Predicting transcription factor binding sites using phylogenetic footprinting and a probabilistic framework for evolutionary turnover

Victor Parmar

Master of Science

School of Computer Science

McGill University

Montréal,Québec

2009-08-15

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

DEDICATION

This document is dedicated to the graduate students of McGill University.

ACKNOWLEDGEMENTS

I would like to thank first and foremost my supervisor, Dr. Mathieu Blanchette for introducing me to the field of bioinformatics and the topic of transcription factor binding site turnover. I am ever grateful for his guidance and support throughout the project and for providing the reconstructed ancestral genome and other data critical to our research. I would also like to thank my parents for their understanding and unwavering faith.

ABSTRACT

Identifying genomic locations of transcription-factor binding sites (TFBS), particularly in higher eukaryotic genomes, has been an enormous challenge. Computational methods involving identification of sequence conservation between related genomes have been the most successful since sites found in such highly conserved regions are more likely to be functional, i.e. are bound and regulate protein production. In this thesis, we present such a probabilistic algorithm for predicting TFBSs which also takes evolutionary turnovers into account. Our algorithm is validated via simulations and the results of its application on ChIP-chip data are presented.

ABRÉGÉ

L'identification des sites de fixation des facteurs de transcription (TFBS), particulièrement sur les génomes eucaryotiques plus élevés, a été un énorme défi. Les méthodes informatiques comportant l'identification de la conservation de séquence entre les génomes de différentes espèces ont eu beaucoup de succès parce que les sites trouvés dans de telles régions fortement conservées sont probablement fonctionnels (les facteurs de transcription se rajoutent sur le génome à ces sites-là et règlent la production de protéine). Dans cette thèse, nous présentons un algorithme probabiliste pour la prédiction de TFBSs qui prend en considération également le remuement évolutionnaire. Notre algorithme est validé par l'intermédiaire des simulations et les résultats de son application sur des données ChIP-chip sont présentés.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

Life is specified by genomes. Every organism, including humans, has a genome that contains all of the biological information needed to build and maintain a living individual of that organism. The biological information contained in a genome is encoded in its deoxyribonucleic acid (DNA) molecules, which is divided into discrete units called *genes* [42]. Genes are segments of DNA and between them they direct the physical development of an organism by coding proteins.

1.1 Physical structure of the human genome

DNA itself is a long molecule that looks like a twisted ladder (Figure 1–1) and is made up of four types of simple units called nucleotides. These are the repeating units in the DNA and form the "rungs" of the DNA ladder. There are four types of nucleotides: Adenine (A), Cytosine (C), Guanine (G), Thymine (T) and it is the sequence of these nucleotides that carries information, just as the sequence of letters carries information on a page.

All DNA in a genome is packaged into chromosomes, each of which contains a single long piece of DNA that is wound up and bunched together into a compact structure. The genome itself is a collection of such chromosomes.

1.1.1 The central dogma of molecular biology

Although DNA is the carrier of genetic information in a cell, proteins do the bulk of the work. Each cell contains thousands of different proteins: enzymes that

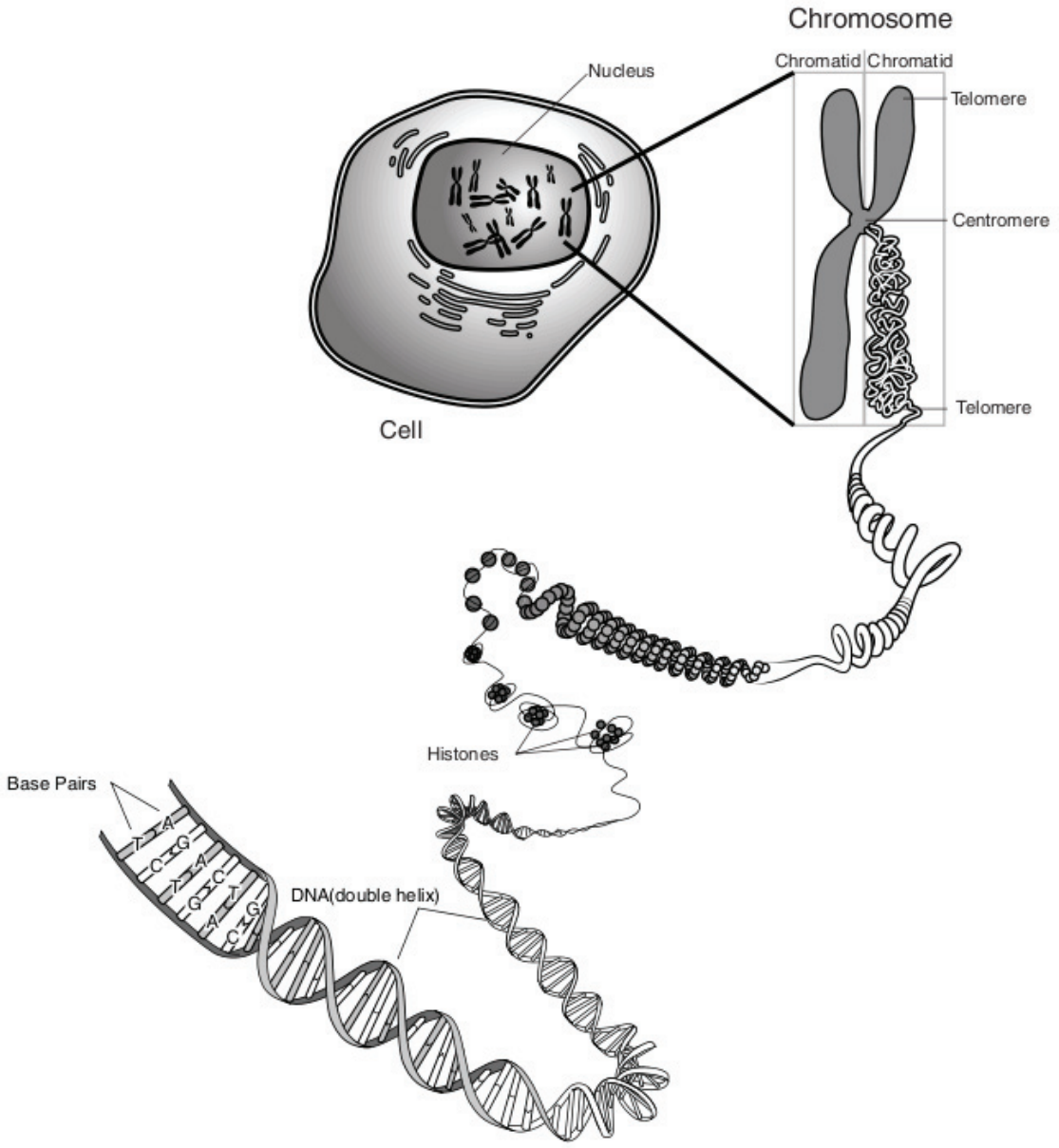

Figure 1–1: DNA Structure [44]

make new molecules and catalyze nearly all chemical processes in cells; structural components that give cells their shape and help them move; hormones that transmit signals throughout the body; antibodies that recognize foreign molecules; and transport molecules that carry other molecules such as oxygen. The genetic code carried by DNA is what specifies the shape and function of the protein.

The "Central Dogma" refers to the flow of genetic information in biological systems. In general, genetic information flows from DNA to RNA[1] to protein [15]. In order to use the information present in DNA to produce proteins, the DNA must first be converted into a messenger RNA (mRNA) through a process called **transcription**. The information carried by the mRNA is then used to construct a specific protein (or polypeptide) through a process called **translation** (Figure 1–2).

Transcription is carried out by an enzyme called the RNA polymerase. This molecule has the job of recognizing the DNA sequence where transcription is initiated, called the "promoter" region. In general, promoter regions are usually found upstream from the beginning of every gene. While the composition of bases varies between promoters, they are recognized by the RNA polymerase complex, which can grab hold of the sequence and start the production of an mRNA [18].

[1] Ribonucleic acid (RNA) is a nucleic acid similar to DNA but is single-stranded, consists of ribose sugar rather than deoxyribose sugar and uracil (U) replaces thymine (T) as one of the bases. RNA plays an important role in protein synthesis and other chemical activities of the cell. There are also several classes of RNA molecules, including messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA) and other small RNAs.

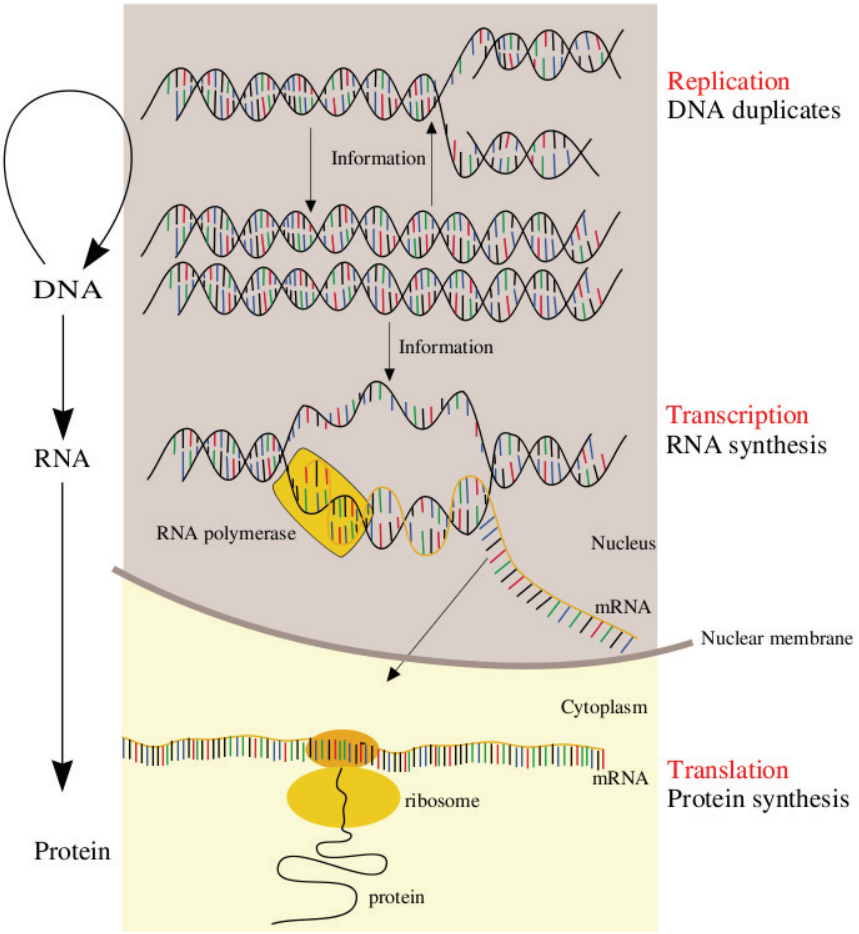

Figure 1–2: Central dogma of molecular biology [56]

1.2 Logical structure of the human genome

Genes make up about 1.5 percent of the total DNA in our genome [32]. In the human genome, the coding portions of a gene, called **exons**, are interrupted by intervening sequences, called **introns** (Figure 1–3). A eukaryotic gene does not code for a protein in one continuous stretch of DNA. Both exons and introns are "transcribed" into pre-mRNA, but before it is exported to the nucleus for translation, the primary mRNA transcript is edited. This editing process is called splicing and it describes the removal of the introns, joining of the exons together, and the addition of unique features to each end of the transcript to make a mature mRNA [42].

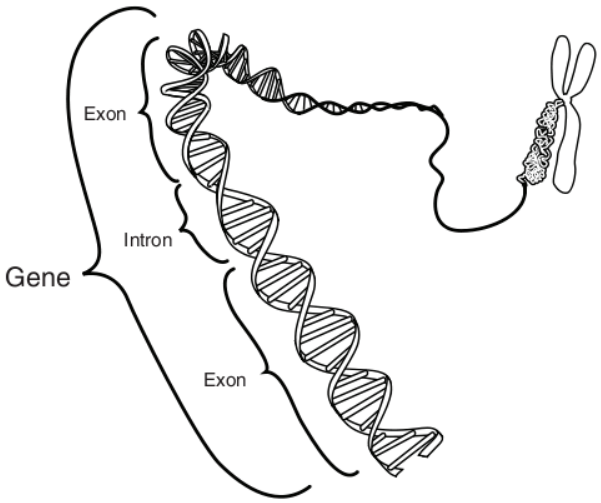

Figure 1–3: Gene [45]

The evolutionary conservation across the mammalian genomes of much more sequence than can be explained by protein-coding regions indicates that many,

and perhaps most, functional sequences in the genome remain unknown [61]. Regulatory sequences (such as promoter regions) are a part of this evolutionary conservation and while they make up only about 2% of the human genome, it is probably more than the percentage of genes in the human genome.

Forty to forty-five percent of our genome is made up of short non-protein coding (or non-coding) sequences that are repeated, sometimes millions of times [12]. There are numerous forms of this "repetitive DNA", and a few have known functions, such as stabilizing the chromosome structure or inactivating one of the two X chromosomes in developing females, a process called X-inactivation [1]. Of this 40-50% of repetitive DNA, most of it consists of transposable elements which are sequences of DNA that can move around to different positions within the genome of a single cell causing mutations and changing the amount of DNA in the genome. Another class of highly repeated sequences found so far in mammals are called "satellite DNA" because their unusual composition allows them to be easily separated from other DNA. These sequences are associated with chromosome structure and are found at the **centromeres** (or centers) and **telomeres** (ends) of chromosomes. Although they do not play a role in the coding of proteins, they do play a significant role in chromosome structure, duplication, and cell division [64].

While the publication of a nearly complete draft sequence of the human genome is an enormous achievement, characterizing the entire set of functional elements encoded in the human and other genomes remains an immense challenge [14, 32].

1.3 Transcription Factors

A transcription factor (TF; sometimes called a sequence-specific DNA binding factor) is a protein that binds to specific sequences of DNA and thereby controls the transfer (or transcription) of genetic information from DNA to RNA [33]. Transcription factors perform this function alone, or with other proteins in a complex, by promoting (as an activator), or blocking (as a repressor) the recruitment of RNA polymerase (the enzyme that performs the transcription of genetic information from DNA to RNA) to specific genes [49, 51] (Figure 1–2).

Since transcription factors bind to the genome, one of the most important functional elements in any genome are the *sites* within the DNA to which they bind. Defects in transcription factors or interferences in their interaction with DNA can contribute to the progression of various diseases. For example, Durbin et al. showed that targeted disruption of the mouse *Stat1* gene results in compromised innate immunity to viral disease [19]. According to Bulyk [9], a more complete understanding of transcription factors, their DNA binding sites, and their interactions, would permit a comprehensive and quantitative mapping of the regulatory pathways within cells, as well as a deeper understanding of the potential functions of individual genes regulated by newly identified DNA-binding sites.

1.4 Transcription Factor Binding Sites

Transcription factor binding sites (TFBSs) are usually short (around 5-15 base pairs) and they are frequently degenerate sequence motifs, that is, a given transcription factor can bind to a family of similar sequences. While they do follow certain patterns, they are often quite degenerate. Thus, the most common

computational representation of a TFBS is by a single consensus sequence. The consensus sequence simply gives the nucleotide that is found the most often in each position. Another representation is via an alternate (or degenerate) consensus sequence which gives the possible nucleotides in each position. For example, the alternate consensus sequence `ARYCGN` means that the first nucleotide is always A, the second could be A or G, similarly the third nucleotide could be C or T, the fourth and fifth bases are C and G respectively, and finally N represents any nucleotide in the last position. The sequence degeneracy of TFBSs has been selected through evolution and is beneficial, because it confers different levels of activity upon different promoters, thus causing some genes to be transcribed at higher levels than others, as may be required by the cell [53]. Another example of a consensus sequence is presented in Figure 1–4a.

From a computational perspective therefore, potential binding sites thus can occur very frequently in larger genomes such as the human genome. Moreover, the experimental identification of regulatory regions in higher eukaryotes is more difficult than in organisms with smaller genomes, partly because of the larger genome size, because a larger portion of higher genomes is non-coding, and because even the general principles governing the locations of DNA regulatory elements in higher eukaryotic genomes remain poorly understood. For example, regulatory elements can be found far upstream of coding regions, within introns, and even far downstream of the genes they regulate, making the search for them difficult [9] (Figure 1–5).

Despite these challenges, many experimental and computational approaches have been developed to identify such sites where transcription factors bind on the genome to regulate transcription. These are described in the following sections.

1.4.1 Identifying candidate TFBSs *in silico*

Although degenerate consensus sequences (also known as motifs), are still frequently used to depict the binding specificities of TFs, they do not contain precise information about the relative likelihood of observing the alternate nucleotides at the various positions of a TFBS. Thus, a common way of representing the degenerate sequence preferences of a DNA-binding protein is by a position weight matrix (PWM). A PWM, again like a consensus sequence, is based on a set of sequences for a TFBS. The difference is that it is a matrix with the rows representing the four nucleotides: A, C, G, T and the number of columns depending on the size of the TFBS. The elements of a PWM correspond to scores reflecting the likelihood of observing that particular nucleotide at that particular position of the known or candidate TFBS. The score for each nucleotide at each position is derived from the observed frequency of that nucleotide at the corresponding position in the input set of promoters. Thus, if in a set of 10 TFBS sequences, the nucleotide in the first position is C in eight sequences with G and A found in the other two sequences, the first column of the PWM would have 1, 8, 1, 0 for A, C, G, T rows respectively. The score for any particular site is the sum of the individual matrix values for that site's sequence (Figure 1–4b).

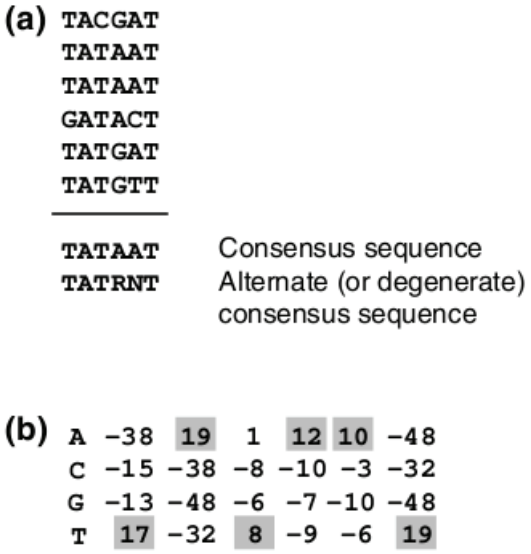

Figure 1–4: Representation of transcription factor binding sites. (a) An example of six sequences and the consensus sequence that can be derived from them. (b) A position weight matrix for the -10 region of *E. coli* TATA-box, as an example of a well-studied regulatory element. The sequence TATAAT thus gets a score of of 85 $(17 + 19 + 8 + 12 + 10 + 19)$. Note that the matrix values in (b) do not come from the example shown in (a) but rather are derived from a much larger collection of sequences. [9, 53].

1.4.2 Motif Scanning

The availability of consensus target sequences for many of the known transcription factors has been used to construct databases that can be used by computer algorithms to search for and identify novel regulatory elements in nucleotide sequences. At present, the most widely used transcription factor database is TRANSFAC [62], which catalogs eukaryotic TFs and their known binding sites, and provides PWMs.

Given a TF sequence motif, a simple approach would be to search the complete genome for sequences (sites) that result in scores above a certain threshold. Unfortunately, given the short length and degenerate nature of TFBSs and the size of the human genome, the output of such a simple approach results in a large number of false-positives. For instance, the unambiguous sequence TATAA is expected once every 1,024 basepairs (bp) by chance, which predicts 3 million potential binding sites in a mammalian genome, thus leading to a vast majority of them being biologically non-functional predictions [46]. The choice of PWM score cutoffs is a critical issue in all predictions of sites from PWMs, as the requirement for a more stringent match (a higher cutoff) is likely to result in fewer false-positive predictions but can potentially result in more sites being missed (false-negatives).

Various improved algorithms have been developed to sift through the output of a transcription factor database search to decrease the number of false-positive returned. Power can be gained by taking advantage of the sequence context in which a predicted binding site is found. In the TATAA example, higher statistical scores can be assigned if the site is found within 25 to 30 bp of a predicted transcription initiation sequence. Predictions can be further strengthened if a transcription factor is known to function as a dimer (Figure 1–5), and two similar adjacent binding sites are found [46].

Along with motif scanning, inter-species sequence comparisons have been used to identify non-coding sequences that have a reasonable likelihood of having gene regulatory properties [20, 34], the most prominent being phylogenetic footprinting.

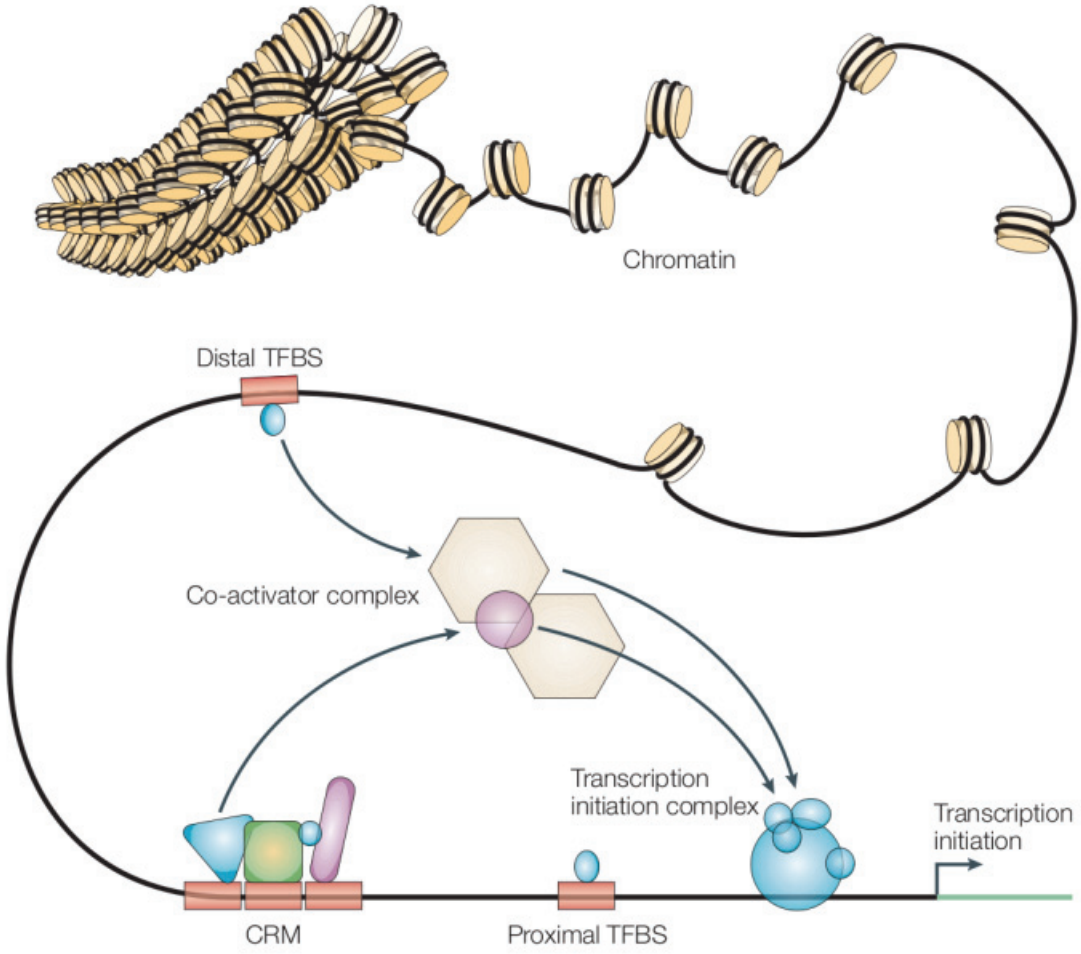

Figure 1–5: Transcription factors bind to specific sites that are either proximal or distal to a transcription start site. Sets of TFs can operate in functional *cis*-regulatory modules (CRMs) to achieve specific regulatory properties [60].

1.4.3 Phylogenetic Footprinting

Phylogenetic footprinting [54] is a method wherein regulatory elements, specifically TFBSs, associated with a given gene are identified by considering a set of orthologous[2] non-coding sequences from a group of related species. If these sequences contain regions that are unusually well conserved, it is reasonable to conclude that these regions have some regulatory function. According to Blanchette et al. [5], functional sequences tend to evolve much more slowly than nonfunctional sequences as they are subject to selective pressure and it is this difference in mutation rates that phylogenetic footprinting exploits. The phylogenetic footprinting approach has proved successful for the discovery of regulatory elements for many genes, including *ϵ-globin* (Tagle et al. [54]; Gumucio et al. [27]), *γ-globin* (Tagle et al. [54]), *rbcL* (Manen et al. [38]), *cystic fibrosis transmembrane conductance regulator* (Vuillaumier et al. [57]), and *interleukin (IL)-4*, *IL-13*, and *IL-5* (Loots et al. [34]). The same idea of using comparative analysis to identify conserved elements, but among only two or three species (particularly human and mouse), was made initially made popular by Hardison et al. [28], Wasserman et al. [59] and Wu et al. [63]. Figure 1–6 presents an example of TFBS identification using phylogenetic footprinting.

[2] Orthologs are sequences in different species that are derived from the same sequence in the last common ancestral species and thus usually have similar functions in the genomes being compared.

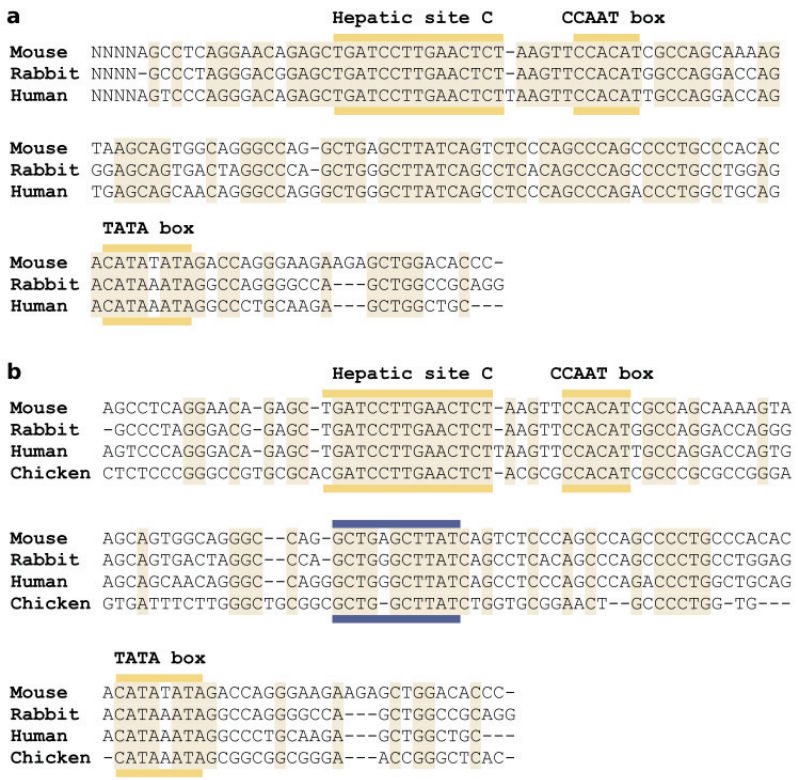

Figure 1–6: An example of TFBS identification of TFBS using multi-species comparative genomic sequence analysis. The region under consideration is upstream of the well-studied apolipoprotein AI (*ApoAI*) gene. (a) Comparison of roughly 150bp upstream of the predicted *ApoAI* transcription start site in human, mouse and rabbit. This comparison indicates high levels of sequence conservation across the entire region in these mammals, making it difficult clearly identify any sequences as TFBSs. (b) Addition of the orthologous region of the chicken *ApoAI* gene. This decreased the level of conservation greatly but this also resulted in the high levels of conservation in regions previously shown to be important in gene regulation (yellow). Both the CCAAT box and the TATA box, important in core promoter activity, are almost perfectly conserved across all four species. In addition, hepatic enhancer site C, experimentally shown to be necessary for *ApoAI* liver expression, reveals strong sequence conservation (14 of 15 bp are conserved across all 4 species). The other novel conserved block (blue) that was revealed by comparative analysis was not assigned a biological function, Pennacchio et al. [46]

The major advantage of phylogenetic footprinting over the motif scanning approach is that the latter requires a reliable method for motif discovery whereas phylogenetic footprinting is capable of identifying regulatory elements as long as they are sufficiently conserved across many of the species considered [6]. Moreover, as sequence data for more species has become available, the accuracy of phylogenetic footprinting methods has greatly increased.

Phylogenetic footprinting works by constructing a global multiple alignment of the orthologous regulatory sequences and the subsequent identification of conserved regions in the alignment. A tool such as CLUSTALW [55] is appropriate for this purpose, as it can take advantage of knowledge of the phylogeny relating the species. Unfortunately this approach is not always successful due to the short lengths of TFBSs. Regulatory elements in general tend to be short (5 to 20 nucleotides long) relative to the entire regulatory region in which the search for them is conducted (typically, a 1000-bp promoter region). If the species are somewhat diverged, chances are that the the noise of the diverged nonfunctional background will overcome the short conserved signal resulting in a mis-alignment of the short regulatory sequences. In this case, the regulatory elements would not appear to belong to conserved regions and would go undetected and thus, for regions that are moderately to highly diverged, global multiple alignment is likely to miss significant signals [6].

Instead of relying on multiple alignment, a more successful approach to phylogenetic footprinting is to use one of the existing motif discovery programs such as `Projection` (Buhler and Tompa [8]), `Consensus` (Hertz and Stormo [30])

or `AlignAce` (Roth et al. [50]). These motif discovery programs take as input multiple sequences along with certain parameters and attempt to find candidate motifs in them. Cliften et al. [13], for instance, reported some successes using `AlignAce` when global multiple alignment tools failed. However, these general motif discovery algorithms do not take into account the phylogenetic relationship of the given sequences since they assume the input sequences to be independent. As explained by Blanchette et al. [6], if the phylogeny underlying the data is ignored in data sets containing a mixture of some closely related species and some distant ones, similar sequences from the set of closely related species will have an unduly high weight in the choice of motifs reported. Even if these methods were modified to weight the input sequences unequally, this would still not capture the information in an arbitrary phylogenetic tree.

Blanchette et al. [6] therefore proposed an algorithmic method designed specifically for phylogenetic footprinting in multiple species which avoids the drawbacks described above of both multiple alignment and general motif discovery algorithms. Given a set of unaligned orthologous sequences, their approach identifies all DNA motifs that appear to have evolved unusually slowly compared with the surrounding sequence (See Figure 1–7 for results from the `FootPrinter` algorithm).

While the above mentioned phylogenetic footprinting methods (including a hybrid approach making use of local multiple sequence alignment blocks when those are available and reliable and also allowing finding motifs in unalignable

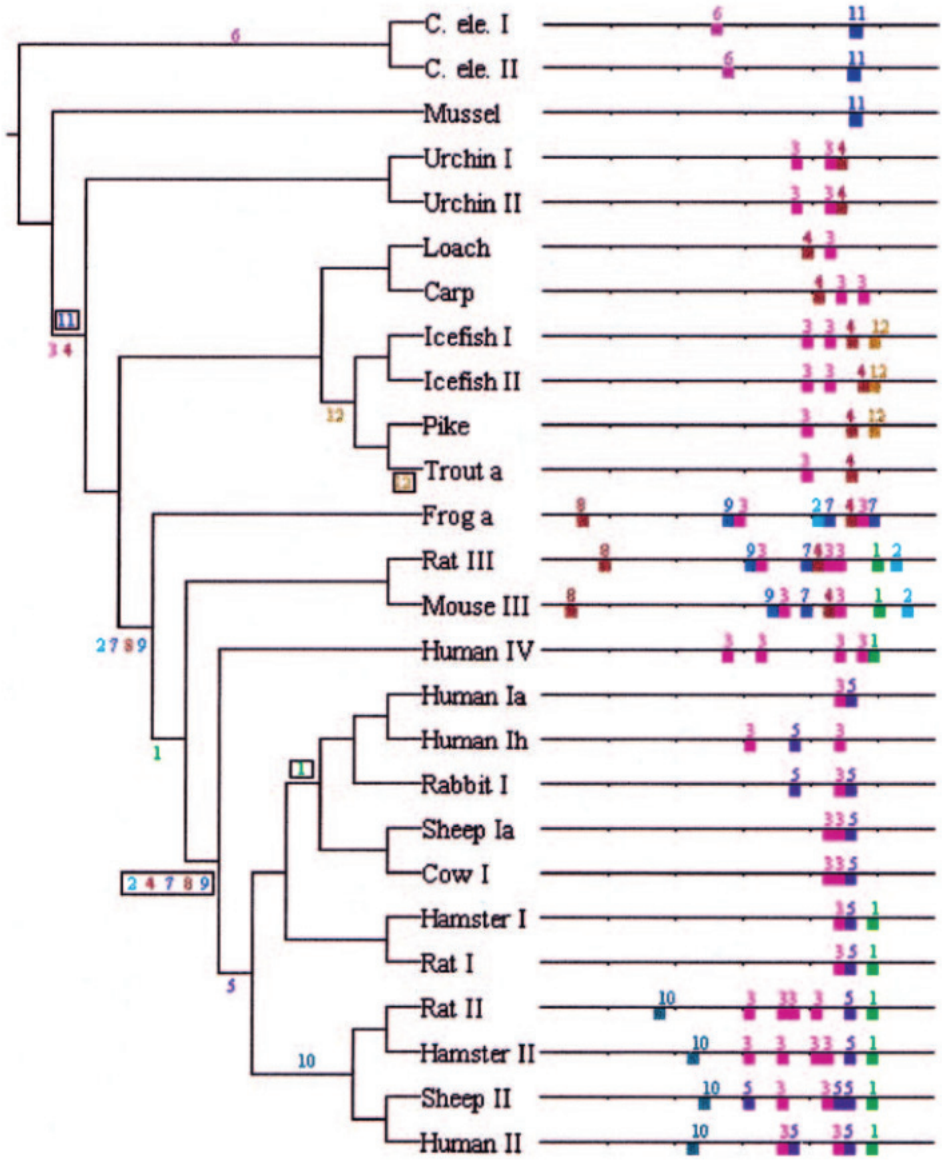

Figure 1–7: Identification of 12 highly conserved motifs using the `FootPrinter` algorithm in the metallothionein gene family. Numbers along branches indicate when each motif was created (unboxed) or lost (boxed), ignoring any less conserved occurrences of the motif not reported by `FootPrinter`, Blanchette et al. [6].

regions [22]), have been successful in the discovery and identification of TFBSs, none of these methods consider binding site turnover.

1.4.4 Binding site turnover

Attempting to identify functional genomic elements via phylogenetic footprinting works well when such elements are relatively long so that they are not drowned out by the non-functional "noise" surrounding them and also when the base composition of these elements is well known. Unfortunately, TFBSs suffer from both problems as they are very short in length and have variable sequences. Due to the combination of these two properties, it is quite likely that TFBSs can be created via random mutations with high frequency [52]. Similarly, since a substitution at a position with high base specificity within a TFBS can abolish its binding activity, mutations that inactivate binding sites should also be frequent. Now, if the inactivation of an existing site follows creation of a new site which is close enough to satisfy any constraints on the position of a bound transcription factor, the new site can take over the function of the previous site, leading to a binding site turnover [48] (Figure 1–8).

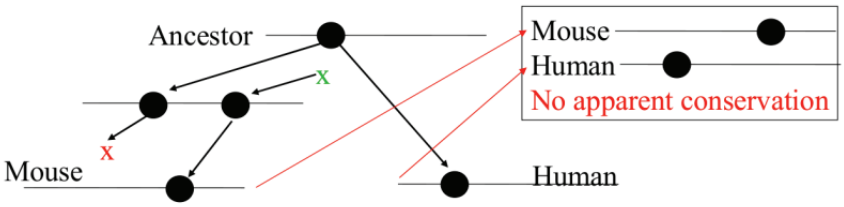

Figure 1–8: Example of binding site turnover: note that a new binding site was created very close to the original in the mouse lineage and took over the function of the original in the mouse genome, whereas the ancestral site was conserved in the human genome.

The most prominent study providing evidence of widespread binding site turnover was done by Dermitzakis et al. [17], who showed that there is extensive divergence within the nucleotide sequence of TFBSs, and by using direct experimental data from functional studies in both human and rodents for 20 of the regulatory regions, they estimated that 32% to 40% of the human functional sites are not functional in rodents. A handful of case studies of binding site turnover show that some binding site gain and loss events are tolerated, or even preferred, by natural selection. Some of these clearly alter regulatory output [25, 37]. For example, the gain of binding sites for the TF engrailed in a preexisting regulatory sequence has led to the emergence of a pigmented spot on the wings of *Drosophila biarmepes* [25], a clear example of binding site gain altering regulatory output [41]. Interestingly, other case studies have found turnover events that do not alter function [47, 35]. The orthologous evenskipped stripe 2 enhancers of *Drosophila* species differ considerably, with many functional sites found in *D. melanogaster* absent in related species. Yet these enhancers function normally in *D. melanogaster* embryos [35, 36, 37], leading to the conclusion by Dermitzakis et al. that "conservation of function can be maintained in the face of fluidity in the exact composition of regulatory regions".

Characterization of binding site turnover on a large scale has only recently been attempted. Moses et al. [41] analyzed the dynamics of sites bound by the transcription factor Zeste using the genome sequences of four species in the melanogaster species group: *D. melanogaster*, *D. simulans*, *D. erecta*, *D. yakuba*, and found that at least 5% of the functional Zeste binding sites in Zeste-bound

regions have been created or lost since the four analyzed species diverged from a common ancestor approximately 10 million years ago. Their approach was to identify functional Zeste binding sites using a genome wide ChIP–chip experiment. This data was then used with the multiple alignment of the 4 melanogaster species to develop an evolutionary model to study binding site turnover. In this thesis, we have approached the problem in reverse, i.e., given a phylogenetic tree with predicted binding sites on its constituent sequences, our aim is to predict which sites are most likely functional. However, there are certain points worth highlighting from the work done by Moses et al:

- Multiple sequence alignment errors did not significantly impact their analysis in closely related species.
- They identified 294 regions of the *D. melanogaster* genome bound *in vivo* by Zeste with orthologous sequences in the other species. These regions were then used to identify 1,406 Zeste binding sites. They note that a comparision to flanking sequences of the bound regions revealed significant number of functional nonconserved binding sites, lending weight to the theory that functional binding sites are not necessarily constrained to promoter regions.
- As in our model, they separate turnover events into binding–site losses and gains and estimate the rate of each process.

Other studies have also involved creating probabilistic models for large scale identification of binding site turnover [16, 40], and while successful at shedding light on the properties of the phenomenon, they also deal with smaller genomes and closely related species. For example, Moses et al. argue in their study that

binding sites for crucial transcription factors do not vary quantitatively between species and go on to present a transcription factor binding site model based on this assumption. Again, in this case their study relies on binding sites obtained from *in vivo* experiments and simulations to make up for lack of sufficient data.

1.5 Thesis outline

This thesis presents a probabilistic framework to predict TFBSs on the human genome using a phylogenetic approach but also taking binding site turnover into account. Chapter 2 details the mathematics behind our probabilistic model: given motifs for known TFs, we treat binding sites as individual entities that can be created, lost or conserved. Our model also incorporates certain constraints that take the biological properties of observed binding site turnover into account.

Our goal is to compare promoter regions of various extant and reconstructed ancestral genomes and predict sites that are most likely to be biologically functional. In chapter 3, we validate our approach via simulations and also present the results of applying our algorithm on ChIP-chip microarray data.

CHAPTER 2
Methodology

In this chapter, we first present the basic notions of our probabilistic model which is used to predict functional TFBSs. Thereafter, we introduce the binding site evolutionary model which forms the basis of the entire promoter evolutionary model. This probability is then used to identify functional sites.

Before presenting our probabilistic model, we need to make an important distinction regarding the functionality and non-functionality of transcription factor binding sites:

- A functional site is one that is bound by a transcription factor and regulates expression.
- A non-functional site is one that may or may not be bound by a given transcription factor. However, a non-functional site does **not** play any role in regulating gene expression or may even play a different role altogether.

Much of the work presented deals with this distinction between functionality (denoted by F) and non-functionality (denoted by N), and our goal is to be able to predict with a certain degree of confidence, the *functional* binding sites in the human genome for a given transcription factor. Also, although a TFBS is essentially a sequence of base-pairs on a given genome, for the purposes of our model, we treat each site as a single unit.

Phylogenetic footprinting, as explained in chapter 1, exploits the relationship between orthologous sequences to identify binding sites that have been well conserved over time. Our algorithm tries to improve upon this approach by labeling sites as functional or non-functional taking turnover into account. Moreover, phylogenetic footprinting originally involved comparison between genomes of related *living* species whereas our model is based on aligned sequence data available from a phylogenetic tree of reconstructed ancestral genomes as well as present genomes. Thus, our algorithm takes as input:

1 A phylogenetic tree T,

2 n aligned input promoter sequences (one for each leaf and internal node of T) containing sites identified for a particular transcription factor. Binding site identification is done via motif scanning using PWMs obtained from a database (Section 3.1).

It labels the sites of each promoter as being non-functional or functional. To describe our probabilistic model, the first question tackled is that of modeling TFBS turnover. Thereafter, the presented model is used in calculating the promoter content transition probability and applied in estimating the labeling over the given phylogenetic tree.

2.1 Binding site evolutionary model

To model binding site evolution, we approached the problem in reverse, that is, given ancestral and descendant sequences annotated with functional and non-functional sites, what would be the simplest way to capture all possible evolutionary events? For example, in Figure 2–1, the following events have taken

place at some point in time separating the two sequences: loss of a site ($F2$), gain of a site ($N2$), conservation of two sites ($F1$, $N1$) and a turnover event ($F3$, $F4$).

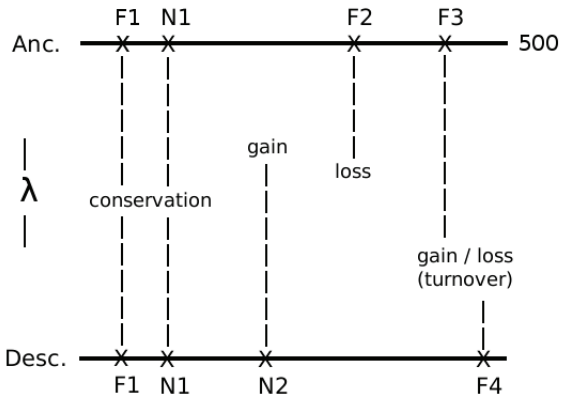

Figure 2–1: An evolutionary scenario between ancestor and descendant promoter regions of 500 bp each, separated by time λ with TFBSs for a given TF marked and labeled as functional (F) or non-functional (N). The first two sites in the ancestor have been conserved in the descendant, a non-functional site has been gained, a functional site has been lost and a functional site ($F3$) has shifted to the right, that is, a turnover has occurred.

We note that by treating the complete set of predicted sites as a single evolving entity, all possible events can be modeled if we only model binding site conservation, gain and loss. A turnover event need not be modeled explicitly as it basically consists of a binding site gain followed by a loss. The evolution of individual binding sites is modeled using a 3-state continuous-time Markov chain, with transition rates as described in Figure 2–2.

Consequently, the probabilities that a site undergoes a particular event over time λ are:

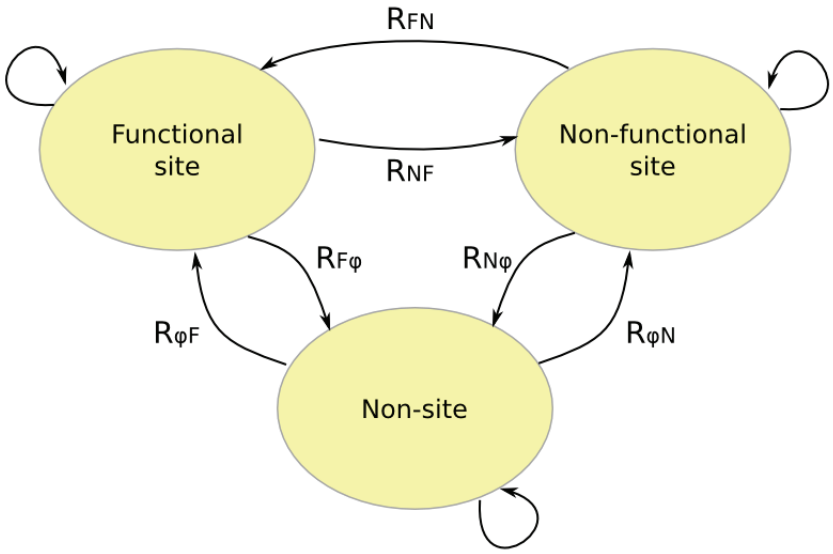

Figure 2–2: Binding site evolutionary model

- The probability of gain from nothing to non-functional site

$$P_{\phi N}(\lambda) = 1 - e^{-\lambda \cdot R_{\phi N}}$$

- The probability of gain from nothing to functional site is mutually exclusive from the gain from nothing to a non-functional site

$$P_{\phi F}(\lambda) = (1 - P_{\phi N}(\lambda)) \cdot (1 - e^{-\lambda \cdot R_{\phi F}})$$

- The probability of loss of a non-functional site to nothing

$$P_{N\phi}(\lambda) = 1 - e^{-\lambda \cdot R_{N\phi}}$$

- The probability of loss of a functional site to nothing

$$P_{F\phi}(\lambda) = 1 - e^{-\lambda \cdot R_{F\phi}}$$

- The probability of gain of a functional site from a non-functional site is mutually exclusive from the loss of a non-functional site to nothing

$$P_{NF}(\lambda) = (1 - P_{N\phi}(\lambda)) \cdot (1 - e^{-\lambda \cdot R_{NF}})$$

- The probability of loss of a functional site to a non-functional site is mutually exclusive from the loss of a functional site to nothing

$$P_{FN}(\lambda) = (1 - P_{F\phi}(\lambda)) \cdot (1 - e^{-\lambda \cdot R_{FN}})$$

- The probability of a non-functional site being conserved is the probability that its functionality did not change nor did it disappear altogether

$$P_{NN}(\lambda) = 1 - P_{NF}(\lambda) - P_{N\phi}(\lambda)$$

- The probability of a functional site being conserved is the probability that it did not lose its functionality nor did it disappear altogether

$$P_{FF}(\lambda) = 1 - P_{FN}(\lambda) - P_{F\phi}(\lambda)$$

- The probability of no event having occurred at a given position is the probability that no new sites were created

$$P_{\phi\phi}(\lambda) = (1 - P_{\phi N}(\lambda)) \cdot (1 - P_{\phi F}(\lambda))$$

At present, we assume that we have accurate values for the rate parameters and continue building the probabilistic framework. Section 2.3.3 explains how these values can be estimated in more detail.

2.1.1 Distance function

An important observation on the characteristics of regulatory sequences is that gene regulation is mediated by cooperative interactions between TFs that bind to clusters of sites within *cis* -regulatory modules (CRMs) [60, 31]. While it is clear that functional binding sites are under selective pressure, this observation leads us to believe in the redundancy of such functional sites. Also, turnover is more likely when there are multiple sites for a single TF, since the selective pressure for all of them to be conserved is low. Hence, we introduce a distance function that increases the variability in the rate of loss of a *functional* site and the rates of functionality changes of a site if there are other *functional* sites close to it. In our model, we hope to capture the phenomenon of binding site turnover by allowing sites close to each other to have a higher variability than those further apart. Consider the following function,

$$f(dist) = min\left(0.9\ ,\ \frac{1}{1+\frac{dist}{c}} + d\right) \tag{2.1}$$

where $dist$ is the number of bp between a site and the closest functional site on the sequence under consideration. If there is no functional site present, $dist = \infty$.

In practice the distance function will be used to modulate the rates of loss and functionality changes of functional sites. The idea behind the distance function is that if there are two functional sites close to each other, then the rates are

decreased by a minimal amount, however for two functional sites far apart the decrease will be substantial resulting in a lower probability (Figure 2–3). This means that functional sites closer together are subject to more frequent changes than those further apart.

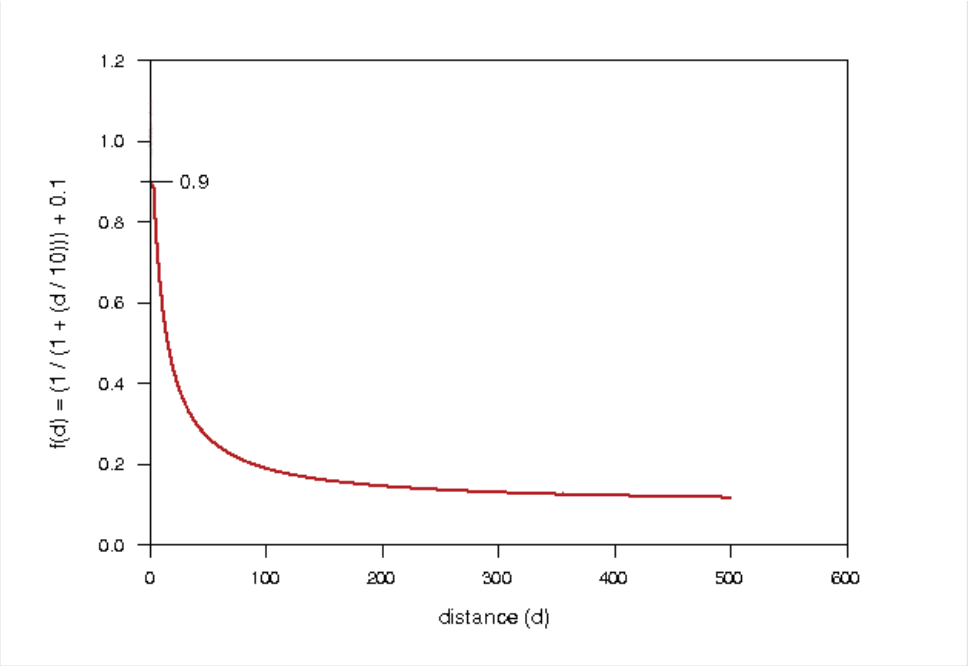

Figure 2–3: Distance function plot with example values for c and d as 10 and 0.1 respectively. Note that the value of this function is capped at 0.9.

Again, for the moment we leave c and d as unknown parameters which will be dealt with in section 2.3.3.

2.1.2 Probabilities revisited

The primary idea of the distance function is to increase the variability of a functional site if there are other functional sites nearby. The $f(dist)$ is thus multiplied to all the rates involving functional sites. Revising the probabilities in

our binding site evolutionary model to incorporate the distance function, we get:

$$
\begin{aligned}
P_{\phi N}(\lambda, dist) &= 1 - e^{-\lambda \cdot R_{\phi N} \cdot (1 - f(dist))} \\
P_{\phi F}(\lambda, dist) &= (1 - P_{\phi N}(\lambda, dist)) \cdot (1 - e^{-\lambda \cdot R_{\phi F} \cdot f(dist)}) \\
P_{N\phi}(\lambda) &= 1 - e^{-\lambda \cdot R_{N\phi}} \\
P_{F\phi}(\lambda, dist) &= 1 - e^{-\lambda \cdot R_{F\phi} \cdot f(dist)} \\
P_{NF}(\lambda, dist) &= (1 - P_{N\phi}(\lambda, dist)) \cdot (1 - e^{-\lambda \cdot R_{NF} \cdot f(dist)}) \\
P_{FN}(\lambda, dist) &= (1 - P_{F\phi}(\lambda, dist)) \cdot (1 - e^{-\lambda \cdot R_{FN} \cdot f(dist)}) \\
P_{NN}(\lambda, dist) &= 1 - P_{NF}(\lambda, dist) - P_{N\phi}(\lambda) \\
P_{FF}(\lambda, dist) &= 1 - P_{FN}(\lambda, dist) - P_{F\phi}(\lambda, dist) \\
P_{\phi\phi}(\lambda, dist) &= (1 - P_{\phi N}(\lambda, dist)) \cdot (1 - P_{\phi F}(\lambda, dist))
\end{aligned}
$$

where λ is time elapsed.

In our model, while the existence of functional sites around a given site affects its probability of loss and change of functionality, the creation of any type of a site is independent of the presence of other sites around. Since the concept of site functionality does not exist biologically, this keeps our model close to biological processes. Thus, to compensate for the effect of the distance function on $R_{\phi F}$, we negate its effect on $R_{\phi N}$.

2.2 Promoter content transition probability

Coming back to the example presented in section 2.1 (Figure 2–1), given the labeling of a promoter sequence under consideration, we would like to calculate the probability of all TFBS gains, losses and conservations that occurred over time λ

between a given ancestral site configuration and a given descendant site configuration. In Figure 2–1 there are 4 events (2 gains, 2 losses) and 2 conservations. We note that there are actually infinitely many possible explanations for the observed pair of configurations. For example, if a site at position x in the ancestral sequence is found at position $x + 10$ in the descendant sequence, then the site at x could have been conserved until the time $\lambda/8$, another site could have been gained and then lost from the time $\lambda/8$ to $\lambda/5$ and finally the observed site at position $x + 10$ in the descendant could have been gained at time $\lambda/5$.

Enumerating all the possible scenarios is thus impossible and whereas the probability of a lot of these scenarios occurring is very small, in principle they do need to be considered. We therefore present a compromise by enumerating all the possible permutations of a given set of events for two sequences and sum the probabilities of each scenario to approximate the promoter content transition probability. Thus, revisiting the example in section 2.1, we slice the time λ elapsed between the two sequences depending on the number of events occurred[1] and enumerate all the possible orderings of events. Calculating the promoter content transition probabilities for each of the time slices, we get the probability of a single scenario. Considering a scenario with k events, occurring at times $\frac{\lambda}{k}, \frac{2\lambda}{k}, \frac{3\lambda}{k}, \ldots, \lambda$, we get figure 2–4 which illustrates one of the permutations: loss of site N2, gain of site F2, gain of site F3, gain of site F4.

[1] Note that an event is when a site is either lost, gained, or changes functionality.

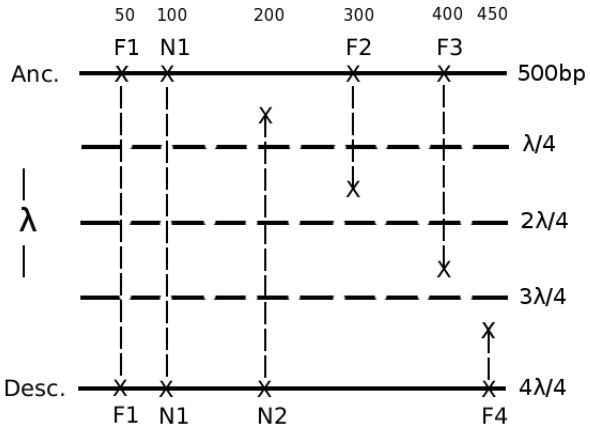

Figure 2–4: Given the data in Fig. 2–1, we split the time λ by the number of events and enumerate all the possible scenarios for the 4 events. As a result, we can generate 4! scenarios and the above figure shows $N2 \rightarrow F2 \rightarrow F3 \rightarrow F4$, i.e. the first two sites were conserved for λ, the third site ($N2$) was gained between time 0 and $\lambda/4$, the fourth site ($F2$) was conserved until $2\lambda/4$, the fifth site ($F3$) was conserved until $3\lambda/4$ and the sixth site ($F4$) was gained and conserved from $3\lambda/4$ to λ. Equation 2.2 formalizes this probability.

The probability of the scenario in Figure 2–4 can be calculated by multiplying the probability of the events having occurred during each time slice with the probability that no other sites were gained or lost in that time period. For example:

$$P(N2, F2, F3, F4) =$$

$$\{P_{FF}(\tfrac{\lambda}{4}, 250) \cdot P_{NN}(\tfrac{\lambda}{4}, 50) \cdot P_{\phi N}(\tfrac{\lambda}{4}, 100) \cdot P_{FF}(\tfrac{\lambda}{4}, 100) \cdot P_{FF}(\tfrac{\lambda}{4}, 100) \cdot$$

$$(P_{\phi\phi}(\tfrac{\lambda}{4}))^{500-5}\} \times$$

$$\{P_{FF}(\tfrac{\lambda}{4}, 250) \cdot P_{NN}(\tfrac{\lambda}{4}, 50) \cdot P_{NN}(\tfrac{\lambda}{4}, 100) \cdot P_{F\phi}(\tfrac{\lambda}{4}, 100) \cdot P_{FF}(\tfrac{\lambda}{4}, 100) \cdot$$

$$(P_{\phi\phi}(\tfrac{\lambda}{4}))^{500-5}\} \times$$

$$\{P_{FF}(\tfrac{\lambda}{4}, 350) \cdot P_{NN}(\tfrac{\lambda}{4}, 50) \cdot P_{NN}(\tfrac{\lambda}{4}, 150) \cdot P_{F\phi}(\tfrac{\lambda}{4}), 350) \cdot$$
$$(P_{\phi\phi}(\tfrac{\lambda}{4}))^{500-4}\} \times$$
$$\{P_{FF}(\tfrac{\lambda}{4}, 500) \cdot P_{NN}(\tfrac{\lambda}{4}, 50) \cdot P_{NN}(\tfrac{\lambda}{4}, 150) \cdot P_{\phi F}(\tfrac{\lambda}{4}, 400) \cdot$$
$$(P_{\phi\phi}(\tfrac{\lambda}{4}))^{500-4}\} \tag{2.2}$$

The total promoter content transition probability is the sum of the probabilities of all the event permutations. Again in our example this is:

$$P(descendant|ancestor) = P(N2, F2, F3, F4) + \ldots + P(F4, F3, F2, N2)$$

Before we continue our discussion, we note that there are certain tools from Markov theory that could be applied to estimating (or bounding) the probabilities of these events. A continuous time Markov process can be used to simulate a dynamical system. These processes work well when the number of variables is small but increase exponentially in complexity when we consider a system with multiple components.

In their conference paper, El-Hay et al. [21] present continuous time Markov networks which are specifically designed for modeling sequence evolution. In their implementation individual mutations specified by different rates are modeled by a continuous-time proposal process. A global fitness or energy function of the entire system determines the probability of a proposed change being accepted, which is captured by a Markov network that encodes the fitness of different states.

While their implementation could be used a complete solution to TFBS evolution (since they also go on to describing a maximum likelihood function and a learning strategy for parameters as we do), it is not focused on incorporating binding site turnover into their predictions. Moreover, a major hurdle lies in proposing a global fitness function for different sequences with TFBS predictions.

Thus, to reduce the complexity of our algorithm, we enumerate scenarios with the same number of events as the observed data and calculate the promoter content transition probability between each.

2.3 Prediction

Given data consisting of aligned mammalian sequences and their ancestral reconstructions, we annotate the promoter regions with TFBS predictions and would like to be able to predict the functionality of these predicted TFBSs. Thus, given a TF and a *binary* phylogenetic tree T of promoter regions with binding site predictions for that TF, we want to find the functional/non-functional labeling L whose probability given the observed sites is maximized. In other words, we would like to find functional/non-functional labeling for the sites on the given phylogenetic tree which maximizes the promoter content transition probabilities over all the branches. More formally we want to maximize the probability:

$$P(Obs, L) = \left(\prod_{(a,b) \in edges(T)} P(Obs(b), L(b) \mid Obs(a), L(a)) \right) \times P(Obs(root) \cdot L(root)) \tag{2.3}$$

where $L(x)$ is the vector of labels of all the sites on node x and $Obs(x)$ is the vector of positions of sites on node x.

2.3.1 Computing the total likelihood of all possible labelings

Now that we have an evolutionary model that allows us to compute probabilities of change of states along the given phylogenetic tree, we present an algorithm very much based on the maximum likelihood algorithm as proposed by Felsenstein [23]. Felsenstein proposed a maximum likelihood technique to estimate evolutionary trees from nucleic acid sequence data. In our case, we would like to calculate the sum of the probability of all possible labelings over the given phylogenetic tree. This value will help us in estimating the values for the unknown parameters as we can use it to find the combination of parameters which maximize the likelihood of observed data (Section 2.3.3).

In order to apply Felsenstein's algorithm, we assume that evolution over a single branch is independent from other branches. To calculate the sum of all possible labelings on a phylogenetic tree, we start from the leaves onwards and cycle through all the possible labelings between the children and their parent sequences, summing them up all the way to the root. To facilitate our algorithm, we introduce a quantity which may be called the conditional likelihood of a subtree, $X_u[L]$. This value is the likelihood of all the labelings for a subtree rooted at node u in which u is labeled with L (note that if u has k sites, there are 2^k possible labelings). This conditional likelihood is first calculated for the leaves for all possible promoter labeling combinations and stored in a data structure. These values are then used to calculate the total likelihoods upwards to the root. More formally, the general formula for calculating the conditional probability at node u

for a particular labeling L is:

$$
\begin{aligned}
X_u[L] = & \\
& (\sum_{L' \in v} (P(L' \mid Obs(u), Obs(v), L) \ \cdot \ X_v[L'])) \times \\
& (\sum_{L'' \in w} (P(L'' \mid Obs(u), Obs(v), L) \ \cdot \ X_w[L'])) \qquad (2.4)
\end{aligned}
$$

where v and w are the children of node u, L' is a labeling of v, L'' is a labeling of w and $P(B \mid Obs(a), Obs(b), A)$ is the promoter content transition probability of going from a parent sequence with labeling A to a child sequence with labeling B, given by equation 2.3. When u is a leaf with known labeling Z, that is, when the labeling of a particular leaf sequence is known, any labeling configuration that does not comprise of the known labeling for that leaf sequence will return a probability of 0:

$$X_u[L] = 1, \text{ if } L = Z$$

$$X_u[L] = 0, \text{ otherwise}$$

and when u is a leaf with unknown labeling, that is, when a leaf sequence has no known labeling, all labelings are equally possible in that case:

$$X_u[L] = 1, \text{ for all } L$$

The calculation of the total conditional likelihood probability at each node is explained via Figure 2–5. In Figure 2–5, the conditional likelihood for each of the

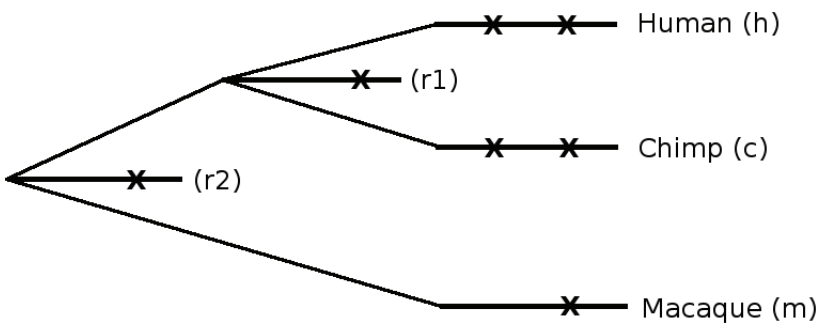

Figure 2–5: The above is a small phylogenetic tree example with 5 promoter sequences along with TFBS predictions for a particular TF. Here, an extra site has cropped up in the human and chimp sequences.

nodes is calculated as follows:

$$X_h[FF] = X_h[FN] = X_h[NF] = X_h[NN] = 1$$

$$X_c[FF] = X_c[FN] = X_c[NF] = X_c[NN] = 1$$

$$X_m[F] = X_m[N] = 1$$

$$\begin{aligned} X_{r1}[F] = \{&P(FF \mid F) \cdot X_h[FF] + P(FN \mid F) \cdot X_h[FN] + \\ &P(NF \mid F) \cdot X_h[NF] + P(NN \mid F) \cdot X_h[NN]\} \times \\ \{&P(FF \mid F) \cdot X_c[FF] + P(FN \mid F) \cdot X_c[FN] + \\ &P(NF \mid F) \cdot X_c[NF] + P(NN \mid F) \cdot X_c[NN]\} \end{aligned}$$

$$\begin{aligned} X_{r1}[N] = \{&P(FF \mid N) \cdot X_h[FF] + P(FN \mid N) \cdot X_h[FN] + \\ &P(NF \mid N) \cdot X_h[NF] + P(NN \mid N) \cdot X_h[NN]\} \times \\ \{&P(FF \mid N) \cdot X_c[FF] + P(FN \mid N) \cdot X_c[FN] + \end{aligned}$$

$$P(NF \mid N) \cdot X_c[NF] + P(NN \mid N) \cdot X_c[NN]\}$$

$$X_{r2}[F] = \{P(F \mid F) \cdot X_{r1}[F] + P(N \mid F) \cdot X_{r1}[N]\} \times$$
$$\{P(F \mid F) \cdot X_m[F] + P(N \mid F) \cdot X_m[N]\}$$

$$X_{r2}[N] = \{P(F \mid N) \cdot X_{r1}[F] + P(N \mid N) \cdot X_{r1}[N]\} \times$$
$$\{P(F \mid N) \cdot X_m[F] + P(N \mid N) \cdot X_m[N]\}$$

The total likelihood for all possible labelings of the tree is $\sum_L X_{root}[L]$. In Figure 2–5, this is $X_{r2}[F] + X_{r2}[N]$.

2.3.2 Labeling the tree

To label the tree with the best likelihood labeling, we again use a modified version of the Felsenstein algorithm [23] as used for computing the total likelihood for all possible labelings over the phylogenetic tree. This algorithm is similar and works in tandem with the conditional likelihood calculation. As $X_u[L]$ for a node at root u and labeling L is calculated, we associate the *best* labeling found when going from node u to each of its children. In this manner, as information flows up the tree, the conditional likelihood at each node (X_u), is associated with the best child labelings, for each of its labelings (L).

The best labeling here is the labeling that results in the maximum likelihood of going from ancestor to descendant, taking the subtree of the descendant into account. More formally, for any $X_u[L]$, the best labelings for its children v and w

are:

$$L' = argmax_{L' \in v}(P(L' \mid Obs(u), Obs(v), L) \cdot X_v[L'])$$
$$L'' = argmax_{L'' \in w}(P(L'' \mid Obs(u), Obs(v), L) \cdot X_w[L'']) \quad (2.5)$$

where v and w are the children of node u labeled with L, and $P(B \mid Obs(a), Obs(b), A)$ is the promoter content transition probability of going from a parent sequence with labeling A to a child sequence with labeling B. Note that if u is a leaf, then L' and L'' do not exist.

Finally, once all the labelings have been generated for each of the conditional likelihood values, we label the root of the entire tree with the labeling of the maximum X_u value and label its children with the associated labelings. Thereafter, each node having been labeled by its parent, propagates the labeling down to its children via its X_u value. In this way, all the information gathered while going up the tree is propagated back down to the leaves. Revisiting the example in Figure 2–5, the maximum likelihood labeling algorithm is explained in Figure 2–6.

2.3.3 Estimating rates and other parameters

Having presented the probabilistic framework, we describe how to obtain the values for the parameters in our binding site evolutionary model from section 2.1.

Estimating $R_{\phi N}$ and $R_{N\phi}$

Given data in the form of regions of a multiple sequence alignment whose constituent species come from a phylogenetic tree, we used the TRANSFAC database to make TFBS predictions for all of its matrices. The regions chosen were

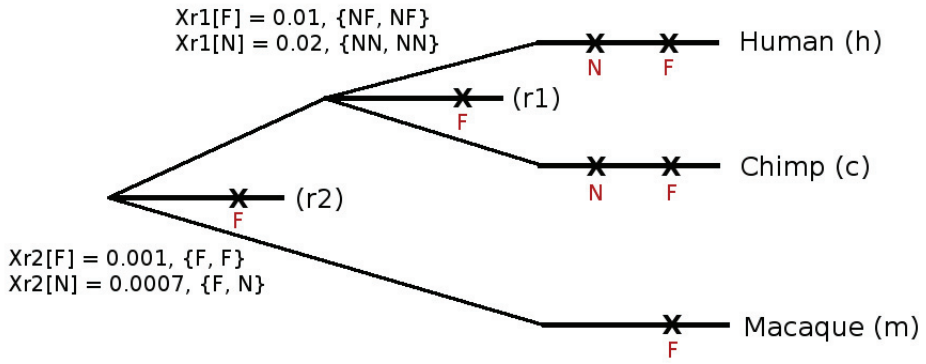

Figure 2–6: Revisiting the example in Figure 2–5, the maximum likelihood labeling algorithm works as follows: The values of X_u are calculated for each of the nodes, as given in equation 2.4 along with the labeling that results in the best probability of going from u to each of its children, as given in equation 2.5. The calculated values for X_{r1} and X_{r2} are shown in the figure. Thereafter, the maximum probability at the root of the entire tree determines the labeling for the root sequence, which in this example is 0.001, and its children, the $r1$ and m sequences are labeled with F and F respectively. Given than $r1$ is labeled as F, we label its children, the h and m sequences as NF, NF respectively.

1000bp upstream from transcription start sites on the human genome and in total we generated 57332 sets of orthologous promoters per matrix.

Since we did not know anything about the functionality of these predicted sites, we made the conservative assumption that they were all non-functional. We then estimated the probabilities of gain and loss of a non-functional site by counting the number of site changes per branch of the phylogenetic tree from the given data. More formally, for a given branch, we get:

$$P_{\phi N} = \frac{num_{\phi N}}{length(l)}$$
$$P_{N\phi} = \frac{num_{N\phi}}{num_{N\phi} + num_{NN}} \tag{2.6}$$

where $length(l)$ is the number of base-pairs on branch l, $num_{\phi N}$, $num_{N\phi}$ and num_{NN} are the number of site gains, losses and conservations respectively observed along branch l.

Given formulae for $P_{\phi N}(\lambda)$ and $P_{N\phi}(\lambda)$ from section 2.1.2, we calculate the average rates of gain and loss of a non-functional site for a particular branch,

$$R_{\phi N} = Avg_{l \in T} \left\{ \frac{-ln(1 - P_{\phi N}(\lambda))}{\lambda(l)} \right\}$$
$$R_{N\phi} = Avg_{l \in T} \left\{ \frac{-ln(1 - P_{N\phi}(\lambda))}{\lambda(l)} \right\} \tag{2.7}$$

where l is a branch in tree T and $\lambda(l)$ is the time elapsed over the branch l.

Unknown rates and parameters

Having estimated values for $R_{\phi N}$ and $R_{N\phi}$ for a given TF, we look to estimate the other rates and the parameters for the distance function in section 2.1.1. Since biologically all sites are supposed to be equal, in our model we can be conservative and assume that all sites are non-functional. Thus, we can assume that the rates of loss for both functional and non-functional sites are the same, that is, $R_{F\phi} = R_{N\phi}$. Similarly, we set the rates of gain of functional and non-functional as same, that is, $R_{\phi F} = R_{\phi N}$. Note that $R_{F\phi}$ exists as a separate entity as it will be modified by the distance function in a different manner for functional and non-functional sites.

To resolve the remaining parameters, which are R_{FN}, R_{NF}, c and d (from the distance function described in section 2.1.1), we try to estimate them as best as possible via our proposed algorithm. Given sets of orthologous promoters with sites predicted for a particular TF, we would like to find the parameters that result

in the best total likelihood. Although before we assumed that all binding sites are non-functional, the reality is that a certain fraction of is functional. We thus seek the combination of parameter values that maximizes the likelihood of the observed data.

Therefore, given a set of values for the parameters, we calculate the total likelihood value over a tree of promoter sequences. We then sum the $X_u[L]$ value at the root for 1000 sets of orthologous promoters (where L is the best labeling as determined by our algorithm) and find parameters that will result in the highest possible sum. The sum M is defined as:

$$M = \sum_{t \in T} (X_{root(t)}[L]) \tag{2.8}$$

where t is a tree in the dataset T and L is the best possible labeling at the root of t.

To find the best parameters we pick a range of values for each of the unknowns (See Table 2–1 for an example starting range), cycle through all the combinations and sort them in descending order of their respective M scores. The combination that achieves the best score is then used as a base for formulating another range by increasing the resolution around it (a concrete example presented in the following chapter). We then conduct another search over the second range to find the set of parameters that yields a better M score than the previous iteration. The cycle of iterations stop when there is no significant improvement or M or a certain number of iterations has been reached. Thus, for a given dataset we are able to identify parameters specific to a particular TF.

Table 2–1: First iteration of parameter estimation

Parameter	**Range**
c	0.1, 1, 10, 100, 1000
d	0.001, 0.01, 0.1, 1, 10, 100
R_{FN}	0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10
R_{NF}	0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10

Note that this method of parameter estimation can be thought of as "drilling down" to obtain the parameters which will yield the best total likelihood over all the labelings, over all the provided data. Unfortunately, this method does not guarantee a global maximum, that is, the best parameter combination that we find may not necessarily be the actual best. However, given the computational complexity involved in the calculation of the total likelihood value of one tree, it is the most feasible option.

2.4 Summary

This chapter presents a probabilistic framework for predicting TFBSs using a phylogenetic approach. We start by presenting our binding site evolutionary model, wherein we treat TFBSs as individual entities which evolve much like individual nucleotides on the genome. A TFBS can thus be gained, lost or conserved and in this manner, we do not explicitly model turnover but instead introduce the concept of site functionality via labels. We then determine the parameters of site functionality and how they differentiate TFBS evolution.

Once the framework for a single site and is laid out, we present the promoter content transition probability, which takes a parent sequence and a child sequence,

both with known TFBS configurations and labelings, and calculates the probability of the parent configuration having evolved to the child configuration. This probability is then used in our prediction algorithm, which takes as input a set of orthologous sequences, each sequence representing a branch of a phylogenetic tree and each sequence composed of sites marking TFBS predictions with unknown functionalities. Given this input, our prediction algorithm determines the correct functionality labeling for the entire tree by using a maximum likelihood approach. Finally, we discuss the estimation of the various parameters introduced in our binding site evolutionary model.

In the following chapter, we validate our methodology and present the results of applying our algorithm on ChIP-chip data.

CHAPTER 3
Results

This chapter presents an overview of the results obtained by our TFBS prediction algorithm. We first describe in detail the generation and sources of data used in evaluating our algorithm. We then provide examples of the parameter estimation for $R_{\phi N}$, $R_{N\phi}$, $R_{\phi F}$, $R_{F\phi}$, R_{FN}, R_{NF}, c and d as presented in chapter 2, for two TF matrices on data generated from multiple sequence alignment (Section 3.1). Thereafter, we evaluate our algorithm on data generated *in silico* and finally, the results of our algorithm on experimental data from published experiments are presented.

3.1 Experimental Data

Our algorithm takes as input a region of a multiple sequence alignment whose constituent species come from the phylogenetic tree as depicted in Figure 3–1. Starting with a set of 57332 human promoter regions (some of which were defined as 1000bp upstream of transcription start sites on the human genome), we used the TBA program (Blanchette et al. [4]) to obtain the corresponding alignments with other extant species. The Threaded Blockset Aligner (TBA) itself is a novel multiple alignment technique which instead of using a reference sequence, produces "blocks" of alignments. As explained in Blanchette et al. [4], these blocks are basically a local alignment of some or all of the given sequences, in which each position in the given sequences appears precisely once. Any detected match

among some or all of the sequences is represented among the blocks, and mutually consistent reference-sequence alignments can be extracted at will. This property is exploited to produce a set of blocks under the assumption that the matching regions occur in the same order and orientation in all species.

This alignment was further used to reconstruct the ancestral sequences as described in Blanchette et. al [3]. Specifically, given the phylogenetic tree and assuming that its topology is correct, the branch lengths were inferred using the HKY model [29] and the PHYML program [26]. Thereafter, the TBA tool [4] was used to obtain a multiple sequence alignment of the following mammalian species: human, chimp, macaque, rat, mouse, rabbit, dog, cow, armadillo and elephant. The algorithm of Blanchette et al. [3] was then used to make predictions regarding which positions of the alignment correspond to ancestral bases, and which correspond to nucleotides inserted after the ancestor. These position specific predictions were made using a greedy algorithm that seeked to explain the observed alignment using a set of insertions and deletions of maximum likelihood. Finally, the identity of the nucleotide at each ancestral position was then predicted using a context-dependent maximum-likelihood estimation.

An actual data file thus consists of a region of the multiple alignment of the extant and the reconstructed ancestral sequences with binding site predictions on all sequences. To generate the dataset of predicted TFBSs, we predicted binding site matches for all 481 TRANSFAC [62] position weight matrices, where each match was scored using a log likelihood ratio score. We only report those matches such that the probability that a random sequence would have a log likelihood

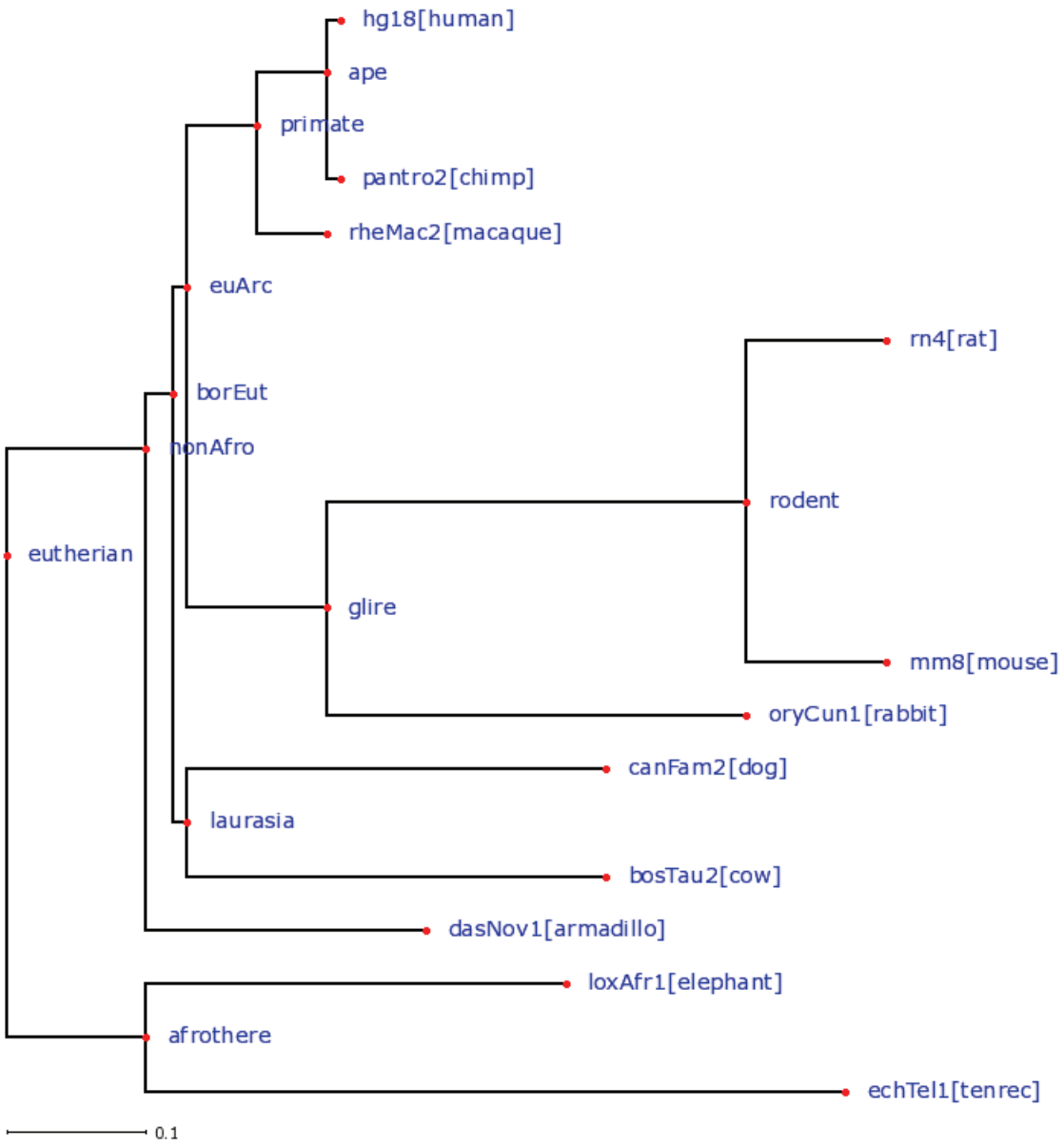

Figure 3–1: The mammalian phylogenetic tree used to generate experimental data and make predictions. Image courtesy Freslund et al. [24].

ratio score that is as good or better than the one observed is very small, i.e. $p - value < 2^{11.1} = 0.0005$. Figure 3–2 illustrates the data generation process.

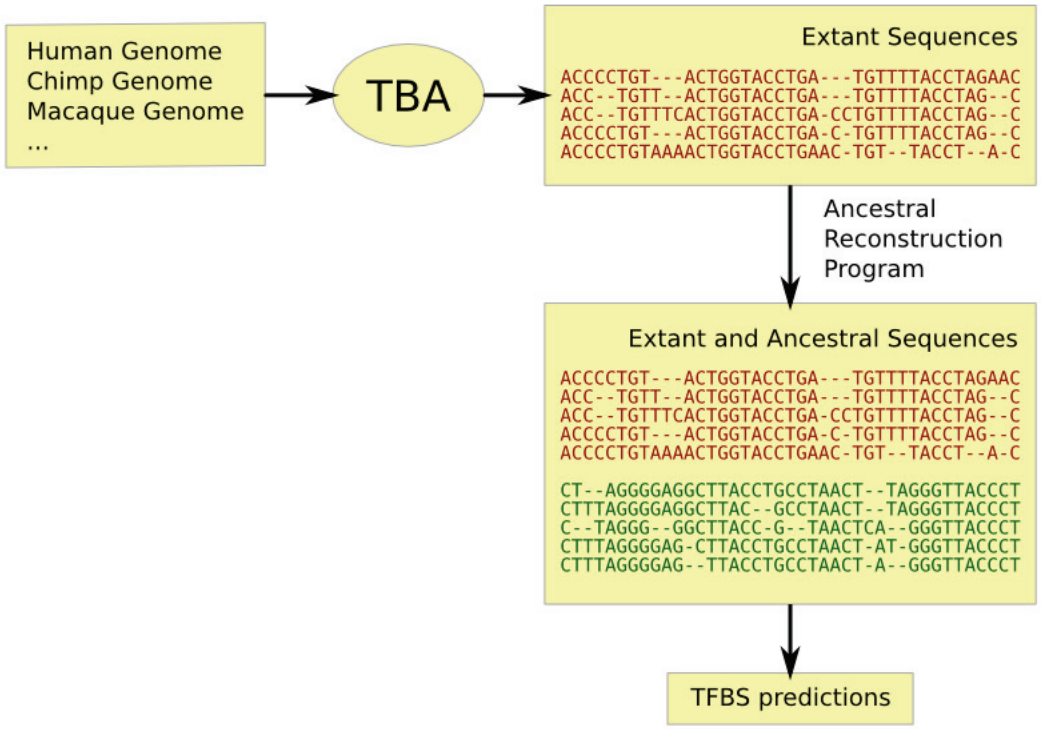

Figure 3–2: Using the TBA program, mammalian genomes from the tree in figure 3–1 are aligned and this alignment is then used to reconstruct the ancestral sequences. The extant and ancestral sequences are re-aligned using the TBA program again, and it on this alignment that TFBS predictions for all TRANSFAC matrices are made.

The generated dataset of predicted TFBSs for a single matrix consists of 57332 sets of orthologous promoters in total over the entire genome. Each set itself contains an average of 10 predicted sites giving us over 500,000 sites per matrix over which to base our calculations on.

3.2 Estimation of rates and parameters

An important aspect of our binding site evolutionary model are the rates and parameters chosen. Using the theory presented in section 2.3.3 and data from section 3.1, we present the results of the rate calculation and parameter estimation for two transcription factors: Estrogen Receptor (TRANSFAC Matrix M00191) and NF-κB (TRANSFAC Matrix M00054).

3.2.1 $R_{\phi N}$ and $R_{N\phi}$ calculation

Using the data from section 3.1, we calculated the number of non-functional site gains, losses and conservations for each branch of the phylogenetic tree, over all the promoters. We used these counts to calculate the probability of the gain and loss of a non-functional site (Equation 2.6). These probabilities were then used to calculate $R_{\phi N}$ and $R_{N\phi}$ (Equation 2.7).

Table 3–1 provides the non-functional site counts per branch for the Estrogen Receptor (ER) transcription factor. We observe that the calculated probabilities correspond to the phylogenetic tree used for their calculation. For example, the highest probability of loss of a site is to be found at the longest branch (between afrothere and echTel1). Also, the branch between glire and rodent presents relatively high probabilities of loss and gain of sites, which is again consistent with the observation that the mouse lineage underwent a higher rate of deletion [61]. Moreover, these probabilities are also consistent with the results of studies on the divergence of transcriptional regulation between the two genomes [43]. Table 3–3 provides the final calculated rates for both the ER and NF-κB transcription factors with the rates for NF-κB having been calculated in a similar manner.

Table 3–1: Per branch rate calculations for ER transcription factor

Ancestor	**Descendant**	$P_{\phi N}$	$P_{N\phi}$	$R_{\phi N}$	$R_{N\phi}$
eutherian	afrothere	0.000180142	0.412879	0.00180158	5.32524
eutherian	nonAfro	0.000129851	0.136802	0.0012986	1.47112
nonAfro	borEut	0.0001362	0.119834	0.00681049	6.38225
nonAfro	dasNov1	0.000318872	0.784657	0.00159461	7.6776
borEut	euArc	0.000101016	0.17138	0.0101021	18.7994
borEut	laurasia	0.000181322	0.305224	0.0181339	36.4166
euArc	glire	0.000161872	0.386653	0.00161885	4.88825
euArc	primate	0.00034406	0.466486	0.00688239	12.5654
primate	ape	9.12852e-05	0.146427	0.00182579	3.16649
primate	rheMac2	0.000160595	0.265394	0.00321215	6.16842
ape	hg18	2.98571e-05	0.0481245	0.00298575	4.9321
ape	panTro2	3.26969e-05	0.0908542	0.00326975	9.52498
glire	oryCun1	0.000268342	0.779211	0.000894593	5.03516
glire	rodent	0.000478442	0.723211	0.00159519	4.28166
rodent	rn4	0.000233828	0.405187	0.00233855	5.19507
rodent	mm8	0.000184511	0.329764	0.00184528	4.00125
laurasia	bosTau2	0.000401474	0.674886	0.00133851	3.74527
laurasia	canFam2	0.000392757	0.601698	0.00130945	3.06848
afrothere	echTel1	0.000319747	0.78587	0.000639597	3.08234
afrothere	loxAfr1	0.000255089	0.550095	0.000850406	2.66239

3.2.2 Estimation of R_{FN}, R_{NF}, c and d

To estimate the unknown parameters, we seeked the values that maximize the likelihood of the given data. As explained in section 2.3.3, these parameters were estimated by a sieving technique. This involved creating initial ranges for each of the 4 parameters and cycling through all the possible combinations. Calculation of $X_u[L]$ for a single set of orthologous promoters takes about 30 seconds and since we had to cycle over 1000 combinations for a single range, we limited the dataset to 1000 sets of orthologous promoters.

After having cycled through all the parameter combinations we got a list of all the combinations and their M-scores (Equation 2.8). From this list we picked the combination that resulted in the highest M-score and created another range around each of the values in the combination. We repeated the calculation of a list of M-scores for this range, creating another range and continued the iterations until there was no significant change in parameter values or an iteration limit of 6 was reached. Table 3–2 presents the unknown rate and parameter estimation ranges for ER transcription factor whilst all the calculated and estimated parameters for ER transcription factor are listed in Table 3–3. Similarly, we estimated the unknown rates and parameters for NF-κB transcription factor (Table 3–3).

Table 3–2: Unknown rate and parameter estimation ranges for ER transcription factor

Iteration	**Parameter**	**Values tested**
1	c	0.1, 1, 10, 100, 1000, 10000
	d	0.001, 0.01, 0.1, 1, 10, 100
	R_{FN}	0.0001, 0.001, 0.01, 0.1, 1, 10
	R_{NF}	0.0001, 0.001, 0.01, 0.1, 1, 10
Best: $c = 10$, $d = 0.1$, $R_{FN} = 0.0001$, $R_{NF} = 10$		
2	c	0.1, 1, 5, 10, 20, 100
	d	0.01, 0.05, 0.1, 0.5, 1
	R_{FN}	0.0001, 0.005, 0.001, 0.005, 0.01
	R_{NF}	0.1, 1, 5, 10, 20
Best: $c = 20$, $d = 0.1$, $R_{FN} = 0.0001$, $R_{NF} = 5$		
3	c	10, 15, 20, 25, 30
	d	0.05, 0.1, 0.2, 0.25
	R_{FN}	0.0001, 0.005, 0.001, 0.05
	R_{NF}	1, 2.5, 5, 7.5, 10
Best: $c = 10$, $d = 0.2$, $R_{FN} = 0.0001$, $R_{NF} = 5$		
4	c	7.5, 10, 12.5
	d	0.15, 0.2, 0.25
	R_{FN}	0.00001, 0.0001, 0.001
	R_{NF}	4, 5, 6
Best: $c = 7.5$, $d = 0.25$, $R_{FN} = 0.0001$, $R_{NF} = 4$		

Table 3–3: All rates and parameters for ER and NF-κB transcription factors

Rate/Parameter	$R_{\phi N} = R_{\phi F}$	$R_{N\phi} = R_{F\phi}$	R_{FN}	R_{NF}	c	d
ER Values	0.00351738	7.41948	0.0001	4	7.5	0.25
NF-κB Values	0.00398662	6.54232	0.0001	0.5	0.0001	0.23

3.3 Algorithm performance on simulated data

To evaluate the efficacy of our prediction algorithm, we needed a well-studied and biologically relevant dataset against which we could compare results. Due to the lack of such a standard, we decided to generate data *in silico* and then by assuming that our proposed model is correct, we could evaluate the accuracy of the predictions over this "gold standard".

To generate data *in silico*, we used the mammalian phylogenetic tree from section 3.1. We set all the sequences to be 1000bp in length and inserted sites at the root, that is, the eutherian ancestral sequence at positions 100, 500, 800 and labeled them as being functional, non-functional and functional respectively. Using the rates and parameters as calculated in section 3.2 for a single TF, we simulated the evolution of these initial sites over the entire set of orthologous promoters using our binding site evolutionary model to obtain a dataset with known site functionality.

We evaluated the performance of our algorithm by generating 100 simulated datasets using the parameters for both ER and NF-κB transcription factors, and then calculating the accuracy of predictions made by our algorithm on each of the simulations. The algorithm was applied in the same way we would test a dataset with unknown site functionalities. Also, in addition to calculating the accuracy of the predictions, for each dataset we made two sets of predictions:

- the first with a constant value for the distance function (Section 2.1.1), $f(dist) = 0.1$,
- and the second with the original parameters used to generate the data.

In our binding site evolutionary model, the distance function is what differentiates functional and non-functional sites, along with constraining binding site evolution depending upon functionality. Therefore, comparing algorithm performance against a fixed turnover parameter gave us a clearer picture of whether our algorithm was actually picking up binding site turnover as opposed to identifying only those sites that have been linearly conserved in the given input.

Table 3–4 presents the results on simulated data for both sets of predictions. The first thing to note is that algorithm performance is quite good in both sets and for both TFs. We note that using a fixed distance function parameter leads to a lower specificity but higher sensitivity in the predictions and vice versa when the original parameters are used. This leads to a rather open interpretation and depending upon the goals of a future experiment, algorithm parameters could be accordingly tweaked.

The results also indicate that our predictor has a tendency in general to over predict functional binding sites. A detailed analysis of the predictions made shows us that most of the false positives are clusters of non-functional sites which were predicted as functional. A very prominent example of this is the gain of 5 non-functional sites around 1 conserved functional site going from laurasia to the bosTau2 sequences in the simulation. Our algorithm predicted all the sites as functional whereas, using the constant distance function, one of the false-positives was correctly predicted as non-functional. Biologically, the fact that our algorithm can capture these clusters is a good thing as functional binding sites rarely act in solitude [2].

Table 3–4: Simulation results for 100 sets of orthologous promoters generated using parameters for ER and NF-κB. Predictions were made on each tree using the original parameters and also by using a constant value for the distance function (represented by an asterisk), for each TF. 'N' and 'F' refer to non-functional and functional labelings respectively. The number of labelings for all 100 sets of orthologous promoters are summed to obtain the sensitivity and specificity.

	ER*	**ER**	**NF-κB***	**NF-κB**
Correct N predictions	1276	866	1519	1245
Total N labelings	1568		1911	
Specificity	81.38%	55.23%	79.49%	65.15%
Correct F predictions	2446	2597	2816	2898
Total F labelings	2671		2920	
Sensitivity	91.58%	97.23%	94.44%	99.23%

Analyzing the simulation results further, we note that our algorithm does a moderately good job of identifying binding site turnover. The caveat here lies in the definition of binding site turnover: our algorithm performs well when we allow the loss and the gain of a functional site very close (less than 100 bp) to an existing functional site and over a long time, that is, over two nodes on the phylogenetic tree or $\lambda > 0.1$. However over small distances ($\lambda < 0.1$), our algorithm is unable to catch binding site turnover, mainly because the parameters rarely generate such situations.

3.4 Results on ChIP-chip experimental data

Presented here are the results of applying our predictor on data obtained from ChIP-chip experiments. ChIP-chip is a technique that combines chromatin immunoprecipitation (ChIP) with microarray technology (chip). Like regular

ChIP, ChIP-chip is used to investigate interactions between proteins and DNA *in vivo*. More importantly, ChIP-chip allows the identification of binding sites of DNA-binding proteins on a genome-wide basis and whole-genome analysis can be performed to determine the locations of binding sites for almost any TF of interest [11].

We picked two papers that documented genome-wide ChIP-chip experiments for a particular transcription factor and given the regions provided, we were able to use the methodology from section 3.1 to create the relevant datasets with TFBS predictions. The first paper by Carroll et al. [10], documents a genome-wide ChIP-chip experiment for the Estrogen Receptor (ER) transcription factor. The second by Martone et al. [39], documents a genome-wide ChIP-chip experiment for the NF-κB transcription factor. The first experiment results in 5782 regions with an average of 1162bp per region and the second results in 1000 regions of 1300bp each.

A disadvantage of ChIP-chip technology is the fact that the size of DNA fragments achieved is at minimum 200bp. Considering that the length of the binding sites is much smaller, such an experiment can only reveal a broad picture of binding site locations. Therefore, when we applied our predictor on the ER regions, our predictor only predicted 33% of the sites found as functional (Table 3–5).

To determine whether the predictions made by our predictor were statistically significant, we applied our predictor with the same ER parameters on data that is not enriched with ER binding sites. Thus, we applied our predictor on the NF-κB

Table 3–5: Results of applying our algorithm on both ER (TRANSFAC Matrix: M00191) and NF-κB (TRANSFAC Matrix: M00054) ChIP-chip experimental data. The ER dataset was obtained from regions provided by Carroll et al. [10] and the NF-κB dataset was obtained from the regions provided by Martone et al. [39].

Predictions	**ER dataset**	**NF-κB dataset**
Functional labelings	3332 (33%)	193 (25%)
Non-functional labelings	5624 (66%)	598 (75%)
Totals	8956	791

dataset with ER parameters. Using a χ^2 test, we conclude that the predictions made on the ER dataset are very significant with $P < 0.001$ (Table 3–6). This leads us to believe that either the ER transcription factor is critical to biological processes leading it to having a large number of redundant sites, or that there exist other functional sites outside of the regions obtained from the *in vivo* experiment.

Table 3–6: χ^2 test to measure the statistical significance of predictions made with ER parameters on the ER dataset versus the NF-κB dataset. o_i and e_i are the observed and expected frequencies respectively.

Category	o_i	e_i	$o_i - e_i$	$\frac{(o_i-e_i)^2}{e_i}$
Functional labelings on ER	3332	3258.23	73.77	1.67
Functional labelings on NF-κB	214	287.77	-73.77	18.91
Non-functional labelings on ER	5624	5697.77	-73.77	0.96
Non-functional labelings on NF-κB	577	503.23	73.77	10.81
	n = 9747	9747		$\chi^2 = 32.35$

3.5 Conclusions

Using a methodology whereby we obtain a multiple sequence alignment of extant genome sequences and their reconstructed ancestral sequences, we generated dataset containing over 50 million bp and used the TRANSFAC database to make binding site predictions for each of its matrices (or TFs). We then used this data to estimate parameters for two well studied transcription factors: ER and NF-κb. We first evaluated our algorithm by calculating the accuracy of its predictions on simulated data. We next applied our algorithm on data obtained from genome-wide ChIP-chip experiments and evaluated its usefulness in making functional site predictions.

The results that we have obtained are a positive indication of the potential of our algorithm. Using the results on simulated data, we were able to successfully validate our model. The results on ChIP-chip data were also positive with functional site predictions on data enriched for ER binding sites being proportionally much higher than those on data not enriched for the same. Also, the number of functional sites predicted by our algorithm on ER regions was only 33% out of all the sites predicted by the data generation process. Taking into consideration the fact that our algorithm has a tendency to over-predict functional sites, it could be quite a useful tool to weed out false-positive predictions in ChIP-chip experiments.

3.6 Further Work

While our initial attempts at modeling binding site turnover have been successful, we note that there are certain improvements that could be made to enhance the speed and accuracy of our algorithm:

Calculation of the promoter content transition probability We note that having to enumerate through all possible event scenarios to calculate this probability is computationally very expensive. A simpler model with more relaxed constraints would speed up this process. For example, we could fix the functionality of sites and not allow them to change once determined at the leaves. Another option is to reduce the number of parameters which would lead to a quicker and a more accurate estimation.

Parameter estimation It is evident that the power of our algorithm is in the parameters that are used for each transcription factor. Hence, the weakest link in the chain is our sieving technique of parameter estimation, which is prone to hit local maxima. A better approach would be to use an Expectation–Maximization algorithm. The EM algorithm is an efficient iterative procedure to compute the Maximum Likelihood (ML) estimate in the presence of missing or hidden data. Each EM iteration is composed of two steps: Estimation (E) and Maximization (M). The M-step maximizes a likelihood function that is further refined in each iteration by the E-step [7]. However, even such a method is prone to hit local maxima.

Confidence score The addition of a confidence score to predictions would would help us in improving our algorithms sensitivity and specificity. At present the algorithm works as follows: the total likelihood for all possible labelings is computed from the leaves to the root; and the labeling is specified by the maximum conditional likelihood value at the root and propagated back to the leaves. This labeling is a simple binary value without any indication of

the confidence of such a prediction. Since we are interested in the human binding sites, we propose the calculation of the total likelihood score when one of the sites on the human branch is fixed to be functional or non-functional. Fixing the functionality of one site on the human promoter thus would enable us to calculate the total likelihood for all the labeling permutations for the remaining sites on the human promoter. These scores could then be used to calculate the level of confidence of our predictions. This addition will however, increase the computational complexity and may render our algorithm infeasible for genome-wide datasets.

CHAPTER 4
Summary and Conclusions

Transcription factor binding sites play an important role in gene regulation and generally, the transcriptional machinery is highly conserved during evolution. However, there are well documented examples in yeast and fly where the loss and gain of binding sites have lead to some dramatic changes in gene expression. At the same time there are other well documented examples where such binding site activity has produced little or no change in the same [48]. Due to their short length and the degeneracy of their binding requirements, binding sites are subject to a lot of evolutionary variation. Fortunately, binding site turnover, which is the coordinated loss of a site and the gain of a new one nearby helps in keeping disruptive changes in gene regulation at bay.

While studies on TFBS turnover have been done on insects and other micro-organisms, it is only recently that studies attempting to confirm the same have been attempted on mammals [17, 58]. Moreover, identification of functional elements has largely only been successful via phylogenetic footprinting and with the availability of accurate reconstructed mammalian ancestral genomes, the accuracy of this method has been greatly improved. Identification of transcription factor binding sites in mammals however, has largely been done via *in vivo* ChIP-chip or ChIP-seq genome-wide sequencing which, while useful, can only be used to

accurately identify regions where binding sites for a given TF are functional for the conditions under which the experiment was done.

In this thesis, we present a model that aims to identifying individual functional binding sites for a given TF as opposed to identifying functionally active regions *in vitro*. As explained in chapter 1, biologically, binding sites are not simple black and white entities that are labeled as functional or non-functional. It is quite possible that a binding site is only at times involved in regulating gene expression. Instances of this scenario include multiple TFs working together to increase production, thereby leading to a lot of sites being bound. However, our goal is to identify with a certain degree of accuracy, those sites that have a high chance of participating in gene regulation taking the concept of binding site turnover into account.

Since this study is a first of its kind, a major hurdle in establishing the accuracy of our predictor was the lack of standardized data and testing protocols. As a result, we relied on simulations to assess its efficacy and we note from our simulation tests that our predictor has a tendency to over predict sites as being functional, especially when encountering clusters of sites. However, this is biologically relevant and a good sign because functional binding sites rarely act in solitude [2]. Our algorithm also does a moderately good job in identifying binding site turnover, however we hope that with the availability of more genome-wide ChIP-chip or ChIP-seq experiments, we will be able to enhance our test suites and improve our algorithm in this regard.

While our presented mathematical model is useful, any further understanding of the actual underlying biological processes will have to be grounded with experimental data from *in vivo* effects of sites with various sequences on gene expression [48]. Thus, the next step will be to experimentally confirm the validity of predicted functional sites. Feedback from *in vivo* testing will also help us refine our model by analyzing the false negatives and false positives.

Finally, it would also be prudent to note that despite the various ways to minimize the number of false-positive functional binding site predictions, even those sites that are perfectly conserved across all species or meet the most stringent criteria for being classified as functional might still be non-functional in a genomic context, for instance, the binding site might be inaccessible owing to chromatin structure or blockage by other proteins [46]. Clearly, this study is but an initial attempt in the iterative refinement of our knowledge of the mammalian regulatory processes.

References

[1] R.C. Allen, H.Y. Zoghbi, A.B. Moseley, H.M. Rosenblatt, and J.W. Belmont. Methylation of HpaII and HhaI sites near the polymorphic CAG repeat in the human androgen-receptor gene correlates with X chromosome inactivation. *American Journal of Human Genetics*, 51(6):1229, 1992.

[2] B.P. Berman, Y. Nibu, B.D. Pfeiffer, P. Tomancak, S.E. Celniker, M. Levine, G.M. Rubin, and M.B. Eisen. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proceedings of the National Academy of Sciences*, 99(2):757–762, 2002.

[3] M. Blanchette, E.D. Green, W. Miller, and D. Haussler. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome research*, 14(12):2412–2423, 2004.

[4] M. Blanchette, W.J. Kent, C. Riemer, L. Elnitski, A.F.A. Smit, K.M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E.D. Green, et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome research*, 14(4):708–715, 2004.

[5] M. Blanchette, B. Schwikowski, and M. Tompa. Algorithms for Phylogenetic Footprinting. *Journal of Computational Biology*, 9(2):211–223, 2002.

[6] M. Blanchette and M. Tompa. Discovery of Regulatory Elements by a Computational Method for Phylogenetic Footprinting. *Genome Research*, 12(5):739–748, 2002.

[7] S. Borman. The Expectation Maximization Algorithm A short tutorial.

[8] J. Buhler and M. Tompa. Finding Motifs Using Random Projections. *Journal of Computational Biology*, 9(2):225–242, 2002.

[9] M.L. Bulyk. Computational prediction of transcription-factor binding site locations. *Genome Biology*, 5(1):201–201, 2004.

[10] J.S. Carroll, C.A. Meyer, J. Song, W. Li, T.R. Geistlinger, J. Eeckhoute, A.S. Brodsky, E.K. Keeton, K.C. Fertuck, G.F. Hall, et al. Genome-wide analysis of estrogen receptor binding sites. *Nature genetics*, 38(11):1289–1297, 2006.

[11] C.I.I.N.Y. CELLS. Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. *Current protocols in molecular biology*, 21:1–21, 2005.

[12] B. Charlesworth, P. Sniegowski, and W. Stephan. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature*, 371(6494):215–220, 1994.

[13] P.F. Cliften, L.D.W. Hillier, L. Fulton, T. Graves, T. Miner, W.R. Gish, R.H. Waterston, and M. Johnston. Surveying Saccharomyces Genomes to Identify Functional Elements by Comparative DNA Sequence Analysis. *Genome Research*, 11(7):1175–1186, 2001.

[14] F.S. Collins, E.D. Green, A.E. Guttmacher, M.S. Guyer, et al. A vision for the future of genomics research. *Nature*, 422(6934):835–847, 2003.

[15] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–3, 1970.

[16] E.T. Dermitzakis, C.M. Bergman, and A.G. Clark. Tracing the Evolutionary History of Drosophila Regulatory Regions with Models that Identify Transcription Factor Binding Sites. *Molecular Biology and Evolution*, 20(5):703–714, 2003.

[17] E.T. Dermitzakis and A.G. Clark. Evolution of Transcription Factor Binding Sites in Mammalian Gene Regulatory Regions: Conservation and Turnover. *Molecular Biology and Evolution*, 19(7):1114–1121, 2002.

[18] J.D. Dignam, R.M. Lebovitz, and R.G. Roeder. Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. *Nucleic Acids Res*, 11(5):1475–1489, 1983.

[19] J.E. Durbin, R. Hackenmiller, M.C. Simon, and D.E. Levy. Targeted disruption of the mouse Stat1 gene results in compromised innate immunity to viral disease. *Cell*, 84(3):443–450, 1996.

[20] L. Duret and P. Bucher. Searching for regulatory elements in human noncoding sequences. *Current Opinion in Structural Biology*, 7(3):399–406, 1997.

[21] T. El-Hay, N. Friedman, D. Koller, and R. Kupferman. Continuous time markov networks. In *Proceedings of the Twenty-second Conference on Uncertainty in AI (UAI)*. Citeseer, 2006.

[22] F. Fang and M. Blanchette. FootPrinter3: phylogenetic footprinting in partially alignable sequences. *Nucleic Acids Research*, 34(Web Server issue):W617, 2006.

[23] J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

[24] J. Fredslund. PHY·FI: fast and easy online creation and manipulation of phylogeny color figures. *BMC bioinformatics*, 7(1):315, 2006.

[25] N. Gompel, B. Prud'homme, P.J. Wittkopp, V.A. Kassner, and S.B. Carroll. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. *Nature*, 433:481–487, 2005.

[26] S. Guindon and O. Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, 52(5):696–704, 2003.

[27] D.L. Gumucio, H. Heilstedt-Williamson, T.A. Gray, S.A. Tarle, D.A. Shelton, D.A. Tagle, J.L. Slightom, M. Goodman, and F.S. Collins. Phylogenetic footprinting reveals a nuclear protein which binds to silencer sequences in the human gamma and epsilon globin genes. *Molecular and Cellular Biology*, 12(11):4919–4929, 1992.

[28] R.C. Hardison, J. Oeltjen, and W. Miller. Long Human-Mouse Sequence Alignments Reveal Novel Regulatory Elements: A Reason to Sequence the Mouse Genome. *Genome Research*, 7(10):959–966, 1997.

[29] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2):160–174, 1985.

[30] G.Z. Hertz and G.D. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7):563–577, 1999.

[31] J.D. Hughes, P.W. Estep, S. Tavazoie, and G.M. Church. Computational identification of Cis-regulatory elements associated with groups of functionally

related genes in Saccharomyces cerevisiae. *Journal of Molecular Biology*, 296(5):1205–1214, 2000.

[32] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[33] D.S. Latchman. Transcription factors: An overview. *International Journal of Biochemistry and Cell Biology*, 29(12):1305–1312, 1997.

[34] G.G. Loots, R.M. Locksley, C.M. Blankespoor, Z.E. Wang, W. Miller, E.M. Rubin, and K.A. Frazer. Identification of a Coordinate Regulator of Interleukins 4, 13, and 5 by Cross-Species Sequence Comparisons. *Science*, 288(5463):136, 2000.

[35] MZ Ludwig. Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. *Development*, 125(5):949–958, 1998.

[36] M.Z. Ludwig, C. Bergman, N.H. Patel, and M. Kreitman. Evidence for stabilizing selection in a eukaryotic enhancer element. *NATURE-LONDON-*, pages 564–566, 2000.

[37] M.Z. Ludwig, A. Palsson, E. Alekseeva, C.M. Bergman, J. Nathan, et al. Functional Evolution of a cis-Regulatory Module. *PLoS Biol*, 3(4):e93, 2005.

[38] J.F. Manen, V. Savolainen, and P. Simone. The atpB and rbcL promoters in plastid DNAs of a wide dicot range. *Journal of Molecular Evolution*, 38(6):577–582, 1994.

[39] R. Martone, G. Euskirchen, P. Bertone, S. Hartman, T.E. Royce, N.M. Luscombe, J.L. Rinn, F.K. Nelson, P. Miller, M. Gerstein, et al. Distribution of NF-κB-binding sites across human chromosome 22. *Proceedings of the National Academy of Sciences*, 100(21):12247–12252, 2003.

[40] A.M. Moses, D.Y. Chiang, D.A. Pollard, V.N. Iyer, and M.B. Eisen. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol*, 5(12):R98, 2004.

[41] A.M. Moses, D.A. Pollard, D.A. Nix, V.N. Iyer, X.Y. Li, and A. Sidow. Large-Scale Turnover of Functional Transcription Factor Binding Sites in Drosophila. *PLoS Comput Biol*, 2(10):e130, 2006.

[42] NCBI. What is a genome? (http://www.ncbi.nlm.nih.gov/about/primer/genetics_genome.html). Online article, March 2004.

[43] D.T. Odom, R.D. Dowell, E.S. Jacobsen, W. Gordon, T.W. Danford, K.D. MacIsaac, P.A. Rolfe, C.M. Conboy, D.K. Gifford, and E. Fraenkel. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature genetics*, 39(6):730–732, 2007.

[44] US National Library of Medicine. DNA Structure. Online image, March 2004.

[45] US National Library of Medicine. Eukaryotic Gene. Online image, March 2004.

[46] L.A. Pennacchio and E. Rubin. Genomic strategies to identify mammalian regulatory sequences. *Nature genetics*, 2(2):100–109, 2001.

[47] F. Piano, M.J. Parisi, R. Karess, and M.P. Kambysellis. Evidence for Redundancy But Not trans Factor-cis Element Coevolution in the Regulation of Drosophila Yp Genes. *Genetics*, 152(2):605–616, 1999.

[48] I. Reid. Transcription factor binding site turnover in mammals. Master's thesis, McGill University, 2007.

[49] R.G. Roeder. The role of general initiation factors in transcription by RNA polymerase II. *Trends in Biochemical Sciences*, 21(9):327–335, 1996.

[50] F.P. Roth1JT, J.D. Hughes, P.W. Estep, and G.M. Church. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnology*, 16:939, 1998.

[51] S.T. Smale. Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes. *BBA-Gene Structure and Expression*, 1351(1-2):73–88, 1997.

[52] J.R. Stone and G.A. Wray. Rapid evolution of cis-regulatory sequences via local point mutations. *Molecular Biology and Evolution*, 18(9):1764–1770, 2001.

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

[54] D.A. Tagle, B.F. Koop, M. Goodman, J.L. Slightom, D.L. Hess, and R.T. Jones. Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol*, 203(2):439–55, 1988.

[55] J.D. Thompson, D.G. Higgins, T.J. Gibson, et al. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22:4673–4673, 1994.

[56] A. Vierstraete. The Central Dogma of Molecular Biology. Online image, 1999.

[57] S. Vuillaumier, I. Dixmeras, H. Messai, C. Lapoumeroulie, D. Lallemand, J. Gekas, F.F. Chehab, C. Perret, J. Elion, and E. Denamur. Cross-species characterization of the promoter region of the cystic fibrosis transmembrane conductance regulator gene reveals multiple levels of regulation. *Biochemical*, 327:651–662, 1997.

[58] G.P. Wagner, W. Otto, V. Lynch, and P.F. Stadler. A stochastic model for the evolution of transcription factor binding site abundance. *Journal of theoretical biology*, 247(3):544–553, 2007.

[59] W.W. Wasserman, M. Palumbo, W. Thompson, J.W. Fickett, and C.E. Lawrence. Human-mouse genome comparisons to locate regulatory sites. *Nature Genetics*, 26:225–228, 2000.

[60] W.W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, 5(4):276–287, 2004.

[61] R.H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J.F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002.

[62] E. Wingender, X. Chen, E. Fricke, R. Geffers, R. Hehl, I. Liebich, M. Krull, V. Matys, H. Michael, R. Ohnhäuser, et al. The TRANSFAC system on gene expression regulation. *Nucleic Acids Research*, 29(1):281–283, 2001.

[63] Q. Wu, T. Zhang, J.F. Cheng, Y. Kim, J. Grimwood, J. Schmutz, M. Dickson, J.P. Noonan, M.Q. Zhang, R.M. Myers, et al. Comparative DNA Sequence Analysis of Mouse and Human Protocadherin Gene Clusters. *Genome Research*, 11(3):389–404, 2001.

[64] J.J. Yunis and W.G. Yasmineh. Heterochromatin, Satellite DNA, and Cell Function. *Science*, 174(4015):1200–1209, 1971.

Key To Abbreviations

A: adenine

bp: basepair

C: cytosine

CRM: *cis* -regulatory module

DNA: deoxyribonucleic acid

ER: Estrogen Receptor

G: guanine

mRNA: messenger ribonucleic acid

Mya: million years ago

NCBI: national center for biotechnology information

RNA: ribonucleic acid

rRNA: ribosomal ribonucleic acid

T: thymine

TF: transcription factor

TFBS: transcription factor binding site

tRNA: transfer ribonucleic acid

U: uracil