Motif Discovery Algorithms Incorporating Nucleosome Positioning Information

Azin Sayad-Rahim

School of Computer Science
McGill University, Montreal

August 2009

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

This thesis is dedicated to my mother, father and sister.

Acknowledgements

First and foremost, this work would never have been possible without the intellectual and financial support of Prof. Mathieu Blanchette. I have benefited greatly from his research direction, insightful discussions, encouragement and patience. Mathieu, you have my sincere and profound thanks.

Prof. Ted Perkins generously gave me the opportunity to work on interesting problems of a more theoretical nature. I gained much from our many discussions and his insightful observations.

I also thank Prof. Michael Hallett for introducing me to bioinformatics through his challenging computational biology courses. The classic papers we discussed provide a foundation I come back to regularly.

Prof. Doina Precup deserves a special mention for her excellent machine learning course, with its integration of theory and hands-on work, whose notes I still use in my work today.

Finally, I learned much from my many discussions and benefited greatly from the collaborative research atmosphere at the McGill Centre for Bioinformatics and its students, including but not limited to Anna, Emmanuel, Francois, Greg, Ian, Julian, Mike, Sarah and Sean.

Abstract

Transcription factor binding sites are essential components of the machinery that controls gene expression. In the absence of experimental data, computational approaches are used to predict binding sites based on promoter DNA sequence. However transcription factor binding depends not just on sequence but also the packaging of the DNA molecule. Nucleosomes, as the smallest unit of DNA packaging, affect transcription factor binding by obstructing protein-DNA interactions.

We use an empirically-derived relationship between binding sites and nucleosome positioning to augment an existing computational approach to predicting transcription factor binding sites. We demonstrate that the inclusion of experimentally-derived nucleosome positioning data improves the prediction capabilities of the basic computational approach using a large dataset of experimentally confirmed transcription factor binding sites.

Abrégé

Les sites de liaison de facteurs de transcription sont des composants essentiels du méchanisme de contrôle de l'expression génique. En l'absence de données expérimentales, les approches informatiques sont utilisées pour prédire les sites de liaison basée sur la séquence d'ADN promoteur. Toutefois la liaison de facteurs de transcription dépend non seulement de la séquence mais également de l'emballage biologique de la molécule d'ADN. Les nucléosomes, en tant qu'unité d'emballage de base de l'ADN, ont un effet marqué sur le positionnement des sites de liaison de facteurs de transcription.

Nous dérivons une relation empirique entre les sites de liaison et le positionnement des nucléosomes pour améliorer un algorithme de prédiction de sites de liaison. Nous démontrons que l'inclusion de données de positionnement de nucléosome améliore la performance de l'algorithme de base en utilisant un ensemble de données de sites de liaison confirmé expérimentalement.

Table of Contents

Table of Figures

Introduction

Elucidating the regulatory mechanisms of cells is one of the great challenges of computational biology. The size of the challenge is staggering. Cells have thousands of genes, each potentially producing several protein products which can potentially interact with dozens if not hundreds of other entities in the cell. All this forms a very complex regulatory code.

One locus of convergence for many of these proteins is at the promoter regions of genes. A myriad of proteins interact among themselves and with DNA to orchestrate the timely and complex regulatory code for each gene. The sheer scale of the challenge makes comprehensive experimental analysis infeasible. Even the most advanced high-throughput experimental techniques can investigate, in an error-prone fashion, the interactions of one protein product with diverse DNA sites. These protein-DNA interactions are referred to as transcription factor binding sites.

The size of the problem and scarcity of experimental data makes computational approaches attractive. The ultimate promise of such approaches is the reliable prediction of DNA docking sites for transcription factors. However, the current state of such approaches is very far from these ultimate ends.

We approach this problem from the premise that the incorporation of additional biologically relevant data is likely to improve the outcome of computational approaches. We use DNA packaging information to enhance basic DNA sequence data as the source of binding site predictions.

We proceed as follows. We start with an overview of the relevant literature related to the biology and experimental approaches used. We continue with the joint analysis of two genome-scale datasets, deriving a detailed relationship between DNA packaging and transcription factor binding sites. We then present a computational approach for predicting binding sites, as well as an extension to accommodate DNA packaging information. We then test the performance of the extended algorithm against the basic version using genome-scale experimental data.

Chapter 1 – Biological Preliminaries and Experimental Processes

In this chapter, we survey the biological foundations and other preliminaries of this work. We begin with an overview of the relevant molecular and cellular biology, before moving on to some more recent comparative and experimental techniques and basic computational representations of sequences.

1.1. Fundamentals of Cellular Organization

The eukaryotic cell is the smallest structural unit of an organism capable of independent functioning. It is structured internally by various organelles including the nucleus, cytoplasm and mitochondria. The nucleus and mitochondria of the cell contain DNA, the encoded blueprint for all aspects of a cell's functioning and components (Lodish, et al. 2004). We will primarily focus on DNA contained in the nucleus.

Nuclear DNA consists of several long, contiguous chains of four distinct nucleotides. These are Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). When considered from an information storage perspective, DNA can, to a first approximation, be conceived of as a string encoding information using a four-letter alphabet. Diverse DNA sequence features encode information. The most prominent of these is a gene, a contiguous stretch of DNA, typically thousands of bases long, encoding for at least one protein molecule.

Proteins are possibly the most common and diverse functional components of a cell. These molecular machines interact with virtually all other aspects of the cell and are critically involved in all functions. They are large molecules consisting of several hundred or thousand sequentially linked building blocks called amino acids. There are most commonly twenty amino acids. At the DNA level, each amino acid is encoded by at least one nucleotide triplet, called a codon.

The functionality of a protein is derived from its three dimensional structure. This structure is a consequence of the specific amino acid chain composition of the protein. However, the process by which the protein spontaneously folds into its structure is highly complex and still poorly understood. The bewildering diversity of proteins provides many different structures, which interact and interlock to form protein-protein complexes. Like the protein, these can be thought of as complex molecular machines that perform different roles in the functioning of the cell.

It is important to note that the DNA chains encoding the above information are highly structured. They are packaged and compacted into chromosomes. For example, Saccharomyces Cerevisiae, the most widely studied unicellular eukaryote, has fourteen chromosomes. Humans have twenty-three pairs. Chromosomes are only the highest level of DNA packaging. At different scales, DNA is intricately organized. We will delve into the details of this organization further in the coming pages.

There is a critical, multi-step process by which proteins are produced from their genetic blueprints, known as the central dogma of molecular biology. The simplest version of this process is that DNA is converted into RNA via transcription. RNA is, in turn, translated into protein. This simple sketch of the process leaves out many complications and caveats, but will suffice for our purposes. We will focus on this first crucial step. Transcription starts when the RNA polymerase protein complex copies a template strand of DNA into a similar molecule called RNA, which is subsequently translated into the amino acid chain forming a protein.

1.1.1. Transcription Regulation

Gene expression is modulated at different stages by diverse mechanisms. However, most expression regulation occurs at the transcription level (Wray et al. 2003). Genes are flanked by DNA regions known as promoters. These regions are most commonly upstream of the gene, and crucially enable the complex regulation of the neighboring gene. The promoter region allows for the integration of information about the status of diverse cellular processes, "encoded" in the environmental presence, absence and variation of molecules and occurrence of reactions. The promoter region can be thought of as the control switch control of a gene.

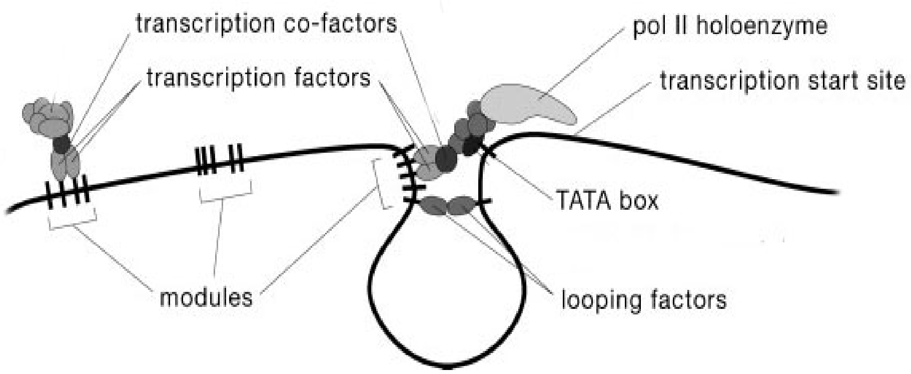


Figure 1 - Regulation of transcription, from the binding of transcription factors to the recruitment of the DNA polymerase II complex and transcription initiation (Adapted from Wray et al. 2003)

Promoter regions have no known fixed delimiters, except perhaps the boundaries of other genomic features such as genes (Wray et al.). They are characterized by the significant enrichment of short, functionally significant sequences affecting the transcriptional regulation of a gene. Lengths vary widely among eukaryotic organisms, ranging from several hundred bases in Yeast to over a hundred kilobases (kb) in Human (Lodish, et al.). In uncommon cases, adjacent genes facing outward on different DNA strands may partially or totally share a promoter region. The immediate promoter region is a few hundred bases long upstream of the transcription start site. Other regulatory regions can be located much further 100 kb upstream in humans. This expanded region containing regulatory elements is known as the enhancer region of a gene.

The genetic basis for transcription specificity is a function of both promoter nucleotide sequence, and other genome segments coding for proteins called transcription factors, which we examine below. Important and frequently observed sequence features of a promoter include the TATA box, named after its signature sequence. This short sequence often occurs a few dozen bases upstream of the transcription start site of many genes. It is the anchor site for the RNA polymerase II complex.

1.1.1.1. Transcription Factors and Binding Sites

Transcription factors (TFs) are proteins that help the RNA polymerase locate promoter regions, initiate transcription and change overall gene transcription rate (Lodish, et al.). These proteins attach to DNA in the vicinity of the gene to be transcribed. The protein-DNA interactions enabling this attachment require nucleotide patterns highly specific to each individual TF, a DNA signature identifying the docking site. These docking sites are called transcription factor binding sites (TFBS).

The nucleotide sequence is the single most important determinant of transcription factor binding site function. It is the specific protein-DNA interactions with these nucleotides that modulate transcription factor binding. Ten to twenty percent of nucleotides in well-studied promoters are part of binding sites, though this may vary depending on the specific organism and gene studied (Lodish, et al.).

Nucleotide sequence is, however, insufficient to identify functionally active binding sites from candidate DNA sites closely matching the pattern. Certain functionally inactive sites may be vestiges of previously functionally active sites modified through evolution. Additionally, given the length of genomic DNA, similar nucleotide sequences may occur randomly, particularly in the case of shorter TFBSs. Finally, binding sites may be functional only under a specific set of environmental circumstances potentially difficult to discover (Lodish, et al., Wray et al.).

For all these reasons, experimental validation of putative binding sites is often necessary to confirm the functional significance of a site. Traditional experimental studies are laborious and time consuming, thus infeasible for large-scale confirmation. Several high-throughput techniques, to be discussed in following sections, address this issue.

It is also important to distinguish between the actual binding site of the transcription factor, which typically ranges between five to eight bases, and the binding site footprint, which typically varies between ten and twenty bases. It is often the binding site footprint that is identified through experimental or computational means, whereas the bound nucleotides form a subset of this footprint (D'Haeseleer 2006).

Furthermore, binding sites may endure one or more nucleotide substitutions without loss of function. Such changes may however change the strength of the overall transcription factor-DNA interaction by changing the quality or number of protein-DNA interactions. The strength of such interactions is broadly associated with the magnitude of the activation or repression effect exerted by the factor. It is a counter-intuitive fact that the magnitude of activation or repression required for cellular functions is far from maximal, often ranging between 1.3 to three-fold (Blanchette 2006). Thus in many biologically relevant binding sites, the strength of binding is significantly below the maximum possible protein-DNA binding energy. We also note in passing that post-translational modifications of the transcription factor in some cases alter its binding strength and specificity (Lodish, et al.).

Transcription factors can act in concert with other proteins to bind DNA, depending on whether two identical or different proteins join in a complex to interact with a binding site. These complexes are called, respectively, homo-dimers and hetero-dimers. As noted in (Giguere 1999), nuclear receptors are important examples of transcription factors functioning as dimers, with different dimer combinations altering the binding specificity of the nuclear receptor.

1.1.1.2. Cis-Regulatory Modules

TFs are fundamental elements of the regulation of transcription, but they are also constituents of a higher order regulatory mechanism. TFBS are organized into functionally discrete modules, typically comprising six to fifteen TFBS for four to eight TFs, clustered across up to several hundred bases (Lodish, et al., Wray et al.).

The TFs of a module interact and form protein complexes that affect transcription in various ways. Some effects include: transcription initiation, increasing or decreasing of transcription rate in response to specific spatial and temporal conditions, mediating extra-cellular signals, and restricting the effects of other modules. The combinatorial nature of modules allows them to encode a vast variety of complex functionally relevant information.

It is significant that module functions are most often discrete. In particular, when one module is deleted via DNA mutations or other experimental technique, other modules continue functioning. Thus they add another layer of combinatorial transcriptional control. We will not explore the topic of modules further here, as they are not a primary focus of our work.

1.1.1.3. Chromatin Structure

In our previous discussion, we implicitly treat DNA as a strand analogous to a string. In the cellular environment, however, it is in a packaged and highly compacted state known as chromatin. Chromatin is bundled into chromosomes. This packaging is organized hierarchically on different levels corresponding roughly to different size scales. At the lowest level, chromatin is composed of single nucleosomes.

A nucleosome is the combination of a fixed length of DNA, 146-147 bases, wrapped approximately 1¾ turns around a bundle of eight proteins called the histone octamer (Lowary and Widom 1997, Richmond and Davey 2003, Khorasanizadeh 2005). In what follows we'll take 147 bases as the length of nucleosomal DNA. The histone octamer consists of two units of each of the histones H2A, H2B, H3 and H4. We'll note in passing that, under certain conditions, variants of certain histones, such as the H2A variant H2A.Z, may constitute be incorporated in the nucleosome instead.

Nucleosomes are separated by stretches of free DNA called linker DNA. Linker lengths vary, but are often between five to one hundred bases (Lodish, et al.). A string of pearls is a good visual analogy for such packaged DNA, with nucleosomes (pearls) separated by lengths of linker (string).

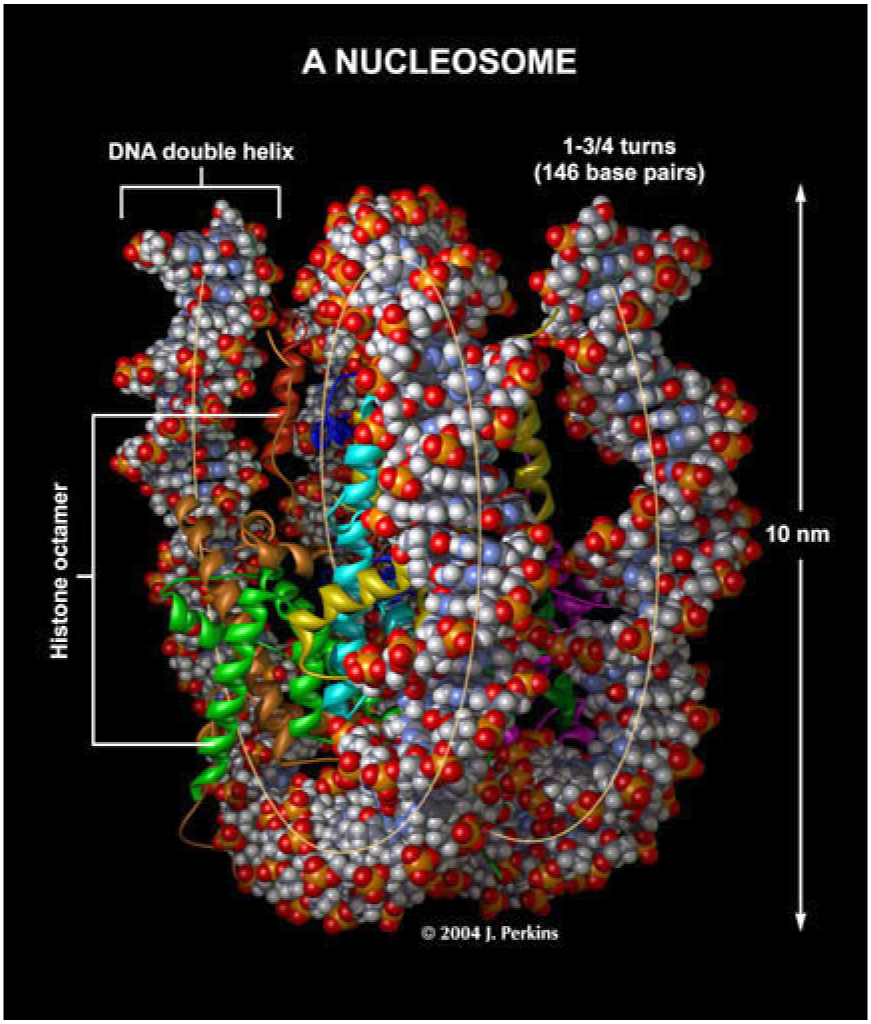


Figure 2 - At the lowest level, DNA packaging consists of 146-147 bp of DNA wrapped around a histone octamer. Each such bundle is separated by short lengths of linker DNA. Adapted from (Perkins 2004).

Given the fundamental importance of the packaging mechanisms, it is not surprising that nucleosome structure is ubiquitous among eukaryotes from yeast to human, as are structurally similar and functionally equivalent histone proteins.

It is important to that histones in a nucleosome actively modulate access to their wrapped DNA. This typically occurs via post-translational modifications of specific histone amino acid, most commonly acetylation or methylation. Acetylation is most often associated with increased transcription levels, while the role of methylation is mixed: certain histone methylation patterns enable gene transcription while others repress it. In all cases these modifications directly or indirectly lead to a loosening or tightening of the DNA packaging (Geiman and Robertson 2002).

1.1.1.3.1. Chromatin Compaction

As previously noted, key function of chromatin is to modulate access to the packaged DNA. Molecules, most notably proteins, have more or less access depending on how tightly nucleosomes are compressed together into higher order packaging, and on how strongly DNA is attached in individual nucleosomes.

At one end of the spectrum, and by default, chromatin is tightly packaged. This prevents most or all molecule access to the DNA, effectively preventing protein-DNA interactions required for transcription. This tightly packaged state is characterized by hyper-methylated DNA and certain specific methylation marks on histone residues. The converse reaction, acetylation, makes DNA more accessible and thus enables higher transcription rates. Methylation and acetylation are mechanisms for durable modification of gene transcription and thus expression levels (Geiman and Robertson).

1.1.1.3.2. Nucleosome Formation

As mentioned, the nucleosome is the fundamental packaging unit of chromatin. Genomic DNA has a natural affinity for nucleosome formation (Lowary and Widom 1997, Lowary and Widom 1998). That is, if histone octamers are introduced into a solution containing random DNA, nucleosomes will spontaneously form. However, of the large variety of sequences are packaged as nucleosomes, only a small fraction (5-10%) demonstrate a significantly higher affinity for spontaneous nucleosome formation when compared t arbitrary DNA (Lowary and Widom 1998). It is thus infeasible that strongly sequence-specific protein-DNA interactions are a pre-requisite for nucleosome formation.

However the observed high-affinity sequences, or similar ones, can be thought of as positioning signals that reliably place nucleosomes at certain DNA positions in-vivo (Lowary and Widom 1998, Segal et al. 2006). Given the impact of nucleosomes on DNA accessibility, completely random positioning would be highly detrimental to organisms given the level of randomness this would introduce into such processes as transcription.

The affinity for of a DNA sequence for nucleosome formation is a function of the energy required to morph this sequence from its equilibrium state to the tightly wrapped nucleosomal state. The relative difference in nucleosome-formation affinity between two stretches of DNA depends on the difference in their respective bending energy requirements. This energy in turn depends on the mechanical properties of the given 147 base length of the DNA molecule: inherent bendedness in a direction, bendability, inherent twist and twistability (Widom 2001, Scipioni et al. 2004).

To a first approximation ignoring higher order effects, mechanical properties of a stretch of DNA molecule as a whole can be reduced to the mechanical properties of its sequence of constituent di-nucleotides (Anselmi et al. 2000, Scipioni et al. 2004, Wiggins et al. 2005). In particular, certain di-nucleotides are over-represented at periodic intervals in sequences that are particularly favorable to nucleosome formation (Lowary and Widom 1998, Widom 2001, Segal et al. 2006). The period of ten or eleven bases corresponds to a complete helical turn of DNA. Examples of these di-nucleotides include AT/TA and AA/TT. These can also be accompanied by a periodic occurrence of CG/GC di-nucleotides, shifted by five bases from the AT. Thus, the mechanical properties of DNA are partially reflected in certain sequence patterns.

1.1.1.3.3. Nucleosome Positioning Signals and the Parking Lot Model

The above considerations inform the following model of nucleosome positioning. The key idea is that favorable nucleosome positioning sequences occur at regular intervals, but not for each individual nucleosome. Certain sequences preferentially form nucleosomes, and these positions constrain the possible positions of other nucleosomes by exclusion. This is referred to as the parking lot model, in analogy to a parking lot where the positions of already-parked cars constrain the position of new cars to free parking spots.

The above model is evident, for example, in the human Beta-Globin locus, inherently curved DNA occurs frequently in periods of 680 bases, or the length of four nucleosomes and the additional linker DNA (Kiyama et al. 1999). This periodicity has been plausibly been linked to positioning of tetra-nucleosomes (Makeev et al. 2003). Additional periods occur in multiples of 170 bases, corresponding to the length of one nucleosome and linker DNA. This pattern suggests that sequences at regular intervals preferentially position nucleosomes, and thus constrain and determine the position of other nucleosomes.

1.1.1.3.4. Types of Nucleosome Positioning and Dislocation

There exist two types of nucleosome positioning: translational and rotational (Widom 2001). Translational, or strongly positioned, nucleosomes occupy a defined position along the DNA strand. In contrast, rotational, or weakly positioned, nucleosomes shift to occupy a range of adjacent positions on the same DNA strand through time.

Given the unspecific behavior of weakly positioned nucleosomes, it is not surprising to find that Yeast promoters regions are enriched for strongly positioned nucleosomes (Yuan et al.). The authors also report enrichment for seemingly weakly positioned nucleosomes in promoter regions of highly transcribed genes. They hypothesize that such nucleosomes appear weakly positioned because they are being transiently dislocated by the transcription machinery as it initiates and proceeds with transcription.

Nucleosomes can be actively shifted or moved off DNA by ATP-powered protein machines (Geiman and Robertson 2002, Miller and Widom 2003, Strohner 2005). Furthermore, although DNA is wrapped around the histone octamer, it unwraps and rewraps spontaneously several times per second (Li et al. 2005). Transient DNA accessibility on such timescales is sufficient for proteins to gain access to the previously wrapped stretch of DNA, and prevent the rewrapping by passive obstruction, without recruiting an active mechanism for nucleosome dislocation.

However, passive access to buried DNA and subsequent prevention of rewrapping still requires energy. Thus there is an entry barrier. In this context, (Miller and Widom 2003) note an interesting mechanism of collaborative competition. In such situations, two proteins cooperate for access to the same buried stretch of nucleosomal DNA, and “compete” with the histone octamer for access to this DNA. For example, if one protein is already wedged in the nucleosomal DNA and holding it open, another protein can access a binding site buried in this DNA with far less energy expenditure. Thus access is increased if, for example, two binding sites are within the same stretch of nucleosomal DNA.

As a final point, it is worth noting that a given nucleosome can occupy a range of positions, according to some probability density function, in a population of cells. Thus, for example, we may often expect to see a well-positioned nucleosome close to the transcription start site of a certain gene, but the exact position may vary somewhat among cells in a population.

1.2. Experimental Data

As described above, transcription factor binding is a highly complex process crucially affected by such factors as chromatin compaction, spatial and temporal TF expression patterns and co-factor expression to name a few.

The ideal scenario for testing the quality of any TFBS prediction algorithm would be to check whether the predicted sites are bound in the cell as it functions. This is not possible quite yet. However, high-throughput experimental techniques and whole genome comparisons provide large-scale data that may be used to evaluate the quality of predictions on biological data.

The experimental techniques used in the generation of this data take cells in a given state, extract the genetic contents and perform analyses on these. Another approach used below, comparative genomics, looks at the similarities and differences between species, using the principle that genetic conservation is a surrogate for functional significance. As we will note, however, such experimental datasets are highly noisy.

Experimental data comes from S. Cerevisiae, or baker’s yeast. It is a widely studied and readily available model eukaryotic organism where many of the same fundamental processes, in particular transcription and DNA packaging, are fundamentally similar to other eukaryotes. The yeast genome is also well sequenced, annotated and studied, as are yeast TFs. Two notable differences are short promoter regions, which are typically less than a thousand bases long, and the lack of modules of binding sites. Indeed, the length of yeast promoters makes them comparable to a single cis-regulatory module containing multiple binding sites.

1.2.1 Techniques Used in Generating Experimental Data

Below we will briefly discuss the three major tools used in the generation of large-scale experimental TFBS and nucleosome positioning data in the yeast genome. This will help us better understand the quality limitations and noise inherent in such data.

1.2.1.1 Comparative Genomics

Comparative genomics relies on the fact that genetic mutations in functional DNA sequences are most often deleterious, and are thus selected against, whereas non-functional DNA can mutate freely. By comparing tracts of DNA between similar species, islands of high conservation are identified. Most recently, with the availability of whole genomes for several species, such comparisons are possible across entire genomes (Rubin et al. 2000).

Such comparisons are particularly powerful if the conserved sequences are short, such as TFBSs in orthologous promoters. While statistically significant cross-species conservation is not in itself proof of in-vivo function, it significantly raises our confidence in the functional importance of such regions. In conjunction with other techniques, this information can significantly reduce false positives and increase predictive accuracy.

1.2.1.2 Microarrays

When two complementary single stranded fragments of DNA collide in solution, they will spontaneously hybridize. If we were aware of the genomic position of one such fragment, and it were sufficiently long as to be probably unique in a genome, we could confidently deduce that the second fragment also originated at the same genomic position.

Microarrays use this powerful idea on a massive scale. They are solid surfaces to which are attached, in ordered fashion, up to tens of thousands of DNA strands with known genomic positions. These are called probes, and many copies of each probe occur on each array. Once an array is ready, solution containing DNA strands to be examined is applied to the array surface, and complementary strands allowed to hybridize. The amount of hybridization at each probe that hybridizes is related to the number of complementary DNA fragments in the solution (Stoughton 2005).

Hybridization intensities are read using optical scanners. The images are processed and intensity data extracted. This data is then analyzed statistically to mitigate errors, biases and variability and arrive at an estimate of the relative abundances of specific DNA fragments of interest in the applied solution.

There are two major types of microarray: dual-hybridization arrays and oligonucleotide arrays. We will focus on the first type.

1.2.1.2.1 Dual-Hybridization Arrays

In dual-hybridization arrays, spots on the array are printed with specific probes. However, two different samples of DNA are hybridized to each. One sample is dyed green (Cye3) and the other dyed red (Cye5). Both samples hybridize with the probes. When the unhybridized portion is washed away, a certain ratio of green vs. red light is read at each probe. This represents the ratio of one sample hybridizing compared to the other. In other words, the relative instead of absolute hybridization are read.

1.2.1.2.2 Problems and Limitations

Although microarrays are widely used, they have significant problems. First is their significant cost. Given that microarray experiments are inherently noisy, we need to reproduce each array experiment. High cost leads to low numbers of replicates, which in turn reduces the statistical power used to determine the hybridization values. Lack of adequate quality control, especially for dual-label arrays, is also an issue leading to increased noise.

Hybridization values are also analyzed based on images, whose processing can also lead to increased error. The dyes also have somewhat different chemical properties leading to differing levels of luminescence for the same number of hybridized strands. Finally, the statistical techniques used to process the raw results have a large impact on the final reported results.

1.2.1.3. ChIP-Chip

Chromatin Immuno-Precipitation on Chip, commonly called ChIP-Chip, is a recent high-throughput technology to identify in-vivo protein-DNA interactions (Buck and Lieb 2004). It is infeasible to observe such interactions in a functioning cell without massively disrupting the environment and hence the interaction. The solution offered by this technique is to freeze transient interactions as they occur and subsequently infer protein-DNA interactions occurring in the cell at that time.

More specifically, we begin by obtaining cells where the appropriate protein-DNA interaction is taking place. This may be performed experimentally, for example, by subjecting cultured cells to appropriate external stimuli.

Next, the cells are cross-linked. They are treated with UV light or Formaldehyde to durably attach the proteins to the DNA strands they transiently interact with. Once complete, the cells are ground to pieces using sonic shearing.

The cellular DNA fragments, with attached proteins, are gathered and immuno-precipitated. For this process, it is a pre-requisite to have an antibody that specifically targets our protein of interest. A column is packed with beads coated with the specific antibody, and the DNA fragments are precipitated through while the beads remain in the column. The anti-bodies will attach to protein cross-linked DNA fragments, preventing them from eluting from the column. Emptying the column and removing the beads provides us with the DNA fragments of interest.

Reversing the cross-linking and removing proteins provides us with a set of DNA strands.

We know these strands interact with the protein of interest. The next major step is to pinpoint their genomic position and thus to position the protein-DNA interactions on the genome. For this purpose dual-hybridization microarrays containing probes from the entire genome are used. The protein-interacting strands are dyed one color while the non-interacting strands that washed out of the column are dyed another. They are then applied to the microarray. Using the procedures previously described, the positions of the protein-interacting strands are discovered.

1.2.1.3.1 Problems and Limitations

Several factors complicate this process. Given that protein binding is ephemeral, and is related to environmental conditions and stage in the cell life cycle, only a fraction of the protein interactions that may occur under a set of environmental conditions will be captured by cross-linking.

The precipitation step is also error-prone, with protein-attached strands eluting from the column and some non-attached strands remaining. This is only compounded by the noise inherent in the microarrays used to determine the origin of the protein-bound DNA fragment.

Finally, even if the DNA fragment is positioned in the genome, we still do not have an exact position for the binding site. This length of the fragment is highly variable due to the random nature of cell and DNA fragmentation. Furthermore, the microarray probes are often several hundred bases in length. All we learn from the microarray is that the fragment is a complementary sub-sequence of the probe. Thus we have a range of several hundred positions within which an experimentally verified protein binding occurs. We must find the precise position by other means.

1.2.2 Experimental Binding Site Positions in Yeast

Having set the preliminaries, we examine how (Harbison et al. 2004) obtained genome wide experimentally supported transcription factor binding sites for yeast. The authors began with thirteen distinct conditions, and two hundred TFs identified from the YPD and MIPS databases.

For each growth condition and TF, a special strain of yeast was created. This strain contained a MYC epitope inserted into the coding region of the TF in question. The epitope is the coding sequence of a recognized protein structure that will be appended to the coding sequence of the TF, thus adding an additional structure to the TF that serves as a recognition tag.

The yeast strains were cultured and analyzed via ChIP-Chip using an anti-body recognizing the MYC. The DNA fragments were then hybridized against microarrays with probes covering all yeast intergenic regions. Such probes ranged between 48 and 1500 bases, with an average length of 480.

Once binding sites were mapped to a limited position range, a computational approach was adopted to identify the exact position. A battery of established motif finding programs were applied to the grouped DNA sequences containing binding sites for a given TF. In addition, the findings of the programs were integrated using a statistical model that assigned a p-value to each set of predictions.

The result of the above process was a copious list of experimentally strengthened predictions of TFBS across the yeast genome. The list contains over two hundred thousand positions for some eighty three factors, each assigned a score based on strength of ChIP-chip evidence, consensus computational motif prediction and evolutionary conservation in closely related species.

1.2.2.1 Issues Regarding Predictions

In addition to the issues encountered with ChIP-chip experiments, the introduction of MYC epitope tags can alter or disrupt TF production and function. Adding the new structure to the protein can change its structural conformation, potentially altering or disrupting interactions with other proteins or binding sites.

In addition, the microarray hybridization experiments had only 3 replicates, which is an insufficient number to provide much statistical power when analyzing highly noisy raw microarray data, and thus significantly lowers prediction quality.

1.2.3 Experimental Nucleosome Position Data in Yeast

Next, we consider how experimental nucleosome position data was obtained by (Yuan et al.). Yeast cells were cross-linked with formaldehyde to cross-link the histone octamer to bound DNA. The cellular DNA was extracted and digested using micrococcus nuclease. This nuclease requires access to DNA in order to digest it, and thus removes linker DNA stretches while leaving DNA obstructed by the histone octamer intact. The cross-links were reversed and histone bundles removed, leaving only the nucleosomal DNA.

This DNA was dyed green, mixed with whole genome DNA fragments dyed red, then competitively hybridized against a microarray covering most of yeast chromosome three, as well as 227 additional promoter regions of 1000 bases. These microarrays had 50 base long probes that tiled the regions every 20 bases. That is, probes from adjacent regions of DNA had 30 bases of overlap.

The raw microarray values were processed to mitigate biases, and used as input to a Hidden Markov Model trained to predict, based on a probability threshold, whether a probe is hybridized with nucleosomal or linker DNA based on the color read from the array. This was supplemented with human examination of ambiguous cases and calls as to their state. Thus 20 base lengths of sequence in the regions were assigned either as nucleosomal or linker DNA.

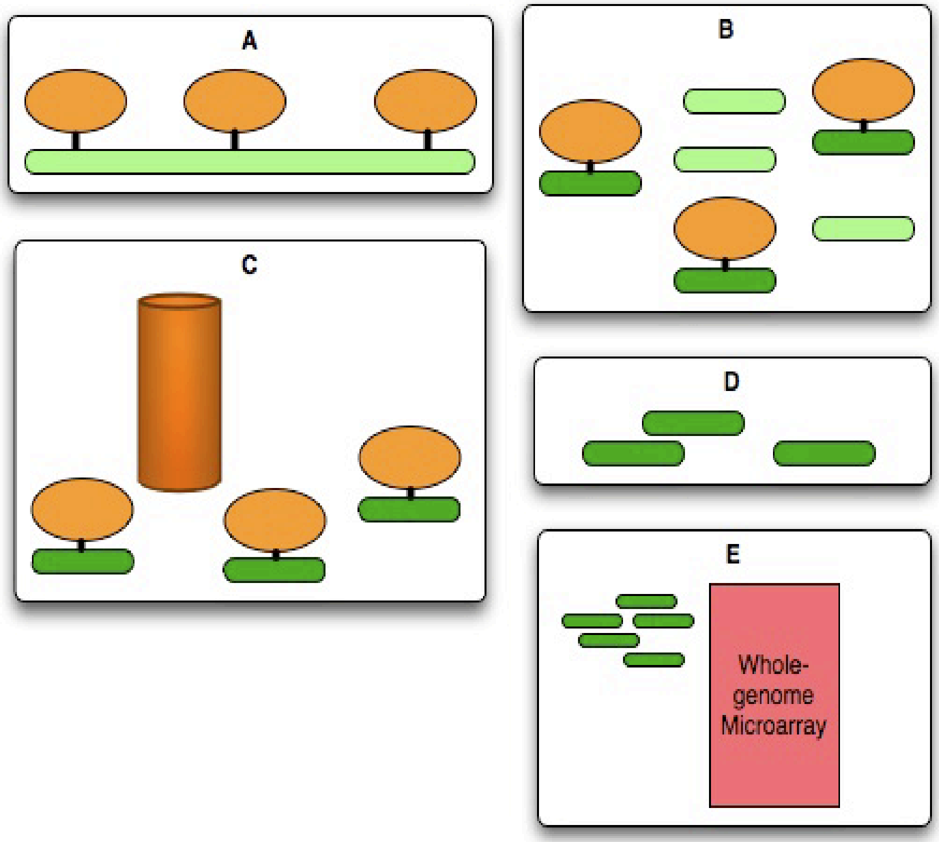


Figure 3 - ChIP-Chip technology allows the discovery of DNA binding sites for a specific protein through a multi-step process. A) Cross-link histone proteins to wrapped nucleosomal DNA using phomeldahyde. B) Fragment DNA using nuclease - DNA wrapped in nucleosomes is highly resistant to fragmentation. C) Run the DNA fragments through an immunoprecipitation column, where beads coated with antibodies engineered to bind to histone proteins retain the nucleosome-wrapped DNA fragments. D) Reverse the cross-linking and remove histone proteins, retaining only the nucleosome-wrapped DNA fragments. E) Amplify the DNA fragments and hybridize them against a tiling microarray to determine the origin of the DNA fragments along the genome.

1.2.3.1 Issues Regarding Predictions

Apart from the normal caveats applied to microarray data, digestion is a stochastic process and does not digest all linker regions, or leave all nucleosomal DNA intact. Also, the use of a predictor to classify nucleosomal and linker DNA introduces certain questions. The model is trained with data allowing it to recognize different DNA types based on hybridization color, which is itself noisy data.

Additionally DNA from a collection of cells is used. These cells are in different stages of the cell cycle, and contain natural variations in nucleosome positions. In particular, at certain times during the cell cycle, DNA is not bundled in nucleosomes. Furthermore, a cell's transcription machinery actively expels histone bundles from the DNA being transcribed. These will appear as linker DNA or delocalized nucleosomes when they are in fact well-positioned nucleosomes actively moved in the normal functioning of the cell.

Finally, human examination and calls based on probabilities emitted by a Hidden Markov Model add further potential for error in the cases where judgment was used. All these factors combined, the predictions of Yuan et al. should be considered as a rough estimate of nucleosome positions rather than a highly accurate map.

1.3. Computational Representations of DNA Sequences

1.3.1. Markov Models

Markov models are stochastic processes whose future state depends only on a finite number of past states. More formally, a Markov model of order m is a sequence of random variables $X_1, X_2, ..., X_n$ where

$$\begin{aligned} &\Pr(X_n = x_n \mid X_{n-1} = x_{n-1}, ..., X_2 = x_2, X_1 = x_1) \\ &= \Pr(X_n = x_n \mid X_{n-1} = x_{n-1}, ..., X_{n-m} = x_{n-m}) \end{aligned}$$

In other words, the next variable depends exclusively on the previous m variables. Markov models are useful as a model for DNA sequences. In such cases, each random variable is a nucleotide, and the occurrence of a nucleotide is dependent on the preceding nucleotides in the sequence. Given a Markov model, it is possible to estimate the probability of the occurrence of an observed nucleotide sequence according to the Markov model. A key step in the definition of a Markov model is the specification or calculation of the conditional properties for the occurrence of each nucleotide.

1.3.2. Regular Expressions

One way to model nucleotide sequences, including binding site motifs, is by representing them as regular expressions. In this case, the regular expression indicates valid patterns for a nucleotide sequence. The matching criterion is binary; either a sequence matches a pattern or not. This can be problematic for binding sites, whose patterns reveal significant levels of degeneracy. Different nucleotides in a binding site are not created equal: positions range from contributing little to the protein-DNA interaction, to being critical to its occurrence. Conservation of nucleotides at the latter positions will be far greater, but still not perfect. However, one can still detect clearly dominant nucleotides. The regular expression approach does not allow for this nuance. Implicitly, all possible nucleotides at a given position are given equal weight – one cannot specify that a critical nucleotide is the overwhelming favorite.

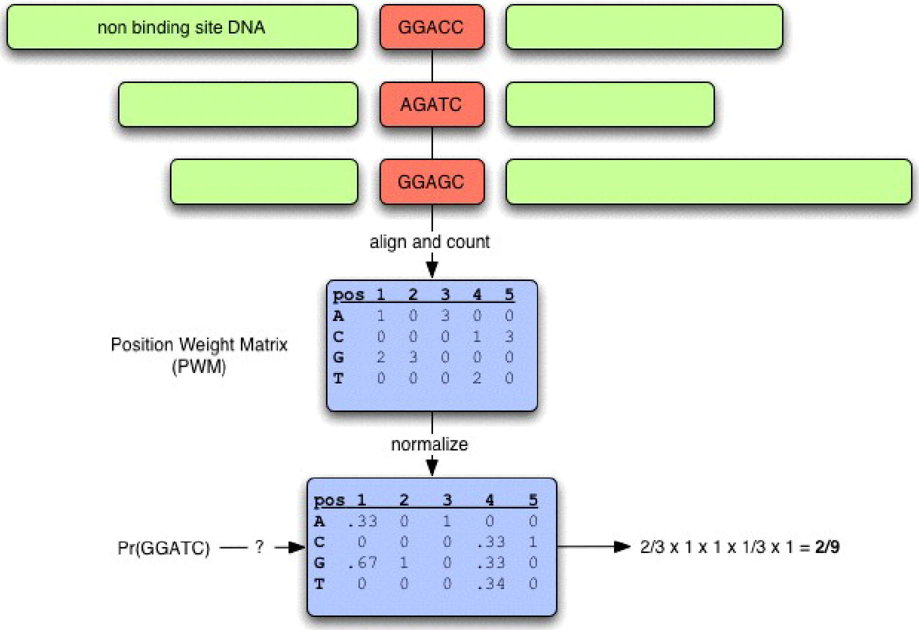
1.3.3. Position Weight Matrices

In the previous section we discussed some of the limitations of a regular expression based approach to sequence modeling. Here we describe an alternative model for DNA sequences, and binding sites in particular, addressing some of these limitations.

A Position Weight Matrix (PWM) assigns a probability to the occurrence of each letter in the alphabet under consideration (in our case, the DNA alphabet *{A,C,G,T}*). Some characteristics of PWMs are:

- Fixed length: the PWM only models sequences of a specified length. In our case, this length *W* is the sequence length of a TFBS.

- Positional independence: each position in the model is assumed independent of all others. So the nucleotide in the second position is assumed independent of the first, and so on. In particular, this does not model certain position-specific dependencies within a TFBS. For example, the first and second nucleotides in a TFBS may be tightly inter-related in-vivo.



pos	1	2	3	4	5
A	1	0	3	0	0
C	0	0	0	1	3
G	2	3	0	0	0
T	0	0	0	2	0

pos	1	2	3	4	5
A	.33	0	1	0	0
C	0	0	0	.33	1
G	.67	1	0	.33	0
T	0	0	0	.34	0

Figure 4 – Deriving a PWM from known TFBS data. Known instances of a TFBS are aligned. The nucleotide frequencies for each column are counted and normalized to one in order to obtain probabilities for each column, or position, in

the model. If certain counts are zero, it is possible to add pseudo-counts, for example 0.1, to that position.

1.4. Computational Approaches to Motif Discovery

1.4.1. Enumerative and Probabilistic Approaches

Computational DNA motif discovery is a difficult but well-studied problem. Broadly speaking, existing approaches can be separated into two categories: those relying on exhaustive enumeration of substrings occurring in the input data, and those applying probabilistic modeling to "learn" a motif from the sequence data. What follows below is a brief description of some common approaches, as well as a short comparison between our approach and similar ones recently appearing in the literature.

Enumeration based approaches, including Phylogenetic Footprinting (Blanchette and Tompa 2002), YMF (Sinha and Tompa 2003) and WEEDER (Pavesi et al. 2001) represent motifs as fixed-length words with a certain allowed number of mismatches between instances. Identified instances are grouped and significance is calculated via various methods, such as simulation-based p-values or Z-score, to determine motif over-representation.

Probabilistic approaches including MEME (Bailey and Elkan 2004) and Gibbs sampling (Lawrence et al. 1999) model input sequences as a combination of motifs and background sequence. Motifs are typically represented by PWMs, whereas background sequence is often modeled using MMs or similar probabilistic model. Additionally, an iterative training method such as expectation maximization or Gibbs sampling is used to sample model starting points and learn the parameters.

Beyond the broad distinctions sketched above, several recent approaches have tackled the problem of motif discovery by incorporating various types of prior information known to be functionally relevant to TF binding in vivo. We discuss two in particular, and compare them to the approach we adopt in the coming chapters.

1.4.2. The BayesMD Algorithm

We first consider BayesMD (Tang et al. 2008). This approach is of particular interest as a flexible means of incorporating various types of prior information. The authors adopt a Bayesian modeling framework, using prior distributions to encode the uncertainty inherent in various model parameters. Bayesian inference techniques are used to derive model predictions. In contrast, as further discussed in the coming chapters, we adopt a frequentist approach, using expectation maximization to derive model parameter estimates.

Of note, the authors apply a mixture-modeling approach to represent a motif as a mix of highly conserved and poorly conserved positions. The model parameters are learned using TFBS data from transcription factor databases. This modeling of structure inherent within a TFBS is in addition to modeling the input sequences as a mix of motif and background sequence. In our approach, the PWM assumes independence between different positions within a motif.

Also interesting is the proposed approach for incorporating prior information, where multiple sources are integrated into a position-specific prior. This is similar to the approach we adopt. However, Tang et al. incorporate nucleosome information following the method of Narlikar et al. (2007), which we discuss below.

1.4.3. The PRIORITY Algorithm

Another interesting approach was recently presented by Narlikar et al. (2007). In their results, the authors use empirically-derived nucleosome-positioning information, in addition to a Gibbs sampling approach, to improve the prediction of TFBS. As we discuss in the coming chapters, we also use nucleosome-positioning information to derive a prior distribution for the position of TFBS, albeit in an expectation maximization framework.

Additionally, the method used to generate the prior and its incorporation into a motif-scoring scheme are different in our respective approaches. Briefly, we use the fact that nucleosomal DNA has a fixed length to derive a likely center-point for the nucleosome, then infer a prior using an empirically-derived frequency of TFBS occurrence. In contrast, Narlikar et al. use empirical nucleosome occupancy data more directly, calculating an average nucleosome occupancy score using a sliding window approach.

A further difference is the Narlikar et al. requirement that input sequences be identified as containing or missing binding sites for the TF in question. While this information is available in the context of ChIP-chip experiments, it is not generally the case that such information is available for input sequences of interest.

Finally, the Narlikar et al. approach can accommodate at most one occurrence of a TFBS in a given input sequence, while our approach can model any number of occurrences. Overall, however, the core ideas behind both approaches are similar.

We conclude this section by noting that there are multiple different but viable approaches to incorporating empirically-derived prior information to augment motif finding efforts. In the coming chapters we describe our approach based on the incorporation of empirical information in a probabilistic framework trained using expectation maximization.

1.5. Conclusion

In the following pages, we will jointly analyze the binding site and nucleosome positioning experimental data (Harbison et al., Yuan et al.) to deduce certain trends regarding the preferential positioning of binding sites in linker regions. We will then introduce a widely used algorithm for binding site prediction, MEME, and modify the basic algorithm to incorporate nucleosome positioning information into the prediction. We will then test the modifications with both simulated and experimental data, before analyzing the results and drawing conclusions regarding the benefit of nucleosome information in binding site prediction.

Chapter 2 – Exploring The Relationship Between Binding Sites And Nucleosome Positioning

Our goal in the following pages is to quantify the relationship between the positioning of active transcription factor binding sites and nucleosome positions, and between adjacent nucleosomes. We proceed by considering the center of a nucleosome as a fixed point, and examine the positioning of binding sites and other adjacent nucleosomes with respect to this fixed point.

2.1. Binding Site Positioning as a Function of Nucleosome Positioning

2.1.1. Distance from Nucleosome Centers

In studying the relationship between TFBS and nucleosomes, we need to always consider the positions of TFBS with respect to those of nucleosomes. As previously noted, the DNA wrapped in a nucleosome is exactly 147 bases in length. The mid-point of this DNA, at position 73, is the center of the nucleosome. It is convenient to consider this nucleosome center as the defining position of each nucleosome.

We can now consider how far a DNA position is from the center of a given nucleosome. Given a series of adjacent nucleosomes, we can also define any position between two nucleosomes as having a certain distance from the closest nucleosome center. We define the "distance from center" of a DNA position as its distance, in bases, from the closest positioned nucleosome center. For a position *i*, we denote this distance as *dfc(i)*.

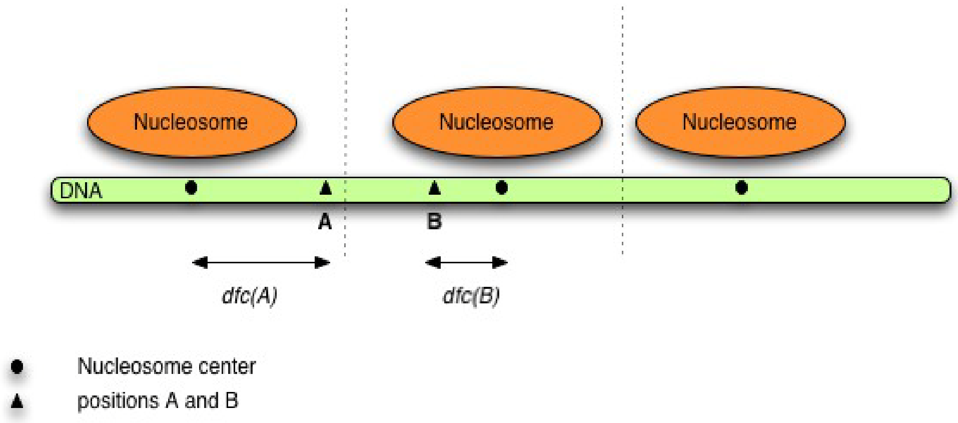


Figure 5 - Defining a genomic position with respect to its distance from the closest nucleosome center. This facilitates the analysis of protein-DNA binding frequency as a function of distance from the closest nucleosome center.

Given the fixed length of Nucleosomal DNA, for any genomic position *i*, *dfc(i)* has the following useful property: the DNA at position *i* is wrapped in a nucleosome if and only if $0 \le dfc(i) \le 73$.

2.1.2. Obtaining Nucleosome Centers From Published Data

The nucleosome positioning data published by Yuan et al. does not contain the coordinates of nucleosome centers. It typically assigns a 100-200 bp region of the genome as wrapped in a nucleosome. How can this be, given that biologically, the length of DNA wrapped in a nucleosome is 147 bp. The answer is that nucleosome positioning experiments provide approximate positions.

To clarify this, we must consider the nature of the experimentally-derived nucleosome positions. The Yuan et al. nucleosome positions are derived via ChIP-Chip technology. Of key importance here is the last part of this procedure.

In the last phase, fragments of DNA previously wrapped in nucleosomes are hybridized against a tiling microarray. This microarray has DNA probes that sequentially cover a significant proportion of the yeast genome. As previously noted, these DNA probes are 50 bases long: each covers 20 new bases, and overlaps with the previous probe over 30 bases. By seeing which DNA probes that show significant hybridization with Nucleosomal DNA fragments, certain Genomic positions are predicted as being wrapped in nucleosomes.

However, even assuming all predictions are accurate, prediction accuracy is limited since the microarray probes are positioned in 20 bp steps. Thus, while we have a good idea of where nucleosomes are positioned, we need to perform additional processing to predict the positions of nucleosome centers.

2.1.2.1. Computational Prediction of Nucleosome Centers

Our goal is to computationally assign nucleosome center positions, given microarray tiling probe data that has either a “linker” or “nucleosome” designation.

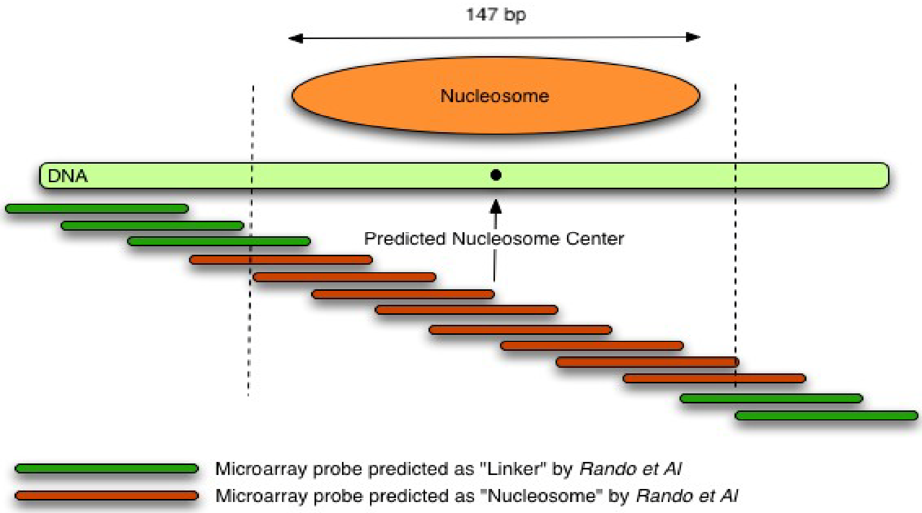

147 bp
Nucleosome
DNA
Predicted Nucleosome Center
Microarray probe predicted as "Linker" by *Rando et Al*
Microarray probe predicted as "Nucleosome" by *Rando et Al*

Figure 6 - Inferring the position of nucleosomes using information from tiling microarrays. Sequential overlapping probes predicted to represent nucleosomal stretches of DNA define nucleosome positions.

We proceed as follows. Given a contiguous sequence of overlapping "nucleosome" DNA probes, we first determine the start/end genomic positions overlapping by at least 2 probes. We then assign a nucleosome center position to the midpoint between start and end.

2.1.3. Exploring the Distance Between Adjacent Nucleosomes

Once nucleosome centers were assigned as described above, we were in a position to determine the distance relationship between adjacent nucleosomes. These distances were performed in a straightforward manner by calculating the difference in center coordinates between adjacent nucleosomes in the Yuan et al. dataset, and normalizing the counts for different distances to sum to one, thus obtaining an empirical probability distribution function.

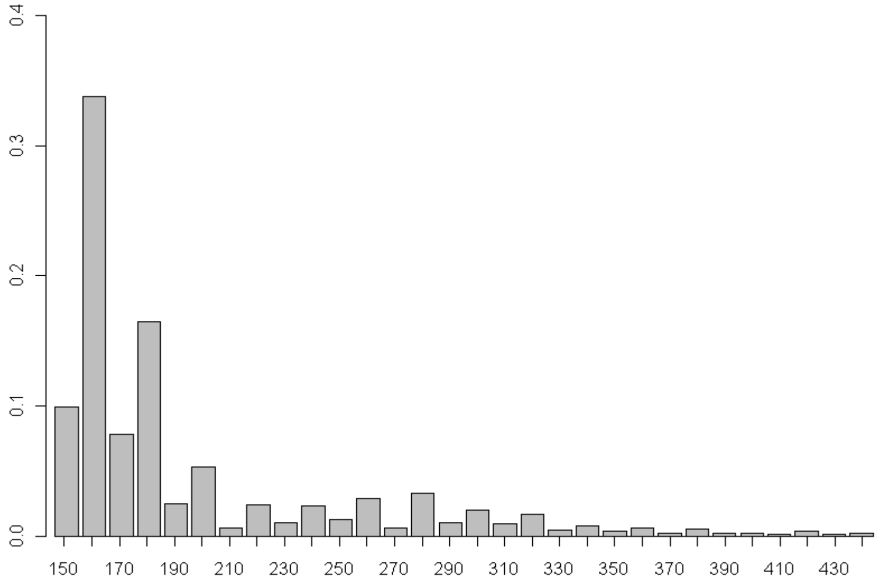


Figure 7 - Probability distribution of distance (number of bases) between two adjacent nucleosome centers from the Yuan et al. dataset.

One observation we make immediately is that the significant majority of nucleosome centers are at most 180 bp apart. Of that, 147 bp is occupied by the nucleosomal DNA. Even accounting for the coarse granularity of the tiling microarray used by Yuan et al., this hints that the majority of linker DNA regions connecting two nucleosomes are very short, on the order of 40 bp or less, with over 40% of all linker regions 20 bp or shorter in length.

This observation is broadly consistent with the observed periodicity of bend sites reported by Kiyama et al. (1999), as well as the distance preferences reported by Makeev et al. It is also further confirmation of the cross-species nature of these observed relationships, consistent with the strong evolutionary conservation of different components of nucleosome packaging from yeast to human.

2.1.4. Computing Probability of TFBS Occurrence in Relation to Nucleosome Center Position

As noted in chapter 1, DNA wrapped in a nucleosome is less accessible for protein-DNA interactions, particularly those involving TFs. We also noted that preferred nucleosome positions were encoded in the DNA sequence. Also, given the ubiquity and importance of chromatin packaging in eukaryotic organisms, preferred nucleosome positions are evolutionarily conserved.

These considerations strongly suggest that TFBS occur disproportionately often in linker DNA regions. Indeed, Yuan et al. note an overall enrichment of TFBS in genomic positions predicted as linker DNA in their study. Our goal here is to quantify the disproportionate TFBS occurrence as a function of the distance to the closest nucleosome center (*dfc*).

Since the *dfc(i)* is defined with respect to the nucleosome center closest to DNA position *i*, there are maximum upstream and downstream values for *dfc(i)* dependent on the positions of adjacent nucleosome centers. Beyond these, the position *i* is closer to another nucleosome center.

Thus we obtain the expected probability of binding site occurrence at a given distance from a nucleosome center as

$$Pr(TFBS(i) = true \mid dfc(i) = d) = \frac{\left|\{i : dfc(i) = d, TFBS(i) = true\}\right|}{\left|\{i : dfc(i) = d\}\right|}$$

We then average the probabilities using a length 10 sliding window.

One additional data-related issue we must contend with is the sparseness of TFBS instances occurring at positions with higher *dfc(i)* values. For *dfc(i)* values beyond 200, we observe few occurring TFBS, typically 0 or 1 instances per *dfc(i)* value.

For this reason, we use

$$\Pr(TFBS(i) = true \mid dfc(i) > 200) = \frac{1}{127}\sum_{k=74}^{200} \Pr(TFBS(i) = true \mid dfc(i) = k)$$

That is, for higher *dfc(i)* values, we use the mean of the probabilities of transcription factors at *dfc(i)* values between 74 and 200, positions outside the nucleosome for which there are a significant number of observed TFBS.

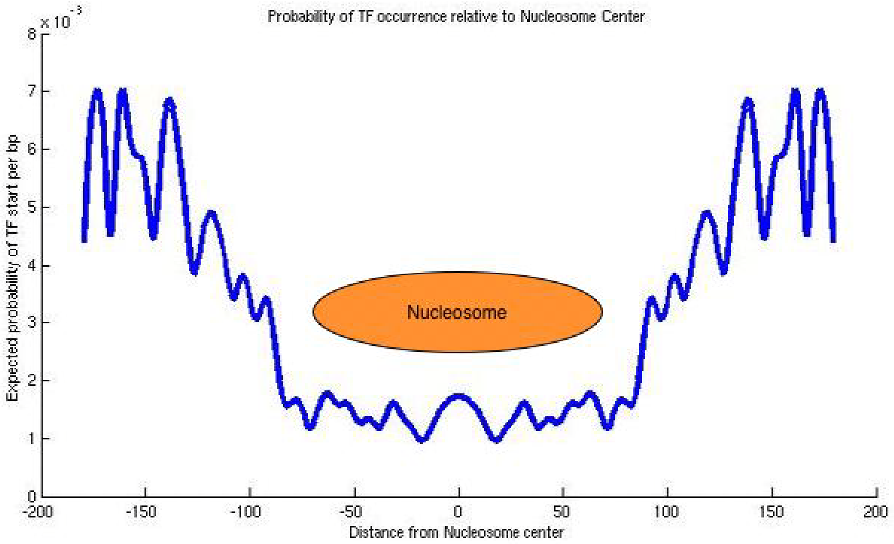


Figure 8 - Probability of binding occurrence conditioned on distance from the closest nucleosome center. For distances beyond 200, the mean of the probabilities for distances 74 to 200 was used.

We use the empirically derived relation represented in Figure 8 in the next chapter.

2.2. Measuring Positioning Affinity for Individual TFs

As seen above, on the whole, TFBS occur predominantly outside nucleosomes. However, the degree of this positioning affinity may vary between different TFs. The more instances of a particular TFBS occur within nucleosomes, the more restricted their access. However, in certain cases, this restricted access may be biologically desirable. This would be the case, for example, if the wayward TF binding and subsequent effect on regulation were highly damaging.

We proceeded as follows to create a measure of Linker positioning affinity for individual TFs. We assume that there are two functionally meaningful positions for a TFBS: inside and outside a nucleosome. We then proceed to count the number of occurrences of a TFBS inside and outside nucleosomes as we did for the overall nucleosome positioning prior.

For each TF, we define two Bernoulli random variables: pIn and pOut. They represent the probability of a DNA position inside or outside a nucleosome being the start of a TFBS of the specified type. We calculate the point estimators for pIn and pOut as follows:

$$pIn = Pr(TFBS(i) = true \mid dfc(i) < 74) = \frac{\left|\{i : dfc(i) < 74, TFBS(i) = true\}\right|}{\left|\{i : dfc(i) < 74\}\right|}$$

$$pOut = Pr(TFBS(i) = true \mid dfc(i) \geq 74) = \frac{\left|\{i : dfc(i) \geq 74, TFBS(i) = true\}\right|}{\left|\{i : dfc(i) \geq 74\}\right|}$$

We can thus quantify the relative occurrence frequency of a specific type of TFBS as $\frac{pOut}{pIn}$. A ratio $\frac{pOut}{pIn} = 1$ indicates that binding sites for a specific TF are expected to occur equally frequency in nucleosomal (pIn) and linker (pOut) DNA regions. A ratio $\frac{pOut}{pIn} > 1$ indicates increased relative frequency of binding site occurrence in linker DNA, while the opposite inequality indicates binding site enrichment in nucleosomal DNA regions. We computed $\frac{pOut}{pIn}$ for all TFs in the Harbison et al. dataset. The ratios, sorted in decreasing order, are shown in figure 9 below.

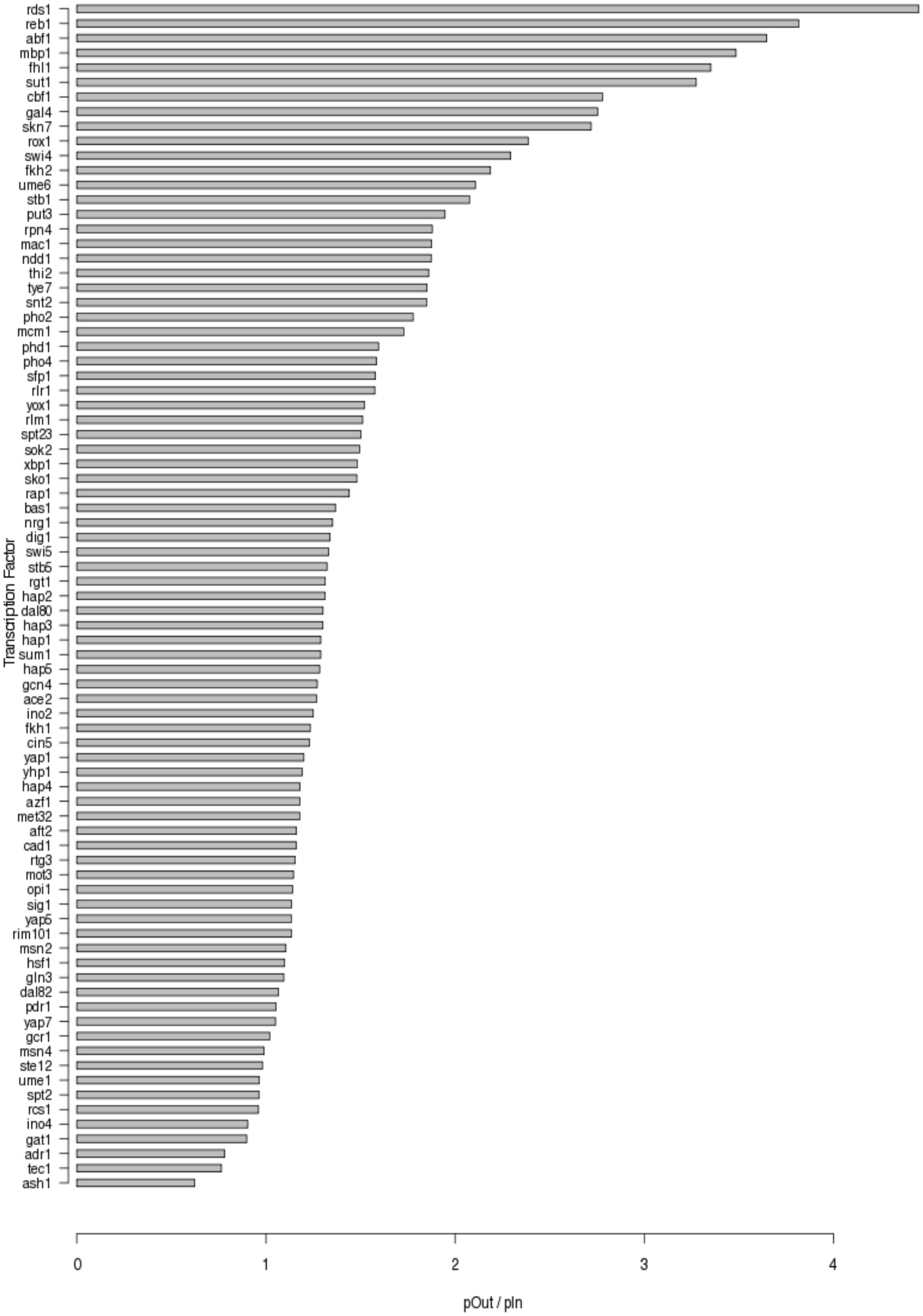


Figure 9 - Relative binding enrichment for different TFBS in linker DNA, as measured by the pOut/pIn ratio. The higher the value, the more frequently the TFBS occurs in linker DNA regions.

As seen in the above figure, the vast majority of binding sites in the Harbison et al. dataset display a relative positioning enrichment in linker DNA regions (pOut/pIn > 1). We further note that a significant subset of these binding sites display a very strong bias towards linker regions (pOut/pIn > 2).

2.3. Conclusion

In the preceding pages we jointly explored genome-scale binding site and nucleosome positioning data. By examining distances of all such elements with respect to the closest nucleosome center, we showed both an interesting distance relationship between adjacent nucleosomes, and confirmed in detail previous reports of functional binding site enrichment in linker DNA regions. Finally we demonstrated that while most transcription factor binding sites occur disproportionately in linker DNA regions, the extent of this enrichment is dependent on the transcription factor under consideration. Specifically, a few binding sites display greater than three-fold enrichment in linker regions.

We now proceed to detail a model for predicting transcription factor binding sites, as well as an extension allowing the incorporation of the above-derived empirical relationships.

Chapter 3 – Modeling and Algorithms

In the previous chapter, we investigated the empirical relationship between transcription factor binding sites and nucleosome positions. Our goal in what follows is to incorporate nucleosome positioning information into a computational binding site prediction model of to improve its predictive qualities. For this purpose, we choose the framework provided by Bailey and Elkan (1994), and adjust it to take into account available nucleosome positioning information.

3.0. Preliminary Notation Conventions

Given the significant number of mathematical formulae to come in the following pages, it may be beneficial to set some notational conventions and definitions beforehand.

i – index on input sequences

j – index on the current position within a sequence

k – index on position within a TFBS or associated Position Weight Matrix ($1 \le k \le W$)

X_i – an input DNA sequence

$x_{i,j}$ –nucleotide at position j of sequence i

M – total number of sequences

N – the nucleotide length of the current sequence under consideration

$X_{i,j}^{W}$ – sub-sequence of length W starting at position j of X_i.

$X_{i,j}^{W} = x_{i,j} \ldots x_{i,j+W-1}$

Σ – set of characters *{A,C,G,T}*

θ_1 – Position weight matrix of the TFBS.

$\theta_{1,c,k}$ – entry for character $c \in \Sigma$, at column k, of the Position Weight Matrix θ_1

3.1. Problem Statement

We are given a set of promoter sequences $X_1, X_2, ..., X_M$ for co-regulated genes. We would like to discover the PWM representing the preferences for the TF binding these promoters. We also want to discover the putative positions of TFBS along the input promoter sequences.

As we detail below, we model the problem as a classifier wherein each position (or nucleotide) of the input sequences can be either background DNA or the start position of a TFBS. The goal of the modeling then becomes determining the class of each position.

More formally, given input DNA sequences $X_1, X_2, ..., X_M$

We wish to find:
A probabilistic model of the TFBS
TFBS positions along the input sequences

Following Bailey and Elkan, we approach this as a classification problem using mixture models.

3.2. Classification Using Mixture Models

Our classifier is an instance of a Two Component Mixture (TCM) model. Each nucleotide in the input sequence is considered to originate from either a background random process, or a “transcription factor binding site start" (TFBS start) random process that positions binding sites along a sequence.

More formally, for each position *(i,j)*,

$$Z_{i,j} = \begin{cases} 1 & \text{if position j is a TFBS start point in X}_\text{i} \\ 0 & \text{otherwise} \end{cases}$$

The indicator variable $Z_{i,j}$ tells us which random process each nucleotide originated from. However, we do not know the values $Z_{i,j}$, and the goal of the classifier is to learn them. This is our missing information. We distinguish between two possible classes: background, or class 0, and start of TFBS, or class 1.

We make the assumption, prior to training the model, that all positions are equally likely to be the start of a TFBS. That is, we assume a uniform prior probability of TFBS start over the input sequences. More formally, we write $Pr(Z_{i,j} = 1) = \lambda$ ($0 < \lambda < 1$).

3.3. Probability of Binding Site Start According to Model

The background DNA sequence is modeled as a 2nd order Markov model defining the probability of a nucleotide occurring in terms of the previous two nucleotides that occurred along the sequence. We denote this background DNA model as θ_0. A PWM is used as the TFBS model. We denote it θ_1, and denote the entry for character c at column/position k as $\theta_{1,ck}$.

Given the prior probability of TFBS start, and models for Background DNA and TFBS (θ_0 and θ_1 respectively), we are able to calculate the probability of a position being the start of a TFBS according to the TCM model.

As noted earlier, we initially assume a uniform prior probability of TFBS start for each position. That is, the prior probability of the indicator variable $Z_{i,j}$ is uniform.

$$Pr(Z_{i,j} = 1) = \lambda$$

Using Bayes' rule, we obtain

$$Pr(Z_{i,j} = 1 \mid \theta_1, X_i) = \frac{Pr(X_i \mid \theta_1, Z_{i,j} = 1)\lambda}{Pr(X_i \mid \theta_0, Z_{i,j} = 0)(1-\lambda) + Pr(X_i \mid \theta_1, Z_{i,j} = 1)\lambda}$$

The value $Pr(Z_{i,j} = 1 \mid \theta_1, X_i)$ is the probability of TFBS start (equivalently, the probability of class 1) according to our model and the sequence data.

3.4. Learning Model Parameters Using Expectation Maximization

Expectation maximization is a class of algorithms applicable to estimating model parameter values in the presence of missing information (Dempster et al. 1977). EM algorithms follow an iterative procedure where each round consists of two distinct steps: an expectation (E) step and a maximization (M) step. During the E-step, the expected values of missing model information are calculated based on the current value of the model parameters. In the M-step, the model parameters are re-estimated based on the missing information estimated in the E-step. The re-estimation is done in such a way as the log-likelihood of the model is maximized. This iterative process is repeated until the model parameters converge to fixed values.

3.3.1. Model Log-Likelihood

The models learned using EM are often highly complex. While in theory there is a globally optimum set of model values that maximize the log-likelihood, in practice the solution space has many local optima. EM is proven to converge to a locally optimal set of model parameters.

Given the input sequence data and a set of model parameters, we can calculate the log-likelihood of the model as follows (Bailey and Elkan)**:**

$$logPr(X,Z \mid \theta_0,\theta_1,\lambda) =$$
$$\sum_{i=1}^{M}\sum_{j=1}^{N}(1-Z_{i,j})log(Pr(X_i \mid \theta_0,Z_{i,j}=0)(1-\lambda)) + Z_{i,j}log(Pr(X_i \mid \theta_1,Z_{i,j}=1)\lambda)$$

Note that, in the above formula, $Z_{i,j}$ is a binary variable. This is fine so long as the values are known. In our case, these values are missing information that the TCM model estimates. Thus, in our calculations and implementation, we use the estimated posterior probability $Pr(Z_{i,j}=1 \mid \theta_1,X_i)$ instead of $Z_{i,j}$.

The EM algorithm for learning the model parameters leads to a solution that is locally optimal, in that the model parameters obtained locally maximize the above log-likelihood function. As we discuss further below, a heuristic based on this criterion is also used to find a good initial set of parameters for the model.

3.3.2. E-Step

In the first half of an iteration, we calculate the expected probabilities of TFBS start for each position in the input sequences. In other words, we calculate the expected probability of each position originating from class 1, given the current model parameters.

More precisely, we calculate the expected posterior probability $Pr(Z_{i,j} = 1 \mid \theta_1, X_i)$ mentioned previously, using the current values for model parameters θ_0, θ_1 and λ. In the second half of the EM iteration, we use these values to re-estimate the model parameters using maximum likelihood estimates for each.

3.3.3. M-Step

We re-estimate the model parameters as follows (Bailey and Elkan 1994)

For θ_1: for character $c \in \sum_{DNA}$, column $k = 1,...,W$

$$n_{c,k} = \sum_{i=1}^{M} \sum_{\{j|x_{i,j+k-1}=c\}} \Pr(Z_{i,j} = 1 \mid \theta_1, X_i)$$

To obtain the PWM entries, which are probabilities, we normalize each column:

$$\theta_{1,ck} = \frac{n_{c,k}}{\sum_{\{d \in DNA\}} n_{d,k}}$$

For θ_0: for each tri-nucleotide $c_1c_2c_3$ ($c_1, c_2, c_3 \in \sum_{DNA}$)

$$n_{c_1c_2c_3} = count(c_1c_2c_3) - \sum_{i=1}^{M} \sum_{\{j|x_{i,j-2}=c_1, x_{i,j-1}=c_2, x_{i,j}=c_3\}} Pr(Z_{i,j} = 1 | \theta_1, X_i)$$

Where $count(c_1c_2c_3)$ denotes the observed count of the tri-nucleotide in our input sequences. Again, we normalize the adjusted tri-nucleotide counts to obtain probabilities for the 2nd order Markov Model:

$$\theta_{0,c_1c_2c_3} = \frac{n_{c_1c_2c_3}}{\sum_{\{d_1,d_2,d_3 \in DNA\}} n_{d_1d_2d_3}}$$

For λ: To obtain the prior probability of TFBS start, we average the expected probabilities of TFBS start for each position:

$$\lambda = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} Pr(Z_{i,j} = 1 | \theta_1, X_i)$$

3.4. Model Initialization

3.4.1. Sensitivity to Initial Conditions

EM is highly sensitive to model initial conditions. The iterative procedure often leads to different final model parameters depending on the initial parameters of the model. This is because the shape of the likelihood (or log-likelihood) surface is highly complex, with many local minima and maxima. Well-chosen initial conditions would be those that provide some information about the missing values we want to estimate in the model (in our case the indicator $Z_{i,j}$ values).

W initialization: we assume that the user knows the correct length of the motif to be found.

λ initialization: we set the initial value for lambda to *1/N*. That is, we initially expect one site occurrence per sequence. This value will be re-estimated by the EM algorithm.

θ_0 initialization: we initialize the background model from the input sequences by counting the number of occurrences of each tri-nucleotide in the input sequences, and normalizing so that the conditional probabilities for $\forall c_1, c_2, c_3 \in \sum_{DNA}$, $Pr(c_3 \mid c_2 c_1)$ sum to one.

3.4.2.1. Start-Point Selection Through Sampling

The initialization of θ_1 critically affects the results of the classifier. If θ_1 resembles the TFBS we wish to find, the results are often good. If it is not, the end-result is poor.

The PWM θ_1 is initialized as follows. Each possible W-mer in the input sequences is sampled, and used as a seed sequence to initialize the PWM. A single E-step is then performed to calculate the posterior probabilities according to the PWM, and the log-likelihood of the model is calculated.

The W-mer yielding the highest log-likelihood from the above sampling is kept as the initialization seed for the PWM, and EM is then performed on the classifier until convergence of the parameters. More formally, to initialize a PWM given a W-mer:

given W-mer $X^W = x_1 x_2 ... x_W$

$$\theta_{i,ck} = \begin{cases} \varepsilon & \text{if } \mathrm{x}_\mathrm{k}\text{=c} \\ \dfrac{1-\varepsilon}{3} & \text{otherwise} \end{cases}$$

In other words, we favor character x_k at each position by setting it to ε, and set all other columns uniformly so the total sums to 1. We have used $\varepsilon = \frac{2}{3}$; values $0.5 \le \varepsilon \le 0.8$ are reported to be effective. The key is to allow for alternate nucleotides, so that the diversity inherent in instances of a motif are captured.

The start-point search is performed by initializing the PWM with successive W-mers from the input sequences, running a single EM iteration, and selecting the sequence with highest log-likelihood.

This is a good heuristic due to the fact that EM converges in a few iterations (typically 4-8), and we expect that a PWM representing the TFBS to be found leads to the highest model log-likelihood. Intuitively, we expect that the instances of the TFBS to be found are the most "surprising" features in the dataset, those that stand out the most against the background. It must be noted that this is not always a realistic assumption, given the degeneracy of in-vivo TFBS and their resilience in the presence of mutations.

3.5. Model Output

Once initial values for model parameters are chosen, the classifier is trained using EM with 100 iterations in our implementation. The number was chosen to ensure convergence of model parameter values, which typically converge in less than 10 iterations.

Once EM has completed, the model has a new λ value, or expected probability of TFBS start per position. For the entire input dataset of M sequences, this means

$$E[\text{number of TFBS in dataset}] = \lfloor \lambda MN \rfloor$$

To determine the positions of the TFBS in the input sequences, the positions are sorted according to decreasing posterior probability $Pr(Z_{i,j} = 1 | \theta, \lambda, X)$, and the top $\lfloor \lambda MN \rfloor$ positions are chosen as TFBS start positions.

3.6. Incorporating Experimental Information Using Position-Specific Priors

As noted earlier, EM is very sensitive to initial model parameter values. It stands to reason that choosing good initial values leads to better trained model parameter values. We also wish to incorporate additional prior information that may be available from experimental and other sources in order to more realistically model the biological phenomenon.

One way of doing this is through the position-specific prior. The default TCM model assumes a uniform prior probability λ of TFBS starts. However, as demonstrated in our exploration of the Harbison et al. and Yuan et al. datasets, TFBS functional in-vivo occur disproportionately in Linker DNA regions.

Thus, given a set of co-regulated sequences $X_1, X_2, ..., X_M$, we also have the positions of nucleosome centers for each sequence: $Y_1, Y_2, ..., Y_M$, where each $Y_i = y_{i,1}y_{i,2}...y_{i,r}...y_{i,R}$ is a set of nucleosome center coordinates positioned along the corresponding input sequence X_i.

3.6.1. Position-Specific Prior Based on nucleosome Positions

For each sequence position *(i,j)*, we associate it with the nucleosome center closest to it. In cases where a position is equidistant from two adjacent nucleosome centers, we assign it to the upstream nucleosome center. Call this distance $d_{i,j}$.

As shown in Figure 8 (section 2.1.4), we compiled an empirical prior probability of TFBS start according to distance from nucleosome center. To obtain a position-specific prior for each position *(i,j)* of our input sequences, we simply do a table lookup of the empirical probability associated with distance $d_{i,j}$.

Thus, the uniform prior probability of TFBS start, $Pr(Z_{i,j} = 1)$, is modified to a probability conditioned on the positions of nucleosome centers $Pr(Z_{i,j} = 1 \mid Y_i)$, which is obtained by first calculating $d_{i,j}$, and then performing a table lookup with the empirically derived prior. However, this causes problems when we wish to use maximum likelihood estimation to re-estimate λ, as we will discuss further below.

3.6.2. Lack of Closed-Form Estimate

Before introducing position-specific priors, we had a closed-form estimate for $\lambda = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} Z_{i,j}$. We have no such formula for the empirical position-specific prior.

We thus modify our approach to consider the position-specific priors not as absolute probabilities, but rather as relative weights. Specific to each sequence, these weights bias the prior probability of TFBS start towards certain empirically plausible positions, but do not alter the overall expected probability of TFBS start λ. This modified prior is calculated at model initialization, and at each M-step, as detailed below.

3.6.3. Modified Model Initialization

Let λ_0 be our initial estimate for the prior probability of TFBS start. For each sequence *i=1, ..., M*:

$$\gamma_i = \sum_{j=1}^{N} \lambda_0$$

$$\rho_i = \sum_{j=1}^{N} Pr(Z_{i,j} = 1 \mid Y_i)$$

Here, for each j, $Pr(Z_{i,j} = 1 \mid Y_i)$ is obtained via a table lookup using the empirical prior probabilities calculated from the Harbison et al. and Yuan et al. data.

Let the modified position-specific prior, for each *(i,j)* position, be

$$v_{i,j} = \frac{\gamma_i}{\rho_i} Pr(Z_{i,j} = 1 \mid Y_i)$$

We have added a sequence-specific scaling factor $\frac{\gamma_i}{\rho_i}$ so the expected probability of TFBS start remains constant, but still incorporates empirically derived TFBS positioning information by assigning a higher prior probability to certain positions.

3.6.4. Modified M-Step

We calculate the scaling factor $\frac{\gamma_i}{\rho_i}$ similarly to the model initialization,

except that $\gamma_i = \sum_{j=1}^{N} Z_{i,j}$

And $v_{i,j} = \frac{\gamma_i}{\rho_i} Pr(Z_{i,j} = 1 | Y_i)$ is calculated with the new γ_i.

This prior has the property that

$$\lambda = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} Z_{i,j} = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} v_{i,j}$$

In other words, by design, the scaled prior allows us to keep our closed-form estimate of λ, and still bias more likely TFBS start sites, while sacrificing some information in estimating expected TFBS occurrence frequency.

3.6.5. Modified E-Step

The modified posterior probability of TFBS start, calculated at each E-step of the EM algorithm, then becomes

$$Pr(Z_{i,j} = 1 | \theta_1, X_i, Y_i) = \frac{Pr(X_i | \theta_1, Z_{i,j} = 1, Y_i) v_{i,j}}{Pr(X_i | \theta_0, Z_{i,j} = 0, Y_i)(1 - v_{i,j}) + Pr(X_i | \theta_1, Z_{i,j} = 1, Y_i) v_{i,j}}$$

With the $v_{i,j}$ calculated as above.

3.7. Conclusion

In this chapter, we laid out a model as well as an algorithm for predicting the occurrence of binding sites. This two-component mixture model represents the input sequences as a product of two competing stochastic processes: the first process responsible for generating the background DNA sequence, and the second process responsible for positioning instances of a given binding site. Using the EM algorithm, we can obtain maximum likelihood estimates for the model parameters. Finally, we modified the algorithm to allow the specification of an arbitrary prior distribution for the occurrence of binding sites. In the next chapter, we will evaluate the performance of this model using large-scale simulated and empirical data.

Chapter 4 – Assessing Algorithm Performance

4.1. Introduction

As previously noted, one of the key roles of prior information based on nucleosome positions is to shift the prior probability of TFBS start. We expect that, in many cases, this will have a beneficial impact on the effectiveness of the start point search portion of the algorithm. This would be especially the case when searching for TF instances preferentially positioned in linker regions.

Two important caveats apply however. The first is that the bias in TF positioning is statistical. As we detail below, each test data set contains 10-16 binding site instances for a specific TF. Thus, even in cases where a marked positioning preference exists, we may expect at most 1-3 instances of the TF to be assigned higher prior probabilities. Several of the instances will also be assigned significantly lower prior probabilities, leading to lower posterior probability values.

The start point search computes the model log-likelihood by adding the posterior probability of each position separately, adding nucleosome prior information has a mixed effect on the log-likelihood calculation, increasing the scores for true sites in most cases, but decreasing them in a significant minority of cases. How this affects the overall log-likelihood is not entirely clear.

A second important caveat is related to the fact that the nucleosome information leads to a higher prior probability for all putative start points located in Linker DNA regions, not just true start sites. Thus while we expect the prior probability of most true instances to be increased, the prior probability of all background DNA in Linker DNA regions is also increased. Thus, to some extent, we strengthen the noise as well as the signal.

4.2. Test Data

4.2.1. Simulated vs. Biological Test Data

The most credible test for any algorithm is its performance on biological datasets. Several limitations must however be considered. The first limitation involves producing datasets from existing annotated data sources. We must deal with the problem of incomplete knowledge and errors. In particular, we must consider biologically active ("real") TFBS that exist in a given stretch of sequence, but are missed in existing annotations. Conversely, certain identified motif positions are false-positives arising from the experimental and computational identification methods that are notoriously error-prone. In order to properly measure application performance, we need "gold-standard" test data with highly reliable annotations.

Another limitation is the amount of data available for a given motif or set of conditions. We would like many similar test data sets having similar characteristics such as common TF and number of TFBS. This would allow us to examine application performance on a large scale, systematically control for variables and thus obtain more reliable results.

For these reasons, we initially resorted to simulated datasets. Such datasets do not reflect all the biologically relevant characteristics of real datasets, but their ease of generation and control of important characteristics, such as number and position of motif instances, provide important advantages.

4.2.2. Description of Test Datasets

We can classify our test data sets into 3 categories:

- Simulated:
    - simulated sequences
    - simulated nucleosomes
    - TFBS positioned according the nucleosome position-specific prior derived from the simulated nucleosome positions
- Experimental nucleosome:
    - Promoter sequences from yeast
    - Experimental TFBS positions (Harbison et al.*)* for these promoters
    - nucleosome positions (Yuan et al.*)* for these promoters
- In-silico nucleosome:
    - Same promoters and TFBS positions as Experimental nucleosome
    - In-silico-predicted nucleosome positions (Segal et al.) for these promoters

Each of the test data sets used in our subsequent tests has the following common characteristics, regardless of data origin:

- 10 DNA sequences, 780-800 bases length each
- 10-16 “real” TFBS instances with known positions, with every sequence containing at least one known TFBS
- the coordinates of nucleosome centers along each sequence

4.3. Production of Test Data Sets

4.3.1. Simulated Sequences

For the purpose of assessing relative performance, we constructed a set of simulated datasets. These simulations incorporated the assumption that

binding sites are positioned according to a nucleosome position biased prior. The construction proceeded as follows.
A simulated dataset containing ten promoter sequences of 800 bases each was constructed for each of the transcription factors under consideration. As a first step, the promoter sequences were generated using a second order Markov model. The conditional probabilities for the model were obtained by using the extracted yeast promoter sequences in our database as training material. A sliding window approach was used to count the different tri-nucleotides in the sequences. These were then grouped based on the first two nucleotides and the frequency of the nucleotides in the third position normalized to obtain the conditional probabilities. An analogous approach yielded the first-order conditional probabilities (using di-nucleotides) and simple nucleotide probabilities.

After generating the promoter sequences, we associated simulated nucleosome positions with each promoter. For this purpose we used the empirical distribution of distances between nucleosome centers introduced in section 2.1.3**.** The initial nucleosome was positioned according to a uniform prior. Subsequent nucleosome centers were positioned by randomly sampling from the empirical distribution above, conditionally dependent on the previously positioned nucleosome center. New nucleosome centers were positioned until one fell beyond the length of the simulated sequence.

Once simulated nucleosome centers were positioned for each promoter, a probability of being a binding site start position was assigned using the empirical *dfc* distribution introduced in chapter 2**.**

Finally, simulated binding sites were positioned as follows. For each transcription factor, the associated PWM was then obtained from the UCSC genome browser database (Kent WJ 2002) and used to generate instances of the binding site sequence. Ten of these binding sites were generated for each dataset, and positioned in the generated promoter sequences. First a random promoter was selected. Then a random position for the binding site start was selected according to the nucleosome-influenced prior distribution of binding site positions described above.

No restriction was placed on the random positioning of the binding sites, in other words multiple sites could be positioned in the same promoter sequence and others could be left empty.

4.3.2. Experimental Test Data Set Production

All non-simulated data used in testing the algorithm is from yeast. Yeast genome sequence and basic annotations from the SGD database were obtained and added to a database. To this were added the TFBS positioning data from Harbison et al. and the Yuan et al. nucleosome positioning predictions.

For each TF in the Harbison et al. data, we created a test data set if the following conditions were satisfied: we located 10 yeast promoter regions of length at approximately 800 bp existed, for which Yuan et al. nucleosome positions existed, each containing 1-2 instances of a TFBS from Harbison et al.. We found 78 TFs for which these conditions were satisfied. In each case, we created a test data set consisting of 3 files: promoter DNA sequences, known TFBS positions and nucleosome positions.

4.3.3. Experimental TFBS/ In-Silico Nucleosome Test Dataset Production

Given the ability to computationally predict nucleosome positions without needing to resort to costly and time-consuming experiments, we wish to investigate whether there's a predictive benefit to incorporating information from in-silico predicted nucleosomes into the basic algorithm.

We wanted to have comparison results directly comparable to the performance effects of experimental nucleosome positions. For this purpose, we used the exact same promoter sequences and TFBS positions as for the 78 Experimental nucleosome test data sets. However, instead of using the Yuan et al. nucleosome positions, we used the in-silico nucleosome predictions of Segal et al. for the same DNA sequences. We are thus able to isolate the differing effects of experimental and in-silico nucleosome positions on algorithm predictive performance.

4.4. Assessing Performance on Test Datasets

We are interested in measuring the difference between the number of sequence positions predicted to be TFBS by the algorithm and the number of positions that are actual biological TFBS. We denote the predicted TFBS sites as predicted True Positive (predicted TP), and the actual TFBS as real True Positive (real TP). Quantifying the discrepancy between predicted and actual sites will help us determine the quality of the predictions made by the algorithm. To achieve this, we describe in the sections that follow two measures based on these two quantities.

4.4.1. Site-Level Sensitivity

Sensitivity is an extremely important characteristic for a TFBS finding algorithm, or any other method dealing with “needle in a haystack” problems. By this, we mean prediction problems where the data to be predicted are surrounded by a great deal of noise or confounding information. In the case of TFBS prediction in particular, a few short, degenerate DNA sites are embedded within a sea of confounding DNA sequence two to three orders of magnitude larger than the TFBS.

$$Sensitivity = \frac{\left|\{predicted.TP\} \cap \{real.TP\}\right|}{\left|\{real.TP\}\right|}$$

Sensitivity is a measure of the algorithm’s ability to detect a real signal, in this case real TFBS positions, in the presence of noise.

4.4.2. Site-Level Positive Predictive Value

Positive Predictive Value (PPV) is a measure of how confident we are that predictions of an algorithm are actual TFBS and not false-positive predictions.

$$PPV = \frac{\left|\{predicted.TP\} \cap \{real.TP\}\right|}{\left|\{predicted.TP\}\right|}$$

A low PPV indicates that any site predicted by the algorithm, regardless of reported prediction confidence, is likely to be false. In other words, most of the predictions of the algorithm are false positives. For our purposes, a predicted binding site was correctly predicted if it overlapped with the actual binding site over at least 1/3 of its length, rounded up.

4.4.3. Performance on Simulated Data

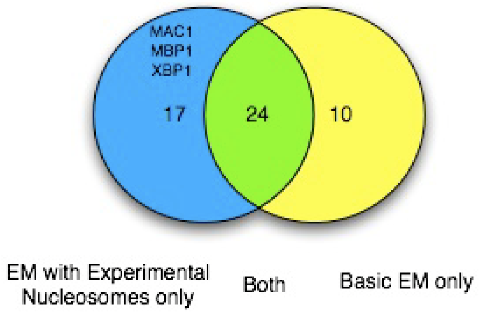
We will not dwell long on the simulated data, except to note that, as expected, the algorithm clearly performed better with the addition of the simulated nucleosome prior. The following table summarizes the results:

	Algorithm	
	Basic	**Including Simulated Prior**
Sensitivity	0.229	0.32
PPV	0.204	0.309

One must note that the improvement here is somewhat circular: if the data is generated using the exact prior one then uses to predict sites in the data, significant improvements are to be expected. In some sense, these numbers provide an upper bound on the performance of the algorithm with or without a prior. As we see below, actual experimental data is much harder to predict.

4.4.4. Experimental Nucleosome Information vs. Basic Algorithm

Test Data Sets for Which True TFBS Predictions Were Made



	Sensitivity	PPV
Basic EM	0.093	0.101
EM with Nucleosomes	0.125	0.139
Difference (%)	**25.6%**	**27.3%**

Figure 10 - Overlap and comparison of predictions made for different TFBS datasets using the default and modified versions of the algorithm. The modified algorithm incorporated experimentally-derived nucleosome positions from Harbison et al.

The basic EM approach detected a true TFBS instance in 34 test data sets. Adding experimental nucleosome information increased this number to 41 test data sets. Of these, 24 test data sets were correctly predicted by both approaches. Thus, in 17 test data sets, a signal was only detected upon inclusion of experimental nucleosome information, and in 10 test data sets, a signal detected using the basic approach disappeared upon inclusion of nucleosome positioning information.

We examined the prediction details to better understand these differences. In all but 3 test data sets correctly predicted using only one approach, only 1-2 real TFBS instances were detected by the algorithm (Sensitivity < 0.17 in all cases). This is a very weak signal, and we summarize that in most such cases the addition of nucleosome positioning information had a marginally positive or negative impact.

In the 3 remaining test data sets, for transcription factors MAC1, MBP1 and XBP1, inclusion of nucleosome information led to the detection of a significant fraction of real binding sites (Sensitivity of 0.32, 0.45, 0.9 respectively). In these cases the prior information proved decisive in detecting a significant number of actual binding sites.

To strengthen this conclusion, we examined the propensity of binding sites for these 3 TFs to occur in Linker DNA regions. They ranked respectively 14th, 4th and 31st among 78 TFBS considered. The rankings lend support to the conclusion that including of nucleosome positioning information helps us detect real TFBS occurrences in cases where the TF occurs mainly outside nucleosomes.

However the one must ask why binding sites that ranked highest in the enriched list provided in section 2.2 were not detected by the addition of nucleosome prior information. In particular, of the six transcription factors at the top of the list, only MBP1 was detected after the addition of nucleosome prior information.

To investigate this question, we first looked at the length of the sequences, using the yeast genome database (Cherry JM 1997). In both cases, no regularity with respect to length emerged. In fact, MAC1 and MBP1 were of length 7 and 6 respectively, making them two of the shortest binding sites in our list. The other top transcription factors had variable length, ranging from 6 to 13 bases.

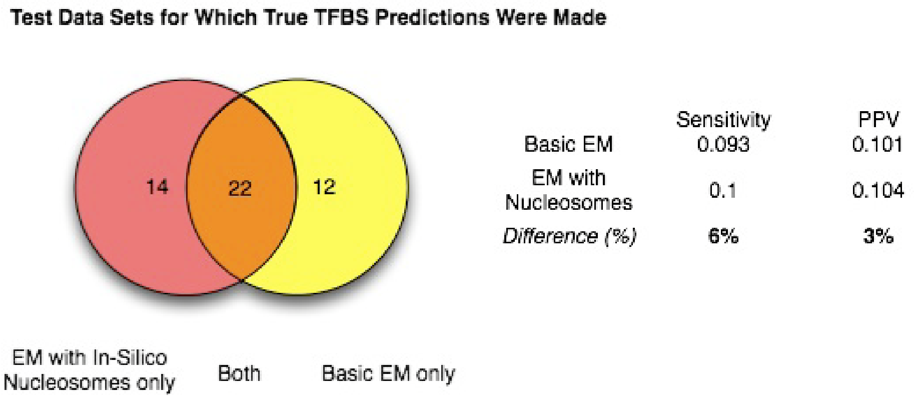
Next we considered the information content (Durbin et al. 1998) of the experimental binding sites found in our datasets. Given that different transcription factors have different lengths, we normalized the information content by dividing it by the length of the binding sites. This gave us the average information content per base, a measure of average conservation of the binding sites. We calculated this average information content for all the datasets and ranked the transcription factors. In this case, MAC1 and MBP1 were at the top of the list, ranking 2nd and 1st respectively. MBP1 was perfectly conserved in our test dataset. What was more interesting is that, of the top ten linker-enriched (section 2.2), the most conserved (REB1) ranked at 17th out of 78 in average information content, with the others ranking substantially lower.

We also examined the average TP and PPV values for both approaches, as shown above. When considering only TFs for which some signal was detected, there is only a marginal improvement in TP and PPV attributable to the additional prior information. However, when the average includes the test data sets for which no signal was detected, the improvement is interesting at 25%. We consider this last number a more representative measure of performance improvement, since in actual applications, we have no way of determining whether a signal was detected as opposed to a false-positive prediction.

As previously noted, there are two ways in which empirical priors may improve the basic EM algorithm: by improving the start point search used to select a candidate motif, and by assigning higher probabilities to instances of the motif. We wish to disentangle the relative contributions of these two factors in the observed improvement in Sensitivity and PPV.

A large fraction of the observed improvement in expected Sensitivity and PPV is attributable to the Sensitivity and PPV values of MAC1, MBP1 and XBP1 raising the average. Much of the rest of the improvement can be attributed to the four additional test data sets for which a weak signal was detected.

For the large majority of test data sets in the intersection, the expected Sensitivity and PPV values were very similar after the addition of nucleosome positioning information. Surprisingly in fact, the improvement in average Sensitivity when considering only the 24 test data sets common to both methods is only 5%, while the improvement in average PPV is 10.7%.



	Sensitivity	PPV
Basic EM	0.093	0.101
EM with Nucleosomes	0.1	0.104
Difference (%)	**6%**	**3%**

Figure 11 - Overlap and comparison of predictions made for different TFBS datasets using the default and modified versions of the algorithm. The modified algorithm incorporated in silico predicted nucleosome positions from Segal et al.

When replacing the experimentally-derived nucleosome positions with in-silico positions predicted by Segal et al., the difference in performance between the two algorithms decreases to a negligible level. The overall difference in average Sensitivity is 6%, whereas the overall difference in PPV is only 3%.

When examining the 14 test data sets for which the inclusion in-silico nucleosome allowed the detection of a weak signal, we notice that the Sensitivity and PPV in all cases is very low (< 0.17), raising the possibility that the changes in prior induced by in-silico nucleosome positions were uninformative. Thus, a novel weak signal would be detected due to mostly random changes in the prior distribution.

This would suggest that the in-silico predicted nucleosomes from Segal et al. are not as accurate as the experimentally derived nucleosome positions of Yuan et al. Examining the testing of the Segal model, it is stated that the model places 54% of nucleosomes within 35 bp of their true positions, versus 39(+/-1)% within 35 bp for randomly determined nucleosome positions. While this does indicate a statistically significant signal, the signal is not biologically relevant in our case. That is, it is not sufficient to induce a clear improvement in the predictive capabilities of the algorithm when included instead of the non-informative uniform prior.

4.4.5. Increased Sensitivity in the Presence of Spurious Sequences

The nature of our problem calls for detecting a weak signal in the presence of large amounts of background "noise" sequence. The longer the promoter sequences we search for TFBS, the weaker the signal. It is therefore important to increase an algorithm's ability to detect the signal in the presence increasing amounts of background sequence.

We wish to test whether adding nucleosome positioning information strengthens the algorithms ability to detect a signal, or a real TFBS instance, against increasingly large amounts of background sequence.

As mentioned previously, the input to the algorithm is a set of promoter sequences, within which we wish to locate real TFBS instances. It is conceivable that a certain fraction of these input sequences will have no instances of the TFBS under consideration. We would consider these to be spurious, or "noise", promoters.

We could systematically measure the number of noise promoters with which a dataset still yields a viable signal when examined using our algorithms. It is desirable that the addition of nucleosome positioning information increases the amount of noise in whose presence a real TFBS instance is still detected, when compared to the basic algorithm.

4.4.5.1. Generating Extended Test Data Sets

The idea behind an extended version of an existing test data set is simple. We start with one of the test data sets discussed previously, containing instances of a given TFBS. To this core test set, we add one or more promoter sequences with no known instances of the TFBS. Thus we end up with increasingly larger test data sets with one, two, and up to twenty additional noise promoter sequences.

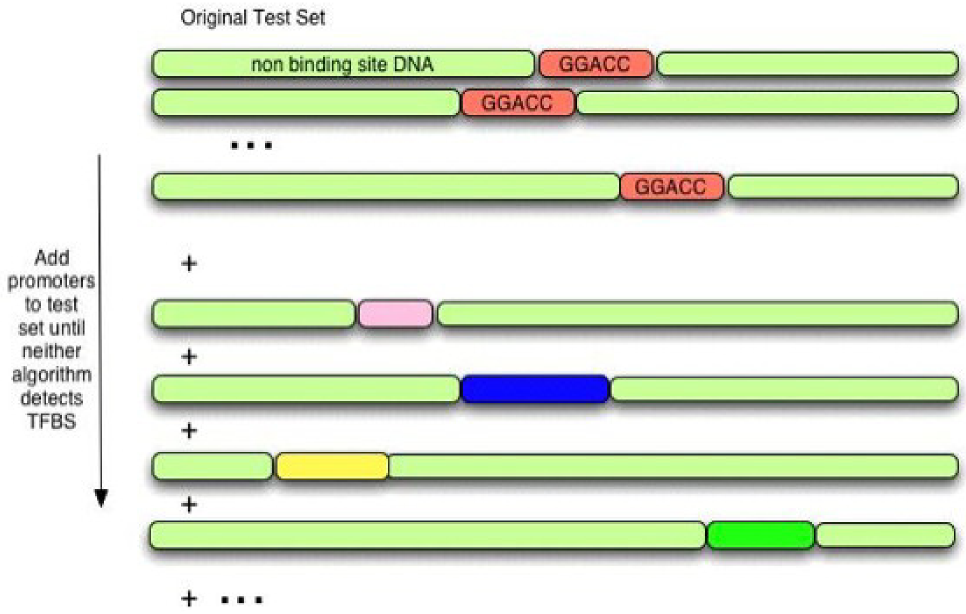


Figure 12 - Adding noise to test data sets by adding an increasing number of promoter sequences not containing the TFBS of interest.

We then compare the number of additional sequences with which the basic and modified versions of the algorithm still detect a viable signal. That is, at least one real instance of the TFBS under consideration. For example, suppose the basic algorithm detects a valid signal with at most four additional noise promoters added to the core test data set. The modified algorithm detects a valid signal with at most six additional promoters. Then the modified algorithm allows the detection of a valid signal, or known TFBS instance, in the presence of two additional noise promoters, or approximately 1600 bp (each promoter being about 800 bp).

4.4.5.2. Test Results

The graph below summarizes the above difference in detection thresholds for the 24 TFBS where both versions of the algorithm detected at least one true TFBS instance.

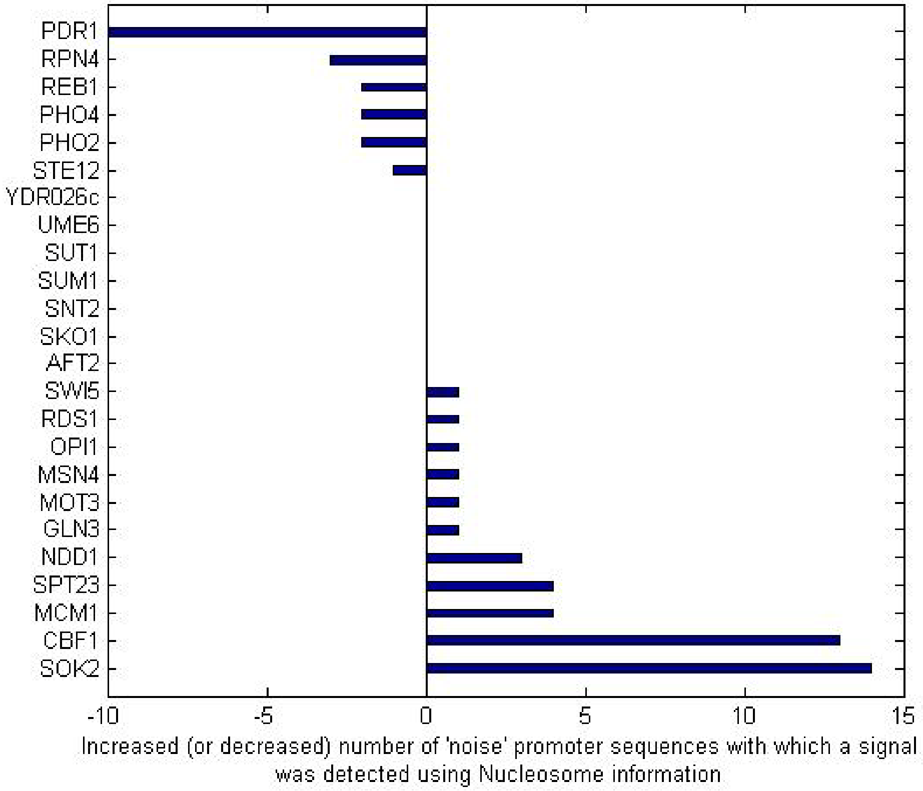


Figure 13 – comparing algorithm ability to predict true binding sites in the presence of increasing numbers of promoters with no true binding sites. A positive x value indicates the algorithm incorporating experimental nucleosome information detected a true binding site in the presence of more spurious promoters, whereas a negative value indicates that the basic algorithm did better.

Here, a positive bar indicates that the modified EM algorithm detected a signal in the presence of more noise promoters than basic EM. Conversely, a negative bar indicates that the basic algorithm performed better in this regard.

One way to objectively assess the difference attributable to the prior information is through the use of Wilcoxon Signed Rank test. We are given a set of paired observations, each of which are assumed generated by two independent random processes. The null hypothesis is that the pairs are generated by processes with identical means. A low p-value indicates that the means of the paired experiments are different. In our case, the random processes are the basic EM algorithm and the version augmented with prior information. Paired observations are the maximum number added noise promoters with which a signal is still detected.

In this case, the test p-value of 0.34 does not rule out the null hypothesis. Thus, while the graph above hints that adding nucleosome information allows improved signal detection in the presence of increasing noise, particularly for CBF1 and SOK2 binding sites, we cannot claim a statistically significant improvement for binding sites of all TFs. It should be noted that, given that the two versions of the algorithm are fundamentally related, the independence assumptions for the signed-rank test do not necessarily hold, given the fact that the prediction algorithm remains mostly unchanged. Thus the random processes cannot truly be said to be independent.

Here is the same experiment as above, except that the nucleosome positions used are the in-silico predictions of Segal et al.

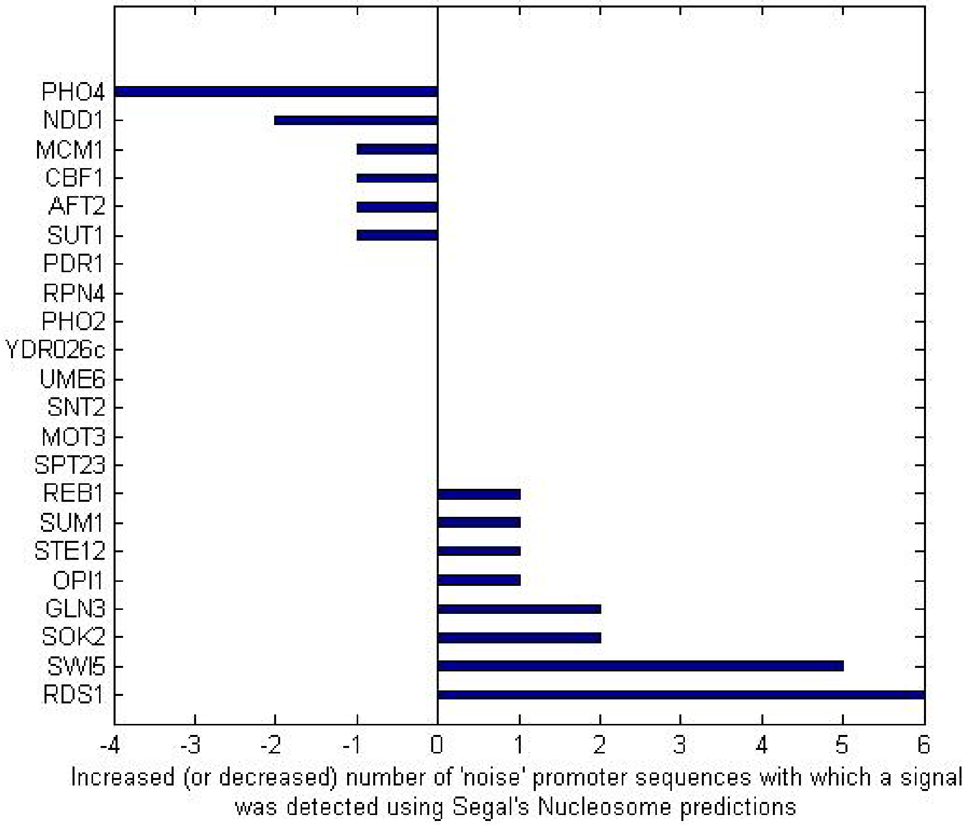


Figure 14 – comparing algorithm ability to predict true binding sites in the presence of increasing numbers of promoters with no true binding sites. A positive x value indicates the algorithm incorporating experimental nucleosome information detected a true binding site in the presence of more spurious promoters, whereas a negative value indicates that the basic algorithm did better.

We see less extreme discrepancies when comparing to the previous graph, and significantly less increased resilience to noise promoter sequences. The graph also looks more symmetric. That is, the addition of information from Segal et al. seems to decrease and increase detection ability for comparable numbers of test data sets, and by matching amounts. Had the nucleosome positions been highly informative, we would have expected to see more improvement, as was seen with the addition of experimental nucleosome information.

The lack of visible improvement is consistent with the lack of marked improvement in expected Sensitivity or PPV associated with the additional information above.

4.6. Conclusion

In this chapter, we used a combination of simulated, experimentally generated and in-silico predicted data to evaluate the effect of incorporating nucleosome-based prior information on the mixture model based motif finding algorithm described in chapter 3. We looked for improvement along two dimensions: the ability to detect valid binding sites in a greater number of datasets, and an improved ability to detect a valid signal as the signal is attenuated by increasing amounts of irrelevant data.

As expected, we found that, with simulated data, the improvement was clear. With experimentally derived data, some improvement was observed in both the number of datasets for which a signal was detected, and the continued ability to detect a signal in the presence of increasing noise. However, this improvement was modest and based on a small number of test datasets in which a significant improvement was observed. Finally, using in-silico predicted nucleosomes fails to improve the predictive ability of the model.

Conclusion

Predicting transcription factor binding sites with precision is a challenging problem. This is because protein-DNA interactions are crucially based on the structures of the protein and the region of the DNA helix to which it binds. The amount of DNA to search is vast, the binding sites short and the sequence motifs degenerate. As a result, algorithms using only sequence information to make predictions only account for part of what identifies and distinguishes biologically relevant transcription factor binding sites. While sequence information is the most widely available and easiest to model, additional, biologically relevant data can and should play a role in predictions.

Genome-scale nucleosome positioning data has rapidly become available over the past several years, and this trend will continue, as it has with other types of genomics and proteomics data. The challenge then is to use it effectively. Incorporating nucleosome positioning information addresses one of the weaknesses of motif finding by taking into consideration biologically relevant information that has a marked effect on the positioning of transcription factor binding sites.

To this end, we jointly evaluated a genome-scale nucleosome positioning dataset and genome wide transcription factor binding sites from yeast to derive detailed, empirically valid relationships between binding site and nucleosome positioning. We then modified an established motif finding algorithm to incorporate nucleosome positioning information. Finally, we evaluated the basic algorithm and the modified version to quantify the improvement due to this additional information.

This work also includes several significant limitations and room for improvement. Firstly, we made simplifying assumptions regarding the positioning of nucleosomes. As mentioned in the first chapter, nucleosomes in vivo are not static structures with necessarily well-defined and invariant genomic positions. While DNA sequence properties predispose certain tracts of DNA to nucleosome formation, such positioning is not guaranteed. Thus, when considering a population of cells, each with identical nuclear DNA, nucleosomes may form at different spots along the same tract of DNA in different cells of this population. DNA-encoded positioning constraints provide at most a preferential positioning rather than an absolute one.

Thus, if nucleosome positions are derived from experimental source such as microarrays, as the data used in the present work, a statistically significant number of microarrays using different samples may be preferable to form a more statistically accurate picture of the relative positioning preferences of nucleosomes along a given stretch of DNA. One could then determine a more accurate empirical prior, albeit a prior with less extreme contrasts between regions of high and low nucleosome affinity.

An additional limitation of this work is the treatment of all TFBS as having an "average" affinity towards linker DNA regions. As demonstrated by the analysis of the Harbison et al. binding site dataset, individual binding site vary widely in their overabundance in linker DNA, with a few even preferentially occurring within the confines of nucleosomal DNA. Thus, additional information about the expected types of TFBS could be used to more accurately construct a prior by indicating the degree of TFBS (under)over -abundance in linker regions.

This would be the case, for example, when the results of a ChIP-Chip experiment are analyzed and the TFBS of interest is known by the experimenters as part of experiment design and antibody selection. This would however require that a previous analysis of the positioning affinities of the TF under consideration be available, which is not necessarily feasible.

It would also be worth investigating whether structurally similar TFBS families have comparable positioning affinities. Such an approach may be fruitful for proteins derived from paralogous genes or structurally similar proteins within a species, as well as orthologous proteins across different species. Thus, it may be possible to derive a more accurate prior by inferring TFBS affinities using data for the binding sites of such related proteins.

One additional type of information that may be readily incorporated is a position-specific prior with respect to the Transcription Start Site (TSS). It has been observed that different TFBS occur preferentially within certain distances of the TSS. This could be easily accommodated by the framework described in the previous chapters.

Computationally, several notable improvements could be implemented. One would be the option to model the palindromic structure of certain classes of TFBS by placing constraints on "paired" beginning/end positions within the PWM. This would begin to address a more general shortcoming of our approach, which is the assumed independence of different positions within a predicted motif. It is often the case that the first and last few positions of a TFBS are highly conserved, while intermediate positions are degenerate. An initial attempt could be made to address this shortcoming by imposing constraints among the first or last few bases using an information theoretic measure such as mutual information. Or, alternatively, a requirement could be imposed that at least one pair of positions, adjacent or not, have high mutual information in the motif model represented by the PWM.

Another improvement would be the use of a higher order MM as background model, or better yet, a more sophisticated model incorporating positional trends in nucleotide content, such as that proposed in the BayedMD approach.

The idea in all cases is to incrementally improve the model to incorporate additional relevant structure, in order to better reflect observed biological regularities that may be exploited to improve prediction performance.

References

Anselmi, C., et al. "A theoretical model for the prediction of sequence-dependent nucleosome thermodynamic stability." *Biophysical Journal*, 2000: 601-613.

Bailey, T. L., and C. Elkan. "Fitting a mixture model by expectation maximization to discover motifs in biopolymers." *Proceedings of the international conference on intelligent systems in molecular biology (ISMB)*, 1994: 28-36.

Bembom, B, et al. "Supervised detection of conserved motifs in DNA sequences with cosmo." *U.C. Berkeley Division of Biostatistics Working Paper Series*, 2006: -.

Blanchette, M. et al. "COMP 618 - Functional Genomics, Class Notes." Notes, Montreal, 2006.

Blanchette, M. and Tompa, M. "Discovery of regulatory elements by a computational method for phylogenetic footprinting" *Genome Research*, 2002, 12: 739-748

Buck, M.J., and Lieb, J. D. "ChIP-chip: considerations for the design, analysis and application of genome-wide immunoprecipitation experiments." *Genomics*, 2004: 349-360.

Chen, L, and J Widom. "Mechanism of transcriptional silencing in Yeast." *Cell*, 2005: 37-48.

Cherry, J.M., et al. "Genetic and physical maps of saccharomyces cerevisiae." *Nature*, 1997: 67-73.

Cloutier, T, and J Widom. "DNA twisting flexibility and the formation of sharply looped protein-DNA complexes." *Proceedings of the National Academy of Sciences (PNAS)*, 2005: 3645-3650.

Dempster, A. P., et al. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistics Society B*, 1977: 1-38.

D'Haeseleer, P. "How does DNA motif discovery work?" *Nature Biotechnology*, 2006: 959-961.

D'Haeseleer, P. "What are DNA sequence motifs?" *Nature Biotechnology*, 2006: 423-425.

Durbin, R, et al. *Biological sequence analysis: probabilistic models of proteins and nucleic acids.* Cambridge: Cambridge University Press, 1998.

Ewens, W.J., and E.R. Grant. *Statistical methods in bioinformatics - an introduction.* New York: Springer, 2001.

Geiman, T.M., and K.D. Robertson. "Chromatin remodeling, histone modification and DNA methylation - how does it all fit together?" *Journal of Cellular Biochemistry*, 2002: 117-125.

Giguere, V. "Orphan Nuclear Receptors: From Gene to Function." *Endocrine Reviews*, 1999: 689.

Harbison, Chrostopher, et al. "Transcriptional regulartory code of a eukaryotic genome." *Nature* 431 (2004): 99-104.

Hu, J., B. Li, and Kihara, D. "Limitations and potentials of current motif discovery algorithms." *Nucleic Acids Research*, 2005: 4899-4913.

Karlin, S., and S. Altschul. "Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes." *Proceedings of the National Academy of Sciences (PNAS)*, 1990: 2264-2268.

Kellis, M., et al. "Methods in comparative genomics: genome correspondence, gene identification and motif discovery." *Journal of Computational Biology*, 2004: 319-355.

Kent, W.J., et al. "The human genome browser at UCSC." *Genome Research*, 2002: 996-1006.

Khorasanizadeh, S. "The nucleosome: from genomic organization to genomic regulation." *Cell*, 2005: 259-272.

Kiyama, R., and Trifonov, E.N. "What positions nucleosomes? - a model." *Federation of European Biochemical Societies Letters*, 2002: 7-11.

Kiyama, Y., et al. "DNA bend sites in the human beta-globin locus: evidence for a basic and universal component of genomic DNA." *Molecular Biology and Evolution*, 1999: 922-930.

Lawrence, C.E., et al. "Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment." *Science* 1993: 208–214.

Li, G., et al. "Rapid spontaneous accessibility of nucleosomal DNA." *Nature Structural and Molecular Biology*, 2005: 46-53.

Lodish, Harvey, Berk, Zipursky, Matsudaira, and Darnell. *Molecular Cell Biology.* New York: W. H. Freeman & Co., 2004.

Lowary, P.T., and Widom, J. "New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning." *Journal of molecular biology*, 1998: 19-42.

Lowary, P.T., and J. Widom. "Nucleosome packaging and nucleosome positioning of genomic DNA." *Proceedings of the National Academy of Sciences (PNAS)*, 1997: 1183-1188.

Makeev, V.J., et al. "Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information." *Nucleic Acids Research*, 2003: 20-31.

Miller, J.A., and J. Widom. "Collaborative competitition mechanism for gene activation in vivo." *Molecular and Cellular Biology*, 2003: 1623-1632.

Narlikar, L., Godran, R., and Hartemink, A.J. "A nucleosome-guided map of transcription factor binding sites in yeast" *PLOS Computational Biology*, 2007: 2199-2208

Pavesi, G. et al. "Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes." *Nucleic Acids Research, 2004* W199–W203

Richmond, T.J., and C.A. Davey. "The structure of DNA in the nucleosome core." *Nature*, 2003: 145-150.

Rubin, G.M., et al. "Comparative genomics of the eukaryotes." *Science*, 2000: 2204-2215.

Saha, A, et al. "Chromatin remodeling through directional DNA translocation from an internal nuclesomal site." *Nature Structural and Molecular Biology* 12, no. 9 (2005): 747-755.

Schalch, T., et al. "X-ray structure of the tetranucleosome and its implcatios for the chromatin fibre." *Nature*, 2005: 138-141.

Scipioni, A., et al. "Dual role of sequence-dependent DNA curvature in nucleosome stability: the critical test of highly bent Crithidia Fasciculata DNA tract." *Biophysical Chemistry*, 2004: 7-17.

Segal, E., et al. "A genomic code for nucleosome positioning." *Nature* 442 (2006): 772-778.

Sinha, S. & Tompa, M. "YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation." *Nucleic Acids Research,* 2003: 3586–3588

Stoughton, R.B. "Application of DNA microarrays in biology." *Annual Reviews Biochemistry* 74 (2005): 53-82.

Strohner, R. e. A. "A loop recapture mechanism for ACF-dependent chromatin remodeling." *Nature Structural and Molecular Biology*, 2005: 683-690.

Tang, M.H., Krogh, A. and Winther, O. "Bayes MD: flexible biological modeling for motif discovery" *Journal of Computational Biology* 2008: 1347:1363

Tompa, M. *et al*. "Assessing computational tools for the discovery of transcription factor binding sites." *Nature Biotechnology,* 2005: 137–144

Widom, J. "Role of DNA sequence in nucleosome stability and dynamics." *Quarterly Reviews of Biophysics*, 2001: 269-324.

Wiggins, T., et al. "Exact theory of kinkable elastic polymers." *Physical Review Letters E*, 2005: 1-19.

Wray, G.A., et al. "The evolution of transcriptional regulation in Eukaryotes." *Molecular Biology and Evolution*, 2003: 1377-1419.

Yuan, G.C., et al. "Genome scale identification of nucleosomes in S. Cerevisiae." *Science*, 2005: 626-630.