PHYSICALLY INTERPRETABLE MACHINE LEARNING METHODS FOR TRANSCRIPTION FACTOR BINDING SITE IDENTIFICATION USING PRINCIPLED ENERGY THRESHOLDS AND OCCUPANCY

by

AMAR MOHAN DRAWID

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Computational Biology and Molecular Biophysics

written under the direction of

Prof. Anirvan Sengupta

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

January, 2009

ABSTRACT OF THE DISSERTATION

Physically Interpretable Machine Learning Methods for Transcription Factor Binding Site Identification Using Principled Energy Thresholds and Occupancy

By AMAR MOHAN DRAWID

Dissertation Director:

Prof. Anirvan Sengupta

Regulation of gene expression is pivotal to cell behavior. It is achieved predominantly by transcription factor proteins binding to specific DNA sequences (sites) in gene promoters. Identification of these short, degenerate sites is therefore an important problem in biology. The major drawbacks of the probabilistic machine learning methods in vogue are the use of arbitrary thresholds and the lack of biophysical interpretations of statistical quantities. We have developed two machine learning methods and linked them to the biophysics of transcription factor binding by incorporating simple physical interactions. These methods

estimate site binding energy, recognizing that it determines a site's function and evolutionary fitness. They use the occupancy probability of a transcription factor on a DNA sequence as the discriminant function because it has a straightforward physical interpretation, forms a bridge between binding energy and evolutionary fitness, and has a natural threshold for classifying sequences into sites that allows establishing the threshold in a principled manner. Our methods incorporate additional characteristics of sites to enhance their identification. The first method, based on a hidden Markov model (HMM), identifies self-overlapping sites by combining the effects of their alternative binding modes. It learns the threshold by training emission probabilities using unaligned sequences containing known sites and estimating transition probabilities to reflect site density in all promoters in a genome. While identifying sites, it adjusts parameters to model site density changing with the distance from the transcription start site. Moreover, it provides guidance for designing padding sequences in experiments involving self-overlapping sites. Our second method, the Phylogeny-based Quadratic Programming Method of Energy Matrix Estimation (PhyloQPMEME), integrates evolutionary conservation to reduce false positives while identifying sites. It learns the threshold by solving an iterative quadratic programming problem to optimize the distribution of correlated binding energies of neutrally evolving orthologous sequences while restricting the values of binding energies of known sites and their orthologs. We have used the NF-κB transcription factor family as a case study for both methods and gained new insights into its biology.

Acknowledgements

I would like to thank:

Anirvan Sengupta for being a fantastic advisor and making my graduate school experience a very pleasant one; Ron Levy for giving me the opportunity to pursue my education while working and for supporting my work all these years; Céline Gélinas for a great collaboration and for her insightful analysis of the biological significance of NF-κB targets identified by the HMM; Gyan Bhanot for his support and advice during the graduate school work; Nupur Gupta for a great collaboration and the experimental work; Viji Nagaraj for the experimental validation work; Paul Ehrlich for helping me during admission and throughout all these years; my fellow BioMaPS graduate students for their help, fun and camaraderie; Sanofi-Aventis tuition reimbursement program for the financial support; Lisa Vawter, Bob Dinerstein, Michael Tocci, Christoph Brockel and Art Williams for encouraging me to join graduate school and for managing my workload that made my going to school possible; Barbara Butler for the moral support and great discussions; Nilofer Jiwani, Tai-he Xia and many of my Sanofi-Aventis colleagues for their advice and for sharing my office workload; Sayali for credulously listening to me say “I am going to submit two papers in the next two months” for the past two years and for everything; and Mom and Dad for everything.

Table of Contents

List of Tables

List of Figures

Chapter 1

Biological Context

"I am among those who think that science has great beauty. A scientist in his laboratory is not only a technician: he is also a child placed before natural phenomena which impress him like a fairy tale."
Marie Curie (1867-1934)

1.1 Regulation of Gene Expression

The fundamental problem in biology is to understand how organisms function. It requires answers to several related questions: How does a single fertilized cell develop into a multicellular organism? How do thousands of molecules in a cell interact with one another? How do hundreds of different cell types form despite the identical genetic content in each cell? How do increasingly complex structures such as tissues and organs work as single units? How do 10^{14} cells in a human body cooperate with each other and function in harmony? How do organisms respond to external signals? How do species adapt to their environment and evolve? Scientists strive to seek answers to these questions not just to satisfy their intellectual curiosity. They expect that these answers will provide a key to two issues important to our welfare: Why do things go wrong? And, how do we correct them? Let's take an example of cancer. How does a physiological disorder in which a flaw in a tiny part of a single cell impairs the function of other cells and ultimately ends up destroying the entire organism? And how do we snuff out the

rogue cell while protecting the good ones? Admittedly, these are problems of tall order. Our knowledge of biological systems still remains pitifully miniscule in spite of the fact that we have taken tremendous strides in our understanding in the last few decades. We have at least established a few basic underlying paradigms.

A cell, the basic unit of an organism, receives an external signal and gives out an appropriate response. One of the important paradigms is that different cells are programmed to respond to specific sets of signals in different manners. This accounts for various cellular behaviors (phenotypes) such as physiological functions, structural changes, growth, differentiation, morphogenesis and death [1]. To give a few examples, the neurotransmitter acetylcholine decreases the rate and force of contraction of heart muscle cells. The cytokine interleukin-2 stimulates growth and differentiation of immune cells. During the development of an organism, cells in its different parts receive different stimuli and some become neurons, some others epithelial cells, some others yet lymphocytes. Interestingly, different cell types sometimes give different responses even to the same stimulus. For example, when treated with glucocorticoid hormones, liver cells increase glucose production, fat cells reduce tyrosine aminotransferase production, while some other cell types do not respond at all. Although we know the nature of external signals and the resulting cellular phenotype, as in many biological experiments, the exact cellular mechanism precipitating the change still eludes us.

What we do know, however, of this mechanism is that regulation of gene expression is pivotal to cell behavior. Gene expression is the process of synthesizing RNA and

proteins, the functional products of a gene. Different cell types have distinct phenotypes on account of differences in the relative abundance of these products. The regulation of gene expression, or gene regulation, consists of controlling the abundance and timing of these products in response to external conditions. Cells typically contain hundreds of complex and interconnected molecular signaling pathways (cascades) starting with receptor proteins which sense external signals. Signals are then propagated, amplified, combined and tuned along the pathway, ultimately resulting in the change in the abundance of RNA and proteins that alters the cell phenotype. The reason why the signals cause varied responses in different cells is that they regulate gene expression differently, i.e. change the abundance of these products differently. Thus, gene regulation is the key to an organism's survival and adaptability, and defects in this process result in pathogenesis.

In eukaryotes, gene regulation is carried out in any of the following ways: (1) Specialized proteins called transcription factors determine the rate of transcription from DNA to mRNA (messenger RNA) in a process called transcriptional regulation. (2) The transcription of some mRNA molecules is prematurely terminated in a process known as transcription attenuation. (3) RNA molecules are alternatively spliced, incorporating different combinations of exons and sometimes of introns. (4) Cleavages at different sites at the 3' end create mRNA molecules of different lengths. (5) Nucleotide sequences of some mRNA molecules are changed in mRNA editing. (6) The addition of a 5' cap and a 3' poly-A tail stabilizes mRNAs. (7) Shortening of a poly-A tail or the presence of certain sequences in the 3' untranslated region (UTR) causes mRNA decay. (8) Export of

mRNAs from nucleus to cytoplasm is regulated. (9) Proteins bindings to the 5' and 3' UTRs of mRNAs can abort translation from RNA to protein at the initiation, elongation or termination steps. (10) Post-translational modifications such as phosphorylation modulate the amount of functional protein.

Of all these, transcriptional regulation, in which transcription factors bind to specific DNA sequences in the promoters of particular genes, is not only chronologically the first but also the most predominant form of gene regulation as it limits synthesis of unwanted products and thereby saves energy. Transcription factors thus play a central role in establishing, maintaining and altering cell behavior [2]. They are to a large extent responsible for the tremendous diversity in cell behavior.

1.2 Transcription Factors

Transcription factors are proteins that, as we have seen, regulate gene expression by binding to specific DNA sequences [3-5]. They activate or repress the expression of certain genes which are their direct targets. These (specific) transcription factors should be distinguished from general transcription factors, a set of extremely well-conserved proteins responsible for initiating transcription of every mRNA in eukaryotes by assembling with RNA polymerase II.

The more complex the organism the greater are the transcription factors, not just in absolute number but also in their ratio to the gene population [6]. To wit, only 5% of the

~6000 yeast genes code for transcription factors; it is 10% in humans. In humans, transcription factors form the largest protein family. As each gene is regulated by multiple transcription factors, the 2000+ transcription factors in humans can potentially yield millions of combinations. Moreover, formation of transcription factor complexes, either of the same type (homodimers) or of different types (heterodimers), further increases the variety. This phenomenon facilitates singular responses to thousands of environmental conditions.

Some examples of transcription factors are: (1) Highly conserved Hox proteins which control body pattern formation during the development of an organism, (2) Tissue-specific transcription factors, such as the liver-specific HNF proteins, which maintain cell phenotypes, (3) p53 which suppresses tumors by activating DNA repair machinery, arresting cell cycle and inducing apoptosis, (4) Heat shock factor (HSF) which induces expression of genes important for survival at high temperatures, and (5) Members of the Jun, Fos and ATF transcription factor families which control cell proliferation, differentiation and transformation.

A transcription factor protein consists of modules, called protein domains. One such domain is the DNA-binding domain (DBD), which binds to specific DNA sequences in its target genes. Another is the transactivation domain, also known as activation function (AF). It forms complexes with auxiliary gene regulatory proteins called cofactors. One other domain is an optional ligand-binding domain (LBD), which binds to signaling molecules.

External signals activate transcription factors in a number of ways [7]. Signaling molecules bind either directly to nuclear transcription factors (e.g. hormones binding to nuclear hormone receptors such as glucocorticoid or estrogen receptors) or to receptors on the cell surface. In the latter case, the resulting molecular signaling cascades activate the nuclear transcription factors such as CREB and ATM family members by phosphorylating their particular serine residues. The signaling cascades also activate cytoplasmic transcription factors using various mechanisms: serine (e.g. SMADs) or tyrosine phosphorylation (e.g. STATs), removal of phosphorylation (e.g. NFAT), proteolytic cleavage of the transcription factor (e.g. Notch) or its binding molecule (e.g. NF-κB, see more below). The activated cytoplasmic transcription factors then translocate to the nucleus. Internal signals also activate transcription factors using similar mechanisms (e.g. DNA damage activates p53).

The activation or repression of gene expression is effected through various mechanisms. In one mechanism, activating transcription factors promote the assembly of RNA polymerase II and general transcription factors. In another mechanism, activating transcription factors simultaneously decondense the chromatin, a packaged complex of DNA and proteins (e.g. histones). Decondensation consists of acetylation of histones and remodeling of nucleosomes (the basic units of chromatin). These together make the DNA more accessible to transcription. Repressing transcription factors work just the reverse way. They inhibit transcription by blocking the assembly of RNA polymerase II and general transcription factors. They also condense the chromatin by deacetylating histones

and remodeling nucleosomes. Repression can sometimes also be carried through by inhibiting functional binding of DNA to activating transcription factors. The same transcription factors that activate gene expression in some cases may repress it in other cases, depending upon target genes, external signals and/or cell types. For example, while NF-κB is known to activate several hundred genes, it also represses the BLNK and BCAP genes in the B-cell receptor pathway in lymphoid cells [8].

Thus, transcription factors are critical in controlling cell behavior. What now remains to be done is the identification of DNA sequences to which a given transcription factor will attach itself, like a piece in a jigsaw puzzle.

1.3 Transcription Factor Binding Sites

A transcription factor binds to a specific DNA sequence called a transcription factor binding site, regulatory element or response element. We shall refer to it simply as a "site." A site is a short stretch of DNA 5-20 bp in length, that binds to the DBD (DNA-binding domain) of a transcription factor by forming non-covalent bonds (such as hydrogen or van der Waals). Naturally, the DBD will try to seek those DNA sequences which contain energetically most favorable nucleotides at each position. In certain positions in a site, the interaction between a nucleotide and the DBD is so strong that a specific nucleotide will always be present at that position. On the other hand, nucleotide variations are acceptable at a site position where the interactions are weak, which accounts for the degeneracy.

In higher eukaryotes, sites are located at various positions along a gene. A vast majority of these are present in the region of –1000 to +250 bp around the transcription start site (TSS), many of them in the proximal promoter (the 250 bp region upstream of the TSS.) A small number of sites may also be found further upstream or downstream of the gene or in an intron. Sites often form complexes, known as *cis*-regulatory modules (CRMs), a few hundred base pairs in length, containing one or more binding sites for multiple transcription factors. The exact order and spacing of sites within a CRM is usually not important. CRMs activating gene expression are called enhancers and those repressing gene expression are called silencers. CRMs can be present in an intron or upstream or downstream of a gene as far as 100 kb away from a gene. Thus, determining the location of a site can be quite difficult.

Identification of transcription factor binding sites is an important problem in biology. Site identification helps determine the direct targets of each transcription factor and enables enumeration of various transcription factors controlling each gene's expression. It is the first step towards understanding the combinatorial effects of transcription factors on gene regulation. If we can then measure the exact nature of gene regulation, that is how transcription factors quantitatively affect a gene's expression independently and collectively, we can combine this information with the existing knowledge about signaling cascades and the functions of transcription factors and target genes, and thus begin to paint a picture of the mechanism by which a cell responds to environmental signals. We can thus decipher the black box in an important paradigm of biology.

Popular experimental methods for identifying sites include traditional low-throughput methods such as gel shift assay, DNA footprinting and ChIP; and high-throughput methods such as SELEX and ChIP-chip [1, 9-12]. In a gel shift assay or EMSA (electrophoretic mobility shift assay), DNA fragments travel through an electric field in a polyacrylamide gel, and are separated by size as larger fragments, facing greater viscous forces, advance more slowly through the gel. Thus, transcription factor-bound DNA fragments separate from unbound DNA due to the slow progression of the former. After removal of the protein, they are amplified using multiple rounds of PCR (polymerase chain reaction) and sequenced. We can identify the exact location of a site in a DNA fragment by DNA footprinting, in which multiple copies of the DNA fragment are labeled at one end, bound to the transcription factor and cleaved at random positions with a nuclease. They are then run on a gel shift assay, where no bands (footprint) are observed at locations corresponding to protection from cleavage due to protein binding. In each round of the high-throughput *in vitro* method SELEX (systematic evolution of ligands by exponential enrichment), DNA fragments from a combinatorial library that are bound to a transcription factor are separated from unbound DNA and PCR amplified, thus detecting even weak-binding sites after multiple rounds of partition and amplification. We can also identify sites occupied by a known protein in the native DNA structure of cells of interest using ChIP (chromatin immunoprecipitation), in which the bound protein is cross-linked with the DNA using formaldehyde, cells are lysed, DNA is broken into fragments, and the fragments bound to the protein are precipitated using an antibody against the protein, after which cross-linking is reversed and the DNA sequence

is identified using PCR amplification. In its high-throughput version called ChIP-chip, DNA sequences identified in ChIP are PCR amplified uniformly and hybridized to microarray chips containing genomic sequences to identify genomic binding locations. After identifying a site, its functional importance can be determined in site-directed mutagenesis by cloning the promoter or enhancer containing the site in front of a reporter gene and studying the change in the readout after modifying some nucleotides in the site. Despite the progress in the above methods, experimental identification of all sites for all transcription factors in higher eukaryotes is still a difficult proposition in view of the facts that the genes number in thousands, that genome sizes are large, and that there is a great variability in the position of regulatory regions with respect to genes. Experimentally, these cause low signal-to-noise ratios.

Because experimental methods have limitations at the present time, we need to resort to computational methods to identify transcription factor binding sites. While many of the currently used computational methods restrict themselves to identifying sites, methods that are also able to predict the occupancy of transcription factors on the sites will take us one step closer to understanding the exact nature of gene regulation. In this thesis, we describe two methods that identify transcription factor binding sites and predict occupancy of transcription factors on these sites.

A model of a transcription factor's binding sites that incorporates the degeneracy is called a "motif." In the next chapter, we will study three important motif models – regular expression (also called consensus sequence), weight matrix and energy matrix. In this

section, we will review the popular visual representations of these models (Figure 1.1) [13-17]. The most basic representation corresponds to a consensus sequence or regular expression. In this, each position is represented by the symbol of the most prevalent nucleotide at that position. If two or more nucleotides are present at a position with nearly equally high frequency, a symbol representing all of them is used (e.g. R for purines A and G). N represents positions with no nucleotide preference. Another representation is a 4-by-ℓ table, where ℓ is the motif length, corresponding to a weight matrix or an energy matrix. Each element of the table contains the relative frequency (for weight matrix) or the binding energy (for energy matrix) of each nucleotide at every position of the site. In the third, perhaps the most visually appealing, representation, known as a sequence logo, the overall height of the nucleotide stack at each position is drawn proportional to the information content at that position, and the height of each nucleotide within the stack is proportional to its relative frequency.

In the next section, we will take up an important family of transcription factors, the NF-κB (nuclear factor-kappa B) family. These transcription factors are critical for the development and well-being of an organism. We have taken NF-κB as the case study in this thesis for reasons that will become apparent in the course of the discussion. We will see what they are, what processes they regulate, how they regulate them, how defects in their regulation leads to pathogenesis, and also how certain viruses manage to cheat the immune system which is the main defense of an organism.

1.4 NF-κB and Its Self-Overlapping Binding Sites

The NF-κB family plays a major balancing role in several cellular processes including cell cycle control, growth, proliferation, apoptosis (programmed cell death) and differentiation [8, 18-49]. It is involved in the nervous, hepatic, epidermal (e.g. hair, tooth, mammary glands) and many other systems. But it is the immune system where NF-κB is of paramount importance. It regulates the cell formation, negative or positive selection leading respectively to either proliferation or apoptosis of cells, cell differentiation and maturation, as well as the survival after maturity of most immune cells. It is also implicated in the development of primary (bone marrow and thymus) and secondary (spleen, lymph nodes, etc.) lymphoid organs.

Various trauma conditions such as invasion by microbes, irradiation, oxidative stress, injury, hemorrhagic shock (excessive blood loss), heavy metals and therapeutic drugs (including chemotherapeutic agents) also activate NF-κB. In the event of an invasion by viruses, bacteria, fungi or parasites, pathogen-associated molecular patterns (PAMPs) activate NF-κB via pattern recognition receptors (PRRs) such as toll-like receptors (TLRs). Activated NF-κB drives an innate (or non-specific) immune response in a process called inflammation. This response consists of (i) the release of antimicrobial proteins, such as defensins and nitrogen and oxygen molecules, (ii) upregulation of cytokines (e.g. tumor necrosis factor and interleukin-1), chemokines, enzymes (e.g. cyclooxygenase-2) and adhesion molecules, and (iii) recruitment and activation of professional immune cells at the site of infection for killing the infectious agents. In a

positive feeback loop, proinflammatory cytokines further activate NF-κB. NF-κB orchestrates the adaptive (or antigen-specific) immune response by modulating the activation, selection and maturation of antigen-presenting cells (APCs) and T and B cells.

The failure of proper regulation of the immune system leading to pathogenesis can also be traced to the NF-κB family. Some viruses use NF-κB to their advantage. HIV, for example, contains DNA sites that are activated by NF-κB in such a way as to cause HIV to proliferate. Another example is that of an avian retrovirus. It carries its own NF-κB homolog, v-Rel, which causes fatal leukemia. Any malfunction of the NF-κB signaling pathways is generally a hallmark of various types of cancer, where NF-κB helps growth of tumor cells, causes them to proliferate and prevents their apoptosis. Furthermore, it causes angiogenesis, local invasion and metastasis, thus spreading cancerous cells to the healthy regions of the organism. In the case of autoimmune diseases, such as chronic inflammation, asthma and rheumatoid arthritis, NF-κB gives out misguided response, and in the case of many infectious diseases inadequate response. The failure of NF-κB in systems other than the immune system can cause neurodegenerative and heart diseases.

The NF-κB family of proteins is among the most studied transcription factors. It is interesting to note that mammals share NF-κB proteins with organisms evolutionarily as distant as jellyfish. Dorsal, Dif and Relish in the model organism fruit fly *Drosophila* are NF-κB proteins important for the species' development and immunity, and have received a great deal of attention in scientific studies. In mammals, the NF-κB family consists of five proteins divided into two subfamilies. All five NF-κB proteins contain a conserved

300 amino acid Rel homology domain (RHD) for DNA binding, dimerization and binding to the inhibitor IκB protein. The RHD also possesses a nuclear localization signal.

The first subfamily of NF-κB, called the Rel subfamily, includes RelA, RelB and c-Rel (with official human gene symbols RELA, RELB and REL, respectively). They contain a C-terminal transcriptional activation domain (TAD) that directly interacts with co-activators and the general transcription factor machinery. The second subfamily consists of p105 and p100 (with official human gene symbols NFKB1 and NFKB2, respectively). Instead of a TAD, they contain C-terminal ankyrin repeats which inhibit their own activity, and which are removed by proteolysis to respectively produce the active proteins p50 and p52.

Like many other transcription factors, all NF-κB proteins form homodimers and heterodimers with other family members *in vivo*, with the exception of RelB, which cannot form homodimers. On the other hand, p50 and p52 can form homodimers, but their homodimers or the p50/p52 heterodimer cannot activate transcription as they lack the necessary TAD. NF-κB dimers have slightly different, not yet well-characterized DNA-binding specificities, expression profiles and target genes. It should be noted that in the entire NF-κB family (including the products and dimers) the p50/RelA heterodimer is so predominant that it is often referred to as NF-κB in the literature.

In the absence of signals, inactive (dormant) NF-κB dimers are sequestered in the cytoplasm. They are bound to an inhibitory protein called IκB. They are exported to the

cytoplasm from the nucleus with the help of a nuclear export signal in their alpha subunit of the IκB protein. Once in the cytoplasm, they cannot enter the nucleus because the IκB protein masks their nuclear localization sequences. The dimers containing p105 or p100 are also kept in the cytoplasm by the ankyrin repeats-containing domain which masks their own nuclear localization signals.

An external signal activates dormant NF-κB through possibly many signal transduction pathways. Two of these are known. One is called classical or canonical, and the other alternative or non-canonical. The classical or canonical pathway is activated primarily during (i) the innate immune response and (ii) the survival of immune cells. In this pathway, PAMPs or proinflammatory cytokines bind to their receptors and phosphorylate the beta subunit of the IKK (IκB kinase) complex. Activated IKK phosphorylates IκB and marks it for degradation, releasing NF-κB for transcription. The alternative or non-canonical pathway is activated primarily during (i) the adaptive immune response, (ii) the development and organization of the secondary lymphoid organs and (iii) B-cell maturation. In this pathway, lymphotoxin, BAFF (B-cell activating factor) or CD40 bind to their receptors and activate NIK (NF-κB inducing kinase), which phosphorylates the alpha subunit of IKK. Activated IKK then phosphorylates the ankyrin-rich domain of p100 in the p100/RelB complex and marks it for degradation, producing p52/RelB.

In the next step of both pathways, the active NF-κB dimer enters the nucleus with the help of the exposed nuclear localization signal and starts regulating the expression of its targets genes. It is sometimes post-transcriptionally modified (e.g. phosphorylated and

acetylated) prior to starting its function. NF-κB activation is often transient due to auto-regulatory feedback loops. For example, one of the genes whose expression is upregulated by NF-κB is its own inhibitory protein IκB-alpha. The newly formed IκB-alpha binds to the nuclear NF-κB and translocates it to the cytoplasm as a dormant dimer with the help of the inhibitory protein's strong nuclear export signal.

Once in the nucleus, NF-κB will bind to a site in its target genes. NF-κB binding sites (κB sites) have the consensus nucleotide sequence GGGRNNYYCC [50]. (As explained before, N stands for no preference, R for either A or G, and Y for either T or C.) As expected from the high evolutionary conservation of the NF-κB proteins, κB sites are also found to be highly conserved in mammals [51]. In Chapter 5, we will develop a biophysical model to identify conserved κB sites.

κB sites often self-overlap because they contain multiple G's at the 5' end and multiple C's at the 3' end (Figure 1.2). For a good κB site, when the sequence window is shifted by one position in either the 5' or the 3' direction, the resulting sequence is often a putative κB site. Moreover, we can have additional possibility of self-overlap because the reverse complement of a κB site is often a κB site, allowing functional binding in the opposite direction.

There are examples in which NF-κB can also bind to sites that deviate significantly from the above κB site [45, 65]. RelA/c-Rel heterodimer is known to bind with high affinity to AGGAAAGTAC in the promoter of murine urokinase plasminogen. Similarly, p52/RelB

binds to AGGAGATTTG. However, the incidence of such sites is rare and hence we will not consider them in this thesis.

It is important that a good computational method should take into account all of the alternative binding modes (i.e. include self-overlapping sites) while scoring a candidate regulatory site, although there are many in literature that fail to do so. Ignoring the self-overlapping nature of the site will obviously lead to an incorrect design and interpretation of *in vitro* experiments. For example, a 3' padding sequence (sequence immediately following the site in the experiment construct) starting with nucleotide C in a gel shift experiment can form a spurious strong κB site, and thus confer a falsely high binding affinity to the experimental sequence. This is in spite of the fact that the test sequence, in the context of the native promoter, may make for a weak κB site. A computational method which identifies the self-overlapping sites and helps in the design of padding sequences will obviously avoid such errors. In Chapter 3, we will design a hidden Markov Model to identify self-overlapping κB sites.

Self-overlapping κB sites is not a feature unique to NF-κB. Several transcription factors bind to self-overlapping sites. When a site consists of highly conserved consecutive positions containing the same nucleotide (although the nucleotide can be different for different sets of highly conserved consecutive positions within the site), the transcription factor can bind in different sequence windows, i.e. to self-overlapping sites. Examples include *Drosophila* developmental transcription factor Hunchback [52], worm PHA-4 [53], human Sp-1, C/EBPalpha, yeast ADR1, MIG1, chicken Cdx-1, *Arabidopsis*

Agamous, etc. [54]. Furthermore, when binding by the transcription factor in either orientation is permissible, the corresponding binding DNA site and its reverse complement will have to be considered as two different types of self-overlapping sites as long as they are not exactly palindromic. Hence, the identification of self-overlapping sites and a proper accounting for them is equally important in the treatment of many other transcription factors. Problems similar to the identification of self-overlapping sites also arise in the context of prediction of nucleosome positioning [55].

Although the NF-κB family is probably the most studied transcription factor family and its several target genes, including the sites, have been identified experimentally [18, 45, 50, 56-64] and computationally ([65-68]; see the next chapter for limitations), our knowledge of this family in many respects is far from complete. We do not know whether pathways in addition to the two discussed above exist. In fact, we do not know very much about even these two signaling pathways. The other features of NF-κB that are still mystery to us are (1) its decision-making process concerning apoptosis versus cell survival, (2) the dynamics of its signaling, (3) the composition of its dimers, (4) mechanisms used by NF-κB to control several cellular and organismal processes, and (5) its roles in a multitude of diseases.

We need to identify many more target genes regulated by NF-κB to tackle the above issues. Knowledge of the NF-κB pathways based upon the identification of NF-κB's direct target genes will help understand the regulatory actions of NF-κB and shed light on the pathogenesis of cancer as well as inflammatory and other diseases. Armed with this

knowledge, we will be in a better position to discover new therapeutic targets present in the NF-κB pathways and to decipher the mechanism of action and side-effects of drugs affecting the NF-κB pathways.

1.5 Thesis Overview

In this chapter, we provided the biological framework for transcription factors and their binding sites on the DNA sequences of their target genes. We discussed the importance of identifying the binding sites. We briefly described experimental methods that are in vogue to achieve the goal of such identification. Because we propose to use NF-κB proteins, one of the most important families of transcription factors, as a case study in the subsequent chapters of this thesis, we developed the necessary background of this family in terms of its biology and its self-overlapping binding sites. We remarked upon how the ignorance of self-overlapping sites can lead to inaccurate experiments. In Chapter 2, we will review the computational methods currently used for identifying transcription factor binding sites with an emphasis on the various issues in computational modeling as well as on the limitations of the current models. In Chapter 3, we will address the issue of self-overlapping sites. We will present the hidden Markov model (HMM) method that we have developed for identifying them. In Chapter 4, we will discuss the incorporation of evolutionary information for site identification. We will review the relevant evolutionary models and computational methods. In Chapter 5, we will identify evolutionarily conserved sites by developing a composite model in which we integrate biophysics with evolutionary conservation. We call this model the Phylogeny-based Quadratic

Programming Method of Energy Matrix Estimation (PhyloQPMEME). Finally, in Chapter 6, we will summarize our findings with a discussion on how the methods we have developed will enhance our understanding of NF-κB biology, and with it that of the biological systems in general in which NF-κB plays such an important role. We will close the final chapter with an outlook of the future.

Figure 1.1: Commonly used visual representations of the motif models corresponding to the binding site of the transcription factor NF-κB.

A. Consensus sequence or regular expression. R represents a purine (A or G), Y represents a pyrimidine (C or T) and N represents any nucleotide.

B. Table. Each element contains the relative frequency (for weight matrix) or the binding energy (for energy matrix) of the nucleotide at the site position. Table for a weight matrix is shown.

C. Sequence logo. The overall height of the nucleotide stack at each position is proportional to the information content at that position and the height of each nucleotide within the stack is proportional to its relative frequency.

A **GGGRNNYYCC**

B

A	.00	.00	.03	.50	.56	.16	.02	.04	.03	.01
C	.09	.00	.00	.01	.25	.02	.09	.40	.93	.95
G	.82	.98	.95	.46	.07	.08	.03	.00	.00	.01
T	.08	.01	.01	.02	.11	.74	.85	.55	.03	.02

C

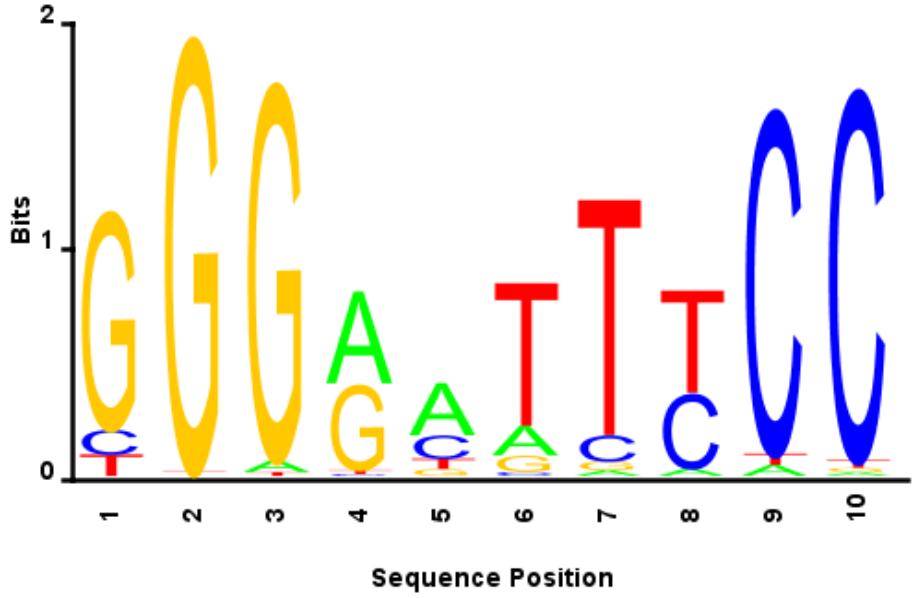
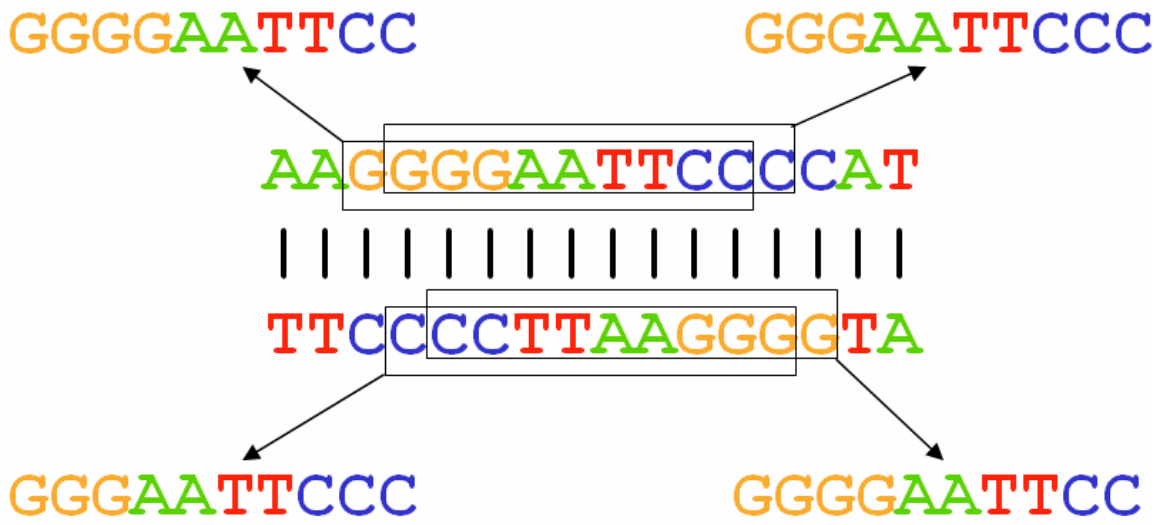


Figure 1.2: Self-overlapping κB sites.

Four self-overlapping κB sites are present on the two strands in three adjacent 10 bp sequence windows.



Chapter 2

Computational Methods for Identifying Transcription Factor Binding Sites

"The important thing in science is not so much to obtain new facts as to discover new ways of thinking about them."
William Bragg (1890-1971)

2.1 Machine Learning for Identifying Binding Sites

Machine learning consists of programming computers to develop the optimum performance criterion based on past examples, which are collectively known as a training set [69]. For this purpose, complex mathematical simulation models are built using statistical theory. Algorithms employing advanced techniques in computer science are then written for efficient execution of programs. Machine learning is either "supervised" or "unsupervised". In supervised learning, each example in the training set contains an input and an output. During the training stage, the algorithm learns the correct mapping (function) of the input to the output while being "supervised" by the output. Then comes the "scoring" stage. In this, the trained algorithm predicts the output of new input data. The two main examples of supervised learning are classification, where the outputs are discrete class labels, and regression, where the outputs are generally continuous numbers. In unsupervised learning, the outputs are not provided in the training set and hence the

algorithm learns only the patterns in the input data during the training stage. It isolates the already learned patterns during the scoring stage.

Machine learning algorithms are used for computationally identifying transcription factor binding sites (referred to as sites). The problem of site identification generally appears in two contexts. In the first context, one wants to identify sites in new sequences based on known examples of experimentally validated sites. This is a supervised machine learning classification problem, where sequences are to be partitioned into two classes – binding sites and non-binding sequences. During the training stage, a learning algorithm (1) constructs a motif (a model of sites) by making certain assumptions, (2) creates a discriminant function that can be evaluated for any sequence, and (3) determines a threshold value (or cutoff, or decision boundary) such that only sequences with the discriminant value on one side of the threshold are considered sites. When scoring a new sequence, it calculates the value of the discriminant function of each subsequence in a sliding window of length equal to that of the site, and compares it to the threshold to determine the positions of potential sites in the sequence. Both algorithms developed in this thesis are supervised machine learning classification algorithms.

In the other context, the training set consists of large sequences, often the promoters of co-expressed or co-regulated genes, which may contain none, one or multiple sites. One wants to determine statistically over-represented motifs in these sequences and then identify corresponding sites in these and new sequences based on a discriminant function and a threshold. This requires unsupervised learning algorithms. We will only review

some of the unsupervised learning algorithms arising in this context, while concentrating almost exclusively on supervised learning algorithms in the thesis.

2.2 Motif Models and Statistical Framework

We now describe motif models used in the machine learning methods for identifying sites. As pointed out earlier, the three important models are regular expression, weight matrix and energy matrix.

The regular expression model is the most basic of the motif models. In the training stage, the most prevalent nucleotide at each position in the training set is assigned that position. (See Section 1.3 for more information.) When scoring a new sequence, its subsequences in a sliding window the size of the site length are declared "hits" if they match the regular expression. Although this model is a good starting point, it fails to take into account statistical variations that are present at sites. Consequently, it is not expected to be highly accurate. Moreover, it has no physical interpretation in terms binding energy or other useful parameters.

Weight matrix [16, 70] and energy matrix [71-74] models, being probabilistic in nature, allow for variations at sites in a meaningful way. The weight matrix model is the most widely used model at present, but, as we will see, it contains certain assumptions that may not always be valid. These assumptions make the mathematics simpler, but the

results are suspect, as they may be fraught with inaccuracies [75]. The energy matrix model attempts to do away with these assumptions, as we will see below.

In both these models, each term in the matrix can be interpreted as a function of the binding energy of each of the four nucleotides at each position in the site, and of the state parameters such as temperature and concentration. The matrix is of dimensions 4 x ℓ, where ℓ is the length of the site. Both models assume that positions in a site are independent. Thus, the total binding energy is just the sum of the binding energies of nucleotides at different positions, quite a good approximation [76] in view of the fact that the second order energy terms are negligibly small. Furthermore, their simple versions also assume that sites in the training set are independent. (Chapter 4 discusses modeling of dependent sites in the training set using evolutionary models.)

In the weight matrix model, the matrix terms are derived with the *a priori* assumption that the occupancy probability, i.e. the probability that a transcription factor occupies a DNA sequence at equilibrium, follows a Boltzmann distribution [17, 75, 77-84]. Each term of the energy matrix, on the other hand, is derived with the consideration that the occupancy probability follows a more general Fermi-Dirac distribution. This distribution reduces to the Boltzmann distribution only in the limiting events when the concentration levels of the transcription factor are very low or when the binding energies are very high (see Section 2.3).

We will now review the general statistical framework for site identification and sophisticated methods of supervised and unsupervised learning based on weight matrix, followed by other methods for site identification. We will discuss the energy matrix model in the next section.

According to the definition of a weight matrix (WM), also called position-specific score matrix (PSSM) or profile and denoted by w, each of its elements is taken to be equal to the probability (i.e. "weight") of each nucleotide at each position of the motif [16, 70]. Let's now see how a weight matrix model is trained. As training sites and positions within each site are assumed to be independent, the likelihood of the training set S is $p(S \mid w) = \prod_{s \in S} p(s \mid w) = \prod_{i=1}^{\ell} \prod_{\alpha} (w_{i\alpha})^{n_{i\alpha}}$, where s is each training site, ℓ is the length of the motif, i is each position in the motif, $w_{i\alpha}$ is the weight (probability) of the nucleotide α at the ith position of the motif and $n_{i\alpha}$ is the frequency of the nucleotide α at the ith position of the training sites. The maximum likelihood estimator (MLE) of $w_{i\alpha}$ is $w_{i\alpha} = \frac{n_{i\alpha}}{n_i}$, where $n_i = \sum_{\alpha} n_{i\alpha}$ is the number of training sites for which the nucleotide at the ith position is known.

One drawback of this formalism is that some small probabilities $w_{i\alpha}$ may get recorded as zero particularly since n_i is small. This happens because the number of known sites of a transcription factor is usually small, and hence the training set may not cover all the possible nucleotides present at a position in a functional site. The probability that a given

sequence is a site is the product of the probabilities of nucleotides at all positions of the sequence. A sequence containing a nucleotide that the training set did not contain at a position will never be identified as a site because the probability of this sequence is zero (a case of false negatives).

This situation is remedied by taking a Bayes estimator. According to the Bayes formula, the posterior probability of the weight matrix given the training set is $p(w|S) \propto p(S|w)p(w)$. The class of the prior probability distribution is chosen such that the distribution of the posterior probability has the same class as that of the likelihood. Such prior probability distribution is said to be conjugate to the likelihood distribution. The Dirichlet distribution, which is the multiple parameter generalization of the beta distribution, is conjugate to the multinomial distribution. Because the nucleotides at a position have a multinomial distribution, a Dirichlet prior is used for weight matrices. The Dirichlet prior at the i th position of the motif has the general form $p(w_i) \propto \prod_{\alpha} (w_{i\alpha})^{\psi_{i\alpha}-1}$, which results in the Bayes estimator $w_{i\alpha} = \frac{n_{i\alpha} + \psi_{i\alpha}}{n_i + \psi_i}$. The term $\psi_{i\alpha}$ is called a pseudocount because increasing its value by one has the same effect on the posterior as adding one to the frequency count. The weight matrix is trained using the Bayes estimator.

Before moving to scoring using a weight matrix, let's see how the background sequence is modeled. One can treat the background the same way as transcription factor binding sites by thinking of it as a special type of site of unit length. The background is thus

modeled using a motif of unit length. A nucleotide's background probability will then be its weight in this motif.

When scoring a new sequence, each subsequence s in the sliding window of length ℓ has the weight matrix score of $\ln\frac{p(s\mid w)}{p(s\mid b)}=\sum_{i=1}^{\ell}\ln\frac{w_{i\alpha}}{b_{\alpha}}$, where b_{α} is the probability of the nucleotide α in the background motif b. The weight matrix score measures the distinctness between the probabilities that the subsequence is generated from the weight matrix or the background model. s is called a site if the score is above an arbitrary threshold.

In practice, $\ln\frac{w_{i\alpha}}{b_{\alpha}}$ terms are used in a weight matrix table for two reasons. First, summing them gives the weight matrix score of the subsequence. Second, each of these logarithmic terms corresponds to the binding energy of the nucleotide α at the ith position of the motif, when the occupancy probability distribution is assumed to be Boltzmann [17, 75, 77-84]. Thus, the higher the weight matrix score, the lower the binding energy (using the convention that binding energy decreases with higher affinity).

Weight matrices have often been used to identify binding sites of a particular transcription factor. For example, we identified NF-κB-binding sites in the promoters of B-cell linker (BLNK) and B-cell adaptor for phosphoinositide 3-kinase (BCAP) in the B-cell receptor (BCR) signaling pathway, which uncovered a possible role of NF-κB in the transcriptional repression of these molecules that results in tumor [8]. Weight matrices

can also be used to identify *cis*-regulatory modules (CRMs). A CRM, as we have seen before, is a cluster of one or more binding sites for multiple transcription factors. To determine if a large sequence is a CRM, algorithms such as ModelInspector and CIS-ANALYST are used to identify binding sites in the sequence using all available weight matrices for multiple transcription factors [85-89]. They declare the sequence to be a CRM if the number of sites present in the sequence exceeds an arbitrary threshold. (This threshold should not be confused with the thresholds used for identifying individual sites.)

The major shortcoming of the weight matrix method is that the weight matrix score does not provide a natural threshold to allow classification of sequences into sites. Boltzmann distribution, by its nature, does not have a natural threshold. Weight matrix scores are shown not to corroborate well with experimental binding data [75].

The use of weight matrices and statistical inference leads to a hidden Markov model (HMM), such that the HMM emission and transition probabilities correspond to the weight matrix and the prior probabilities of motifs, respectively [90]. While we will discuss HMMs in detail in the next chapter, let's review the statistical framework here because it applies to many supervised and unsupervised learning algorithms. Given a motif's weight matrix w, its prior probability z and background probabilities b (b is thought of as a one nucleotide long weight matrix of the background, as explained before), the joint probability of new sequence s and configuration c of motifs in the sequence takes the general form

$p(s,c \mid W,Z) = p(s \mid c,W,Z)\,p(c \mid W,Z) = p(s \mid c,W)\,p(c \mid Z)$. Here, the set W contains the weight matrix w and the background probabilities b; the set Z contains the prior probabilities of the motif z and the background $(1-z)$; configuration c may contain one or more motifs at various positions in the sequence or none at all. The latter equality in the above equation follows because the probability of the sequence given a configuration is independent of Z (and hence $p(s \mid c,W,Z) = p(s \mid c,W)$) and the probability of a configuration depends only on Z (and hence $p(c \mid W,Z) = p(c \mid Z)$). The first term on the equation's right hand side is $p(s \mid c,W) = \prod_{j_m}\prod_{i=1}^{\ell} w_{i\alpha}^{j} \prod_{j_b} b_{\alpha}^{j}$, where j_m and j_b are the start positions of the instances of the motif and the background in the configuration respectively, and b_{α}^{j} and $w_{i\alpha}^{j}$ are the probabilities of the nucleotide α at the background and the i th position of the motif starting at the j th position of the sequence, respectively. The second term is $p(c \mid Z) = z^{n_m}(1-z)^{n_b}$, where n_m and n_b are the number of instances of the motif and the background in the configuration. Note that the above formalism can be extended easily for multiple types of motifs, where W contains multiple motif types and the background, and Z contains their prior probabilities.

Let's consider supervised learning, i.e. when known examples of sites are provided. The training stage consists of estimating Z and W. We described W estimation above. Z can be estimated from the equation $p(Z \mid S,W) \propto p(S \mid W,Z)\,p(Z \mid W)$. Because the term $p(Z \mid W)$ does not contain any prior information, Z is estimated by maximizing the

likelihood $p(S \mid W, Z) = \sum_c p(S \mid W, Z, c)\, p(c \mid W, Z) = \sum_c p(S \mid W, Z, c)\, p(c \mid Z)$. This requires summing over all configurations and hence an analytical solution of Z is not possible. Using expectation maximization (EM), $z = \frac{\langle n_m \rangle}{\sum_{m'} \langle n_{m'} \rangle}$, where $\langle n_m \rangle$ is the expected number of instances of the motif and m' are the different motif types including the background. While the HMM program called Stubb [91, 92] estimates Z using EM, another HMM program called Ahab [93] uses the conjugate gradient method to arrive at similar estimates. Both Stubb and Ahab estimate W using the weight matrix training method described above. While scoring, both programs divide a sequence into segments of length L, determine Z for each segment and call a segment a CRM if the likelihood ratio $\frac{p(s \mid W, Z)}{p(s \mid b, z_b)}$ is above a threshold, where z_b and $p(s \mid b, z_b)$ are the prior probability and the likelihood of the background, respectively. Stubb also takes into account the order and correlation of motifs in a sequence.

The major disadvantage of many HMM-based methods is that an arbitrary threshold needs to be chosen while scoring. Furthermore, these methods do not explicitly discuss the physical meaning of a motif's prior probability as the transcription factor's concentration. They also do not train W while training Z. We will discuss all these issues in the next chapter.

In unsupervised learning, W of statistically over-represented motifs needs to be estimated when sites are not known and the training set S consists of long sequences that

may contain sites. The two popular methods are (1) expectation maximization (EM) and (2) Gibbs sampling. EM focuses on $p(W \mid S,Z) \propto p(S \mid W,Z)\,p(W \mid Z)$, where the prior $p(W \mid Z) = p(W)$ is the pseudocounts. Similar to the EM estimate of Z above, EM estimate of W is $w_{i\alpha} = \frac{\langle n_{i\alpha;m} \rangle}{\langle n_m \rangle}$, where $\langle n_{i\alpha;m} \rangle$ is the expected number of motif instances containing the nucleotide α at the ith position and $\langle n_m \rangle$ is the expected number of motif instances. Algorithms such as MEME [94], MAST [95] and MDscan [96] use EM. The Gibbs sampling procedure determines the best configuration first before estimating W by focusing on $p(c \mid S,Z) \propto p(S \mid Z,c)\,p(c \mid Z)$, where $p(S \mid Z,c) = \int_W p(S \mid W,Z,c)\,p(W \mid Z,c)\,dW = \int_W p(S \mid W,Z,c)\,p(W)\,dW$. The posterior $p(c \mid S,Z)$ is a complicated equation in terms of beta functions and polynomials and thus cannot be solved analytically. However, a Gibbs sampler samples from it to determine the configuration c that maximizes the posterior. Gibbs sampling procedures include the original algorithms [97-99] as well as extensions such as AlignACE [100], Motif Sampler [101], BioProspector [102] and Gibbs Recursive Sampler [103]. Unlike the above methods, CONSENSUS identifies sites that have weight matrices with the best p-values [104]. Furthermore, unsupervised site identification methods that enumerate all possible sequences of a particular length based on regular expressions or word "dictionaries" have also been developed [105-109].

Udalova *et al.* used regression instead of classification for identification of κB sites. They developed a principal coordinate model to specifically determine relative binding

affinities of κB sites using experimental quantitative binding data [67]. They selected a subset of the 256 possible variants of the fully palindromic NF-κB binding consensus sequence GGRRNNYYCC such that no variant differed from the selected sequences or their reverse complements by more than one nucleotide. They mapped these sequences to a Euclidean space using metric scaling by defining the distance between two sequences as the number of positions with different nucleotides. They then used the largest principal components in the mapped space as features for least-squared linear regression of the logarithm of binding affinity in a gel shift assay. This model automatically incorporated effects of interactions between base pair positions in the binding motif, its predictions were highly correlated with experimental binding data, and it identified motif positions responsible for differential binding of homodimeric p50 versus p50p65 based upon their gel shifts. Moreover, its results regarding differential binding between homodimeric p50 and p52 were consistent with crystallographic studies [66]. The authors subsequently devised an algorithm to optimize selection of sequences for experimental testing, and used microarrays for high-throughput quantitative binding assays [65]. The model's disadvantages are that (1) it requires experimental quantitative binding data of all selected sequences and (2) includes variants of only the consensus sequence. Because several known κB sites do not fit the consensus sequence [54], inclusion of all possible 10-mer variants for this model will require binding experiments with a large number of sequences, making this model infeasible.

2.3 Threshold, Occupancy Probability and a Biophysical Model

Threshold determination is a major roadblock in identifying sites. In a typical two-class classification problem, the training set consists of examples of both classes, and many standard algorithms exist to determine the "best" threshold in some sense. When identifying sites, however, one usually does not know experimentally validated non-binding sequences, and hence the training set consists of examples of only one class, requiring new techniques to determine the threshold. A good threshold should have a palatable biophysical interpretation and should offer insight into the biological function of sequences classified as sites. Most site identification methods proclaim that a sequence is a site if its score is above an arbitrary threshold or statistically significant compared to the score of a background sequence. These criteria do not have a biophysical interpretation and do not offer any information about the importance of a site in modulating gene expression. Moreover, an arbitrary stringent threshold misclassifies true sites whereas a lenient threshold classifies random sequences as sites.

A good threshold with a biophysical interpretation can be determined in the following ways. We will see in the next chapter that an HMM trained on all binding sites and non-binding sequences in a genome, which the HMMs described above are not, learns the true threshold in terms of the transcription factor's concentration as its transition probability to the motif. Another way of finding a good threshold is to use a discriminant function whose physical interpretation offers a natural threshold.

Occupancy probability makes an excellent discriminant function for site identification due to the following three reasons. First, unlike a purely statistical entity, it has a straightforward biophysical interpretation. A transcription factor's occupancy on the promoter determines gene expression. Second, highly occupied sites may be physiologically more significant. Thus, occupancy probability not only helps in classifying sequences as sites but also offers insight into their influence on gene expression. We will see in Chapter 5 that evolutionary fitness can be thought of as a linear function of occupancy probability. Third, occupancy probability has a natural threshold at 0.5 as we will see below.

Occupancy probability of a sequence by a particular transcription factor depends upon (1) its sequence and (2) the transcription factor's concentration. At a particular transcription factor concentration, only sequences with favorable bonds and low binding energy qualify as sites. However, as the transcription factor's concentration increases, it also binds to sequences forming less favorable bonds and higher binding energy, the equilibrium shifts toward more bound product and the threshold shifts to assign more sequences as sites.

Let's derive an equation for the occupancy probability of a sequence and study its behavior as a function of binding energy (which depends upon the sequence) and the transcription factor concentration [71-74]. When a transcription factor protein P binds to DNA D to form a complex DP ($P + D \rightleftharpoons DP$), the dissociation constant is

$K_d = e^{\frac{\Delta G^\circ}{K_b T}} = e^{\beta E(s)} = \frac{[P][D]}{[DP]}$, where K_b is the Boltzmann constant, T is the absolute temperature, $\beta = \frac{1}{K_b T}$, $\Delta G^\circ = E(s)$ is the binding free energy of DNA sequence s under standard conditions, and $[P]$, $[D]$ and $[DP]$ are the equilibrium concentrations of the free protein, free DNA and protein-DNA complex, respectively. The occupancy probability of sequence s is then

$$p^{bound}(s) = \frac{[DP]}{[DP]+[D]} = \frac{1}{1+[D]/[DP]} = \frac{1}{1+K_d/[P]} = \frac{1}{1+e^{\beta E(s)}/[P]} = \frac{1}{1+e^{\beta(E(s)-\mu)}},$$

where $\mu = \frac{\ln[P]}{\beta}$ is the chemical potential and K_d is the dissociation constant. Thus, occupancy probability has the well-known Fermi-Dirac distribution (Figure 2.2). It has the following features:

- When the binding energy $E(s)$ of a sequence is very low or the transcription factor concentration $[P]$ (and hence μ) is very high, $\beta(E(s)-\mu) \to -\infty$, $e^{\beta(E(s)-\mu)} \approx 0$ and hence the sequence is always occupied as seen in the top left part of the figure.
- When the transcription factor concentration $[P]$ (and hence μ) is very low or the binding energy $E(s)$ of a sequence is very high, $e^{\beta(E(s)-\mu)} \gg 1$ and hence the occupancy probability has an approximate Boltzmann distribution $p^{bound}(s) \approx e^{-\beta(E(s)-\mu)}$, as seen in the bottom right part of the figure. Logarithms of the weight matrix terms correspond to binding energies when occupancy

probability has a Boltzmann distribution [17, 75, 77-84], i.e. when the transcription factor concentration is very low or binding energy is very high.

- Occupancy probability of 0.5 can be thought of as a natural threshold in classifying sequences as sites. It corresponds to binding energy equal to the chemical potential μ. Sequences with binding energy less than the chemical potential ($E(s) < \mu$) have occupancy probability greater than 0.5 and hence should be classified as sites.

The biophysical model QPMEME (Quadratic Programming Method of Energy Matrix Estimation) estimates binding energies based upon occupancy probability [71, 73, 74]. Assuming that the binding energy at each position of a sequence is independent of other positions, the binding energy $E(s)$ of a sequence is the sum of the binding energies at individual positions. Binding energies of all possible sequences of length ℓ are approximately normally distributed when $\ell \gg 1$ because binding energy at each position is a random variable and $E(s)$ is the sum of these ℓ random variables. When this normal distribution of binding energies and the Fermi-Dirac distribution of occupancy probability are compared on the same scale, the Fermi-Dirac distribution can be approximated by the step function for $\ell \gg 1$ because the standard deviation of the binding energy distribution is much greater that $K_b T$ [110-113]. Thus, a true site has binding energy $E(s) < \mu$ and hence is occupied, while a random sequence with $E(s) > \mu$ is not occupied (the threshold binding energy μ is at the far left of the mean of the normal distribution). The likelihood of the training sequences consisting of only true

sites is maximized by minimizing the probability of a random sequence with binding energy below μ while ensuring that all known sites have binding energy below μ. The probability of a random sequence with energy below μ is minimized by minimizing the variance of the normal distribution. Let $\boldsymbol{\varepsilon}$ be the binding energy vector of length 4ℓ containing binding energy of each nucleotide at each position ($\varepsilon_{i\alpha}$), and $\mathbf{S}$ be the sequence vector of length 4ℓ such that each element $s_{i\alpha}$ equals one if the sequence has the nucleotide α at the ith position and zero otherwise. Then the sequence's binding energy is $E(s) = \boldsymbol{\varepsilon} \cdot \mathbf{S} = \sum_{i=1}^{\ell} \sum_{\alpha=1}^{4} \varepsilon_{i\alpha} s_{i\alpha}$. If we set mean binding energy $E(s)$ to zero and $\mu = -1$, binding energy is expressed in terms of μ. Let's assume that the probability of a nucleotide (p_α) is the same at all positions of a random sequence. The above problem is a quadratic programming problem (quadratic objective function with linear constraints; explained in Chapter 5) in 4ℓ $\varepsilon_{i\alpha}$ variables with Lagrangian $\frac{1}{2}\sum_{i=1}^{\ell}\sum_{\alpha=1}^{4} p_\alpha (\varepsilon_{i\alpha})^2 + \sum_a \lambda_a \left(\sum_{i=1}^{\ell}\sum_{\alpha=1}^{4} \varepsilon_{i\alpha} s_{i\alpha}^a + 1 \right) - \sum_{i=1}^{\ell} \nu_i \left(\sum_{\alpha=1}^{4} p_\alpha \varepsilon_{i\alpha} \right)$, where i is any position in the sequence. The first constraint is that the binding energy of each training sequence is less than μ: $E(s^a) = \sum_{i=1}^{\ell}\sum_{\alpha=1}^{4} \varepsilon_{i\alpha} s_{i\alpha}^a \le -1$. The second constraint is that at each position i, the average binding energy is set to 0: $\sum_{\alpha=1}^{4} p_\alpha \varepsilon_{i\alpha} = 0$. λ_a and ν_i are Lagrange multipliers and $\lambda_a \ge 0$. QPMEME estimates the binding energy of each nucleotide at each position of the motif by solving the above problem.

2.4 Limitations of Conventional Methods and Addition of Heterogeneous Data

The conventional computational methods for site identification described above have several limitations [17]. (1) They identify several false positives because sites are short (5-15 bases) and degenerate (the nucleotides at each position are not unique). (2) They assume that adjacent positions in a site are independent. (3) In higher eukaryotes, sites are present upstream or downstream of genes, in introns or even in regions far away from genes. (4) Genomic backgrounds around sites can differ widely. (5) Moreover, they do not explicitly identify overlapping motifs such as for NF-κB. As mentioned in the last section, most methods other than QPMEME (6) use an arbitrary threshold lacking a straightforward biophysical interpretation for classifying sequences into sites and (7) fail to estimate occupancy probability that helps elucidate a site's role in gene expression modulation.

Furthermore, these primary sequence-based methods have limitations in predicting functioning sites *in vivo*. A predicted site may be inaccessible to a transcription factor *in vivo* due to DNA methylation or condensation of chromatin structure. Moreover, a transcription factor may not be able to bind to a site if another transcription factor has occupied a sequence overlapping the site. Even if a transcription factor binds to a site, it may be unable to initiate transcription due the unavailability of cofactors.

Due to the above limitations of classical methods, new methods have begun incorporating various types of data (heterogeneous data) to identify sites more accurately. The following heterogeneous data have been included: (1) Conservation of sites in orthologs in related species (see Chapter 4 for more information), (2) Experimental identification in high-throughput binding experiments (e.g. ChiP-chip) [114], (3) Occurrence of sites in the promoters of genes with a similar expression pattern (e.g. microarray gene expression data) [102, 115], (4) Occurrence of sites in the promoters of genes with a similar function (e.g. Gene Ontology categories), (5) Presence of multiple sites in a promoter [116], (6) Occurence of a site within a cluster of binding sites for different transcription factors [116], (7) Network-level conservation, i.e. over-representation of sites in the promoters of genes present in the transcription factor's network in closely related species [117] .

A composite model is required to take full advantage of heterogeneous data. Some methods analyze each type of data separately to identify sites that have a score above a threshold for each type of data, and declare sites that have scores higher than thresholds for all data types as the true sites [116] . They fail to identify sites that have scores lower than the individual thresholds, but have a good overall composite score (figure 2.3). Composite models that incorporate sequence conservation form the focus of Chapters 4 and 5. Bussemaker *et al.* have built a composite model of sequence composition and high-throughput expression data that fits the logarithm of the expression ratio to the sum of activating and inhibitory contributions of motifs and thus finds statistically significant motifs [118].

We have developed two novel methods that identify sites by addressing the threshold issue and estimating occupancy probability. In Chapter 3, we describe the HMM that identifies overlapping sites and offers many more improvements to the existing HMMs. In Chapter 5, we describe a method that combines the QPMEME biophysical model with sequence conservation data.

Figure 2.1: Schematic of the weight matrix method to identify transcription factor binding sites.

WM is trained using the known sites and assuming that the training sites as well as the positions in each site are independent. Each element of the WM $w_{i\alpha}$ is estimated using the Bayes estimator as shown, where $n_{i\alpha}$ is the frequency of the nucleotide α at the i th position of the training sites, n_i is the number of training sites for which the nucleotide at the i th position is known, and the ψ terms are the pseudocounts. When scoring a new sequence, weight matrix score of each subsequence in a sliding window of length ℓ equal to that of the site is calculated, where b_α is the background probability of the nucleotide α. A subsequence is declared a hit if its weight matrix score is above a threshold (10 in this figure).

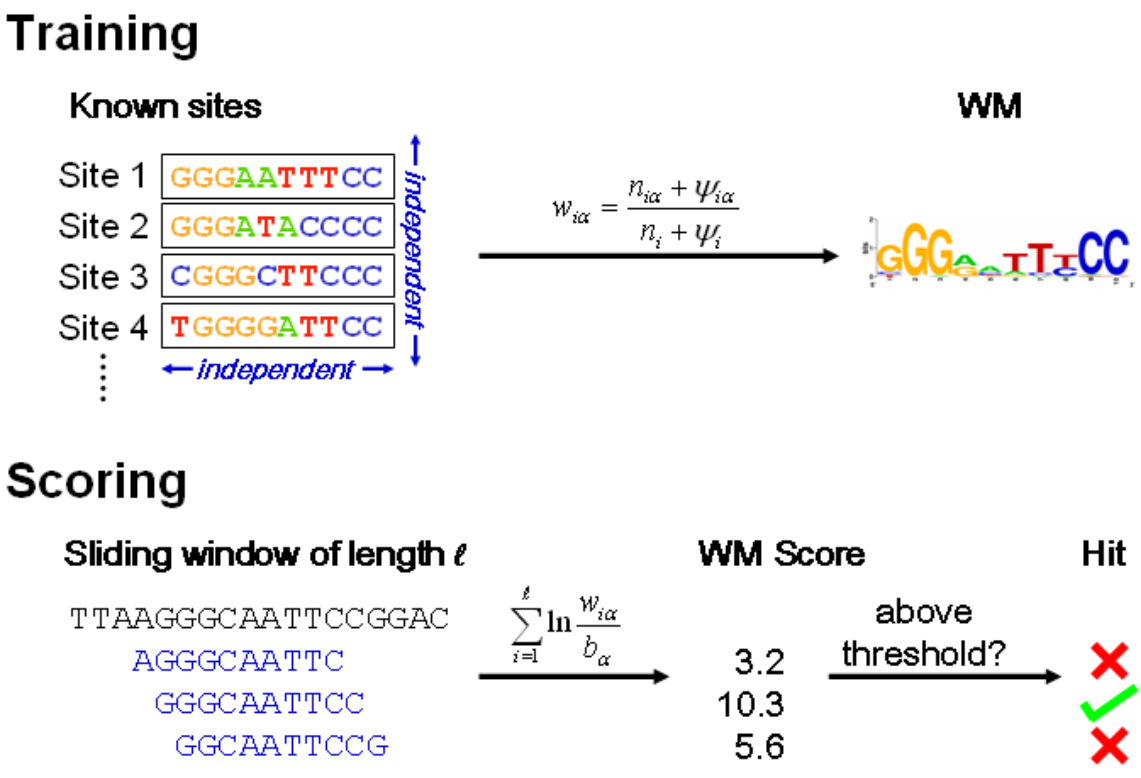


Figure 2.2: Occupancy probability as a function of binding energy of a sequence. Occupancy probability has an overall Fermi-Dirac distribution. A sequence is always occupied by the transcription factor protein if its binding energy is very low or the protein concentration (and hence the chemical potential) is very high. If the protein concentration (and hence the chemical potential) is very low or binding energy is very high, occupancy probability has an approximate Boltzmann distribution. In classifying sequences as sites, the natural threshold is the occupancy probability of 0.5, which corresponds to binding energy equal to the chemical potential.

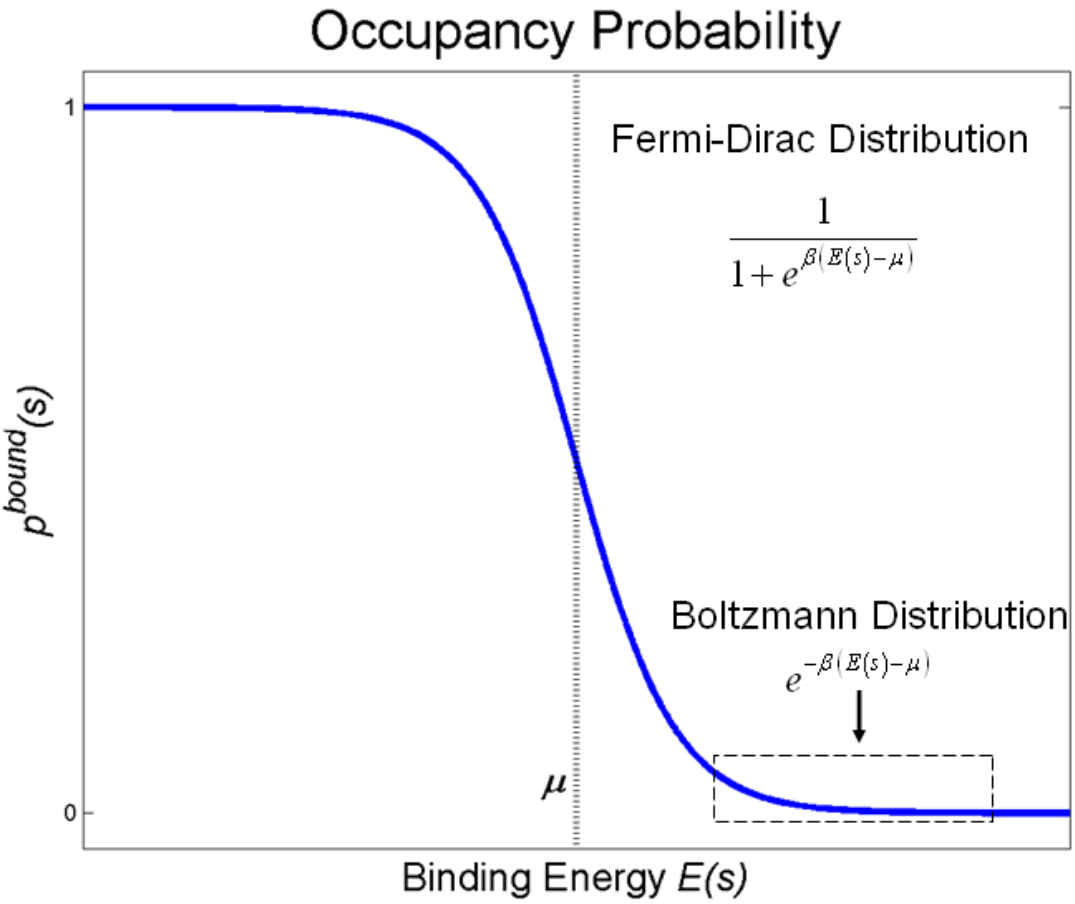
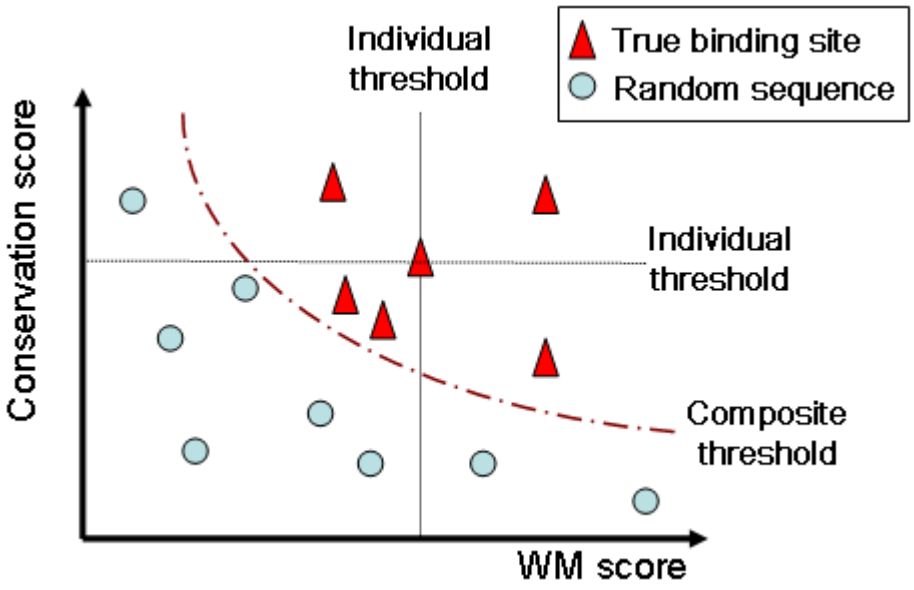


Figure 2.3: Illustration showing the need for a composite model when analyzing heterogeneous data.

Red triangles are true sites, whereas gray circles are random sequences. The x coordinates are weight matrix (WM) scores and the y coordinates are conservation scores, i.e. how well a sequence is conserved in orthologs. Dotted lines show the thresholds for each type of data such that a score higher than the threshold in each dimension corresponds to a putative hit. Dashed brown arc shows the threshold used in a composite model. Methods that declare only those sequences with the weight matrix and the conservation scores above the corresponding thresholds as sites fail to identify true sites with individual scores lower than the thresholds but a good composite score.



Chapter 3

Hidden Markov Model to Identify Self-Overlapping Sites

"All truths are easy to understand once they are discovered; the point is to discover them."
Galileo Galilei (1564-1642)

3.1 Markov Model and Hidden Markov Model

We briefly describe discrete Markov models before proceeding to discrete hidden Markov models, which form the focus of this chapter [69, 119, 120] (a continuous Markov model requires a slightly different treatment). Consider a system which changes from one state to another using a stochastic process as it moves forward in space or time. If the process is a homogeneous Markov process, the state at a particular instance (1) depends only upon the previous state and (2) is independent of states prior to the previous state. Such a system is called a Markov chain and the associated probabilistic model is called a Markov model. In a Markov model, the probability of transition from one state (i) to another (j) in unit interval is called a transition probability (a_{ij}), and the probability of a state at the first instance is called its initial probability (π_j for state j). Thus, the probability of state j at any instance t is given by $p_t(j) = \sum_i p_{t-1}(i) a_{ij}$. (For the special case of $t = 1$, $p_t(j) = \pi_j$.)

As an example, one can think of a Markov model as generating a string of DNA sequence as it moves 5' to 3'. If the four types of nucleotides are considered as the four states, the Markov model can generate a sequence based on its transition and initial probabilities (Figure 3.1A). A Markov model can generate many different nucleotide sequences with different probabilities. Each sequence can be thought of as a state path, i.e. a path in which particular states are present at particular positions. An observed sequence is one state path, and one can easily calculate its likelihood in terms of initial and transition probabilities given a Markov model. While this example describes change in a sequence in space (5' to 3'), we will see another example of Markov models in the next chapter – an evolutionary model that describes change in a sequence with time as a species evolves.

In the above "observable" Markov model, nucleotides are considered as states (such that a nucleotide at a position depends only on the nucleotide at its 5' position) and thus states are observable in the data. However, what if a DNA sequence has a hidden underlying structure that actually determines the chain of nucleotides? For example, a promoter sequence if often made of motifs corresponding binding sites of transcription factors and the background. One can model such a sequence by considering the motif and the background as the two states (the simplest case; Figure 3.1B). Nucleotides in such a model are not states themselves but merely the observed symbols emitted by these states. The problem is that the states – background and motif – cannot be observed. Only the nucleotides emitted by these states can be observed. Of course, the two hidden states have different emission probabilities according to which they emit these nucleotides (otherwise,

there is no need of considering two separate hidden states). Such a Markov model is called a hidden Markov model (HMM). Its hidden states emit symbols that one can observe using a stochastic process and the associated probabilities are called emission probabilities. Incidentally, just like the states in an "observable" Markov model, states in an HMM change from one to another based on their transition probabilities. Thus, an HMM has two sources of randomness that correspond to transition and emission probabilities.

Such an HMM can also be thought of as a sequence generating model [90]. It generates a sequence from 5' to 3' as follows. At any position in the sequence, the HMM (1) determines the probabilities of the motif and the background states at the previous position, (2) calculates the probability of either state at the current position using the transition probabilities and (3) generates the new nucleotide based upon the states' emission probabilities. An HMM can generate many hidden state paths each with a specific probability, and each hidden state path can generate the observed sequence with a specific probability.

An HMM is trained using a dynamic programming expectation-maximization (EM) algorithm called the Baum-Welch algorithm. An HMM consists of five elements: (1) number of states, (2) symbols (nucleotides in the above example) emitted by states, (3) initial probabilities, (4) transition probabilities and (5) emission probabilities. The first two elements are hyper-parameters and are determined before training. Given a training

set of observation sequences, the Baum-Welch algorithm iteratively learns the other three elements as model parameters by maximizing the likelihood of generating the training set.

In this chapter, we pay extra attention to one of the intermediate variables, called the gamma (γ) variable, computed during the training procedure. This variable corresponds to the probability of each state at each position. It is actually the normalized product of the forward (α) and backward (β) variables for each state at that position. That is, the probability of state i at position t is given by $\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{n}\alpha_t(j)\beta_t(j)}$, where n is the number of states and j denotes each of these states during normalization. The forward and backward variables are in turn calculated using the standard HMM recursion relations. The probability of observing the partial sequence $O_1 \cdots O_t$ until position t and being in state j at position t is given by the recursion relation $\alpha_t(j) = \left(\sum_{i=1}^{n}\alpha_{t-1}(i)a_{ij}\right)b_j(O_t)$, where a_{ij} is the transition probability from state i to state j, and $b_j(O_t)$ is the emission probability of state j that generates the nucleotide at position t. (It is initialized as $\alpha_1(j) = \pi_j b_j(O_1)$, where π_j is the initial probability of state j.) Similarly, the probability of being in state i at position t and observing the partial sequence $O_{t+1} \cdots O_T$ until the end position T is given by the recursion relation $\beta_t(i) = \sum_{j=1}^{n} a_{ij}\beta_{t+1}(j)b_j(O_{t+1})$. (It is initialized as $\beta_T(i) = 1$.) The gamma variable is

important because it gives a transcription factor's occupancy probability when the HMM is used as a physical binding model for site identification (see below).

There are two popular ways of scoring a new sequence using an HMM. We can calculate its likelihood using a dynamic programming algorithm that takes into account all hidden state paths. We can also calculate the state path that has the greatest probability of generating the new sequences using a similar dynamic programming algorithm called the Viterbi algorithm.

As indicated in the previous chapter, emission and transition probabilities of HMMs used in site identification generally correspond to the weight matrix and the prior probabilities of motifs, respectively. If a motif is assumed to have the same prior probability z irrespective of the state in the previous position, transition probability to the motif state from any state is the same. While this simplifying assumption does not require a model as complex as an HMM, the HMM framework is often used to take advantage of its standard training and scoring procedures.

3.2 HMM as a Physical Binding Model for Site Identification

Even though in the context of site identification an HMM is usually interpreted as a sequence generative model, we focus on its somewhat obscure interpretation as a physical binding model of a transcription factor on DNA. This interpretation leads us to transform the statistical HMM model into a biophysical one. We can then determine the

occupancy probability of a transcription factor on a DNA sequence, and think of the prior probability of the motif (transition probability to the motif z) as a measure of transcription factor concentration and the weight matrix (motif emission probabilities) as a measure of binding energies. More importantly, the biophysical model offers a principled threshold for classifying sequences into sites.

We discussed in the last section that an HMM is commonly used to generate spatial or time series sequences in the machine learning field. This pedestrian approach to an HMM, however, misses the role played by the occupancy of a transcription factor. Therefore, we wish to emphasize the interpretation of an HMM as a physical binding model that estimates the occupancy probability of a transcription factor on a particular position of a DNA sequence, i.e. how often the transcription factor is bound to that position of the DNA sequence. When an HMM has a background and a motif state, the occupancy probability at a position is the probability of the motif state at that position. It is given by the gamma (γ) variable of the motif state at that position. As the sum of the probabilities of the two states at a position is one, this allows us to conclude that the probability of the background state is the probability of unbound DNA at that position.

The above two interpretations of an HMM, it must be noted, are not really that different if we bear in mind that the generative probability of the motif state is in fact the occupancy probability of the transcription factor.

Although the occupancy probability of a transcription factor at a particular position in a sequence can easily be calculated using the standard HMM techniques, we still review the method of calculating it from first principles [90]. This is done primarily for two reasons. First, the comparison of these two methods will show how the HMM is computationally a great deal easier to use, particularly since (i) the factors determining occupancy probability are actually trained and (ii) occupancy probability is calculated as an intermediate variable using HMM techniques. Secondly, it will reveal the underlying physical connection between the occupancy probability obtained from the basic thermodynamics principles and the statistical quantities associated with an HMM. For the purpose of demonstration, we only consider the case of non-self-overlapping sites and assume that binding is allowed only on one strand. (A more general case is discussed in Appendix A.)

In calculating occupancy probability using first principles [90], let's denote the motif state, representing the entire site, as m and the background state as b. The emission probability of the motif state corresponds to its weight matrix, whereas the transition probability to the motif state from either state is the prior probability of the motif z. Let $p(b)$ be the probability that a long sequence s does not contain any motifs (i.e. it is all background), and $p_j(m)$ be the probability that the sequence has one motif m starting at the j th position. This latter probability can be written as

$p_j(m) = \ldots(1-z).w_\alpha^{b(j-1)}.z.\prod_{i=1}^{\ell} w_{i\alpha}^{mj}.(1-z).w_\alpha^{b(j+\ell+1)}\ldots,$ where the transition probability to the motif state at any particular position in a sequence is small ($z \approx 0$), ℓ is the length of

the motif, $w_{i\alpha}^{mj}$ is the probability that the nucleotide α at the $(j+i-1)$th position of the sequence is emitted by the ith position of the motif state, and $w_{\alpha}^{b(j-1)}$ is the probability that the nucleotide α at the $(j-1)$th position of the sequence is emitted by the background state. Note that this formulation fits into the general Bayesian probabilistic framework described in Section 2.2 such that the product of the w terms is the likelihood and the product of the z terms is the prior. We can write

$p_j(m) = p(b).\frac{z}{(1-z)^{\ell}}\prod_{i=1}^{\ell}\frac{w_{i\alpha}^{mj}}{w_{i\alpha}^{bj}} \approx p(b).z.E_j$, where $(1-z)^{\ell} \approx 1$ and $w_{i\alpha}^{bj}$ is the probability that the nucleotide α at the $(j+i-1)$th position of the sequence is emitted by the background state. Also, $E_j = e^{W_j}$ such that $W_j = \ln\left(\prod_{i=1}^{\ell}\frac{w_{i\alpha}^{mj}}{w_{i\alpha}^{bj}}\right)$ is the weight matrix score of the motif starting at the jth position of the sequence. Then, the occupancy probability at the jth position of the sequence is $p_j^{bound}(s) = \frac{p_j(m)}{p(b)+p_j(m)} \approx \frac{z.E_j}{1+z.E_j} = \frac{z.e^{W_j}}{1+z.e^{W_j}}$.

Thus, the two factors determining occupancy probability are (i) the transition probability to the motif z and (ii) the measure of distinctness of the emission probabilities of the motif from that of the background (weight matrix). Both these factors need to be high for the transcription factor to be bound to a particular position in a DNA sequence with high probability. For example, even if the weight matrix score is high, occupancy probability cannot reach one if z is really small. A site identification method based only on a weight matrix has no way of dealing with this interplay with z.

The HMM training techniques offer two advantages over the calculations made using first principles. First, the HMM Baum-Welch procedure trains the transition and emission probabilities. This produces optimized values of z and the weight matrix, which are essential for an accurate estimation of occupancy probability. Secondly, as mentioned above, the HMM training procedure also calculates an intermediate variable called the gamma (γ) variable. The occupancy probability at a particular position is given simply by the gamma variable of the motif state at that position: $p_j(s) = \gamma_j^m = 1 - \gamma_j^b$, where γ_j^m and γ_j^b are the gamma values of the motif and background states at that position, respectively. Thus, the calculation of occupancy probability in a new sequence requires a simple extension of the scoring procedure in which the gamma variable is computed just like during the training procedure.

We now transform the statistical HMM model into a biophysical model. Based on the thermodynamics principles described in Section 2.3, occupancy probability can be written as $p_j^{bound}(s) = \frac{[P]e^{-\beta E_j(s)}}{1+[P]e^{-\beta E_j(s)}}$, where $[P]$ is the concentration of a free transcription factor at equilibrium, $E_j(s)$ is the binding energy at position j, $\beta = \frac{1}{K_b T}$ where K_b is the Boltzmann's constant, and T is the absolute temperature [73]. A comparison of this equation with the equation of occupancy probability obtained from first principles gives us $W_j = -\beta E_j(s)$ and $z = [P]$. Thus, the weight matrix represents binding energy [17]. In addition, the transition probability to the motif z corresponds to the free transcription factor concentration. As the transcription factor concentration

increases, the transition probability to the motif increases and we expect higher occupancy by the transcription factor on the DNA.

The greatest benefit of this transformation of an HMM to a biophysical model is that it enables the HMM to learn the threshold for classification of sequences into sites in a principled way. We have seen in the last chapter that occupancy probability offers a natural threshold at 0.5 when used as a discriminant function. We just saw that the occupancy probability depends upon the transition probability to the motif z and the weight matrix, both of which are trained by an HMM. Thus, when an HMM uses occupancy probability as a discriminant function, it learns the natural threshold based on the training sequences. Because of this accurate estimation of the threshold, an HMM is expected to identify weak sites much more accurately with a fewer false positives. There is, however, one caveat. The method requires that the training sets are chosen carefully so that the emission probabilities reflect the weight matrix of the motif, and that the transition probability to the motif z represents the density of sites in the promoters of all genes in a genome. This is by no means an easy task. We will discuss the training of such an HMM in a subsequent section.

While the use of an HMM is thus advantageous in identifying sites, we have discovered that an HMM offers tremendous benefits for the special case of identifying self-overlapping sites.

3.3 HMM Advantage in Identifying Self-Overlapping Sites

We mentioned in Chapter 1 that binding sites of many transcription factors self-overlap, the effects of which are quite complicated. We know of no method that explicitly takes these effects into account for identifying self-overlapping sites. We will show here how a simple extension of an HMM described in the previous section can be used simply and appropriately for this purpose.

When sites self-overlap, the occupancy probability at a position in a DNA sequence depends upon the strength of all sites containing that position. Thus, the effects of binding in all sequence windows containing the position need to be integrated. Moreover, the binding strength of the reverse strand sequence needs to be taken into account by incorporating an additional motif type whose emission probabilities are flipped from 5' to 3' with respect to the original motif type. As described in detail in Appendix A, the resulting occupancy probability at the j th position based on first principles has a rather complicated formula of $p_j^{bound}(s) \approx \frac{\sum_{k=j-\ell_m+1}^{j} \sum_m z_m . E_{mk}}{1+\sum_{k=j-\ell_m+1}^{j} \sum_m z_m . E_{mk}}$, where k corresponds to the first position of each sequence window containing position j, m is the motif type (including the one corresponding to the site on the other strand), ℓ_m is the length the m th motif type, and z_m is the transition probability to the m th motif type. While many site identification methods do not estimate the transition probability to the motif z, the above equation shows that calculation of the occupancy probability at a position using these methods is

tedious even when *z* is known. The calculation of the overall occupancy probability over the entire sequence is even harder because a simple equation using the *z* and the weight matrix does not exist.

The use of a gamma variable in the HMM method that we have developed enables us to compute the occupation probability for multiple types of self-overlapping sites. This is a distinct advantage over the two HMM scoring methods in vogue at the present time, viz. the likelihood method and the Viterbi method. Both fail to calculate occupancy probability at a particular position in a sequence and identify self-overlapping sites. The likelihood of a sequence is greater when more sites are present in the sequence. Still, its exact correspondence with the occupancy probability of a transcription factor is obscure. In addition, the likelihood score does not indicate the location of sites in a promoter. In contrast, the Viterbi method proclaims the presence of sites at positions where the state path with the highest probability contains the motif state. It, however, fails to consider self-overlapping sites because self-overlapping sites cannot be present in the same state path. It also fails to estimate occupancy probability.

The trick in using HMM gamma variables is to break up each motif type into ℓ states each corresponding to one position in the motif. The emission probabilities of each state are the weights corresponding to that position of the weight matrix. When the site contains no insertions or deletions, the transition probabilities between the states associated with the consecutive positions within the motif are equal to one. Such an HMM calculates the occupancy probability at each position by combining the strengths

of all overlapping motifs (in different sequence windows) in a natural way, even when multiple types of motifs exist. The occupancy probability of any transcription factor at a position is simply the sum of the γ's of all motif states corresponding to the transcription factor (including those corresponding to binding on the opposite DNA strand) at that position. Thus, its occupancy probability at the j th position is $p_j^{bound}(s) = \sum_{m \in M} \sum_{i=1}^{\ell_m} \gamma_j^{m_i}$, where M is the set of motif types corresponding to the transcription factor, and $\gamma_j^{m_i}$ is the γ of the state corresponding to the i th position in motif m at the j th position of the sequence. When the binding of only one transcription factor is considered, the formula simplifies further to $p_j^{bound}(s) = 1 - \gamma_j^b$, as in the case of non-self-overlapping sites, where γ_j^b is the γ of the background state at that position. We can also easily calculate transcription factor occupancy over the entire sequence for non-overlapping or overlapping sites using an HMM as $p^{bound}(s) = \sum_{j=1}^{L} \sum_{m} \gamma_j^{m_1}$, where $\gamma_j^{m_1}$ is the γ of the state corresponding to the first position of the m th motif type at the j th position of the sequence.

Even though HMMs serve well for identifying sites, we will see below the difficulties involved in using existing HMMs for identifying self-overlapping sites, and a need to develop a new HMM to address these difficulties.

3.4 Need for a New HMM for Identifying Self-Overlapping Sites

An HMM can be considered a good method for classifying sequences into self-overlapping sites only if it can be made to learn the classification threshold in a principled manner. Its transition probabilities should therefore be trained to reflect the density of sites in promoters in an entire genome. Its emission probabilities should be trained using known sites in their native promoters to capture the relationship between emission and transition probabilities properly. Moreover, it should take into account the alternative binding modes of self-overlapping sites.

HMMs, now popular for identification of sites for more than a decade [121], have been used in two different ways.

(1) For identification of one or more occurrences of non-overlapping sites: 'Profile HMMs' are generally used for this purpose. Profile HMMs were originally designed to model protein domains such as a kinase domain or a serine protease domain [122-124]. More recently, they have been used to identify binding sites of transcription factors, for example, of cAMP receptor protein in cyanobacterium *Anabaena* [125], liver X receptor [126] and CREB [127]. A profile HMM library was built using TRANSFAC sequences to classify transcription factors [128]. In a profile HMM, each position within a motif has three states. A match state is associated with a nucleotide being present at that position and has corresponding emission probabilities. A deletion state corresponds to absence of any nucleotide at that position. An insertion state allows for insertion of nucleotides

between the current position and the next position within the motif, and has its own emission probabilities. A profile HMM becomes much simpler when the motif is known not to contain any insertions or deletions. Such a profile HMM does not contain insertion or deletion states, and the transition probabilities between the match states of successive positions within the motif are equal to one. The HMM described in the previous section for identifying self-overlapping sites belongs to this category.

(2) For identification of *cis*-regulatory modules (CRMs) that contain multiple sites of different types: This is usually performed using 'motif HMMs' [92, 93, 129-132]. In a motif HMM, the entire motif, i.e. all positions within the motif, is represented by one state. Different states correspond to different motif types (i.e. motifs associated with different transcription factors). The statistical framework of motif HMMs is described in detail in Section 2.2. Phylogenetic conservation has been incorporated in motif HMMs to reduce false positives [92, 133-137].

Both the above HMMs, however, have a number of disadvantages. A profile HMM has a complicated architecture and requires a large number of parameters as a consequence. Because the number of known sites is generally small, training of a profile HMM using known sites in their native promoters (which effectively requires their simultaneous alignment) is usually not possible. Therefore, a profile HMM is generally trained using pre-aligned sites. Because the whole promoters containing the training sites are not used, transition probability to the motif z from background is not trained, and the relationship between transition and emission probabilities is likewise not captured. Motif HMMs, on

the other hand, are more focused on identifying motifs of multiple types. But while they attempt to estimate the transition probability to the motif z, they generally use only the promoters containing the motifs for training z and thus overestimate z. Moreover, they usually train emission probabilities of motifs separately using pre-aligned training sites, thus ignoring the effect of z on emission probabilities.

Many of these HMMs use the likelihood or the Viterbi algorithm for scoring, and thus end up using an arbitrary classification threshold. Furthermore, they leave their relationship with biophysical models rather obscure and thus fail to calculate the occupancy probability. These shortcomings are in addition to their basic failure to explicitly consider the combinatorial effects of self-overlapping sites.

We have therefore developed a new HMM to identify self-overlapping sites based on the theoretical framework described in the previous sections [138]. It uses occupancy probability as the discriminant function. It trains the threshold in a principled manner by training emission probabilities using known sites in their native promoters and training transition probabilities using promoters in an entire genome. The NF-κB family of transcription factors is a prominent example of transcription factors with self-overlapping sites, and we will use identification of its binding sites (κB sites) as a case study.

We will begin by describing the HMM that we have developed and its training procedure.

3.5 Two-Step Training and Behavior of HMM Parameters

Our HMM consists of 21 states: one background state and a state corresponding to each of the ten positions within the κB motif on the two DNA strands (Figure 3.2). Because the κB motif is not known to contain insertions or deletions, the transition probabilities between the states corresponding to successive positions within the motif on a strand are fixed to one. The nine transition probabilities available for training are the transition probabilities from (i) the motif states corresponding to the last position in the motif on both strands and (ii) the background state to (i) the motif states corresponding to the first position in the motif on both strands and (ii) the background state. The rest of the transition probabilities are fixed to zero. The emission probabilities of the motif states on the two strands are flipped from 5' to 3' so as to represent identical binding irrespective of the motif strand. Because initial probabilities are a special case of the transition probabilities at one edge of the sequences, we will not mention them separately from now on.

The transition probabilities were initiated using the transition probability to the motif z chosen by us. The motif emission probabilities (motif profile) were initiated using the 97 κB sites generated in unbiased experiments and obtained from TRANSFAC 9.3 [54, 56-59]. We will refer to this motif profile as the initial motif profile. The promoters were defined as the regions starting at 800 bp upstream of the TSS (transcription start site) and ending at 100 bp downstream of the TSS. The background state's emission probabilities were assigned from the nucleotide distribution of the promoters corresponding to the

reference sequences of all human genes in RefSeq Release 19 [139] associated with human assembly hg18, NCBI Build 36.1 available at the University of California Santa Cruz (UCSC) genome bioinformatics site (http://genome.ucsc.edu/) [140, 141]. The background probabilities were also used as pseudocounts when generating the motif profile initially and during subsequent training.

The HMM parameters can be divided into two sets: (1) the emission probabilities of each motif state (motif profile) and the background, and (2) the transition probabilities that depend upon the transition probability to the motif (z). While the HMM needs to be trained using site-rich sequences to learn the motif profile, training on random sequences (promoters of randomly selected genes) is required to learn z reflecting the site density in the promoters of all genes in the human genome.

We therefore trained the HMM, using the Baum-Welch algorithm [69], in two steps. In the first step, we trained both the emission and transition probabilities using short sequences rich in known sites with the aim of accurately estimating the motif profile. In the second step, we kept the emission probabilities constant, and trained the transition probabilities on promoters containing known sites as well as random promoters to accurately estimate the transition probability to the motif z reflecting the site density in the promoters of all genes in the human genome. We estimated z separately for the TSS-800:TSS (upstream 800 bp) and TSS:TSS+100 (downstream 100 bp) regions because the κB site density is different in these regions. Training an HMM to discover motif locations in unlabeled promoter sequences would generally be regarded as an unsupervised

learning algorithm. We, however, initiated the motif profile based on known sites and trained it using promoters enriched in known sites. Thus, our HMM can be considered as a semi-supervised learning algorithm.

In the first step, we used two types of site-rich sequences of different lengths as well as various initial *z*'s to train all HMM parameters and determine the emission probabilities to be used in the second step. We used the following two types of promoters. (1) The "TSS-n promoters" consist of n nucleotides upstream of the TSS of the 42 human genes known to contain a κB site [54, 61, 62]. (2) The "Surround-n promoters" consist of 34 promoters containing the 36 known κB sites whose exact genomic locations were identified (two promoters each contained two closely located known κB sites) and the surrounding regions. Each surround-n promoter is n nucleotides long. The HMMs trained on these promoters were called "TSS-n HMMs" and "surround-n HMMs," respectively. After each training iteration, the emission probabilities of the motif states of the corresponding motif positions on both strands were averaged to ensure that the learned motif profiles on both strands were exactly flipped 5' to 3'. After training, the sum of the transition probabilities from the background state to the motif states corresponding to the first motif position on both strands was estimated as z (the transition probabilities to the motif states on the two strands are near identical).

Trained motif profiles of TSS-n HMMs appear similar to the background emission probabilities regardless of the promoter length and initial *z*. On the other hand, trained motif profiles of surround-50 or surround-100 HMMs with a reasonable initial *z* (between

0.0001 and 0.01) are distinct from the background (Figure 3.3A). They are also distinct from the initial motif profile, as their symmetrical Kullback-Leibler (KL) divergences (defined as $\sum_{i=1}^{l} \frac{1}{2}(D_{KL}(P_i \,||\, Q_i) + D_{KL}(Q_i \,||\, P_i))$, where P_i and Q_i are the emission probability distributions of the ith motif positions of motif profiles P and Q, ℓ is the motif length and D_{KL} is the log e-based KL divergence) with respect to the initial motif profile are high (0.49 and 0.5, respectively; in comparison, the KL divergences between the initial motif profile and 100 multinomial distributions simulated from the initial motif profile have a normal distribution with mean 0.0015 and standard deviation 0.00038). The trained motif profiles are slightly weaker than the initial motif profile, i.e. more similar to the background. In a surround-50 or surround-100 HMM, any initial *z* between 0.0001 and 0.01 results in the same trained motif profile, indicating that perhaps a local optimum is reached. Trained motif profiles of surround-200 HMMs, however, appear more and more like the background as the initial *z* increases above 0.001. Trained motif profiles of surround-400 HMMs appear similar to the background regardless of the initial *z*. We used the trained motif profile of the surround-50 HMM for further analysis (Figure 3.3A).

The above results show that successful training of the motif profile requires a high density of sites in training promoters and a reasonable transition probability to the motif at the beginning of the training (initial z). Training of the motif profile using TSS-n promoters failed due to the low density of κB sites in these promoters – long promoters contained too few sites as compared to the total number of nucleotides and many short promoters did not contain any κB sites. Training of the motif profile using surround-400

promoters also failed because the density of the κB sites was too low even though each training promoter contained a κB site. Motif profile was successfully trained only in the case of surround-50 and surround-100 promoters, and surround-200 promoters with initial z less than 0.1. High initial z, which presupposes high density of sites, forces the motif profile to appear like the background because many background sequences are characterized as sites during training. As expected, this failure to properly train the motif profile with a high initial z is more pronounced for promoters with lower density of sites – for initial z of 0.1, training of the motif profile of the surround-200 HMM failed while that of the surround-50 HMM did not.

We used unaligned sequences containing known sites rather than pre-aligned known sites to estimate the motif profile. As we have seen before, the motif profile and the transition probability to the motif z corresponding to a training set are competing parameters, i.e. higher value of z corresponds to a weaker motif profile. To avoid arbitrarily strengthening or weakening the motif profile, we used the known sites and their surrounding sequences, and trained the motif profile and z simultaneously to get the best estimates based on expectation maximization. When sequences surrounding the known sites are not used in training, z cannot be trained and thus an arbitrary z is used while training the motif profile.

Unlike most HMMs in the literature, we successfully trained the HMM emission probabilities without requiring pre-alignment of training sites. Because a κB site does not contain any insertions or gaps, we did not need to model the insertion and deletion

probabilities as in profile HMMs. This reduced the number of training parameters substantially, allowing us to train the emission probabilities using unaligned sites.

Before we discuss the second training step, it is instructive to note that trained z is proportional to site density. To see this, we examined the effect of the nature and the length of training promoters on trained z. We used the initial motif profile in this study and kept it fixed during training to isolate the effect on z. When trained on TSS-n promoters, z is inversely proportional to the training promoters' length in the range between 500-3000 bp. Hence, the quantity $z * promoter\ length$ is relatively constant at ~0.9 (Figure 3.4). It drops slightly between 500 to 200 bp and then substantially after 200 bp due to the lack of κB sites in the shorter training promoters. When trained on surround-n promoters, the trained z is again inversely correlated to the training promoters' length and the above quantity is higher at ~1.8 (z = 0.0347, 0.0175 and 0.0087 when n = 50, 100 and 200, respectively). Trained z is a great deal higher when the HMM is trained on surround-n promoters as opposed to TSS-n promoters because all surround-n promoters are guaranteed to contain κB sites. Incidentally, the initial z of TSS-n or surround-n HMMs in the range between 0.0001 and 0.1 does not affect the trained z probably because a global optimum is reached after a few expectation maximization (EM) iterations during training.

Another interesting feature is that trained z is inversely related to the strength of the motif profile. In other words, when the motif profile is kept constant and only z is trained, the weaker the motif profile used (e.g. closer to the background), the higher the

trained z. To see this, we compared the quantity $z * promoter\ length$ after training z on the initial motif profile vs. the motif profiles of surround-n HMMs, which are weaker. Again, the motif profiles were fixed during training. Surround-n promoters were used for training. The quantity $z * promoter\ length$ is ~1.9 when trained on motif profiles of surround-n HMMs (z = 0.0363 and 0.02 when n = 50 and 100, respectively) as opposed to ~1.8 when trained on the initial motif profile. This is probably the result of the compensating mechanism between the motif profile and z discussed above. As one can recall, this competition between z and the motif profile also determines occupancy probability.

The goal of the second training step was to estimate the transition probability to the motif z reflecting the site density in the promoters of all genes in the human genome. This z corresponds to the appropriate threshold when identifying sites in all human promoters. As noted above, the trained z is proportional to the site density in the training sequences. The problem is that we do not know how many sites are present in the promoters of all genes in the human genome. Obviously, z trained in the first step was not appropriate due to the high site density in the site-rich training sequences. We need to train z on all human promoters, which is computationally expensive, or train it on a sufficient number of promoters to get a reasonable estimate of z. Therefore, we began with the human promoters containing the known sites as the training set and progressively added promoters of randomly selected human genes to the training set, training z each time. As expected, the trained z decreased with the addition of random promoters until the training set reached a few thousand promoters and then stabilized (Figure 3.3B). We

estimated this *z* as representing sites in all human promoters. The motif profile of the surround-50 HMM was used during this training. It was not trained in this step because the trained motif profile would have appeared similar to the background thanks to the low site density in the training sets, as discussed above.

We estimated the *z* separately for the upstream 800 bp and downstream 100 bp regions. The estimated *z* for the upstream 800 bp regions was slightly higher (0.00017 vs. 0.00012), reflecting the fact that the proximal promoter (up to 200 bp upstream of the TSS) has a high density of κB sites.

With these trained parameters in hand, we will now turn to our scoring scheme.

3.6 Scoring with Location-Dependent Transition Probabilities

A novel feature of our scoring scheme is the tuning of parameters with the distance from the TSS in accordance with the varying density of sites.

The majority of known κB sites are located just upstream of the TSS in gene promoters and the number of known κB sites decreases further upstream. Specifically, of the 36 known κB sites upstream of the TSS, 16 are located within 100 bp and 28 are located within 200 bp of the TSS. Liu *et al.* have also made a similar observation in the promoters of NF-κB-regulated immune genes [68]. To counter the claim that such an observation for binding sites may be due experimental bias, we cite Tabach *et al.* who

showed in a wide-scale bioinformatic study that functional binding sites are more likely to be present in the 200 bp region upstream of the TSS than any other upstream region for most human transcription factors and specifically for NF-κB [142]. They defined functional binding sites as those over-represented in functionally related genes (in the same Gene Ontology categories) and conserved in related species. To bolster their conclusion, they showed location dependence of binding sites for the transcription factor Mycardin in a controlled experiment. Xie *et al.* also arrived at a similar conclusion based on binding site conservation [143]. Even though the exact reasons for the occurrence of such a phenomenon are not known at the present time, a better interaction of the transcription factor with the transcription machinery if it is bound close to the TSS and the low density of nucleosomes near the TSS [55] heuristically explain why this phenomenon may occur.

While we agree that a great number experiments need to be conducted to definitively prove the location dependence of functional κB sites, we feel that a site identification method must be able to take it into account. Therefore, the commonly used site identification methods are certainly deficient in that they assume equal probability of a site everywhere in the gene structure and thus fail to adjust according to the location within the gene structure.

Noticing that site density decreases sharply with the upstream disease from the TSS and that transition probability to the motif z is proportional to site density, we modeled z using an exponential functional form such that a region close to the TSS had a higher site

density than a region further upstream. We estimated the mean of this exponential functional form as follows. z can be written as $z = \frac{z_0}{\theta} e^{-x/\theta}$ at x positions upstream of the TSS, where θ is the mean distance of κB sites upstream of the TSS (in number of nucleotide positions) and z_0 is the scale factor. Based on the position of known upstream κB sites, the maximum likelihood estimate of the mean for the exponential form was 170, and the estimate of the mean using the median was quite close at $\theta = \frac{median}{\ln 2} \approx 169$. The maximum likelihood estimate was used in further analysis. To determine the scale factor z_0, we noted that the site density per promoter is $800 * z$ if z has a uniform functional form and $\frac{z_0}{\theta} \int_0^{800} e^{-x/\theta} dx$ if z has an exponential functional form. Equating these two expressions for the uniform $z = 0.00017$ obtained from training and $\theta = 170$ results in $z_0 = 0.137$.

We used location-dependent transition probabilities based on the above calculations to compute occupancy probability (γ variable) and identify sites in the upstream promoter regions. The value of z was calculated at each upstream position. Accordingly, a different transition probability matrix was generated at each upstream position as follows. We assigned the transition probability from (i) the background and (ii) the motif states corresponding to the last motif position on either strand to the background state as $1 - z$, and the transition probability from (i) the background and (ii) the motif states corresponding to the last motif position on either strand to the motif states corresponding to the motif first position on either strand as $z/2$. Because the transition probabilities a_{ij}

are position-specific, the forward (α) and backward (β) variables and the resulting occupancy probability (γ variable) also depend upon the position in the promoter.

On the other hand, the sites in the downstream 100 bp regions (of the TSS) were identified using location-independent transition probabilities based on the uniform *z* of 0.00012 obtained in training. This choice was made due to the paucity of evidence of positional dependence of site density in these regions.

The benefit of the exponential form of *z* is that the probability of identifying a site decreases further upstream but never reaches zero (and the transition probability matrix varies accordingly). Even though this approach may fail to identify sites in distal promoters and enhancers, we believe that it allows site search in large upstream regions without identifying too many false positives. Whereas we used the exponential form by observing the locations of the known κB sites, a better functional form for location dependence can be incorporated if positions of a larger number of known sites are available.

We now examine some properties of the occupancy probabilities calculated by our HMM. A κB site exerts influence on the occupancy probabilities at the positions surrounding the site in either direction. Because self-overlapping binding sites are usually present when the scoring window is moved by one position, occupancy probability stays high at all positions in that window. When the window is moved by more positions, the occupancy probability at the new positions dips slightly below the average due to the high motif

probability at the site. The occupancy probability returns to the average background value when the scoring window moves by ten positions.

Moreover, the relationship between the occupancy probabilities of sites located close to each other is quite instructive. When two sites are in tandem without any space between them, occupancy probabilities of both of them are lower due to the small motif-to-motif transition probability. Occupancy probabilities are not very high even when the sites are one position apart because a window shift by one position from a κB site usually contains an overlapping κB site. The sites need to be at least two positions apart so that they do not exert significant influence on each other's occupancy probabilities. Please note that in any case, the overall occupancy in a region containing two nearby sites is quite high.

We will now see that all the above efforts taken to identify self-overlapping sites indeed pay off.

3.7 Our HMM Performs Better than a Weight Matrix

We compared the performance of our HMM to that of a weight matrix (WM) as follows. WM scoring was performed with the motif profile used to initialize the HMM. All overlapping windows on both strands were considered and the highest WM score was recorded. Positive examples consist of the 36 known human κB sites present in upstream 800 bp regions (in their native promoters). Negative examples consist of all 10-mers in the upstream 800 bp regions in 100 randomly selected human genes that have no

association with inflammation or cancer. Leave-one-out cross-validation was performed, where each site was scored using an HMM trained on the surround-50 promoters of the other 35 known κB sites. The lists of HMM and WM scores of the negative examples were compressed by taking the maxima of the consecutive scores above a threshold (0.03 for HMM, 4 for WM) to ensure that self-overlapping binding sites were represented by the score of the strongest site.

The ROC analysis shows that our HMM performs better than the weight matrix (Figure 3.5). While both the HMM and the WM are highly accurate when identifying strong sites, the HMM is more accurate in identifying weak sites. The segregation of weak sites from site-like sequences is quite difficult due to degeneracy and provides a crucial test. In this respect, our model far outperforms the WM. We believe that this superior performance of our HMM is the result of training the threshold in a principled manner to minimize false positives and false negatives.

We will now focus on the predictions made by our HMM.

3.8 Validation using Conservation and Expression

We predicted κB sites in the upstream 800 bp and downstream 100 bp regions of all genes in the human genome and calculated their occupancy probabilities. Two types of data suggest that they may be functional sites. Many predicted κB sites are (i) evolutionarily conserved and (ii) regulated after NF-κB over-expression

First, evolutionary conservation scores of κB sites predicted by our HMM are higher than those of 1000 10-tuples randomly selected from human promoters, and κB sites with higher HMM occupancy probability have higher evolutionary conservation scores (Figure 3.6). To calculate the conservation score of a site, its multiple alignment was retrieved from UCSC. Only mammalian sequences with at least five nucleotides present in the alignment were included. Consensus nucleotides were determined at all positions in the alignment where the human sequence did not contain a gap, and the number of sequences containing the consensus nucleotide was counted for each position. The conservation score was calculated as the ratio of the sum of these counts at all positions to the product of the number of sequences in the alignment and the number of nucleotides in the site (generally 11 or 12 for self-overlapping κB sites, 10 for non-overlapping κB sites), multiplied by 100. The perfect score, when all aligned sequences are identical, is 100. Kernel-smoothing density estimates of the conservation scores of sets of κB sites were calculated using default MATLAB parameters.

Secondly, the chicken genes regulated by over-expressed NF-κB proteins in a microarray experiment [8] and their human orthologs are enriched with κB sites predicted by our HMM (Figure 3.7). Notably, genes regulated in a higher number of comparisons are more enriched with HMM-predicted sites. Also, our HMM predicted more κB sites per regulated gene among genes predicted to contain κB sites probably because true NF-κB targets contain multiple sites. Interestingly, human orthologs of regulated chicken genes are more enriched with predicted NF-κB targets than the chicken genes themselves

probably due to the availability of higher quality sequences for humans. In this experiment, seven NF-κB proteins from different species were over-expressed in chicken DT40 pre-B cell lines, and regulated genes were identified by comparing the expression level for each experimental condition against the control. Human orthologs of the regulated chicken genes were obtained using Ensembl [144].

We will now see that the occupancy probabilities predicted by our HMM are indeed quite accurate.

3.9 Correlation with Gel Shift Experiment Results

As shown in Figure 3.8, a strong correlation exists between the occupancy probabilities predicted by our HMM and the *in vitro* binding affinity of the NF-κB family members c-Rel and RelA for oligonucleotides in a gel shift experiment (correlation coefficients of 0.91 and 0.92, respectively). This validates our physical binding model because the HMM-predicted occupancy probabilities appear to correspond to observed binding affinities.

Gel shift experiments were performed by our collaborators using double-stranded radiolabeled oligonucleotide probes containing 10-mers derived from several chicken promoters to determine if our HMM accurately predicted occupancy probabilities.

For comparison with gel shift binding affinities, occupancy probabilities need to be calculated based on an accurate transition probability to the motif (z), which corresponds to the protein concentration as we have seen above. However, protein concentration in a gel is higher than in the cellular context, and is difficult to determine. We therefore estimated z as follows: (1) calculate the occupancy probabilities of all the sequences in the gel shift experiment using various z's, (2) compute the sum of KL divergences of the occupancy probabilities of all the sequences with their binding affinities in the gel shift experiment, and (3) estimate z as the one corresponding to the minimum sum. The rationale behind this procedure is that the occupancy probabilities resulting from the correct z should be in the same ballpark range as gel shift binding affinities. KL divergence can be used as a measure to determine if they are indeed in the same ballpark range. The estimated z is 0.001 for both RelA and c-Rel.

We take this opportunity to show the effect of z on HMM-predicted occupancy probability. We plotted the HMM-predicted occupancy probability with respect to z while keeping the same motif profile (Figure 3.9). Three characteristics of the dependence between z and occupancy probability stand out: (1) Occupancy probability increases sigmoidally and then saturates as z increases. (2) Occupancy probability of a stronger site (e.g. *itm2b* vs. *bcap* κB site in the figure) saturates at lower z, and therefore occupancy probability of the stronger site is greater at a particular z. (3) Occupancy probability is influenced by surrounding sequences due to the formation of spurious self-overlapping sites (e.g. it is higher when the 3' padding sequence of a κB site in a gel shift construct starts with a C than with a T).

3.10 Take-Away for Scientists when Designing Experiments

The current practice for performing gel shift experiments in the NF-κB community consists of using particular padding sequences around the 10-mer corresponding to a potential κB site (for example see [19, 25, 29, 40]). The padding sequences, however, need to be chosen carefully. As we have discussed in Chapter 1, certain nucleotides in the padding sequences in positions adjacent to the 10-mer can form spurious κB sites due to the self-overlapping nature of κB sites, and hence the experiment cannot represent binding of the 10-mer in the native promoter. Choice of such padding sequence may lead to incorrect results.

Our HMM reveals the dependence of occupancy probability on padding sequences of self-overlapping sites, and thus offers guidance on the selection of correct padding sequences when designing experiments. We observed that the occupancy probabilities calculated using the sites in their native chicken promoters did not correlate as well with the experimental binding affinities as those calculated in the previous section using the sites and their padding sequences in the gel shift construct. When we observed that the difference was due to a C in the padding sequence 3' of the predicted κB sites in the oligonucleotides used for gel shift, we performed a systematic combinatorial analysis using HMM to determine the padding sequences that did not form spurious self-overlapping binding sites and hence affected native binding the least. We found that the use of the padding sequences in the above experiment (the 5' padding sequence is

gatctgaattcgt and the 3’ padding sequence is cacctctcctta) may misrepresent NF-κB binding. The predicted occupancy probabilities suggest that a gel shift oligonucleotide containing an A 5’ to the 10-mer and a T 3’ to the 10-mer in the padding sequence has the least chance of forming spurious binding sites (e.g. A*GGGAATTCCC*T, where the 10-mer is shown in italics). Any other nucleotide forms a spurious site shifted one position from the 10-mer, and in some cases may even change the binding occupancy by more than 50%. Any of the C, G or T in the 5’ end creates a site beginning with CGG…, GGG… or TGG…. An A, C or G in the 3’ end creates a site on the opposite strand beginning with TGG…, GGG… or CGG…. In addition, a C at the 3’end also creates a site on the same strand ending with …CCC. We have already seen in Figure 3.9 that the predicted occupancy probability of the *bcap* and *itm2b* oligonucleotides used in the gel shift experiment oligonucleotides is greater when the 3’ padding sequence begins with a C than with a T.

Based on the HMM analysis, we recommend that a gel shift oligonucleotide should have an A 5’ to the κB site and a T 3’ to the κB site for minimum interference, and that a 3’ C should be avoided at all cost. Ideally, oligonucleotides containing a few bases corresponding to those surrounding the κB site in the promoter of the gene should be used to pad the site. This will capture the effects of all the neighboring self-overlapping binding sites in the native promoter and will avoid creation of artifacts based on the nucleotides present in the padding sequence.

3.11 Biological Insights from Identification of NF-κB Targets

Although hundreds of NF-κB-regulated genes have been identified (see a review in [45]), it is not known whether many of them are direct or indirect targets of NF-κB. We therefore predicted κB sites in the promoters of all genes in the human genome. Genes containing at least one κB site with predicted occupancy probability greater than or equal to 0.5 are designated as putative direct targets of NF-κB. We also identified cellular pathways, biological functions and diseases in which our predicted NF-κB targets were over-represented through the use of Ingenuity Pathways Analysis (Ingenuity® Systems, www.ingenuity.com) and DAVID [145-147].

We were able to ascertain known NF-κB targets in the pathways NF-κB is known to regulate. We also identified novel targets in the pathways affected by NF-κB. Most interestingly, we discovered novel roles of NF-κB in the pathways in which it is not known to be involved.

As expected, we found many known NF-κB target genes with roles in B or T cell receptor signaling, NF-κB signaling, cytokine and chemokine signaling, antigen presentation, acute phase response, or in death receptor and apoptosis signaling among others (Table 3.1).

Importantly, our HMM also located several novel candidate NF-κB targets in these and other pathways that have not yet been described in literature to be regulated by NF-κB

[18, 45]. For example, (1) activation of DIABLO, which potentiates some forms of apoptosis, and the TRAF family-associated NF-κB activator TANK, which activates cell death signals and inhibits survival signals, may shed light on the less well-characterized but nonetheless important pro-apoptotic activity of NF-κB. (2) Our HMM can also illuminate novel roles of NF-κB in the ubiquitination pathway, which is responsible for the control of protein activity through proteolytic degradation. While NF-κB is known to activate the expression of deubiquitinating enzymes CYLD and A20 (TNFAIP3) that negatively regulate NF-κB signaling [148-151], our HMM identified the E2 ubiquitin-conjugating and ubiquitination-promoting enzymes UBE2H, UBE2D3 and UBE2M as putative NF-κB targets. (3) BTRC (better known as beta-TrCP) facilitates degradation of the NF-κB inhibitor IκB proteins. If BTRC is a true target of NF-κB as predicted, this would suggest that upregulation of BTRC by NF-κB could set up a positive feedback loop for amplifying NF-κB signaling. (4) Potential positive feedback loops can also be uncovered if IKBKB (IKKBeta; IKK2), which is the key NF-κB activating kinase in the canonical NF-κB signaling pathway, and Toll-like receptor 7 (TLR7), which participates in the innate immune response to microbial agents, are genuine NF-κB targets. (5) Moreover, our HMM predicted κB sites for deacetylases HDAC8 and SIRT1, as well as for transcriptional corepressor SIN3A, which may uncover a new mode of action for NF-κB in gene-specific transcriptional repression [8, 23, 36].

Our HMM also provided new insights into the influence of NF-κB on signaling pathways in which its role has not been established. This is exemplified by (1) the Notch signaling pathway, which is involved in cell-cell communications to regulate a broad array of cell-

fate determinations. While the role of the Notch signaling pathway in the activation of NF-κB pathways is known (reviewed in [152]), impact of NF-κB on the Notch signaling pathway is not. We identified κB sites in several key mediators in the Notch signaling pathway, including the delta-like 1 ligand for Notch receptors DLL1, the Notch2 receptor, transcriptional regulator RBP that acts as a transcriptional repressor in the absence of Notch but is a transcriptional activator when associated with activated Notch, and mastermind-like 2 (MAML2) that serves as a transcriptional coactivator for Notch. Thus, NF-κB may modulate Notch signaling to influence cell fate determination during development, immunity and cancer. (2) Based on the putative target list that includes the ras-related G-protein RRAS, ribosomal protein S6 kinase 1 RPS6KA1, phosphatidylinositol 3-kinase regulatory subunit PIK3R2 and transcription factor cyclic AMP-responsive element-binding protein 1 (CREB1), NF-κB may be involved in neurotrophin/Trk signaling cascade and thus in the regulation of neuronal survival and function in the nervous system. (3) Our HMM also pinpointed candidate targets in xenobiotic metabolism, including sulfotransferase SULT1C2, aldehyde dehydrogenase 3 family gene ALDH3B2, and transcription factor nuclear factor erythroid derived 2-like 2 (NFE2L2) that regulates the oxidative stress response, suggesting a possible role for NF-κB in drug metabolism, multidrug resistance and detoxification of poisonous compounds.

As we have seen, in addition to identifying many known NF-κB target genes, our HMM identified several novel candidate NF-κB targets that have not yet been shown to be controlled by NF-κB. Further studies are needed to determine which of them are genuine NF-κB targets. They will likely shed light on the role of NF-κB in its less-characterized

or novel pathways mentioned above. The experimental evidence (reviewed in [20, 31, 153, 154]) suggests a role for NF-κB in immunological and inflammatory diseases, cancer and therapy-resistance, in skeletal myogenesis and cachexia, as well as in cognition, behavior and neurological disorders. In accordance, our HMM identified κB sites in genes with roles in disease conditions ranging from immune and inflammatory disorders to infectious diseases, cancer, skeletal muscular disorders and neurological diseases (Table 3.1). In view of all of these, we believe that our HMM is a powerful tool with a potential to uncover various biological functions of NF-κB.

3.12 Summary

Our model successfully identifies self-overlapping transcription factor binding sites, besides having a straightforward physical interpretation. It has the following advantages and unique features:

- It is the first model to deal specifically with identifying self-overlapping sites. It takes into account all of the alternative binding modes of such sites.
- It provides guidance in the selection of padding sequences in gel shift experiments.
- When considered as a biophysical model, our HMM estimates a transcription factor's occupancy probability on a site. The high correlation with experimental binding affinities justifies the use of such an estimate.
- Estimation of occupancy probability offers more biological insight than most of the current site identification methods do.

- The use of occupancy probability as a discriminant function allows our HMM to learn the threshold in a principled manner. To learn the threshold, the HMM trains emission probabilities using unaligned sequences containing known sites, and estimates transition probabilities to reflect site density in all promoters in a genome. An accurate threshold thus leads to accurate identification of weak sites.
- While identifying sites, it adjusts its parameters to reflect the change in the density of sites with respect to the distance from the TSS.

On the other hand, our method has the following limitations. It requires a complicated two-step training procedure. Moreover, we have considered only κB sites, which do not contain insertions or deletions. Consequently, the architecture of the HMM is simple and well-suited for the simultaneous training of emission and transition probabilities. We do not yet know if such training is possible if a site contains insertions or deletions since it will require a full-fledged profile HMM. In addition, we have considered the transition probability to the motif z in the upstream region to decrease exponentially during the scoring procedure. This density function may not be accurate, especially for enhancers. We will need many more sites to establish a more accurate density function. Identification of sites in the introns or downstream of genes may also be difficult because few known sites are present in these regions, and therefore z in these regions cannot be estimated.

On the whole, we expect that the high evolutionary conservation scores and enrichment in experimentally regulated genes suggest that κB sites predicted by our method might be

functional. Our results may provide important new insights into the function and regulation of NF-κB and uncover possible new biological roles for this important transcription factor family.

The biophysical model described so far focuses on site identification in a single species. In the next chapter we will focus our attention to the conservation of sites in similar species. Such evolutionary conservation can be profitably used to identify the sites more accurately. We will then develop a biophysical model for site identification that incorporates evolutionary conservation.

Table 3.1: Selected pathways, functions and diseases enriched with NF-κB targets predicted by the HMM.

Selected cellular pathways, biological functions and diseases in which our predicted NF-κB targets were over-represented are shown. The associated predicted NF-κB targets are represented by official human gene symbols. Genes containing κB sites with predicted occupancy probability greater than 0.5 were used in this analysis. Genes known in the literature to be regulated by NF-κB (although not necessarily directly) [18] are denoted with *.

Pathway/Function/Disease	Gene Symbols
NF-κB Signaling	NFKB2*, CD40*, IL1F9, IKBKB, RRAS, TNFAIP3*, BCL3*, TLR7, TRAF5, NFKBIB, NFKB1*, LTA*, PIK3C3, NFKBIA*, RELB*, BTRC, PIK3R2, ZAP70, TRAF3, IL1RN*, PLCG2, MAP3K8
Glucocorticoid Receptor Signaling	VCAM1*, ICAM1*, MED1, SMAD3, IKBKB, RRAS, MAPK12, BCL3*, IL13*, CCL5*, NFKBIB, NFKB1*, IL8*, PIK3C3, NFKBIA*, NR3C1*, STAT1, CXCL3*, CREB1, PIK3R2, JAK3, SELE*, IL1RN*, IL6*
Antigen Presentation Pathway	B2M*, PSMB9*, HLA-A, CD74, HLA-B*, HLA-DQA1, TAPBP*
Acute Phase Response Signaling	SAA1*, IL1F9, RBP1, IKBKB, RRAS, MAPK12, BCL3*, SERPINA3*, NFKBIB, CFB*, NFKBIA*, NR3C1*, PIK3R2, NOLC1, SAA2*, SOCS2, IL1RN*, IL6*
B Cell Receptor Signaling	IKBKB, RRAS, MAPK12, BCL3*, NFKBIB, CALML5, NFATC1, PTPN6, NFKBIA*, PIK3C3, CREB1, MAP3K11, PIK3R2, PLCG2, MAP3K8
Death Receptor Signaling	NFKBIA*, BIRC3, DIABLO, IKBKB, BCL3*, TANK, NFKBIB, TNFSF15*
Apoptosis Signaling	NFKBIA*, BIRC3, DIABLO, IKBKB, RRAS, BCL3*, MAPK6, TP53*, NFKBIB, RPS6KA1, PLCG2, MAP3K8
Cell Cycle: G1/S Checkpoint Regulation	BTRC, SMAD3, SIN3A, TP53*, HDAC8, E2F6
Chemokine Signaling	CCL4*, RRAS, CCR3, MAPK12, CCL5*,

	PLCG2, CALML5
T Cell Receptor Signaling	NFATC1, PIK3C3, NFKBIA*, IKBKB, RRAS, PIK3R2, ZAP70, CALML5
Notch Signaling	DLL1, NOTCH2, RBPJ, MAML2
P53 Signaling	BBC3, PIK3C3, SIRT1, PPP1R13B, MED1, PIK3R2, TP53*
Xenobiotic Metabolism Signaling	IL4I1, SULT1C2, MED1, RRAS, MAPK12, NFKB1*, NFKB2*, GSTP1*, PIK3C3, PPP2CB, ALDH3B2, EIF2AK3, PIK3R2, NFE2L2, IL6*, IL1RN*, GSTA5
Neurotrophin/TRK Signaling	PIK3C3, CREB1, RRAS, PIK3R2, RPS6KA1
Protein Ubiquitination Pathway	UBE2H, UBE2D3, B2M*, UBE2M*, BIRC3, BTRC, PSMB9*, HLA-A, HLA-B*
Skeletal and Muscle Development and Function	CD40*, CSF1*, CXCL11*, DLL1, IKBKB, IL6*, IL13*, IL1RN*, MED1, NFATC1, NFKB1*, NFKB2*, NFKBIA*, RBPJ, SMAD3, STAT1, VCAM1*, WNT10B*
Infection of Virus	CCL4*, CCL5*, CLEC4M, DEFA1, ICAM1*, IL13*, IRF8, XPO1
Cancer	ACACA, AIM2, B2M*, BBC3, BCL2L10, BIRC3, BTRC, C6ORF66, CARD8, CD40*, CREB1, CTGF, CYLD, DBC1, DIABLO, DLL1, DPP4, DUT, EGR2, EIF2AK3, GNB1, GNB2L1*, HINT1, HUWE1, IER3*, IFNB1*, IGFBP6, IL6*, IL8*, IL13*, IL1RN*, IRF1*, IRF8, ITGA5, LCN2*, LTA*, LTB*, MAML2, MAP3K11, MAPK12, MEN1, MIA, MSX1, MYB*, NFKB1*, NFKB2*, NFKBIA*, NFKBIZ, NR3C1*, OAS3, PLCG2, PPP1R13B, PPP5C*, PTPN6, RBM17, REL*, RHOC, RPS6KA1, RUNX1T1, SMPD2, STAT1, THOC1, TNFAIP3*, TNFSF13, TP53*, TRAF3, TWIST1*
Rheumatoid Arthritis	ACAN, ACTA1, ADAMTS7, B2M*, BLR1*, CARD8, CCL1*, CCL4*, CCL5*, CCL19*, CD40*, CD69*, CD70, CD74, CD83*, CD86*, CD274*, CFB*, CXCL1*, CXCL2*, CXCL3*, CXCL5*, CXCL6*, CXCL10*, DEFA1, DPP4, GP1BA, HLA-A, HLA-DQA1, HPRT1, ICAM1*, IFNB1*, IL6*, IL8*, IL13*, LTA*, LTB*, MAPK12, NFKB1*, NFKBIA*, NR3C1*, PSMB9*, SAA1*, SAA2*, TNFAIP3*, TNFRSF13B, TNFSF15*, TP53*, TPM2, VIM*, WNT10B*
Experimental Autoimmune Encephalomyelitis	B2M*, CD40*, CD86*, CXCL10*, DPP4, HLA-DQA1, IFNB1*, IKBKB, IL6*, LTA*, LTB*, NR3C1*, REL*, STAT1

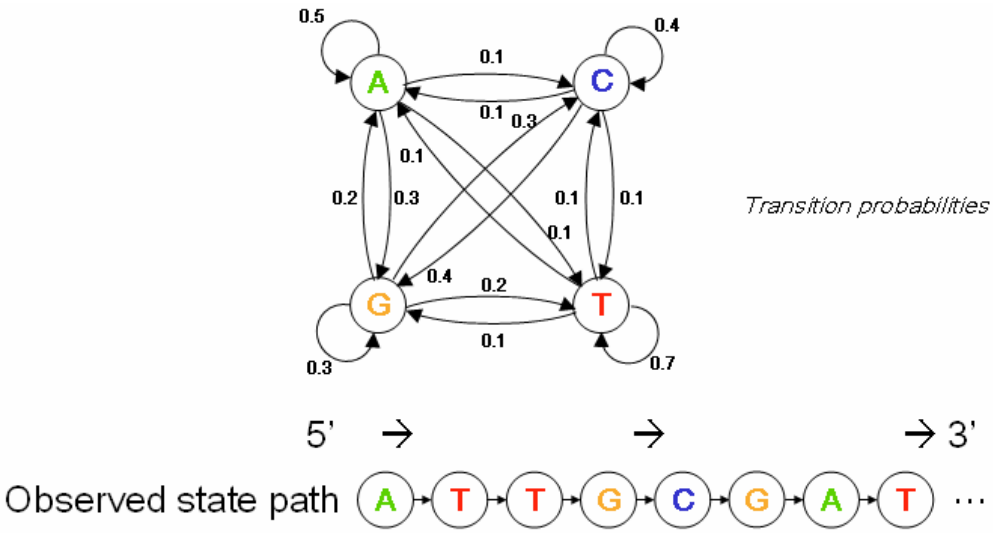
Figure 3.1: A Markov model and a hidden Markov model of a DNA sequence.

A. Markov model. The four states corresponding to nucleotides A, C, G and T are observable. The observed sequence at the bottom can be interpreted as one state path generated by the Markov model.

B. Hidden Markov model (HMM). The hidden states are the background (red) and the motif (yellow), both emitting one of the four nucleotides. One can think of an HMM as generating many hidden state paths with different probabilities, each of which can generate the observed sequence with different probabilities.

Circles represent states. Transition probabilities and emission probabilities are shown in black and blue, respectively. Initial probabilities of the states are not shown for simplicity.

A



B

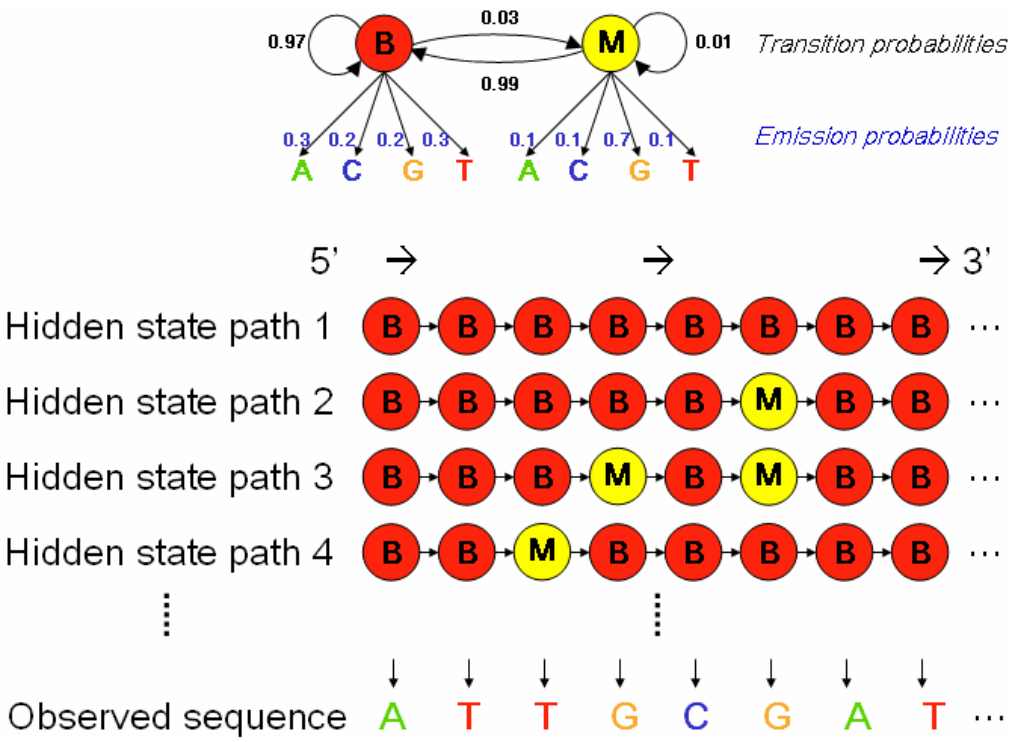

0.97
B
0.03
M
0.01
0.99
Transition probabilities
0.3 0.2 0.2 0.3
A C G T
0.1 0.1 0.7 0.1
A C G T
Emission probabilities
5'
3'
Hidden state path 1
Hidden state path 2
Hidden state path 3
Hidden state path 4
Observed sequence
A T T G C G A T

Figure 3.2: Our HMM.

Our HMM consists of 21 states. The background state is colored red and designated by B. Each of the 20 motif states corresponds to each of the ten positions within the κB motif on the two DNA strands. The motif states are colored yellow and designated using M, the position within the motif and the strand. The emission probabilities of the motif states on the two strands are flipped from 5' to 3' so as to represent identical binding irrespective of the motif strand. The transition probabilities between the states corresponding to successive positions within the motif on a strand are fixed to one. They are represented with black arrows and the transition probability values are shown. The nine transition probabilities available for training are also represented with black arrows. The sum of the transition probabilities from the background state to the states representing the first position of the motif on the two strands (z_1 and z_2) is estimated as the transition probability to the motif z . The rest of the transition probabilities are fixed to zero and are not shown.

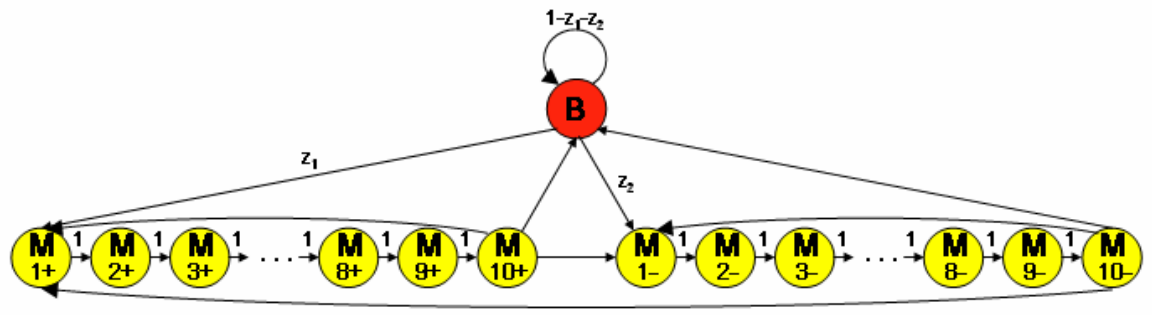
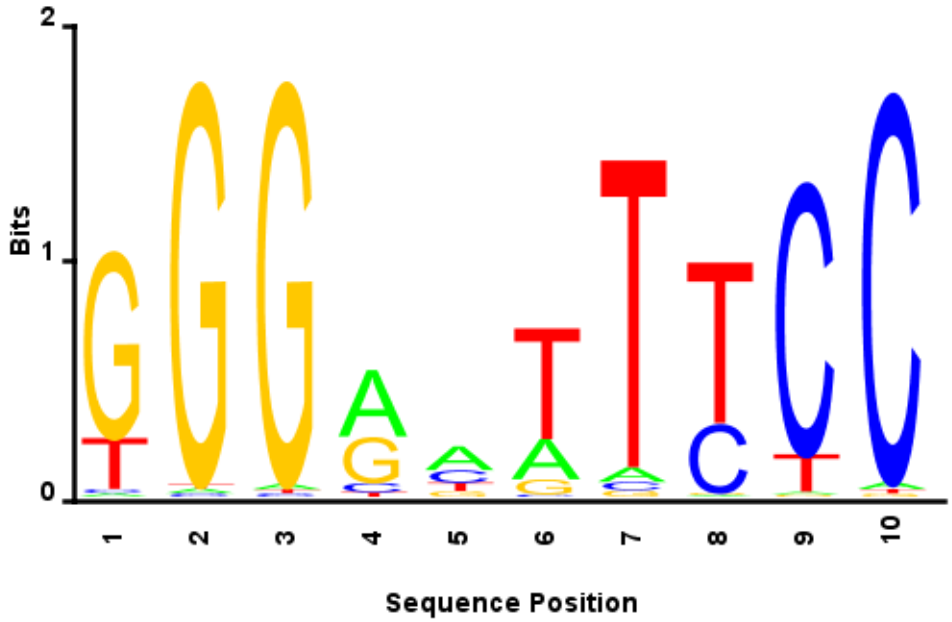


Figure 3.3: Trained HMM Parameters.

A. Sequence logo of the motif profile of the HMM trained on 50 bp sequences each consisting of a known κB site and surrounding region (surround-50 HMM) with initial transition probability to the motif (*z*) equal to 0.02. The overall height of the nucleotide stack at each position is proportional to the information content at that position and the height of each nucleotide within the stack is proportional to its frequency.

B. The estimated transition probability to the motif (*z*) for upstream 800 bp and downstream 100 bp regions with respect to the transcription start site (TSS) as the number of randomly selected training genes increases. The estimated *z* stabilizes after the addition of a few thousand genes. Each training set for estimating *z* in the upstream 800 bp region contains sequences consisting of the 20 κB sites known to be present in this region. Similarly, each training set for estimating *z* in the downstream 100 bp region contains sequences consisting of the 4 κB sites known to be present in this region.

A



B

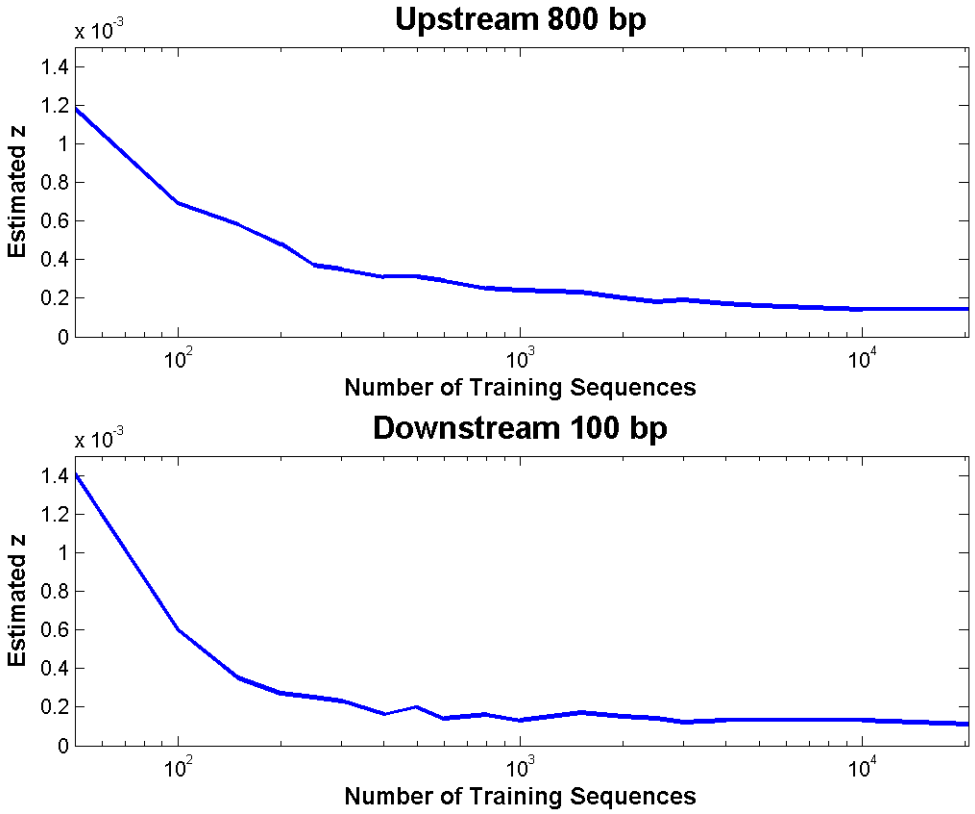

Upstream 800 bp
x 10^-3
Estimated z
1.4
1.2
1
0.8
0.6
0.4
0.2
0
10^2
10^3
10^4
Number of Training Sequences
Downstream 100 bp
x 10^-3
Estimated z
1.4
1.2
1
0.8
0.6
0.4
0.2
0
10^2
10^3
10^4
Number of Training Sequences

Figure 3.4: Trained *z* is inversely proportional to the length of the training promoter. HMMs were trained on TSS-n promoters keeping the initial motif profile fixed. The transition probability to the motif (*z*) is inversely proportional to the training promoters' length in the range between 500-3000 bp and hence $z * promoter\ length$ is constant around 0.9. This quantity drops slightly between 500 to 200 bp and then substantially after 200 bp due to the lack of κB sites in the shorter training promoters.

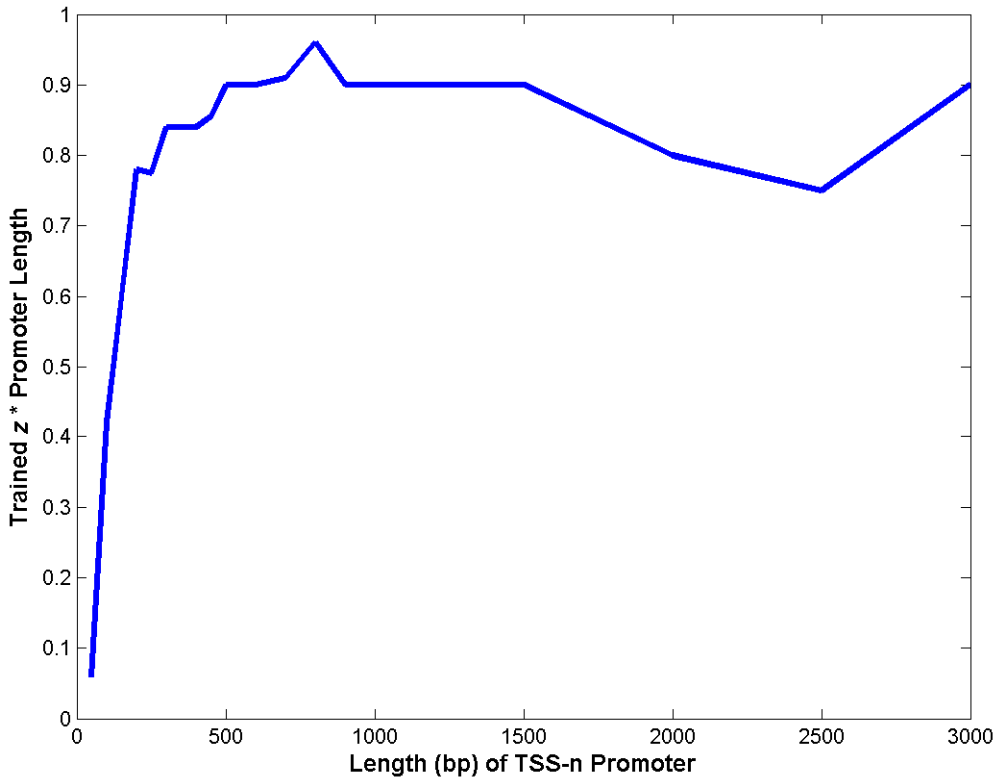


Figure 3.5: ROC analysis shows that our HMM performs better than a weight matrix.

The performances of the HMM and the weight matrix (WM) are represented by the green and the blue curves, respectively. Whereas the HMM and the WM perform similarly for strong sites, the HMM is more accurate in identifying weak sites. The positive examples consist of the 36 known human κB sites present in upstream 800 bp regions (in their native promoters), and the negative examples consist of all 10-mers in the upstream 800 bp regions in 100 randomly selected human genes as described in the text. Leave-one-out cross-validation was performed. ROC: Receiver Operating Characteristic curve.

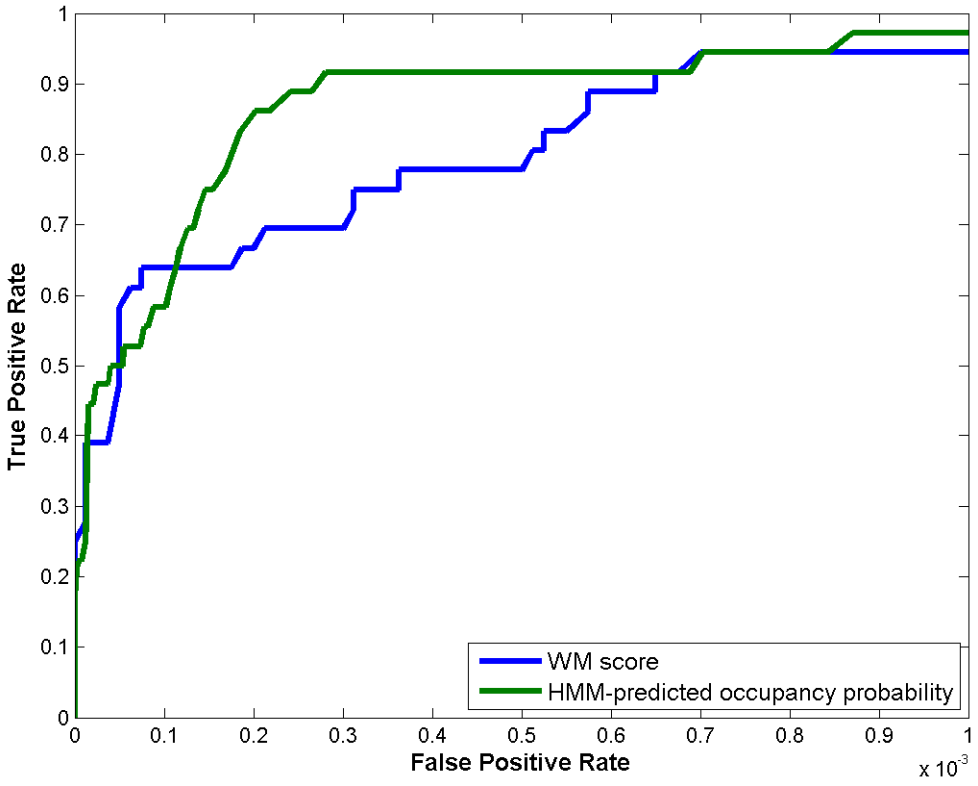


Figure 3.6: κB sites with greater HMM occupancy probability are conserved better. Each curve represents the kernel-smoothing density estimate of the evolutionary conservation scores of a set of κB sites. Each set consists of κB sites predicted by our HMM to have occupancy probability above a threshold shown in the legend. The "random" set consists of 1000 10-tuples randomly selected from the human promoters. Conservation scores of κB sites predicted by our HMM are higher than those of the random sequences. Moreover, κB sites with higher HMM occupancy probability have higher conservation scores. Conservation scores and kernel-smoothing density estimates were calculated as described in the text.

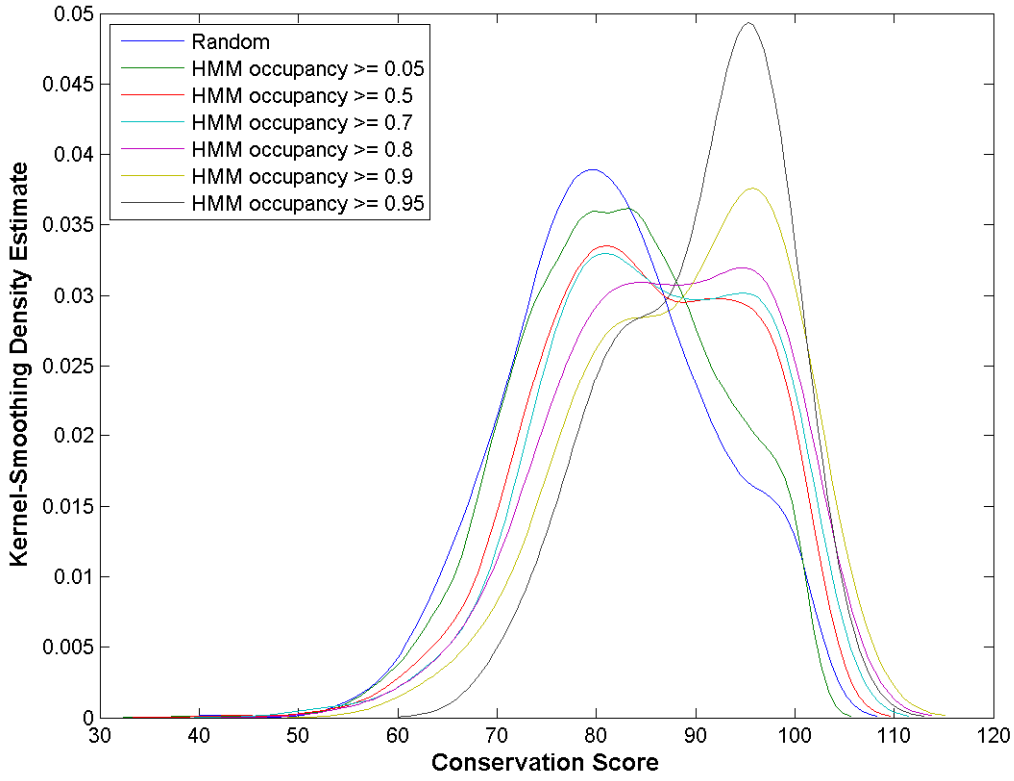
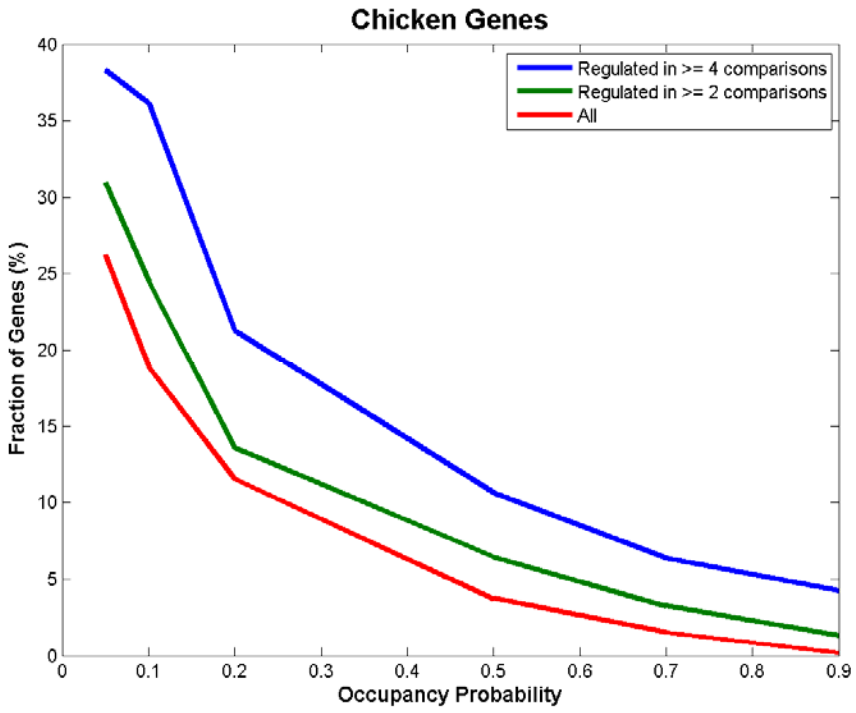


Figure 3.7: Regulated genes are enriched with HMM-predicted κB sites.

The chicken genes regulated by over-expressed NF-κB proteins in a microarray experiment and their human orthologs are enriched with κB sites predicted by our HMM. **A.** The y-axis shows the fraction of chicken genes that contains at least one κB site with HMM-predicted occupancy probability above the thresholds shown on the x-axis. The data is shown for three sets of genes: (i) genes regulated in at least four of the seven comparisons in the experiment, (ii) genes regulated in at least two comparisons and (iii) all genes.

B. The y-axis shows the fraction of the human orthologs of the chicken genes in part (A). Genes regulated in a higher number of comparisons are more enriched with HMM-predicted sites. Human orthologs of regulated chicken genes are more enriched with predicted NF-κB targets than the chicken genes themselves probably due to the availability of higher quality sequences for humans. In this experiment, seven NF-κB proteins from different species were over-expressed in chicken DT40 pre-B cell lines, and regulated genes were identified by comparing the expression level for each experimental condition against the control [8].

A


Chicken Genes
Regulated in >= 4 comparisons
Regulated in >= 2 comparisons
All
Fraction of Genes (%)
Occupancy Probability
0
5
10
15
20
25
30
35
40
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

B

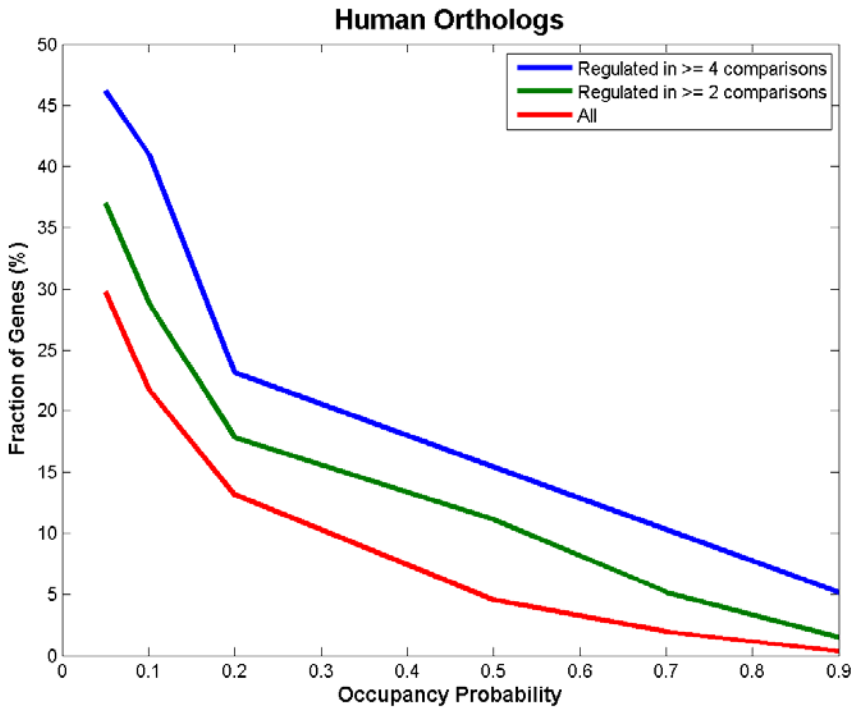
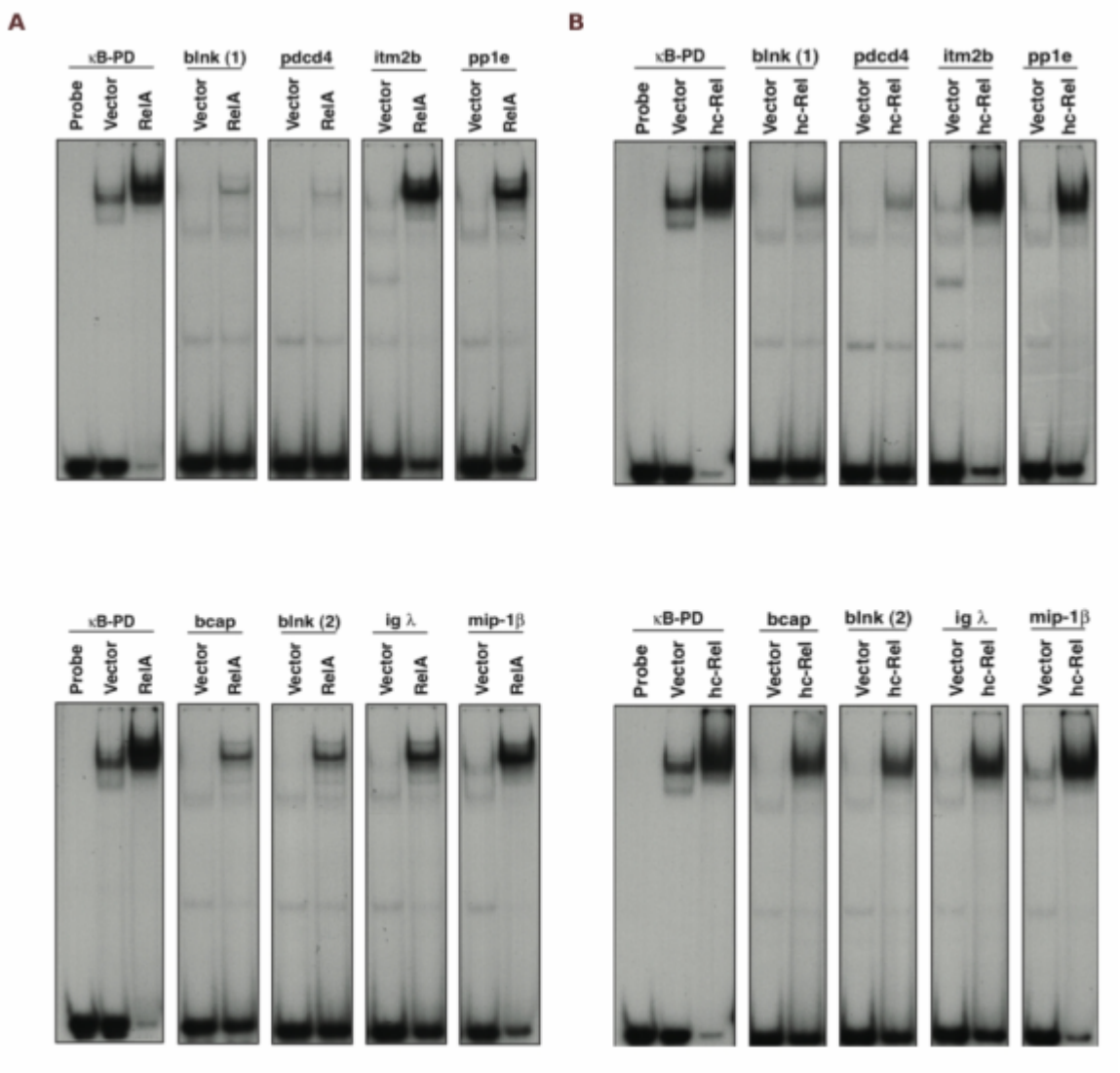

Human Orthologs
Regulated in >= 4 comparisons
Regulated in >= 2 comparisons
All
Fraction of Genes (%)
Occupancy Probability
0
5
10
15
20
25
30
35
40
45
50
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

Figure 3.8: *In vitro* binding affinity of NF-κB's RelA and c-Rel proteins to κB sites correlates well with HMM-predicted binding occupancy probability.

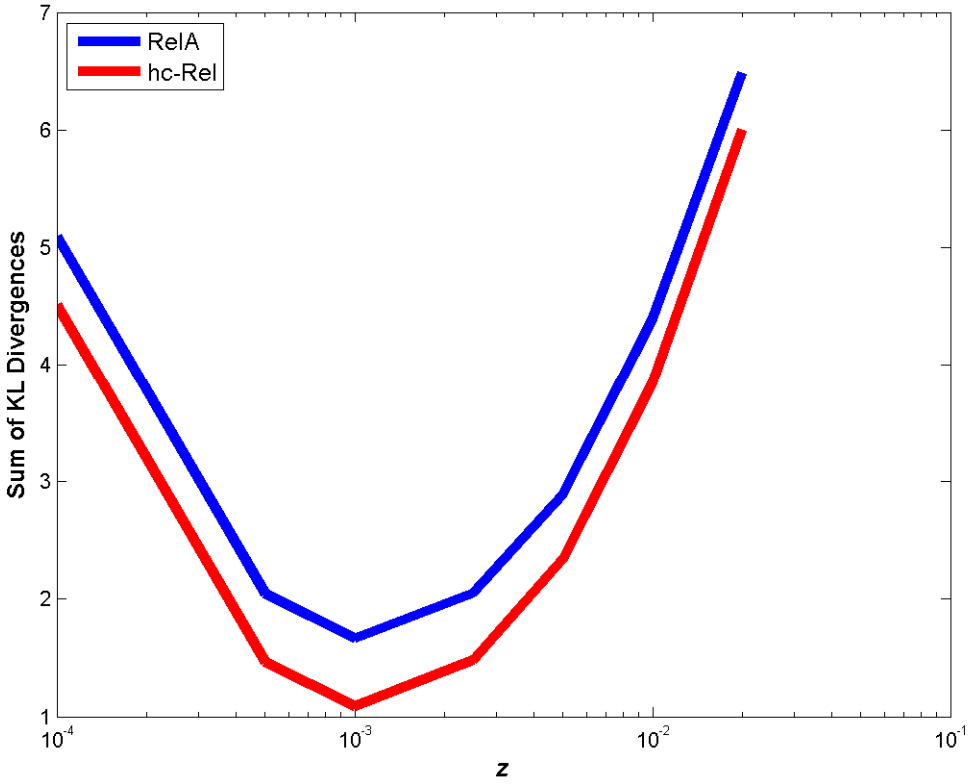
A, B. Gel shift assays with extracts from 293T cells transiently transfected with either CMV-hRelA (**A**), CMV-hc-Rel (**B**) or empty CMV vector as control (vector) and radiolabeled double-stranded oligonucleotide probes containing the predicted NF-κB sites derived from chicken *blnk* site 1 or site 2, *pdcd4*, *itm2b*, *pp1e, bcap*, *igλ*, or *mip-1β*, or a palindromic NF-κB DNA site as control (κB-PD). Reactions containing the κB-PD probe alone, in absence of cell extract, were loaded as control (probe). DNA/protein complexes were resolved from unbound DNA probes in native 5% polyacrylamide gels.

C. Sum of Kullback-Leibler (KL) divergences of the HMM-predicted occupancy probabilities of the above sequences (in the gel shift constructs) with their binding affinities in the gel shift experiments, as a function of the transition probability to the motif z. The sum of the KL divergences is minimum at z equal to 0.001 for both NF-κB proteins.

D. Correlation between the gel shift binding affinities of the above sequences and their occupancy probabilities predicted by the HMM at z equal to 0.001. The correlation coefficients are 0.91 and 0.92 in the case of RelA and c-Rel, respectively. The dashed lines are linear least square fits.


A
κB-PD
blnk (1)
pdcd4
itm2b
pp1e
Probe
Vector
RelA
bcap
blnk (2)
ig λ
mip-1β
B
hc-Rel

C



D

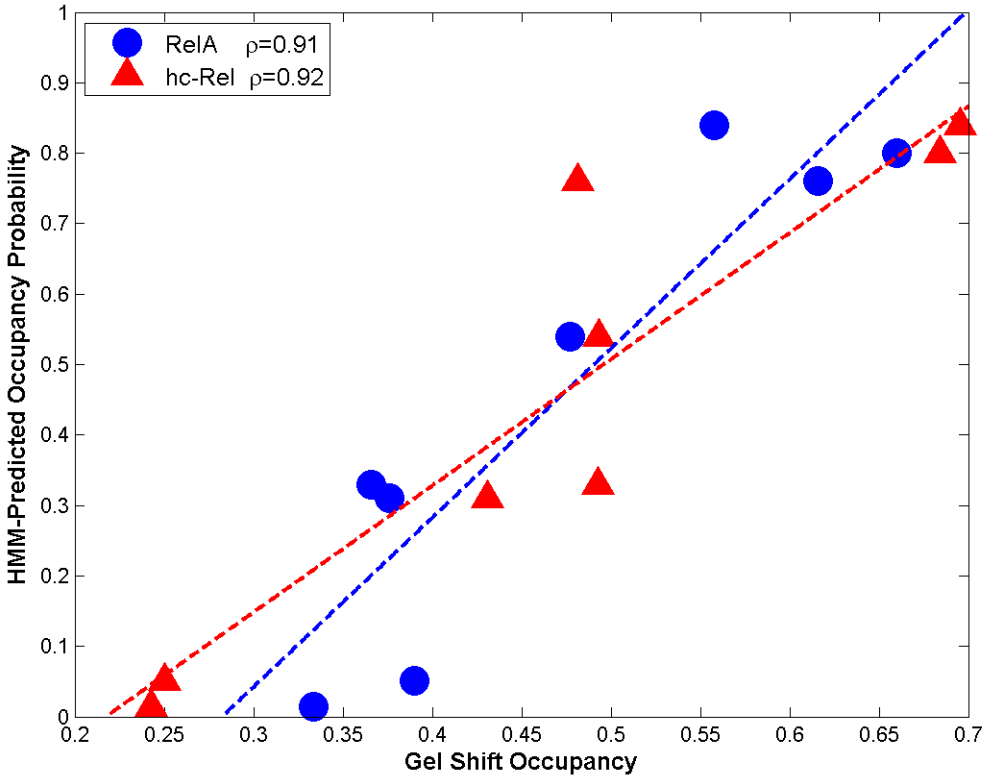
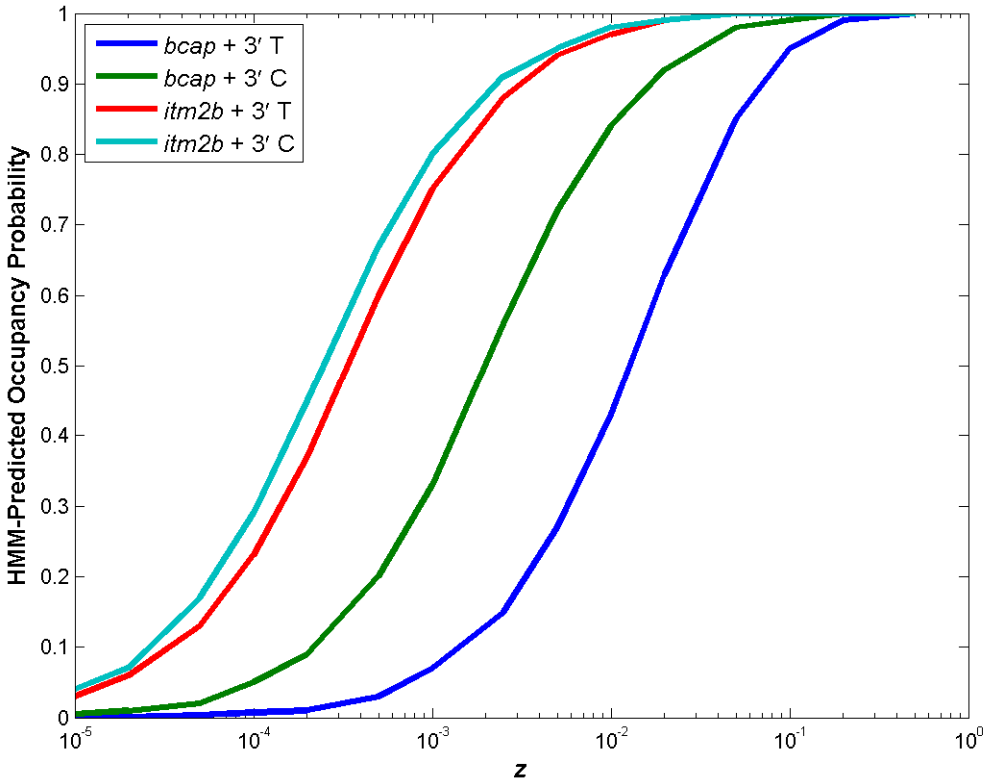

RelA ρ=0.91
hc-Rel ρ=0.92
HMM-Predicted Occupancy Probability
Gel Shift Occupancy
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7

Figure 3.9: Occupancy probability increases sigmoidally with respect to *z*, is greater for stronger κB sites and depends upon the padding sequences in the case of self-overlapping sites.

Occupancy probability of the *bcap* and *itm2b* oligonucleotides used in the gel shift experiment, with either a C or a T at the beginning of the 3' padding sequence, was predicted using an HMM with different *z*'s. The HMM's motif profile was the same in all instances. The predicted occupancy probability rises as a sigmoidal function of *z*. The occupancy probability of the stronger κB site (*itm2b* vs. *bcap*) saturates at lower *z*, and therefore the occupancy probability of the stronger site is greater at a particular *z*. Moreover, the occupancy probability of oligonucleotides is greater when the 3' padding sequence begins with a C (resulting in a stronger spurious self-overlapping site) than a T.



Chapter 4

Phylogeny, Sequence Conservation and Transcription Factor Binding Sites

"We all grow up with the weight of history on us. Our ancestors dwell in the attics of our brains as they do in the spiraling chains of knowledge hidden in every cell of our bodies."
Shirley Abbott (1934-)

4.1 Phylogeny and Evolution

Charles Darwin revolutionized biology with the theory of evolution, which asseverates that different species on earth have evolved from a common ancestor [155]. The idea that we the humans are related to all the life forms that have ever existed on earth, from albatross to algae, from Bactrian camels to bacteria, from chicken to chickpeas, from crabs to cockroaches, from dinosaurs to dandelions, from elephants to eels, from frogs to fruit flies, from kangaroos to kiwis, from mice to mosquitoes, from redwood trees to radish, from saber-toothed tigers to snakes, from wolves to worms, from yellowtail fish to yeast, from psychrophiles that inhabit arctic soils and dark oceanic depths at temperatures below 15 ˚ C to hyperthermophiles reigning at temperatures above 80 ˚ C and emanating the brilliant colors of the hot springs at the Yellowstone National Park, all originating

from a single fountainhead – a primitive microbe of over some three billion years ago, is mind-boggling, fantastic and romantic.

The phylogeny of species (*phylo-* means "tribe" or "race" and *-geny* means "origin"), i.e. their evolutionary history and relationships, can be depicted using a phylogenetic tree, also called an evolutionary tree. Each node and leaf of the tree represents a species. A node signifies the most recent common ancestor of the descendent species. A leaf denotes a contemporary species or an extinct species with no descendents. The length of each branch corresponds to the evolutionary distance or divergence between the ancestor and the descendent to which it connects. Figure 4.1 shows four examples of phylogenetic trees. The first example is a schematic of the tree of life, showing the relationships between many types of life forms on earth [156]. The second phylogenetic tree shows the evolutionary history of baker's yeast *Saccharomyces cerevisiae*, commonly used as a model organism in molecular and cell biology, and species closely related to it [157]. The third example consists of a phylogenetic tree of animals, where the divergence between different species in terms of millions of years is indicated [158]. A phylogenetic tree of the animals representing major groups of placental mammals comprises the fourth example [159]. While no consensus exists on the exact topology of these trees, they serve as good examples in demonstrating how evolutionary relationships can be depicted graphically as well as providing some quantitative measure for the relative divergence between species.

We now briefly describe the phylogeny of mammals [159-165]. The divergence estimates of periods, in million years ago (Ma), are somewhat approximate. It must also be noted that some group assignments are still subjects of controversy. Mammal-reptile divergence occurred ~310 Ma, after which placental mammals diverged from Monotremes (e.g. platypus) ~210 Ma and from Marsupials (e.g. opossum) ~180 Ma. Placental animals are divided into four superorders. While most scientists agree that the superorders Euarchontoglire and Laurasiatheria can be grouped into Boreoeutheria, the branching pattern of placental mammals into Boreoeuteria and the superorders Afrotheria and Xenarthra is not yet clearly established. The controversial Euarchontoglire superorder consists of Euarchonta, which contain primates, treeshrews and flying lemurs, and Glires, which are divided into rodents and Lagomorphs (e.g. rabbit). The Laurasiatheria superorder is a diverse group that consists of Carnivores, further divided into dog-like (e.g. wolf, seal, walrus, bear, panda, raccoon, weasel, skunk) and cat-like (e.g. lion, hyena, mongoose) mammals, Cetartiodactyla, further divided into Cetacea (e.g. whale, dolphin) and Artiodactyla (even-toed ungulate; e.g. hippo, pig, camel, llama, cattle, sheep, deer, giraffe), Chiroptera (e.g. bat) and Perissodactyla (odd-toed ungulate; e.g. horse, tapir, rhino), among others. The examples of the Afrotheria superorder, evolved mainly in Africa, are elephant, tenrec and sea cow manatee. Finally, the Xenarthra superorder evolved in Central and South America, and consists of species such as armadillo and sloth. Genomic sequences of the mammals shown in the phylogenetic tree in Figure 4.1D are available, and therefore these mammals are used in the site identification method developed in the next chapter.

Phylogenetic trees have been traditionally constructed with the help of fossil records [166]. Physiological and morphological attributes have been used as characters, or features that quantitatively determine the similarity between species and thus establish evolutionary relationships. This approach has many shortcomings. Physiological and morphological attributes may not accurately reflect the divergence between species. A common example of this are bats, which are not classified as birds in spite of the fact that they have wings. Moreover, these attributes are complex and hard to model. They are sometimes unfit to determine relationships between distant groups of species, such as between mammals and bacteria. Paucity of data is another major issue for this approach. Fossil records of these attributes corresponding to ancestor species are often not available. Thus, the number of physiological and morphological attributes that can be used for comparison is limited.

In recent years, DNA sequences are increasingly being used to create phylogenetic trees. Because DNA of a species is its blueprint, comparison of DNA sequences of different species is expected to reflect their evolutionary history accurately. The DNA sequences of all life forms on earth have the same four types of nucleotides, making determination of the evolutionary relationship between any two species possible. Moreover, DNA evolution follows a pattern and hence can be modeled mathematically. Finally, DNA sequences are long and contain a much larger amount of information than physiological and morphological attributes. These long sequences have now been available to researchers thanks to the advent of high-throughput DNA sequencing techniques in the last two decades.

The simple principle of sequence conservation is used while comparing the DNA sequences of different species. If two species are closely related, that is they diverged from a common ancestor in the recent past, their sequences have not had a great deal of time to evolve separately. Therefore, their sequences are expected to be similar. In other words, they are expected to have a high degree of sequence conservation. In contrast, sequences of distantly related species will show a low degree of conservation.

Before we discuss how to determine sequence conservation, let's briefly review the mechanism of evolution of a sequence. The first step in a sequence evolution is the mutation in the sequence of one individual. The three types of mutations at a particular position in a sequence are (1) substitution of the nucleotide by a different type of nucleotide, (2) insertion of one or more nucleotides or (3) deletion of the nucleotide. The major causes of a mutation are (1) mistakes made by the DNA replication machinery during replication, (2) intrusion by mobile genetic elements (e.g. transposons that move to different positions in a sequence), and (3) environmental factors such as radiation. The second step in a sequence evolution is fixation, in which a mutation spreads through the population of the species so that, after several generations, the entire population contains the mutation. In other words, only the descendents of the mutated individual survive after many generations and they comprise the entire population of the species. While many mutations disappear from the population, some are fixed, either by sheer chance (genetic drift) or because they offer a selective advantage to the individuals possessing them, who as a result have a greater fitness for surviving in the contemporaneous environment

(natural selection). Sometimes a mutation gets fixed in a subpopulation that is reproductively isolated from the rest of the population. If this sub-population evolves separately from the rest of the population and accumulates a large number of different mutations for many generations, it is unable to breed with the rest of the population to produce fertile offspring. It thus evolves into a separate species. Sequences in the descendent species that have evolved from the same sequence in a common ancestor by speciation (formation of new species) are called orthologous sequences.

Establishing evolutionary relationships is not the only use of sequence conservation data. As we have mentioned in Chapter 2, evolutionary conservation of sequences can be used for accurate identification of transcription factor binding sites.

4.2 Phylogenetic Footprinting

Phylogenetic footprinting is the forensic tool to identify functional sequences such as transcription factor binding sites in the non-coding regions of orthologous genes [9, 51, 158, 167-171]. The basic premise of phylogenetic footprinting is that better conserved orthologous sequences are more likely to be functional sequences (Figure 4.2). The reasoning is that functional sequences have come under a greater selective pressure than non-functional sequences during the long periods of evolution and are therefore better conserved. It is argued that a random non-functional sequence may appear like a site in one species by pure chance, but the probability of its orthologous sequences (in other species) appearing like sites is extremely low; phylogenetic footprinting thus reduces

false positives. Because genome sequences of different species have become available only in the last few years, phylogenetic footprinting is a relatively recent research field.

Phylogenetic footprinting consists of the following steps:

1. Selection of species for identifying conserved functional sequences. One generally wants to identify functional sequences in a particular species, sometimes referred to as the reference species (e.g. humans), that are conserved in other related species. The choice of related species depends upon their divergence from the reference species and upon the type of functional sequences to be identified. The selected species must have sufficient divergence from the reference species so that the non-functional sequences are not conserved (reducing false positives), but not so much divergence that even the desired type of functional sequences are not conserved (reducing false negatives). Different species, their divergence from humans (in terms of Ma – million years ago) and regions conserved with humans are listed below:
    - Bony fish (~450 Ma): Only coding sequences.
    - Birds (~310 Ma): Coding sequences and a small subset of transcription factor binding sites.
    - Mammals (<210 Ma): Coding sequences and a great number of transcription factor binding sites.
    - Primates (<40 Ma): Most functional sequences and even some non-functional sequences.

Therefore, using the above criteria, non-primate mammals appear to be the ideal choice for identifying conserved transcription factor binding sites in humans.

2. Identification of orthologous regions. This involves identification of orthologous genes, determination of their regulatory regions and removal of repeated elements. Identification of promoter regions is especially difficult in higher eukaryotes. (See the discussion in Chapter 2.) Anchors other than the annotated transcription start site (TSS) are often needed because the TSS may not be annotated correctly and the distance between the TSS and regulatory regions may vary in different species.
3. Alignment of orthologous regions. While local alignment programs have higher specificity in general, global alignment programs have higher sensitivity in aligning conserved regions, and hence they are known to perform slightly better [172]. Furthermore, programs that align multiple species simultaneously (e.g., MLAGAN, MAVID using global alignment, and MultiPipMaker, Multiz using local alignment) are better than those that align pairwise (e.g. LAGAN and AVID).
4. Calculation of a score based on sequence conservation and determination of a threshold. Local neutral substitution rates are sometimes needed to be taken into account while calculating this score because different regions of a genome evolve at different rates.

The sequences with a score above the threshold are identified as sites.

Phylogenetic footprinting methods in the literature identify sites in at least three different ways. (1) Use conservation as the sole criterion. (2) Treat site specificity (i.e. motif,

modeled by a weight matrix or energy matrix) and conservation as two separate axes of data (Figure 2.3). A site needs to have specificity and conservation scores above the respective thresholds. (3) Build a composite model of site specificity and conservation. Methods falling into the first two categories generally ignore the evolutionary relationships among the selected species and treat orthologous sequences as independent. On the other hand, a composite model incorporates the dependence of orthologous sequences using evolutionary models.

Methods that use conservation as the sole criterion typically calculate the conservation score of a putative site as the ratio of the number of identical nucleotides at the corresponding positions in orthologous sequences to the site length [51, 173, 174]. The putative site is declared a site if the score exceeds a threshold or is statistically significant. Other methods in this category forgo alignment and instead use unsupervised learning methods, designed to find over-represented motifs in independent sequences (e.g. Gibbs Sampler), on the orthologous sequences [175, 176]. The only method that takes the phylogenetic relationships into account is FootPrinter [177, 178].

Methods that treat specificity and conservation as two separate axes of data identify sites whose motif is either known (supervised learning) or not known (unsupervised learning). Supervised learning methods calculate the motif score first and the conservation score second, or vice-versa. For example, they first identify a sequence in the reference species as a putative site if its weight matrix score is above a threshold [116]. They then calculate the weight matrix scores of its orthologs and declare the putative site as a site if a certain

number of its orthologs have weight matrix scores above the threshold. This approach has been extended to identify CRMs [134, 179-181]. For example, eCIS-ANALYST first identifies putative CRMs in a reference species as regions with at least a certain number of sites with weight matrix scores above a threshold [89] and determines CRMs as those with a high number of aligned (or overlapped) and preserved (not aligned but present in the orthologous region) sites. Unsupervised learning methods use training sets that contain orthologous sequences as additional independent instances to find statistically over-represented motifs. For example, PhyloCon extends the CONSENSUS algorithm to include other species [182].

However, any method that ignores the evolutionary relationships between the species and treats orthologous sequences as independent is bound to lead to inaccuracies. As pointed out earlier, close species will tend to increase the incidence of false positives, and divergent species will tend to increase the incidence of false negatives. Put it differently, the conservation of orthologous sequences between divergent species is a better indication of sites than that between close species. Therefore, we need some quantitative measure that accounts for the closeness of species. We must account for the fact that orthologous sequences in a species pair with a larger divergence have had more time to evolve separately and are less dependent upon each other. For example, a mouse sequence should be weighed more heavily than a chimp sequence when calculating the conservation score of human sequences.

Evolutionary models take into account the dependence between orthologous sequences based upon the phylogenetic tree and the divergence between the species. In the next section, we will review evolutionary models and methods that use evolutionary models to identify transcription factor binding sites. We will focus primarily on the evolutionary models dealing with the substitution of nucleotides.

4.3 Evolutionary models

Evolution of a DNA sequence can be modeled using a Markov model consisting of the nucleotides as discrete states that can change continuously in time [120]. (See Chapter 3 for an introduction to Markov models.) An evolutionary model typically assumes that a nucleotide at each position in a sequence evolves independently of nucleotides at other positions. Its parameters are a set of $q_{\alpha\beta}$, the instantaneous substitution rates between nucleotides α and β (in this section, we use the term "substitution" instead of the standard "transition" to denote a change in Markov model states because the term transition in the biological context is reserved to denote a change from a purine (A or G) to a purine or from a pyrimidine (C or T) to a pyrimidine). By definition, sum of the instantaneous substitution rates from a nucleotide to all nucleotides including itself is 0 ($\sum_{\beta} q_{\alpha\beta} = 0$). For an evolutionary model, substitution probabilities $p_{\alpha\alpha}(t)$ and $p_{\alpha\beta}(t)$ for any time t are calculated using a system of differential equations, called the Forward Kolmogorov equations. The stationary distribution at a given site, which is essentially the probabilities of nucleotides φ_{α} at $t = \infty$, can in principle be determined from the initial

distribution. In practice, however, it is a formidable task in mathematics, as it involves sixteen coupled differential equations. Different models therefore make many simplifying assumptions, as we will see below, by placing "reasonable" constraints on $q_{\alpha\beta}$.

One such constraint is to demand that the solutions to Kolmogorov equations be time reversible. This greatly simplifies the analysis of related contemporary sequences in phylogenetic footprinting. A Markov model is considered reversible if the Markov chain running forward in time is the same as the Markov chain running backward in time. Thus, an observer watching a Markov chain cannot tell if it is going forward or backward in time. When analyzing contemporary sequences from different species, a reversible evolutionary model allows reaching one sequence from another by going to back in time to the presumed common ancestor and then forward in time to the other sequence. A model is reversible if it satisfies the detailed balance equation $\varphi_\alpha p_{\alpha\beta}(t) = \varphi_\beta p_{\beta\alpha}(t)$ for all pairs of nucleotides α and β – the equation implies that the amount of change from any nucleotide α to any nucleotide β when moving forward in time is the same as that from β to α when moving backward in time.

Several evolutionary models for DNA sequences are available in the literature (Table 4.1). They can be roughly divided into two categories. (1) Models in which the instantaneous substitution rate $q_{\alpha\beta}$ depends upon the nature of both α and β. The simplest of these models is the Jukes-Cantor model, which is a one-parameter model. Here, substitution rates from any nucleotide to any other nucleotide are assumed to be the same. Thus,

$q_{\alpha\beta} = q$ for all $\alpha \neq \beta$ and $q_{\alpha\alpha}$ is determined from $\sum_{\beta} q_{\alpha\beta} = 0$. One drawback of this model is that all nucleotides in the stationary distribution are equiprobable due to symmetry. A better model is the Kimura K2P model, which has two instantaneous substitution rates: one for transition and the other one for transversion (substitution from a purine to a pyrimidine or vice-versa). This model, too, suffers from the same drawback that the nucleotides have equal probability in the stationary distribution. (2) Models that use a back-door approach, and assign substitution rates proportional to the known (present) stationary states in such a manner as to get the stationary state independent of the initial state. In the Felsenstein F81 model, for example, the instantaneous substitution rate is proportional to the stationary probability of the substituting nucleotide φ_β, i.e. $q_{\alpha\beta} = u\varphi_\beta$, where the proportionality constant u, called multiplier, is the same for all substitutions. On solving Kolgamarov forward equations, we get substitution probabilities as $p_{\alpha\beta}(t) = \delta_{\alpha\beta}e^{-ut} + \left(1 - e^{-ut}\right)\varphi_\beta$, where $\delta_{\alpha\beta}$ is the Kronecker delta ($\delta_{\alpha\beta} = 1$ if $\alpha = \beta$, and 0 otherwise). Thus, at $t = 0$, the substitution probability matrix is an identity matrix, as it must, and at $t = \infty$, $p_{\alpha\beta}(\infty) = \varphi_\beta$, resulting in the stationary state $\left[\varphi_A, \varphi_C, \varphi_G, \varphi_T\right]$ for any initial state.

No single model is suited to represent the evolution of both sites and background sequences due to the different ways they evolve. During the evolution of a sequence at a position, first a random mutation occurs. The mutated nucleotide is fixated in a population because of genetic drift or selection. In the background (i.e. in non-functional sequences), fixation after mutation occurs only through genetic drift as they are not under

selection pressure. Therefore, one can assume that sequences at all positions in a background sequence evolve at the same rate if the mutation rates at all positions are assumed to be identical. Thus, the models that use the same parameter values for all positions are perfectly suited for background sequence evolution. The Jukes-Cantor, K2P and F81 models are among these and are more often used than the others because they have fewer parameters and they are reversible. In contrast to the background sequences, sites evolve more slowly, and different positions in sites evolve at different rates because of the functional constraint and the consequent selection pressure [183]. A "bad" mutation, for example, will be quickly weeded out through the process of selection. Therefore, evolutionary models of sites should not only have different parameter values but should also take into account position-specific variation.

Evolution of sites is generally modeled after one of the two reversible models that allow position-specific variation: the "adapted F81" model and the more complex Halpern-Bruno (HB) model. The F81 model adapted for the evolution of a site assumes that the stationary distribution at each position consists of the weight matrix probabilities at that position, and hence the substitution probability is the weight matrix probability of the substituting nucleotide at that position. The HB model separates the mutation and fixation processes [184, 185]. It assumes that mutation is identical at all positions but fixation at each position in all species considered in the phylogenetic tree occurs with a probability specific to that position. Moreover, it assumes that the time of fixation after mutation is a great deal shorter than the time between mutations, thus ignoring polymorphisms. The

substitution rate is proportional to the product of the position-invariant mutation rate and the position-specific fixation rate.

Let us discuss a model, which is a composite of weight matrix and evolutionary models, for the identification of sites when aligned orthologous sequences are provided. Let's assume that the topology of the phylogenetic tree T of the species whose sequences comprise the training set is known and the branch lengths of the tree are also known. Unlike the simple formula for the likelihood in the case of independent training sequences ($p(S \mid w) = \prod_{s \in S} p(s \mid w)$), the likelihood $p(S \mid T, w)$ at a position for orthologous sequences is a much more complicated equation with no easy analytical solution for the maximum likelihood estimators of w [90]. The likelihood is usually calculated as follows: (1) traverse the tree from the leaves (contemporary sequences) to the root, and determine the probability that each node has a particular nucleotide and its descendent leaves have the observed nucleotides using a recursion relation; (2) sum over the probabilities of all nucleotides at the root node (with the weight matrix values used as prior probabilities) to obtain the likelihood. In these calculations, the conditional probability of the child node's nucleotide given the parent node's nucleotide at each branch is the substitution probability according to one of the evolutionary models described above.

We will now review supervised and unsupervised learning methods that use a composite weight matrix-evolutionary model to identify sites. The supervised learning methods do not generally train the weight matrix explicitly using example sites from different species.

They get the weight matrix as an input, either based on the weight matrix training method described in Chapter 2 or using an unsupervised learning method described in the next paragraph. The supervised learning methods focus on scoring orthologous sequences. MONKEY uses the HB model for sites and the Jukes-Cantor or the HKY model for the background sequence [186]. It assumes that either all or none of the branches of the phylogenetic tree evolve according to motif's weight matrix (i.e. they are under selection pressure). MotEvo uses the adapted F81 model and determines the branches leading to contemporary sequences with high weight matrix scores as being under selection pressure [187]. eSimAnn aligns orthologous sequences in two species using an extended Smith-Waterman algorithm that also takes the weight matrix into account, and simultaneously identifies conserved sites in the sequences [183]. It uses the adapted F81 or HB model for sites and the Jukes-Cantor model for the background. Based on long sequence windows containing one or more aligned blocks, the HMM Stubb identifies CRMs by computing the overall likelihood as the sum of the likelihood of each block treated as one unit and the likelihood of the unaligned sequences [91, 92]. While it has reported results for two species, it can incorporate more than two species by assuming star topology (i.e. the ancestor has more than two direct descendents).

The unsupervised methods identify statistically over-represented motifs when motifs are not known by incorporating orthologous sequences using a composite weight matrix-evolutionary model. OrthoMEME, EMnEM and PhyMe extend the Expectation-Maximization (EM) algorithm [188-190]. While OrthoMEME extends MEME to two species, EMnEM assumes that sites evolve more slowly than the background under the

Jukes-Cantor model, and PhyMe uses the adapted F81 model and permits any topology of the phylogenetic tree. On the other hand, PhyloGibbs, CompareProspector and Li *et al.* extend the Gibbs sampling algorithm to include multiple species [191-193]. PhyloGibbs uses the adapted F81 model as in PhyME, but assumes star topology of the phylogenetic tree. While CompareProspector biases site search in conserved regions based on conservations scores, Li *et al.* assume that sites evolve more slowly than the background and find motifs without requiring ortholog alignments. MultiModule discovers CRMs using a coupled HMM while assuming star topology in its current form and using the adapted F81 and K2P models for the sites and the background, respectively [135].

The methods described above have the same problems of thresholds and occupancy probability as described in Chapter 2. In addition, they fail to consider site loss and turnover, which is the topic of the next section, as well as the fitness interactions between the positions within a site. This latter is discussed in the next chapter.

4.4 Site Loss and Turnover

In earlier sections we discussed methods based on sequence evolution models that deal only with substitutions within binding sites. However, sites are not always conserved across species. Even closely related species sometimes have different regulatory networks, probably owing to adaptation to different environments. This is especially true in higher eukaryotes. While many related species have a similar number of genes, the differences arise mainly due to variations in their regulatory networks. If the expression

patterns of genes in species are different, the corresponding ancestral sites are no longer under selection pressure. They evolve according to a neutral substitution rate. Even in cases in which a regulatory network is conserved, sites are not necessarily conserved. This occurs for a number of reasons. First, orthologous transcription factors can have different binding specificities due to mutations in the DNA-binding domain, subjecting sites to different selection pressures. Second, the concentrations of orthologous transcription factors and their cofactors may also be different in different species, changing the selection pressure on sites. Third, some promoters have multiple sites with redundant functions, permitting loss of a particular site without changing the function. Conserved sites sometimes falsely appear to be lost because (i) alignment programs fail to align them, particularly if the sites occur within long stretches of non-conserved sequences, or (ii) sequences in some species are simply not available in the current draft versions of their genomes. Systematic genome-wide estimates about site loss are available only in a few cases. According to one estimate, more than 30% of experimentally identified sites in *Drosophila* are not conserved [134].

An associated phenomenon is site turnover. A site at a particular location of the promoter is lost while a new site is formed at another location, thus keeping the function intact. Another variation of site turnover is that a strong site can be lost and its function is taken up by multiple weak sites in multiple locations. Moreover, sometimes the entire promoter is rearranged and thus the site cannot be properly aligned with those of the related species. The flexibility of site location within a promoter facilitates site turnovers. Site turnovers

cause problems in site identification methods that use alignments to determine site conservation.

Models that take site loss/gain into account have begun to appear in the literature only recently [194, 195]. Doniger *et al.* calculated the likelihood of semi-conserved sites by integrating over loss of site events in the phylogenetic tree of four yeast species, and compared it to the likelihoods of conserved sites or of neutral evolution [194]. They found that a great number of sites were lost in closely related yeast species and that only about half of the loss events could be explained by site turnover. Lässig and colleagues have developed a model based on binding energy distributions of sites and background sequences (more in the next chapter) that considers site loss or gain to identify sites in three bacterial species [195].

In this chapter, we have seen how models combining weight matrix and sequence evolution have been used to identify sites. In the next chapter, we will first show that modeling the evolution of the energy of an entire site is more important than that of sequences at individual positions, and then we will build a composite energy matrix-evolutionary model.

Table 4.1: Simple evolutionary models for DNA sequences.

They fall into two categories: (1) models in which an instantaneous substitution rate $q_{\alpha\beta}$ depends upon the nature of both α and β, and (2) models in which an instantaneous substitution rate is proportional to the stationary probability of the substituting nucleotide φ_β and has a corresponding proportionality constant (multiplier). The general model of the first category has 12 parameters because the constraint $\sum_\beta q_{\alpha\beta} = 0$ allows at most 3 free parameters for substitution rates from any particular nucleotide. In the general time reversible (GTR) model of the second category, multipliers for reverse substitutions are the same; for example, $q_{AG} = D\varphi_G$ if $q_{GA} = D\varphi_A$ where the multiplier D is specific to these substitutions. Transition means substitution from a purine (A or G) to a purine or from a pyrimidine (C or T) to a pyrimidine. Transversion means substitution from a purine to a pyrimidine or vice-versa. Pu = purine and Py = pyrimidine.

1. $q_{\alpha\beta}$ depends upon the nature of both α and β			
Name	**Number of Parameters**	**Description of Instantaneous Substitution Rates $q_{\alpha\beta}$**	**Reversibility**
Jukes-Cantor	1	All rates are equal	Yes
Kimura K2P	2	One rate for transition and the other for transversion	Yes
Kimura 3ST	3	One transition rate & two transversion rates: (1) A↔T/G↔C, (2) A↔C/G↔T	Yes
Kimura (3)	3	One transition rate & two transversion rates: (1) Pu→Py, (2) Py→Pu	Yes
Blaisdell	4	Two transition rates: (1) A→G/T→C, (2) G→A/C→T & two transversion rates: (1) Pu→Py, (2) Py→Pu	No
Schadt	8	Four transition rates & four transversion rates	Conditional
General	12	All rates are different	No
2. $q_{\alpha\beta}$ is proportional to β's stationary probability			
Name	**Number of Multipliers**	**Description of Multipliers**	**Reversibility**
Felsenstein F81	1	Same multiplier for all substitutions	Yes
HKY	2	Different multipliers for transition and transversion (combination of the F81 and K2P models)	Yes
General time reversible (GTR)	6	Multipliers for reverse substitutions are the same	Yes
General	12	All multipliers are different	No

Figure 4.1: Examples of phylogenetic trees.

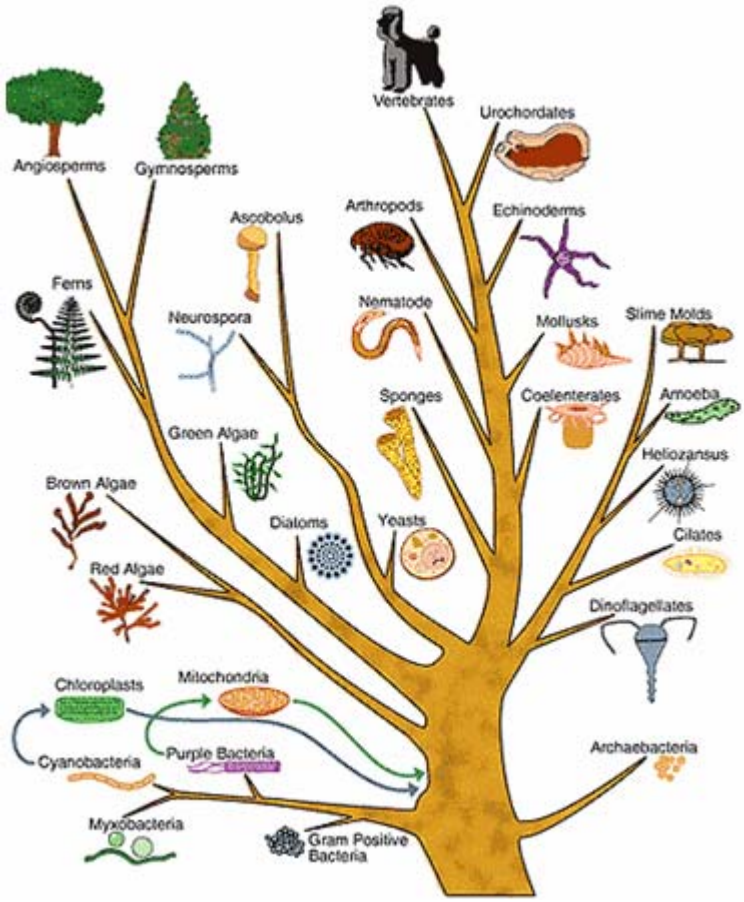
A. Tree of life.

B. Phylogenetic tree of yeast.

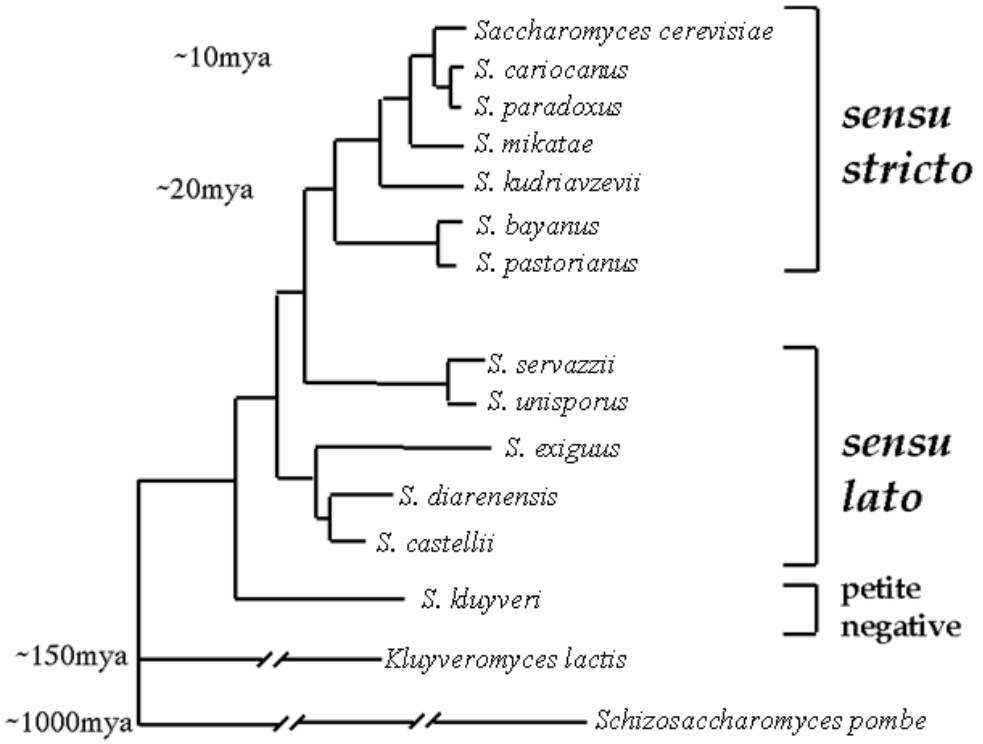
C. Phylogenetic tree of animals. Divergences in terms of million years are shown at each branch division. Common names of species are indicated in parentheses.

D. Phylogenetic tree of animals representing the major groups of placental mammals. Chicken and western clawed frog are used as outgroups; numbers represent bootstrap supports in likelihood calculations and Bayesian posterior probabilities, respectively. The trees were adapted from [156], [157], [158] and [159], respectively. They are for illustrative purposes only, as no consensus exists in the scientific community about their exact topology.

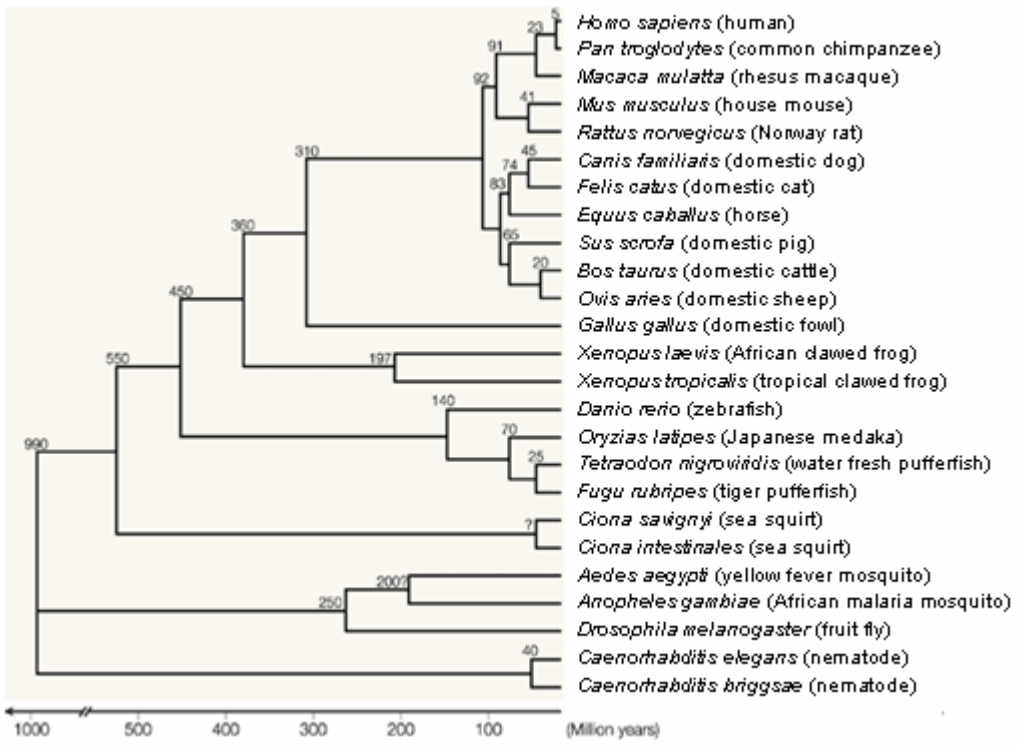
A


Vertebrates
Urochordates
Angiosperms
Gymnosperms
Arthropods
Echinoderms
Ascobolus
Ferns
Nematode
Neurospora
Mollusks
Slime Molds
Sponges
Coelenterates
Amoeba
Green Algae
Heliozansus
Brown Algae
Diatoms
Yeasts
Ciliates
Red Algae
Dinoflagellates
Chloroplasts
Mitochondria
Archaebacteria
Cyanobacteria
Purple Bacteria
Myxobacteria
Gram Positive Bacteria

B


~10mya
~20mya
~150mya
~1000mya
Saccharomyces cerevisiae
S. cariocanus
S. paradoxus
S. mikatae
S. kudriavzevii
S. bayanus
S. pastorianus
S. servazzii
S. unisporus
S. exiguus
S. diarenensis
S. castellii
S. kluyveri
Kluyveromyces lactis
Schizosaccharomyces pombe
sensu stricto
sensu lato
petite negative

C


Homo sapiens (human)
Pan troglodytes (common chimpanzee)
Macaca mulatta (rhesus macaque)
Mus musculus (house mouse)
Rattus norvegicus (Norway rat)
Canis familiaris (domestic dog)
Felis catus (domestic cat)
Equus caballus (horse)
Sus scrofa (domestic pig)
Bos taurus (domestic cattle)
Ovis aries (domestic sheep)
Gallus gallus (domestic fowl)
Xenopus laevis (African clawed frog)
Xenopus tropicalis (tropical clawed frog)
Danio rerio (zebrafish)
Oryzias latipes (Japanese medaka)
Tetraodon nigroviridis (water fresh pufferfish)
Fugu rubripes (tiger pufferfish)
Ciona savignyi (sea squirt)
Ciona intestinales (sea squirt)
Aedes aegypti (yellow fever mosquito)
Anopheles gambiae (African malaria mosquito)
Drosophila melanogaster (fruit fly)
Caenorhabditis elegans (nematode)
Caenorhabditis briggsae (nematode)
5
23
91
92
41
310
45
74
83
360
65
20
450
550
197
140
70
990
25
?
200?
250
40
1000
500
400
300
200
100
(Million years)

D

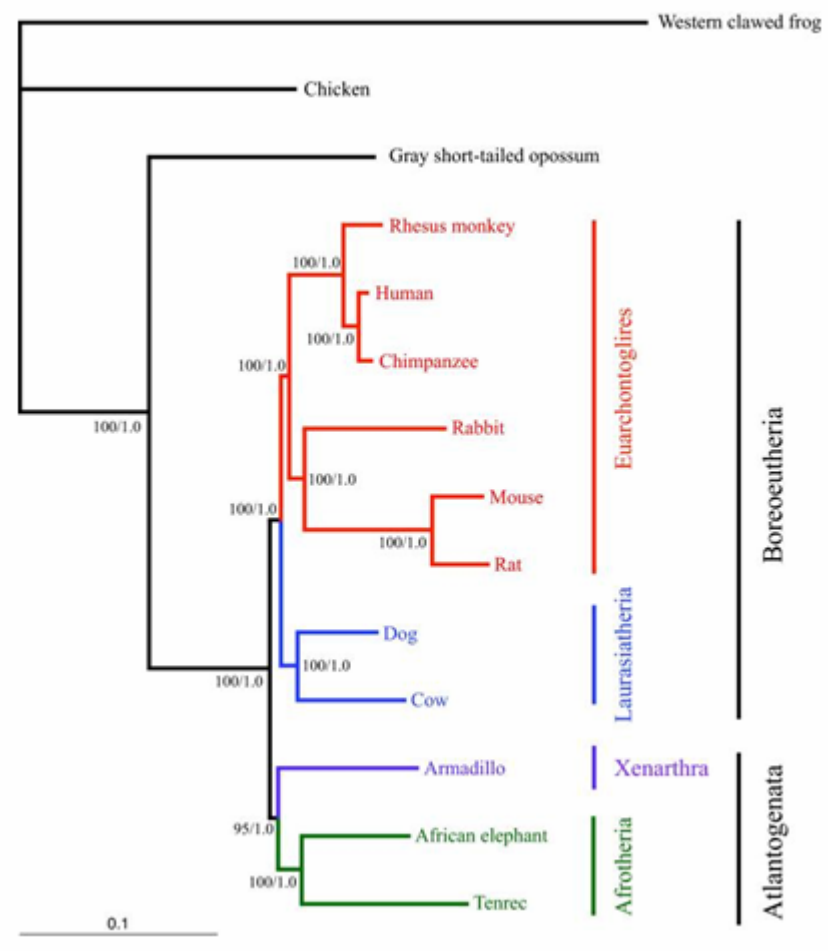
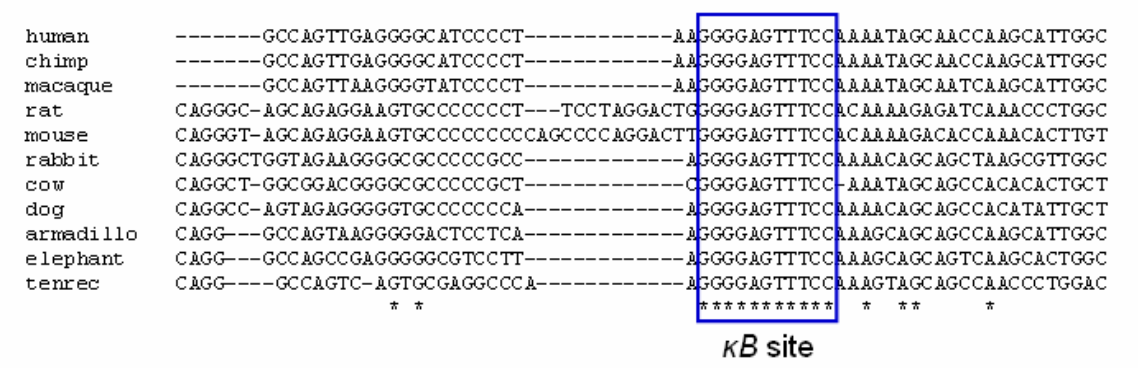

Western clawed frog
Chicken
Gray short-tailed opossum
Rhesus monkey
Human
Chimpanzee
Rabbit
Mouse
Rat
Dog
Cow
Armadillo
African elephant
Tenrec
100/1.0
95/1.0
0.1
Euarchontoglires
Laurasiatheria
Xenarthra
Afrotheria
Boreoeutheria
Atlantogenata

Figure 4.2: Illustration of the phylogenetic footprinting principle.

Sequence alignment of a regulatory region of the CCL5 gene in eleven mammals is shown. A binding site for the transcription factor NF-κB (κB site) is more conserved than the surrounding sequences. Sequence positions containing the same nucleotide in all the species are indicated with an asterisk in the bottom row. The sequences were obtained from UCSC [141] and the alignment was created using Clustal W [196].



Chapter 5

Phylogeny Based Biophysical Model to Identify Conserved Sites

"In time of test, family is best."
Burmese Proverb

5.1 Site Energy, Occupancy and Fitness

The ultimate measure of a site's functional significance is the evolutionary fitness of the individual possessing the site [197-201]. Fitness is the central concept in evolutionary biology that describes the ability of an individual of a particular genotype, or genetic makeup, to reproduce. Evolving sites are subject to two opposing forces, mutation and selective pressure. While a mutation in a site's sequence tends to weaken its fitness, selection pressure maintains or even increases its fitness [72, 111, 112, 197]. A site is lost when mutation destroys its functionality, and it is conserved when the selection pressure wins. But what aspect of a site is under selection pressure, and how does this aspect determine the site's fitness?

A site's binding energy, and not its sequence, is under selection pressure [110-113, 195, 202, 203]. An entire site is the functional unit, its basic phenotype is the binding energy,

and its fitness depends entirely on the binding energy. A site imparts its function entirely through its binding energy, regardless of its sequence. In other words, two sites with different sequences but the same binding energy are functionally equivalent. It has been shown that while the sequences of many orthologous sites are quite different, their energies are quite similar. The change in binding energy caused by the substitution in one position is negated by compensatory substitutions in the other positions in the site. Moreover, nucleotides only at energetically important site positions are highly conserved because substitutions there would change binding energies drastically in the absence of strong compensatory substitutions at the other positions. Thus, a function is conserved through the conservation of energy, not of sequences, and therefore a proper model of a site's energy is essential to understand its evolution and to identify conserved sites.

The connection between a site and fitness has not been well established. Lässig and colleagues have showed that fitness is a non-linear function of energy. They have proposed that fitness is related to the log ratio of binding energy distributions of sites and background sequences. The resulting fitness has a mesa landscape. Fitness decreases non-linearly with a negative curvature as energy increases to a threshold value. The energy above the threshold value corresponds to background sequences and hence the corresponding fitness value is flat.

The non-linear relationship between energy and fitness causes fitness interactions, or epistasis, between nucleotides at different positions in a site [203]. When the nucleotide at a position in a site is substituted, the binding energy change is independent of

nucleotides at other positions in the site because binding energies of individual nucleotides in a site are approximately additive. However, the change in fitness depends upon the initial energy and thus on all the other nucleotides in the site. Thus, the evolution of nucleotides at any two positions in a site is correlated.

Models in the previous chapter assumed that nucleotides at different positions in a site evolve independently of each other. Fitness in these models therefore is a linear function of the binding energy [185, 194]. A linear relationship between energy and fitness, however, cannot explain the observed evolutionary correlations at various positions in a site. Thus, these models are not adequate to account for site evolution.

Although we agree in principle that fitness should be a non-linear function of energy and that a model incorporating this fact is necessary to capture the evolution of a site, we claim that fitness can be better approximated as a linear function of occupancy probability. One can easily visualize this when the transcription factor is active in only one cellular state. Let a cellular state be characterized by a particular concentration of the transcription factor. Now consider two cellular states, the inactive one in which the transcription factor is absent (and hence there is no activity associated with it), and the active one in which the transcription factor is present with a certain concentration, responds to a stimulus and imparts its function. One can intuitively understand that the occupancy of a site in the active state, i.e. the amount of time the transcription factor sits on a site, will determine the increase or decrease (in the case of activation or repression, respectively) in the number of mRNA copies it makes of the target gene. Thus, the

occupancy probability is better related to fitness. It is easy to see that the non-linear relationship between fitness and binding energy is a built-in feature of our model which considers a linear relationship between fitness and occupancy probability. (Note that occupancy probability itself is a non-linear function of binding energy since it has the Fermi-Dirac form for a particular concentration of the transcription factor.)

We are aware that certain objections to this approximation are possible at this point. It may be contended that in the case of a graded response by a transcription factor to a stimulus, that is when the transcription factor is active in many cellular states, fitness will be a function of the combinations of occupancy probabilities in these multiple states. This will result in a very complicated model. Moreover, when constitutive activity of the transcription factor is not desired, as in the case of NF-κB, a strong binder site with very low binding energy may have a low fitness. Such a site can bind even in the inactive cellular state when the concentration of the transcription factor is negligible and thus transcription factor activity is not controlled. However, this lack of control does not necessarily have to be programmed in binding energy. We know that external controls such as tethering NF-κB with IκB in the inactive cellular state have been used by nature. Another possible objection is that there appears to be a wide gap between the control of abundance of the mRNA of a gene and the evolutionary fitness of the individual, and that there is very little by way of theoretical understanding and available experimental data to link these two at the present juncture.

These objections can adequately be addressed as more research becomes available. It should be clear, however, that occupancy probability is a more logical choice than, say, binding energy or an arbitrary function of the site sequence to establish site fitness. This is one of the reasons for using occupancy probability as the discriminant function for identifying sites. In the next section, we will develop a model of site binding energy evolution that wields occupancy probability to identify conserved sites.

5.2 PhyloQPMEME: Using Covariance of Energies of Orthologous Sequences

PhyloQPMEME (Phylogeny-based Quadratic Programming Method of Energy Matrix Estimation) integrates the biophysical model QPMEME, reviewed in Chapter 2, with evolutionary conservation to accurately identify sites of a transcription factor using a principled threshold [204]. It constructs a model of binding energies of orthologous sequences. For a particular transcription factor's binding sites, it estimates the energies of nucleotides at each position of the sites by optimizing the distribution of binding energies of orthologs of neutrally evolving sequences while restricting the values of binding energies of experimentally validated sites and their orthologs. PhyloQPMEME performs quadratic programming, a special type of constrained optimization, iteratively to arrive at the solution. The training set consists of experimentally validated sites as well as their orthologs. In the scoring stage, PhyloQPMEME identifies evolutionarily conserved sites by calculating the binding energies and occupancy probabilities of orthologous sequences in all the considered species.

Our motif model consists of an energy matrix as in QPMEME described in Chapter 2. In brief, the energy matrix has dimensions 4 x ℓ, where ℓ is the length of the site. Each matrix element corresponds to the binding energy of a nucleotide at a position in the site. Binding energies of the nucleotides at various positions in the site are assumed to be independent of the other positions, and are added to give a good approximation of the total binding energy of the site.

The binding energy of any sequence s of length ℓ can be calculated using the following vector notation. Let $\mathbf{S}$ be the sequence vector of length 4ℓ such that each element $s_{i\alpha}$ equals one if the sequence has nucleotide α at the i th position and zero otherwise. Thus, the sequence vector has the form $\mathbf{S} = \begin{pmatrix} s_{1A} & s_{1C} & s_{1G} & s_{1T} & s_{2A} & s_{2C} & s_{2G} & s_{2T} & s_{3A} & \cdots \end{pmatrix}$. For example, sequence CGA… can be represented using the sequence vector $\mathbf{S} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & \cdots \end{pmatrix}$. The energy matrix can be similarly written as energy vector $\boldsymbol{\varepsilon}$ of length 4ℓ of the form $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_{1A} & \varepsilon_{1C} & \varepsilon_{1G} & \varepsilon_{1T} & \varepsilon_{2A} & \varepsilon_{2C} & \varepsilon_{2G} & \varepsilon_{2T} & \varepsilon_{3A} & \cdots \end{pmatrix}$, and its each element $\varepsilon_{i\alpha}$ is the binding energy of a nucleotide α at the i th position. Then, the binding energy of the sequence is $E(s) = \boldsymbol{\varepsilon} \cdot \mathbf{S} = \sum_{i=1}^{\ell} \sum_{\alpha=1}^{4} \varepsilon_{i\alpha} s_{i\alpha}$.

A unique feature of PhyloQPMEME is the model of binding energies of neutrally evolving orthologous sequences. A sequence and its orthologs are collectively defined as

the orthologous set of the sequence. Let $\mathbf{E}$ be the vector of binding energies of an orthologous set in d species. The binding energy of a sequence in any species is calculated using the energy matrix described above. We assume that the binding energy of a nucleotide at a position ($\varepsilon_{i\alpha}$) is equal in all species. We also approximate the set of neutrally evolving sequences to all possible random sequences, and hence the terms "neutrally evolving" and "random" are used interchangeably. The binding energy distribution of all sequences of length $\ell \gg 1$ in one species is approximately normal because binding energy of a sequence is the sum of the binding energy at each position, each of which is a random variable. Analogously, we assume that the binding energies of orthologous sets of all possible sequences have the multivariate normal distribution

$p(\mathbf{E}) = \dfrac{e^{-\frac{1}{2}\mathbf{E}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{E}}}{(2\pi)^{d/2}|\mathbf{C}|^{1/2}}$, where the mean of $\mathbf{E}$ is assumed to be 0 and $\mathbf{C}$ is the covariance matrix of binding energies of orthologous sets. The covariance matrix is given as

$$\mathbf{C} = \begin{pmatrix} \sum_{i=1}^{\ell}\sum_{\alpha=1}^{4} p_{\alpha}^{A}\varepsilon_{i\alpha}^{2} & \sum_{i=1}^{\ell}\sum_{\alpha=1}^{4}\sum_{\beta=1}^{4} p_{\alpha\beta}^{BA}\varepsilon_{i\alpha}\varepsilon_{i\beta} & \sum_{i=1}^{\ell}\sum_{\alpha=1}^{4}\sum_{\beta=1}^{4} p_{\alpha\beta}^{CA}\varepsilon_{i\alpha}\varepsilon_{i\beta} & \cdots \\ \sum_{i=1}^{\ell}\sum_{\alpha=1}^{4}\sum_{\beta=1}^{4} p_{\alpha\beta}^{AB}\varepsilon_{i\alpha}\varepsilon_{i\beta} & \sum_{i=1}^{\ell}\sum_{\alpha=1}^{4} p_{\alpha}^{B}\varepsilon_{i\alpha}^{2} & & \\ \sum_{i=1}^{\ell}\sum_{\alpha=1}^{4}\sum_{\beta=1}^{4} p_{\alpha\beta}^{AC}\varepsilon_{i\alpha}\varepsilon_{i\beta} & & \ddots & \\ \vdots & & & \ddots \end{pmatrix},$$

where α and β are nucleotides, the capital letters in the superscripts denote species, p with one subscript and one superscript is the probability of a nucleotide in a species, and p with two subscripts and superscripts is the joint probability of two nucleotides in two species. These probabilities are assumed to be identical at all positions of a random sequence, and they can be readily obtained from aligned sequences.

The covariance matrix of binding energies of orthologous sets captures the evolutionary relationships of the selected species. As we saw above, it is constructed using the joint probabilities of all pairs of nucleotides for all pairs of species. The joint probabilities decrease as the divergence between the species pairs increases. The shape of the resulting multivariate normal distribution of binding energies correctly reflects the species divergence. The greater the divergence between two species, the less the correlation between the binding energies of the corresponding orthologous sequences. In effect, this model of binding energies gives more importance to the sequence conservation in highly diverged species. The representation of evolutionary relationships by the covariance matrix allows any number of species for sequence comparison and any topology of the phylogenetic tree. The only restriction is that a sufficient number of aligned sequences are available for accurate calculation of joint probabilities of nucleotides. Incidentally, because PhyloQPMEME explicitly calculates the joint probabilities of nucleotides in the contemporary species, it does not need to assume an evolutionary model.

The principled threshold used by PhyloQPMEME, as by QPMEME, comes from the distribution of occupancy probability, which it uses as a discriminant function for classifying sequences as sites. From basic thermodynamics, occupancy probability has a Fermi-Dirac distribution with a natural threshold at the chemical potential μ. Because the standard deviation of the distribution of binding energies of all sequences of length $\ell \gg 1$ is much greater than $K_b T$ [110-113], where K_b is the Boltzmann constant and T is the absolute temperature, the Fermi-Dirac distribution can be approximated by the step

function when compared on its scale. Thus, a site has binding energy $E(s) < \mu$ and hence it is occupied, whereas a random sequence with $E(s) > \mu$ is not occupied. The threshold binding energy μ is at the far left of the mean of the normal distribution, and hence a few sequences have energy less than μ. The threshold binding energy can be different in different species.

The goal of PhyloQPMEME is to estimate $\varepsilon_{i\alpha}$ such that all known sites and their orthologs have binding energies below a threshold and occupancy probabilities of one, whereas the probability that a random sequence and its orthologs have binding energies below the threshold and are thus occupied is as small as possible (Figure 5.1).

Before proceeding, let's note from the figure that a random sequence may have a binding energy below the threshold by chance, but the probability that all its orthologs also have binding energies below the corresponding thresholds is considerably small. PhyloQPMEME thus reduces false positives by considering multiple species.

To estimate $\varepsilon_{i\alpha}$, PhyloQPMEME optimizes the joint probability distribution of binding energies of orthologous sets of random sequences. Our assumption that this distribution is multivariate normal facilitates the formulation of this problem. If we set the vector of binding energy thresholds in different species $\boldsymbol{\mu} = [\mu_1 \quad \mu_2 \quad \cdots \quad \mu_d]$ to $[-1 \quad -1 \quad \cdots]^d$, binding energies of sequences in each species will be determined in the units of the corresponding thresholds. The probability that the orthologous set of a random sequence

has binding energies below the threshold (shaded area in Figure 5.1) is given by the integral $I = \frac{1}{(2\pi)^{d/2}} \int_{-\infty}^{\boldsymbol{\mu}} \frac{e^{-\frac{1}{2}\mathbf{E}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{E}}}{|\mathbf{C}|^{1/2}} d\mathbf{E} \approx cofactor.|\mathbf{C}|^{1/2} e^{-\frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{C}^{-1}\boldsymbol{\mu}}$. This probability is determined by the two covariance terms. Because the exponential term dominates, we can write $I \approx e^{-\frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{C}^{-1}\boldsymbol{\mu}}$. Thus, this probability is minimized by solving the optimization problem $\max_{\boldsymbol{\varepsilon}} \boldsymbol{\mu}^{\mathrm{T}}\mathbf{C}^{-1}\boldsymbol{\mu}$. Although the covariance matrix $\mathbf{C}$ is a quadratic function of $\varepsilon_{i\alpha}$, $\mathbf{C}^{-1}$ is not, and hence the objective function (see definition below) is hard to optimize. We can bypass this issue by solving an equivalent optimization problem of $\min_{\mathbf{y},\boldsymbol{\varepsilon}} \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{C}\mathbf{y} - \boldsymbol{\mu}^{\mathrm{T}}\mathbf{y}$. We prove the equivalence as follows. Let $f = -\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{C}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\boldsymbol{\mu}$. At its maximum value, $\frac{\partial f}{\partial \mathbf{y}} = 0 = -\mathbf{C}\mathbf{y} + \boldsymbol{\mu}$ and hence $\mathbf{y} = \mathbf{C}^{-1}\boldsymbol{\mu}$. Substituting the values of $\mathbf{y}$ in f, $f = -\frac{1}{2}\left(\mathbf{C}^{-1}\boldsymbol{\mu}\right)^{\mathrm{T}}\mathbf{C}\left(\mathbf{C}^{-1}\boldsymbol{\mu}\right) + \left(\mathbf{C}^{-1}\boldsymbol{\mu}\right)^{\mathrm{T}}\boldsymbol{\mu} = \frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{C}\mathbf{C}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^{\mathrm{T}}\mathbf{C}^{-1}\boldsymbol{\mu} = \frac{1}{2}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{C}^{-1}\boldsymbol{\mu}$. The reason for using numerical optimization to solve the problem is that no analytical solution exists.

We now proceed to deal with the constraints associated with this optimization problem, such as the constraint placed on the values of binding energies of the training sites and their orthologs. The constrained optimization problem requires the use of Lagrange's unknown multipliers, as outlined in the next section.

5.3 Constrained Optimization and Lagrange Multipliers

Optimization, also called mathematical programming, is a branch of mathematics devoted to finding the extremum (minimum or maximum) value of a function, called objective function, and the corresponding values of its variables $\{x_1, x_2, \ldots\}$, collectively written as x [205]. We will focus on function minimization in this section, since minimizing any function is equivalent to maximizing its negative function (or maximizing its reciprocal, if the function does not change sign). In an optimization problem, a local minimum is the point in the variable space at which the value of the function is smaller than that of any other point in the close neighboring region. A global minimum, on the other hand, is the point at which the value of the function is smaller than at any other point in the entire variable space. One is generally interested in finding the global minimum, which is either the smallest of minima or some value on the boundary of the variable space. When solving an optimization problem, one then has to find all local solutions first. A major issue in optimization is that for complicated objective functions, the number of local solutions is usually unknown and hence it is difficult to ascertain that the global solution has indeed been arrived at.

There are two main types of optimization problems: unconstrained and constrained. In an unconstrained optimization problem, the solution can lie anywhere in the variable space. A constrained optimization problem, however, consists of constraints on what values each variable or the combination of variables can take. One needs to consider two types constraints: (i) equality constraints $c(x) = 0$ and (ii) inequality constraints $c(x) \geq 0$

(constraints of the type $c(x) \leq 0$ are equivalent to $-c(x) \geq 0$). The part of the variable space that satisfies the constraints is called the feasible region. An inequality constraint is said to be active at point x if $c(x) = 0$. This point lies on a boundary of the feasible region. An inequality constraint is inactive at a point inside the feasible region, i.e. when $c(x) > 0$. In constrained optimization, the objective function needs to be minimized while satisfying the constraints on its variables, and the solutions lie in the feasible region.

We will now briefly describe the three important special cases of constrained optimization: convex programming, linear programming (LP) and quadratic programming (QP). In convex programming, the objective function is convex, the equality constraints are linear and the inequality constraints are concave. Therefore, local solutions are global solutions. This feature is a great advantage because one only needs to find one local solution to find the global solution, which is relatively easy as compared to finding an unknown number of local solutions as in the general case. The second special case is LP, used most commonly in practice, which consists of a linear objective function and linear constraints. QP is the third special case and consists of a quadratic objective function and linear constraints. QP, when the objective function's Hessian is positive semi-definite, and LP are in fact examples of convex programming. The QPMEME algorithm described in Chapter 2 uses convex QP [73]. Moreover, the PhyloQPMEME algorithm solves a constrained quartic optimization problem (quadratic in two sets of variables) using iterative convex QP, as we will see in the next section.

Let's see the general conditions for a local solution in a constrained optimization problem (and global solution in convex programming). At any point x in the feasible set, let d be a direction in which the objective function $f(x)$ decreases. According to the Taylor series, $f(x+d) - f(x) \approx \nabla f(x) \cdot d < 0$, and hence the angle between the gradient $\nabla f(x)$ and d is greater than $90°$, or equivalently, d is in the open half-space opposite of $\nabla f(x)$ (Figure 5.2A). x is a solution only if no d exists at x. This is possible if (i) $\nabla f(x) = 0$ or (ii) the open half-space opposite of $\nabla f(x)$ lies outside the feasible region.

We will now see the conditions for a local solution when (i) it lies inside the feasible region, (ii) an equality constraint is active, or (iii) an inequality constraint is active. Inside the feasible region (when an inequality constraint is inactive, or in unconstrained optimization where the entire variable space is the feasible region), d will not exist at a point only if $\nabla f(x) = 0$, and hence a solution fulfills the condition $\nabla f(x) = 0$. A point where an equality or inequality constraint is active can be a solution if it fulfills either of the conditions described above. Let's now focus on the second condition for equality and inequality constraints. We discuss them separately because the details of the condition differ for each constraint.

In the case of an equality constraint $c(x) = 0$ (Figure 5.2B), direction d satisfies the constraint (i.e. $c(x+d) = 0$) if $c(x+d) \approx c(x) + \nabla c(x) \cdot d = \nabla c(x) \cdot d = 0$, and thus d is perpendicular to $\nabla c(x)$. We noted above that the angle between d and $\nabla f(x)$ is

greater than $90°$. From these two statements, we can see that d does not exist only when $\nabla c(x)$ is in the same or opposite direction of $\nabla f(x)$, leading to the contradiction $\nabla f(x) \cdot d = 0$ and $\nabla f(x) \cdot d < 0$. Thus, when an equality constraint is active, a solution fulfills the condition $\nabla f(x) = \lambda \nabla c(x)$, where the Lagrange multiplier λ can have either sign.

When an inequality constraint $c(x) \geq 0$ is active (Figure 5.2C), $c(x) = 0$. In this case, direction d satisfies the constraint (i.e. $c(x + d) \geq 0$) if $\nabla c(x) \cdot d \geq 0$, and the angle between d and $\nabla c(x)$ is less than or equal to $90°$. As the angle between d and $\nabla f(x)$ must be greater than $90°$, the only scenario where d does not exist is when $\nabla c(x)$ and $\nabla f(x)$ have the same direction, resulting in the contradiction $\nabla f(x) \cdot d \geq 0$ and $\nabla f(x) \cdot d < 0$. Thus, a point is a solution only when $\nabla f(x) = \lambda \nabla c(x)$, where λ is strictly positive.

Based on the above discussion, a constrained optimization problem with multiple equality and inequality constraints can be solved using the Lagrange multiplier method. Consider the Lagrangian function $\Lambda(x, \lambda) = f(x) - \sum_{i \in E \cup I} \lambda_i c_i(x)$, where E and I are the sets of equality and inequality constraints, respectively. A solution satisfies the following Karush-Kuhn-Tucker (KKT) conditions: (1) $\nabla_x \Lambda(x, \lambda) = 0$, (2) $c_i(x) = 0$ for all $i \in E$, (3) $c_i(x) \geq 0$ for all $i \in I$, (4) $\lambda_i \geq 0$ for all $i \in I$ and (5) $\lambda_i c_i(x) = 0$ for all $i \in I$. The

last condition (complementarity condition) implies that the Lagrange multiplier can be positive only when the constraint is active and is always zero when the constraint is inactive. Whether the solution is actually a minimum (as opposed to a maximum or a saddle point) is determined by using second derivatives.

The algorithm we will develop in the next section is very similar to a one-class support vector machine (SVM) [206]. An SVM is a classification method using a linear discriminant in a high-dimensional space. It is a special case of constrained optimization method. Let's illustrate it with a simple example, where the training instances $\mathbf{x}_i$ of the negative ($y_i = -1$) and the positive ($y_i = +1$) classes are separable. Let the equation of the separating hyperplane be $\mathbf{w} \cdot \mathbf{x} + b = 0$, where $\mathbf{w}$ is normal to the hyperplane, $\frac{|b|}{||\mathbf{w}||}$ is the perpendicular distance of the hyperplane from the origin and $||\mathbf{w}||$ is $\mathbf{w}$'s Euclidean norm. For each instance, one can write the inequality constraints of the training data as $\mathbf{x}_i \cdot \mathbf{w} + b \geq +1$ for $y_i = +1$ and $\mathbf{x}_i \cdot \mathbf{w} + b \leq -1$ for $y_i = -1$, or $y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0$ in a combined form. The instances for which the constraints are active (i.e. $y_i(\mathbf{x}_i \cdot \mathbf{w} + b) = 1$) are called support vectors. The margin is the distance between the hyperplanes of the support vectors of the negative and the positive classes (these hyperplanes are parallel to the separating hyperplane) and turns out to be $\frac{2}{||\mathbf{w}||}$. The SVM finds a hyperplane that maximizes the margin, i.e. minimizes its reciprocal, given the constraints. Hence, the Lagrangian is $\frac{1}{2}||\mathbf{w}||^2 - \sum_i \left(\lambda_i y_i (\mathbf{x}_i \cdot \mathbf{w} + b) - 1\right)$, where the

Lagrange multipliers are greater than 0 ($\lambda_i \geq 0$). This is a convex QP problem. Because examples of the negative class are generally not known when identifying transcription factor binding sites, we use a different objective function below. The general structure of the problem, however, is similar.

Armed with the above knowledge of constrained optimization problems, let's proceed with the formulation of the PhyloQPMEME problem.

5.4 Constrained Optimization Problem to Identify Conserved Sites

Constraints need to be added to the optimization problem due to the restriction on the binding energies of the sequences in the training set and due to the designation of the average binding energies of random sequences.

All experimentally validated sites and their orthologs need to have binding energies below (or equal to) the thresholds. Let $s^{a,d}$ denote a sequence in the training set corresponding to the known site *a* in species *d*. Thus, each training sequence introduces the constraint $E(s^{a,d}) = \sum_{i=1}^{\ell} \sum_{\alpha=1}^{4} \varepsilon_{i\alpha} s_{i\alpha}^{a,d} \leq -1$. Because training sequences are treated equally, the number of above constraints equals the number of unique training sequences.

This model is similar to a one-class SVM with non-separable data [206] and is shown graphically in Figure 5.3. The training sequences with binding energy exactly equal to the threshold determine the separating hyperplane in the sequence space and are thus like support vectors.

Even though we would like to constrain the binding energy of every sequence in the training set, we do not know if all orthologs of the experimentally validated sites are functional binding sites. We have seen in the last chapter that site loss occurs frequently due to the lack of selection pressure and, as a result, the orthologous sequence evolves according to the neutral rate. In our case study of NF-κB, we observed that some sequences had evolved quite far away from the consensus (e.g. GAGGGATCTG), and the hard constraint that the binding energy of such sequences had to be below the threshold resulted in a great number of false positives while identifying conserved sites (~1.8% of all possible unique sequences). Unfortunately, there is no principled way of removing such "erroneous" sequences from the training set.

PhyloQPMEME therefore uses a "soft margin" during training. It keeps the potentially erroneous sequences in the training set, and allows them to have binding energies above the threshold only at a cost. The constraint in the above "hard margin" PhyloQPMEME model is modified to $E(s^{a,d}) = \sum_{i=1}^{\ell}\sum_{\alpha=1}^{4} \varepsilon_{i\alpha} s_{i\alpha}^{a,d} \leq -1 + \xi^{a,d}$ in this new soft margin model, where $\xi^{a,d} \geq 0$ is a positive slack variable (or error) for each sequence. $\xi^{a,d} > 0$ for an erroneous training sequence. PhyloQPMEME penalizes such a sequence by adding a

positive penalty $C\xi^{a,d}$ to the objective function, which it is actually trying to minimize. The positive cost parameter C is chosen by the user. Higher C increases the penalty and thus reduces the number of erroneous training sequences. Due to the constraints described here, the optimization process tries to balance the implicit cost of restricting the binding energy of an erroneous sequence below the threshold with the penalty of allowing its binding energy to have a value above the threshold.

The average binding energies of random sequences also need to be specified because binding energies can take arbitrary values. The average binding energy at each position *i* is set to zero, adding the constraint $\sum_{\alpha=1}^{4} p_\alpha \varepsilon_{i\alpha} = 0$ for every position. It is not necessary to add this constraint for every species because p_α in each of these species is assumed to be the same (this is a reasonable assumption for our case study, as described in the next section).

Thus, the following Lagrangian equation needs to be optimized.

$$\min_{\mathbf{y},\boldsymbol{\varepsilon}} \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{C}\mathbf{y} - \boldsymbol{\mu}^{\mathrm{T}}\mathbf{y} + C\sum_{a,d}\xi^{a,d} + \sum_{a,d}\lambda^{a,d}\left(\sum_{i=1}^{\ell}\sum_{\alpha=1}^{4}\varepsilon_{i\alpha}s_{i\alpha}^{a,d} + 1 - \xi^{a,d}\right) - \sum_{a,d}\tau^{a,d}\xi^{a,d} \cdots$$
$$-\sum_{i=1}^{\ell}\nu_i\left(\sum_{\alpha=1}^{4} p_\alpha \varepsilon_{i\alpha}\right)$$

In this equation, $\lambda^{a,d}$, $\tau^{a,d}$ and ν_i are the Lagrange multipliers for the three types of constraints. The KKT conditions are as follows.

- ℓ linear equality constraints specifying the average binding energy at each position *i* is to be 0: $\sum_{\alpha=1}^{4} p_\alpha \varepsilon_{i\alpha} = 0$.
- Two types of linear inequality constraints due to the restrictions of the binding energies of the training sequences: $-1 + \xi^{a,d} - \sum_{i=1}^{\ell}\sum_{\alpha=1}^{4} \varepsilon_{i\alpha} s_{i\alpha}^{a,d} \geq 0$, $\xi^{a,d} \geq 0$. The number of each type of constrains equals the number of training sequences.
- The Lagrange multipliers corresponding to the inequality constraints are required to be non-negative: $\lambda^{a,d} \geq 0$, $\tau^{a,d} \geq 0$.
- The complementarity conditions: $\lambda^{a,d}\left(\sum_{i=1}^{\ell}\sum_{\alpha=1}^{4} \varepsilon_{i\alpha} s_{i\alpha}^{a,d} + 1 - \xi^{a,d}\right) = 0$, $\tau^{a,d}\xi^{a,d} = 0$.

The derivative of the Lagrangian with respect to $\xi^{a,d}$ gives $C = \lambda^{a,d} + \tau^{a,d}$, showing that these Largrange multipliers are bounded by the cost parameter C. This equation, along with the second complementarity condition, implies that $C > \lambda^{a,d}$ when the binding energy of a sequence is not above the threshold ($\xi^{a,d} = 0$).

The above problem is a constrained quartic optimization problem. The objective function is quartic – quadratic in both $\varepsilon_{i\alpha}$ (because the covariance matrix $\mathbf{C}$ is quandratic in $\varepsilon_{i\alpha}$) and $\mathbf{y}$. It can be solved using iterative quadratic programming (QP). Each iteration consists of two steps. In the first step, the binding energies $\varepsilon_{i\alpha}$ are estimated by QP while keeping $\mathbf{y}$ fixed. The global solution is found because the objective function is convex, the Hessian of the covariance matrix $\mathbf{C}$ is always positive semi-definite and the constraints are linear. In the second step, $\mathbf{y}$ is updated by the simple formula $\mathbf{y} = \mathbf{C}^{-1}\boldsymbol{\mu}$.

While the condition number of $\mathbf{C}$ can be used to assess whether this linear system is well-conditioned, the generalized inverse (or pseudoinverse) of $\mathbf{C}$ is used to calculate $\mathbf{y}$ as it avoids any problems associated with the possible singularity of $\mathbf{C}$. Constrained quadratic optimization is performed using the "fmincon" solver with the active set algorithm in MATLAB (version R2008a, The MathWorks, Inc.). The final output of the training procedure is the energy matrix.

We have so far focused on the training of the PhyloQPMEME model. We now move our attention to the scoring procedure for the identification of conserved sites.

5.5 Scoring Procedure

During scoring, PhyloQPMEME (i) identifies a putative site in the reference species, (ii) determines the binding energies of its orthologs while allowing for some misalignment, and (iii) counts the number of species in which it is conserved, i.e. its orthologs have binding energy below the threshold. The conservation score equals this count plus one (to account for the reference species).

In the first step, PhyloQPMEME scans all ℓ-mer windows (where ℓ is the site length) on both strands in a promoter sequence in the reference species and calculates the binding energies of the sequences in these windows with the help of the trained energy matrix. A sequence with binding energy below the threshold (-1) is a putative site. When sites overlap, only the one with the lowest binding energy is considered.

To determine if the putative site is conserved in a related species, PhyloQPMEME retrieves the sequence for that species from the alignment corresponding to the site and a certain number of flanking nucleotides on its either side. Flanking nucleotides allow the capture of the orthologous site even if the local sequence alignment is not accurate. However, the number of flanking nucleotides is smaller than the site length so as not to include the ortholog of a potential tandem site from the reference species. (For example, it was chosen to be seven for identifying conserved κB sites of length ten described in the next section.) PhyloQPMEME removes gaps from the retrieved sequence and counts the number of nucleotides. If this number is less than the site length, the site is considered to be "unaligned" and thus lost in this species. Otherwise, PhyloQPMEME scans all ℓ-mer windows in this sequence and calculates binding energies as described above. If the lowest binding energy is below the threshold, the corresponding sequence is assigned to be the conserved orthologous site. Or else, the site is considered to be aligned but lost in this species. Thus, a site is considered to be lost in another species when (i) it is not aligned or (ii) it is aligned but has binding energy above the threshold. After repeating this procedure for all the related species, the conservation score is calculated by counting the number of conserved sites in the orthologous set of the putative site.

This scoring scheme has the following shortcomings: (i) it does not have a built-in model to allow for site loss in divergent species, and (ii) it uses a hard threshold for classification, declaring sequences with binding energy just above the threshold as lost sites. Other scoring schemes can be used to overcome these shortcomings. Based on the

phylogeny of the considered species and the number of lost sites in the training set, a site loss rate can be calculated as a function of species divergence using a maximum likelihood method. Then a composite score that penalizes site loss in the inverse proportion of species divergence can be calculated. Alternatively, a composite score can be calculated based on the likelihoods of selection constraint loss at various branches of the phylogenetic tree. If the Fermi-Dirac distribution of occupancy probabilities is not approximated by a step function, sequences with binding energies just above the threshold may be considered to be partially conserved by constructing a scoring model based on their occupancy probabilities. Finally, a likelihood scoring model integrating site loss probabilities and occupancy probabilities can be generated. These scoring schemes, however, have their own set of problems. For example, the exact phylogeny is usually unknown, a simple relationship between site loss rate and divergence may not exist, the link between site loss rate and penalty amount in a composite score or between occupancy probability and the amount of partial conservation does not have strong theoretical underpinnings, and the training data may be too sparse to estimate the additional parameters. Moreover, these scoring schemes are expected to increase false positives. We have therefore decided to use the simple scoring scheme.

Leave-one-out cross-validation of training sequences is performed as follows. A training set consists of orthologous sets of known sites. For each orthologous set, PhyloQPMEME estimates the energy matrix by training on all the other orthologous sets in the training set and scores the orthologous set to determine its conservation.

5.6 Identification of κB Sites Conserved in Mammals

Now that we have explained the PhyloQPMEME model, we will describe a case study of its application. Let's recall from Chapter 1 that the binding sites of the transcription factor family NF-κB, called κB sites, are highly conserved [51]. With human chosen as the reference species for obvious reasons, we have used PhyloQPMEME to identify conserved human κB sites.

PhyloQPMEME required the following input: (1) related species for determining conservation, (2) alignment of human promoters with orthologous sequences in these species for locating conserved sites, (3) single-species probabilities and joint probabilities in each pair of the selected species associated with neutrally evolving promoter sequences for construction of the covariance matrix, (4) experimentally validated κB sites and their orthologs that comprise the training set and (5) the cost parameter for penalty assignment to erroneous training set sequences.

Mammals were deemed to be the appropriate choice of species for determining conserved sites. As we have discussed in the last chapter, the selected species should neither be too closely related nor be too divergent. Whereas closely related species conserve even non-functional sequences and cause false positives, divergent species fail to conserve even functional sites and result in false negatives. Non-primate mammals, with divergence from humans greater than 40 Ma (million years ago) but less than 210 Ma, appear to be the ideal choice for identifying conserved human sites [168]. More than 30% of the

human sequence is aligned with that of another mammal and many known human κB sites are conserved in mammals [140, 141]. Primates were also included in this case study because the covariance matrix of PhyloQPMEME appropriately gives them low weight. Non-mammals, on the other hand, were not used because their high divergence from humans allows conservation of few transcription factor binding sites. Less than 10% of the human sequence is aligned with that of a non-mammal, and less than 10% of the known human κB sites are conserved in a non-mammal.

The twelve mammals used in this case study are human (*Homo sapiens*), chimp (*Pan troglodytes*), rhesus macaque (*Macaca mulatta*), rat (*Rattus norvegicus*), mouse (*Mus musculus*), rabbit (*Oryctolagus cuniculus*), cow (*Bos taurus*), dog (*Canis familiaris*), armadillo (*Dasypus novemcinctus*), elephant (*Loxodonta africana*), tenrec (*Echinops telfairi*) and opossum (*Monodelphis domestica*). The primary reason for this particular choice of mammals was the ready availability of the sequence alignment of their genomes with the human genome. Moreover, these species cover a wide range of divergence within the mammalian class. They represent all four super-orders of primates – Euarchontoglire (human, chimp, rhesus macaque, rat, mouse and rabbit), Laurasiatheria (cow and dog), Xenarthra (armadillo) and Afrotheria (elephant and tenrec) – while opossum is a marsupial (see Figure 4.1D).

We defined a promoter as the region from 800 bp upstream to 200 bp downstream of the transcription start site (TSS) of a gene. More than 85% of the known human κB sites fall into this region. Inclusion of additional regions is thus expected to yield few more sites.

On the flip side, the covariance matrix, whose terms are calculated based on the nucleotide composition of promoters, may change substantially with the inclusion of additional regions due to different nucleotide compositions in different parts of the genome. For example, promoter regions are GC rich, whereas intergenic regions are AT rich.

PhyloQPMEME takes the alignments of promoters with orthologous sequences as an input. While alignment of multiple sequences is a complex problem, a number of good algorithms are available in the literature [172, 207, 208], and hence PhyloQPMEME does not focus on this problem. However, its success depends on the accuracy of the alignments.

17-way multiple sequence alignments of the promoters corresponding to all human reference sequences (RefSeq Release 19 [139]) with the sequences of other vertebrates were retrieved from the University of California Santa Cruz (UCSC) genome bioinformatics site (http://genome.ucsc.edu/) [140, 141]. Human sequences in these multiple sequence alignments correspond to human assembly hg18, NCBI Build 36.1. Sequences of promoters corresponding to all human reference sequences were also retrieved from the UCSC site. Duplicate entries were removed. Moreover, only mammalian sequences in the multiple sequence alignments were retained.

Nucleotide probabilities in each species and joint probabilities of two nucleotides in all pairs of species corresponding to neutrally evolving sequences are required in the

covariance matrix. They were calculated using the aligned promoter sequences retrieved above.The assumption made by PhyloQPMEME that the single-species distributions of nucleotides in all the considered species are identical seems to be reasonable. These distributions were indeed found to be similar. They are close to the vector $[p_A, p_C, p_G, p_T] = [0.22, 0.28, 0.28, 0.22]$, which was used as the single-species distribution for all species. Furthermore, as one would expect, the joint probability of the same nucleotide in two species decreases as the divergence between the species increases. For example, it is ~0.98 between human and chimp, ~0.94 between human and rhesus macaque, in the range 0.7-0.75 between human and other placental mammals, and ~0.63 between human and opossum.

The training set consists of 50 experimentally validated (known) κB sites from TRANSFAC 9.3 [54, 56-59] that are present in the promoters defined above, as well as the mammalian orthologs of these known sites according to the UCSC multiple sequence alignments. 43 of the known sites are from human and the other seven are from mouse. An orthologous sequence the same length of a κB site was retrieved only if it was available from the alignment. A total of 473 sequences were obtained, 114 of which are unique.

The final input, i.e. the cost parameter, required an elaborate procedure which we describe in the next section.

5.7 Determination of the Cost Parameter

The cost parameter needs to be set by considering the trade-off between false positives and false negatives. When the value of the cost parameter is high, the penalty for allowing the binding energy of potential erroneous training sequences to be above the threshold is high. Therefore, many erroneous training sequences are forced to have binding energy below the threshold, creating a lenient threshold. Even though such a model correctly classifies most genuine sites (few false negatives), the probability that a random sequence has binding energy below (or equal to) the threshold is substantial, resulting in many false positives. The extreme case is the cost parameter value of positive infinity, when all training sequences are forced to have a binding energy below the threshold. This is the hard margin PhyloQPMEME noted above. In contrast, a low value of the cost parameter creates a stringent threshold. Even though it restricts the number of false positives, it misclassifies many genuine sites (many false negatives) due to the low penalty of allowing binding energies to be above the threshold.

To determine the cost parameter, we trained PhyloQPMEME on all sites in the training set using different cost parameters (Figure 5.4). We then scored the training sites as well as all possible unique sequences of length ten (4^{10} sequences) with the trained energy matrix and determined their binding energies. The number of training sites with binding energy above the threshold (an approximate measure of potential false negatives) decreases precipitously as the cost parameter increases from 0.4 to 0.6 and flattens out above 0.6 (blue line in Figure 5.4A). On the other hand, the number of all possible unique

sequences with binding energy below the threshold (an approximate measure of potential false positives) increases substantially as the cost parameter increases above 0.8 (Figure 5.4B). Thus, the cost parameter range of 0.6-0.8 appears to be reasonable. We chose 0.6 as the cost parameter for our model as it appears to have a good balance between potential false negatives (~20%) and potential false positives (0.2%). Even though the potential false negative rate appears to be high, we have to keep in mind that experimentally not validated orthologs of the known sites, in addition to the known sites, were used to calculate it. Thus, it may be misleading. In fact, all the known sites had binding energy below the threshold, while all the training sequences that had binding energy above the threshold were orthologs of the known sites. Thus, the potential false negatives seen above may have been lost sites. Incidentally, Figure 5.4 shows that using cost parameter values above 1.2 is equivalent to training with a hard margin PhyloQPMEME.

The energy matrix obtained after training PhyloQPMEME on the above data is shown in Figure 5.5. As expected, G has the lowest binding energies at positions 1-3 and C has the lowest binding energies at positions 9-10 (see Chapter 1). While T has the lowest binding energies at positions 6-7, A and G have comparable low binding energies at position 4, and C and T have comparable low binding energies at position 8. Nucleotide at position 5 does not appear to be important for specific binding to the transcription factor, as binding energies of all nucleotides at this position are close to zero.

5.8 Conservation and Loss of κB Sites

We performed leave-one-out cross-validation on the training set to assess the performance of PhyloQPMEME as well as to understand the characteristics of κB site conservation and loss.

All known human and mouse κB sites have binding energies below the threshold, showing that PhyloQPMEME is quite sensitive. Moreover, about one third of the κB sites are conserved in at least eleven of the twelve mammals in this case study, and about two thirds of the κB sites are conserved in at least nine mammals (Figure 5.6). This high conservation rate validates our approach of using related species for identifying conserved κB sites.

As expected, the site loss rate increases with species divergence from human (Figure 5.7). However, the correlation between the site loss rate and the divergence is not exact. Species can be roughly divided into three groups based on the fraction of lost sites. (1) Less than one eighth of the sites lost: For example, primates (chimp, rhesus macaque). All sites, including the orthologs of the known mouse sites, are conserved in human. (2) About one fourth of the sites lost: For example, the rest of the Boreoeutheria, which consists of the placental mammal super-orders Euarchontoglire (e.g. rat, mouse and rabbit, in addition to primates) and Laurasiatheria (e.g. cow and dog). Site loss rates in the non-primate Euarchontoglires and Laurasiatheria appear to be similar probably because the divergence between human and the rest of the Euarchontoglires considered in this

analysis is not much less than the divergence between human and Laurasiatheria (see Figure 4.1D). (3) Between one third and one half of the sites lost: This includes the rest of the mammals. Although the site loss rate is expected to be proportional to the divergence of mammals from humans, it is not always the case. Using this criterion, the marsupial opossum, for example, should have a much larger site loss rate than all placental mammals. But this is not true. The species in which the most sites are lost is elephant, not opossum, in spite of a better overall alignment of human promoters with the elephant's than with the opossum's sequences. The site loss rate does not appear to be a simple function of divergence.

Figure 5.7 also shows that when a site is lost, that is when selection constraint is no longer applicable, a sequence often changes so much that it is difficult to align with a conserved site. Majority of the lost κB sites cannot be aligned with the conserved sites. Only a few lost sites align but have binding energy above the threshold.

We also identified conserved κB sites in the promoters of all genes in the human genome. The number of mammals in which a predicted κB site is conserved increases from one to three and then drops offs exponentially (Figure 5.8). When a predicted κB site is conserved in three mammals, the species with conservation are usually human, chimp and rhesus macaque. Therefore, the peak of conservation at three species can be attributed to the high conservation of non-functional sequences in primates, and most of these predicted κB sites are false positives. Finally, 302 κB sites are conserved in all mammals and 884 are conserved in eleven mammals.

What is a reasonable threshold for the conservation score? Or, how many species does a predicted κB site need to be conserved in to enable us to declare that it is a real κB site? According to Figure 5.9A, as the threshold decreases from twelve to nine, about two thirds of the known sites are recovered (declared to be conserved; also seen in Figure 5.6.) However, as this threshold decreases further, the gain in true positives (recovered known sites) is quite small compared to the additional number of predicted sites, most of which are expected to be false positives. Figure 5.9B is more informative. It shows that the gain in true positives, in comparison with the number of predicted κB genes (genes whose promoters contain the predicted κB sites), decreases noticeably if the conservation score threshold is reduced below ten. The figure also shows that while a sizable number of κB sites are conserved in a small number of species, their identification using conservation as the sole classification criterion is made quite difficult in view of the occurrence of a large number of false positives.

Figure 5.10 shows that most of the 1186 predicted κB sites conserved in 11 or more mammals lie close to the TSS, and that their number decreases in the regions further upstream away from the TSS. This fact is in agreement with the observed distribution of the known κB sites, discussed at length in Chapter 3 and supported by a number of studies [68, 142, 143]. We are therefore confident that many of the sites conserved in 11 or more mammals are indeed real functional κB sites.

5.9 Site Energy is Better Conserved than Site Sequence

We began this chapter with the theme that a site's binding energy rather than its sequence needs to be conserved to maintain its fitness. In one of the key results of this chapter, we show that this indeed is the case with κB sites.

To start out, we calculated the Hamming distance between each pair of orthologous sites in the orthologous sets of the 302 predicted κB sites that are conserved in all mammals. Hamming distance between two sequences is simply the number of positions where their nucleotides differ. As expected, the Hamming distance between orthologous site pairs has an exponential-like distribution in which most pairs are identical (Figure 5.11). We then calculated the maximum pairwise Hamming distance in each orthologous set (i.e. the maximum of the Hamming distances calculated for all site pairs in an orthologous set.)

The distribution of the maximum pairwise Hamming distance (MPHD) in orthologous sets is striking (Figure 5.12). The MPHD within an orthologous set is often quite high. The largest fraction of the orthologous sets has the MPHD of five. Over 36% of the orthologous sets have the MPHD of five or more. In other words, over one third of the orthologous sets have at least one pair of sites that has different nucleotides in at least half the positions (the κB site length is ten). Four orthologous sets have the MPHD of seven! For example, the human site in the promoter of the gene DTNA has the sequence GCGAAATCCC, whereas its orthologous site in cow has the sequence TGGGCTTTCG.

Despite these vastly different sequences, the binding energies of the human and cow sites are -1.04 and -1.08, respectively.

These data clearly demonstrate that the selection pressure during site evolution works on binding energies rather than on sequences, at least for κB sites. They also highlight the superiority of evolutionary models of site binding energies, like the PhyloQPMEME model, over evolutionary models of site sequences described in the last chapter. An evolutionary model of site sequences will have a great deal of difficulty in explaining the fact that sites in which seven out of ten nucleotides differ are actually conserved.

After showing the merits of the PhyloQPMEME model, we now discuss the biological significance of conserved κB sites as identified by PhyloQPMEME.

5.10 Biological Insights from Conserved NF-κB Targets

Because the conservation threshold of 11 recovers one third of the known sites while predicting a reasonable number of NF-κB target genes (972 genes corresponding to 1186 sites; Figure 5.9), we decided to ascertain the biological significance of the resulting gene set. While we are aware that this gene set misses many NF-κB target genes, we believe that most genes included in the gene set are true positives. We identified cellular pathways, biological functions and diseases in which these putative NF-κB targets were over-represented by using Ingenuity Pathways Analysis (Ingenuity® Systems, www.ingenuity.com) and DAVID [145-147].

The NF-κB targets in this gene set fall into three categories: (i) known targets in the pathways regulated by NF-κB, (ii) previously unknown targets in the NF-κB pathways and (iii) targets in the pathways not known to be regulated by NF-κB (Table 5.1). PhyloQPMEME predicted many well-known NF-κB targets, such as chemokines, integrins and interleukins, associated with the antigen presentation, glucocorticoid receptor signaling, G-protein coupled receptor signaling, MAP kinase signaling and B cell receptor signaling pathways, among others. Because NF-κB influences the inflammation-fibrosis-cancer axis in liver [209], a lot of genes involved in hepatic fibrosis are also highlighted. The Wnt-beta catenin signaling cascade plays quite an important role in many aspects of development. While the role of NF-κB in this pathway is still emerging [210], PhyloQPMEME has pinpointed several genes, including WNT5A and WNT8B, as NF-κB targets. As discussed in Chapter 3, NF-κB's role in the regulation of ubiquitination is not well characterized. The targets identified by PhyloQPMEME may help elucidate this role. Moreover, the predicted NF-κB targets in the actin cytoskeleton signaling may shed light on the exact role of NF-κB in this particular pathway. Just like the HMM, PhyloQPMEME has also identified several potential targets responsible for xenobiotic metabolism. Interestingly, our analysis also suggests roles for NF-κB in various nervous system pathways, including circadian rhythm signaling, synaptic long term potentiation and dopamine receptor signaling.

It remains to be determined if these and other genes identified using PhyloQPMEME are genuine NF-κB transcriptional targets. If confirmed, they could yield important new

insights into the roles of NF-κB in numerous pathways and associated diseases. The comparison of the genes predicted by the HMM and PhyloQPMEME is discussed in the next chapter.

5.11 Summary

We have successfully built a composite model integrating binding energies from biophysics with evolutionary conservation to identify transcription factor binding sites. PhyloQPMEME has a number of advantages and unique features:

- Consideration of multiple species substantially reduces the number of false positives.
- Modeling evolution of a site's binding energy is more appropriate than modeling evolution of its sequence because binding energy of a site is conserved better than its sequence.
- PhyloQPMEME uses occupancy probability as the discriminant function, which itself provides a number of benefits including a clear biophysical interpretation.
- Occupancy probability can be linked to evolutionary fitness in a logical manner, especially if a transcription factor functions during one cellular state.
- Based on the thermodynamics principles, PhyloQPMEME assumes the Fermi-Dirac distribution of occupancy probability, which is applicable in any range of transcription factor concentrations.
- This occupancy probability distribution also offers a non-arbitrary threshold, unlike most other methods in vogue.

- Because the covariance matrix captures the correlations between the binding energies of orthologous sequences, PhyloQPMEME places no restrictions on the number of species or the topology of the phylogenetic tree.
- Unlike the other supervised learning methods, PhyloQPMEME takes the evolutionary relationships into account during the training procedure. While the covariance matrix deals with neutrally evolving sequences, binding energies of known sites and their orthologs are used explicitly as constraints in the optimization problem for estimating binding energies.
- PhyloQPMEME uses a soft margin during training, which enables the incorporation of sequence information of experimentally unverified orthologs of known sites, while allowing for the loss of selection constraint in some of them.

PhyloQPMEME also has the following limitations. Its success depends on the accuracy of the sequence alignments provided to it. Moreover, its scoring scheme does not incorporate site loss, and uses a hard threshold that may potentially misclassify genuine sites with binding energy just above the threshold as lost sites. It also makes the following assumptions:

- Binding energies at different positions in a site are independent and additive. This assumption is quite common, and is shown to be very reasonable [76].
- Binding energy of a nucleotide at a position in a site ($\varepsilon_{i\alpha}$) is identical in all species, which is expected to be true if the transcription factor is highly conserved (as is the case of NF-κB in mammals).

- Binding energies of orthologous sets have a multivariate normal distribution and the Fermi-Dirac distribution of the occupancy probability can be approximated by a step function. Both these assumptions are valid only if the site length is substantially greater than one (e.g. ten in the case of κB sites).
- The distribution of neutrally evolving nucleotides is identical in the selected species, which again is a reasonable assumption for species that are not too divergent (e.g. mammals).
- The set of neutrally evolving sequences can be approximated by all possible random sequences because a very small proportion of all possible sequences are functional and thus under selection pressure.

Through the application of PhyloQPMEME, we demonstrated that (i) the majority of functional sites are indeed conserved in many species, (ii) the site loss increases roughly with species divergence, (iii) the sequences of lost sites often change so much that their alignment with conserved sites becomes difficult, and (iv) the site binding energy rather than its sequence is under selection pressure. Moreover, pathway analysis shows that the predictions made by PhyloQPMEME are biologically significant.

Table 5.1: Selected pathways, functions and diseases enriched with NF-κB targets predicted by PhyloQPMEME.

Selected cellular pathways, biological functions and diseases in which our predicted NF-κB targets were over-represented are shown. The associated predicted NF-κB targets are represented by official human gene symbols. Genes containing κB sites with the conservation score of eleven or more were used in this analysis. Genes known in the literature to be regulated by NF-κB (although not necessarily directly) [18] are denoted with *.

Pathway/Function/Disease	Gene Symbols
Wnt/beta-catenin Signaling	GJA1, GSK3B, WNT5A, TLE3, PPP2R5C, CSNK1A1, CSNK2A1, SOX17, SOX4, SOX12, PPP2CB, PPP2R1A, PPP2R2B, MARK2, SOX14, WNT8B, RARA, TGFB3, NLK
PI3K/AKT Signaling	NFKB2*, YWHAQ, GSK3B, HSP90AA1*, PPP2R5C, PIK3CB, NFKBIA*, YWHAE, PPP2CB, NRAS, SHC1, PPP2R2B, PPP2R1A, FOXO3, MAP3K8
Chemokine Signaling	CCL4*, CCL2*, PPP1R12A, NRAS, CCL5*, MAPK11, PTK2B, PPP1CC, PLCB2, CALML5, CAMK2A
G-Protein Coupled Receptor Signaling	NFKB2*, PDE1A, ADCY3, DUSP6, RASA1, CAMK2A, RGS14, PDE4D, PIK3CB, NFKBIA*, PDE11A, GRM2*, CREB1, NRAS, SHC1, CREB5, PTK2B, PLCB2, PDE7A*, MAP3K8
Hepatic Fibrosis / Hepatic Stellate Cell Activation	MYL1, NFKB2*, EGF, VCAM1*, CCL5*, CTGF, CYP2E1*, CCL2*, CXCL1*, MYH6, IGF1, IL4, EDN1*, TGFB3, IL6*
Glucocorticoid Receptor Signaling	HSPA1A, VCAM1*, IL13*, CCL5*, MNAT1, HSP90AA1*, MAPK11, TAF9, HSPA1L, POU2F1, PIK3CB, CCL2*, CXCL1*, NFKBIA*, POU2F2, IL4, CREB1, NRAS, SHC1, GTF2A2, SELE*, TGFB3, IL6*, IL2*
Antigen Presentation Pathway	HLA-G*, CALR, HLA-C, PSMB9*, HLA-B*, HLA-F, HLA-DMB
B Cell Receptor Signaling	NFKB2*, GSK3B, MAPK11, CALML5, CAMK2A, PIK3CB, POU2F2, NFKBIA*,

	CREB1, NRAS, CREB5, SHC1, EGR1*, MAP3K8, PIK3AP1*
Actin Cytoskeleton Signaling	MYL1, MSN, EGF, NCKAP1L, ACTN1, VCL, FGF7, FGF18, PIK3CB, FGF14, PPP1R12A, MYH6, PAK3, GRLF1, NRAS, TTN, ARPC5, SHC1, PPP1CC, PAK4
Circadian Rhythm Signaling	BHLHB3, CREB1, CREB5, NR1D1, GRIN2A*
Synaptic Long Term Potentiation	PPP1R12A, GRIA3, GRM2*, CREB1, NRAS, CREB5, PPP1CC, PLCB2, CALML5, GRIN2A*, CAMK2A
Protein Ubiquitination Pathway	PSMC6, PAN2, UBE2D3, PSME2*, HLA-C, PSMB9*, USP15, PSMD3, CDC20, HSP90AA1*, PSMB10, USP48, HLA-B*, PSMD1, USP2, PSMB3
Dopamine Receptor Signaling	PPP1R12A, ADCY3, PPP2CB, SPR, PPP2R2B, PPP2R1A, PPP1CC, PPP2R5C
Xenobiotic Metabolism Signaling	NFKB2*, NR1I3, HSP90AA1*, MAPK11, PPP2R5C, GSTM3, CAMK2A, PIK3CB, ALDH1L2, AIP, ALDH6A1, CYP2C19, PPP2CB, NRAS, PPP2R2B, PPP2R1A, AHR, NDST1, IL6*
Rheumatoid Arthritis	ADAMTS4, CCL2*, CCL4*, CCL5*, CCL19*, CD68, CD69*, CD86*, CFB*, CSF3R, CXCL1*, CXCL2*, CXCL3*, CXCL5*, CXCL6*, CXCL9*, CXCL10*, EIF1B, FOSB, HLA-C, HLA-DMB, HLA-G*, HSPA1A, HSPB8, IGF1, IL6*, IL9*, IL13*, LCP1, LTA*, LTB*, MAPK11, MME, NFKBIA*, NR4A2*, OSM, PDE4D, PSMB9*, PTPN22, RUNX1, STAT4, TNFSF4, TPM2, USP15, ZNF143

Figure 5.1: Illustration of the basic idea of PhyloQPMEME.

The illustration is shown for two species. E_1 and E_2 are the binding energies of orthologous sequences (orthologous set) in the two species. μ_1 and μ_2 are the corresponding thresholds. A sequence with binding energy below the threshold has occupancy probability of one. The shaded area contains orthologous sets such that both sequences in the set have binding energies below the corresponding thresholds and hence are occupied. PhyloQPMEME determines the binding energy of each nucleotide at each position so as to minimize the probability that orthologous sets of random sequences fall into the shaded area, while confining orthologous sets of all known sites to the shaded area. This is shown graphically by the black arrows compressing the ellipse or equivalently by the black arrows moving the shaded area away from the center. PhyloQPMEME assumes that the binding energies of orthologous sets of random sequences have a multivariate normal distribution and maximizes $\boldsymbol{\mu}^T\mathbf{C}^{-1}\boldsymbol{\mu}$ subject to constraints, where $\boldsymbol{\mu}$ is the vector of threshold binding energies and $\mathbf{C}$ is the covariance matrix of binding energies of orthologous sets of random sequences. Consideration of multiple species reduces false positives because the probability that the orthologs of a random sequence have energies below the thresholds is miniscule, even though it may have energy less than the threshold by sheer chance. The joint distribution of the energies has the shape of an ellipse unaligned with the axes, indicating the correlation in the binding energies of orthologous sequences. The blue dots represent the binding energies of random sequences the same length as the known sites. The red plus signs denote the binding energies of the known sites.

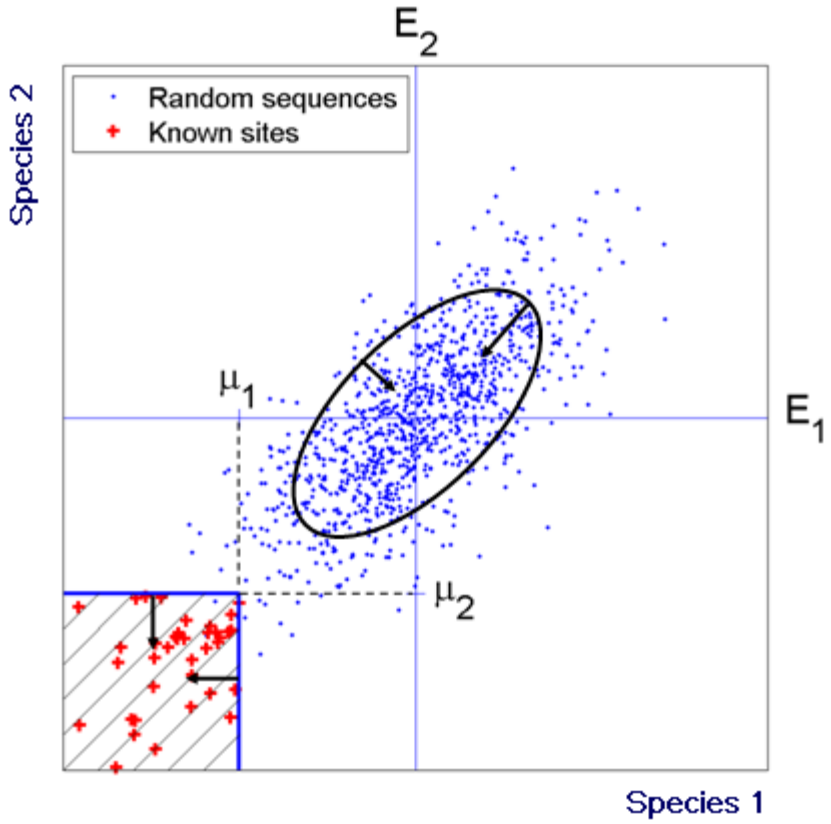

E2
Species 2
Random sequences
Known sites
μ1
E1
μ2
Species 1

Figure 5.2: Explanation of the constrained optimization problem.

A. Any direction d in the open half-space opposite of the gradient $\nabla f(x)$ (i.e. with angle greater than $90°$) decreases the value of the objective function $f(x)$. Many such directions are shown. $\nabla f(x)$ in parts (B) and (C) is assumed to be in the same direction as in part (A) and equal in the entire space.

B. Equality constraint: feasible region is assumed to be a circle (e.g. $c(x) \equiv x_1^2 + x_2^2 - 1 = 0$). At any point A, one of the two directions lies in the open half-space opposite of $\nabla f(x)$ and thus decreases the function value. Constraint gradient $\nabla c(x)$ at point B is in the same direction as $\nabla f(x)$ and $\nabla c(x)$ at point C is in the opposite direction. Direction d does not exist at either point, and both points are potential minima. Thus, a local solution satisfies the condition $\nabla f(x) = \lambda \nabla c(x)$, where the Lagrange multiplier λ can be either positive or negative.

C. Inequality constraint: feasible region is assumed to be the interior of a circle (shaded region; e.g. $c(x) \equiv 1 - x_1^2 - x_2^2 \geq 0$). The constraint is active at points D and E. At point D, $\nabla c(x)$ and $\nabla f(x)$ are in the same direction and d does not exist; D is a local solution. At point E, $\nabla c(x)$ and $\nabla f(x)$ are in the opposite direction and d exists as shown; thus, E is not a local solution. At any point F where the constraint is inactive (inside the feasible region), d exists as long as $\nabla f(x)$ is not zero. Points with a star are local solutions.

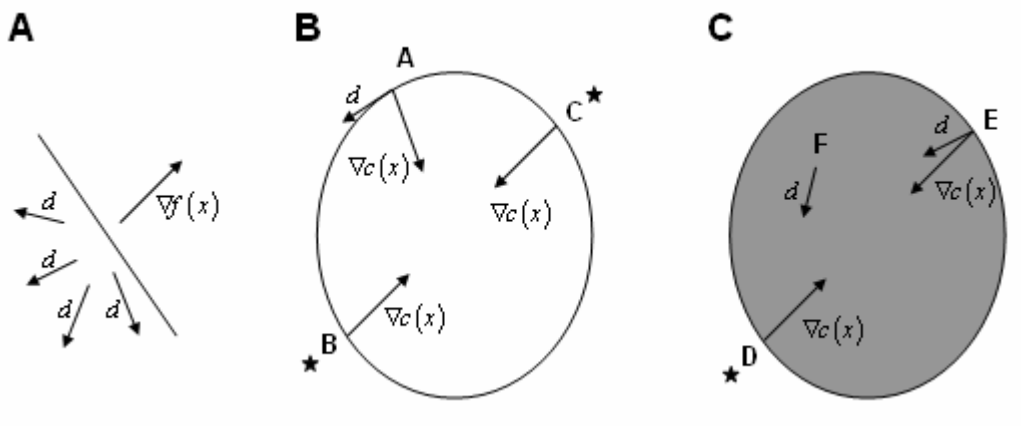

A
B
C
$\nabla f(x)$
d
d
d
d
A
d
C
$\nabla c(x)$
$\nabla c(x)$
$\nabla c(x)$
B
F
d
d
E
$\nabla c(x)$
$\nabla c(x)$
D

Figure 5.3: Classification in sequence space.

A sequence of length ℓ (site length) can be represented by sequence vector $\mathbf{S}$ of length 4ℓ such that each element $s_{i\alpha}$ equals one if the sequence has nucleotide α at the ith position and zero otherwise. Such sequence vectors occupy the surface of a sphere in a 4ℓ-dimensional space. PhyloQPMEME constructs a separating hyperplane or threshold in this space such that sequences on its one side have binding energy below the threshold and those one the other side have binding energy above the threshold. Because only known sites are usually available for training, PhyloQPMEME estimates binding energies of individual bases in such a way as to obtain the fewest random sequences on the side of the separating hyperplane corresponding to binding energy below the threshold. This model is similar to a one-class support vector machine (SVM). The known sites with binding energy equal to the threshold are responsible for defining the separating hyperplane and are thus like support vectors. Sequence vectors corresponding to known sites with binding energy below the threshold, known sites with binding energy equal to the threshold and random sequences with binding energy above the threshold are shown with black, red and blue arrows, respectively.

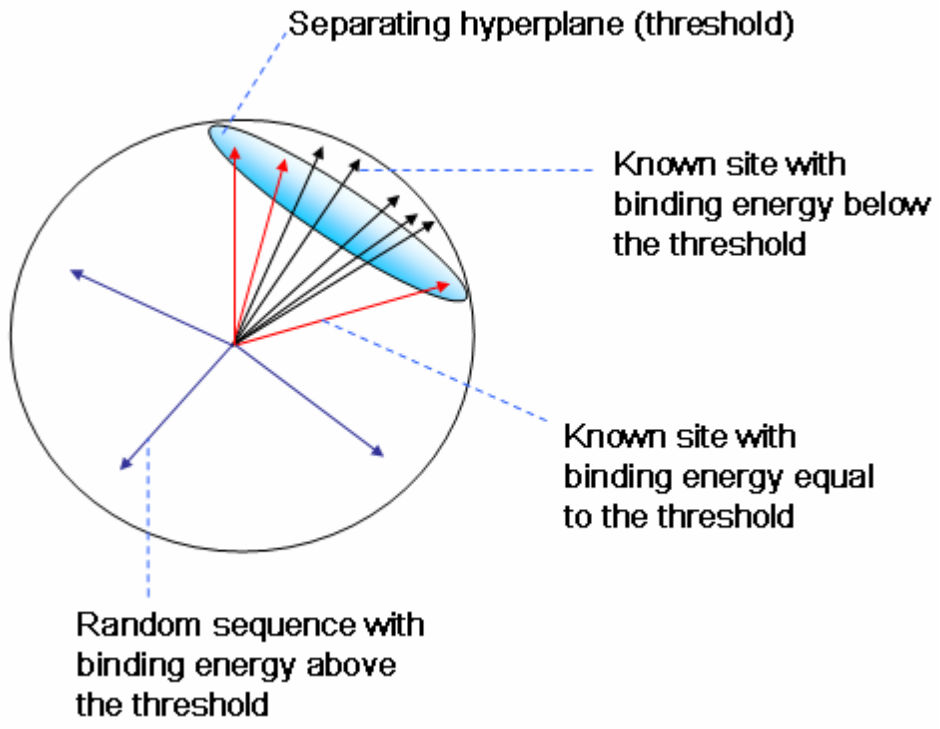

Separating hyperplane (threshold)
Known site with binding energy below the threshold
Known site with binding energy equal to the threshold
Random sequence with binding energy above the threshold

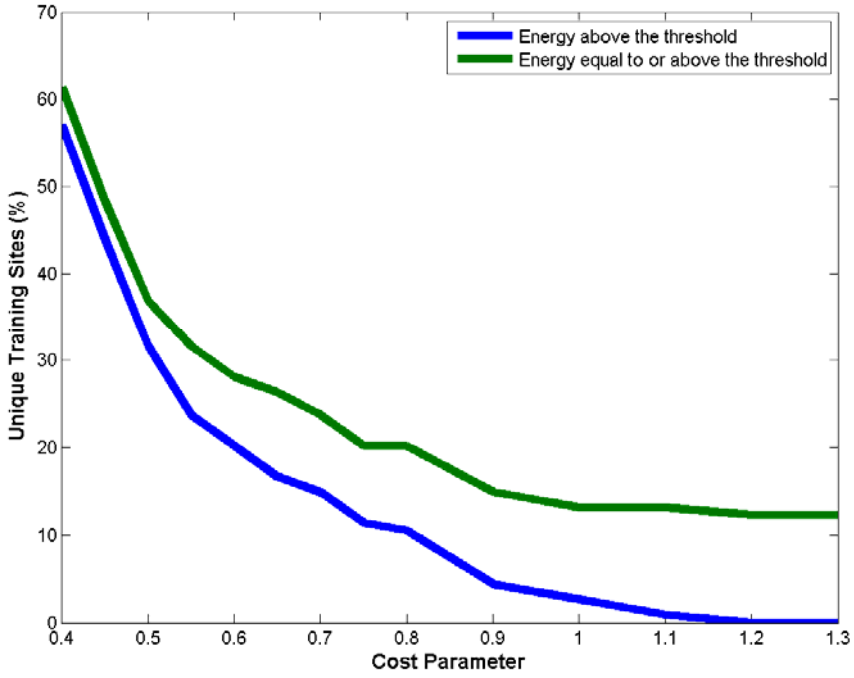
Figure 5.4: Binding energy distribution as a function of the cost parameter.

A. The fraction of unique training sites with binding energy above the threshold (an approximate measure of potential false negatives) decreases as the cost parameter used to train PhyloQPMEME increases. The fraction decreases sharply when cost parameter increases from 0.4 to 0.6 and then flattens out. It reaches zero when the cost parameter is 1.2. The fractions with binding energy above the threshold and with binding energy equal to or above the threshold are shown with the blue and green lines, respectively. The fraction of unique training sequences with binding energy exactly equal to the threshold (difference between the y positions of the blue and green lines), which is responsible for determining the classification boundary, is ~7-13% when the cost parameter value is above 0.6.

B. The fraction of all possible unique sequences (4^{10} sequences) with binding energy above the threshold (an approximate measure of potential false positives) increases with the cost parameter. Its slope is high in the cost parameter range of 0.8-1.2, after which it is flat because the hard margin PhyloQPMEME limit is reached.

Based on parts (A) and (B), the cost parameter range of 0.6-0.8 appears to have a good tradeoff between potential false negatives and potential false positives.

A


Energy above the threshold
Energy equal to or above the threshold
Unique Training Sites (%)
Cost Parameter
0
10
20
30
40
50
60
70
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3

B

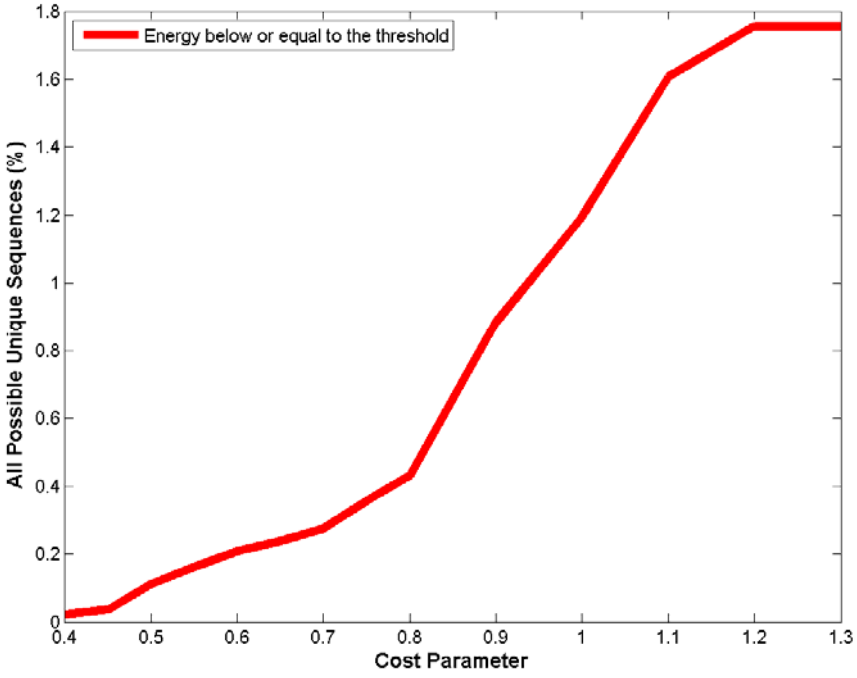

Energy below or equal to the threshold
All Possible Unique Sequences (%)
Cost Parameter
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3

Figure 5.5: Trained energy matrix.

The negative values of the trained binding energies of nucleotides at each position of a κB site are shown. At each position, nucleotides with high negative energies, or high positive values in the figure, have strong affinities to the DNA-binding domain of the transcription factor. Thus, the nucleotides with the strongest affinities are G at positions 1-4, T at positions 6-8 and C at positions 9-10.

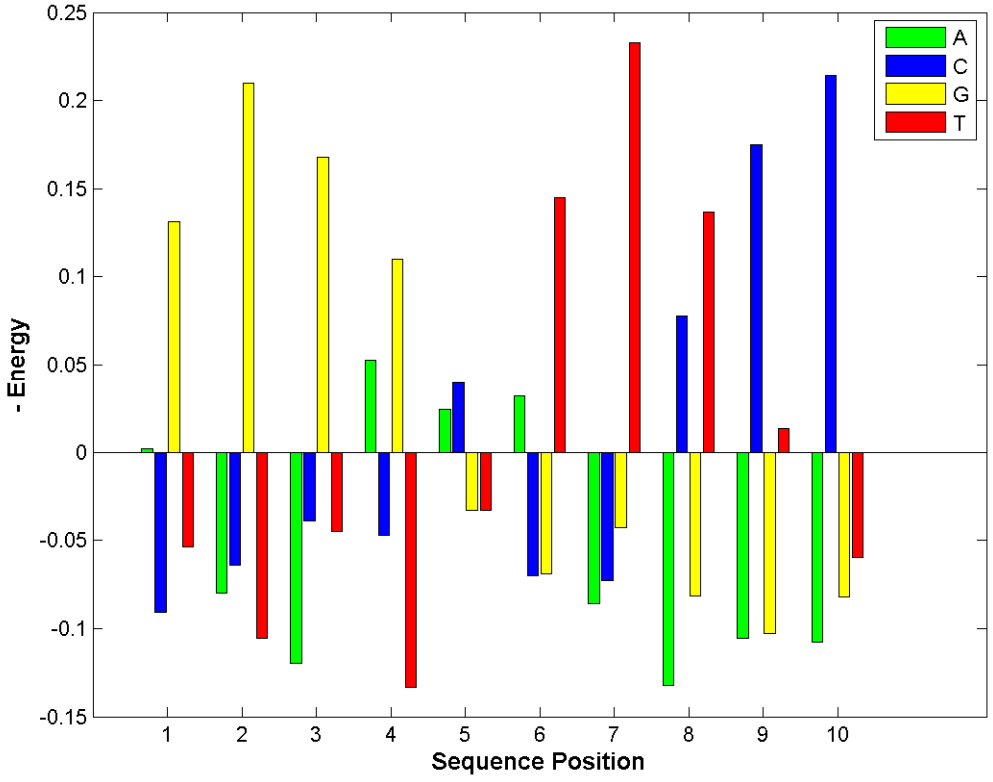


Figure 5.6: Conservation of the known κB sites.

A histogram of the number of considered mammals in which the known κB sites are conserved is shown. While about one third of the κB sites are conserved in at least eleven species, about two thirds of the κB sites are conserved in at least nine species. The fraction of sites conserved in a particular number of species is displayed at the top of the corresponding bar. The 50 known κB sites present in the PhyloQPMEME training set were considered. Conservation was determined by PhyloQPMEME using the leave-one-out cross-validation procedure described in Section 5.5.

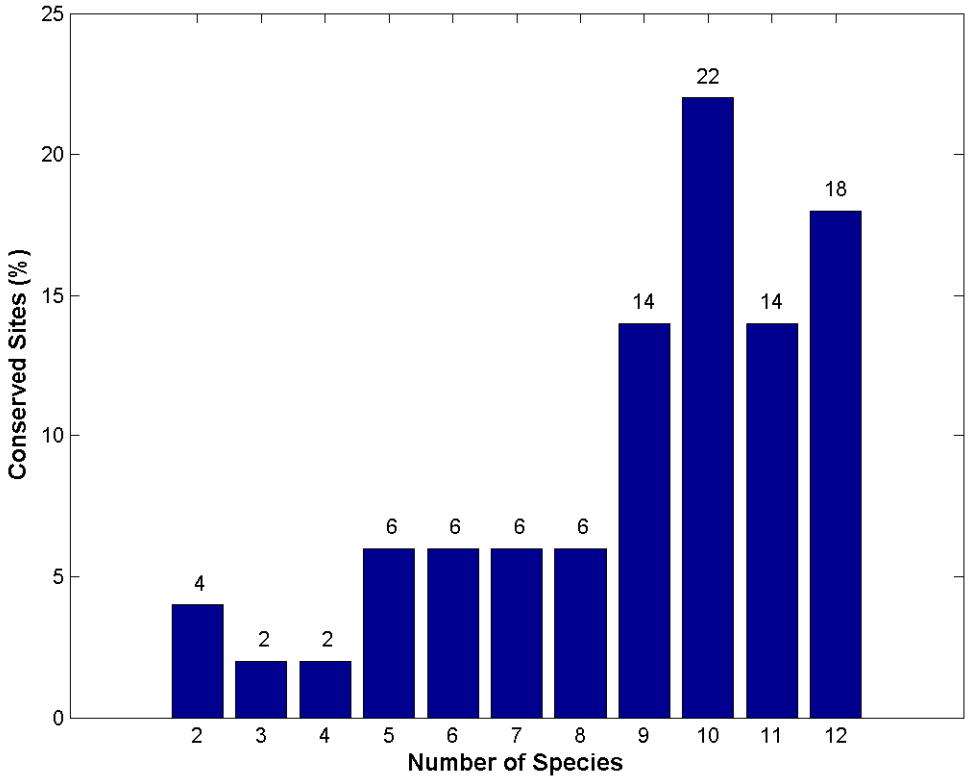


Figure 5.7: Species-wise loss rates of the known κB sites.

The fraction of the known κB sites lost in each species is shown. A site is considered to be lost (i.e. not conserved; green line) in a species (i) if it cannot be aligned (blue line) or (ii) if the binding energy of the aligned sequence is above the threshold. The majority of the lost sites fall in the first category. Site loss rate increases with species divergence from humans, although the correlation is only approximate. Primates (chimp, rhesus macaque) have the lowest site loss rates of less than 15%, the rest of the Boreoeutheria (consisting of the placental mammal super-orders Euarchontoglire and Laurasiatheria; rat through dog in the figure) have site loss rates of ~20-30%, and the rest of the mammals have site loss rates of ~35-50%. Site loss rate in human is not shown because all sites, including the orthologs of the known mouse sites, are conserved in human. The 50 known κB sites present in the PhyloQPMEME training set were considered. Loss rate was determined by PhyloQPMEME using the leave-one-out cross-validation procedure described in Section 5.5.

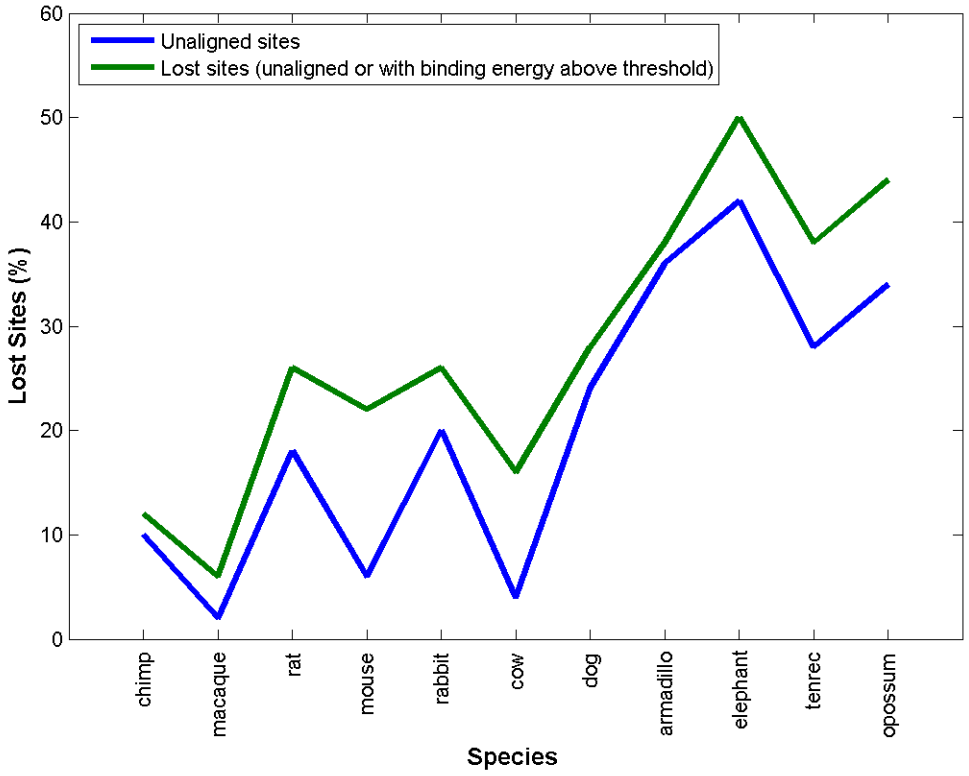

Unaligned sites
Lost sites (unaligned or with binding energy above threshold)
Lost Sites (%)
0
10
20
30
40
50
60
chimp
macaque
rat
mouse
rabbit
cow
dog
armadillo
elephant
tenrec
opossum
Species

Figure 5.8: Conservation of the predicted κB sites.

A histogram of the number of considered mammals in which the predicted κB sites are conserved is shown. PhyloQPMEME predicted the conserved κB sites by performing a genome-wide search. The reason for conservation of the largest fraction of predicted κB sites in three mammals is that most of these sites are false positives which, like many other non-functional sequences, are conserved in human, chimp and rhesus macaque. The number of sites conserved in a particular number of species is displayed at the top of the corresponding bar.

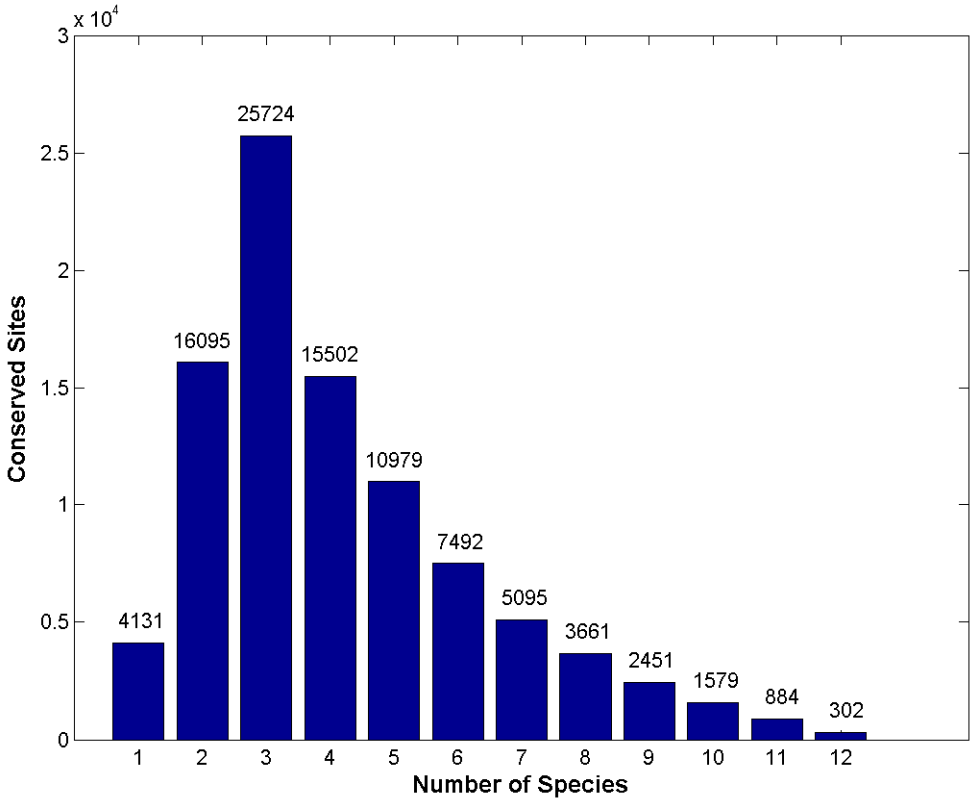


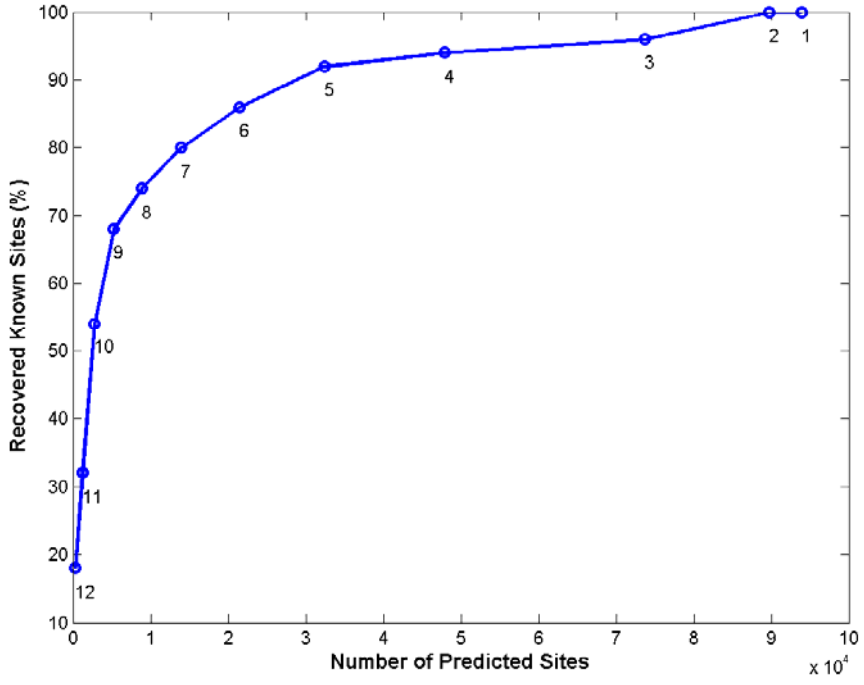
Figure 5.9: Comparison of the known κB sites with predictions as a function of the conservation score threshold.

A. The x- and y-axes correspond to the number of predicted κB sites and the fraction of recovered known κB sites (i.e. declared to be conserved) at a particular conservation score threshold, respectively.

B. The x- and y-axes correspond to the number of predicted κB genes (i.e. genes whose promoters contain the predicted κB sites) and the fraction of recovered known κB sites at a particular conservation score threshold, respectively.

As the conservation score (the number of species in which an orthologous set is conserved) decreases from twelve to one, the fraction of recovered known sites increases. The number of predicted κB sites in a genome-wide search and the corresponding number of κB genes also increase. After the conservation score of nine, the number of predicted κB sites increases much faster than the fraction of recovered known sites. After the conservation score of ten, the number of predicted κB genes increases much faster than the fraction of recovered known sites. Each conservation score threshold is shown and the corresponding point is depicted by a circle in the plots. This figure is based partly on Figure 5.6 and Figure 5.8.

A


Recovered Known Sites (%)
Number of Predicted Sites
x 10^4
1
2
3
4
5
6
7
8
9
10
11
12

B

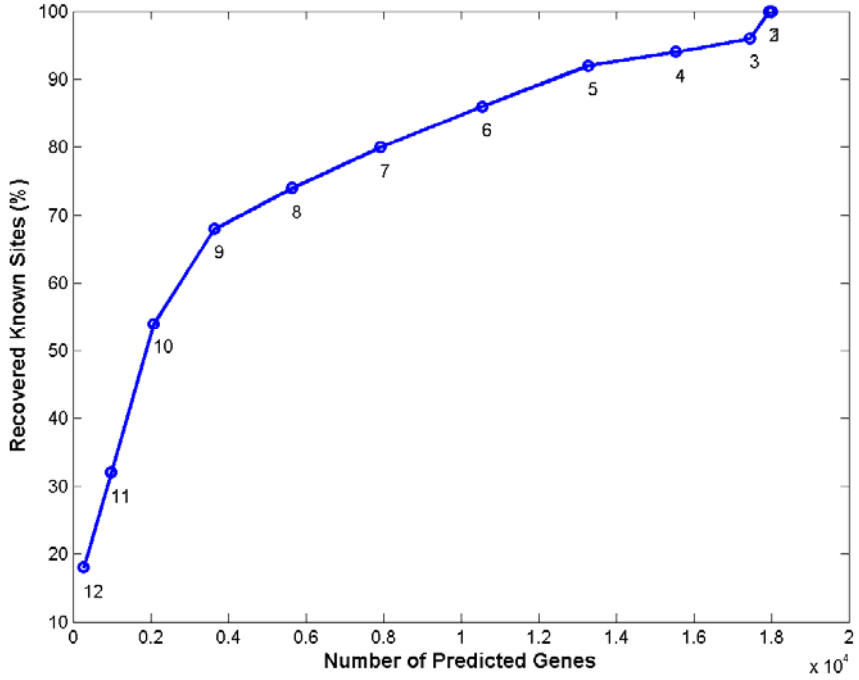

Recovered Known Sites (%)
Number of Predicted Genes
x 10^4
3
4
5
6
7
8
9
10
11
12

Figure 5.10: Distribution of location of the predicted conserved κB sites in promoters.

Data is shown for the 1186 predicted κB sites conserved in 11 or more mammals. The majority of sites are located near the transcription start site (TSS). Their number decreases in the regions further upstream of the TSS. Location in promoters is shown with respect to the TSS. Upstream regions have negative coordinates and downstream regions have positive coordinates.

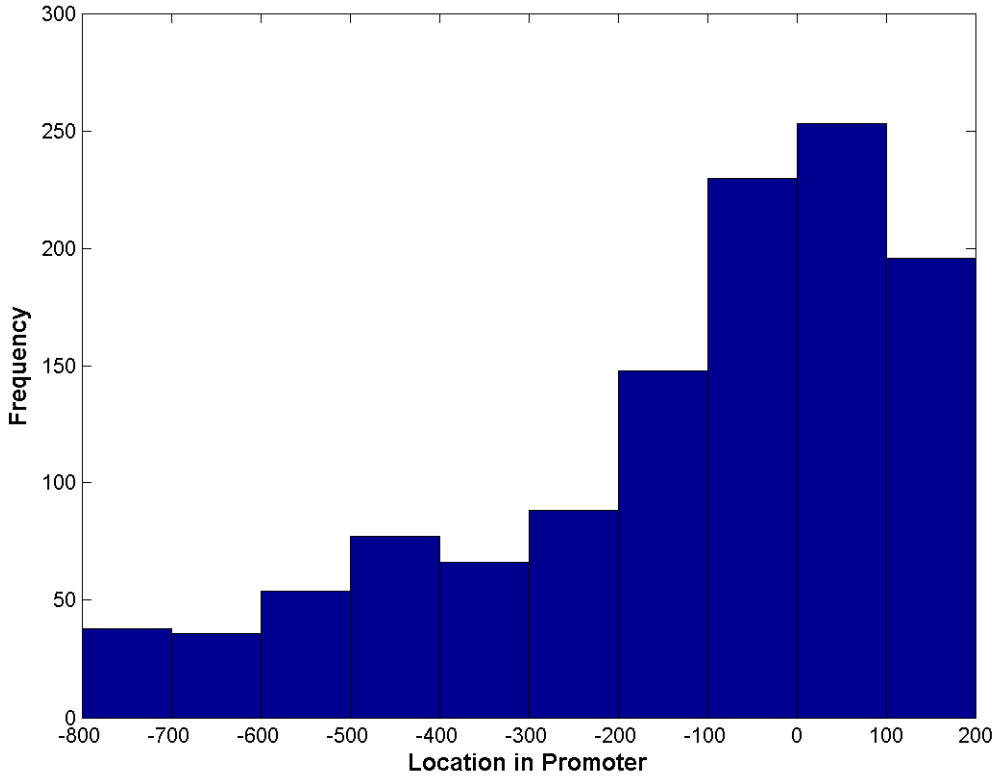


Figure 5.11: Pairwise Hamming distance in conserved κB sites.

Distribution of the Hamming distance between each pair of orthologous sites in conserved orthologous sets is shown. It has an exponential-like form. Most site pairs are identical, and the number of site pairs decreases sharply as the Hamming distance increases. The orthologous sets correspond to the 302 predicted κB sites conserved in all mammals.

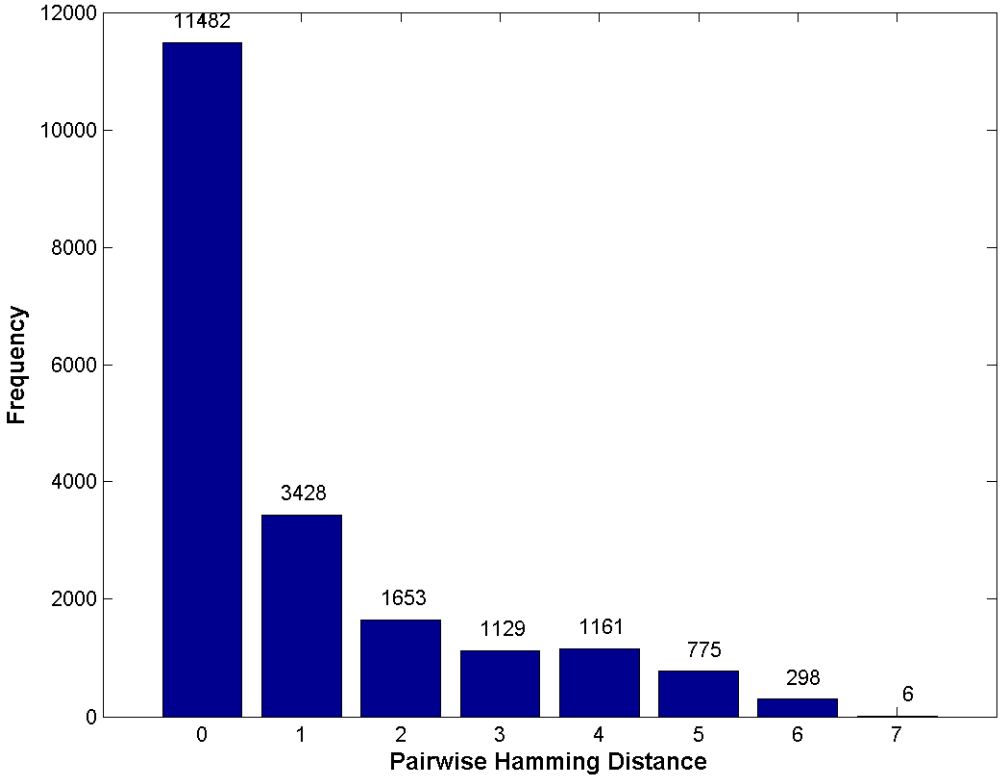
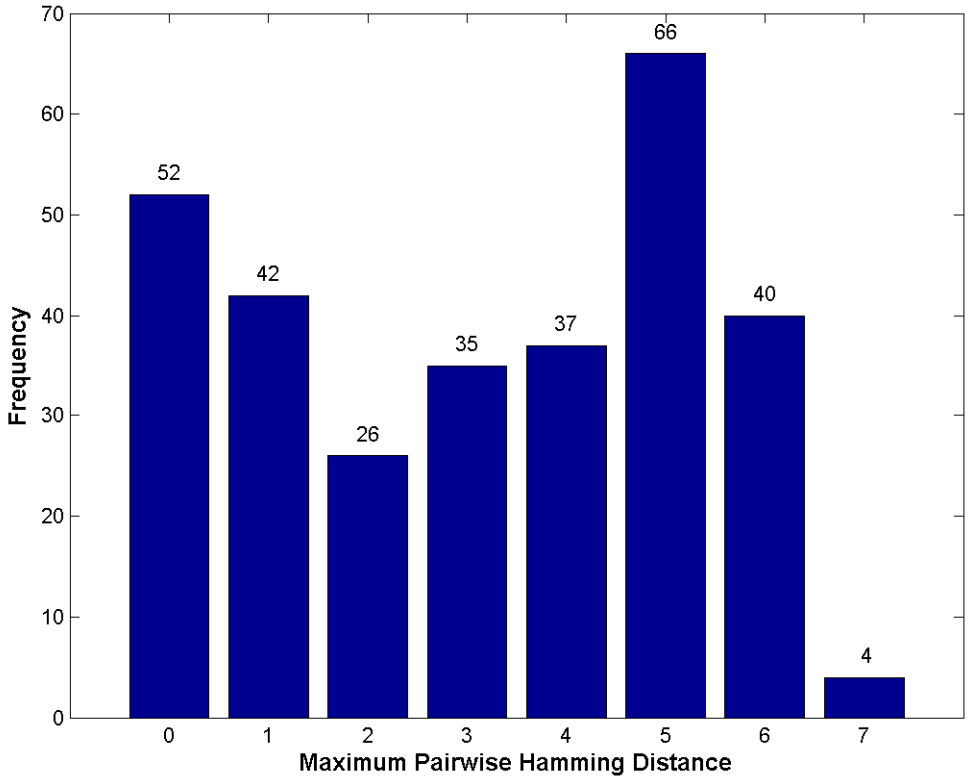


Figure 5.12: Distribution of the maximum pairwise Hamming distance in conserved orthologous sets of κB sites shows conservation of binding energy.

The high maximum pairwise Hamming distance (MPHD) values for a large number of conserved orthologous sets demonstrate that site energy rather than site sequence is conserved. The largest fraction of the orthologous sets has the MPHD of five. Over 36% of the orthologous sets have the MPHD of five or more, which corresponds to quite a large difference in orthologous κB sites that are only ten nucleotides long. Four orthologous sets have the MPHD of seven. The orthologous sets correspond to the 302 predicted κB sites conserved in all mammals. MPHD in an orthologous set is the maximum of the Hamming distances calculated for all site pairs in that set.



Chapter 6

Conclusions and Outlook

"I do not know what I may appear to the world; but to myself I seem to have been only like a boy playing on the seashore, and diverting myself in now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of truth lay all undiscovered before me."
Isaac Newton (1643-1727)

6.1 Conclusions

Unraveling the biological significance of a small degenerate sequence that constitutes a transcription factor's binding site is a formidable task requiring sophisticated knowledge in a variety of scientific disciplines. It is indeed a long road from a seemingly simple site sequence to the species survival. The numerous milestones along the way can be briefly outlined. A site's sequence determines its binding energy, which governs its occupancy probability, which affects the mRNA expression of the target gene, which in turn determines the resulting protein's abundance, which influences cellular pathways, which shape a cell's function, which affects the survival of an organism, which corresponds to the organism's evolutionary fitness, which ultimately determines the survival of its progeny.

Two types of models can in principle be used for site identification. At one extreme, there is a purely biophysical model, in the realm of structural biology or molecular modeling,

which emulates the physical binding interaction of a transcription factor protein with a DNA sequence in atomistic details. It models the bond lengths, bond angles and torsion angles within each molecule as well as the van der Waals forces and electrostatic interactions between neighboring atoms. (The associated set of parameters of such a model is called a force field.) It aims to minimize the potential energy, also called the potential function, of the interaction between the two molecules using simulations and optimization techniques. Because calculation of the binding energy of a transcription factor with even a single DNA sequence requires huge computational power, this approach has had limited success in identifying sites.

At the other extreme is a plethora of probabilistic machine learning models. They use limited physical intuition but they are able to identify sites based on the sequences of a few known sites.

Our approach falls in between these models. While we do not model each atomic interaction in the three-dimensional space, we do not confine ourselves to purely statistical quantities either. We have set up our machine learning models by incorporating simple physical interactions, modeling binding energies and using occupancy as a way of scoring things.

To be able to develop a stochastic model with biophysical underpinnings, one needs to be conversant with various fields, from computer science to genetics, from statistics to systems biology and from thermodynamics to evolutionary biology. Computational

identification of a transcription factor's binding site alone requires a confluence of such diverse fields as machine learning (which itself is an amalgamation of statistics and computer science), biophysics, systems biology and evolutionary biology. Site identification is a classification problem in machine learning that partitions sequences into binders and non-binders through the use of a discriminant function. Transcription is a biophysical process dealing with a site's binding energy and occupancy of a transcription factor on DNA. Systems biology takes up the study of gene expression on the cell function. Evolutionary biology is concerned with how a site's composition and direct function affects evolutionary success, and its subfield of population genetics focuses on the relationship between the fitness of an organism with the survival of its progeny.

In this work, we have linked machine learning methods of site identification to biophysical models of transcription factor binding in a single species and then extended these techniques to a model of site evolution involving a number of species.

One part of our work is thus devoted to the biophysical interpretation of machine learning methods to calculate occupancy probability of a transcription factor on a site and to establish the classification threshold in a principled manner. We recognize that binding energy of a site, and not its sequence – as is commonly assumed –, is its key property that determines its function and evolutionary fitness. Occupancy probability in a sense forms a bridge between binding energy and fitness. It has a sigmoidal shape (Fermi-Dirac distribution) with a natural threshold at 0.5 when viewed as a function of binding energy. Therefore, if binding energy of a sequence, at a particular transcription factor

concentration, corresponds to occupancy probability of greater than 0.5, it can be considered a functional site.

The hidden Markov model (HMM) that we have developed interprets the weight matrix as binding energy, the transition probability to the motif as transcription factor concentration and the gamma variable as occupancy probability. PhyloQPMEME determines binding energies and occupancy probability after solving a constrained optimization problem. Because they use occupancy probability as the discriminant function, they learn the associated natural threshold in a principled manner during the training procedure. These features highlight the distinction between our methods and most other machine learning methods for site identification which deal with statistical quantities that are not immediately interpretable as the biophysical variables. These latter methods, therefore, are forced to resort to arbitrary, often non-physical, thresholds. In contrast, linking our machine learning models to biophysics not only allows a rigorous interpretation of the statistical quantities but also improves their performance.

The second part of our work deals with specific characteristics of sites to enhance their identification. The HMM combines the effects of alternative binding modes of self-overlapping sites. It then biases site identification in regions with high site density. Our HMM analysis provides guidance on the design of padding sequences in experiments associated with self-overlapping sites. On the other hand, PhyloQPMEME integrates evolutionary conservation of sites into a model of binding energies. Conservation in

multiple species is a hallmark of continuing selection pressure due to functional constraints, and hence the use of the conservation criterion helps reduce false positives.

By identifying the direct target genes of the NF-κB transcription factor family using both the HMM and PhyloQPMEME, we have been able to learn a great deal about NF-κB biology, the evolution of κB sites and the predictive nature of these two methods. With the aim of determining the biological significance of the sets of κB target genes predicted by these methods at the various thresholds, we used pathway analysis to discover the pathways enriched in these κB target genes (Figure 6.1). Two types of measures were applied to determine the biological significance of the gene sets: (1) the sum of the negative logarithm of the p-values of the top 25 enriched pathways, and (2) the number of pathways enriched with a p-value less than 0.01. It is useful to note that only about 50-70% of the genes in each gene set are available for pathway analysis because the rest of the genes are not adequately annotated. The numbers in Figure 6.1, however, correspond to the number of genes in the entire gene sets.

The gene sets predicted by the HMM at various occupancy probability thresholds are much more biologically significant than randomly selected genes. Moreover, the biological significance reaches a peak at the occupancy probability threshold of 0.5 (corresponding to ~800 genes). This implies that the gene sets corresponding to the thresholds greater than 0.5 have many false negatives, because of which these gene sets do not have enough key target genes to attain high significance. On the contrary, the gene sets corresponding to the thresholds less than 0.5 have many more false positives which

dilute these sets and lower their biological significance. Thus, the HMM appears to identify sites most accurately in a short window around the threshold of 0.5. This observation ties excellently with our justification for training the HMM threshold and the use of occupancy probability as the discriminant function.

While the PhyloQPMEME-predicted gene sets at various conservation score thresholds have much greater biological significance as compared to randomly selected genes, their biological significance has a much broader peak going down to the conservation scores of 7-8 (corresponding to a few thousand genes). This suggests that many genuine sites are conserved in only some of the considered mammals. In other words, a great number of κB sites have been lost during evolution (although we do not know their turnover rate), and hence a significant number of sites can be recovered even when the conservation score is low. This observation is in agreement with the fact that one third of the known sites are conserved in fewer than 9 of the considered mammals (Figure 5.6). The above discussion does not imply that higher conservation thresholds are useless. Even though gene sets corresponding to these thresholds contain many false negatives, we have already seen that several well-known NF-κB pathways are enriched with these genes. In addition, one should not forget their significance in terms of evolutionary conservation. Therefore, these gene sets should serve as non-comprehensive but reliable sets of κB targets.

The comparison of the gene sets predicted by the HMM and PhyloQPMEME shows that both these methods are useful in investigating the role played by κB sites and NF-κB

target genes. We compared the gene set corresponding to the HMM occupancy probability threshold of 0.5 (~800 genes) with that corresponding to the PhyloQPMEME conservation score threshold of 11 (~900 genes). Even though only ~15% of the genes in these sets overlap, the enriched pathways associated with these gene sets are remarkably similar. Many of the well-known NF-κB pathways are shared by these gene sets, including glucocorticoid receptor signaling, antigen presentation, B cell receptor signaling, chemokine signaling, and so on. Interestingly, both gene sets highlighted the less well-characterized roles of NF-κB in the ubiquitination and xenobiotic metabolism signaling pathways. However, the most extraordinary feature of the comparison was the suggestion by both gene sets of the potential NF-κB-regulated pathways in the nervous system, an idea still in its infancy among the biologists at the current time.

In conclusion, in this work we have developed machine learning methods to enhance identification of transcription factor binding sites, emphasized the paradigm of bridging probabilistic models with biophysics, and gained new insights into NF-κB biology. We believe that our work will uncover hidden truths about the regulation of gene expression and contribute to a better understanding of biology.

6.2 Outlook

One logical extension of our work is to combine the HMM of self-overlapping sites with a model of site energy evolution like PhyloQPMEME. However, our understanding of site evolution is very deficient at the present time. While little information is gained from

species with a small divergence, we believe that PhyloQPMEME has adequately dealt with it. With species having large divergence, however, phenomena such as lost sites, site turnovers and promoter reorganizations complicate the use of species evolution as a way to decide about binding sites. A comprehensive treatment of these phenomena and the determination of a principled conservation threshold are still a distant goal.

Another possible extension to the current work is the discrimination between the sites of slightly different specificities associated with binding to different configurations of a transcription factor or to related transcription factors. For example, members of the NF-κB transcription factor family form various homo- and hetero-dimers that have slightly different binding specificities [56]. As some functions of these dimers are known to be different, identification of sites that prefer one dimer to another (i.e. marginal κB sites) will help us to better understand their differential gene regulation effects. The problem here is that their binding specificities (modeled by a weight matrix or an energy matrix) are quite similar, and as a result, most sites are expected to bind to many dimers without discrimination. There is not even sufficient experimental data which distinguishes the binding of different dimers and which can conceivably be used to fine-tune theoretical models. In this work, therefore, we have treated all κB sites as a single set. In the future, one can tease out different affinities of these NF-κB dimers once the experimental data on the marginal sites becomes available and computational modeling is focused on these sites. In particular, data about the abundance of various dimers on marginal sites are needed as the training set. Computational models like PhyloQPMEME or support vector machines can be made to establish classification thresholds based on marginal sites. They

are more appropriate in this case than weight matrices or HMM, which are based on the affinities of all sites and thus dilute the effect of marginal sites.

Figure 6.1: Biological significance of predicted target gene sets using pathway analysis.

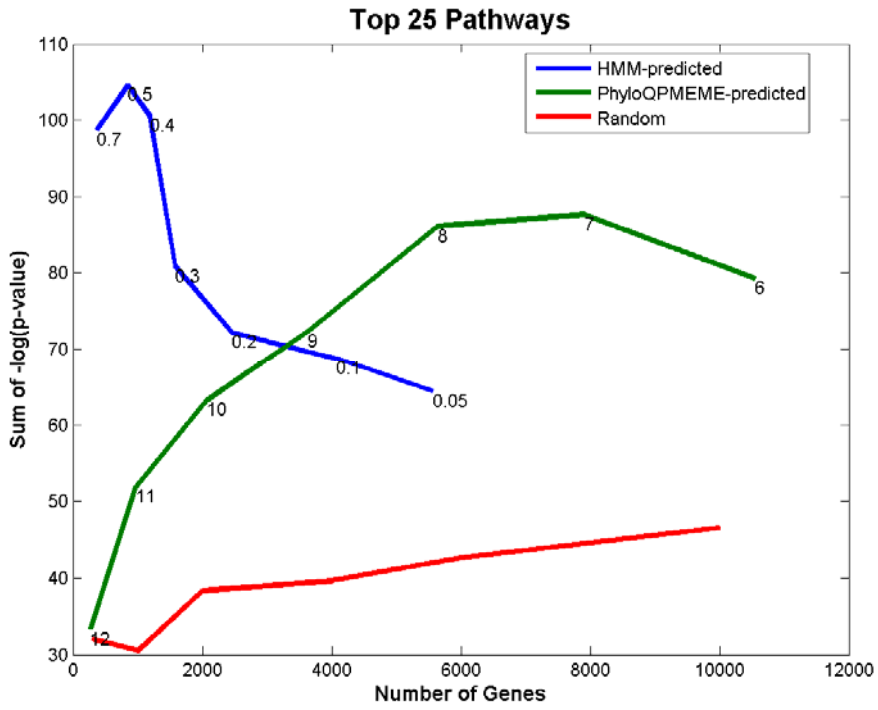
Biological significance is shown with the help of the pathways enriched in the κB target gene sets predicted by the HMM and PhyloQPMEME at various thresholds. The two statistics used are:

A. The sum of the negative logarithm of the p-values of the top 25 enriched pathways.

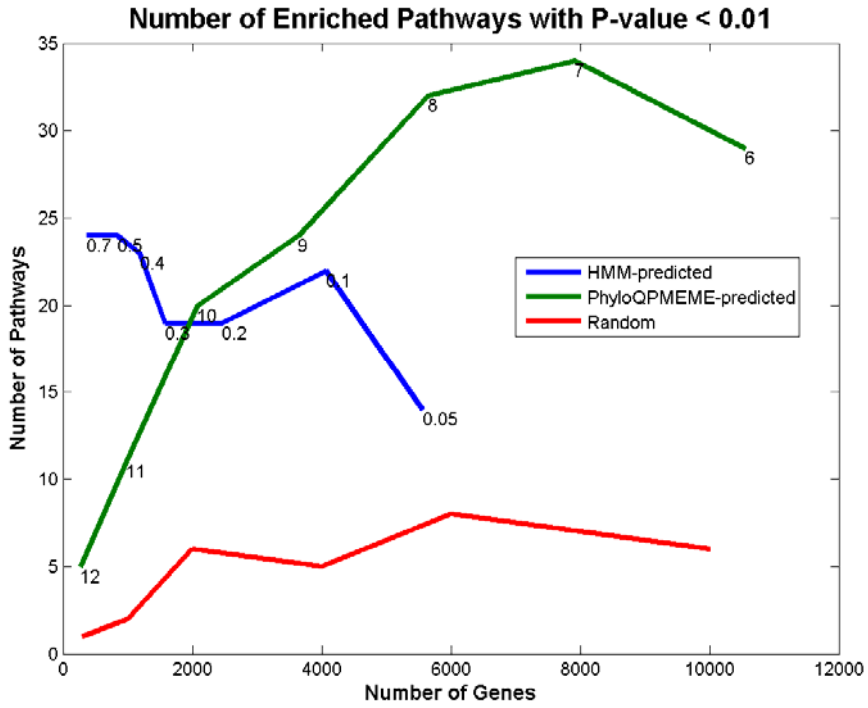
B. The number of pathways enriched with a p-value less than 0.01.

Gene sets associated with both methods are biologically significant as compared to randomly selected genes. While the HMM-predicted genes show a peak at the threshold occupancy probability of 0.5 (~800 genes), PhyloQPMEME-predicted genes show a much broader peak at the conservation score thresholds of 7-8 (a few thousand genes). The thresholds used for obtaining the gene sets for the pathway analysis (occupancy probability threshold between 0.05 and 0.7 for HMM and conservation score threshold between 6 and 12 for PhyloQPMEME) are indicated. HMM-predicted gene sets, PhyloQPMEME-predicted gene sets and randomly selected gene sets are indicated by blue, green and red curves, respectively. Only about 50-70% of the genes in each gene set are available for pathway analysis because the rest of the genes are not adequately annotated. The numbers in the figure, however, correspond to the number of genes in the entire gene sets.

A


Top 25 Pathways
HMM-predicted
PhyloQPMEME-predicted
Random
Sum of -log(p-value)
Number of Genes
0.7
0.5
0.4
0.3
0.2
0.1
0.05
12
11
10
9
8
7
6

B


Number of Enriched Pathways with P-value < 0.01
HMM-predicted
PhyloQPMEME-predicted
Random
Number of Pathways
Number of Genes
0.7
0.5
0.4
0.3
0.2
0.1
0.05
12
11
10
9
8
7
6

Appendix A. Derivation of Occupancy Probability of Overlapping Sites

In this Appendix, we will calculate the occupancy probability of sites at any position in a sequence as well as over the entire sequence using both first principles method and standard HMM techniques. We will consider the cases of (i) one site, (ii) non-overlapping sites of the same type, (iii) exactly overlapping sites of multiple types, and finally, (iv) the most general case of overlapping sites of multiple types. We will also show that even though a conventional weight matrix and an HMM are closely related in principle, an HMM is more appropriate to determine occupancy probability when self-overlapping sites exist.

We will use the following symbols in the derivation below: *b* for the background state; *m* for the motif state, where the motif is a representation of a type of binding sites; α for a nucleotide; ℓ for the length of the motif; *i* for the position in the motif; *z* for the transition probability to the motif; *s* for a sequence; *L* for the entire length of the sequence; *j* for the position in the sequence; w_{α}^{bj} for the probability that nucleotide α at the *j*th position of the sequence is emitted by the background state; $w_{i\alpha}^{mj}$ and $w_{i\alpha}^{bj}$ for the respective probabilities that nucleotide α at the $(j+i-1)$th position of the sequence is emitted by the *i*th position of the motif state or by the background state ; $p_j^{bound}(s)$ for the occupancy probability of transcription factors at the *j*th position of the sequence; $p^{bound}(s)$ for the

occupancy of transcription factors over the entire sequence. Because most of the promoter sequence is the background, transition probability to the motif $z \approx 0$ and hence $\ell.z \approx 0, (1-z) \approx 1 \,\&\, (1-z)^{\ell} \approx 1$.

The weight matrix score corresponding to the motif starting at the *j*th position of the sequence is defined as

$$W_j = \ln\left(\prod_{i=1}^{\ell} \frac{w_{i\alpha}^{mj}}{w_{i\alpha}^{bj}}\right) \text{ (A1)}$$

The motif's strength compared to the background at that position is

$$E_j = \left(\prod_{i=1}^{\ell} \frac{w_{i\alpha}^{mj}}{w_{i\alpha}^{bj}}\right) = e^{W_j} \text{ (A2)}$$

A.1 One Site

When a sequence is scored using an HMM, the likelihood of the sequence is the sum of all configurations, i.e. combinations of the background and motif states at all positions of the sequence. The configuration that has the background state at all positions is given by the probability

$$p(b) = \ldots(1-z).w_{\alpha}^{bj}.(1-z).w_{\alpha}^{b(j+1)} \ldots = (1-z)^{L}.\prod_{j=1}^{L} w_{\alpha}^{bj} \ \ldots \text{ (A3)}$$

The configuration with motif *m* at the *j*th position has the probability

$$p_j(m) = \ldots(1-z).w_{\alpha}^{b(j-1)}.z.\prod_{i=1}^{\ell} w_{i\alpha}^{mj}.(1-z).w_{\alpha}^{b(j+\ell+1)} \ldots$$

and can be expressed in terms of the probability of the configuration of the background state at all positions as follows:

$$p_j(m) = p(b).\frac{z}{(1-z)^{\ell}}\prod_{i=1}^{\ell}\frac{w_{i\alpha}^{mj}}{w_{i\alpha}^{bj}} = p(b).\frac{z}{(1-z)^{\ell}}.e^{W_j} \approx p(b).z.e^{W_j} = p(b).z.E_j \quad \text{(A4)}$$

Thus, the two factors z and $E_j = e^{W_j}$, one the transition probability to the motif and the other a measure of distinctness of the emission probabilities of the motif (motif profile) from that of the background, determine occupancy probability. Occupancy probability at the *j*th position in terms of the transition probability to the motif and the weight matrix score is given by

$$p_j^{bound}(s) = \frac{p_j(m)}{p(b)+p_j(m)} = \frac{\frac{z}{(1-z)^{\ell}}.e^{W_j}}{1+\frac{z}{(1-z)^{\ell}}.e^{W_j}} = \frac{z.e^{W_j}}{(1-z)^{\ell}+z.e^{W_j}} \approx \frac{z.e^{W_j}}{1+z.e^{W_j}} = \frac{z.E_j}{1+z.E_j} \ldots \quad \text{(A5)}$$

As long as we know z, we can calculate the occupancy probability at a sequence position using the weight matrix, and hence the weight matrix threshold for classifying sequences into sites can be easily determined from the occupancy probability threshold. For example, the occupancy probability threshold of 0.5 (corresponding to $p_j(m) = p(b)$ and $z.e^{W_j} = 1$) results in the weight matrix threshold of $W = -\ln z$.

We can also calculate occupancy probability using an HMM. The HMM gamma variable corresponds to the probability that position *j* of the sequence is in a certain state. For example, γ_j^m and γ_j^b correspond to the probabilities of the motif and background states at the *j*th position, respectively. $\gamma_j^m + \gamma_j^b = 1$ if only these two states are considered. Hence,

gamma of the motif state at a sequence position is the occupancy probability at that position ($p_j^{bound}(s) = \gamma_j^m = 1 - \gamma_j^b$).

A.2 Non-overlapping sites of the same type

A configuration containing two non-overlapping sites of the same type at positions j_1 and j_2 has the probability $p_{j_1 j_2}(m) = \ldots z. \prod_{i=1}^{\ell} w_{i\alpha}^{mj_1} \ldots z. \prod_{i=1}^{\ell} w_{i\alpha}^{mj_2} \ldots = \frac{1}{(1-z)^{2\ell}} p(b).z.E_{j_1}.z.E_{j_2}$

To explain the case of non-overlapping sites better, we can think of the motif state at a sequence position to be emitting ℓ nucleotides, and thus each sequence position has one of the two states *m* or *b*.

Let's calculate the overall likelihood of the sequence to understand the relationships between the different terms. The likelihood is the sum of the probabilities of all configurations: $e^{\mathcal{I}} = p(b_1 b_2 \ldots) + p(b_1 m_2 \ldots) + p(m_1 b_2 \ldots) + \ldots$, where $\mathcal{I}$ is the log likelihood. When the transition probability to the background or to the motif state is independent of the previous state (which is assumed for the HMMs in this text),

$$e^{\mathcal{I}} = p(b_1).p(b_2 \ldots) + p(b_1).p(m_2 \ldots) + p(m_1).p(b_2 \ldots) + \ldots$$

$$e^{\mathcal{I}} = \big(p(b_1) + p(m_1)\big).\big(p(b_2 \ldots) + p(m_2 \ldots)\big)$$

$$e^{\mathcal{I}} = \prod_{j=1}^{L} \big(p(b_j) + p(m_j)\big)$$

$$e^{\mathcal{I}} = p(b).\prod_{j=1}^{L}\left(1+\frac{z}{(1-z)^{\ell}}.E_j\right)\ldots \text{(A6)}$$

$$e^{\mathcal{I}} = (1-z)^{L}.\prod_{j=1}^{L} w_{\alpha}^{bj}.\prod_{j=1}^{L}\left(1+\frac{z}{(1-z)^{\ell}}.E_j\right)\ldots \text{From equation (A3)}$$

$$e^{\mathcal{I}} \approx (1-z)^{L}.\prod_{j=1}^{L} w_{\alpha}^{bj}.\prod_{j=1}^{L}(1+z.E_j)\ldots \text{Because } (1-z)^{\ell} \approx 1$$

$$e^{\mathcal{I}} \approx (1-z)^{L}.\prod_{j=1}^{L} w_{\alpha}^{bj}.\left(1+\sum_{j}^{L} z.E_j + \sum_{j}^{L}\sum_{k\neq j}^{L} z.E_j.z.E_k + \ldots\right)\ldots \text{(A7)}$$

In equation (A7), the first term corresponds to the configuration with only background, the second term corresponds to all configurations with one site, the third term corresponds to all configurations with two non-overlapping sites, etc. Note that the summation terms in equation (A7) do not take overlapping sites into account, and hence the above equations are inaccurate for overlapping sites.

We see from equation (A6) that the likelihood is dominated by high E_j's. If there is only one strong weight matrix score ($z.E_j = z.e^{W_j} >> 1$), the likelihood is in the order of magnitude of its exponent. If there are multiple strong weight matrix scores, the likelihood is in the order of magnitude of the product of their exponents (equivalently, the log likelihood is in the order of magnitude of the sum of the weight matrix scores). However, if there are many moderate weight matrix scores ($z.E_j \approx 1$), the likelihood will also increase slightly. Most weight matrix scores are very low ($z.E_j << 1$) and thus do not contribute significantly to the likelihood.

To determine the occupancy over the entire sequence, let's calculate the maximum likelihood estimate (MLE) of z by taking the derivative of log likelihood.

$$\mathcal{L} \approx L.\ln(1-z) + const + \sum_{j=1}^{L} \ln(1 + z.E_j)$$

$$\frac{\partial \mathcal{L}}{\partial z} = -\frac{L}{1-z} + \sum_{j=1}^{L} \frac{E_j}{1 + z.E_j}$$

$$L = \sum_{j=1}^{L} \frac{E_j}{1 + z.E_j}$$

Therefore, occupancy over the entire sequence, i.e. the product of the sequence's length and the transition probability to the motif, is given by

$$p^{bound}(s) = L.z = \sum_{j=1}^{L} \frac{z.E_j}{1 + z.E_j} \quad \ldots \text{(A8)}$$

As with the case of calculating the occupancy probability at a position, the knowledge of z allows us to calculate the occupancy over the entire sequence with the help of a weight matrix. In the HMM context, this is simply the sum of the occupancy probabilities at all positions. For the case of non-overlapping sites, it is the sum of the gammas of the motif states at all positions ($p^{bound}(s) = \sum_{j=1}^{L} \gamma_j^m$).

A.3 Exactly overlapping sites of multiple types

When different types of sites are present such that they overlap exactly, for example when we consider the κB site on both strands, the occupancy probability of any type of site at a position is

$$p_j^{bound}(s) = \frac{p(m1) + p(m2)}{p(b) + p(m1) + p(m2)} \approx \frac{z_1.E_{1j} + z_2.E_{2j}}{1 + z_1.E_{1j} + z_2.E_{2j}} = \frac{\sum_m z_m.E_{mj}}{1 + \sum_m z_m.E_{mj}} \ldots \text{(A9)}$$

where m indicates the motif type, z_m is the transition probability to the m th motif type, and E_{mj} is the exponent of the weight matrix of motif type m starting at sequence position *j*. In this case, calculation of occupancy probability using weight matrices requires knowledge of multiple *z*'s. In the HMM context, however, the occupancy probability is simply the sum of gammas of all motif states, i.e. $p_j^{bound}(s) = \sum_m \gamma_j^m$, or alternatively $p_j^{bound}(s) = 1 - \gamma_j^b$, where γ_j^b is the gamma of the background state at that position.

The likelihood of the entire sequence is then

$$e^{\mathcal{I}} = \prod_{j=1}^{L} \left(p(b_j) + p(m1_j) + p(m2_j) \right)$$

$$e^{\mathcal{I}} \approx p(b).\prod_{j=1}^{L} \left(1 + z_1.E_{1j} + z_2.E_{2j}\right) = p(b).\prod_{j=1}^{L} \left(1 + \sum_m z_m.E_{mj} \right)$$

The occupancy of any type of site over the entire sequence is $p^{bound}\left(s\right) = \sum_{j=1}^{L} \frac{\sum_m z_m.E_{mj}}{1 + \sum_m z_m.E_{mj}}$.

Its calculation using weight matrices is tedious. However, it can be easily calculated using HMMs if we divide up each motif type into ℓ states each corresponding to one position in the motif (this HMM architecture is described in detail for the next case). The

occupancy over the entire sequence is then $p^{bound}(s) = \sum_{j=1}^{L} \sum_{m} \gamma_j^{m_1}$, where $\gamma_j^{m_1}$ is the gamma of the first position of the *m*th type of motif at the *j*th position of the sequence.

A.4 Overlapping sites of multiple types

Finally, we consider the most general case of overlapping sites of multiple types. The probabilities of configurations of two self-overlapping sites are shown below:

$$p_j(m) = \ldots(1-z).w_\alpha^{b(j-1)}.z.\prod_{i=1}^{\ell} w_{i\alpha}^{mj}.(1-z).w_\alpha^{b(j+\ell+1)}.(1-z).w_\alpha^{b(j+\ell+2)}\ldots$$

$$p_{(j+1)}(m) = \ldots(1-z).w_\alpha^{b(j-1)}.(1-z).w_\alpha^{bj}.z.\prod_{i=1}^{\ell} w_{i\alpha}^{m(j+1)}.(1-z).w_\alpha^{b(j+\ell+2)}\ldots$$

They show that calculation of the occupancy probability at a position requires consideration of all windows containing the position. When only one type of motif is considered, the occupancy probability at a sequence position is given by an equation similar to equation (A9):

$$p_j^{bound}(s) \approx \frac{\sum_{k=j-\ell+1}^{j} z.E_k}{1+\sum_{k=j-\ell+1}^{j} z.E_k} \quad \ldots \text{(A10)}$$

where *k* is the first position of each sequence window containing sequence position *j*. We can extend equation (A10) to calculate the occupancy probability at a position in overlapping sites of multiple types as follows:

$$p_j^{bound}(s) \approx \frac{\sum_{k=j-\ell_m+1}^{j} \sum_m z_m.E_{mk}}{1+\sum_{k=j-\ell_m+1}^{j} \sum_m z_m.E_{mk}} \quad \ldots \text{(A11)}$$

where m is the motif type (including the motif on the other strand corresponding to the same transcription factor), and ℓ_m is the length of the m th motif type. Because this occupancy probability depends upon multiple sequence windows, its calculation using weight matrices is not straightforward even when we know the z's.

The HMM architecture described briefly at the end of the discussion of the last case allows easy calculation of occupancy probability. In this HMM, each motif type actually corresponds to ℓ_m states, each state associated with one position in the motif. Each state's emission probabilities are equal to the weights corresponding to that position of the motif's weight matrix. If the motif does not contain insertions or deletions, the transition probabilities between successive states within the motif are equal to one. This HMM's gamma variable automatically takes the overlaps into account. The occupancy probability of a particular transcription factor at the j th position is $p_j^{bound}(s) = \sum_{m \in M} \sum_{i=1}^{\ell_m} \gamma_j^{m_i}$, where M is the set of motif types corresponding to the transcription factor, and $\gamma_j^{m_i}$ is the gamma of the state corresponding to the i th position in motif m at the j th position of the sequence. If the HMM contains multiple motif types corresponding to only transcription factor, the occupancy probability at a position is simply

$$p_j^{bound}(s) = 1 - \gamma_j^b \quad \ldots \text{(A12)}$$

where γ_j^b is the gamma of the background at that position. This is the same formula as for exactly overlapping sites.

Calculation of the occupancy over an entire sequence in the above case is quite difficult using weight matrices and quite easy using an HMM. Because a single configuration cannot contain overlapping sites, no simple formula exists for the overall likelihood of the sequence (for example, the third term in equation (A7) is invalid). Hence, the knowledge of the *z* and the weight matrix does not allow us to calculate the overall occupancy of the sequence quickly. However, as for exactly overlapping sites, we can easily calculate the overall occupancy using an HMM as

$$p^{bound}(s) = \sum_{j=1}^{L} \sum_{m} \gamma_j^{m_1} \quad \ldots \text{(A13)}$$

where $\gamma_j^{m_1}$ is the gamma of the motif corresponding to the first position of the *m*th motif type at the *j*th position of the sequence.

In the absence of overlapping sites, occupancy probability at a sequence position calculated using the weight matrix score and the *z* (and thus not calculated using an HMM), the HMM gamma of the first state of the motif at that position and the HMM gamma of the entire motif at that position (i.e. the sum of the HMM gamma's of the various states of the motif at that position) are identical. However, in the case of overlapping sites, the HMM gamma of the entire motif is greater than the occupancy probability calculated using the weight matrix score and the *z* because the latter fails to consider overlapping sites. This occupancy probability is in turn greater than the HMM

gamma of the first state of the motif, which has a low value due to the presence of a site in an overlapping window.

References

1. Alberts B: **Molecular biology of the cell**, 4th edn. New York: Garland Science; 2002.

2. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI**. *Nucleic Acids Res* 2007, **35**(Database issue):D26-31.

3. Latchman DS: **Transcription factors: an overview**. *Int J Biochem Cell Biol* 1997, **29**(12):1305-1312.

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

5. Smale ST, Kadonaga JT: **The RNA polymerase II core promoter**. *Annu Rev Biochem* 2003, **72**:449-479.

6. van Nimwegen E: **Scaling laws in the functional content of genomes**. *Trends Genet* 2003, **19**(9):479-484.

7. Brivanlou AH, Darnell JE, Jr.: **Signal transduction and the control of gene expression**. *Science* 2002, **295**(5556):813-818.

8. Gupta N, Delrow J, Drawid A, Sengupta AM, Fan G, Gelinas C: **Repression of B-cell linker (BLNK) and B-cell adaptor for phosphoinositide 3-kinase (BCAP) is important for lymphocyte transformation by rel proteins**. *Cancer Res* 2008, **68**(3):808-814.

9. Bulyk ML: **Computational prediction of transcription-factor binding site locations**. *Genome Biol* 2003, **5**(1):201.

10. Elnitski L, Jin VX, Farnham PJ, Jones SJ: **Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques**. *Genome Res* 2006, **16**(12):1455-1464.

11. Gopinath SC: **Methods developed for SELEX**. *Anal Bioanal Chem* 2007, **387**(1):171-182.

12. Sikder D, Kodadek T: **Genomic studies of transcription factor-DNA interactions**. *Curr Opin Chem Biol* 2005, **9**(1):38-45.

13. Schneider TD, Stephens RM: **Sequence logos: a new way to display consensus sequences**. *Nucleic Acids Res* 1990, **18**(20):6097-6100.

14. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator**. *Genome Res* 2004, **14**(6):1188-1190.

15. Gribskov M, McLachlan AD, Eisenberg D: **Profile analysis: detection of distantly related proteins**. *Proc Natl Acad Sci U S A* 1987, **84**(13):4355-4358.

16. Rahmann S, Muller T, Vingron M: **On the power of profiles for transcription factor binding site detection**. *Stat Appl Genet Mol Biol* 2003, **2**:Article7.

17. Stormo GD: **DNA binding sites: representation and discovery**. *Bioinformatics* 2000, **16**(1):16-23.

18. **Rel/NF-kB Transcription Factors** [http://www.nf-kb.org]

19. Ballard DW, Walker WH, Doerre S, Sista P, Molitor JA, Dixon EP, Peffer NJ, Hannink M, Greene WC: **The v-rel oncogene encodes a kappa B enhancer binding protein that inhibits NF-kappa B function**. *Cell* 1990, **63**(4):803-814.

20. Basseres DS, Baldwin AS: **Nuclear factor-kappaB and inhibitor of kappaB kinase pathways in oncogenic initiation and progression**. *Oncogene* 2006, **25**(51):6817-6830.

21. Boersma MC, Meffert MK: **Novel roles for the NF-kappaB signaling pathway in regulating neuronal function**. *Sci Signal* 2008, **1**(6):pe7.

22. Bonizzi G, Karin M: **The two NF-kappaB activation pathways and their role in innate and adaptive immunity**. *Trends Immunol* 2004, **25**(6):280-288.

23. Campbell KJ, Rocha S, Perkins ND: **Active repression of antiapoptotic gene expression by RelA(p65) NF-kappa B**. *Mol Cell* 2004, **13**(6):853-865.

24. Chiao PJ, Miyamoto S, Verma IM: **Autoregulation of I kappa B alpha activity**. *Proc Natl Acad Sci U S A* 1994, **91**(1):28-32.

25. Doerre S, Sista P, Sun SC, Ballard DW, Greene WC: **The c-rel protooncogene product represses NF-kappa B p65-mediated transcriptional activation of the long terminal repeat of type 1 human immunodeficiency virus**. *Proc Natl Acad Sci U S A* 1993, **90**(3):1023-1027.

26. Dutta J, Fan Y, Gupta N, Fan G, Gelinas C: **Current insights into the regulation of programmed cell death by NF-kappaB**. *Oncogene* 2006, **25**(51):6800-6816.

27. Fan Y, Dutta J, Gupta N, Fan G, Gélinas C: **Regulation of programmed cell death by NF-kB and its role in tumorigenesis and therapy**. In: *Programmed cell death in cancer progression and therapy.* Edited by Khosravi-Far R, White E, vol. XIV: The Netherlands: Springer; 2007: 223-250.

28. Fan Y, Rayet B, Gelinas C: **Divergent C-terminal transactivation domains of Rel/NF-kappa B proteins are critical determinants of their oncogenic potential in lymphocytes**. *Oncogene* 2004, **23**(5):1030-1042.

29. Ganchi PA, Sun SC, Greene WC, Ballard DW: **I kappa B/MAD-3 masks the nuclear localization signal of NF-kappa B p65 and requires the transactivation domain to inhibit NF-kappa B p65 DNA binding**. *Mol Biol Cell* 1992, **3**(12):1339-1352.

30. Hayden MS, Ghosh S: **Signaling to NF-kappaB**. *Genes Dev* 2004, **18**(18):2195-2224.

31. Hayden MS, West AP, Ghosh S: **NF-kappaB and the immune response**. *Oncogene* 2006, **25**(51):6758-6780.

32. Karin M: **How NF-kappaB is activated: the role of the IkappaB kinase (IKK) complex**. *Oncogene* 1999, **18**(49):6867-6874.

33. Karin M, Ben-Neriah Y: **Phosphorylation meets ubiquitination: the control of NF-[kappa]B activity**. *Annu Rev Immunol* 2000, **18**:621-663.

34. Karin M, Lin A: **NF-kappaB at the crossroads of life and death**. *Nat Immunol* 2002, **3**(3):221-227.

35. Kucharczak J, Simmons MJ, Fan Y, Gelinas C: **To be, or not to be: NF-kappaB is the answer--role of Rel/NF-kappaB in the regulation of apoptosis**. *Oncogene* 2003, **22**(56):8961-8982.

36. Majid SM, Liss AS, You M, Bose HR: **The suppression of SH3BGRL is important for v-Rel-mediated transformation**. *Oncogene* 2006, **25**(5):756-768.

37. Rocha S, Martin AM, Meek DW, Perkins ND: **p53 represses cyclin D1 transcription through down regulation of Bcl-3 and inducing increased association of the p52 NF-kappaB subunit with histone deacetylase 1**. *Mol Cell Biol* 2003, **23**(13):4713-4727.

38. Scott ML, Fujita T, Liou HC, Nolan GP, Baltimore D: **The p65 subunit of NF-kappa B regulates I kappa B by two distinct mechanisms**. *Genes Dev* 1993, **7**(7A):1266-1276.

39. Sun SC, Ganchi PA, Ballard DW, Greene WC: **NF-kappa B controls expression of inhibitor I kappa B alpha: evidence for an inducible autoregulatory pathway**. *Science* 1993, **259**(5103):1912-1915.

40. Walker WH, Stein B, Ganchi PA, Hoffman JA, Kaufman PA, Ballard DW, Hannink M, Greene WC: **The v-rel oncogene: insights into the mechanism of transcriptional activation, repression, and transformation**. *J Virol* 1992, **66**(8):5018-5029.

41. Chen LF, Greene WC: **Shaping the nuclear action of NF-kappaB**. *Nat Rev Mol Cell Biol* 2004, **5**(5):392-401.

42. Ghosh S, May MJ, Kopp EB: **NF-kappa B and Rel proteins: evolutionarily conserved mediators of immune responses**. *Annu Rev Immunol* 1998, **16**:225-260.

43. Gilmore TD: **Introduction to NF-kappaB: players, pathways, perspectives**. *Oncogene* 2006, **25**(51):6680-6684.

44. Natoli G: **Tuning up inflammation: how DNA sequence and chromatin organization control the induction of inflammatory genes by NF-kappaB**. *FEBS Lett* 2006, **580**(12):2843-2849.

45. Pahl HL: **Activators and target genes of Rel/NF-kappaB transcription factors**. *Oncogene* 1999, **18**(49):6853-6866.

46. Busse MS, Arnold CP, Towb P, Katrivesis J, Wasserman SA: **A kappaB sequence code for pathway-specific innate immune responses**. *Embo J* 2007, **26**(16):3826-3835.

47. Huang DB, Phelps CB, Fusco AJ, Ghosh G: **Crystal structure of a free kappaB DNA: insights into DNA recognition by transcription factor NF-kappaB**. *J Mol Biol* 2005, **346**(1):147-160.

48. Leung TH, Hoffmann A, Baltimore D: **One nucleotide in a kappaB site can determine cofactor specificity for NF-kappaB dimers**. *Cell* 2004, **118**(4):453-464.

49. Sanjabi S, Williams KJ, Saccani S, Zhou L, Hoffmann A, Ghosh G, Gerondakis S, Natoli G, Smale ST: **A c-Rel subdomain responsible for enhanced DNA-binding affinity and selective gene activation**. *Genes Dev* 2005, **19**(18):2138-2151.

50. Baeuerle PA: **The inducible transcription activator NF-kappa B: regulation by distinct protein subunits**. *Biochim Biophys Acta* 1991, **1072**(1):63-80.

51. Sauer T, Shelest E, Wingender E: **Evaluating phylogenetic footprinting for human-rodent comparisons**. *Bioinformatics* 2006, **22**(4):430-437.

52. Gupta M, Liu JS: **De novo cis-regulatory module elicitation for eukaryotic genomes**. *Proc Natl Acad Sci U S A* 2005, **102**(20):7079-7084.

53. Vavouri T, Walter K, Gilks WR, Lehner B, Elgar G: **Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans**. *Genome Biol* 2007, **8**(2):R15.

54. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV *et al*: **TRANSFAC: transcriptional regulation, from patterns to profiles**. *Nucleic Acids Res* 2003, **31**(1):374-378.

55. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JP, Widom J: **A genomic code for nucleosome positioning**. *Nature* 2006, **442**(7104):772-778.

56. Kunsch C, Ruben SM, Rosen CA: **Selection of optimal kappa B/Rel DNA-binding motifs: interaction of both subunits of NF-kappa B with DNA is required for transcriptional activation**. *Mol Cell Biol* 1992, **12**(10):4412-4421.

57. Schreck R, Zorbas H, Winnacker EL, Baeuerle PA: **The NF-kappa B transcription factor induces DNA bending which is modulated by its 65-kD subunit**. *Nucleic Acids Res* 1990, **18**(22):6497-6502.

58. Urban MB, Baeuerle PA: **The role of the p50 and p65 subunits of NF-kappa B in the recognition of cognate sequences**. *New Biol* 1991, **3**(3):279-288.

59. Zabel U, Schreck R, Baeuerle PA: **DNA binding of purified transcription factor NF-kappa B. Affinity, specificity, Zn2+ dependence, and differential half-site recognition**. *J Biol Chem* 1991, **266**(1):252-260.

60. Martone R, Euskirchen G, Bertone P, Hartman S, Royce TE, Luscombe NM, Rinn JL, Nelson FK, Miller P, Gerstein M *et al*: **Distribution of NF-kappaB-binding sites across human chromosome 22**. *Proc Natl Acad Sci U S A* 2003, **100**(21):12247-12252.

61. Hoffmann A, Leung TH, Baltimore D: **Genetic analysis of NF-kappaB/Rel transcription factors defines functional specificities**. *Embo J* 2003, **22**(20):5530-5539.

62. Zhou A, Scoggin S, Gaynor RB, Williams NS: **Identification of NF-kappa B-regulated genes induced by TNFalpha utilizing expression profiling and RNA interference**. *Oncogene* 2003, **22**(13):2054-2064.

63. Bunting K, Rao S, Hardy K, Woltring D, Denyer GS, Wang J, Gerondakis S, Shannon MF: **Genome-wide analysis of gene expression in T cells to identify**

targets of the NF-kappa B transcription factor c-Rel. *J Immunol* 2007, **178**(11):7097-7109.

64. Lim CA, Yao F, Wong JJ, George J, Xu H, Chiu KP, Sung WK, Lipovich L, Vega VB, Chen J *et al*: **Genome-wide mapping of RELA(p65) binding identifies E2F1 as a transcriptional activator recruited by NF-kappaB upon TLR4 activation**. *Mol Cell* 2007, **27**(4):622-635.

65. Linnell J, Mott R, Field S, Kwiatkowski DP, Ragoussis J, Udalova IA: **Quantitative high-throughput analysis of transcription factor binding specificities**. *Nucleic Acids Res* 2004, **32**(4):e44.

66. Nijnik A, Mott R, Kwiatkowski DP, Udalova IA: **Comparing the fine specificity of DNA binding by NF-kappaB p50 and p52 using principal coordinates analysis**. *Nucleic Acids Res* 2003, **31**(5):1497-1501.

67. Udalova IA, Mott R, Field D, Kwiatkowski D: **Quantitative prediction of NF-kappa B DNA-protein interactions**. *Proc Natl Acad Sci U S A* 2002, **99**(12):8167-8172.

68. Liu R, McEachin RC, States DJ: **Computationally identifying novel NF-kappa B-regulated immune genes in the human genome**. *Genome Res* 2003, **13**(4):654-661.

69. Alpaydin E: **Introduction to machine learning**. Cambridge, Mass.: MIT Press; 2004.

70. Quandt K, Frech K, Karas H, Wingender E, Werner T: **MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data**. *Nucleic Acids Res* 1995, **23**(23):4878-4884.

71. Djordjevic M, Sengupta AM: **Quantitative modeling and data analysis of SELEX experiments**. *Phys Biol* 2006, **3**(1):13-28.

72. Sengupta AM, Djordjevic M, Shraiman BI: **Specificity and robustness in transcription control networks**. *Proc Natl Acad Sci U S A* 2002, **99**(4):2072-2077.

73. Djordjevic M, Sengupta AM, Shraiman BI: **A biophysical approach to transcription factor binding site discovery**. *Genome Res* 2003, **13**(11):2381-2390.

74. O'Flanagan R: **Detecting Gene Regulation**. New Brunswick, NJ: Rutgers, The State University of New Jersey; 2005.

75. Cui Y, Wang Q, Stormo GD, Calvo JM: **A consensus sequence for binding of Lrp to DNA**. *J Bacteriol* 1995, **177**(17):4872-4880.

76. Benos PV, Bulyk ML, Stormo GD: **Additivity in protein-DNA interactions: how good an approximation is it?** *Nucleic Acids Res* 2002, **30**(20):4442-4451.

77. Berg OG, von Hippel PH: **Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters**. *J Mol Biol* 1987, **193**(4):723-750.

78. Heumann JM, Lapedes AS, Stormo GD: **Neural networks for determining protein specificity and multiple alignment of binding sites**. *Proc Int Conf Intell Syst Mol Biol* 1994, **2**:188-194.

79. Schneider TD: **Information content of individual genetic sequences**. *J Theor Biol* 1997, **189**(4):427-441.

80. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A: **Information content of binding sites on nucleotide sequences**. *J Mol Biol* 1986, **188**(3):415-431.

81. Stormo GD: **Probing information content of DNA-binding sites**. *Methods Enzymol* 1991, **208**:458-468.

82. Stormo GD: **Information content and free energy in DNA--protein interactions**. *J Theor Biol* 1998, **195**(1):135-137.

83. Stormo GD, Fields DS: **Specificity, free energy and information content in protein-DNA interactions**. *Trends Biochem Sci* 1998, **23**(3):109-113.

84. Stormo GD, Schneider TD, Gold L: **Quantitative analysis of the relationship between nucleotide sequence and functional activity**. *Nucleic Acids Res* 1986, **14**(16):6661-6679.

85. Halfon MS, Grad Y, Church GM, Michelson AM: **Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model**. *Genome Res* 2002, **12**(7):1019-1028.

86. Frech K, Danescu-Mayer J, Werner T: **A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter**. *J Mol Biol* 1997, **270**(5):674-687.

87. Klingenhoff A, Frech K, Quandt K, Werner T: **Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity**. *Bioinformatics* 1999, **15**(3):180-186.

88. Markstein M, Markstein P, Markstein V, Levine MS: **Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo**. *Proc Natl Acad Sci U S A* 2002, **99**(2):763-768.

89. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB: **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome**. *Proc Natl Acad Sci U S A* 2002, **99**(2):757-762.

90. van Nimwegen E: **Finding regulatory elements and regulatory motifs: a general probabilistic framework**. *BMC Bioinformatics* 2007, **8 Suppl 6**:S4.

91. Sinha S, Schroeder MD, Unnerstall U, Gaul U, Siggia ED: **Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila**. *BMC Bioinformatics* 2004, **5**:129.

92. Sinha S, van Nimwegen E, Siggia ED: **A probabilistic method to detect regulatory modules**. *Bioinformatics* 2003, **19 Suppl 1**:i292-301.

93. Rajewsky N, Vergassola M, Gaul U, Siggia ED: **Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo**. *BMC Bioinformatics* 2002, **3**:30.

94. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers**. *Proc Int Conf Intell Syst Mol Biol* 1994, **2**:28-36.

95. Bailey TL, Gribskov M: **Combining evidence using p-values: application to sequence homology searches**. *Bioinformatics* 1998, **14**(1):48-54.

96. Liu XS, Brutlag DL, Liu JS: **An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments**. *Nat Biotechnol* 2002, **20**(8):835-839.

97. Liu JS, Lawrence CE: **Bayesian inference on biopolymer models**. *Bioinformatics* 1999, **15**(1):38-52.

98. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment**. *Science* 1993, **262**(5131):208-214.

99. Neuwald AF, Liu JS, Lawrence CE: **Gibbs motif sampling: detection of bacterial outer membrane protein repeats**. *Protein Sci* 1995, **4**(8):1618-1632.

100. Roth FP, Hughes JD, Estep PW, Church GM: **Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation**. *Nat Biotechnol* 1998, **16**(10):939-945.

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

102. Liu X, Brutlag DL, Liu JS: **BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes**. *Pac Symp Biocomput* 2001:127-138.

103. Thompson W, Rouchka EC, Lawrence CE: **Gibbs Recursive Sampler: finding transcription factor binding sites**. *Nucleic Acids Res* 2003, **31**(13):3580-3585.

104. Hertz GZ, Stormo GD: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences**. *Bioinformatics* 1999, **15**(7-8):563-577.

105. Brazma A, Jonassen I, Vilo J, Ukkonen E: **Predicting gene regulatory elements in silico on a genomic scale**. *Genome Res* 1998, **8**(11):1202-1215.

106. Bussemaker HJ, Li H, Siggia ED: **Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis**. *Proc Natl Acad Sci U S A* 2000, **97**(18):10096-10100.

107. Sinha S, Tompa M: **A statistical method for finding transcription factor binding sites**. *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:344-354.

108. van Helden J, Andre B, Collado-Vides J: **Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies**. *J Mol Biol* 1998, **281**(5):827-842.

109. van Helden J, Rios AF, Collado-Vides J: **Discovering regulatory elements in non-coding sequences by analysis of spaced dyads**. *Nucleic Acids Res* 2000, **28**(8):1808-1818.

110. Berg J, Willmann S, Lassig M: **Adaptive evolution of transcription factor binding sites**. *BMC Evol Biol* 2004, **4**(1):42.

111. Gerland U, Hwa T: **On the selection and evolution of regulatory DNA motifs**. *J Mol Evol* 2002, **55**(4):386-400.

112. Gerland U, Moroz JD, Hwa T: **Physical constraints and functional characteristics of transcription factor-DNA interaction**. *Proc Natl Acad Sci U S A* 2002, **99**(19):12015-12020.

113. Lassig M: **From biophysics to evolutionary genetics: statistical aspects of gene regulation**. *BMC Bioinformatics* 2007, **8 Suppl 6**:S7.

114. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I *et al*: **Transcriptional regulatory networks in Saccharomyces cerevisiae**. *Science* 2002, **298**(5594):799-804.

115. Hughes JD, Estep PW, Tavazoie S, Church GM: **Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae**. *J Mol Biol* 2000, **296**(5):1205-1214.

116. Holloway DT, Kon M, DeLisi C: **Integrating genomic data to predict transcription factor binding**. *Genome Inform Ser Workshop Genome Inform* 2005, **16**(1):83-94.

117. Pritsker M, Liu YC, Beer MA, Tavazoie S: **Whole-genome discovery of transcription factor binding sites by network-level conservation**. *Genome Res* 2004, **14**(1):99-108.

118. Bussemaker HJ, Li H, Siggia ED: **Regulatory element detection using correlation with expression**. *Nat Genet* 2001, **27**(2):167-171.

119. Rabiner LR: **A tutorial on hidden Markov models and selected applications in speech recognition**. *Proc IEEE* 1989, **77**(2):257-286.

120. Ewens WJ, Grant GR: **Statistical Methods in Bioinformatics: An Introduction**, 1 edn: Springer; 2001.

121. Eddy SR: **Hidden Markov models**. *Curr Opin Struct Biol* 1996, **6**(3):361-365.

122. Eddy SR: **Profile hidden Markov models**. *Bioinformatics* 1998, **14**(9):755-763.

123. Krogh A: **An introduction to Hidden Markov Models for biological sequences**. In: *Computational methods in molecular biology*. Edited by Salzberg SL, Searls DB, Kasif S. New York: Elsevier; 1998: 45-63.

124. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: **Hidden Markov models in computational biology. Applications to protein modeling**. *J Mol Biol* 1994, **235**(5):1501-1531.

125. Suzuki T, Yoshimura H, Ehira S, Ikeuchi M, Ohmori M: **AnCrpA, a cAMP receptor protein, regulates nif-related gene expression in the cyanobacterium Anabaena sp. strain PCC 7120 grown with nitrate**. *FEBS Lett* 2007, **581**(1):21-28.

126. Varga G, Su C: **Classification and predictive modeling of liver X receptor response elements**. *BioDrugs* 2007, **21**(2):117-124.

127. Conkright MD, Guzman E, Flechner L, Su AI, Hogenesch JB, Montminy M: **Genome-wide analysis of CREB target genes reveals a core promoter requirement for cAMP responsiveness**. *Mol Cell* 2003, **11**(4):1101-1108.

128. Stegmaier P, Kel AE, Wingender E: **Systematic DNA-binding domain classification of transcription factors**. *Genome Inform* 2004, **15**(2):276-286.

129. Frith MC, Hansen U, Weng Z: **Detection of cis-element clusters in higher eukaryotic DNA**. *Bioinformatics* 2001, **17**(10):878-889.

130. Frith MC, Spouge JL, Hansen U, Weng Z: **Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences**. *Nucleic Acids Res* 2002, **30**(14):3214-3224.

131. Grundy WN, Bailey TL, Elkan CP, Baker ME: **Meta-MEME: motif-based hidden Markov models of protein families**. *Comput Appl Biosci* 1997, **13**(4):397-406.

132. Bailey TL, Noble WS: **Searching for statistically significant regulatory modules**. *Bioinformatics* 2003, **19 Suppl 2**:ii16-25.

133. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S *et al*: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes**. *Genome Res* 2005, **15**(8):1034-1050.

134. Emberly E, Rajewsky N, Siggia ED: **Conservation of regulatory elements between two species of Drosophila**. *BMC Bioinformatics* 2003, **4**:57.

135. Zhou Q, Wong WH: **Coupling hidden Markov models for the discovery of Cis-regulatory modules in multiple species**. *The Annals of Applied Statistics* 2007, **1**(1):36-65.

136. Siepel A, Haussler D: **Combining phylogenetic and hidden Markov models in biosequence analysis**. *J Comput Biol* 2004, **11**(2-3):413-428.

137. Wong WS, Nielsen R: **Finding cis-regulatory modules in Drosophila using phylogenetic hidden Markov models**. *Bioinformatics* 2007, **23**(16):2031-2037.

138. Drawid A, Gupta N, Nagaraj V, Gelinas C, Sengupta AM: **A hidden Markov model with location-dependent transition probabilities accurately predicts the occupancy of a transcription factor with self-overlapping binding sites**. *In preparation*.

139. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Res* 2007, **35**(Database issue):D61-65.

140. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.

141. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ: **The UCSC Table Browser data retrieval tool**. *Nucleic Acids Res* 2004, **32**(Database issue):D493-496.

142. Tabach Y, Brosh R, Buganim Y, Reiner A, Zuk O, Yitzhaky A, Koudritsky M, Rotter V, Domany E: **Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site**. *PLoS ONE* 2007, **2**(8):e807.

143. Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES: **Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites**. *Proc Natl Acad Sci U S A* 2007, **104**(17):7145-7150.

144. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T *et al*: **Ensembl 2007**. *Nucleic Acids Res* 2007, **35**(Database issue):D610-617.

145. Sherman BT, Huang da W, Tan Q, Guo Y, Bour S, Liu D, Stephens R, Baseler MW, Lane HC, Lempicki RA: **DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis**. *BMC Bioinformatics* 2007, **8**:426.

146. Huang da W, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, Guo Y, Stephens R, Baseler MW, Lane HC *et al*: **DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W169-175.

147. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery**. *Genome Biol* 2003, **4**(5):P3.

148. Brummelkamp TR, Nijman SM, Dirac AM, Bernards R: **Loss of the cylindromatosis tumour suppressor inhibits apoptosis by activating NF-kappaB**. *Nature* 2003, **424**(6950):797-801.

149. Kovalenko A, Chable-Bessia C, Cantarella G, Israel A, Wallach D, Courtois G: **The tumour suppressor CYLD negatively regulates NF-kappaB signalling by deubiquitination**. *Nature* 2003, **424**(6950):801-805.

150. Krikos A, Laherty CD, Dixit VM: **Transcriptional activation of the tumor necrosis factor alpha-inducible zinc finger protein, A20, is mediated by kappa B elements**. *J Biol Chem* 1992, **267**(25):17971-17976.

151. Trompouki E, Hatzivassiliou E, Tsichritzis T, Farmer H, Ashworth A, Mosialos G: **CYLD is a deubiquitinating enzyme that negatively regulates NF-kappaB activation by TNFR family members**. *Nature* 2003, **424**(6950):793-796.

152. Osipo C, Golde TE, Osborne BA, Miele LA: **Off the beaten pathway: the complex cross talk between Notch and NF-kappaB**. *Lab Invest* 2008, **88**(1):11-17.

153. Bakkar N, Wang J, Ladner KJ, Wang H, Dahlman JM, Carathers M, Acharyya S, Rudnicki MA, Hollenbach AD, Guttridge DC: **IKK/NF-kappaB regulates skeletal myogenesis via a signaling switch to inhibit differentiation and promote mitochondrial biogenesis**. *J Cell Biol* 2008, **180**(4):787-802.

154. Memet S: **NF-kappaB functions in the nervous system: from development to disease**. *Biochem Pharmacol* 2006, **72**(9):1180-1195.

155. Darwin C: **On the origin of species by means of natural selection, or The preservation of favoured races in the struggle for life**. London: J. Murray; 1859.

156. **The National Biological Information Infrastructure** [http://www.nbii.gov/]

157. **The Genome Sequencing Center at Washington University in St. Louis School of Medicine** [http://www.genome.wustl.edu/]

158. Ureta-Vidal A, Ettwiller L, Birney E: **Comparative genomics: genome-wide analysis in metazoan eukaryotes**. *Nat Rev Genet* 2003, **4**(4):251-262.

159. Wildman DE, Uddin M, Opazo JC, Liu G, Lefort V, Guindon S, Gascuel O, Grossman LI, Romero R, Goodman M: **Genomics, biogeography, and the diversification of placental mammals**. *Proc Natl Acad Sci U S A* 2007, **104**(36):14395-14400.

160. Kriegs JO, Churakov G, Kiefmann M, Jordan U, Brosius J, Schmitz J: **Retroposed elements as archives for the evolutionary history of placental mammals**. *PLoS Biol* 2006, **4**(4):e91.

161. Murphy WJ, Pringle TH, Crider TA, Springer MS, Miller W: **Using genomic data to unravel the root of the placental mammal phylogeny**. *Genome Res* 2007, **17**(4):413-421.

162. Springer MS, Meredith RW, Eizirik E, Teeling E, Murphy WJ: **Morphology and placental mammal phylogeny**. *Syst Biol* 2008, **57**(3):499-503.

163. Springer MS, Murphy WJ, Eizirik E, O'Brien SJ: **Placental mammal diversification and the Cretaceous-Tertiary boundary**. *Proc Natl Acad Sci U S A* 2003, **100**(3):1056-1061.

164. Springer MS, Stanhope MJ, Madsen O, de Jong WW: **Molecules consolidate the placental mammal tree**. *Trends Ecol Evol* 2004, **19**(8):430-438.

165. Waddell PJ, Kishino H, Ota R: **A phylogenetic foundation for comparative mammalian genomics**. *Genome Inform* 2001, **12**:141-154.

166. Nei M, Kumar S: **Molecular evolution and phylogenetics**. 2000:xiv, 333 p.

167. Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC: **Cross-species sequence comparisons: a review of methods and available resources**. *Genome Res* 2003, **13**(1):1-12.

168. Mahony S, Corcoran DL, Feingold E, Benos PV: **Regulatory conservation of protein coding and microRNA genes in vertebrates: lessons from the opossum genome**. *Genome Biol* 2007, **8**(5):R84.

169. Dubchak I, Frazer K: **Multi-species sequence comparison: the next frontier in genome annotation**. *Genome Biol* 2003, **4**(12):122.

170. Cooper GM, Brudno M, Green ED, Batzoglou S, Sidow A: **Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes**. *Genome Res* 2003, **13**(5):813-820.

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

172. Pollard DA, Bergman CM, Stoye J, Celniker SE, Eisen MB: **Benchmarking tools for the alignment of functional noncoding DNA**. *BMC Bioinformatics* 2004, **5**:6.

173. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M: **Finding functional features in Saccharomyces genomes by phylogenetic footprinting**. *Science* 2003, **301**(5629):71-76.

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

175. McCue L, Thompson W, Carmack C, Ryan MP, Liu JS, Derbyshire V, Lawrence CE: **Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes**. *Nucleic Acids Res* 2001, **29**(3):774-782.

176. McGuire AM, Hughes JD, Church GM: **Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes**. *Genome Res* 2000, **10**(6):744-757.

177. Blanchette M, Schwikowski B, Tompa M: **Algorithms for phylogenetic footprinting**. *J Comput Biol* 2002, **9**(2):211-223.

178. Blanchette M, Tompa M: **Discovery of regulatory elements by a computational method for phylogenetic footprinting**. *Genome Res* 2002, **12**(5):739-748.

179. Grad YH, Roth FP, Halfon MS, Church GM: **Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in Drosophila melanogaster and D.pseudoobscura**. *Bioinformatics* 2004, **20**(16):2738-2750.

180. Pierstorff N, Bergman CM, Wiehe T: **Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA**. *Bioinformatics* 2006, **22**(23):2858-2864.

181. Berman BP, Pfeiffer BD, Laverty TR, Salzberg SL, Rubin GM, Eisen MB, Celniker SE: **Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura**. *Genome Biol* 2004, **5**(9):R61.

182. Wang T, Stormo GD: **Combining phylogenetic data with co-regulated genes to identify regulatory motifs**. *Bioinformatics* 2003, **19**(18):2369-2380.

183. Bais AS, Grossmann S, Vingron M: **Incorporating evolution of transcription factor binding sites into annotated alignments**. *J Biosci* 2007, **32**(5):841-850.

184. Halpern AL, Bruno WJ: **Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies**. *Mol Biol Evol* 1998, **15**(7):910-917.

185. Moses AM, Chiang DY, Kellis M, Lander ES, Eisen MB: **Position specific variation in the rate of evolution in transcription factor binding sites**. *BMC Evol Biol* 2003, **3**:19.

186. Moses AM, Chiang DY, Pollard DA, Iyer VN, Eisen MB: **MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model**. *Genome Biol* 2004, **5**(12):R98.

187. Erb I, van Nimwegen E: **Statistical features of yeast's transcriptional regulatory code**. *IEEE Proceedings Systems Biology ICCSB* 2006.

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

189. Prakash A, Blanchette M, Sinha S, Tompa M: **Motif discovery in heterogeneous sequence data**. *Pac Symp Biocomput* 2004:348-359.

190. Moses AM, Chiang DY, Eisen MB: **Phylogenetic motif detection by expectation-maximization on evolutionary mixtures**. *Pac Symp Biocomput* 2004:324-335.

191. Li X, Wong WH: **Sampling motifs on phylogenetic trees**. *Proc Natl Acad Sci U S A* 2005, **102**(27):9481-9486.

192. Liu Y, Liu XS, Wei L, Altman RB, Batzoglou S: **Eukaryotic regulatory element conservation analysis and identification using comparative genomics**. *Genome Res* 2004, **14**(3):451-458.

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

194. Doniger SW, Fay JC: **Frequent gain and loss of functional transcription factor binding sites**. *PLoS Comput Biol* 2007, **3**(5):e99.

195. Mustonen V, Lassig M: **Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies**. *Proc Natl Acad Sci U S A* 2005, **102**(44):15936-15941.

196. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic Acids Res* 1994, **22**(22):4673-4680.

197. Berg OG: **The evolutionary selection of DNA base pairs in gene-regulatory binding sites**. *Proc Natl Acad Sci U S A* 1992, **89**(16):7501-7505.

198. Kimura M: **On the probability of fixation of mutant genes in a population**. *Genetics* 1962, **47**:713-719.

199. Kimura M, Ohta T: **The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population**. *Genetics* 1969, **61**(3):763-771.

200. Ohta T: **Near-neutrality in evolution of genes and gene regulation**. *Proc Natl Acad Sci U S A* 2002, **99**(25):16134-16137.

201. Ohta T, Tachida H: **Theoretical study of near neutrality. I. Heterozygosity and rate of mutant substitution**. *Genetics* 1990, **126**(1):219-229.

202. Brown CT, Callan CG, Jr.: **Evolutionary comparisons suggest many novel cAMP response protein binding sites in Escherichia coli**. *Proc Natl Acad Sci U S A* 2004, **101**(8):2404-2409.

203. Mustonen V, Kinney J, Callan CG, Jr., Lassig M: **Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites**. *Proc Natl Acad Sci U S A* 2008, **105**(34):12376-12381.

204. Drawid A, Sengupta AM: **PhyloQPMEME: Integrating evolutionary conservation with a biophysical model to identify transcription factor binding sites**. *In preparation*.

205. Nocedal J, Wright SJ: **Numerical optimization**. New York: Springer; 1999.

206. Burges CJC: **A tutorial on support vector machines for pattern recognition**. *Data Mining and Knowledge Discovery* 1998, **2**:121--167.

207. Morgenstern B, Frech K, Dress A, Werner T: **DIALIGN: finding local similarities by multiple sequence alignment**. *Bioinformatics* 1998, **14**(3):290-294.

208. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S: **LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA**. *Genome Res* 2003, **13**(4):721-731.

209. Sun B, Karin M: **NF-kappaB signaling, liver disease and hepatoprotective agents**. *Oncogene* 2008, **27**(48):6228-6244.

210. Suh Y, Afaq F, Johnson JJ, Mukhtar H: **A plant flavonoid fisetin induces apoptosis in colon cancer cells by inhibition of COX2 and Wnt/EGFR/NF-{kappa}B signaling pathways**. *Carcinogenesis* 2008.

Curriculum Vita

AMAR MOHAN DRAWID

EDUCATION

Yale University 1996-2000
B.S., M.S. in Molecular Biophysics and Biochemistry

Rutgers University 2005-2009
Ph.D. in Computational Biology and Molecular Biophysics

PROFESSIONAL EXPERIENCE

Sanofi-Aventis, Bridgewater, NJ 2000-Present
Research Investigator

PUBLICATIONS

Drawid A, Gerstein M. A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. J Mol Biol. 2000 Aug 25; 301(4): 1059-75.

Drawid A, Jansen R, Gerstein M. Genome-wide analysis relating expression level with protein subcellular localization. Trends Genet. 2000 Oct; 16(10): 426-30.

Kumar A, Agarwal S, Heyman JA, Matson S, Heidtman M, Piccirillo S, Umansky L, Drawid A, Jansen R, Liu Y, Cheung KH, Miller P, Gerstein M, Roeder GS, Snyder M. Subcellular localization of the yeast proteome. Genes Dev. 2002 Mar 15; 16(6): 707-19.

Daheshia M, Tian N, Connolly T, Drawid A, Wu Q, Bienvenu J, Cavallo J, Jupp R, De Sanctis G, Minnich A. Molecular characterization of antigen induced lung inflammation in a murine model of asthma. Ann N Y Acad Sci. 2002 Dec; 975: 148-59.

Gupta N, Delrow J, Drawid A, Sengupta AM, Fan G, Gelinas C: Repression of B-cell linker (BLNK) and B-cell adaptor for phosphoinositide 3-kinase (BCAP) is important for lymphocyte transformation by rel proteins. Cancer Res 2008, 68(3):808-814.

Drawid A, Gupta N, Nagaraj V, Gelinas C, Sengupta AM: A hidden Markov model with location-dependent transition probabilities accurately predicts the occupancy of a transcription factor with self-overlapping binding sites. In preparation.

Drawid A, Sengupta AM: PhyloQPMEME: Integrating evolutionary conservation with a biophysical model to identify transcription factor binding sites. In preparation.