) this is not the case.

Second, Boltzmann chains provide a convenient structure for explicitly accounting for the system temperature parameter. This is an aspect wholly ignored by HMMs, although a similar effect may be achieved via selective rescaling of transition probabilities (but not identical due to the constraint that the transitions be normalized).

3

Practical derivation, development, and implementation of our model

In the previous chapters, we have motivated and described a view of DNA occupancy that satisfies a set of desirable criteria, including incorporation of multiple of factors, accounting for effects of DBFs on one another, consideration of all possible binding configurations and continuous description of the likelihood of all possible configurations. In this chapter, we turn to the practical implementation of this view in statistical and computational terms, and then investigate its behavior with respect to local and global phenomena.

3.1 Model structure

In section 2.4, we introduced the ideas and mathematics necessary to model spatio-temporal processes via arbitrary Boltzmann chains. To apply this framework to a given problem, though, a graphical model must be constructed to reflect one's beliefs about the nature and behavior of the system that produced the output sequence of interest.

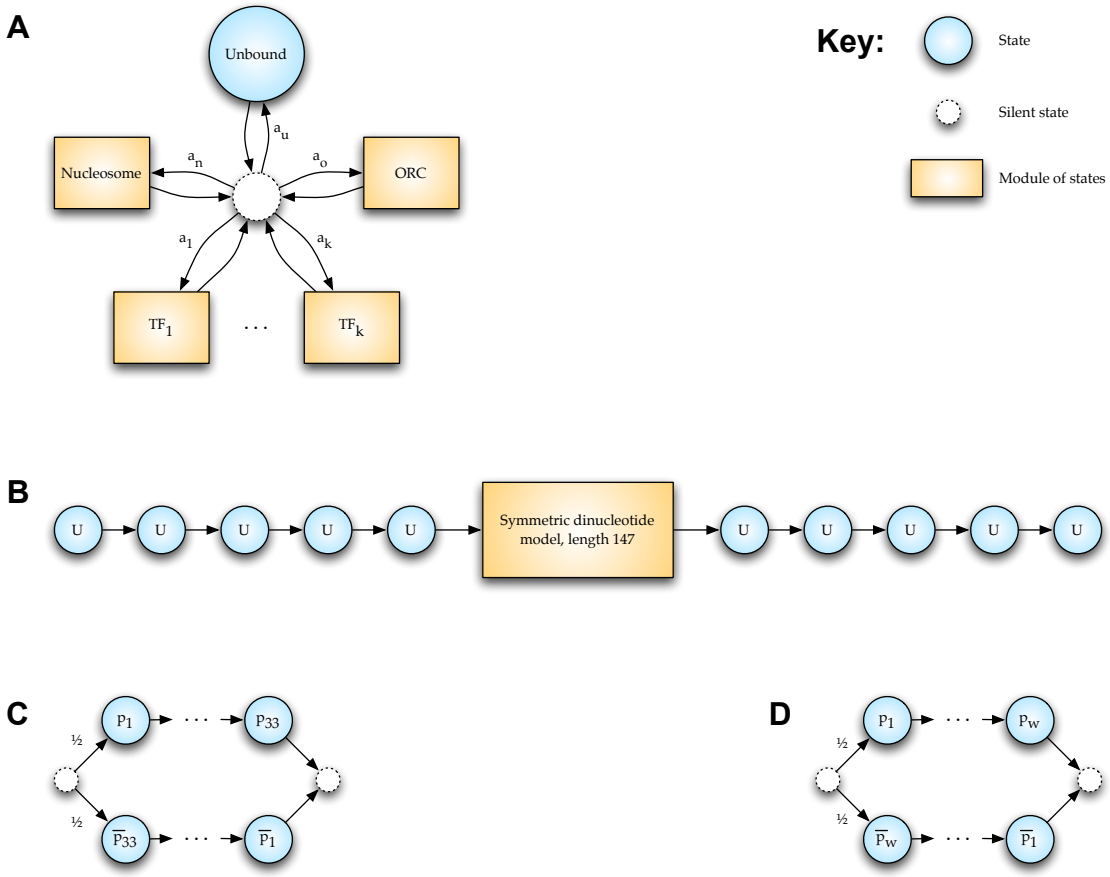

FIGURE 3.1: State transition diagram for Boltzmann chain implementation in COMPETE. Blue circular nodes are states and correspond to single nucleotides. Dashed circular nodes are silent states because they do not correspond to any nucleotide. Orange rectangular nodes are modules of states and correspond to a sequence of nucleotides of some length. Edges represent probabilistic transitions between states, with transition probabilities labeled; to reduce clutter, if a node has only one outbound edge, it is unlabeled (the probability is exactly 1). (**A**) At the highest level, the model represents each position in the genome either as being in an unbound state or as being bound by one of any number of DNA binding proteins or protein complexes. For instance, we might choose to model the genome as being bound by nucleosomes, the origin recognition complex (ORC), and k different transcription factors (TFs). Each DBF is represented by a module of an appropriate length. Transition probabilities from the central silent state are proportional to concentrations of the respective DBFs. (**B**) The nucleosome module consists of a symmetric dinucleotide model of length 147 flanked by 5 unbound nucleotides on either end to enforce a spacing between nucleosomes of at least length 10. To reduce clutter, we do not represent the states within the central module in this figure. (**C**) The ORC module consists of a PSSM-based EACS+B1 motif of length 33 (Xu et al., 2006) in either a forward or reverse-complement orientation (the origin might be on either strand and we assume that each occurs with probability 1/2). (**D**) A TF module consists of a PSSM-based motif of length w, a value which varies for each TF. Because a TF can bind to either strand, the w nucleotides arise either from the PSSM or from its reverse complement, each with probability 1/2.

As illustrated in Fig. 3.1, our model structure is conceptually simple, and quite flexible and extensible. It consists of a central silent state, from which transitions are possible to states that emit DNA not bound by any factor or bound by one of any number of modeled transcription factors, ORC, or a nucleosome, each of which in turn return to the central silent state. Unbound DNA is represented by a single state with emission probabilities reflecting genomic DNA content. TFs, ORC, and nucleosomes are "modules" of several states. TFs are represented as one silent state transitioning into two linear sets of states that proceed through the factor's motif in forward or reverse-complement orientations. Our TF motifs are taken from position-specific scoring matrices (PSSMs) learned from protein binding microarray data and trimmed according to an information content threshold of 0.3 (Zhu et al., 2009), except for the analyses of the *GTR1* promoter, in which the PSSMs are taken from MacIsaac et al. (2006). The ORC module consists of a PSSM-based EACS+B1 motif (Xu et al., 2006) in either forward or reverse-complement orientation. Our nucleosome model is a 147-position dinucleotide HMM flanked by five positions of unbound sequence on either side to enforce a minimum length of linker regions, as described in Segal et al. (2006). However, we have constructed the parameters of the dinucleotide model from recent *in vitro* nucleosome binding data (Kaplan et al., 2009), using over 4.7 million sequences instead of the 199 used originally in Segal et al. (2006). Following Kaplan et al. (2009), we replace the outermost 10 positions of either side of the 147-position model with background distribution to avoid potential micrococcal nuclease sequence specificity biases. Thus the final model has a 127-position dinucleotide core, flanked by 15 positions of background sequence on either side.

Given the modular nature of our model, any combination of DBFs may be represented, so our approach generalizes a wide range of previous techniques. Inclusion of only one TF at a specific concentration can reproduce results from GOMER (Granek

and Clarke, 2005). Using only a few TFs results in a model similar to that of Sinha (2006). A model with only nucleosomes recapitulates the model described in Segal et al. (2006).

3.2 COMPETE software implementation

We can represent the probability over valid binding configurations in the thermodynamic ensemble using a graphical model called a Boltzmann chain (section 2.4). A Boltzmann chain is a generalization of a hidden Markov model (HMM) in which values associated with state transitions and sequence emissions need not be normalized probabilities, but can be arbitrary nonnegative numbers. This generalization allows for easier incorporation of information related to binding energy and concentration, but retains the convenient property of HMMs that exact posterior inference of genome occupancy can be performed efficiently using dynamic programming. Consequently, running time scales linearly with the length of the input sequence.

We have implemented this exact posterior inference of genome occupancy in our COMPETE software package. COMPETE can process entire yeast chromosomes being bound by hundreds of DBFs in a matter of minutes, due to a collection of optimizations. As with HMMs, posterior decodings of Boltzmann chains can be calculated by the forward-backward algorithm (Rabiner, 1989). COMPETE (Competitive Occupancy: Multi-factor Profile Evaluation of a Thermodynamic Ensemble) implements this algorithm to compute its occupancy profiles in a highly efficient manner. The optimizations take place at several interconnected levels: as previously discussed, the model structure leverages silent states to reduce the complexity of the model (section 3.1, and the mathematics involved in calculating the posterior distributions of interest are tuned to allow these arbitrary-located silent states within the model, to be parallelizable (via breaking the traditional dependence between calculation of the forward and backward tables that arises during the usage of probability

scaling), and to retain a high degree of numerical precision (via probability scaling).

However, the optimizations also exist to a large extent within the implementation of the software. The Forward-Backward algorithm as described in section 2.4.3 involves, for each state in the model, performing calculations for every incoming (in the forward table) or outgoing (in the backward table) edge. In our implementation, written in C for speed and efficiency, we pre-scan the model for non-zero entries, to ensure that calculations are only performed for edges that actually exist between states. In a sparsely connected model with a high number of states (particularly in comparison with the average number of edges) such as ours, this optimization alone accounts for a several-hundredfold speedup in practice.

We also use multithreading on multi-core systems to calculate the forward and backward tables simultaneously. In terms of performance, occupancy profiles of entire yeast chromosomes can be computed in tens of minutes even when simultaneously using nucleosomes and all TFs for which we have PSSMs (around 125). The memory required to store the forward and backward tables is the only practically limiting resource, and this can easily be alleviated by periodically writing the tables to disk as they are being computed. The COMPETE software is available, and has been released with source code.

3.2.1 Software suite hierarchy structure

The COMPETE executable is the most important part of our software implementation, but in practice we use a suite of software to automate the execution via human-readable configuration files. We usually run COMPETE indirectly through a Ruby wrapper program (`wrapper.rb`) that takes YAML input files. Along with being human-readable, the YAML files are easily rewritable on the fly for automatic adjustment of which DBFs to include, the concentrations for the DBFs, which genomic sequence to run on, and where to produce output. We adjust all of these

at some point, either via `sed` or other general stream-editing tools or via specially written Ruby programs to, for example, set TF concentrations to multiples of K_d (section 3.4.7).

3.3 Datasets used

The primary datasets used in our work consist of the genomic sequence of *Saccharomyces cerevisiae* and sets of sequence preferences of numerous transcription factors as well as nucleosomes. The sequence data is taken from the May 27, 2006 genome build, curated by (and downloaded from) the Saccharomyces Genome Database (Cherry et al., 1998).

The transcription factor preferences come from two sources: an *in vivo* dataset and an *in vitro* dataset. The *in vivo* sequence preferences, derived from ChIP-chip experiments from Harbison et al. (2004), are taken from a re-analysis of this data published in MacIsaac et al. (2006). The *in vitro* sequence preferences, derived from PBM experiments, are taken from Zhu et al. (2009).

The nucleosome sequence preferences are derived from ChIP-seq data in Kaplan et al. (2009) and calculated using methodology described in Segal et al. (2006).

The specific manners in which these datasets are leveraged are discussed in section 3.1.

3.4 Results

To illustrate the usefulness and generality of our method and our software, in what follows we present a series of examples illustrating distinctive features of our model of genome binding, along with demonstrations of how our model of DBF competition improves both TF positioning and nucleosome positioning genome-wide. These are all computed using COMPETE applied to the genome sequence of the yeast *S. cerevisiae*, with binding specificities of DBFs determined by recent *in vitro* methods (Zhu

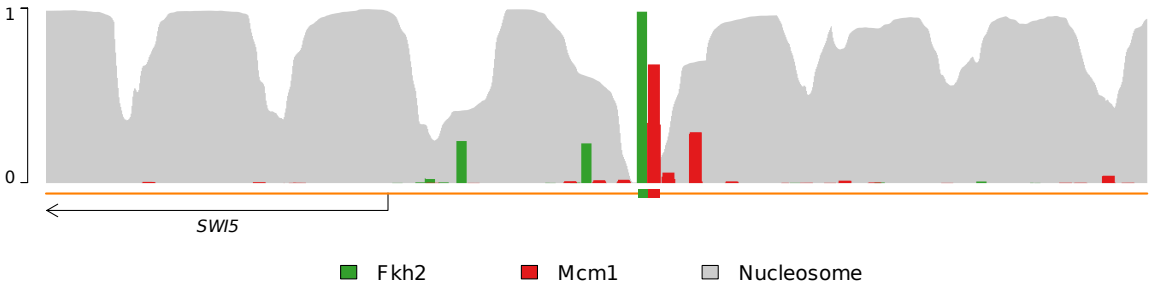

FIGURE 3.2: Occupancy profiles. Shown is a DBF occupancy profile. X-axis is genome position, y-axis is probability of occupancy. Plots with each legend color indicate the probability of the respective DBF occupying each position. The orange line underneath each plot represents the genome; oriented arrows thereon demarcate coding regions of genes according to SGD (Cherry et al., 1998); colored boxes thereon denote strong TF binding site matches according to MacIsaac et al. (2006).

et al., 2009; Kaplan et al., 2009).

This last point is quite important: by using binding specificities of TFs (Zhu et al., 2009) and nucleosomes (Kaplan et al., 2009) bound to naked DNA at low concentrations and in isolation of other factors, our model is able to distinguish between the inherent binding specificities of DBFs (which COMPETE takes as input) and the consequent locations that DBFs take up along the genome when in competition with one another (which COMPETE produces as output). In particular, methods that use TF or nucleosome specificities based on *in vivo* binding locations (such as those of Harbison et al. (2004) or Segal et al. (2006)) will be unable to resolve the relative contributions of inherent sequence specificity versus the consequences of competitive binding.

3.4.1 Occupancy profiles

The results shown in this chapter are frequently illustrated by *occupancy profiles*, with an example shown in Figure 3.2. An occupancy profile is a summary of the posterior decoding of a given stretch of DNA using a given model and set of parameters provided to COMPETE. To a first approximation, the occupancy profile is a bar

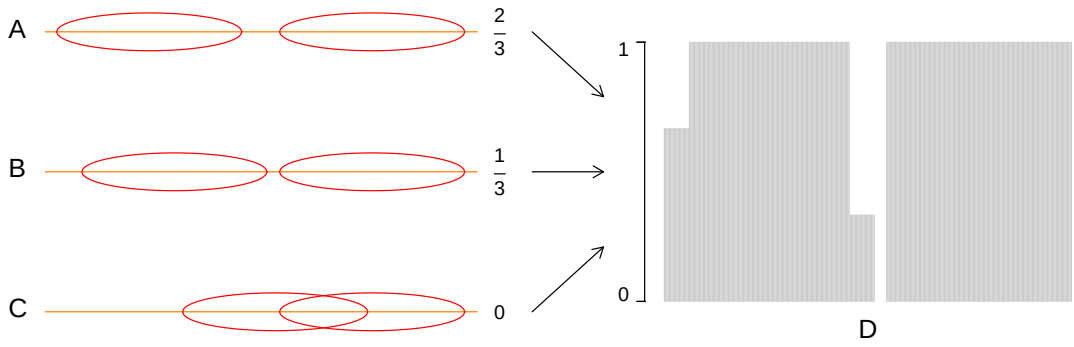

FIGURE 3.3: Example distribution of per-site occupancy. In this simple scenario consisting of only nucleosomes and unbound DNA, nucleosomes are found in configuration **A** $\frac{2}{3}$ of the time, configuration **B** $\frac{1}{3}$ of the time, and are never found in the illegal configuration **C**. **D** shows the output of our method as applied to this set of data. Note the "steps" on either side of the left-hand nucleosome with a central "plateau", suggesting a central core occupied at a constant rate with some variability in the starting, and hence ending, position of the nucleosome. While only nucleosome is shown here, this scenario applies to any DNA binding factor in consideration.

plot of the probability of occupancy per DBF per sequence position. As opposed to a stacked bar plot, the occupancy profile plot places the base of each DBF bar at zero, allowing ease of comparison between heights of different bars. Additionally, the probability of a position being unbound is implicitly represented as empty space instead of being explicitly represented by a colored bar.

Because the posterior decoding gives probabilities at a finer resolution than per DBF per sequence position, being per *state* per sequence position, and this level of detail would clutter a plot instead of adding utility, the occupancy profile summarizes DBF occupancy by presenting the probability of binding by a DBF as the sum of the probabilities of each state in the forward and reverse-complement orientations of the DBF.

The occupancy profile is given with annotations of the genomic region that it corresponds to. The orange line below the plot represents the genome, with colored

boxes corresponding to literature-curated transcription factor binding sites and the arrows corresponding to the gene coding regions, with the directionality of the arrow indicating which strand the gene lies on.

Interpretation of occupancy profiles

In practice, occupancy profiles often depict relatively rounded or ill-defined shapes for nucleosome positions and relatively well-defined positions for many TF positions. In interpreting these plots, it is important to recognize that the posterior distributions are reflecting a weighted average of possible binding configurations. A toy example illustrating how the configurations and their weights result in occupancy profiles is given in Figure 3.3.

The degeneracy of nucleosome positions and the lack thereof of the positions of many TFs is a function of their respective weak and strong sequence specificities. While all binding configurations are considered and weighted, the fact that TF positions show high specificity is a result of the extremely low probabilities of configurations with the TFs positioned *away* from these particular binding locations. Similarly, the degenerate nucleosome positions suggest the relative lack of weighting against nucleosome shifts; that is, nucleosomes are allowed to move around a fair amount more than many types of TFs.

Importantly, this is not a refutation of our earlier assertion of the lack of discreteness of DBF binding. On the contrary, nucleosomes and TFs are indeed binding all sequence with some probability; that probability is just much lower for some TFs than for nucleosomes (and other TFs, for that matter).

3.4.2 Competition with transcription factors affects nucleosome occupancy

Direct competition between DBFs is an important guiding force in determining genome occupancy. The result of this competition depends on several factors, includ-

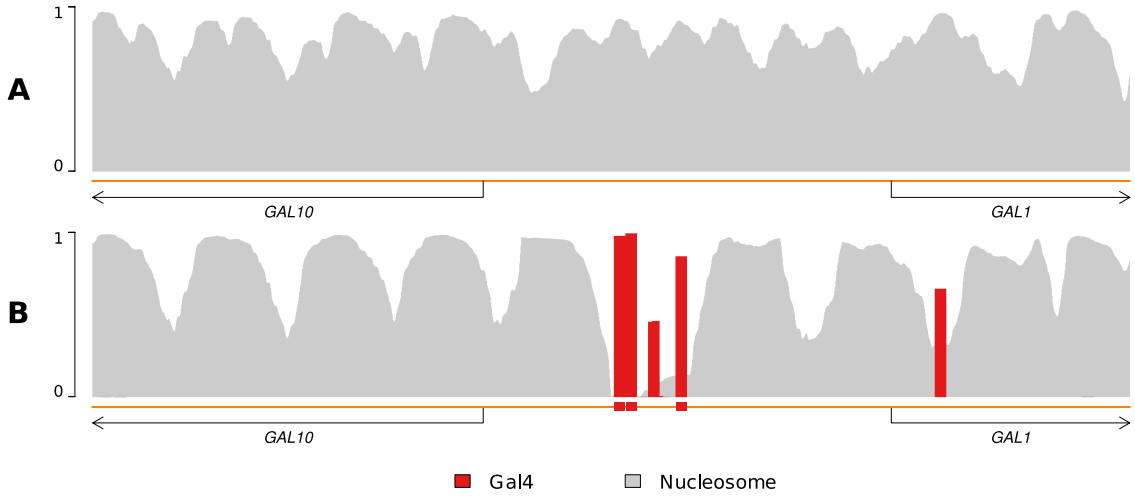

FIGURE 3.4: Transcription factor binding can stabilize nucleosomes. (**A**) Including only nucleosomes and unbound genomic sequence yields an occupancy profile similar to existing nucleosome-only positioning models like that of Segal et al. (2006). In this genomic region, the degenerate nature of nucleosome sequence preferences leads to significant uncertainty of nucleosome positioning. (**B**) Addition of Gal4 yields a strikingly different nucleosome occupancy profile, revealing stable nucleosome positioning. As Gal4 strongly binds its annotated binding sites, nucleosomes are outcompeted at those locations and 'boundaries' are established. These boundaries reduce the feasibility of nucleosome translation, so nucleosome positions coalesce into more stable locations in response.

ing concentration and sequence preferences of the DBFs involved, along with steric hindrance between nearby DBFs, all of which are accounted for in our approach. As one example, although nucleosomes and TFs may in some cases simultaneously bind the same stretch of DNA, their binding is primarily mutually exclusive. A consequence of this mutual exclusion is that nucleosome positions along on the genome are a result not only of inherent nucleosomal sequence preferences, but also competition with TFs and other DBFs.

In Fig. 3.4, we present an example showing how the binding of TFs can position nucleosomes more stably along the genome (by stable, we mean that a nucleosome appears in a well-defined location, as opposed to in one of a few different trans-

lated locations). Depicted is the region surrounding the *GAL1-GAL10* promoter on Chromosome II. In the top panel is the occupancy profile in which nucleosomes compete for positions along the genome without any other DBF competitors (as in the models of Segal et al. (2006) and Kaplan et al. (2009)). As is evident, nucleosome positions are largely ambiguous in this region, owing to a significant degree of uncertainty regarding how the nucleosomes are translated along the genome. In contrast, the bottom panel reveals how nucleosomes are positioned far more stably once the binding of Gal4 is taken into account. The strong binding of Gal4 to its cognate sites in the *GAL1-GAL10* promoter (and even to a site within the coding region of *GAL1*) helps to establish nucleosome 'boundaries'. These boundaries in turn reduce the probability of the various translational modes of the nucleosomes and thus increase their stability at specific locations.

In Fig. 3.5, we present another example showing how the binding of TFs can displace nucleosomes from the genome. Depicted is the region surrounding the *GTR1* promoter on Chromosome XIII. In the top panel again is the occupancy profile in which nucleosomes compete for positions along the genome without any other DBF competitors. Nucleosome positions at the left are somewhat ambiguous, but become more stably positioned to the right. Five to six reasonably stably positioned nucleosomes can be seen, starting near the start codon of *YML122C*. In contrast, the bottom panel reveals how nucleosomes are positioned once the binding of Pho4, Pho2, and Nrg1 is taken into account. The binding of Pho4, Pho2, and Nrg1 to sites in the *GTR1* promoter displaces the two nucleosomes that would otherwise reside there. It also helps to stably position a nucleosome near the 3' end of the coding region of *YML122C*, and hems in a single nucleosome near the 5' end. The strong binding of the TFs form boundaries that prevent nucleosomes from translating upstream or downstream; instead, overall nucleosome occupancy in the region drops dramatically, illustrating the importance of modeling direct competition between nucleosomes and

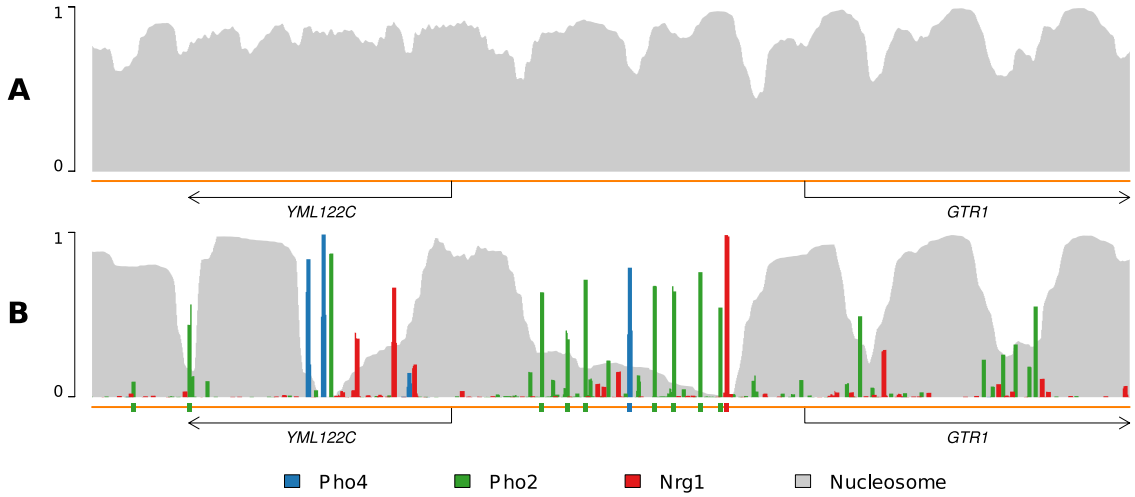

FIGURE 3.5: Transcription factor binding can displace nucleosomes. The promoter region of *GTR1* is known to contain high-affinity binding sites of Pho4 and Nrg1, as well as several of Pho2. (**A**) An occupancy profile including only nucleosomes depicts five to six reasonably well-positioned nucleosomes beginning atop the start codon of *YML122C*. (**B**) Addition of Pho4, Pho2, and Nrg1 to the model results in the displacement of several of the previously bound nucleosomes, and the reduction of binding frequency of those remaining, to various extents. Additionally, a well-positioned nucleosome is established near the 3' end of *YML122C*. The strong TF binding in this region precludes nucleosome repositioning upstream or downstream along the sequence; the strongly-bound TFs displace some nucleosomes and hem in the remaining others, resulting in their well-defined positions.

TFs.

3.4.3 Competition with nucleosomes affects transcription factor occupancy

Just as existing nucleosome positioning models do not generally take into account the binding of other DBFs, existing TF positioning models generally do not take other kinds of DBFs into account. In particular, existing TF positioning models do not generally consider competition with nucleosomes as ours does.

In Fig. 3.6, we present an example demonstrating the role that nucleosomes play in reducing the non-specific binding of TFs across the genome. Depicted is the region surrounding the *SWI5* promoter on Chromosome IV. In the top panel is the

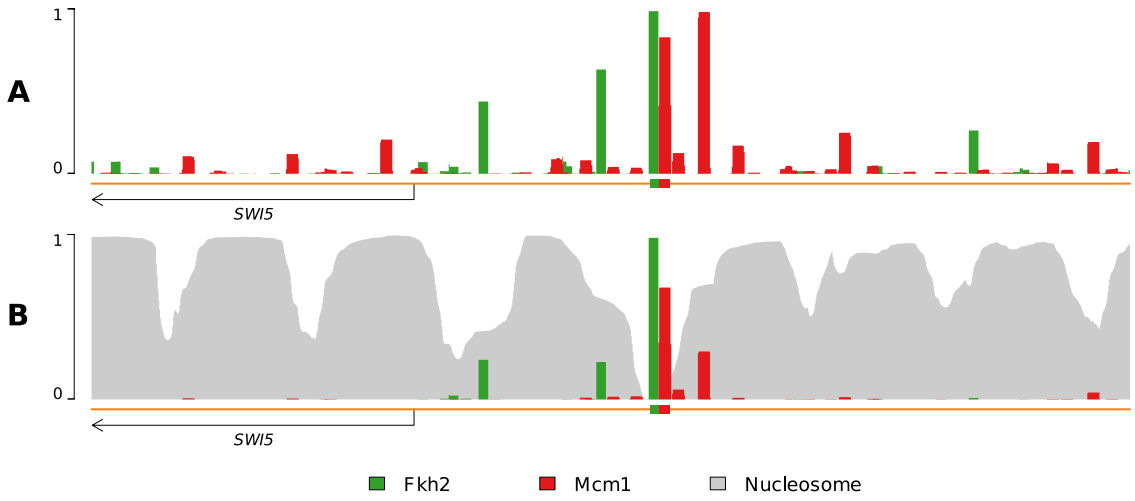

FIGURE 3.6: Nucleosome binding attenuates non-functional transcription factor binding. (**A**) A profile using a model of only TFs and DNA yields many (presumably non-functional) binding predictions throughout the genome. The binding of Fkh2 and Mcm1 is significant even away from their annotated binding sites; Mcm1 even displays an upstream (to the right) binding occurrence stronger than at its annotated site. However, Fkh2 and Mcm1 form a ternary complex with DNA, and are known to bind adjacently to one another, making the observation of this additional, stronger, binding occurrence somewhat curious. (**B**) Addition of nucleosomes yields a dramatically altered TF occupancy profile. Fkh2 and Mcm1 binding is notably diminished at sites other than annotated binding sites, though not entirely eliminated. Additionally, the strong upstream Mcm1 binding event occurs at a much-reduced level, now approximately half as frequently as that of the annotated Mcm1 binding site.

occupancy profile when only Fkh2 and Mcm1 are included in the model. Although the factors bind strongly to their annotated binding sites within the promoter, a great deal of weaker binding can also be observed. Of particular note is the fact that Mcm1 has a stronger match to a binding site just upstream (to the right) of its annotated site than it does to the annotated site itself. Because Fkh2 and Mcm1 form a ternary complex with the DNA, we know that Mcm1 should bind immediately adjacent to Fkh2 more frequently. How does the cell enforce this? Consider what happens when nucleosomes are included in the model, as shown in the bottom panel. Now, the vast majority of non-specific binding of both Fkh2 and Mcm1 has been eliminated.

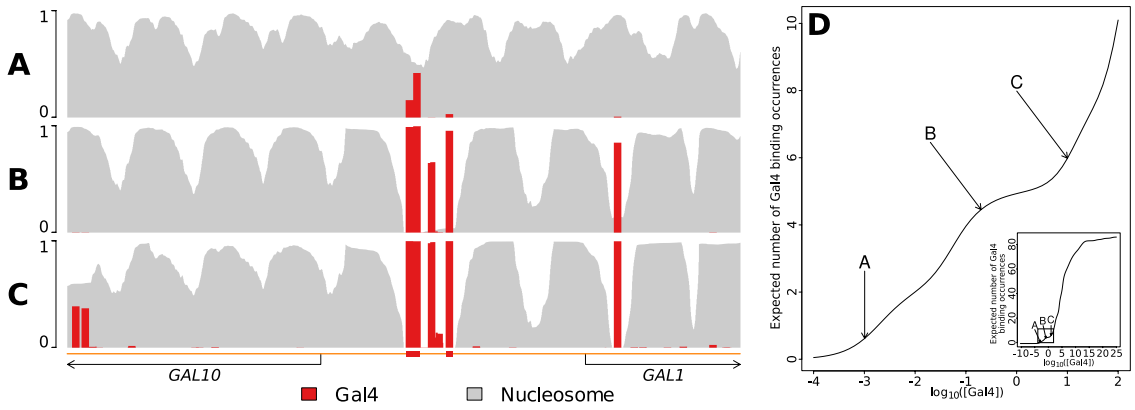

FIGURE 3.7: Binding occupancy is highly dependent on factor concentration. (**A**) At low concentration, Gal4 binding is virtually nonexistent apart from low levels of binding at its annotated binding sites. (**B**) As concentration increases, Gal4 binds strongly to its annotated sites, displacing nucleosomes there and leading to the stabilization of flanking nucleosome positions, as also illustrated in Fig. 3.4. This stabilization and reorganization allows Gal4 to bind at two additional sites. (**C**) Increasing Gal4 concentration further eventually leads to strong binding even away from its annotated sites. However, even at this level, Gal4 cannot outcompete many of the strongly-positioned nucleosomes, reiterating the importance of including many different DBFs concurrently in the model. (**D**) Gal4 binding occupancy as a function of Gal4 concentration. Note how occupancy changes more slowly around the seemingly plausible concentration used in (**B**) than in any other part of the overall occupancy versus concentration curve (inset), suggesting that while occupancy can be fairly sensitive to concentration on the whole, the cell may sometimes be operating in ranges that are relatively less sensitive.

In particular, because of direct competition between Mcm1 and nucleosomes, the upstream Mcm1 binding site is now occupied much less frequently than the site immediately adjacent to Fkh2, as expected.

3.4.4 Varying transcription factor concentration affects occupancy

The binding frequency of a DBF to a particular segment of DNA will depend on its concentration. In Fig. 3.7, we show an example in which Gal4 and nucleosomes compete for binding sites in the region surrounding the *GAL1-GAL10* promoter, at varying concentrations of Gal4. Gal4 is known to bind strongly to several binding

sites in this intergenic region (Giniger et al., 1985; Harbison et al., 2004), and our occupancy profiles recapitulate that behavior. At very low concentrations (panel A), Gal4 only binds at low levels, and essentially only at annotated binding sites, those being by definition the strongest in terms of sequence preference. An increased value of the concentration parameter for Gal4 results in an increased level of occupancy at the annotated binding sites (panel B), along with occupancy at two additional sites that can now be bound because of nucleosomal rearrangements introduced by strong binding to the three annotated sites. At even higher concentrations, additional sites start to become occupied (panel C). Throughout the process, as Gal4 binds more sites with increasing concentration, the nucleosome occupancy profile is dramatically reorganized. We can plot the total level of Gal4 occupancy in the promoter as a function of log concentration parameter (panel D). The range of plausible concentrations is quite small, including B but not likely extending beyond the highlighted region of the overall occupancy versus concentration curve (shown in the inset). Interestingly, several well-positioned nucleosomes continue to occupy their positions with little change even as the increased concentration of Gal4 (C and beyond) yields strong binding away from its annotated sites, reiterating the importance of including nucleosomes in the model.

3.4.5 Binding cooperativity emerges implicitly by considering competition explicitly

In addition to competing with one another, TFs are also known to bind cooperatively in some instances. A few models have shown the importance of TF cooperativity in establishing sharp expression patterns, and have included explicit energetic 'bonuses' to model this cooperativity (Segal et al., 2008). These explicit bonus terms may be added in our model, but binding cooperativity can also arise in our model implicitly, even without the inclusion of such terms. This can occur when multiple TFs are competing against the same nucleosome for nearby binding sites. In such a situation,

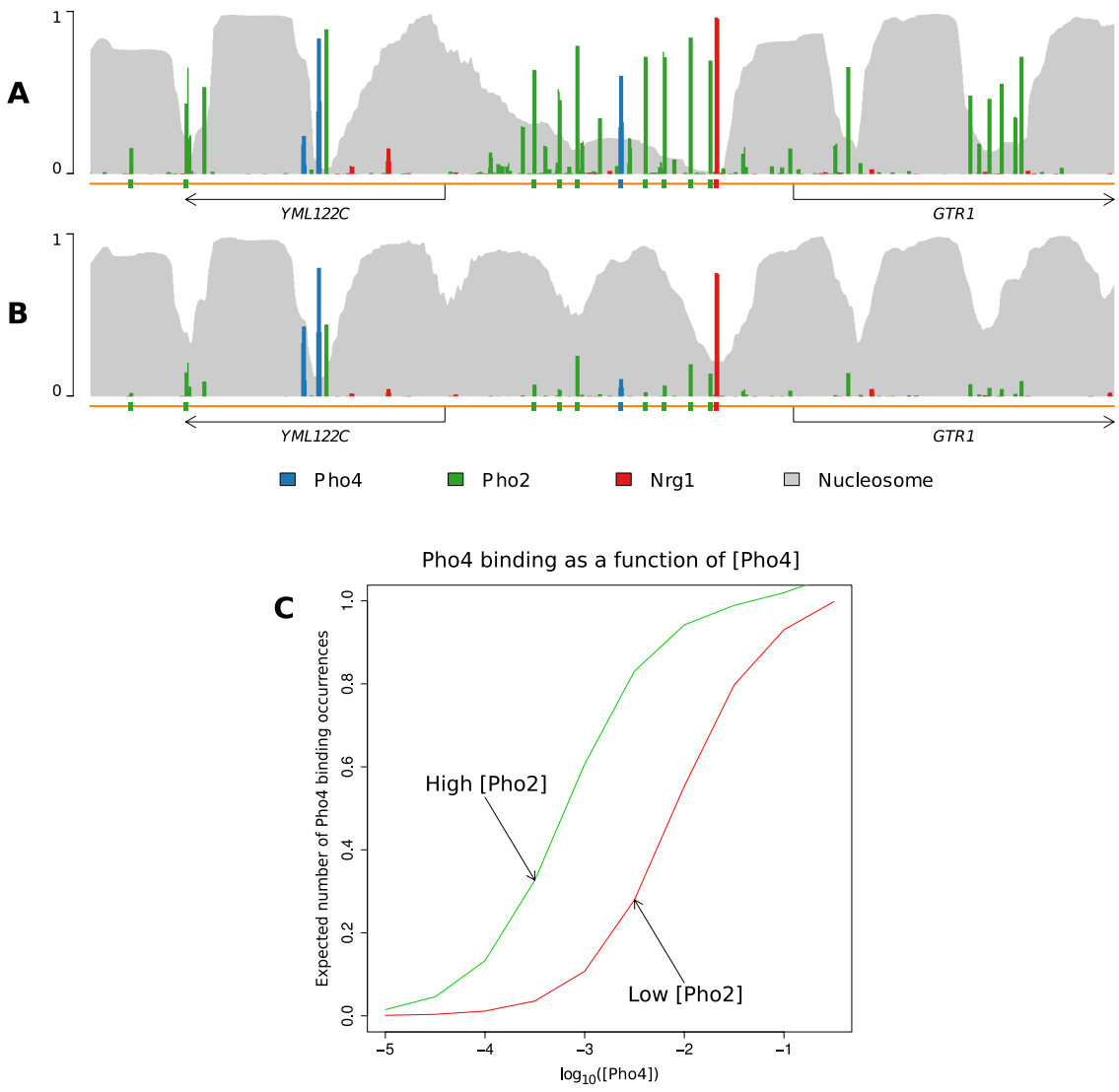

FIGURE 3.8: Cooperative binding emerges implicitly from explicit competition between DBFs. Revisiting the *GTR1* promoter, cooperativity amongst Pho4 and Pho2 can be observed via mutual competition with nucleosomes, particularly the nucleosome positioned atop the annotated Pho4 site and Pho2 sites adjacent to the strong Nrg1 binding site. (**A**) A profile of this region with a high concentration of Pho4 results in high occupancy of both Pho4 and Pho2. (**B**) A similar profile but with decreased concentration of Pho4 results in decreased occupancy of both Pho4 and Pho2, and a resulting increase in the occupancy of nucleosomes. (**C**) A plot of Pho4 binding occupancy as a function of Pho4 concentration, at both high and low concentrations of Pho2. The plot reveals that a higher concentration of Pho2 leads to higher occupancy of Pho4 at all concentrations of Pho4. Importantly, this cooperative effect is purely a consequence of inclusion of nucleosomes into the model. That is, Pho4 and Pho2 have an implicitly cooperative effect due to joint competition with the same nucleosome, as opposed to some explicit energetic cooperativity term.

the more frequently one of the factors is able to bind to a site on the genome, the less often the competing nucleosome will be present to occlude the binding of the other factor. This behavior has been termed 'collaborative competition' and observed experimentally (Miller and Widom, 2003) and characterized mathematically (Mirny, 2009).

In Fig. 3.8, we revisit the region surrounding the *GTR1* promoter, this time focusing on the binding of Pho4 at different concentrations of Pho2. The top occupancy profile is computed at a low Pho2 concentration; the bottom profile is computed at a higher Pho2 concentration. Clearly the higher Pho2 concentration will increase the overall occupancy of Pho2, but what is especially interesting is that it also increases the overall occupancy of Pho4, even though the concentration of Pho4 is unchanged. This is because as Pho2 increasingly binds sites in the promoter, it displaces nucleosomes from the promoter with greater frequency, thereby providing greater access for Pho4 to bind, even at the same level of Pho4. Indeed, we can plot Pho4 occupancy as a function of Pho4 concentration, at each of the two Pho2 concentrations. The occupancy is consistently higher across the entire range of Pho4 concentrations when Pho2 is present at the higher concentration. This upward shift of the Pho4 curve with increasing concentration of Pho2 is a form of binding cooperativity that has arisen entirely implicitly.

3.4.6 Competitive occupancy of ARS sequences near origins of replication

Our examples thus far have focused on TFs and nucleosomes binding the genome, but DNA is occupied by a host of other proteins and protein complexes as well. Among these is the origin recognition complex (ORC), which is instrumental in the initiation of DNA replication. Like TFs, ORC binds exclusively of other DBFs and has a DNA sequence preference that guides its binding in yeast. However, ORC's sequence preferences are rather degenerate; scanning the ORC motif across the entire

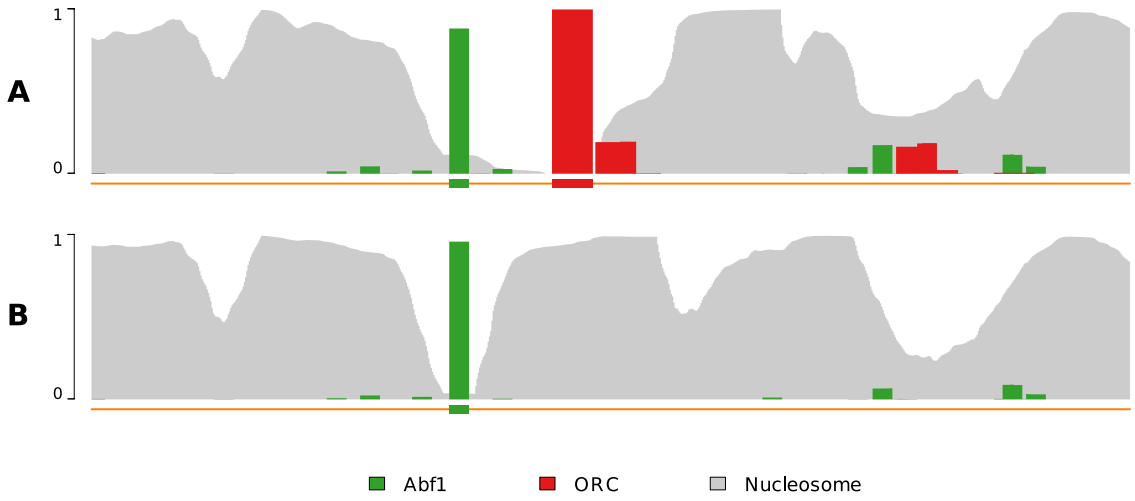

FIGURE 3.9: ORC and nucleosome binding at *ARS1*. The origin recognition complex (ORC) and chromatin play strong roles in the initiation of DNA replication. *ARS1*, an origin of replication in yeast, has a well-characterized DNA occupancy structure including binding by ORC, nucleosomes, and the transcription factor Abf1. This structure is observed to change in the absence of ORC binding, as can be seen in Fig. 2A of Lipford and Bell (2001). The Abf1 binding site annotation is taken from Marahrens and Stillman (1992). The ORC binding site and sequence preferences are derived from Xu et al. (2006). (**A**) The wild-type occupancy profile recapitulates the strong binding of nucleosomes adjacent to the Abf1 and ORC binding sites, as well as the nucleosome free region between them. (**B**) Without ORC present, the occupancy profile depicts a shifted nucleosome over top of the ORC binding site, as is the case in the experimental observation of Lipford and Bell (2001).

genome without consideration of other competing DBFs yields tens of thousands of matches (Breier et al., 2004), though only a few hundred sites seem to be bound by ORC or function as origins of replication *in vivo* (Raghuraman et al., 2001; Wyrick et al., 2001). This suggests a role for competition with other DBFs, including TFs and nucleosomes, in guiding ORC to functional replication origins. In particular, chromatin is known to play a role in pre-replicative complex formation at origins of replication (Lipford and Bell, 2001). Therefore, ORC and nucleosomes must be considered together, along with other DBFs, to provide a full view of binding behavior at origins of replication.

Fig. 3.9 provides occupancy profiles for the region of the genome surrounding the well-studied *ARS1* origin of replication. *ARS1* has experimentally verified binding sites for ORC and Abf1 (Marahrens and Stillman, 1992). Chromatin arrangement in *ARS1* is fairly well characterized, and changes depending on the binding of ORC (Lipford and Bell, 2001). Our model is able to recapitulate this experimentally observed occupancy of *ARS1* in both the presence and absence of ORC (Fig. 3.9A,B). In the presence of ORC, the strong experimentally observed nucleosome occupancy adjacent to the Abf1 and ORC binding sites is recapitulated, as is the nucleosome free region between them. In the absence of ORC, however, a nucleosome has now shifted to cover the ORC binding site, as was also observed experimentally.

3.4.7 DBF competition improves transcription factor binding predictions genome-wide

To this point, we have shown examples of various phenomena that are captured by our modeling framework, and we have done so by focusing on specific loci. However, we expect that predictions of genome-wide DBF positioning should also be improved by the inclusion of competition between nucleosomes and large numbers of TFs in our framework. To demonstrate this, we first examine the effect of DBF competition on the prediction of TF binding genome-wide; in the next section, we consider the effect on the prediction of nucleosome binding genome-wide.

We analyzed the agreement between the binding of TFs to yeast promoters genome-wide as determined experimentally *in vivo* by ChIP-chip data (Harbison et al., 2004), and as predicted computationally by COMPETE under two settings: each TF is modeled to bind by itself versus doing so in competition with nucleosomes and 88 other TFs. The improvement in agreement with experimental data under these two settings across 135 ChIP-chip experiments is shown in Fig. 3.10A, with examples of how this improvement is computed for two particular factors shown

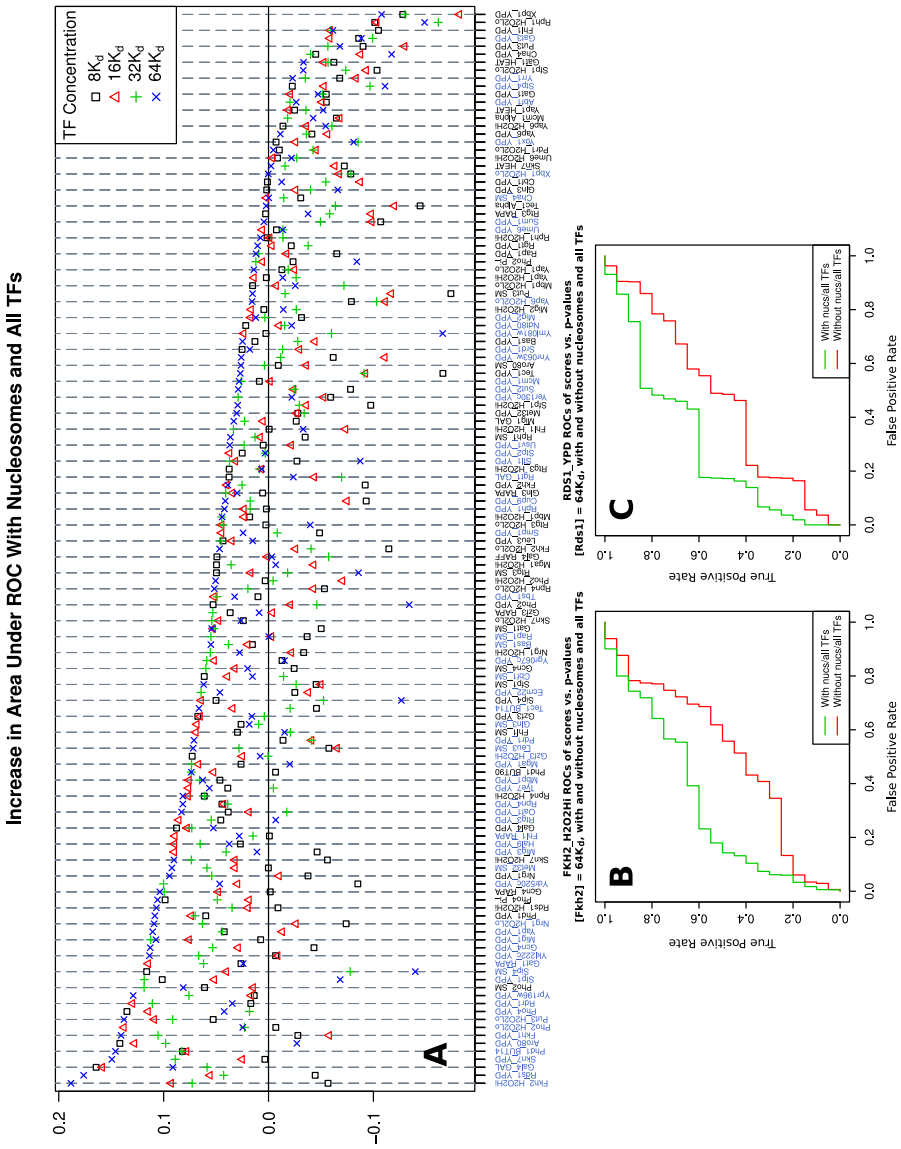

FIGURE 3.10: DBF competition improves transcription factor positioning predictions genome-wide. COMPETE is used to predict mean probabilities of TF binding in regions corresponding to probes from whole-genome ChIP-chip TF binding data (Harbison et al., 2004). The top and bottom 10% of regions as scored by COMPETE are respectively labeled as positively and negatively bound promoters. ROCs are then constructed for probe p-values using these labels. The change in the area under the curve (AUC) of the ROCs is calculated between a model including only one TF and a model including that TF and nucleosomes and all other TFs, at various TF concentrations. The concentrations of all TFs are set to the same multiple of their respective K_d; the value of K_d for each TF is different and calculated from its PSSM, as per Granek and Clarke (2005). (**A**) Change in ROC AUC between TF-only and all TF and nucleosome models, across concentrations. Each point represents a pair of whole-genome analyses, with and without nucleosomes and all other TFs, totaling 272 decodings of the entire genome and 1080 individual analyses. Positive values signify improvement. Experimental conditions for each ChIP-chip experiment are given on the x-axis, with the best performing condition for each TF highlighted in blue. Most TF binding predictions either improve or behave somewhat similarly when including nucleosomes. As it is likely that different concentration choices are appropriate for each TF, several concentrations are included. Values of all points are given in Table A.1, in Appendix A. (**B**) ROCs of Fkh2 ChIP-chip p-values with labels determined by binding probability as predicted by COMPETE in the FKH2_H2O2Hi experiment, with a model including Fkh2, nucleosomes, and all other TFs (green) and a model including Fkh2 alone (red). (**C**) Same as (**B**), but using the RDS1_YPD experiment with models including Rds1 in place of Fkh2.

in panels B and C.

A clear trend can be observed in this plot, in that the majority of TF binding predictions either improve or behave similarly when competition with nucleosomes and other TFs is introduced. Because the ChIP-chip data are known to be somewhat noisy and might not be always a faithful representation of *in vivo* binding, small changes are perhaps not that conclusive. Calling inconclusive those experiments where the absolute change for the best of the four global concentration scalings is less than 0.025, approximately 69% of the 135 experiments showed a positive change, while only about 6% showed a negative change. Changes are unlikely to always be positive because of numerous complications involved in modeling actual *in vivo* DBF binding, including active chromatin remodeling, explicit binding cooperativity, and importantly, the challenge of correctly specifying 90 different DBF concentrations. Noticeably, the change is maximized at different concentrations for each TF, highlighting the difficulty of solving the issue of robustly determining appropriate DBF concentrations.

Methodology for genome-wide analysis of transcription factor binding

Several steps are involved in carrying out our analysis of the genome-wide effect of DBF competition on TF binding. First, the mean probability of starting a binding site for a given TF is calculated for every region corresponding to the ChIP-chip probes of Harbison et al. (2004) across the genome. Under the premise that high probability regions should correspond to bound promoters and low probability regions should correspond to unbound promoters, the top and bottom 10% of regions by mean predicted binding probability are labeled as bound and unbound, respectively. These top and bottom 10% sets are selected by setting thresholds scaled against the maximum value in the genome, and contain either the points above (or below, for the bottom set) that threshold or the 20 highest (or lowest) points, whichever is

larger. The receiver operating characteristic (ROC) curve is then constructed for the probe p-values using the aforementioned labels. The area under the curve (AUC) of the ROCs is used as a measure of binding prediction. The change in these AUCs is calculated (by subtraction) between a model including only one TF, and a model including that TF in competition with nucleosomes and 88 other TFs, at various TF concentrations. Intersecting the set of TFs from the Harbison et al. (2004) ChIP-chip data set with those for which PBM binding motifs exist in Zhu et al. (2009) yields 67 TFs and 135 binding experiments for which this change can be computed. Though only these 67 TFs are analyzed for this metric, all 89 TFs from the PBM data of Zhu et al. (2009) are included in the competition model.

The choice of TF concentrations is a difficult challenge, in that allowing 89 TFs to compete for binding sites along the genome at once implies the need for selection of 89 different concentration parameters. Selecting these parameters independently can result in excessive parameter tuning or over-fitting. However, cross-validation strategies for learning these parameters from data present their own complications because they would require searching in an 89-dimensional real concentration parameter space, even setting aside the many additional parameters relating to binding specificity that could be optimized. To address these challenges and simplify the TF concentration parameter space to a 1-dimensional subspace, we first calculate the K_d of each TF's PSSM, as detailed in Granek and Clarke (2005).

As a brief aside, the motivation for using K_d as a unit is largely related to a desire to express concentration in units that are meaningful independent of the identity of the TF at hand. Because TF concentrations are considered multiplicatively with the score of the TF's PSSM, and the score is dependent on length (being the product of all columns of the PSSM), concentrations of factors of different lengths are not directly comparable. The probability of a DBF X binding its consensus site at concentration

$[X]$ in isolation may be expressed in terms of its K_d as $P(\text{binding}) = \frac{[X]}{K_d+[X]}$. In this way, usage of units of K_d normalize concentrations to be in units that *are* directly comparable. By setting concentrations for all TFs to be the same multiple of their respective K_ds, the probability of each TF binding its consensus site is equal under a simple scan of the motifs across the genome without any competition.

To complete our analysis, each TF's concentration is then set to its K_d multiplied by a single global scaling factor, using the same factor for all TFs. To simplify the space of concentrations further, only six discrete values ($\{2, 4, 8, 16, 32, 64\}$) are used instead of the full range of real values for this global scaling factor.

3.4.8 DBF competition improves nucleosome binding predictions genome-wide

As we expect that nucleosomes and TFs will mutually improve one another's positioning predictions, we also analyzed the agreement between the binding of nucleosomes across the yeast genome as recently determined experimentally *in vivo* (Kaplan et al., 2009), and as predicted computationally by COMPETE under two settings: nucleosomes are modeled to bind by themselves versus doing so in competition with 89 TFs. To analyze our predictions, we consider the improvement under these two settings of the Spearman correlation between a ranked list of genomic regions enriched and depleted for nucleosome occupancy with a list of those regions ranked by predicted occupancy according to COMPETE.

As shown in Fig. 3.11, inclusion of competition with TFs is clearly beneficial in predicting nucleosome occupancy across a range of TF concentrations. This effect is demonstrated for various thresholds used for extraction of enriched and depleted regions, and increases with the stringency of the threshold, as expected. In all cases, as TF concentrations are increased, the improvement reaches a plateau and then diminishes; for low-stringency thresholds, adding all 89 TFs at extremely high concentrations is even deleterious. This is not surprising because at these high concen-

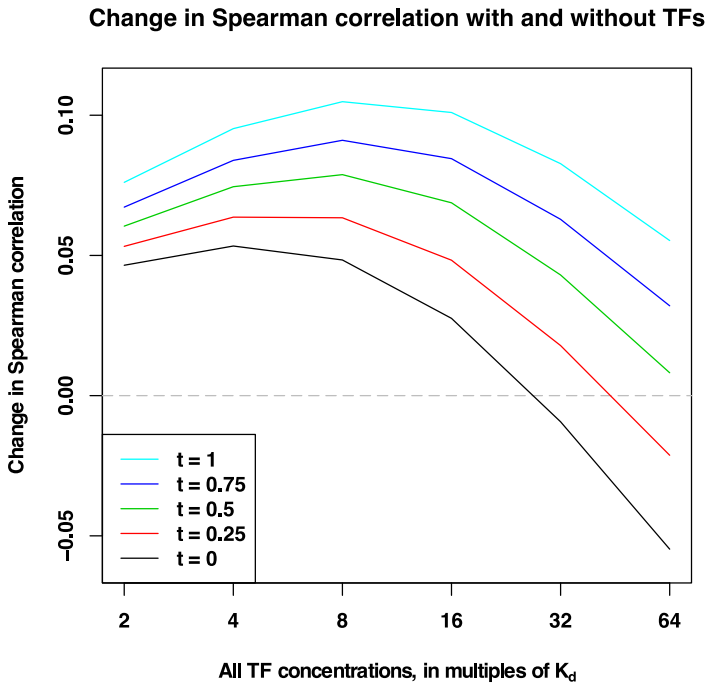

FIGURE 3.11: DBF competition improves nucleosome positioning predictions genome-wide. Enriched and depleted regions across the genome are extracted from the *in vivo* experimental map of Kaplan et al. (2009) using different stringency thresholds t, with all positions in enriched regions having experimental nucleosome occupancy greater than t and all positions in depleted regions having experimental nucleosome occupancy less than $-t$. The Spearman correlation is computed between the experimentally measured nucleosome occupancy of the regions and the occupancy predicted by COMPETE using nucleosome models with and without all 89 TFs of Zhu et al. (2009). The TF concentrations are all set to the same multiples of their respective K_d. Each point corresponds to a pair of whole-genome analyses, with and without all 89 TFs, totaling 12 decodings of the genome and 60 individual analyses. The changes between Spearman correlation of experimental data and COMPETE predictions with and without nucleosomes are shown here, where positive values signify improvement. Inclusion of TFs is clearly beneficial to nucleosome positioning across TF concentrations. This effect ranges in extent and is demonstrated for various TF concentrations and various thresholds t. Raising TF concentrations to high levels imposes deleterious effects on nucleosome positioning, likely due to diminished nucleosome occupancy in competition with many different highly concentrated TFs. When TF concentrations are kept below these high levels, TF inclusion aids significantly in nucleosome binding predictions. Values of all points are given in Table A.2, in Appendix A.

trations, TFs are able to outcompete nucleosomes even at many enriched locations. Nevertheless, the vast majority of combinations of concentrations and stringency thresholds show improved prediction of nucleosome occupancy genome-wide when competition with TFs is incorporated.

Methodology for genome-wide analysis of nucleosome binding

The experimental map of *in vivo* nucleosome positions is the YPD map taken from Kaplan et al. (2009). Following Kaplan et al., we identify maximal regions along the genome of at least 50 consecutive nucleotides in which the nucleosome occupancy of each position in the region is either greater than some threshold t (for enriched regions) or less than $-t$ (for depleted regions). These regions are then sorted by their mean nucleosome occupancy as experimentally determined, as well as by their mean nucleosome binding probability as computed by COMPETE, to produce a pair of ranked lists. The Spearman correlation between the experimental ranking and the ranking produced by COMPETE is then computed under two settings: one in which COMPETE models nucleosomes to bind the genome in isolation, and one in which it models them to bind in competition with 89 TFs whose *in vitro* specificities are known (Zhu et al., 2009). The TF concentrations in the competition setting are globally scaled by multiples of their respective K_d, as discussed above. The improvement in Spearman correlation under the two settings for various global scalings and various choices of threshold parameter t is computed in each case by subtraction.

3.5 Discussion

The extended 'musical chairs' contest view of DNA occupancy represents a fundamental shift in how DNA binding is typically modeled, while conforming to how it is currently understood. The simplifying assumptions often made in the past, such as discrete binding, have not harmonized well with our understanding of the

thermodynamic nature of molecular interactions. By viewing binding as dynamic, probabilistic, and competitive, we move closer to our current understanding of the process, as revealed by numerous recent studies.

In particular, comprehension of binding behavior at an aggregate scale will facilitate mapping of binding to expression. Expression is a complex continuous signal resulting from many kinds of input, and is difficult to explain on a large scale by a small number of strong binding targets alone (Tanay, 2006). Capturing pervasive weak binding instances may aid in our understanding of this control, allowing discovery of contributions of transcriptional regulators for genes whose involvement has been hitherto unknown.

Some improvements may be made to our approach that might boost its accuracy and effectiveness without fundamentally altering its structure or applications. The largest immediate gain may be made by developing a more robust technique for determining appropriate values of DBF concentrations. Our concentration parameters are in fact a product of a true concentration and an unknown energetic constant unique to each DBF. Binding motifs (including those we use) are usually represented by position-specific scoring matrices (PSSMs), and each PSSM is normalized to a probability by a normalization constant. This normalization constant scales energies into probabilities. Experimentally recovering the normalization constants so as to convert the probabilities of PSSMs into energies is prohibitive.

Hence we are developing a procedure to learn these constants automatically, though doing so is nontrivial in the absence of sufficient informative data, and we present an early version of this procedure in chapter 4. In the examples we presented, concentration parameters were chosen to recapitulate expected binding behaviors; namely, strong binding at literature annotated binding sites without overwhelmingly strong binding elsewhere. A suitable automated technique would attempt to likewise fit concentrations in such a way as to recapitulate existing nucleosome oc-

cupancy maps as well as annotated TF binding locations. As discussed elsewhere (Tanay, 2006), ChIP-chip data contains information throughout the range of p-values and may be helpful to guide TF concentration parameter selection. Additionally, a number of experimentally determined or computationally predicted nucleosome positioning data sets exist at both genome-wide and locus-specific resolutions (Kaplan et al., 2009; Shivaswamy et al., 2008; Whitehouse et al., 2007; Lee et al., 2007; Segal et al., 2006). Training on nucleosome data is difficult given the noisy nature of existing nucleosome positioning data in conjunction with the mobility of the nucleosomes themselves. However, taking nucleosome and ChIP-chip data into account together may help to mitigate much of this uncertainty and yield the robust estimates of concentration that we desire.

Our modeling framework would allow wholly different methods of representing DBFs to be employed. As the various DBFs are themselves modules of states, their specific implementation need not be Boltzmann chains (see Methods). Their replacement may be quite extreme, and our model could easily be converted into a generalized Boltzmann chain. For example, our current nucleosome model could be replaced with any of a number of existing models of nucleosome occupancy (Yuan and Liu, 2008; Field et al., 2008; Peckham et al., 2007). When considering this type of modification, it should be noted again that models of *in vitro* sequence specificities are more appropriate than those obtained from *in vivo* data. Our approach focuses on understanding the resultant genomic binding occupancy given a set of DBFs and their inherent sequence specificities. *In vitro* models are likely to be more reflective of inherent specificity (without the influence of competition and other positioning pressures) than *in vivo* models. Indeed, our approach may be understood to be attempting to understand (and output) *in vivo* behavior in terms of *in vitro* inputs. For this reason, the sequence preferences used in our analyses are all *in vitro* where available.

Because our approach yields a full posterior distribution, we can sample from this distribution to answer increasingly complex questions. For example, it is possible to determine the frequency with which a given number of DBFs occupy sequence in a particular orientation, either with respect to one another or specified sequence elements, such as transcription start sites or TATA boxes.

More generally, our model of competitive multi-factor binding will be able to guide biological experiments by providing testable hypotheses. Mutations may be introduced to both sequence and DBFs in our model to predict novel and experimentally verifiable binding behavior. Experiments such as these may help elucidate the true roles of TFs whose functionality is not entirely known, or disambiguate the mechanism by which these roles arise. As one such example, we can test whether a set of TFs that are known to have a mutually cooperative relationship achieve this cooperativity through explicit interactions with one another or implicit cooperativity achieved via nucleosome displacement.

4

Learning biologically relevant values for model parameters

Given that DNA binding is a stochastic process, DNA occupancy models must incorporate various types of modeling decisions to properly represent and account for this stochasticity. At a high level, these choices can be described as pertaining to either the structure of a model, or to its parameter values. In chapters 2 and 3 we present such a model and focus primarily on its structure; in this chapter we explore how to make sensible choices for its parameters. We focus specifically on determining how to choose concentrations of transcription factors that make sense biologically. In particular, we consider "plausible" concentration values to be those that are neither absurdly high (i.e., result in a TF completely dominating genomic binding) nor absurdly low (i.e., result in a TF not binding the genome at all). Determining concentration values is a difficult problem: it has been recently shown that even within single promoters with as few as two strong TF binding sites, many different dynamic and stochastic behaviors can emerge as a function of varying TF concentrations entirely *within* the biologically plausible range of values (Coulon et al., 2010).

One obvious approach to determining plausible TF concentration levels is to directly measure the concentrations of the proteins (TFs) of interest within the cell. However, this type of data is not widely available for many types of transcription factors, for a number of reasons. Performing these measurements is extremely difficult due to the necessity to gather measurements from just a single cell, and without the benefits of the type of amplification techniques available in, for example, gene expression microarray experiments; there is no simple analogue for proteins to PCR amplification of mRNA. This is compounded by the fact that the value actually desired is the concentration of *active* protein, and specifically only the active protein localized to the nucleus. Furthermore, it is not clear just how to map physiological concentrations to the concentration parameters required for most models even if they *were* gathered, and beyond that, measurement of such a concentration may introduce enough noise that it would impact the resulting measurements enough to significantly reduce their utility even if all of these other problems were solved.

Gene expression data *is* widely available, however. Many approaches use mRNA expression data as a proxy for protein level because it indicates the activity of a gene. This is problematic, though, because expression is an indication of transcription, and protein concentrations are a result of translation and myriad other factors. To clarify, at a high level, there are two primary problems with the assumption of a strong relationship between transcription and translation: first, the quantity of transcript does not necessarily correlate to the rate of protein production, and second, the rate of protein production does not necessarily correlate to the concentration of active protein in the nucleus.

Because the concentration values cannot be reliably measured, our focus turns to inferring them. We are far from alone in this dilemma: the difficulty of choosing parameters for models in systems biology in particular is well-known within the field, but it may not actually be much of a problem in many predictive models. Gutenkunst

et al. (2007) showed that the issue of difficult-to-measure parameters in models in systems biology is not only ubiquitous, but that these models furthermore share the characteristic that upon inferring the parameters in question it becomes apparent that the parameters have a "sloppy" spectrum of sensitivities. The message is clear: for numerous reasons, it makes more sense to focus efforts on satisfactory predictive outputs of a model, even if they imply "sloppy" parameter choices, instead of on perfecting parameter selection of inputs to the model.

When attempting to optimize parameters, the community has embraced this approach (either explicitly or implicitly), but there remain many studies related to our focus in which the parameters are simply not optimized at all. Past approaches to set TF concentrations of models of either binding or transcriptional regulation have involved manually picking parameter choices, without automated tuning (Bintu et al., 2005b; Segal et al., 2006, 2008; Raveh-Sadka et al., 2009; Teif, 2010).

However, many studies have made attempts to infer concentration parameters. Recently, various advances have been made in determining concentration parameters for transcription factors in assorted types of modeling paradigms. Many approaches have attempted to tune parameters by way of only considering the effect of one type of DBF, for example nucleosomes in the case of Lubliner and Segal (2009) or transcription factors in the case of Sinha (2006). Gao et al. (2008) have developed a technique for quickly determining TF concentrations via Gaussian processes, but these results apply specifically to models that represent rates of transcriptional activation and repression via differential equations considering only a predetermined set of TFs. Fernandez-Lopez et al. (2010) have a method focused more on experimentally exploring concentration effects than numerically exploring them, although it is also targeted for a similar model, built around differential equations. Liu (2010) uses a gradient descent method to infer the concentrations as part of a Bayesian network, but again, the application is very much model specific. The application-specificity is

not necessarily a bad thing; the problem is difficult enough to address in a specific paradigm, let alone solving it in the general case.

In summary, proper concentration values for our modeling framework cannot feasibly be determined through biological experiments, and there presently is no widely-applicable general purpose technique for automatically determining TF concentration from gene expression or ChIP-chip microarray data. Our concentration parameters must therefore be inferred, and done so in a way tailored to our specific modeling framework. This inference should be concerned in large part with the effects of parameter choices on the resulting predictions of our model and their relation to observed biological behavior. In this chapter, we present such an inference technique.

Specific parameter-fitting considerations with respect to our model

Akin to the above frameworks, inferring concentration parameters for our model requires a customized approach. Often with hidden Markov models, parameter optimization is performed with Expectation-Maximization via the Baum-Welch algorithm. We choose not to follow this approach for three primary reasons. First, because our model is actually a Boltzmann chain, exact E-M updating equations cannot be calculated and hence Baum-Welch does not apply. Second, even if we use another approach to approximate the E-M updating steps, this general approach is sensitive to local optima and hence initialization values for our parameters; this is problematic with our model and we would do better to use an approach that can escape local optima.

Third, and most important, is that Baum-Welch (and any analogue) is maximizing the wrong function for our purposes. The likelihood function used in Baum-Welch is maximized when the input sequence is best explained by our model structure and parameters. It implicitly places much importance on the resemblance of the model

structure to the actual underlying phenomenon that generated the sequence, and disregards other outside information. In our specific scenario, this is problematic. We recognize that our model is substantially incomplete if taken in comparison with our knowledge of all of the underlying factors involved in guiding the resultant DNA occupancy we observe in the cell, but our goal is to learn what we can from just this simple model. Additionally, we have large amounts of other revealing data about DNA occupancy, such as ChIP-chip data of transcription factor binding and nucleosome occupancy maps, and it makes more sense to use this type of data as a standard against which our model's predictions are measured.

We choose to use a scoring metric that takes into consideration these experimental data sets. Because of our lack of knowledge about the structure of the space of the concentration parameters with respect to this function, we focus on a *Markov chain Monte Carlo* inference technique known as the *Metropolis-Hastings* algorithm.

4.1 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) algorithms allow sampling from complex and often high-dimensional probability distributions. There is much statistical theory underpinning MCMC, detailing mathematical characteristics and requirements of different aspects of the framework, but here we focus primarily on a high-level discussion of its use in practice. For a detailed discussion, see Gamerman and Lopes (2006).

MCMC algorithms are Markovian in the sense that each draw from the distribution of interest depends only on the value of the previous draw. That is, for random quantities x in a p-dimensional state space $\mathcal{X}$ and a transition probability distribution $P(x|x')$,

$$x^{(t)} \sim P(x|x^{(t-1)}) \text{ with } x^{(t)} \perp\!\!\!\perp x^{(t-k)}|x^{(t-1)}, (k > 1) \tag{4.1}$$

Letting $p(x|x')$ be the density of $P(x|x')$, then beginning with an initial value $x^{(0)}$, the vector x is constructed by iterative sampling, for $t > 0$, of:

$$x^{(t)} \sim p(x|x^{(t-1)}) \tag{4.2}$$

When constructed appropriately, $p(x|x')$ will converge to the desired (*target*) density over time.

4.1.1 The Metropolis-Hastings algorithm

The Metropolis-Hastings (M-H) algorithm is an MCMC method for sampling from target distributions under which points can be evaluated relatively easily, but direct sampling is difficult. The algorithm is constructed as follows:

- $\pi(x)$ is the p.d.f. of the target distribution on $\mathcal{X}$
- $x^{(1)}, \ldots, x^{(t)}, \ldots$ will be generated given $x^{(0)}$
- $g(x|x')$ is the p.d.f. of the *proposal* distribution for all $x, x' \in \mathcal{X}$
- At step $t-1$, the current state is $x^{(t-1)}$. A *candidate state* x^* is generated from the proposal distribution such that $x^* \sim g(x^*|x^{(t-1)})$.
- Calculate $\alpha_t = \alpha(x^{(t-1)}, x^*)$, where

$$\alpha(x', x) = \min \left\{ 1, \frac{\pi(x)}{\pi(x')} \cdot \frac{g(x'|x)}{g(x|x')} \right\} \tag{4.3}$$

- The chain now moves to the proposed point x^* with probability α_t, otherwise remaining in $x^{(t-1)}$. Formally,

$$x^{(t)} = \begin{cases} x^* & \text{with probability } \alpha_t \\ x^{(t-1)} & \text{otherwise} \end{cases} \tag{4.4}$$

4.2 Our sampler

4.2.1 Score function

For construction of our score function, we refer to the analysis of transcription factor positioning improvement under full models (including all TFs and nucleosomes) versus simpler, or null models (including only one TF) described in section 3.4.7. The measure described in that analysis is an indicator of the satisfactory recapitulation of known measurable biological phenomena (specifically, transcription factor binding) by our model with a given set of TF concentrations. We propose using an extension of this metric as a score of a set of concentrations, which we seek to optimize via M-H. In this section we more formally define the algorithm and metric used in the previous analysis, as it will provide the basis for our score function going forward.

We define the following variables:

T : vector of TFs

θ : vector of concentration of TFs T in units K_d

m_F : full model with all TFs and nucleosomes

$m_{0,i}$: null model with only TF i

P_i : vector of ChIP-chip p-values of TF i for all probes

$P_{i,r}$: ChIP-chip p-value of TF i in probe r

R : vector of ChIP-chip probes

One comment about notation: due to the nature of the intermingling of mathematical and algorithmic notation and convention necessary to fully describe our approach, consider notation such as $P_{i,r}$ and `P(i, r)` to be equivalent.

We define the following utility functions:

`C(m, theta, i, r)` returns the vector of COMPETE probabilities under model m with concentrations θ of beginning TF i at each position across genomic re-

gion r.

`ROC(a, b)` constructs the ROC curve for list of means a and list of p-values b as described in section 3.4.7.

`AUC(r)` calculates the area under the ROC curve r.

Let the function `A(m, theta, i)` for model m with concentrations θ and TF i be defined by the following pseudocode:

```
function A(m, theta, i)
begin
  let means = vector(length(R))
  for each probe in 1 to length(R) do
    means[probe] = mean(C(m, theta, i, R[probe]))
  next probe

  a = AUC(ROC(means, P(i)))
  return a
end
```

Now the difference in area under the curve under the full model and the null model is:

$$D(i) = A(m_F, \theta, i) - A(m_{0,i}, \theta, i) \tag{4.5}$$

For clarity, $D(i)$ is the function described in section 3.4.7 and illustrated in Figure 3.10.

Finally, our score function can be defined in terms of sum of these differences for all TFs:

$$S(\theta) = e^{\sum_i D(i)} \tag{4.6}$$

The summing of the scores of the $D(i)$ for each factor is done to reflect the contribution of each TF to the overall "goodness" of the metric. In this way, a concentration that benefits one TF at the detriment of others can yield an overall net score decrease. We implicitly do not consider any one factor to be more important than the others with this construction, seeking to maximize the group performance at the potential cost of individual members' maximal performance.

We exponentiate the sum because $D(i)$ (and hence $\sum_i D(i)$) may be negative. This is a problem because of the structure of the Metropolis-Hastings rule (Equation (4.3)), which expects non-negative values of $\pi(x)$ and $\pi(x')$. By exponentiating, we still maximize $D(i)$ but avoid issues associated with its potential negativity.

4.2.2 Proposal distributions

Given the score function describing how to evaluate a given move through the parameter space, we now discuss the proposal distribution for generating moves. Because we know relatively little about the covariance structure of the model parameters, we cannot do blocked sampling. Hence, we sample concentrations one-at-a-time.

We have a relatively poor understanding of the shape of the distribution of individual concentrations, though we know the support of the distribution should be non-negative. Hence, we use log-normal distributions for proposals. Additionally, we dynamically resize the standard deviations of the proposal distributions to maintain a target acceptance rate; in practice, this target rate is $\frac{1}{3}$.

In addition to the variables defined above, we define the following variables:

σ_i : standard deviation for proposal distribution for TF i

r : target acceptance rate for proposal distribution

n : number of full iteration rounds between adjustment of σ_i

$n_{a,i}$: number of acceptances for TF i in last n iterations between adjustment of σ_i

A practical consideration is that the M-H rule is slightly simplified in the case

Table 4.1: Transcription factors and ChIP-chip experiments for parameter inference.

TF	ChIP-chip experiment
Abf1	ABF1_YPD
Fhl1	FHL1_YPD
Fkh1	FKH1_YPD
Fkh2	FKH2_YPD
Gat1	GAT1_YPD
Mbp1	MBP1_YPD
Mcm1	MCM1_YPD
Met32	MET32_YPD
Rap1	RAP1_YPD
Ume6	UME6_YPD
Yox1	YOX1_YPD

of symmetric proposal distributions. Because of this, we implement our log-normal proposal distributions by proposing log-concentrations under a normal distribution.

Our proposal distribution for TF i is:

$$\log(\hat{\theta}_i) \sim N\left(\log(\theta_i), \sigma_i^2\right) \tag{4.7}$$

where θ_i is the previous concentration and $\hat{\theta}_i$ is the proposed new concentration.

Every n full iteration rounds (i.e., updates of all factors), σ_i is rescaled as follows:

$$\hat{\sigma_i} = \sigma_i \cdot \frac{n_{a,i}/n}{r} \tag{4.8}$$

This expression scales σ_i by the ratio of the actual acceptance rate to the desired acceptance rate. For example, if the actual acceptance rate is twice that of the desired acceptance rate, the standard deviation is doubled, causing proposed moves to be drawn from further away and hence increasing the chance of rejection.

4.3 Results

For this analysis, we focus on a set of transcription factors and genes thought to play significant regulatory roles in the yeast cell division cycle. The set of TFs was

Table 4.2: ChIP-chip probes for parameter inference.

Chromosome	Probe ID
III	iYCR063W
III	iYCR065W
IV	iYDR451C
IV	iYDR500C
IV	iYDR501W
IV	iYDL057W
IV	iYDL056W
IV	iYDR146C
IV	iYDR207C
IV	iYDR253C
V	iYER111C
VI	iYFL022C
VI	iYFL021W-0
VI	iYFL021W-1
IX	iYIL131C
XI	iYKL113C
XI	iYKL186C
XI	iYKL185W
XII	iYLR131C
XII	iYLR181C
XII	iYLR182W
XII	iYLR183C
XIII	iYML028W
XIII	iYML027W
XIII	iYMR042W
XIII	iYMR043W
XIV	iYNL217W
XIV	iYNL216W
XIV	iYNL104C
XIV	iYNL103W
XIV	iYNL068C
XIV	iYNL309W
XIV	iYNR008W
XIV	iYNR009W
XV	iYOR372C
XV	iYOR083W
XV	iYOR327C
XVI	iYPR104C

Table 4.3: Optimized concentrations and improvements corresponding to Figure 4.1. The average improvement, measured by $D(i)$, is 0.03167901. Note the extreme values of several of the resulting concentrations.

TF	K_d	$D(i)$
ABF1	1.301387e+11	0.0177519118308591
FHL1	6.040883e+04	0.152741228070176
FKH1	3.955875e+42	0.0227477716468589
FKH2	3.367888e+18	0.116127351750417
GAT1	9.835577e+81	-0.0048029102570695
MBP1	4.054059e+73	0.179128289473684
MCM1	2.573102e+14	0.076327537593985
MET32	9.972664e-01	0.0050497866287341
RAP1	9.832629e-01	-0.0970128022759596
UME6	3.679131e+54	0.0055863426451662
YOX1	3.013229e-06	-0.125175438596491

intersected with the set of PBM-derived PSSMs available, resulting in those given in Table 4.1. The set of probes was determined by first taking all cell cycle regulated genes and concatenating the list of TFs from Table 4.1. Given this larger set of genes, the probes were selected by comparing literature-curated genomic coordinates (adjusted for our version of the yeast genome) with the coordinates from the ChIP-chip probes (likewise adjusted). While generally these coordinates agreed, there were occasionally mismatches. In this case, both probes were included under the rationale that including spurious probes is unlikely to pose a major detriment in overall performance of the inference technique. The resulting probe set is given in Table 4.2.

4.3.1 Optimizing transcription factor performance only

Running our M-H sampler as described above yields improvement in the score (Equation (4.6)) as illustrated in Figure 4.1. The score function varies quite a bit (Figure 4.1A), and we focus on the maximum value, denoted by a hashed vertical line.

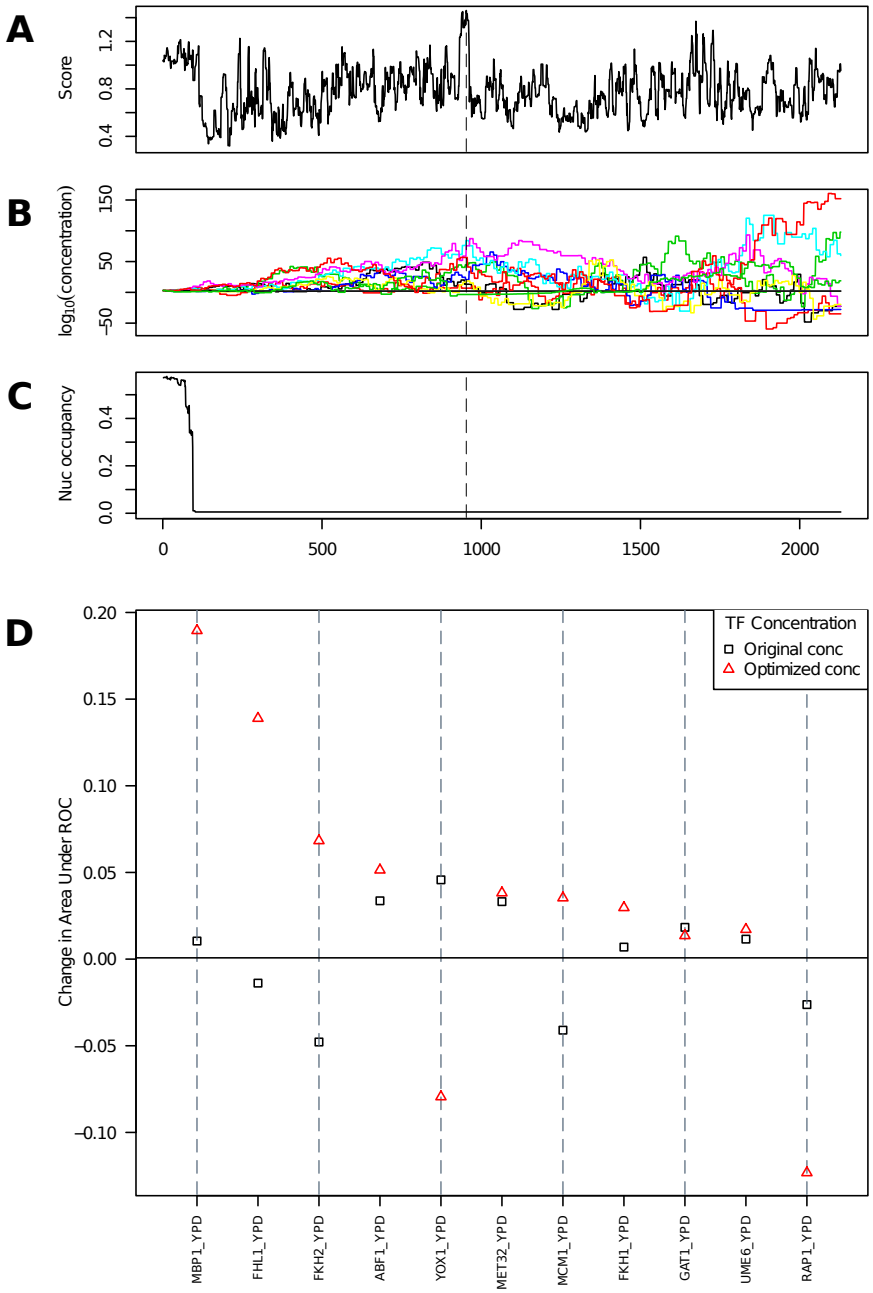

FIGURE 4.1: Optimized concentrations vs. original concentrations. (**A**) The score across iterations. We are only interested in the maximum score, marked with a vertical hashed line. (**B**) Concentration parameters over iterations. Note that with no constraints against it, concentrations are free to wander to unrealistic levels. (**C**) Nucleosome occupancy over iterations. It quickly drops to zero due to the lack of constraint on biological plausibility. (**D**) Analogous to Figure 3.10A, TFs with optimized concentrations give generally better predictions as compared to ChIP-chip data than the same TFs at the initial concentrations ($8K_d$). However, nucleosome occupancy at these high TF concentrations is almost completely eradicated.

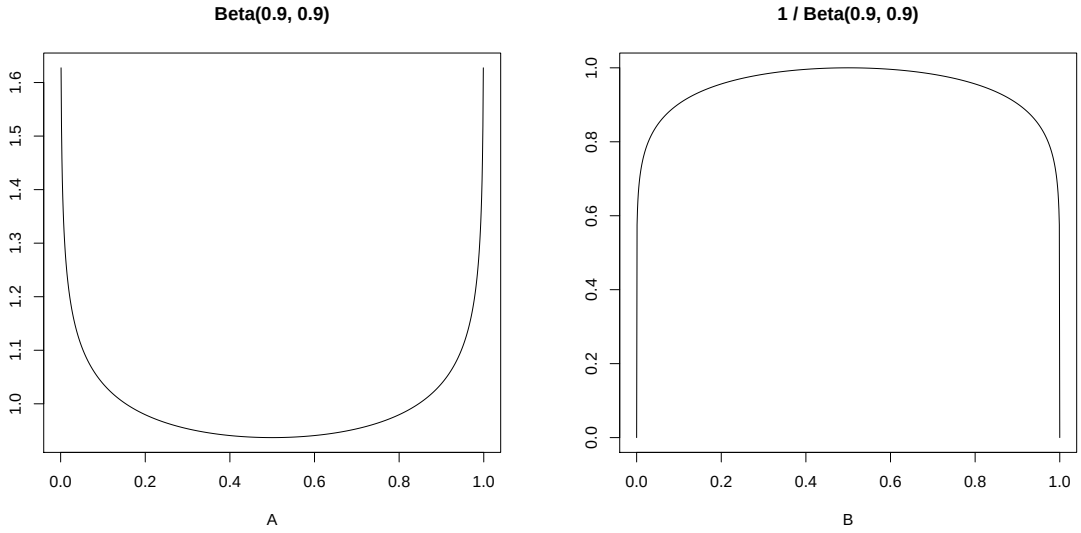

FIGURE 4.2: Nucleosome penalty function. (**A**) The density of a Beta(0.9, 0.9) distribution. This function conveniently suits our needs of existing in $[0, 1]$ with extreme values toward 0 and 1, but tends toward infinity at both. (**B**) Hence, we use 1/Beta(0.9, 0.9) and normalize it to peak at 1. This becomes our penalty function, with the function being evaluated at the point of nucleosome occupancy under a current parameter set and used to scale the score, as depicted in Equation (4.9).

At this point, almost all of the TFs in the model show improvement over their initial concentrations, which are all set to $8K_d$ (Figure 4.1D). The concentrations and individual scores for each TF ($D(i)$) resulting from this optimization are given in Table 4.3. Even though three factors actually perform worse, the aggregate performance outweighs their diminished values, with an average improvement of 0.0317.

Unfortunately, the score as formulated contains no penalty against biologically implausible concentrations. Concentrations are free to meander, and tend to stray to extreme values (Figure 4.1B). As a result, nucleosome occupancy is quickly quashed to near-nonexistence (Figure 4.1C). This characteristic is undesirable, so we sought to address it in a simple way, as described below.

4.3.2 Penalizing against biological implausibility

Because we want concentrations that make sense in a biological setting, we introduce a penalty against extreme nucleosome occupancy. We know that nucleosomes *in vivo* occupy up to 90% of the genome at any given time. However, existing computational models, whether multi-factor or nucleosome-only, have had difficulty achieving satisfactory performance of both nucleosome and transcription factor positioning simultaneously. Indeed, peak performance in the nucleosome-only study depicted in Segal et al. (2008) took place with predicted nucleosome occupancy at approximately 30% genome-wide. Our own work, as depicted in Figure 3.11, peaks at approximately 40% to 50% nucleosome occupancy. This is not to say that controlling for nucleosome occupancy cannot provide utility, though; this merely means that these models are imperfect, and that we must be reasonable with respect to construction of a property penalty function. Hence, we compose a penalty that is not so encouraging *toward* a particular nucleosome occupancy (like 90%), but one that is discouraging *against* extremely low or high nucleosome occupancy.

We desire a penalty function to scale our existing score, so the range of the function should be $[0, 1]$. The function should drop to 0 near the endpoints of its range, and be somewhat ambivalent to ranges in between. A convenient candidate with behavior related to this is the Beta distribution density function, as depicted in Figure 4.2. Although the Beta(0.9, 0.9) density acts in reverse to our desired behavior (Figure 4.2A), this is easily correctable by taking its reciprocal and normalizing it (Figure 4.2B).

To leverage the penalty, given a set of concentrations we set k to be the normalizing constant, $k = f_{Beta(0.9,0.9)}(0.5)$, and calculate a new value, $\gamma =$ percentage

Table 4.4: Optimized concentrations and improvements corresponding to Figure 4.3. The average improvement, measured by $D(i)$, is 0.02024953. Although this is worse than the optimization without constraints, the range of concentration values is much less extreme.

TF	K_d	$D(i)$
ABF1	1.107915e+01	0.0078347578347576
FHL1	9.610127e+02	0.0645833333333334
FKH1	7.419857e-03	-0.0130929791271348
FKH2	4.663677e-02	0.0600961538461539
GAT1	1.630020	-0.006231671554252
MBP1	2.457867	-0.00856335532746821
MCM1	1.182537	0.0202727733195614
MET32	1.410896	0.0094594594594595
RAP1	1.020012	0.0394736842105264
UME6	8.572080e-01	0.0089126559714796
YOX1	2.442437e+02	0.04

nucleosome occupancy across the genome. Our new score function becomes:

$$\begin{aligned} penalty =& \frac{1/f_{Beta(0.9,0.9)}(\gamma)}{k} \\ S(\theta) =& e^{\sum_i D(i)} \cdot penalty \end{aligned} \tag{4.9}$$

Given this new score function, our approach yields positive results again (Figure 4.3, Table 4.4). Although the increase in individual ROC AUC is less dramatic, it is still systematically improved with an average increase of $D(i)$ of 0.0202 (Figure 4.3D, initial concentrations set to $8K_d$), but while maintaining approximately 42% genomic nucleosome occupancy (Figure 4.3C).

Controlling against extreme concentrations

Constraining against extreme nucleosome occupancy improves our parameter selection, but our approach remains sensitive to deleterious effects of accepting extreme parameter values, with large values posing particular issues. Specifically, because of the nature of our one-at-a-time approach to sampling concentrations, the inevitable

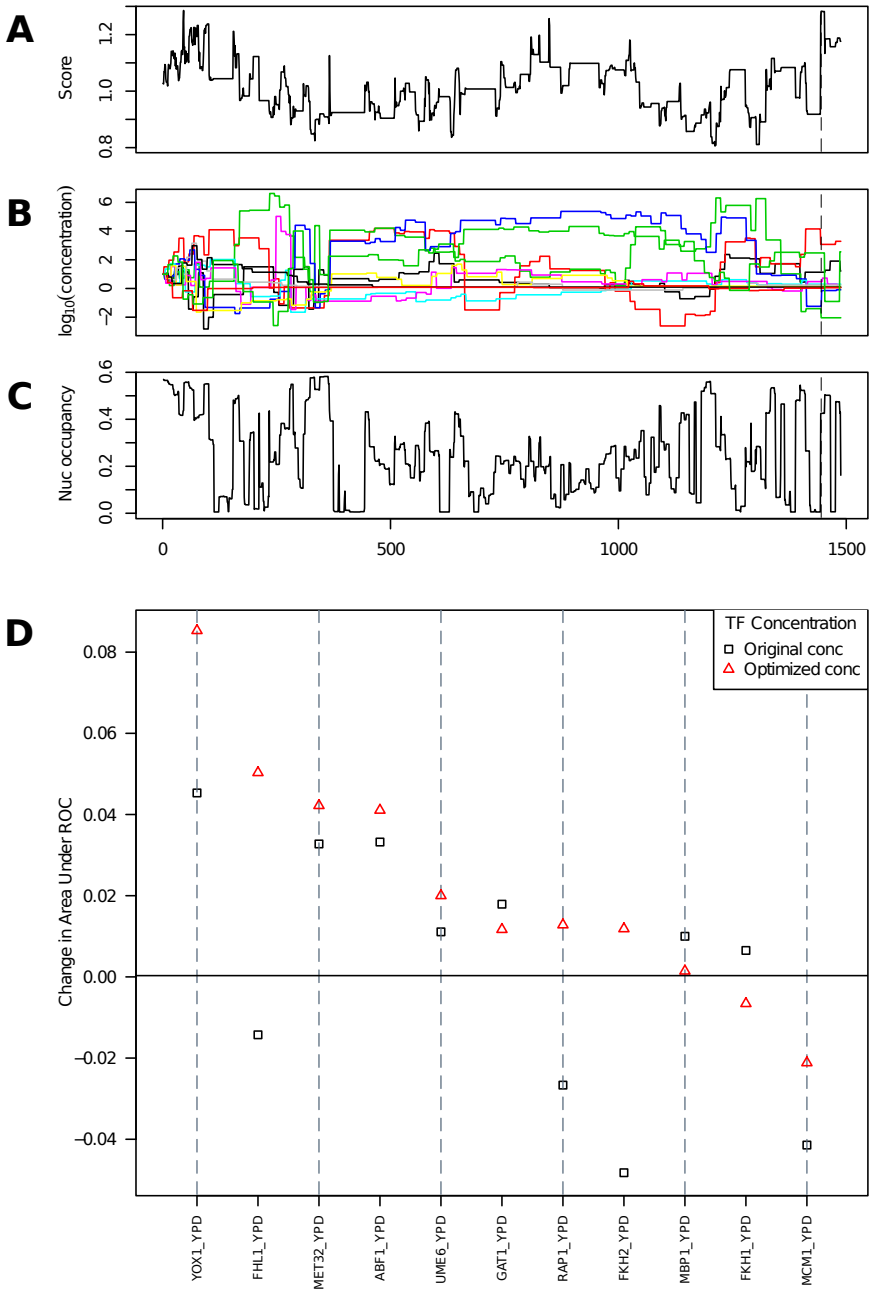

FIGURE 4.3: Optimized concentrations vs. original concentrations, with nucleosome occupancy constraints. (**A**) The score across iterations. As in Figure 4.1, we are only interested in the maximum score, marked with a vertical hashed line. (**B**) Concentration parameters over iterations. With constraints in place to discourage extremely low nucleosome occupancy, concentrations do not wander nearly as far over iterations. (**C**) As a result, nucleosome occupancy remains in a more reasonable range over iterations. (**D**) Compared to Figure 4.1D, increases in AUC difference are more modest. However, the nucleosome occupancy at the point depicted here is approximately 42%, where in the previous figure it was virtually 0. Initial concentrations are again set to $8K_d$.

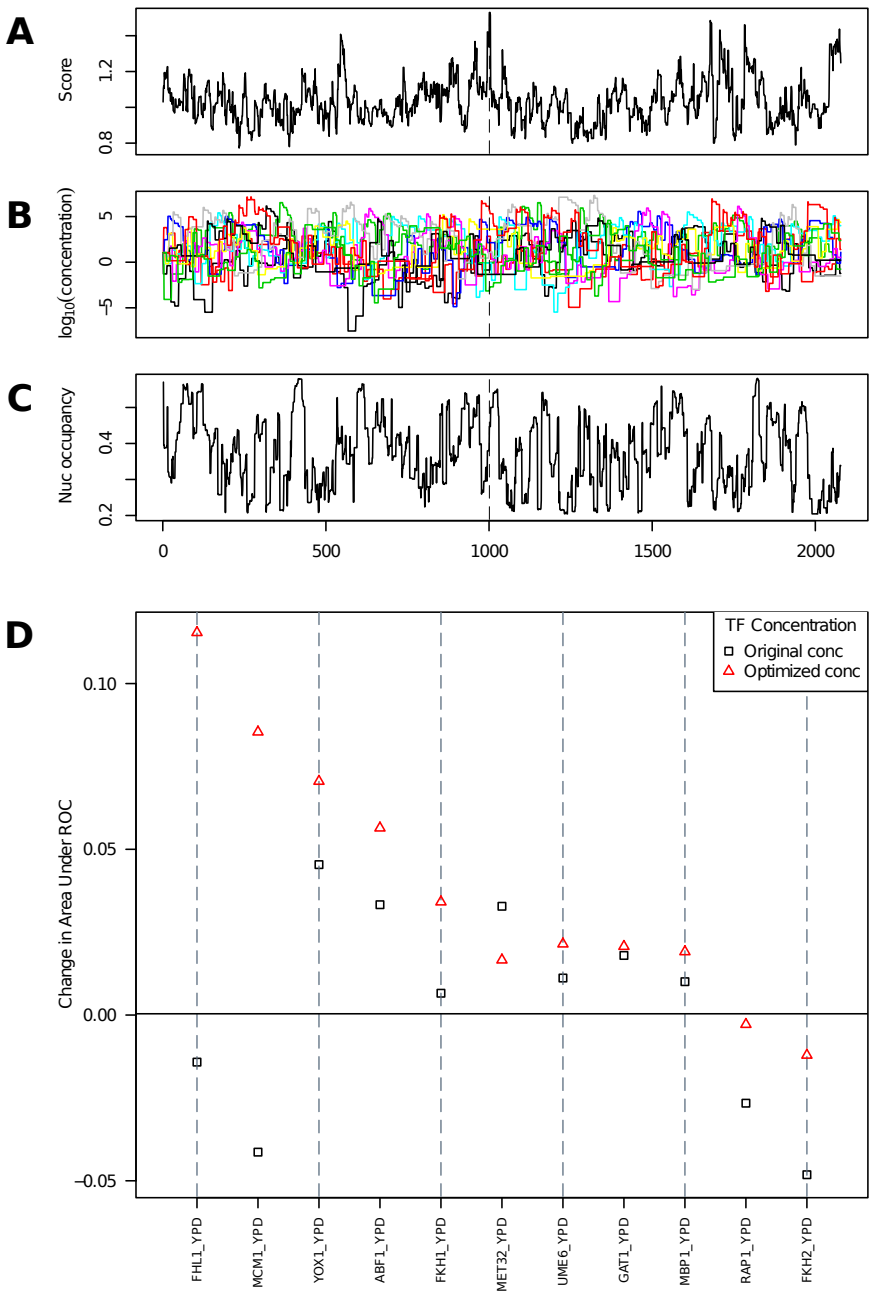

FIGURE 4.4: Optimized concentrations vs. original concentrations, with nucleosome occupancy constraints and controlling for extreme parameter values. (**A**) The score across iterations. (**B**) Concentration parameters over iterations. Explicitly bringing extreme concentrations back toward more reasonable values results in concentrations that still vary quite a bit (as opposed to the lengthy flat regions denoting rejected proposals in Figure 4.3B) but do so within a constrained range. (**C**) Nucleosome occupancy. (**D**) AUC improvements are still systematically present. Initial concentrations are again set to $8K_d$.

Table 4.5: Optimized concentrations and improvements corresponding to Figure 4.4. The average improvement, measured by $D(i)$, is 0.03615779. This is better than either of the previous optimization results, but in this case is also achieved with concentration values of a reasonable magnitude.

TF	K_d	$D(i)$
ABF1	1.011798e-01	0.0231356837606838
FHL1	5.042924e+02	0.129583333333333
FKH1	6.237803e-01	0.0275388291517324
FKH2	1.001596e+02	0.0360576923076924
GAT1	5.789939e-01	0.0027171920821114
MBP1	8.162280	0.00902217741935504
MCM1	2.887178e+03	0.126741071428571
MET32	9.461147	-0.0162162162162161
RAP1	3.346840e-02	0.0237593984962407
UME6	2.994025e+05	0.0102683178534571
YOX1	5.078556e-04	0.0251282051282052

acceptance of a single large parameter value can lead to easing of the constraints against other large parameter values. When one value is extremely large, the resulting score of the parameter set is quite low. Even though the resulting set is usually rejected, eventually one of these low scoring parameter sets will be accepted by the probabilistic nature of the Metropolis-Hastings criterion. Once this has occurred, further extreme parameter choices often result in scores that are not much different than this already poor score, and are therefore often accepted. Not until a full round-robin iteration of all parameter sampling has been completed can the original parameter be adjusted, and by this time it may be the case that sufficiently poor choices have been accepted in other parameters to lead to a particularly poor portion of the parameter space without the ability to quickly escape it. A simpler way of describing this is that due to the nature of our approach, a logical "or" of poor parameter choices is required to enter a poor portion of the parameter space but a logical "and" of good parameter choices is required to escape it.

An especially satisfying way of addressing this problem would be to jointly sample all concentrations at each step, but due to the lack of clarity regarding their covariance structure this is not currently feasible. Instead, we address the issue by attempting to proactively reign in large concentration choices. Each full iteration of all parameter choices, we propose halving or doubling (in linear space) all extremely large or small parameter values, respectively, accepting this move with the same Metropolis-Hastings rule as for other proposed moves through the space. Calling extreme values those above some threshold t or below $-t$, we achieve the results illustrated in Figure 4.4 and detailed in Table 4.5, for $t = 10^4$, with initial concentrations again set to $8K_d$. The resulting genomic nucleosome occupancy is approximately 44%, and the average increase in $D(i)$ is 0.0362.

4.4 Further analysis

Although the purpose of this sampling technique is not to explore the posterior distribution of the concentration parameters, but instead to find high points within this distribution, it nevertheless can be improved through further constraints on movement through the space. In particular, although the current techniques penalize against high concentrations, they do not do enough to control the arbitrary decreasing of concentrations. Any single high concentration will displace nucleosomes, and therefore be discouraged by the nucleosome occupancy penalty; thus the difference in scores between a parameter set containing a single value of $10^4 K_d$ and one containing $10^7 K_d$ is extreme. Because extreme lowering of concentrations is not explicitly discouraged, the difference between sets containing values of $10^{-4} K_d$ and $10^{-7} K_d$ is virtually nonexistent.

Additionally, when considering what constitutes an acceptable set of parameters, further validation is desirable. In some sense, the metric we maximize in the approach described above provides confidence in our parameter selection, in that the

measure we maximize is indicative of the ability of our model to generate occupancy predictions in line with experimentally observed phenomena. However, as touched upon in our discussion of whole-genome validation of our model, good transcription factor modeling choices should lead to benefits in nucleosome positioning predictions as well. Hence, a validation step involving the effect of optimization of transcription factor concentrations on nucleosome positioning predictions is a logical next step, with a reasonable starting point being the type of analysis we performed in section 3.4.8.

5

Future directions

Our approach as it currently exists is powerful, flexible, extensible, and high performance. However, our work is intended as just an early step in the effort to better understand DNA occupancy and its role in processes it affects. Moving forward from here, there are two primary avenues of future work, which are complementary to one another: improving upon our model, and applying our model to novel applications.

5.1 Improving our model

5.1.1 Improvements in the theory underlying our model

We break the following ideas down into groups relating to transcription factors and relating to nucleosomes, but because of the construction of our model an improvement in any one type of DBF prediction should lead to consequential improvements for other DBFs as well.

Transcription factor related improvements

Our TF binding preferences are presently modeled via PSSMs, which are imperfect binding representations in several ways. PSSMs make the assumptions that a

transcription factor occupies exactly a fixed number of nucleotides, and that the contribution of each position in the binding pocket is independent of the other positions. It is well-understood that these assumptions are faulty (Bulyk et al., 2002; Berger et al., 2006), but they have been serviceable simplifications (Benos et al., 2002) that in the past have contributed to useful and interesting studies. However, there exist more complex models that may yield more accurate representations of sequence affinity than PSSMs (Agarwal and Bafna, 1998; Barash et al., 2003; Zhou and Liu, 2004). The implementations vary, but share the characteristics that they overcome the independence and fixed-width assumptions. Through leveraging variants on HMMs, discussed later in this chapter, any mixture of these models could be introduced to our framework in place of our current TF models.

Regardless of the representation of sequence preferences, our description of a DBF by only the nucleotides it physically binds implies steric hindrance only at the site of the binding pocket itself. Many TFs are known to be bulky and preclude occupancy of nucleotides adjacent to them. A similar approach to our current padding of nucleosomes with unbound sequence can allow for this sort of hindrance, but construction of the padding would need to be custom-tailored for each TF.

Finally, better quality binding data is being generated all the time. High-throughput SELEX experiments have recently been used to infer *in vitro* TF binding energies from large numbers of different sequences (Zhao et al., 2009). This HT-SELEX data is similar to the PBM data we use in our work, in that it describes the inherent sequence preferences of the TFs in the absence of other molecules, but it can determine binding site characteristics of sites up to 30 base pairs in length, as compared to the 10 base pair limit of PBM techniques. More accurate information about binding preferences and behaviors from new data sets such as this can only improve our approach.

Nucleosome related improvements

Inclusion of nucleosomes into our competitive model of DNA binding yields clear improvements in transcription factor positioning predictions, but there are several additions that may further benefit the model. As depicted in Figure 1.3, nucleosomes are but the first level of a hierarchy of chromatin compaction. DNA densely packed into higher levels of compaction is largely inaccessible to transcription factors; indeed, this is one of the mechanisms that drives gene silencing. Related to both higher and lower-level chromatin compaction is the separate notion of histone modifications and histone variants. While we do not account for these, they are known to play roles in preventing higher-level compaction and show correlation with DNA free regions, such as gene promoters. Explicitly accounting for higher-level compaction and/or histone modifications through additional states of our model may better reflect overall DNA occupancy.

A separate issue is that linker regions between nucleosomes have a known distribution (Wang et al., 2008), which could be explicitly incorporated into the model via conversion to a hidden semi-Markov model (HSMM) (Murphy, 2002b), as mentioned further in section 5.1.2.

Explicit context-dependent up- or down-weighting terms

Our model as currently posed considers genomic occupancy to be a result strictly of competition amongst mutually exclusive DNA binding proteins and protein complexes, and ignores the effects altogether of other outside forces, known to play a role *in vivo*. For instance, although our model includes no explicit cooperativity-encouraging term to allow occupancy by a factor to promote binding by other factors adjacent to it, such terms have been shown to be useful in at least the case of nucleosome-only models (Lubliner and Segal, 2009). Nor are the effects of ATP-dependent nucleosome remodeling complexes taken into account, though they are

known to affect inter-nucleosome spacing (Fyodorov and Kadonaga, 2002), nucleosome positioning (Whitehouse et al., 2007), and nucleosome sequence affinity (Rippe et al., 2007).

Experimentally observed results of nucleosome remodeling like those mentioned above suggest that in large part the remodeling acts as if the sequence preferences for nucleosomes have not been significantly changed, but merely reduced in strength. Our model could easily reflect this behavior by lowering the concentration parameter of nucleosomes, but this would take effect across the entire sequence being decoded. However, this could also be effected via context-dependent multiplicative terms akin to the manner in which cooperativity could be encouraged: in conditions where ATP-dependent nucleosome remodeling is thought to be active (for example, near certain bound TFs), the nucleosome occupancy parameter could be downweighted in the same manner that cooperativity-encouraging occupancy parameters could be upweighted. These kinds of arbitrary context-sensitive adjustments can be difficult (or, at least, can add prohibitively high model complexity) to construct in HMMs via model structure alone, but are relatively straightforward under the related frameworks of *hierarchical hidden Markov models* (HHMMs), further mentioned below.

Accounting for temporal processes

The improvements to this point have clear paths of implementation, even if tedious or non-trivial. However, there exists a set of alterations to the model that require the introduction of temporal concepts, which may or may not be feasible but should be recognized regardless. For example, RNA polymerase has been shown to play a key role in eviction of nucleosomes in gene promoters (Weiner et al., 2010). Hence, accounting for the traversal of the genome by polymerases may lead to improved understanding of *in vivo* binding behaviors, particularly in regions of special interest like gene promoters. As another example, it is known that nucleosomes and tran-

scription factors do not instantly associate and disassociate with DNA. Furthermore, they require different *amounts* of time for these processes to take place, and the dissociation of a nucleosome from the 147 base pairs it binds to can take significantly longer than a typical TF releasing from its approximately 10 base pairs.

Accounting for the temporal aspects of these types of behaviors implies breaking our assumption of equilibrium, which has significant ramifications. The core of the statistical mechanics canonical ensemble on which Boltzmann chains rely requires the assumption of equilibrium. The system in a canonical ensemble is considered to be closed, and the probability of assuming a microstate is independent of what microstate the system has currently assumed, disallowing any kind of imposed ordered progression through microstates (like polymerase traversal or per-base nucleosome association). As such, families of equilibrium-challenging additions like these may be presently out of reach, without a dramatic overhaul of our modeling framework.

5.1.2 Improvements on the practical implementation of our model

The major restrictions on increasing the complexity of our model are the implication of a large state space on processing time and memory requirements. As our model exists now, while TFs account for only dozens of states, the nucleosome model accounts for over 2000 states. For a model of k states and an input sequence L characters long, processing time takes $O(Lk^2)$ and requires $O(Lk)$ memory, so our whole-chromosome computations with our model can already require many tens of gigabytes of storage. Even though our computational optimizations make the practical running time quite fast, it could still be further improved through reducing the size of the state space.

We can alleviate these conditions, increase performance, and even add new capabilities such as energetic bonuses or penalties, inter-nucleosome length distributions, and alternative TF representations, by re-implementing our model as a Hierarchical HMM (HHMM) (Fine et al., 1998) consisting of sub-units of pseudo-2D HMMs (Kuo

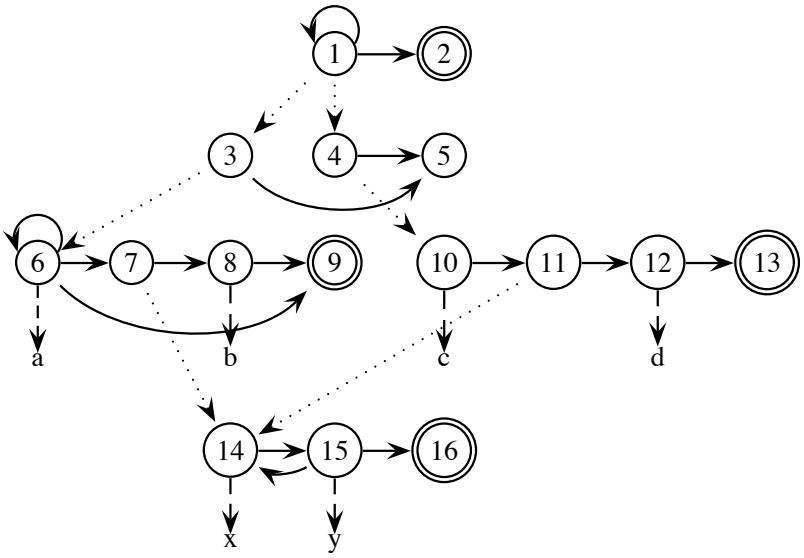

FIGURE 5.1: Hierarchical hidden Markov model. State transition diagram for a four-level HHMM representing the regular expression $a^+|a^+(xy)^+b|c(xy)^+d$. Solid arcs represent horizontal transitions between states; dotted arcs represent vertical transitions, i.e., calling a sub-HMM. Double-ringed states are accepting (end) states; we assume there is at least one per sub-HMM; when we enter such a state, control is returned to the parent (calling) state. Dashed arcs are emissions from production (concrete) states; in this example, each production state emits a single symbol, but in general, it can have a distribution over output symbols. Figure and caption taken from Murphy (2002a).

and Agazzi, 1994) and segment models (Gales and Young, 1993; Ostendorf et al., 1996). HHMMs allow model structures with hierarchical states that may themselves be arbitrary HMMs, or simpler constructions (Figure 5.1). We could conceivably construct an HHMM with subcomponents representing linker sequence, modeled with a length distribution via hidden semi-Markov models or segment models (which allow emission of multiples of characters under arbitrary length distributions), and nucleosome-occupied sequence. Unbound sequence would be a special case of the linker regions, which would be composed also of other bound DBFs like TFs and ORC. If we retained fixed-width DBF representations, all of our DBFs could easily be modeled by pseudo-2D HMMs, which are a special case of segment models where

the length of sequence emitted is known in advance.

A description of the relationship between these, along with inference rules, is described in Murphy (2002a). Re-engineering the model in this way is nontrivial, but gains in both performance and flexibility should be substantial.

5.2 New applications for our model

The binding of DNA by myriad proteins and protein complexes plays a fundamental role in numerous processes within the cell, regulatory and otherwise. Our model can represent any DNA or RNA binding interaction that can be described in terms of sequence preference, concentration, and competition with other like-minded actors, regardless of condition, cell type, tissue, organism, or even domain. As such, the application of our approach to various aspects of intracellular process is limited only by available data and our imagination of how to leverage it.

Our model could be used to understand how chaperones bind to RNA molecules, or the role competition plays in microRNA targets or other non-coding RNA binding. Similar models have recently been used to understand evolutionary pressures on gene promoter regions (Hoffman and Birney, 2010). The most obvious next step, however, is using the output of our model as it is currently constructed to understand how gene expression arises from the binding of various transcription factors in gene promoters.

5.2.1 Linking DNA occupancy to transcription

Although other paths of inquiry may lead to important and high-impact discoveries about the nature of numerous biological phenomena, we give special attention to the potential application of our model of DNA binding to the elucidation of the relationship between binding of transcription factors in promoter regions and the resulting mRNA transcription. The relationship between the dynamics of transcription factor binding and the resultant rate of transcription, by way of what is commonly referred

to as a *gene regulatory function* (GRF), has been an area of active investigation for some time. The scope of the investigations have varied greatly, from understanding broad relationships between modules of transcription factors and modules of genes (Segal et al., 2003) to understanding the intricacies of single loci (Lam et al., 2008; Kim and O'Shea, 2008; Teif, 2010). For a recent review of modeling paradigms employed for this problem, see Kim et al. (2009).

Recent studies have been somewhat successful in related TF binding to expression patterns, but have done so in simplified systems. For instance, Segal et al. (2008) performs this mapping in the fly embryo via a logistic regression of normalized expression (reduced to $[0, 1]$) against a few TFs, where the physiological TF concentrations are known ahead of time. Similar models have been applied to synthetic promoters in yeast, again in toy scenarios (Raveh-Sadka et al., 2009; Gertz et al., 2009). Stating that these scenarios are simplistic is certainly not to disparage them; we merely seek to illustrate the difficulty of this problem by mentioning the limited success in addressing it thus far. The construction of a more generalized approach, capable of predicting *in vivo* expression levels from an arbitrary set of TFs, remains a challenge.

Appendix A

Supplemental Tables

A.1 Table A.1

Table A.1: Values associated with Figure 3.10.

DBF	K_d multiple	ROC AUC with nuc	ROC AUC without nuc	ROC AUC diff
ABF1_YPD	8	0.90458	0.95971	-0.05513
ABF1_YPD	16	0.91456	0.96529	-0.05073
ABF1_YPD	32	0.93565	0.95606	-0.02041
ABF1_YPD	64	0.92685	0.95326	-0.02640
ARO80_SM	8	0.52648	0.53582	-0.00934
ARO80_SM	16	0.51087	0.54668	-0.03581
ARO80_SM	32	0.48164	0.47774	0.00389
ARO80_SM	64	0.48509	0.45887	0.02622
ARO80_YPD	8	0.61012	0.46838	0.14174
ARO80_YPD	16	0.59617	0.46825	0.12793
ARO80_YPD	32	0.56306	0.46493	0.09813
ARO80_YPD	64	0.40827	0.43528	-0.02701
BAS1_SM	8	0.69424	0.67887	0.01537
BAS1_SM	16	0.69883	0.67883	0.02000
BAS1_SM	32	0.71712	0.67898	0.03814
BAS1_SM	64	0.73394	0.67898	0.05496
BAS1_YPD	8	0.75657	0.74362	0.01295
BAS1_YPD	16	0.69954	0.74360	-0.04406
BAS1_YPD	32	0.71593	0.74382	-0.02789
BAS1_YPD	64	0.76862	0.74382	0.02480
CBF1_SM	8	0.97775	0.91648	0.06128
CBF1_SM	16	0.95149	0.91219	0.03930
CBF1_SM	32	0.92452	0.93885	-0.01434
CBF1_SM	64	0.95593	0.94063	0.01530
CBF1_YPD	8	0.76147	0.76041	0.00106

CBF1_YPD	16	0.70686	0.79416	-0.08730
CBF1_YPD	32	0.68842	0.74300	-0.05458
CBF1_YPD	64	0.69044	0.70298	-0.01254
CHA4_SM	8	0.40226	0.43297	-0.03071
CHA4_SM	16	0.43501	0.43291	0.00210
CHA4_SM	32	0.41838	0.43289	-0.01451
CHA4_SM	64	0.43266	0.43289	-0.00022
CHA4_YPD	8	0.34058	0.38558	-0.04499
CHA4_YPD	16	0.29792	0.38549	-0.08757
CHA4_YPD	32	0.34547	0.38551	-0.04005
CHA4_YPD	64	0.26804	0.38538	-0.11734
CUP9_YPD	8	0.49822	0.59134	-0.09313
CUP9_YPD	16	0.51395	0.58863	-0.07469
CUP9_YPD	32	0.60564	0.58820	0.01744
CUP9_YPD	64	0.59450	0.55328	0.04122
ECM22_YPD	8	0.65467	0.67974	-0.02507
ECM22_YPD	16	0.64189	0.67971	-0.03782
ECM22_YPD	32	0.71681	0.65252	0.06429
ECM22_YPD	64	0.68812	0.64174	0.04638
FHL1_H2O2Hi	8	0.45317	0.45373	-0.00057
FHL1_H2O2Hi	16	0.41599	0.48931	-0.07332
FHL1_H2O2Hi	32	0.47002	0.43655	0.03347
FHL1_H2O2Hi	64	0.47850	0.51154	-0.03304
FHL1_RAPA	8	0.50502	0.50585	-0.00084
FHL1_RAPA	16	0.54285	0.45346	0.08939
FHL1_RAPA	32	0.47958	0.46475	0.01483
FHL1_RAPA	64	0.51975	0.49183	0.02792
FHL1_SM	8	0.46727	0.43765	0.02962
FHL1_SM	16	0.48566	0.41635	0.06932
FHL1_SM	32	0.36188	0.38276	-0.02087
FHL1_SM	64	0.42475	0.44007	-0.01532
FHL1_YPD	8	0.60154	0.70639	-0.10485
FHL1_YPD	16	0.66584	0.72457	-0.05873
FHL1_YPD	32	0.64995	0.70949	-0.05954
FHL1_YPD	64	0.63944	0.70082	-0.06138
FKH1_YPD	8	0.72346	0.75140	-0.02794
FKH1_YPD	16	0.68230	0.74003	-0.05773
FKH1_YPD	32	0.76277	0.65745	0.10532
FKH1_YPD	64	0.73869	0.59810	0.14059
FKH2_H2O2Hi	8	0.64709	0.70383	-0.05674
FKH2_H2O2Hi	16	0.68873	0.59579	0.09294
FKH2_H2O2Hi	32	0.64099	0.56821	0.07278
FKH2_H2O2Hi	64	0.68195	0.49356	0.18839
FKH2_H2O2Lo	8	0.55176	0.66649	-0.11472
FKH2_H2O2Lo	16	0.52421	0.54979	-0.02557
FKH2_H2O2Lo	32	0.50744	0.54796	-0.04052
FKH2_H2O2Lo	64	0.55361	0.50691	0.04670
FKH2_YPD	8	0.55539	0.64758	-0.09219
FKH2_YPD	16	0.62669	0.58690	0.03979
FKH2_YPD	32	0.59519	0.56993	0.02526
FKH2_YPD	64	0.62585	0.58710	0.03875

GAL4_GAL	8	0.82965	0.66530	0.16435
GAL4_GAL	16	0.80540	0.64680	0.15860
GAL4_GAL	32	0.71007	0.65149	0.05859
GAL4_GAL	64	0.68007	0.58882	0.09125
GAL4_RAFF	8	0.76445	0.71524	0.04921
GAL4_RAFF	16	0.73748	0.73626	0.00122
GAL4_RAFF	32	0.71428	0.77163	-0.05735
GAL4_RAFF	64	0.75155	0.75470	-0.00315
GAL4_YPD	8	0.75276	0.66503	0.08773
GAL4_YPD	16	0.72193	0.64418	0.07774
GAL4_YPD	32	0.70361	0.63039	0.07321
GAL4_YPD	64	0.66544	0.61285	0.05260
GAT1_HEAT	8	0.56112	0.62342	-0.06229
GAT1_HEAT	16	0.55648	0.60904	-0.05257
GAT1_HEAT	32	0.54326	0.60000	-0.05674
GAT1_HEAT	64	0.56749	0.60084	-0.03335
GAT1_RAPA	8	0.67807	0.65179	0.02628
GAT1_RAPA	16	0.68016	0.56536	0.11480
GAT1_RAPA	32	0.62183	0.55987	0.06196
GAT1_RAPA	64	0.63489	0.61074	0.02415
GAT1_SM	8	0.52767	0.57793	-0.05026
GAT1_SM	16	0.54141	0.48854	0.05288
GAT1_SM	32	0.53048	0.47831	0.05217
GAT1_SM	64	0.53167	0.47751	0.05416
GAT1_YPD	8	0.56432	0.61939	-0.05507
GAT1_YPD	16	0.59660	0.61720	-0.02060
GAT1_YPD	32	0.59166	0.64312	-0.05147
GAT1_YPD	64	0.60627	0.65347	-0.04719
GAT3_YPD	8	0.40573	0.49167	-0.08594
GAT3_YPD	16	0.42598	0.48426	-0.05827
GAT3_YPD	32	0.39019	0.48941	-0.09922
GAT3_YPD	64	0.40755	0.49610	-0.08854
GCN4_RAPA	8	0.88492	0.88656	-0.00163
GCN4_RAPA	16	0.90404	0.85608	0.04796
GCN4_RAPA	32	0.91231	0.81292	0.09938
GCN4_RAPA	64	0.88490	0.78129	0.10362
GCN4_SM	8	0.79848	0.82292	-0.02444
GCN4_SM	16	0.81752	0.78528	0.03224
GCN4_SM	32	0.83483	0.77512	0.05971
GCN4_SM	64	0.79978	0.77983	0.01996
GCN4_YPD	8	0.82720	0.87055	-0.04335
GCN4_YPD	16	0.83209	0.80299	0.02910
GCN4_YPD	32	0.80759	0.75438	0.05321
GCN4_YPD	64	0.79110	0.67830	0.11280
GLN3_RAPA	8	0.66681	0.66120	0.00561
GLN3_RAPA	16	0.69572	0.66119	0.03453
GLN3_RAPA	32	0.66578	0.62492	0.04086
GLN3_RAPA	64	0.70497	0.67504	0.02994
GLN3_SM	8	0.51157	0.48536	0.02621
GLN3_SM	16	0.55415	0.48570	0.06845
GLN3_SM	32	0.49031	0.48090	0.00940

GLN3_SM	64	0.50088	0.48255	0.01834
GLN3_YPD	8	0.41933	0.41762	0.00171
GLN3_YPD	16	0.39169	0.41733	-0.02564
GLN3_YPD	32	0.41096	0.45104	-0.04008
GLN3_YPD	64	0.41104	0.47706	-0.06602
GZF3_H2O2Hi	8	0.56448	0.49171	0.07278
GZF3_H2O2Hi	16	0.51988	0.49479	0.02509
GZF3_H2O2Hi	32	0.49502	0.49461	0.00040
GZF3_H2O2Hi	64	0.50828	0.50032	0.00796
GZF3_RAPA	8	0.54787	0.51131	0.03655
GZF3_RAPA	16	0.50505	0.50851	-0.00346
GZF3_RAPA	32	0.56161	0.50845	0.05315
GZF3_RAPA	64	0.54115	0.53220	0.00895
GZF3_YPD	8	0.63754	0.57012	0.06742
GZF3_YPD	16	0.64320	0.57781	0.06539
GZF3_YPD	32	0.58196	0.57796	0.00399
GZF3_YPD	64	0.60086	0.58497	0.01589
HAL9_YPD	8	0.42704	0.40022	0.02682
HAL9_YPD	16	0.53436	0.44458	0.08978
HAL9_YPD	32	0.45207	0.38689	0.06518
HAL9_YPD	64	0.45543	0.41795	0.03748
LEU3_SM	8	0.52251	0.58007	-0.05756
LEU3_SM	16	0.51801	0.58331	-0.06529
LEU3_SM	32	0.57445	0.54625	0.02819
LEU3_SM	64	0.56128	0.48942	0.07186
LEU3_YPD	8	0.51710	0.47370	0.04340
LEU3_YPD	16	0.49320	0.45811	0.03509
LEU3_YPD	32	0.49669	0.45153	0.04516
LEU3_YPD	64	0.46529	0.45015	0.01514
MBP1_H2O2Hi	8	0.90008	0.88210	0.01799
MBP1_H2O2Hi	16	0.88142	0.85806	0.02336
MBP1_H2O2Hi	32	0.87987	0.83515	0.04472
MBP1_H2O2Hi	64	0.88067	0.83642	0.04425
MBP1_H2O2Lo	8	0.83189	0.81670	0.01519
MBP1_H2O2Lo	16	0.80367	0.81088	-0.00720
MBP1_H2O2Lo	32	0.76441	0.83632	-0.07191
MBP1_H2O2Lo	64	0.81108	0.83659	-0.02551
MBP1_YPD	8	0.85917	0.81281	0.04635
MBP1_YPD	16	0.81935	0.74344	0.07591
MBP1_YPD	32	0.80542	0.73968	0.06574
MBP1_YPD	64	0.80284	0.74015	0.06269
MCM1_Alpha	8	0.85429	0.91899	-0.06469
MCM1_Alpha	16	0.82789	0.89572	-0.06782
MCM1_Alpha	32	0.84363	0.86187	-0.01824
MCM1_Alpha	64	0.82117	0.86373	-0.04256
MCM1_YPD	8	0.84204	0.83329	0.00875
MCM1_YPD	16	0.80965	0.81097	-0.00132
MCM1_YPD	32	0.80516	0.77925	0.02591
MCM1_YPD	64	0.79245	0.76440	0.02804
MET32_SM	8	0.64342	0.64328	0.00014
MET32_SM	16	0.67556	0.64341	0.03215

MET32_SM	32	0.66694	0.60824	0.05869
MET32_SM	64	0.68218	0.59004	0.09214
MET32_YPD	8	0.56247	0.59043	-0.02796
MET32_YPD	16	0.56390	0.59065	-0.02675
MET32_YPD	32	0.54517	0.57915	-0.03398
MET32_YPD	64	0.56109	0.53071	0.03038
MGA1_H2O2Hi	8	0.56155	0.51194	0.04961
MGA1_H2O2Hi	16	0.56535	0.60830	-0.04294
MGA1_H2O2Hi	32	0.64469	0.60902	0.03567
MGA1_H2O2Hi	64	0.65509	0.66193	-0.00684
MGA1_YPD	8	0.42903	0.40285	0.02618
MGA1_YPD	16	0.47360	0.40668	0.06692
MGA1_YPD	32	0.48369	0.41028	0.07340
MGA1_YPD	64	0.41574	0.43601	-0.02027
MIG1_GAL	8	0.56047	0.58691	-0.02644
MIG1_GAL	16	0.56080	0.55539	0.00541
MIG1_GAL	32	0.57822	0.55547	0.02275
MIG1_GAL	64	0.58870	0.55555	0.03315
MIG1_YPD	8	0.50714	0.49967	0.00747
MIG1_YPD	16	0.53898	0.46312	0.07586
MIG1_YPD	32	0.57549	0.46316	0.11233
MIG1_YPD	64	0.57039	0.46303	0.10736
MIG2_H2O2Hi	8	0.39894	0.39449	0.00445
MIG2_H2O2Hi	16	0.41072	0.39400	0.01672
MIG2_H2O2Hi	32	0.40981	0.43630	-0.02649
MIG2_H2O2Hi	64	0.43989	0.45401	-0.01412
MIG2_YPD	8	0.50206	0.53349	-0.03143
MIG2_YPD	16	0.54989	0.53281	0.01708
MIG2_YPD	32	0.53620	0.53272	0.00348
MIG2_YPD	64	0.56292	0.55081	0.01212
MIG3_YPD	8	0.51620	0.56251	-0.04631
MIG3_YPD	16	0.62625	0.53631	0.08994
MIG3_YPD	32	0.58911	0.54866	0.04044
MIG3_YPD	64	0.59411	0.58308	0.01103
NDT80_YPD	8	0.46746	0.44573	0.02173
NDT80_YPD	16	0.44116	0.45104	-0.00989
NDT80_YPD	32	0.46033	0.47568	-0.01535
NDT80_YPD	64	0.48255	0.50460	-0.02205
NRG1_H2O2Hi	8	0.76131	0.79482	-0.03351
NRG1_H2O2Hi	16	0.79842	0.82011	-0.02169
NRG1_H2O2Hi	32	0.80419	0.74870	0.05549
NRG1_H2O2Hi	64	0.76533	0.73771	0.02762
NRG1_H2O2Lo	8	0.59506	0.66906	-0.07400
NRG1_H2O2Lo	16	0.62838	0.65443	-0.02605
NRG1_H2O2Lo	32	0.64786	0.58487	0.06299
NRG1_H2O2Lo	64	0.65026	0.54114	0.10912
NRG1_YPD	8	0.62224	0.65987	-0.03763
NRG1_YPD	16	0.62551	0.63540	-0.00989
NRG1_YPD	32	0.63858	0.58445	0.05413
NRG1_YPD	64	0.66031	0.56586	0.09444
OAF1_YPD	8	0.50258	0.46416	0.03842

OAF1_YPD	16	0.52943	0.51070	0.01873
OAF1_YPD	32	0.47940	0.49684	-0.01744
OAF1_YPD	64	0.55904	0.47610	0.08294
PDR1_H2O2Lo	8	0.54471	0.55503	-0.01031
PDR1_H2O2Lo	16	0.52591	0.57108	-0.04518
PDR1_H2O2Lo	32	0.54180	0.58427	-0.04247
PDR1_H2O2Lo	64	0.55445	0.55902	-0.00458
PDR1_YPD	8	0.51961	0.53343	-0.01382
PDR1_YPD	16	0.50936	0.55033	-0.04097
PDR1_YPD	32	0.51065	0.55170	-0.04105
PDR1_YPD	64	0.53866	0.46760	0.07106
PHD1_BUT14	8	0.46992	0.38795	0.08197
PHD1_BUT14	16	0.46616	0.38795	0.07820
PHD1_BUT14	32	0.47060	0.38789	0.08271
PHD1_BUT14	64	0.53365	0.38773	0.14593
PHD1_BUT90	8	0.54662	0.55333	-0.00671
PHD1_BUT90	16	0.60584	0.55333	0.05251
PHD1_BUT90	32	0.62689	0.55324	0.07366
PHD1_BUT90	64	0.62705	0.55343	0.07362
PHD1_YPD	8	0.60627	0.54640	0.05987
PHD1_YPD	16	0.62022	0.54640	0.07382
PHD1_YPD	32	0.61637	0.54639	0.06998
PHD1_YPD	64	0.65487	0.54645	0.10842
PHO2_H2O2Hi	8	0.36626	0.36308	0.00318
PHO2_H2O2Hi	16	0.33385	0.40427	-0.07042
PHO2_H2O2Hi	32	0.31258	0.31695	-0.00437
PHO2_H2O2Hi	64	0.32399	0.27351	0.05049
PHO2_H2O2Lo	8	0.38370	0.39062	-0.00693
PHO2_H2O2Lo	16	0.45861	0.32061	0.13800
PHO2_H2O2Lo	32	0.35660	0.33377	0.02284
PHO2_H2O2Lo	64	0.39671	0.37214	0.02457
PHO2_Pi-	8	0.34569	0.36909	-0.02340
PHO2_Pi-	16	0.39096	0.38491	0.00605
PHO2_Pi-	32	0.41689	0.40481	0.01208
PHO2_Pi-	64	0.35098	0.43520	-0.08422
PHO2_SM	8	0.43450	0.37350	0.06100
PHO2_SM	16	0.40041	0.38565	0.01476
PHO2_SM	32	0.46659	0.34797	0.11862
PHO2_SM	64	0.45757	0.37628	0.08129
PHO2_YPD	8	0.46624	0.41342	0.05282
PHO2_YPD	16	0.38454	0.40521	-0.02067
PHO2_YPD	32	0.40019	0.44623	-0.04604
PHO2_YPD	64	0.37226	0.50644	-0.13417
PHO4_Pi-	8	0.63397	0.53538	0.09859
PHO4_Pi-	16	0.57291	0.55419	0.01872
PHO4_Pi-	32	0.56919	0.52040	0.04878
PHO4_Pi-	64	0.56121	0.45542	0.10579
PHO4_YPD	8	0.56581	0.43083	0.13498
PHO4_YPD	16	0.51918	0.40423	0.11494
PHO4_YPD	32	0.41337	0.39525	0.01812
PHO4_YPD	64	0.40554	0.36325	0.04229

PUT3_H2O2Lo	8	0.36164	0.30889	0.05275
PUT3_H2O2Lo	16	0.37119	0.26214	0.10905
PUT3_H2O2Lo	32	0.35110	0.25977	0.09133
PUT3_H2O2Lo	64	0.39680	0.25897	0.13784
PUT3_SM	8	0.50293	0.67683	-0.17390
PUT3_SM	16	0.53508	0.65207	-0.11699
PUT3_SM	32	0.61390	0.62970	-0.01580
PUT3_SM	64	0.63387	0.61860	0.01526
PUT3_YPD	8	0.50866	0.59876	-0.09010
PUT3_YPD	16	0.47800	0.60769	-0.12969
PUT3_YPD	32	0.52260	0.57920	-0.05660
PUT3_YPD	64	0.50697	0.57500	-0.06803
RAP1_SM	8	0.90940	0.94620	-0.03680
RAP1_SM	16	0.94212	0.94485	-0.00273
RAP1_SM	32	0.95751	0.90257	0.05494
RAP1_SM	64	0.94289	0.94323	-0.00034
RAP1_YPD	8	0.89958	0.96464	-0.06506
RAP1_YPD	16	0.94032	0.95733	-0.01700
RAP1_YPD	32	0.96041	0.95115	0.00927
RAP1_YPD	64	0.96131	0.94938	0.01193
RDR1_YPD	8	0.47655	0.45964	0.01692
RDR1_YPD	16	0.48871	0.35886	0.12985
RDR1_YPD	32	0.45937	0.34863	0.11074
RDR1_YPD	64	0.40779	0.37322	0.03456
RDS1_H2O2Hi	8	0.78381	0.79284	-0.00903
RDS1_H2O2Hi	16	0.79665	0.77698	0.01968
RDS1_H2O2Hi	32	0.77742	0.74264	0.03478
RDS1_H2O2Hi	64	0.78446	0.67752	0.10694
RDS1_YPD	8	0.63989	0.68439	-0.04450
RDS1_YPD	16	0.68825	0.63229	0.05595
RDS1_YPD	32	0.66747	0.62469	0.04278
RDS1_YPD	64	0.70534	0.52905	0.17628
RGT1_GAL	8	0.74274	0.70514	0.03760
RGT1_GAL	16	0.67041	0.71396	-0.04355
RGT1_GAL	32	0.65007	0.71980	-0.06974
RGT1_GAL	64	0.66623	0.68992	-0.02370
RGT1_YPD	8	0.58148	0.60317	-0.02170
RGT1_YPD	16	0.60048	0.60358	-0.00309
RGT1_YPD	32	0.58664	0.62444	-0.03780
RGT1_YPD	64	0.66381	0.65338	0.01043
RPH1_H2O2Hi	8	0.41509	0.41370	0.00139
RPH1_H2O2Hi	16	0.38135	0.38235	-0.00100
RPH1_H2O2Hi	32	0.38890	0.40269	-0.01379
RPH1_H2O2Hi	64	0.41459	0.40691	0.00768
RPH1_H2O2Lo	8	0.50645	0.60777	-0.10132
RPH1_H2O2Lo	16	0.44856	0.55164	-0.10307
RPH1_H2O2Lo	32	0.40767	0.56962	-0.16195
RPH1_H2O2Lo	64	0.44783	0.59666	-0.14883
RPH1_SM	8	0.49858	0.53352	-0.03494
RPH1_SM	16	0.55109	0.54296	0.00813
RPH1_SM	32	0.51618	0.50328	0.01290

RPH1_SM	64	0.55053	0.51424	0.03629
RPH1_YPD	8	0.46481	0.46281	0.00200
RPH1_YPD	16	0.52139	0.49788	0.02351
RPH1_YPD	32	0.49312	0.47561	0.01751
RPH1_YPD	64	0.49514	0.45312	0.04201
RPN4_H2O2Hi	8	0.56064	0.49941	0.06123
RPN4_H2O2Hi	16	0.57613	0.49938	0.07675
RPN4_H2O2Hi	32	0.56044	0.49965	0.06080
RPN4_H2O2Hi	64	0.58344	0.50177	0.08167
RPN4_H2O2Lo	8	0.71598	0.76929	-0.05331
RPN4_H2O2Lo	16	0.72668	0.76939	-0.04272
RPN4_H2O2Lo	32	0.76061	0.74094	0.01968
RPN4_H2O2Lo	64	0.81978	0.76806	0.05172
RPN4_YPD	8	0.42453	0.38012	0.04441
RPN4_YPD	16	0.42293	0.37997	0.04296
RPN4_YPD	32	0.41550	0.37663	0.03887
RPN4_YPD	64	0.46107	0.37939	0.08168
RTG3_H2O2Hi	8	0.61206	0.57456	0.03750
RTG3_H2O2Hi	16	0.54039	0.53358	0.00681
RTG3_H2O2Hi	32	0.57022	0.56158	0.00864
RTG3_H2O2Hi	64	0.55979	0.55313	0.00666
RTG3_H2O2Lo	8	0.45979	0.45724	0.00255
RTG3_H2O2Lo	16	0.53495	0.49022	0.04473
RTG3_H2O2Lo	32	0.48418	0.44090	0.04328
RTG3_H2O2Lo	64	0.37330	0.41305	-0.03975
RTG3_RAPA	8	0.49463	0.49187	0.00276
RTG3_RAPA	16	0.46633	0.56412	-0.09780
RTG3_RAPA	32	0.52407	0.58234	-0.05827
RTG3_RAPA	64	0.52650	0.56422	-0.03772
RTG3_SM	8	0.52738	0.47774	0.04964
RTG3_SM	16	0.51072	0.49360	0.01712
RTG3_SM	32	0.47081	0.48902	-0.01821
RTG3_SM	64	0.38434	0.47017	-0.08583
RTG3_YPD	8	0.50332	0.45780	0.04552
RTG3_YPD	16	0.52353	0.43787	0.08566
RTG3_YPD	32	0.45905	0.40445	0.05461
RTG3_YPD	64	0.38236	0.38925	-0.00689
SFL1_YPD	8	0.46048	0.48758	-0.02710
SFL1_YPD	16	0.52088	0.48921	0.03166
SFL1_YPD	32	0.51773	0.48046	0.03727
SFL1_YPD	64	0.43509	0.52255	-0.08746
SFP1_H2O2Hi	8	0.41725	0.51469	-0.09744
SFP1_H2O2Hi	16	0.45990	0.49569	-0.03579
SFP1_H2O2Hi	32	0.42979	0.45910	-0.02931
SFP1_H2O2Hi	64	0.46401	0.43462	0.02939
SFP1_H2O2Lo	8	0.31960	0.42305	-0.10345
SFP1_H2O2Lo	16	0.29875	0.39179	-0.09305
SFP1_H2O2Lo	32	0.34811	0.42168	-0.07356
SFP1_H2O2Lo	64	0.37893	0.41205	-0.03312
SFP1_SM	8	0.37128	0.41666	-0.04538
SFP1_SM	16	0.35537	0.40514	-0.04978

SFP1_SM	32	0.42204	0.44840	-0.02636
SFP1_SM	64	0.46283	0.40097	0.06186
SFP1_YPD	8	0.50606	0.40473	0.10133
SFP1_YPD	16	0.43151	0.37970	0.05182
SFP1_YPD	32	0.49695	0.37843	0.11852
SFP1_YPD	64	0.34587	0.41420	-0.06833
SIP4_SM	8	0.41477	0.29858	0.11620
SIP4_SM	16	0.41903	0.37868	0.04035
SIP4_SM	32	0.37505	0.45291	-0.07786
SIP4_SM	64	0.31652	0.45646	-0.13995
SIP4_YPD	8	0.57118	0.52112	0.05005
SIP4_YPD	16	0.58085	0.51647	0.06438
SIP4_YPD	32	0.53504	0.58748	-0.05244
SIP4_YPD	64	0.48610	0.61277	-0.12667
SKN7_H2O2Hi	8	0.45922	0.51534	-0.05612
SKN7_H2O2Hi	16	0.57216	0.53995	0.03221
SKN7_H2O2Hi	32	0.56583	0.49226	0.07357
SKN7_H2O2Hi	64	0.56557	0.47541	0.09016
SKN7_H2O2Lo	8	0.51257	0.48851	0.02406
SKN7_H2O2Lo	16	0.55431	0.50655	0.04776
SKN7_H2O2Lo	32	0.51726	0.46314	0.05412
SKN7_H2O2Lo	64	0.48677	0.46042	0.02635
SKN7_HEAT	8	0.46873	0.54102	-0.07229
SKN7_HEAT	16	0.50880	0.57184	-0.06304
SKN7_HEAT	32	0.54535	0.56113	-0.01578
SKN7_HEAT	64	0.52748	0.52986	-0.00239
SKN7_YPD	8	0.53984	0.53635	0.00349
SKN7_YPD	16	0.55720	0.53193	0.02526
SKN7_YPD	32	0.62851	0.53957	0.08894
SKN7_YPD	64	0.65242	0.50307	0.14935
SMP1_YPD	8	0.43799	0.48668	-0.04869
SMP1_YPD	16	0.49477	0.44992	0.04486
SMP1_YPD	32	0.48245	0.49079	-0.00834
SMP1_YPD	64	0.54912	0.52483	0.02429
SRD1_YPD	8	0.54332	0.51812	0.02521
SRD1_YPD	16	0.51334	0.54268	-0.02934
SRD1_YPD	32	0.55181	0.56520	-0.01339
SRD1_YPD	64	0.54257	0.52466	0.01791
STP2_YPD	8	0.51822	0.49311	0.02510
STP2_YPD	16	0.54046	0.50322	0.03724
STP2_YPD	32	0.50676	0.50359	0.00317
STP2_YPD	64	0.51835	0.51515	0.00321
STP4_YPD	8	0.67766	0.70033	-0.02267
STP4_YPD	16	0.64756	0.70031	-0.05275
STP4_YPD	32	0.63021	0.72695	-0.09674
STP4_YPD	64	0.60328	0.71438	-0.11111
SUM1_YPD	8	0.64326	0.75022	-0.10696
SUM1_YPD	16	0.68597	0.78434	-0.09836
SUM1_YPD	32	0.72421	0.77373	-0.04952
SUM1_YPD	64	0.74283	0.73837	0.00446
SUT2_YPD	8	0.43698	0.51517	-0.07819

SUT2_YPD	16	0.49377	0.51765	-0.02388
SUT2_YPD	32	0.49592	0.52033	-0.02441
SUT2_YPD	64	0.52963	0.50073	0.02889
TBS1_YPD	8	0.57137	0.56131	0.01007
TBS1_YPD	16	0.57314	0.52108	0.05206
TBS1_YPD	32	0.57946	0.53048	0.04898
TBS1_YPD	64	0.59778	0.56548	0.03231
TEC1_Alpha	8	0.57795	0.72250	-0.14455
TEC1_Alpha	16	0.56074	0.68058	-0.11984
TEC1_Alpha	32	0.55078	0.61457	-0.06379
TEC1_Alpha	64	0.57347	0.57119	0.00228
TEC1_BUT14	8	0.71512	0.76083	-0.04571
TEC1_BUT14	16	0.73141	0.69718	0.03422
TEC1_BUT14	32	0.66538	0.68597	-0.02059
TEC1_BUT14	64	0.65655	0.59023	0.06632
TEC1_YPD	8	0.61708	0.78335	-0.16627
TEC1_YPD	16	0.66049	0.75277	-0.09227
TEC1_YPD	32	0.65470	0.74577	-0.09106
TEC1_YPD	64	0.68126	0.65439	0.02687
TYE7_YPD	8	0.90381	0.86496	0.03885
TYE7_YPD	16	0.92555	0.84920	0.07636
TYE7_YPD	32	0.87240	0.87707	-0.00467
TYE7_YPD	64	0.84991	0.79349	0.05642
UME6_H2O2Hi	8	0.91089	0.91950	-0.00861
UME6_H2O2Hi	16	0.91762	0.92126	-0.00364
UME6_H2O2Hi	32	0.90593	0.93284	-0.02691
UME6_H2O2Hi	64	0.90844	0.93036	-0.02192
UME6_YPD	8	0.95763	0.96519	-0.00756
UME6_YPD	16	0.96632	0.96028	0.00603
UME6_YPD	32	0.95530	0.96907	-0.01376
UME6_YPD	64	0.95768	0.97114	-0.01346
USV1_YPD	8	0.42809	0.42302	0.00507
USV1_YPD	16	0.44408	0.46560	-0.02152
USV1_YPD	32	0.50050	0.47726	0.02324
USV1_YPD	64	0.50003	0.46307	0.03696
XBP1_H2O2Lo	8	0.56854	0.64697	-0.07842
XBP1_H2O2Lo	16	0.59980	0.66703	-0.06723
XBP1_H2O2Lo	32	0.60369	0.68222	-0.07853
XBP1_H2O2Lo	64	0.62766	0.62743	0.00023
XBP1_YPD	8	0.29567	0.42368	-0.12801
XBP1_YPD	16	0.26553	0.44741	-0.18188
XBP1_YPD	32	0.32394	0.45405	-0.13011
XBP1_YPD	64	0.36887	0.47670	-0.10783
YAP1_H2O2Hi	8	0.37261	0.37054	0.00207
YAP1_H2O2Hi	16	0.40882	0.39483	0.01399
YAP1_H2O2Hi	32	0.39765	0.42399	-0.02635
YAP1_H2O2Hi	64	0.41742	0.43094	-0.01352
YAP1_H2O2Lo	8	0.68761	0.70033	-0.01272
YAP1_H2O2Lo	16	0.64285	0.66740	-0.02455
YAP1_H2O2Lo	32	0.60852	0.62731	-0.01879
YAP1_H2O2Lo	64	0.59876	0.58479	0.01397

YAP1_HEAT	8	0.39587	0.42072	-0.02485
YAP1_HEAT	16	0.38353	0.40225	-0.01872
YAP1_HEAT	32	0.42013	0.45583	-0.03570
YAP1_HEAT	64	0.44032	0.49242	-0.05210
YAP1_YPD	8	0.57921	0.53697	0.04224
YAP1_YPD	16	0.54182	0.55478	-0.01295
YAP1_YPD	32	0.56468	0.52051	0.04417
YAP1_YPD	64	0.57985	0.46937	0.11049
YAP6_H2O2Hi	8	0.59400	0.60777	-0.01377
YAP6_H2O2Hi	16	0.53254	0.56838	-0.03584
YAP6_H2O2Hi	32	0.51959	0.58001	-0.06043
YAP6_H2O2Hi	64	0.51682	0.57135	-0.05453
YAP6_H2O2Lo	8	0.56811	0.64713	-0.07902
YAP6_H2O2Lo	16	0.52926	0.64033	-0.11107
YAP6_H2O2Lo	32	0.52497	0.62808	-0.10311
YAP6_H2O2Lo	64	0.63972	0.62407	0.01565
YAP6_YPD	8	0.58325	0.62458	-0.04133
YAP6_YPD	16	0.56478	0.62094	-0.05616
YAP6_YPD	32	0.54863	0.58473	-0.03610
YAP6_YPD	64	0.56153	0.57280	-0.01127
YDR520C_YPD	8	0.47293	0.55810	-0.08517
YDR520C_YPD	16	0.57848	0.54879	0.02968
YDR520C_YPD	32	0.60129	0.50121	0.10008
YDR520C_YPD	64	0.57285	0.52630	0.04655
YER130C_YPD	8	0.50915	0.56839	-0.05925
YER130C_YPD	16	0.48651	0.53897	-0.05246
YER130C_YPD	32	0.54836	0.51935	0.02901
YER130C_YPD	64	0.47072	0.49293	-0.02221
YGR067C_YPD	8	0.50795	0.52077	-0.01283
YGR067C_YPD	16	0.57272	0.52077	0.05195
YGR067C_YPD	32	0.57965	0.52075	0.05890
YGR067C_YPD	64	0.50547	0.52066	-0.01519
YKL222C_YPD	8	0.53025	0.53700	-0.00674
YKL222C_YPD	16	0.51984	0.52851	-0.00867
YKL222C_YPD	32	0.49234	0.42579	0.06655
YKL222C_YPD	64	0.51616	0.40266	0.11351
YML081W_YPD	8	0.45135	0.44882	0.00253
YML081W_YPD	16	0.50183	0.47806	0.02377
YML081W_YPD	32	0.51750	0.57748	-0.05998
YML081W_YPD	64	0.42151	0.58758	-0.16607
YNR063W_YPD	8	0.48028	0.54195	-0.06168
YNR063W_YPD	16	0.44563	0.55645	-0.11082
YNR063W_YPD	32	0.51238	0.52400	-0.01162
YNR063W_YPD	64	0.55456	0.52857	0.02599
YOX1_YPD	8	0.57920	0.58637	-0.00717
YOX1_YPD	16	0.56164	0.58723	-0.02559
YOX1_YPD	32	0.52865	0.61397	-0.08533
YOX1_YPD	64	0.53104	0.61217	-0.08113
YPR196W_YPD	8	0.44849	0.43490	0.01359
YPR196W_YPD	16	0.44432	0.42777	0.01655
YPR196W_YPD	32	0.46987	0.39397	0.07589

YPR196W_YPD	64	0.49858	0.36955	0.12904
YRR1_YPD	8	0.43422	0.50210	-0.06789
YRR1_YPD	16	0.46032	0.54362	-0.08330
YRR1_YPD	32	0.49881	0.53388	-0.03508
YRR1_YPD	64	0.50994	0.53290	-0.02295

A.2 Table A.2

Table A.2: Values associated with Figure 3.11.

t	K_d multiple	Spearman with TFs	Spearman without TFs	Spearman diff
0.0	2	0.43925	0.39275	0.04650
0.0	4	0.44611	0.39275	0.05336
0.0	8	0.44113	0.39275	0.04838
0.0	16	0.42031	0.39275	0.02756
0.0	32	0.38343	0.39275	-0.00932
0.0	64	0.33801	0.39275	-0.05474
0.25	2	0.46911	0.41587	0.05323
0.25	4	0.47953	0.41587	0.06365
0.25	8	0.47930	0.41587	0.06343
0.25	16	0.46417	0.41587	0.04830
0.25	32	0.43373	0.41587	0.01786
0.25	64	0.39461	0.41587	-0.02126
0.5	2	0.50040	0.43995	0.06045
0.5	4	0.51446	0.43995	0.07450
0.5	8	0.51877	0.43995	0.07881
0.5	16	0.50872	0.43995	0.06877
0.5	32	0.48298	0.43995	0.04302
0.5	64	0.44811	0.43995	0.00816
0.75	2	0.53971	0.47248	0.06723
0.75	4	0.55638	0.47248	0.08390
0.75	8	0.56354	0.47248	0.09106
0.75	16	0.55699	0.47248	0.08451
0.75	32	0.53536	0.47248	0.06288
0.75	64	0.50452	0.47248	0.03204
1.0	2	0.58584	0.50978	0.07606
1.0	4	0.60500	0.50978	0.09522
1.0	8	0.61461	0.50978	0.10483
1.0	16	0.61076	0.50978	0.10098
1.0	32	0.59250	0.50978	0.08272
1.0	64	0.56509	0.50978	0.05531

Bibliography

Agarwal, P. and Bafna, V. (1998), "Detecting non-adjoining correlations with signals in DNA," in *Proceedings of the Second Annual International Conference on Computational Molecular Biology*, pp. 2–8, New York, New York, United States, ACM.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002), *Molecular Biology of the Cell, Fourth Edition*, Garland.

Bammler, T., Beyer, R. P., Bhattacharya, S., Boorman, G. A., Boyles, A., Bradford, B. U., Bumgarner, R. E., Bushel, P. R., Chaturvedi, K., Choi, D., Cunningham, M. L., Deng, S., Dressman, H. K., Fannin, R. D., Farin, F. M., Freedman, J. H., Fry, R. C., Harper, A., Humble, M. C., Hurban, P., Kavanagh, T. J., Kaufmann, W. K., Kerr, K. F., Jing, L., Lapidus, J. A., Lasarev, M. R., Li, J., Li, Y., Lobenhofer, E. K., Lu, X., Malek, R. L., Milton, S., Nagalla, S. R., O'malley, J. P., Palmer, V. S., Pattee, P., Paules, R. S., Perou, C. M., Phillips, K., Qin, L., Qiu, Y., Quigley, S. D., Rodland, M., Rusyn, I., Samson, L. D., Schwartz, D. A., Shi, Y., Shin, J., Sieber, S. O., Slifer, S., Speer, M. C., Spencer, P. S., Sproles, D. I., Swenberg, J. A., Suk, W. A., Sullivan, R. C., Tian, R., Tennant, R. W., Todd, S. A., Tucker, C. J., Houten, B. V., Weis, B. K., Xuan, S., and Zarbl, H. (2005), "Standardizing global gene expression analysis between laboratories and across platforms," *Nature Methods*, 2, 351–356, PMID: 15846362.

Barash, Y., Elidan, G., Friedman, N., and Kaplan, T. (2003), "Modeling dependencies in protein-DNA binding sites," in *Proceedings of the Seventh Annual International Conference on Research in Computational Molecular Biology*, p. 37.

Bartel, D. P. (2004), "MicroRNAs: genomics, biogenesis, mechanism, and function," *Cell*, 116, 281297.

Bartel, D. P. (2009), "MicroRNAs: target recognition and regulatory functions," *Cell*, 136, 215233.

Baum, L. E. (1972), "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities*, 3, 18.

Benos, P. V., Bulyk, M. L., and Stormo, G. D. (2002), "Additivity in protein-DNA interactions: how good an approximation is it?" *Nucleic acids research*, 30, 4442.

Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., and Bulyk, M. L. (2006), "Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities," *Nature biotechnology*, 24, 14291435.

Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U., Hwa, T., Kondev, J., Kuhlman, T., and Phillips, R. (2005a), "Transcriptional regulation by the numbers: applications," *Curr. Opin. Genet. Dev.*, 15, 125–135.

Bintu, L., Buchler, N., Garcia, H., Gerland, U., Hwa, T., Kondev, J., and Phillips, R. (2005b), "Transcriptional regulation by the numbers: models," *Curr. Opin. Genet. Dev.*, 15, 116–124.

Bloch, F. and Walecka, J. D. (2001), *Fundamentals of Statistical Mechanics: Manuscript and Notes of Felix Bloch*, World Scientific Publishing Company.

Bolstad, B., Irizarry, R., Astrand, M., and Speed, T. (2003), "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias," *Bioinformatics*, 19, 185–193.

Breier, A. M., Chatterji, S., and Cozzarelli, N. R. (2004), "Prediction of *Saccharomyces cerevisiae* replication origins," *Genome Biol.*, 5, R22.

Bulyk, M. L. (2004), "Computational prediction of transcription-factor binding site locations," *Genome Biology*, 5, 201201.

Bulyk, M. L., Johnson, P. L., and Church, G. M. (2002), "Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors," *Nucleic acids research*, 30, 1255.

Carninci, P. (2005), "The transcriptional landscape of the mammalian genome," *Science*, 309, 1559–1563.

Cheng, J. (2005), "Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution," *Science*, 308, 1149–1154.

Chern, T. M., van Nimwegen, E., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., and Zavolan, M. (2006), "A simple physical model predicts small exon length variations," *PLoS Genet.*, 2, e45.

Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., and Botstein, D. (1998), "SGD: Saccharomyces Genome Database," *Nucleic Acids Res.*, 26, 73–79.

Coulon, A., Gandrillon, O., and Beslon, G. (2010), "On the spontaneous stochastic dynamics of a single gene: complexity of the molecular interplay at the promoter," *BMC Systems Biology*, 4, 2.

Crick, F. H. (1958), "On protein synthesis," *Symposia of the Society for Experimental Biology*, 12, 138–163, PMID: 13580867.

Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004), "WebLogo: a sequence logo generator," *Genome Research*, 14, 1188.

Dempster, A. P., Laird, N. M., Rubin, D. B., et al. (1977), "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, 138.

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997), "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science*, 278, 680.

Djordjevic, M., Sengupta, A. M., and Shraiman, B. I. (2003), "A biophysical approach to transcription factor binding site discovery," *Genome Res.*, 13, 2381–2390.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998), *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, UK.

Fernandez-Lopez, R., del Campo, I., Ruiz, R., Lanza, V., Vielva, L., and de la Cruz, F. (2010), "Numbers on the edges: a simplified and scalable method for quantifying the gene regulation function," *BioEssays*, 32, 346–355.

Field, Y., Kaplan, N., Fondufe-Mittendorf, Y., Moore, I., Sharon, E., Lubling, Y., Widom, J., and Segal, E. (2008), "Distinct modes of regulation by chromatin encoded through nucleosome positioning signals," *PLoS Comput. Biol.*, 4, e1000216.

Fine, S., Singer, Y., and Tishby, N. (1998), "The hierarchical hidden Markov model: analysis and applications," *Machine learning*, 32, 4162.

Foat, B., Morozov, A., and Bussemaker, H. (2006), "Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE," *Bioinformatics*, 22, e141–149.

Fyodorov, D. V. and Kadonaga, J. T. (2002), "Dynamics of ATP-dependent chromatin assembly by ACF," *Nature*, 418, 896–900.

Gales, M. J. and Young, S. J. (1993), *The theory of segmental hidden Markov models*, University of Cambridge, Department of Engineering.

Gamerman, D. and Lopes, H. F. (2006), *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition*, Chapman and Hall/CRC, 2 edn.

Gao, P., Honkela, A., Rattray, M., and Lawrence, N. D. (2008), "Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities," *Bioinformatics*, 24, i70–75.

Gertz, J., Siggia, E. D., and Cohen, B. A. (2009), "Analysis of combinatorial cis-regulation in synthetic and genomic promoters," *Nature*, 457, 215–218.

Giniger, E., Varnum, S., and Ptashne, M. (1985), "Specific DNA binding of GAL4, a positive regulatory protein of yeast," *Cell*, 40, 767–774.

Granek, J. and Clarke, N. (2005), "Explicit equilibrium modeling of transcription-factor binding and gene regulation," *Genome Biol.*, 6, R87.

Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., and Sethna, J. P. (2007), "Universally sloppy parameter sensitivities in systems biology models," *PLoS Comput Biol*, 3, e189.

Halley, J. W. (2007), *Statistical Mechanics: From First Principles to Macroscopic Phenomena*, Cambridge Univ Pr.

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

Hoffman, M. M. and Birney, E. (2010), "An effective model for natural selection in promoters," *Genome Research*, 20, 685–692.

Hogan, G. J., Lee, C., and Lieb, J. D. (2006), "Cell cyclespecified fluctuation of nucleosome occupancy at gene promoters," *PLoS Genet*, 2, e158.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003), "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostat*, 4, 249–264.

Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M., and Brown, P. O. (2001), "Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF," *Nature*, 409, 533538.

Jacob, F. and Monod, J. (1961), "Genetic regulatory mechanisms in the synthesis of proteins," *Journal of molecular biology*, 3, 318356.

Johnson, J. M., Edwards, S., Shoemaker, D., and Schadt, E. E. (2005), "Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments," *Trends Genet.*, 21, 93–102.

Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Tillo, D., Field, Y., LeProust, E. M., Hughes, T. R., Lieb, J. D., Widom, J., and Segal, E. (2009), "The DNA-encoded nucleosome organization of a eukaryotic genome," *Nature*, 458, 362–366.

Kharchenko, P. V., Woo, C. J., Tolstorukov, M. Y., Kingston, R. E., and Park, P. J. (2008), "Nucleosome positioning in human HOX gene clusters," *Genome Research*, 18, 1554 –1561.

Kim, H. D. and O'Shea, E. K. (2008), "A quantitative model of transcription factor-activated gene expression," *Nat Struct Mol Biol*, 15, 1192–1198.

Kim, H. D., Shay, T., O'Shea, E. K., and Regev, A. (2009), "Transcriptional regulatory circuits: predicting numbers from alphabets," *Science*, 325, 429–432.

Kuo, S. and Agazzi, O. (1994), "Keyword spotting in poorly printed documents using pseudo 2-D hidden markov models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16, 842–848.

Lam, F., Steger, D., and O'Shea, E. (2008), "Chromatin decouples promoter threshold from dynamic range," *Nature*, 453, 246–250.

Lee, C., Shibata, Y., Rao, B., Strahl, B., and Lieb, J. (2004), "Evidence for nucleosome depletion at active regulatory regions genome-wide," *Nat. Genet.*, 36, 900–905.

Lee, W., Tillo, D., Bray, N., Morse, R., Davis, R., Hughes, T., and Nislow, C. (2007), "A high-resolution atlas of nucleosome occupancy in yeast," *Nat. Genet.*, 39, 1235–1244.

Li, C. and Wong, W. H. (2001), "Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection," *Proceedings of the National Academy of Sciences of the United States of America*, 98, 31 –36.

Lieb, J. D., Liu, X., Botstein, D., and Brown, P. O. (2001), "Promoter-specific binding of Rap1 revealed by genome-wide maps of proteinDNA association," *Nature genetics*, 28, 327334.

Lipford, J. R. and Bell, S. P. (2001), "Nucleosomes positioned by ORC facilitate the initiation of DNA replication," *Mol. Cell*, 7, 21–30.

Liu, H. (2010), "Quantifying transcriptional regulatory networks by integrating sequence features and microarray data," *Bioprocess and biosystems engineering*, 33, 495505.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. (1996), "Expression monitoring by hybridization to high-density oligonucleotide arrays," *Nature Biotechnology*, 14, 1675–1680, PMID: 9634850.

Lubliner, S. and Segal, E. (2009), "Modeling interactions between adjacent nucleosomes improves genome-wide predictions of nucleosome occupancy," *Bioinformatics*, 25, i348–i355, PMID: 19478009 PMCID: 2687970.

MacIsaac, K., Wang, T., Gordon, D., Gifford, D., Stormo, G., and Fraenkel, E. (2006), "An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*," *BMC Bioinformatics*, 7, 113.

Marahrens, Y. and Stillman, B. (1992), "A yeast chromosomal origin of DNA replication defined by multiple functional elements," *Science*, 255, 817–823.

Mavrich, T. N., Jiang, C., Ioshikhes, I. P., Li, X., Venters, B. J., Zanton, S. J., Tomsho, L. P., Qi, J., Glaser, R. L., Schuster, S. C., Gilmour, D. S., Albert, I., and Pugh, B. F. (2008), "Nucleosome organization in the Drosophila genome," *Nature*, 453, 358–362.

Miller, J. and Widom, J. (2003), "Collaborative competition mechanism for gene activation *in vivo*," *Mol. Cell. Biol.*, 23, 1623–1632.

Mirny, L. A. (2009), "Nucleosome-mediated cooperativity between transcription factors," *arXiv.org online preprint*, arXiv:0901.2905v1 [q-bio.BM].

Miura, F., Kawaguchi, N., Sese, J., Toyoda, A., Hattori, M., Morishita, S., and Ito, T. (2006), "A large-scale full-length cDNA analysis to explore the budding yeast transcriptome," *Proc. Natl. Acad. Sci. U.S.A.*, 103, 17846–17851.

Morozov, A. V., Fortney, K., Gaykalova, D. A., Studitsky, V. M., Widom, J., and Siggia, E. D. (2008), "Extrinsic and intrinsic nucleosome positioning signals," *arXiv.org online preprint*, arXiv:0805.4017v1 [q-bio.GN].

Mukherjee, S., Berger, M. F., Jona, G., Wang, X. S., Muzzey, D., Snyder, M., Young, R. A., and Bulyk, M. L. (2004), "Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays," *Nature genetics*, 36, 13311339.

Murphy, K. P. (2002a), "Dynamic bayesian networks: representation, inference and learning," Ph.D. thesis, Citeseer.

Murphy, K. P. (2002b), "Hidden semi-markov models (HSMMs)," *unpublished notes*.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008), "The transcriptional landscape of the yeast genome defined by RNA sequencing," *Science*, 320, 1344–1349.

Nasheuer, H., Smith, R., Bauerschmidt, C., Grosse, F., and Weisshart, K. (2002), "Initiation of eukaryotic DNA replication: regulation and mechanisms," *Progress in Nucleic Acid Research and Molecular Biology*, 72, 41–94, PMID: 12206458.

Ostendorf, M., Digalakis, V., and Kimball, O. A. (1996), "From HMM's to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Trans. on Speech and Audio Processing*, 4, 360378.

Peckham, H., Thurman, R., Fu, Y., Stamatoyannopoulos, J., Noble, W., Struhl, K., and Weng, Z. (2007), "Nucleosome positioning signals in genomic DNA," *Genome Res.*, 17, 1170–1177.

Prasanth, K. V. and Spector, D. L. (2007), "Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum," *Genes Dev.*, 21, 11–42.

Ptashne, M. and Gann, A. (2002), *Genes & Signals*, Cold Spring Harbor Laboratory Press.

Rabiner, L. R. (1989), "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, 77, 257–286.

Raghuraman, M. K., Winzeler, E. A., Collingwood, D., Hunt, S., Wodicka, L., Conway, A., Lockhart, D. J., Davis, R. W., Brewer, B. J., and Fangman, W. L. (2001), "Replication dynamics of the yeast genome," *Science*, 294, 115–121.

Raveh-Sadka, T., Levo, M., and Segal, E. (2009), "Incorporating nucleosomes into thermodynamic models of transcription regulation," *Genome Research*, 19, 1480 –1496.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T. L., Wilson, C. J., Bell, S. P., and Young, R. A. (2000), "Genome-wide location and function of DNA binding proteins," *Science*, 290, 2306–2309.

Rippe, K., Schrader, A., Riede, P., Strohner, R., Lehmann, E., and Längst, G. (2007), "DNA sequence- and conformation-directed positioning of nucleosomes by chromatin-remodeling complexes," *Proceedings of the National Academy of Sciences*, 104, 15635 –15640.

Robison, K., McGuire, A. M., and Church, G. M. (1998), "A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome," *Journal of molecular biology*, 284, 241254.

Saul, L. K. and Jordan, M. I. (1995), "Boltzmann chains and hidden Markov models," in *Advances in Neural Information Processing Systems*, eds. G. Tesauro, D. Touretzky, and T. Leen, vol. 7, pp. 435–442, The MIT Press.

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

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I., Wang, J., and Widom, J. (2006), "A genomic code for nucleosome positioning," *Nature*, 442, 772–778.

Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., and Gaul, U. (2008), "Predicting expression patterns from regulatory sequence in *Drosophila* segmentation," *Nature*, 451, 535–540.

Sellke, T., Bayarri, M. J., and Berger, J. O. (2001), "Calibration of p values for testing precise null hypotheses," *The American Statistician*, 55, 6271.

Shivaswamy, S., Bhinge, A., Zhao, Y., Jones, S., Hirst, M., and Iyer, V. R. (2008), "Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation," *PLoS Biol*, 6, e65.

Sinha, S. (2006), "On counting position weight matrix matches in a sequence, with application to discriminative motif finding," *Bioinformatics*, 22, e454–463.

Steinmetz, E. J., Warren, C. L., Kuehner, J. N., Panbehi, B., Ansari, A. Z., and Brow, D. A. (2006), "Genome-wide distribution of yeast RNA polymerase II and its control by Sen1 helicase," *Mol. Cell*, 24, 735–746.

Stowe, K. S. (2007), *An introduction to thermodynamics and statistical mechanics*, Cambridge Univ Pr.

Suzuki, Y., Taira, H., Tsunoda, T., Mizushima-Sugano, J., Sese, J., Hata, H., Ota, T., Isogai, T., Tanaka, T., Morishita, S., Okubo, K., Sakaki, Y., Nakamura, Y., Suyama, A., and Sugano, S. (2001), "Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites," *EMBO Rep.*, 2, 388–393.

Tanay, A. (2006), "Extensive low-affinity transcriptional interactions in the yeast genome," *Genome Res.*, 16, 962–972.

Teif, V. B. (2007), "General transfer matrix formalism to calculate DNA-protein-drug binding in gene regulation: application to OR operator of phage lambda," *Nucleic Acids Res.*, 35, e80.

Teif, V. B. (2010), "Predicting gene-regulation functions: lessons from temperate bacteriophages," *Biophysical Journal*, 98, 1247–1256.

Van Holde, K. (1989), *Chromatin*, Springer series in molecular biology, Springer-Verlag, New York.

Wang, J., Fondufe-Mittendorf, Y., Xi, L., Tsai, G., Segal, E., and Widom, J. (2008), "Preferentially quantized linker DNA lengths in Saccharomyces cerevisiae," *PLoS Comput Biol*, 4, e1000175.

Wasson, T. and Hartemink, A. J. (2009), "An ensemble model of competitive multi-factor binding of the genome," *Genome Research*, 19, 2101.

Watson, J. D. and Crick, F. H. (1953), "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid," *Nature*, 171, 737–738, PMID: 13054692.

Weiner, A., Hughes, A., Yassour, M., Rando, O. J., and Friedman, N. (2010), "High-resolution nucleosome mapping reveals transcription-dependent promoter packaging," *Genome Research*, 20, 90.

Whitehouse, I., Rando, O. J., Delrow, J., and Tsukiyama, T. (2007), "Chromatin remodelling at promoters suppresses antisense transcription," *Nature*, 450, 1031–1035.

Wodicka, L., Dong, H., Mittmann, M., Ho, M. H., and Lockhart, D. J. (1997), "Genome-wide expression monitoring in Saccharomyces cerevisiae," *Nature Biotechnology*, 15, 13591367.

Wyrick, J. J., Aparicio, J. G., Chen, T., Barnett, J. D., Jennings, E. G., Young, R. A., Bell, S. P., and Aparicio, O. M. (2001), "Genome-wide distribution of ORC and MCM proteins in *S. cerevisiae*: high-resolution mapping of replication origins," *Science*, 294, 2357–2360.

Xu, W., Aparicio, J. G., Aparicio, O. M., and Tavar, S. (2006), "Genome-wide mapping of ORC and Mcm2p binding sites on tiling arrays and identification of essential ARS consensus sequences in *S. cerevisiae*," *BMC Genomics*, 7, 276.

Yuan, G. and Liu, J. (2008), "Genomic sequence is highly predictive of local nucleosome depletion," *PLoS Comput. Biol.*, 4, e13.

Yuan, G., Liu, Y., Dion, M. F., Slack, M. D., Wu, L. F., Altschuler, S. J., and Rando, O. J. (2005), "Genome-scale identification of nucleosome positions in S. cerevisiae," *Science*, 309, 626–630.

Zhao, Y., Granas, D., and Stormo, G. D. (2009), "Inferring binding energies from selected binding sites," *PLoS Comput Biol*, 5, e1000590.

Zhou, Q. and Liu, J. S. (2004), "Modeling within-motif dependence for transcription factor binding site predictions," *Bioinformatics*, p. 61.

Zhu, C., Byers, K. J., McCord, R. P., Shi, Z., Berger, M. F., Newburger, D. E., Saulrieta, K., Smith, Z., Shah, M. V., Radhakrishnan, M., Philippakis, A. A., Hu, Y., De Masi, F., Pacek, M., Rolfs, A., Murthy, T., Labaer, J., and Bulyk, M. L. (2009), "High-resolution DNA-binding specificity analysis of yeast transcription factors," *Genome Res.*, 19, 556–566.

Biography

Todd Steven Wasson was born on April 3, 1979 in Peoria, IL. At the age of five, he moved to Hudson, Florida, where he subsequently attended Hudson Elementary, Middle, and High Schools. He earned a Bachelors of Science in Computer Science with a minor in Mathematics from the University of South Florida in December of 2001. During his undergraduate career he focused on AI robotics research and computer vision. After graduation, he worked for some time for Lockheed Martin in Burlington, Massachusetts, before returning to a university setting to pursue his Ph.D. at Duke University in the Fall of 2004.

He entered graduate school in only the second class of the Bioinformatics and Genome Technologies Ph.D. program, which became the Program in Computational Biology and Bioinformatics. He also pursued a Masters of Science in Statistical Science, which he intends to receive concurrently with his Doctorate.

On August 1, 2010, he married Ana Paula de Oliveira Sales in the Sarah P. Duke Gardens on the Duke University campus.

He has accepted a position as a Postdoctoral Researcher in Knowledge Systems and Informatics at Lawrence Livermore National Laboratory in Livermore, California.

Publications:

Wasson, T. and Hartemink, A. J. (2009), "An ensemble model of competitive multi-factor binding of the genome," *Genome Research*, 19, 2101.