Tel-Aviv University

Raymond and Beverly Sackler

Faculty of Exact Sciences

School of Computer Science

A probabilistic model for the evolution of promoters

Thesis submitted in partial fulfillment of the requirements for

M.Sc. degree in the School of Computer Science, Tel-Aviv University

by

Daniela Raijman

The research work for this thesis has been carried out at

Tel-Aviv University under the supervision of

Prof. Ron Shamir

MAY 2007

Acknowledgments

First and foremost, I would like to thank my advisor, Prof. Ron Shamir, for introducing me to the exciting world of Computational Biology and guiding me through this world. His neverending support, advice and patience have made this work possible. A great thanks goes out to Amos Tanay, my collaborator on this work, for his quidance and ideas. Thank you to my lab members: Irit Gat-Viks, Igor Ulitsky, Chaim Linhart, Adi Maron-Katz, Israel Steinfeld, Ofir Davidovich, Yonit Halperin, Michal Ozery-Flato, Michal Ziv-Ukelson, Rani Elkon, Seagull Shavit and Michael Gutkin, for giving advice, a listening ear, and most of all for being my friends. I would like to thank my parents, for raising me to be curious, to love learning and appreciate education, and for always being there for me. Last but not least I would like to thank my brother and my friends, for making this journey a lot more fun.

Abstract

In comparative genomics one analyzes jointly evolutionarily related species in order to identify conserved and diverged sequences and to infer their function. While such studies enabled the detection of conserved sequences in large genomes and discovered specific motifs that are surprisingly conserved in those regions, the evolutionary dynamics of regulatory regions as a whole remains poorly understood. Here we present a probabilistic, process-based model for the evolution of promoter regions, combining the effects of regulatory interactions of multiple transcription factors into one principled model. The model expresses explicitly the selection forces acting on transcription factor binding sites, and its likelihood provides a method for scoring the goodness of fit between a regulatory model and a set of promoter alignments. We develop algorithms to compute likelihood and to learn *de-novo* collections of transcription factor binding motifs and their selection parameters. Using the new techniques, we explore the evolutionary dynamics in the *Saccharomyces* yeast species promoters. We validate our approach using previously reported regulatory motifs and infer new motifs. Additionally, we examine our methods using artificial simulations. Our study demonstrates how detailed understanding of the complex evolutionary dynamics in non-coding regions emerges from proper formalization of the evolutionary consequences of known regulatory mechanisms.

Contents

1 Introduction

Genomic regulatory regions harbor complex control schemes that collectively allow the genome to operate in a flexible and dynamic fashion. Such control schemes are encoded into the DNA sequence in a way that is not yet fully understood. Important elements of such regulatory code are short DNA sequences that are bound by transcription factors (TFs). TFs bind regulatory DNA specifically, by recognizing short motifs, and contribute to the assembly of complex switches that govern the transcription of a gene given various environmental or internal signals. Much of the current understanding on the way by which DNA determines the regulatory program of a gene is based on identification of TF binding sites (TFBSs) and their association with TFs of known function.

Despite remarkable progress in functional genomics technologies, and the ability to experimentally profile TF-DNA interactions on a genomic scale [19, 9], the understanding of function in regulatory regions remains a major challenge. At the same time, the complete sequencing of evolutionarily proximal genomes has made the detailed comparative study of regulatory regions possible. Consequently, comparative genomics has emerged as one of the central ways by which regulatory signals are computationally detected and studied. All comparative methods assume (explicitly or implicitly) an evolutionary model that distinguishes neutral sequences from functional ones. Most commonly [14, 26, 50, 13, 6], comparative studies focus on conservation, classifying sequence to be functional or non-functional by assuming that evolution in functional loci is slower. In yeast, many conserved loci were shown to correspond to TFBSs, allowing detection of novel sites that were not identifiable using functional methods.

As more species are sequenced, a desirable challenge is to extend the simple conservation-based studies by adding more structure to the function-evolution relationship in regulatory regions. In coding regions, our understanding of the genetic code makes sophisticated evolutionary predictions possible, e.g., by identifying cases of positive selection [12], correlated residues [11] and more. It is hoped that, by acquiring better, more detailed understanding of the function encoded by regulatory loci, one can greatly extend the utility of comparative studies in a similar way.

In this study, building on simple assumptions on the mechanisms of transcriptional regulation, we formalize an evolutionary model combining a neutral mutational process with selection on multiple heterogeneous TFBSs. We develop techniques for computing the likelihood of such a model given pairwise alignments and for learning maximum-likelihood model parameters. Using the new techniques, we can express a substantial part of the current functional knowledge on gene regulation in evolutionary terms and evaluate observed patterns of divergence and conservation based on this model. Applying our method to promoter sequences of yeast *Saccharomyces* species, we validate our approach and exemplify its use. Specifically, we discuss how the selection on BSs of different TFs vary in intensity, and how some families of similar TFBSs are in fact divided to subgroups that are separated by selection. Most interestingly, we compute the fraction of promoter sequence that is under selection due to characterized TFBSs, demonstrating a significant gap between earlier global estimations of selection in yeast promoters and the selection that can be directly attributed to TFBSs. This gap suggests that a significant part of the selection on yeast promoters is due to effects other than classical binding sites, e.g., chromatin organization effects and low affinity transcriptional interactions.

This thesis is organized as follows: In chapter 2 we present the necessary background for this work, and review some of the relevant literature. In chapter 3 we present the results of applying our model to biological data. The model is then presented in detail in the chapter 4. In chapter 5 we discuss our results and their implications, and present some possible future directions. In appendix A we present the results of applying our model to artificial data.

2 Background

2.1 Regulation of gene expression

DNA, which is comprised a double strand of nucleotides (Adenine (A), Cytosine (C), Guanine (G) and Thymine (T)), codes for almost every building block and function of the cell. In eukaryotic cells, the DNA is packed in units called *chromosomes*. In every chromosome there are many *genes*. Genes code for *proteins*, and these proteins are both the building blocks of cells, as well as the machines that keep the cell running. The cell makes proteins from the DNA blueprint by a series of processes: First the DNA is *transcribed* into *RNA*, then this RNA is processed into *mature RNA*, and then this mature RNA is *translated* into a sequence of *amino acids*, which together constitute a protein. The term *gene expression* refers to the first step, in which the DNA is transcribed to RNA. This process is performed by an enzyme called *DNA polymerase*, and aided by regulatory proteins, which we discuss later.

Although all cells in a single organism contain the same DNA, the proteins in the cells differ qualitatively and qualitatively depending on the developmental stage, organ, environmental condition and more. How is this possible? The answer is, to a large extent, *regulation of gene expression*. Every gene is preceded by a regulatory sequence region, which is believed to be several hundreds or several thousands of nucleotides (or *base pairs*) upstream of the gene, depending on the organism. These regions are called *promoters* (**Fig 1**). Promoters contain many short subsequences, which are bound by proteins called *transcription factors (TFs)*, which are themselves gene products. The sequences of DNA bases that are bound by these proteins are called *transcription factor binding sites (TFBSs)*. A typical TFBS will be $6 - 20$ base pairs (bp) long. A single regulatory protein will often recognize a variety of similar sequences. In some cases, only a small fraction of the occurrences of a particular sequence will actually be bound by the TF. These and other reasons make the discovery of TFBSs extremely challenging. Despite extensive biochemical, molecular and computational analysis, it remains unknown, for most if not all TFs, exactly where in the genome they will

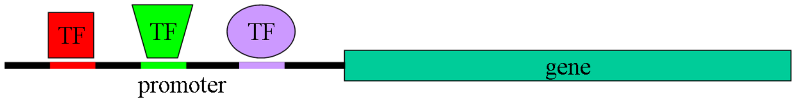


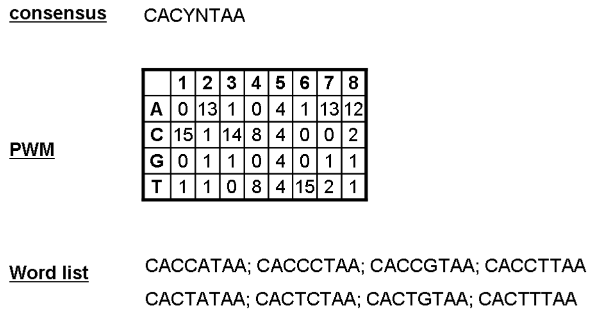
Figure 1: **An illustration of a promoter region**. The region upstream of the gene is termed the *promoter region*. In this example, the promoter region bears binding sites for three TFs. In the figure, we see all three TFs bound to their respective binding sites.

bind, and under what conditions.

In recent years a lot of evidence implied that regulation has a very large role in diversity - between species, between individuals, between developmental stages and between different cell types. Studying the evolution of TFBSs can help us to gain a better understanding of gene expression regulation. Knowing what selection forces are acting on these regions will give us an insight into their function.

2.1.1 Motif Models

A ‘typical’ TFBS, or motif, can be represented in several ways (**Fig 2**). The simplest way is using a consensus sequence, either of nucleotides, or allowing degenerate positions using the IUPAC nomenclature [2]. Another representation is using a *position weight matrix (PWM)* or *position specific score matrix (PSSM)*. In these representations, a score is given for the occurrence of each nucleotide in each position, and the total match score between a motif and a given sequence is calculated by multiplying these scores. A threshold is then used to decide if a particular sequence is a match to this matrix. Both representations mentioned above cannot represent dependencies between sites. More complex models, e.g, Markov models, can be used to represent these dependencies. The simplest model that can represent dependencies is a list of all words bound by the TF.



consensus CACYNTAA

PWM

	1	2	3	4	5	6	7	8
A	0	13	1	0	4	1	13	12
C	15	1	14	8	4	0	0	2
G	0	1	1	0	4	0	1	1
T	1	1	0	8	4	15	2	1

Word list
CACCATAA; CACCCTAA; CACCGTAA; CACCTTAA
CACTATAA; CACTCTAA; CACTGTAA; CACTTTAA

Figure 2: Different representations of the same motif. In the consensus representation, N represents any nucleotide, Y represents either C or T. The PWM gives the relative frequency of each nucleotide in each position, and must be used with a score threshold. The word list contains all the hits of the motifs, and generalizes a consensus.

2.2 Understanding transcription regulation

Many studies have been dedicated to understanding the regulation of gene expression in different organisms. The first studies focused on the regulation by a specific TF, but in recent years there has been a shift towards genome-wide analysis. This problem can be approached both biologically and computationally.

2.2.1 Biological approaches

Classical experiments: The first biological studies in this field focused on trying to find TFBSs for one specific TF in one promoter. The initial approach for this was by deletion experiments. In this method different segments of the promoter are deleted, and the effect on the regulation is examined. In this manner, one is able to determine what parts are essential for proper regulation. Another approach is by DNAse footprinting [28]. This method uses the fact that a protein bound to DNA will usually protect it from enzymatic cleavage. The enzyme deoxyribonuclease (DNAse) is used to cut the DNA, and then gel electrophoresis is used to detect the resulting pattern of cleavage. This pattern can then be compared to the pattern in the absence of the DNA-binding protein. The fragments missing in the first gel are referred to as the TF's 'footprint', which identifies the TF's binding site.

High-throughput experiments: Chromatin immuno-precipitation (ChIP) is a procedure used to determine in-vivo, for a given protein, what DNA sequences are bound by that protein. In this procedure, the protein is allowed to bind DNA sequences, and then cross-linked to the DNA using formaldehyde. The DNA is then sheared, and antibodies specific to the DNA binding protein are used to isolate the protein-DNA complexes. The unbound DNA is removed, and the cross-linking is reversed. At the end of this procedure, we can isolate the sequence fragments that were bound by the protein. A more advanced procedure, ChIP-on-chip (also called location analysis), takes this one step further. The extracted fragments, which were bound by the TF, are hybridized to a DNA microarray (chip), containing sequences specific to each promoter, thereby identifying the loci that are bound by the TF.

The most comprehensive ChIP-on-chip study in the yeast species *S. cerevisiae* to date is due to Harbison et. al. [19]. In this experiment, genome-wide location analysis was used to determine the genomic occupancy of 203 DNA-binding TFs in rich media conditions. For 84 of these, additional experiments were performed in different environmental conditions. The microarrays used in this experiment consist of spotted PCR products representing the intergenic regions of the *S. cerevisiae* genome. In total, the experiment revealed some $11,000$ unique interactions between TFs and promoter regions at high confidence ($P \leq 0.001$). These results were then processed by six motif discovery programs (some examples of which are explained below) and motif PWMs were identified for many TFs. In a follow-up study by MacIsaac et. al. [32], data for conservation between different yeast species were used to improve on these results.

2.2.2 Computational approaches

The computational methods for motif discovery can be divided into two classes. In the first class, ChIP-on-chip or expression data are used to find groups of genes that are likely to be co-regulated due to a response to a particular TF or due to co-expression. These groups are then searched for over-represented motifs. In the second class, comparative genomics is used to find, in alignments of promoter regions of orthologous genes from different species, elements that are much more conserved than expected, and therefore suspected to be functional. For a recent review of computational methods, see [25].

Using a candidate gene group: Many computational tools for finding TFBS motifs receive as their input a group of regulatory regions of genes that are believed to be co-regulated. The co-regulated sets are often obtained using clustering methods on expression profiles, or using ChIP data. These tools identify short sequence motifs that are statistically overrepresented in these regulatory regions. In [16], many of these methods are reviewed and compared. Two main categories of these algorithms are enumerative and model-based. An example of an enumerative algorithm is Weeder [17]. In this study, suffix trees are used to discover short sequences of nucleotides that appear in at least q promoter sequences, with at most $\epsilon * m$ mismatches, where m is the length of the pattern, and ϵ and q are a pre-specified parameters. One of the most widely used model-based methods is MEME [46]. This method assumes that sequences are derived from a mixture model of two components - a probability matrix representing a motif, and a Markovian background model. The probability matrix for the motif is optimized using Expectation-Maximization.

Using Comparative Genomics: Computational tools from the second class do not use groups of co-regulated genes, but rather sets of two or more aligned orthologous promoters. Some of the first studies in comparative genomics approached the problem of finding regulatory elements using sequence conservation by searching for highly conserved regions in promoter alignments. The Phylogenetic Footprinting approach [15] takes advantage of the fact that non-functional sequences are likely to diverge during the course of evolution, while functional sequences will tend to remain conserved due to selective pressure to maintain their function. In this approach, one searches for ”footprints” of evolution - regions in the genome that remained conserved during the course of evolution. For example, Cliften et. al. [13] searched for such footprints in multiple alignments of six *Saccharomyces* species. A similar approach is Phylogenetic Shadowing [6], which can be used on very closely related species, such as human and chimp.

Kellis et. al. [26] also used a comparative analysis of four *Saccharomyces* species to identify regulatory elements. In this study, rather than separating out conserved areas and then looking for enrichment in these areas, they approached the problem from an enumerative angle. The focus was on content, searching for motifs whose appearances are conserved beyond a certain conservation threshold. The analysis revealed 72 well-conserved sequence motifs that occur frequently throughout the yeast genome. Most known regulatory motifs are represented in this set. Many additional, previously uncharacterized motifs were discovered.

Combined approaches: Some motif finders use a combination of both approaches. Early studies pooled together orthologs from all species of the genes in the co-expressed group, and searched for motif enrichment in this pool. Another, two-step strategy searches for motifs that meet both criteria (enrichment and conservation) in a serial manner. Tools that use both conservation and over-representation together give much better results than each approach separately. PhyME [41] and PhyloGibbs [39] are two examples. In these methods, evolutionary models are used to describe orthologous sequence alignments across species. A group of orthologous sequences is viewed as generated from an ancestral sequence. The ancestral sequence is modeled as a mixture of background and TFBS segments, and its evolution into the observed sequences is assumed to be different for neutral sites and TFBS sites.

Using conservation to distinguish functional sites from non-functional sites: Even when we know, with some degree of confidence, the binding specificity of a TF (i.e., its PWM and the corresponding threshold, or its consensus sequence), it is still not straightforward to determine where in the DNA that TF is bound, as not all sites matching the motif with the required specificity will actually be bound. A common approach used in many recent studies (e.g., [3]) is to combine known TFBS specificities with conservation information. A site with the required specificity will be declared functional only if it also satisfies certain conservation criteria.

2.3 Evolutionary Models

An evolutionary substitution model describes the process by which one DNA (or amino acid) sequence evolves into another. Such mathematical models are used, for example, towards the construction of phylogenetic trees, or for the simulation of evolutionary processes.

2.3.1 Simple Evolutionary models

The models we describe in this section assume neutrality and independence of sites. There are many known models with different complexities for DNA site substitution. The first and simplest model was proposed in 1969 by Jukes and Cantor [45]. This model assumes that all four bases are equally likely and that all mutations are equally likely and occur at a constant rate across time, and therefore the model only has one parameter - the overall rate of mutation. A more complex model, due to Kimura [30], distinguishes between transitions (A/G or C/T mutations) and transversions (A/C, A/T, G/C, G/T mutations), which have different rates in nature. A further improvement to this model, due to Hasegawa, Kishino and Yano [29], also considers unequal base frequencies.

A strong assumption made by these models is the existence of a 'molecular clock', i.e., that evolution rate is approximately constant over time. Another assumption is that this rate is the same for all positions in the sequence. While it is clear that these assumptions are unrealistic, it is hard to create a model that will be simple and general, and at the same time incorporate complex biological effects that occur during evolution. Many attempts were made to formulate more complex models, that take additional factors into account. The tradeoff is clear: the more complex the model is, the harder it is to learn its parameters, and the higher the risk of overfitting.

2.3.2 Physical dependency

There are many reasons to believe that the mutational rates in neighboring sites are not independent (For reviews on the causes for this phenomenon see [40] and [5]). Sites that are not directly next to each other can also be dependent. For this reason, it is necessary to develop evolutionary models that allow dependence of evolutionary rates among sites.

Several studies developed evolutionary models allowing context dependency. Schoniger and von Haeseler [31] presented a model for autocorrelated sequences of nonoverlapping doublets of nucleotides, incorporating dependencies between pairs of nucleotides (which do not necessarily have to be adjacent). Felsenstein and Churchill [20] proposed a model that allows variation in the evolution rates of different sites and correlation between the rates of adjacent sites. Siepel and Haussler [40] introduced a context dependent model that allows the dependency of a site on its N-1 predecessors. They used a phylogenetic model composed of four elements - an equilibrium frequency vector for N-mers, a tree topology, branch lengths and an N-mer substitution rate matrix. An EM approach was used to find a maximum likelihood solution for the transition probabilities and branch lengths.

2.3.3 Functional dependency

The models discussed above use physical context to establish richer evolutionary models. Consequently, they focus on location rather than content. The models reviewed in this section focus on functional context. The evolution of a single site is examined in light of its effect on the *function* of the sequence. In other words, these models assume that the phenotype affects the evolution of the genotype. For example, in the evolution of protein-coding sequences, the effect of a single nucleotide mutation depends on how it changes the amino acid coded by the *codon* to which that nucleotide belongs. In many cases, a mutation in the third nucleotide of a codon will not change the amino acid coded by it. Goldman and Yang [35] modeled dependency between bases of the same codon. This was done by introducing a substitution matrix not for nucleotides or amino acids, but rather for codons. This model gave much greater accuracy in phylogenetic estimations, at the cost of an increased computational complexity. Still, this model assumes independence of neighboring entities - only in this case the entities are codons. This independence allows the model likelihood to be factored into a product of distinct codons. A similar model due to Muse and Gaut [42], with less parameters, imposed a penalty for substitutions between codons that do not code for the same amino acid.

Extending the dependencies beyond the scope of a single codon, Pedersen et. al. [4] introduced a more complex model for lentiviral (HIV) evolution, allowing the rate of substitution at a specific site to depend on the sequence in the neighborhood of the site at the time of the mutation. Lentiviral genes have an extremely low level of CpG nucleotides, and a high bias toward A nucleotides. These facts are taken into consideration in the model - a mutation resulting in the generation of a CpG has a reduced probability, while a mutation eliminating a CpG has an increased probability, even in cases where the CpG crosses codon boundaries. Mutations that alter the amino acid sequence also have a reduced probability. The same approach was later extended to include additional effects, e.g. overlapping reading frames [21, 36].

Another example of an evolutionary model that accounts for functional context is the study of Yu and Thorne [51]. Since RNA folds to secondary structures in a mechanism that is highly dependent on base-pairing, the model assumes that selection on mutations is highly dependent on the free energy of the secondary structure before and after the mutation. This model creates correlation between sites that are paired in the secondary structure, and hence are some distance apart.

2.4 Global approaches to the evolution of gene regulation

In recent years, several studies addressed the issue of evolution of gene regulation. Most of these studies focused on the regulation of a specific gene or by a specific TF. In this section, we survey some studies that address the evolution of promoter regions on a genome-wide scale.

Chin et. al. [10] focused on quantifying the amount of negative selection in yeast promoters. In this study, two approaches were taken. The first approach used a hidden Markov model with two states, for the purpose of dividing the promoter regions into neutral and high conservation regions. To estimate the amount of promoter DNA under selection, the authors used a second approach, counting the numbers of blocks of n consecutive conserved bases in the promoter sequences, for different values of n, and comparing this number to the expectation computed using simulations with a uniform model. This approach led to the conclusion that about 30% of *Saccharomyces* promoter regions are under selection.

Taking a different approach, Tanay et. al. [44] studied selection forces acting on individual k-mers, using whole-genome multiple alignments of promoter regions for four yeast species. The authors focused on octamer conservation and on single nucleotide substitutions between octamers, and used those to construct a network of the selection forces acting on octamers. For every alignment, starting with the known tree of the four species with the aligned sequences at the leaves, they inferred ancestral sequences using maximum parsimony, and computed for each branch in the tree the observed and expected counts for every possible single-nucleotide substitution between two octamers, as well for octamer conservation. The expected substitution count was computed using a position independent, 16 parameter substitution matrix. For each single-nucleotide substitution between two octamers, and for the conservation of each octamer, they computed the logarithm of the ratio between the observed and expected counts, and defined it as the normalized substitution rate. Using these rates, they constructed a selection network, in which each node represents an octamer, and an edge exists between two octamers if they differ by a single substitution. The node weights were defined to be the normalized conservation rates, while edge weights were defined to be the normalized substitution rates. In this network, they searched for clusters, i.e., sets of highly conserved nodes, each forming a connected component with non-negative edge weights between set members. Each cluster is a collection of octamers which together correspond to a motif. The motifs were then compared to known TFBSs in order to annotate them. Indeed, many of these motifs corresponded to known TFBSs.

The analysis gave rise to new models for binding site activity, identified families of related binding sites, and characterized the functional similarities among them. This study provided a whole-genome view of the selection forces acting on promoter regions.

2.5 Markov Chains

In this section we briefly review some of the basic probability models that we shall use. For more details see, e.g., [23]. A *random process*, or *stochastic process*, is a collection of indexed random variables, taking values from some *state space* S. The index set I may be discrete or continuous, and the state space may be finite, countable or uncountable, discrete or continuous. Here we focus on finite state space models. Stochastic processes are of interest for describing the behavior of a system operating over some period of time. In that case the index set I is referred to as the time, which can be discrete or continuous.

We say that a random process is a *Markov process* if is satisfies the *Markov property*, namely, given the current state of the system, its future states are independent of its past states (i.e., the process is *memoryless*). In *discrete-time Markov chains*, the system is observed at discrete points in time, $t = 0, 1, 2, ...$, and X_t represents the state of the system at time t and takes values in the finite set $S = 0, .., N-1$. A probabilistic *Markov model* for such a process is define as follows. The parameters of that model are the *initial probability distribution* $p(i) = P(X_0 = i), i = 0, 1, .., N-1$, and the *transition probabilities* - a stochastic matrix P in which $P_{ij} = P(X_t = j|X_{t-1} = i)$. The matrix satisfies $\sum_j P_{ij} = 1$. This formulation assumes that the process is *time homogenous*, i.e., transition probabilities remain constant throughout the process. An example is shown in **Fig 3**. Given a Markov model and a sequence of observations, we can compute the probability for the observations: $P(X_0 = x_0, ..., X_n = x_n) = p(x_0) * \prod_{i=1}^{n} P_{x_{i-1}x_i}$.

There are several properties of Markov chains that are of particular interest:

1. **Reducibility**: A Markov chain is said to be *irreducible* if it is possible to get from every state to every other state, using any number of transitions, with a non-zero probability.

2. **Periodicity**: A state i is said to have *period* k if it is only possible to return to state i after exactly kl time steps for some $l = 1, 2, ...$, and k is the maximal number with such property. When $k = 1$ the state is said to be *aperiodic*. An irreducible Markov chain is said to be *ergodic* if all its states are aperiodic.

3. **Recurrence**: A state i is said to be *recurrent* or *persistent* if, given that we start at state i, the system must return to that state after some finite number of time steps. A state that is not recurrent is said to be *transient*. A state is said to be *positive recurrent* if the expected return time to that state is finite.

4. **Ergodicity**: A state i is said to be *ergodic* if it is aperiodic and positive recurrent. A Markov chain is ergodic if all of its states are ergodic.

In some systems, the time between transitions can vary continuously. While in the discrete-time example (**Fig 3**) we observed the weather at different days, in a continuous-time scenario we may want to observe the weather over a continuous time scale. The Markovian assumption here is that the process is memoryless - Given X_t, the distribution of X_{t+h} is independent of the history previous to time t, for any real-values $h > 0$. When discussing *continuous-time Markov chain*, the index set I for the variables is continuous, and therefore the model must be defined differently. Instead of using transition probabilities, we need to define a probability model for $P(X_{t+h} = j|X_t = i)$, for any $h > 0$. Continuous-time Markov processes are usually defined using a *transition rate matrix* Q, where the element Q_{ij} describes the transition *rate* from state i to state j in one time unit, and the i-th diagonal element is given by $Q_{ii} = -\sum_{j \neq i} Q_{ij}$. It can be shown that the transition probabilities in time t are given by the matrix $exp(tQ) = \sum_{n=0}^{\infty} \frac{(tQ)^n}{n!}$. Hence $exp(tQ)_{ij}$ is the probability of transition from state i to state j in time t. The properties discussed above apply for continuous-time Markov chains as well.

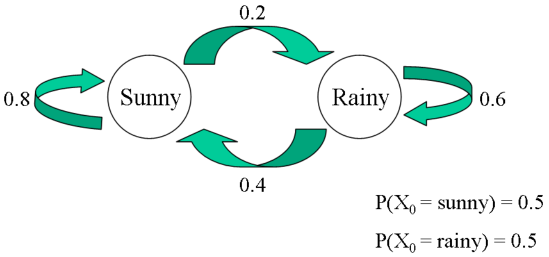


Figure 3: A basic example of a discrete-time Markov model representing the weather in different days. X_i represents the weather in the i-th day. The weather can either be sunny or rainy. If a day is sunny, there is a probability of 0.8 that the following day will also be sunny, and of 0.2 that the following day will be rainy. If a day is rainy, with probability 0.6 and 0.4 the following day being rainy and sunny, respectively. The initial probabilities at time 0 are also part of the model.

3 Results

We developed an integrated model for the evolution of promoters under the influence of a *regulatory code* (see chapter 4). Briefly, the code is a collection of distinct DNA motifs, where each motif (corresponding to a TF) is represented by a set of nucleotide k-mers. We assume that each set (termed *target set*) contains all k-mers recognized by the TF (See **Fig 4A** for an illustration), and any appearance of a k-mer from the target set is declared to be a BS. All k-mers in the same target set are of the same length, but are otherwise unlimited by physical constraints. In practice target sets are usually variations over a consensus sequence. Our model represents a simplification of a much more complex biological reality, by assuming that binding at each locus is completely determined by the existence of a motif, and is either perfect or non-existing (therefore ignoring differences in binding affinity between k-mers of the same target set).

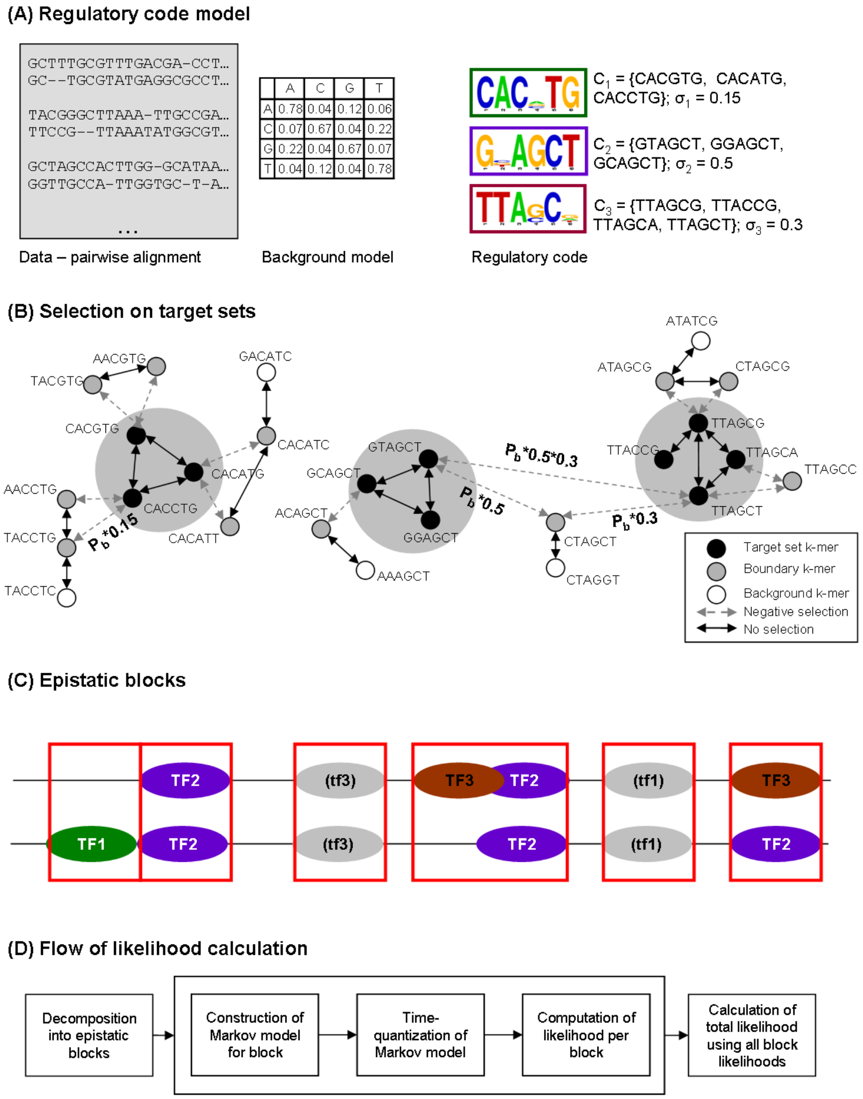
To model the evolution of a promoter region, we assume that sequences are evolving neutrally, except for loci affected by selection on TFBSs. Each target set (and therefore each TF) is associated with a *selection coefficient* $0 \leq \sigma \leq 1$, which represents the fixation probability of a mutation introducing or eliminating a BS. Smaller σ values represent stronger selective pressure on loci bearing k-mers that belong to a given target set. Our model assumes that each appearance or loss of a TFBS is selected against. The replacement of a k-mer from a target set by another k-mer from the same target set is not selected against, since according to our model all k-mers in the same target set are equivalent BSs for the corresponding TF. This simple functional model allows us, once equipped with a regulatory code, to write down a Markov model representing the evolutionary process for an entire promoter sequence.

The evolutionary forces outlined in the above model affect the evolutionary rate at a single base if it is in the context of a TFBS. The evolution of one base can therefore depend on several adjacent bases, and the model formalizes this type of *epistasis* using simple assumptions on TFBSs and their function. Although the epistasis considered by the model is simple and spatially limited (including only BS k-mers), principled computation of the likelihood of a regulatory code given a multiple or even pairwise alignment is very difficult and requires several approximations and heuristics (see chapter 4). We developed algorithms for approximate calculation of the likelihood of a model, which provide us with a method for evaluating how much our model agrees with the patterns of divergence in the alignment. We score models by comparing their likelihood to that of a null model containing no target sets, deriving a log-likelihood ratio (LLR). We developed algorithms for maximum likelihood estimation of selection coefficients, and for learning target sets with optimal likelihood (see chapter 4).

We focused on the evolutionary dynamics of *Saccharomyces* gene regulatory regions. The yeast system has the advantage of many well documented TFBS motifs and clearly identifiable promoters, and was used before in many studies of transcriptional regulation and its evolution [44, 26, 13]. We extracted pairwise alignments from multiple alignments of *Saccharomyces sensu stricto* species (see chapter 4). The resulting alignments for *S. cerevisiae-S. mikatae* consisted of over $900,000$ aligned bases from the upstream regions of $3,503$ genes, with 74.2% identity. We derived similar alignments for *S. cerevisiae-S. paradoxus*, and for *S. cerevisiae-S. bayanus*.

3.1 Literature based regulatory code

We started by constructing an evolutionary model from known TF binding models. We used the compendium of TFBSs composed by MacIsaac et al. based on an extensive ChIP-on-chip data set and literature review [32]. Out of 124 consensus sequences reported by the authors (in IUPAC format), we chose those 94 that translated to target sets containing at most 512 k-mers each, and had at least 5 matches in the *S. cerevisiae-S. mikatae* promoter alignments. We formed an integrated model by starting from an empty model, and incrementally attempting to add each of the 94 target sets (in a pre-specified order). For each candidate target set in turn, we tentatively added it to the model and inferred an optimal selection coefficient σ for it. We then tested whether expanded model with the added target set has a higher model likelihood. If it does, the target set is accepted to the model. Otherwise it is rejected. In total, we accepted 74 target sets



(A) Regulatory code model
Data – pairwise alignment
Background model
Regulatory code
CAC_TG
G_AGCT
TTAgC
C1 = {CACGTG, CACATG, CACCTG}; σ1 = 0.15
C2 = {GTAGCT, GGAGCT, GCAGCT}; σ2 = 0.5
C3 = {TTAGCG, TTACCG, TTAGCA, TTAGCT}; σ3 = 0.3
(B) Selection on target sets
AACGTG
TACGTG
GACATC
CACGTG
CACATC
CACATG
AACCTG
CACCTG
TACCTG
CACATT
TACCTC
Pb*0.15
GTAGCT
GCAGCT
ACAGCT
GGAGCT
AAAGCT
Pb*0.5*0.3
Pb*0.5
CTAGCT
CTAGGT
Pb*0.3
ATATCG
ATAGCG
CTAGCG
TTAGCG
TTACCG
TTAGCA
TTAGCC
TTAGCT
Target set k-mer
Boundary k-mer
Background k-mer
Negative selection
No selection
(C) Epistatic blocks
TF2
TF1
TF2
(tf3)
(tf3)
TF3
TF2
TF2
(tf1)
(tf1)
TF3
TF2
(D) Flow of likelihood calculation
Decomposition into epistatic blocks
Construction of Markov model for block
Time-quantization of Markov model
Computation of likelihood per block
Calculation of total likelihood using all block likelihoods

```
GCTTTGCGTTTGACGA-CCT…
GC--TGCGTATGAGGCGCCT…

TACGGGCTTAAA-TTGCCGA…
TTCCG--TTAAATATGGCGT…

GCTAGCCACTTGG-GCATAA…
GGTTGCCA-TTGGTGC-T-A…

…
```

	A	C	G	T
A	0.78	0.04	0.12	0.06
C	0.07	0.67	0.04	0.22
G	0.22	0.04	0.67	0.07
T	0.04	0.12	0.04	0.78

Figure 4: **An integrated model for evolution under the effect of multiple TFs. (A) The regulatory code model** is defined by a set of motif target sets identified by specific TFs, and selection coefficients that quantify the intensity of selection on substitutions affecting them. In this example the model consists of three target sets for three TFs. The background substitution probabilities are also part of the model. **(B) The implications of the model regarding selection on k-mers.** Substitutions between k-mers that belong to a target set (large grey circles) and k-mers that do not belong to that target set are under negative selection. The probability for that substitution is the background probability multiplied by the selection coefficient. For substitutions between k-mers that belong to two different target sets, their two coefficients are multiplied. **(C) Epistatic blocks.** To compute the likelihood of a model given alignments, we first decompose the sequence into blocks (epistatic blocks) that evolve almost independently of each other. Epistatic blocks include overlapping TFBSs or k-mers that are one substitution away from a target set (boundary k-mers). Because we model motifs of limited size, and since binding site density in yeast is not very high, we derive blocks of reasonable sizes (less than 20bp) for further analysis. Ovals marked TFX represent an TFBS for TF X. Ovals marked (tfX) represent boundary k-mers for TF X. **(D) Computing likelihood.** A schematic representation of the likelihood calculation. The large rectangle in the center is repeated for each block.

(79%), obtaining a final LLR of 1139.28. Similar results were obtained for the other two alignment datasets.

We next studied possible factors that contribute to acceptance or rejection of literature target sets to our model. According to our model, every appearance of a motif is considered to be functional and under selection. In reality, not all appearances are necessarily functional, and some may be functionally different than others. We examined the correlation between the number of appearances of k-mers from a target set in the data and the model acceptance rate (**Fig 5**). While 87% of target sets with $0 - 149$ hits were accepted, only 77% of target sets with $150 - 499$ hits were accepted, and only 47% of target sets with over 500 hits were accepted. These results suggest that the specificity of some the literature target sets may be too low to allow acceptance by our rather stringent model.

To try and control for motif specificity in a systematic way, we next examined, for each TF, a model constructed using a limited data set, containing only pairwise alignments of promoters that were found to be bound by that TF in ChIP-on-chip experiments (using a $p < 0.005$ cutoff) [32]. Since the set of ChIP bound promoters is different for each TF, we could not construct an integrated model in this case, but simply computed the LLR for each TF. We call the resulting model *the ChIP model*. Out of 62 TFs with at least 5 hits in the ChIP bound promoters, 52 target sets had positive LLR (84%), of which 6 were not accepted in the original model (Spt2, Ndd1, Swi5, Bas1, Hap2 and Met31). All of the target sets that were accepted in the original model but not in the ChIP model (27 in total) did not have enough hits in the ChIP data to be considered for the model.

In summary, although our model assumptions are deliberately simplistic, they prove to be enough for capturing a large fraction of the known binding sites in the yeast genome, and they appear to be only marginally biased by a large number of false positives (as confirmed by having similar results in the ChIP dataset).

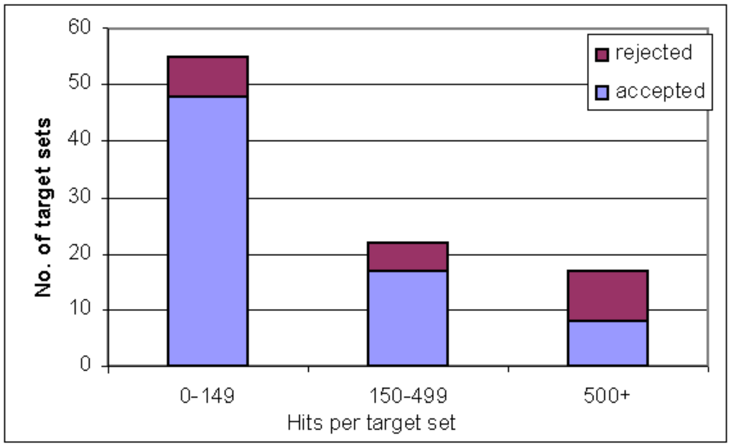


Figure 5: **Effect of target set size and hit number on acceptance** The histogram shows the number of accepted and rejected target sets for different target set sizes, for the known model (section 3.1). A target set was accepted if it had a positive contribution to the likelihood. Target sets with a large number of hits (> 500) were accepted with a much lower probability than target sets with less than 150 hits. An extremely high hit number may indicate low specificity of the BS, so a low fraction of those hits are actually functional BSs and thus under selection.

3.2 Learning a regulatory code de-novo

We next applied our model learning algorithm to construct a regulatory model de-novo, limiting the motif width to $6-12$, the typical range of well-conserved bases in a binding site. The learning algorithm produced a model containing 62 target sets when executed on the *S. cerevisiae-S. mikatae* alignments (See **Fig 6** and **Table 1**). The LLR for this model was 2456.9. We annotated the target sets in the resulting model by searching for matches to known TFBSs (using data from TRANSFAC [49], SCPD [52], SGD [1] and the MacIsaac et al. collection [32]). A target set was declared to match a known TFBS if one of the k-mers in the target set was a substring of, or contained as a substring, the TFBS consensus sequence. We manually annotated a few additional motifs (e.g., PAC and RRPE). After this process, only 11 of the 62 target sets remained unmatched. Our target sets match around 45 known motifs. We note that most of the target sets that we learned are relatively specific, with limited redundancy.

When modeling the evolutionary effects of a large set of TFs in an integrated fashion, we can study relations between different target sets. Most importantly, we can assume that each inferred target set represents a distinct function and that no redundant target sets would be inferred. The reason for this is that substitutions between k-mers in two redundant target sets of the same TF would be predicted by the model to be selected against (multiplying the background probability of substitution by the selection coefficients of each of the target sets), while in fact, substitution between redundant target sets must behave neutrally. This discrepancy should result in lower likelihood for a model that includes two redundant target sets. The residual conservation of erroneous target sets that partially overlap a real binding sequence is therefore unlikely to affect our learning process. Looking at the results, we indeed see only few cases of seemingly redundant target sets, each of which we show to be either a computational artifact or a biologically significant effect.

In the first type of model redundancy, one target set contains substrings of another (e.g., GCGATGAGATG and CGATGAG in the PAC motif). This can be accounted for by the inaccuracy of the discrete binding assumption. If one target set represents strong (more specific) binding sites and the other represents weaker sites, then having two target sets with different selection coefficients improves likelihood. In this case, according to our model the selection on a locus with the more specific version is calculated as if it were part of binding sites for both target sets (implying a de-facto stronger selection on it). In contrast, the selection on a locus with the less specific version would be affected only by the selection coefficient of one target set. This is exemplified in **Fig 7A**.

In the second type of model redundancy, two k-mers from distinct target sets differ in one position (e.g., TTACCCG and TTACCCT in the Reb1 motif, TATTTATA and TATTTACA in the Rlm1 motif). In this case, the likelihood of a model in which the two target sets are combined into one is lower than the likelihood of the redundant model. This suggests that the separation between the two target sets is selected for, possibly since BSs from each set are functioning differently (**Fig 7B**). Examples for such separation were argued for heuristically and demonstrated experimentally before [48, 44], but now we are equipped with the computational means to quantify such selection.

As shown in **Fig 7C**, substitutions between target sets that are seemingly redundant can be directly shown to occur at lower rate than expected using conditional Z-score statistics (see chapter 4), as well as using the LLR of the redundant and combined models. The cases we observed include the previously discussed Reb1 motifs [44] and separation among variants of the still cryptic PAC motif. PAC targets are highly enriched in stress response genes [18], but the mechanisms of PAC based regulation are not well characterized. We discovered two separated PAC-like families (GCGATGAG and GAGATGAG) that are significantly separated from each other. Interestingly, both variants of the PAC model tend to co-occur in the same promoters with the RRPE motifs (co-occurrence Z scores of 15.5 and 15.6), suggesting that they share a common mechanism rather than representing two distinct factors.

3.3 TFBS selection coefficients

By studying the inferred TF selection coefficients (σ parameters), we tried to characterize variability in the selection intensity among targets of specific TFs and between different lineages. We note that our model uses background substitution rates to model neutral evolution, and that we learned these rates separately for each pairwise alignment of two species. We can therefore attribute differences in σ values between lineages to changes in selective pressure or to other TF-specific effects (like divergence of the TF itself), rather than to different evolutionary distances or other background effects. This is demonstrated by comparing the σ values for the same target sets in three sets of pairwise alignments. According to the results (**Fig 8**) the σ values are similar between the different data sets (spearman correlation values ranging around 0.9), suggesting that, by and large, our model can separate background effects from selective pressures on TFBSs.

We observed significant variability in the inferred selection coefficient on known TF models for different TFs. Many of the well-known TFs with small target sets (smaller than 8) had strong selection (small σ) values, suggesting specific binding and tight selection. Some examples are Reb1 (0.18), Rpn4 (0.16), Ume6 (0.03) and Leu3 (0.12). However, for other well-known motifs we derived much higher σ values. These

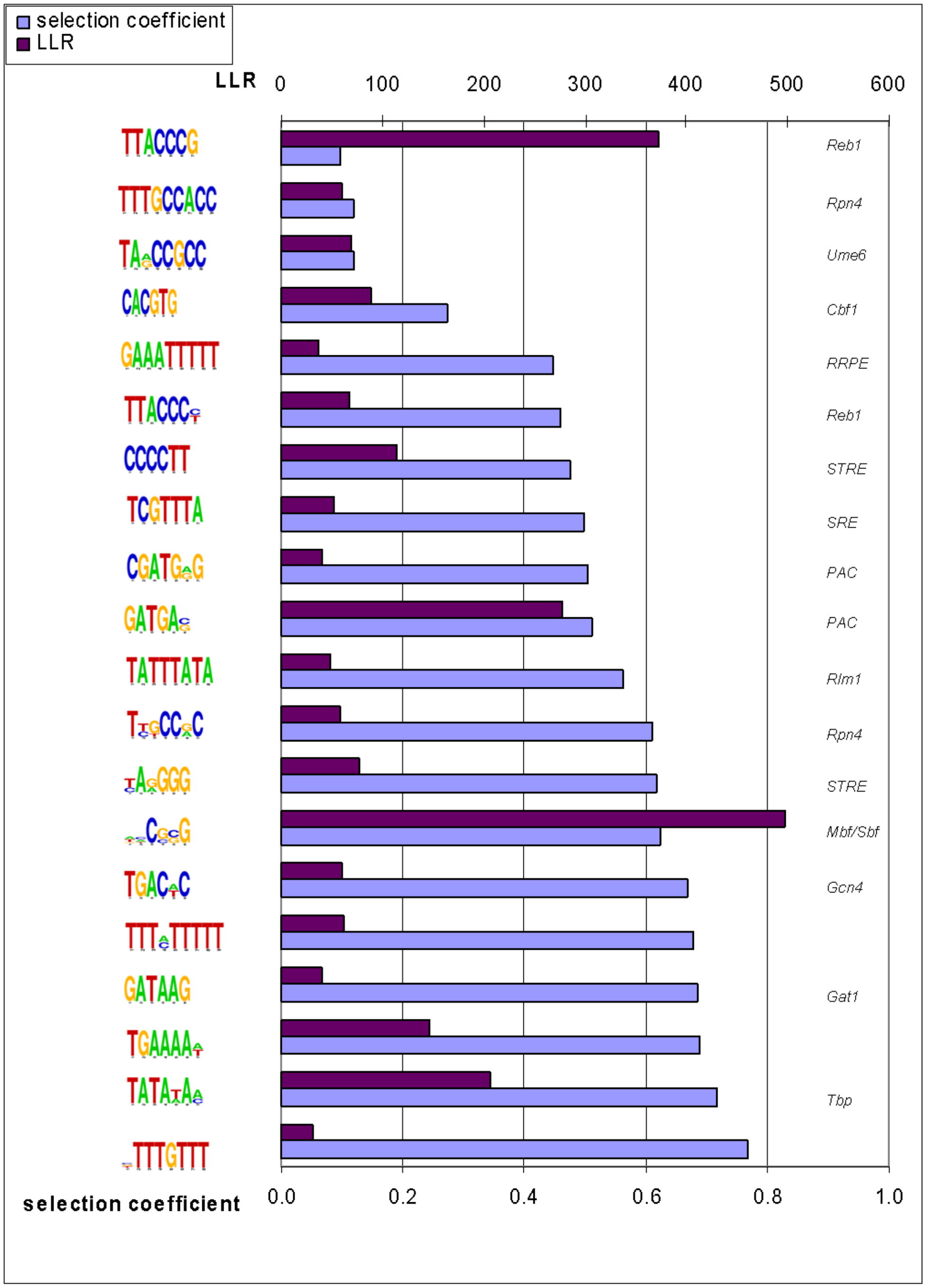
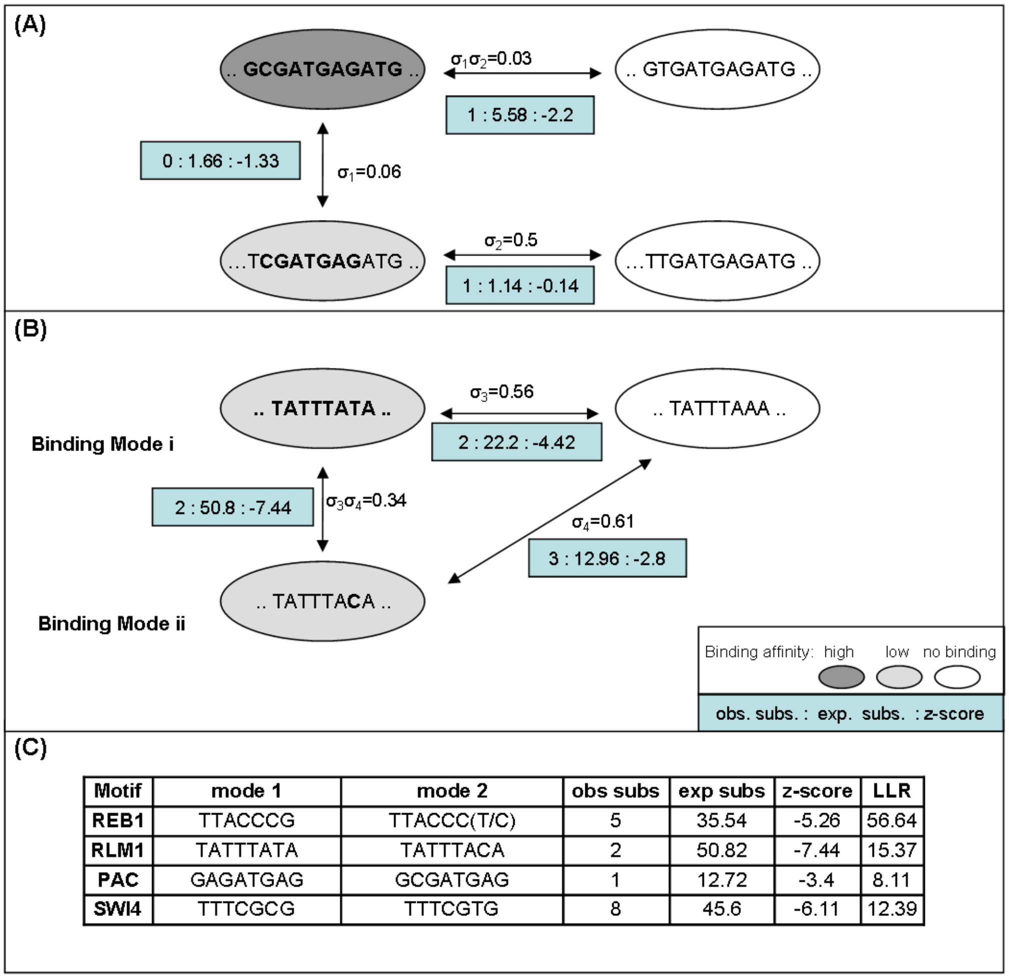


Figure 6: **De-novo model.** Shown are the log likelihood ratio (LLR) and selection coefficient values for the 20 target sets with the highest LLR values as discovered in the de-novo model. The full model is presented in the **Table 1**. For each target set we present its motif logo and, when the set matches a published binding site motif, its most likely TF. Target sets are sorted according to selection coefficient. While the target set with the most stringent selection coefficient, which matches the known Reb1 motif, has a very high LLR value, some other target sets that also match well-known motifs show lower LLR values. This is probably a result of a lower hit number (Rpn4, Ume6), weaker selection or lower sequence specificity (Cbf1).

	target set	z	hits	sigma	LLR	known motif match
1	AACGCG ; GACGCG ; TACGCG ; ACCGCG ; ATCGCG ; AACCCG ; AACGGG ; ACCGGG ; ACCCGG ; ACCCCG ; GCCGGG ; GCCGCG ; TACGGG ; TACCCG ; TCCGGG ; TCCGCG ; TTCGGG ; TTCGCG ; TTCCGG	38.72	2937	0.63	498.69	Cell-cycle box
2	TTACCCG	27.90	248	0.10	372.44	REB1
3	GATGAG ; GATGAC	22.23	984	0.51	277.40	PAC
4	TATATAA ; TATAAAA ; TATATAC	20.95	1950	0.72	205.91	TBP
5	TGAAAAA ; TGAAAAT	16.00	1553	0.69	145.73	
6	CCCCTT	11.70	317	0.48	114.44	STRE MSN2/4
7	CACGTG	17.77	204	0.27	89.15	MET28 TYE7 PHO4 CBF1
8	TAGGGG ; CAGGGG ; TAAGGG	16.02	605	0.62	77.81	MSN2 STRE
9	TAGCCGCC ; TAACCGCC	12.52	50	0.12	70.02	UME6
10	TTACCCT ; TTACCCC	12.59	193	0.46	67.60	REB1
11	TTTCTTTTT ; TTTATTTTT	9.98	546	0.68	62.65	
12	TTTGCCACC	13.66	36	0.12	59.82	RPN4
13	TGACTC ; TGACAC	9.46	537	0.67	59.58	GCN4
14	TTGCCGC ; TTGCCAC; TTTCCGC ; TCGCCGC ; TCGCCAC	13.03	406	0.61	58.42	RPN4
15	TCGTTTA	12.66	156	0.50	52.12	SRE
16	TATTTATA	11.72	222	0.56	48.78	RLM1
17	CGATGGG ; CGATGAG	20.80	268	0.50	40.53	PAC
18	GATAAG	10.27	462	0.68	39.79	GZF3 GATA-box GLN3 GAT1
19	GAAATTTTT	10.63	101	0.45	37.00	RRPE
20	GTTTGTTT ; CTTTGTTT ; TTTTGTTT	9.91	584	0.77	31.45	
21	GGCATCG ; CGCATCG	8.97	101	0.51	28.28	INO2/4
22	TGTGGC ; GGTGGC ; TGCGGC ; TGTCGC ; GGCGGC ; GGCCGC ; TGCCGC	15.05	1126	0.83	28.19	
23	GCGATGAGATG ; GCGATGAGATA	16.22	27	0.06	27.62	PAC
24	GTAAACAA ; GTAAACAG	8.09	136	0.58	24.88	SFF FKH1/2
25	ACACAAAA ; CCACAAAA	9.91	113	0.57	23.59	MSE
26	GTGTGG ; GTGAGG ; GTGGGG	8.53	500	0.78	23.36	
27	GGACCCT ; AGACCCT	7.68	107	0.54	23.25	NRG1
28	CCGAGA ; CCGACA ; CCGAGG ; CCGAGC ; CCGACG	9.54	550	0.76	23.20	
29	ACAATAG	6.30	250	0.70	23.06	Mat1-Mc ROX1
30	TTTGTGTC	8.52	45	0.39	23.00	MATa1/2 MCM1 MIG1 UASH TUP
31	GCGATGG	5.87	67	0.50	19.81	PAC
32	AAATCCG ; ACATCCG ; AACTCCG	9.34	178	0.68	18.14	
33	AGAGATGAG ; TGAGATGAG	17.17	94	0.50	15.78	PAC
34	TGGGTGC ; GGGGTGC ; TGGTTGC	6.14	116	0.65	15.67	AFT2
35	TATTTACA	7.81	110	0.61	15.39	RLM1
36	GTGGGGTA	9.42	11	0.10	15.35	TUP
37	TTTGTGAC	6.36	38	0.46	13.65	MATa1/2 MCM1 MIG1 UASH TUP
38	GAAATTTCA ; CAAATTTCA ; GATATTTCA ; GAAATCTCA ; GAAATTCCA ; GATATCTCA ; CATATTCCA	8.70	126	0.69	12.95	HSTF
39	CCGATA	6.82	165	0.67	12.72	HAP1
40	CGCAAAA; CACAAAA	8.03	359	0.79	11.28	MSE
41	GGCGCTA	6.61	48	0.53	11.01	SNT2
42	CACGCAA ; CACGTAA ; CACGCAG	6.37	156	0.74	9.95	
43	TTTCGTGT	9.15	50	0.57	9.65	SWI4 SCB
44	CGCGGGG	9.33	24	0.36	8.25	SUT1
45	AAGGGAG	6.07	103	0.71	8.24	
46	CGGAAAAA ; CGGAGAAA	7.32	116	0.71	7.83	RGT1 ECB
47	AAAGTGAAAAA	7.47	37	0.54	7.83	
48	TCGGCC ; TCGGGC ; TCGGCA	8.59	397	0.81	7.80	RDS1
49	AAATAGAG ; AAATAGGG	5.53	118	0.74	7.76	
50	ATGACT ; ATGGCT	7.15	728	0.88	7.73	ARG81
51	GCGCGGG	8.75	27	0.49	7.58	PDR3
52	AAAAGAAAC ; AAAACAAAC	9.70	137	0.77	7.51	STE11
53	GCCACCG	9.29	42	0.57	7.47	RPN4
54	AAGATCAA	6.19	73	0.72	7.27	
55	TTCTAGAA	6.45	80	0.62	6.81	HSTF HSF1
56	CCCCAA ; GCCCAA	6.43	354	0.84	6.50	
57	TTTCGCGA	6.51	43	0.49	6.35	SWI4
58	TTCTTCGA ; TTCTTCGG	5.66	100	0.75	6.21	STE11
59	TCAACAA	5.95	235	0.83	5.25	SFF
60	AAAGGGTT ; AAAGGGTA	6.90	105	0.77	5.09	MSN2/4
61	TGAAACA	6.67	211	0.82	4.49	STE12
62	CAGCAGC ; CAGCATC	5.17	161	0.84	4.18	ACE2 SWI5

Table 1: Model learned de-novo on *cerevisiae-mikatae* pairwise alignments



Motif	mode 1	mode 2	obs subs	exp subs	z-score	LLR
REB1	TTACCCG	TTACCC(T/C)	5	35.54	-5.26	56.64
RLM1	TATTTATA	TATTTACA	2	50.82	-7.44	15.37
PAC	GAGATGAG	GCGATGAG	1	12.72	-3.4	8.11
SWI4	TTTCGCG	TTTCGTG	8	45.6	-6.11	12.39

Figure 7: **Overlapping and bimodal target sets.** We demonstrate the effect of similar motifs in two types of redundancies, using examples from the de-novo model. In one type of redundancy **(A)**, two motifs are overlapping (here GCGATGAGATG and CGATGAG). This organization corresponds to one effective motif with two levels of affinity. The strong affinity BS appears as a hit for both motifs, and therefore the selection on it is the combined selection of both motifs. The weak affinity BS appears as a hit for just one motif (CGATGAG). A different type of redundancy **(B)** associates two similar motifs that differ in one base (here TATTTATA and TATTTACA), suggesting selection is preserving the separation between two variants of the same motif. We quantify the intensity of separation by comparing the likelihoods of models with merged and separated target sets, or by directly assessing the rate of substitutions between motifs in the two target set variants (see chapter 4). **(C) Inferred bimodal motifs.** Shown are cases of similar but separated target sets in the de-novo model, indicating the number of observed and expected substitutions between them, a conditional z-score (see chapter 4) and the LLR of the merged and separated models.

include the CACGTG motifs (Cbf1, Pho4, Tye7 and Met28) ($0.4 - 0.91$), Mbp1 (0.46), Swi6 (0.54) and Msn2/4 (0.54). Interestingly, we inferred high σ values (> 0.35) for these TFs in the ChIP model too. This suggests that the mild selection coefficients for these TFBSs are not primarily a side effect of low sequence specificity, since it is widely assumed that motifs in promoters that are also ChIP targets are very likely to be bound *in vivo*.

One possible explanation for the reduced selection on some of the target sets may be that k-mers from these sets tend to appear in multiple copies in each of the promoters they regulate. We examined the percentage of promoters in which these k-mers appear more than once (out of the promoters that have at least one hit). All the motifs mentioned above as having tight selection (low σ) appear only once in all of the promoters, while motifs with less tight selection are occasionally repeated in promoter regions (Swi6 is repeated in 11% of its promoters, Pho4 and Tye7 in 7%, Msn2 and Mbp1 in 5%, and Msn4 in 5%). While the number of cases is too limited to reach a clear statistical conclusion on the relation between redundancy and selection, it can be hypothesized that for many TFs, actual redundancy may be higher (including e.g., low specificity binding sites), and that such redundancy can alleviate some of the selective pressure on individual loci.

Another possible explanation to low selective pressure on the targets of some critical TFs may be that while some of the physical targets of these TFs are functionally essential and therefore under strong selection, other targets are evolutionarily transient and do not have a major functional role. This hypothesis should be further explored once more experimental data on TF binding for additional yeast species become available.

3.4 Quantifying selection on target sets

The regulatory code model implies three evolutionary regimes on motifs: k-mers can a) be functional sites (part of a target set), b) be one substitution away from becoming a functional site (boundary k-mers), or c) at distance of two substitutions or more from any target set, and thus behave in a neutral manner (background

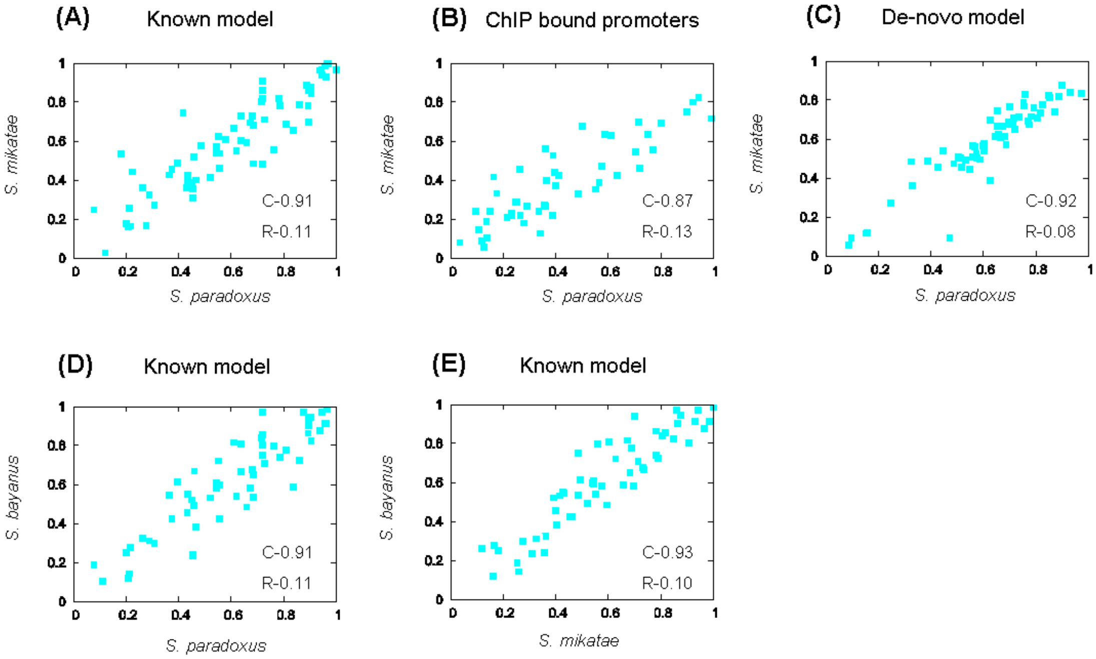


Figure 8: **Selection coefficients.** Shown are inferred selection coefficients, compared for pairs of similar target sets on alignment of different species. All alignments are with *S. cerevisiae*. C - the correlation between the σ values in the two models. R - the root mean square difference value. The high correlation among the inferred values indicates that our model successfully decomposes the background mutation rate (which is different for each species) and the selection on each TFBS (which is quite stable as seen here). **(A,D,E)** Comparison of values in the literature-based model for each pair of species. **(B)** The literature-based model using ChIP-bound promoters only on *cerevisiae-paradoxus* alignments compared to the same model on *cerevisiae-mikatae* alignments. **(C)** The de-novo model on *cerevisiae-paradoxus* alignments compared to the same model on *cerevisiae-mikatae* alignments.

k-mers) (**Fig 9A**). According to our basic assumptions, only substitutions between target set k-mers and boundary k-mers are subject to selection. Consequently, we predict functional sites to be highly conserved, and boundary k-mers to be slightly conserved - not due to functionality but due to selection against binding site emergence. As discussed above, in cases where our modeling assumptions are too restrictive, we may be classifying as boundaries those k-mers that are in fact weak binding sites. In these cases, we expect some selection to act on substitutions between such boundary k-mers and background k-mers.

To try and characterize the global effects of selection on boundary k-mers, we compared the degree of conservation of target set, boundary and neutral k-mers in the literature-based model. This was done by testing how often do motifs from each of these groups appear conserved, compared to what is expected given a neutral model. As shown in **Fig 9B**, the observed conservation of target set k-mers is far above what we expect from a neutral model. A weaker but still significant increase in conservation is observed for boundary k-mers, possibly due to weak selection on binding site appearance, or to mild selection on weak but functional sites. We next examined the substitutions between target set and boundary k-mers, and between boundary and background k-mers, using again the number of observed substitutions compared to the number expected by a neutral model. As shown in **Fig 9C**, substitutions between target-set and boundary k-mers are occurring much less than expected given the neutral model. We observe a slightly weaker, yet similar pattern for substitutions from boundary to background k-mers. At least some of the boundary k-mers in our model may therefore be functional and under some weak selection, forming together with a target set a TFBS recognition model that is more complex than our simple model.

3.5 The fraction of sequence under selection

Using our model, we can approximate the fraction of yeast promoter sequences that are under selection due to TFBSs. Previously, Chin and colleagues [10] addressed the more general question of overall selection intensity by computing the length distribution of runs of consecutive conserved bases and comparing it to

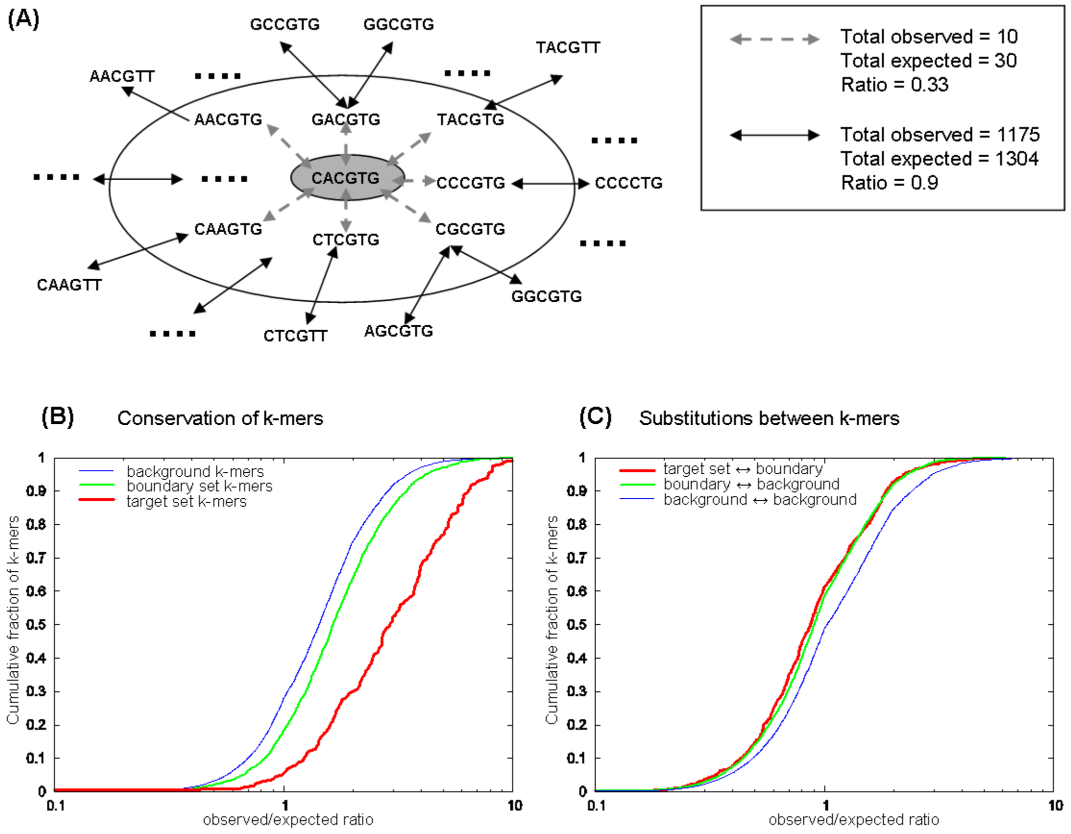


Figure 9: **Selection on target sets. (A) The selection on single nucleotide substitutions.** Our model predicts substitutions between target set k-mers (inner circle) and boundary k-mers (ring) to be under negative selection, and substitutions between boundary k-mers and background k-mers (outside the ring) or between two background k-mers not to be under selection. The right hand box shows observed and expected counts for the substitutions in one example (CACGTG in the cerevisiae-mikatae alignment), as well as the ratio between the observed and expected counts. **(B) Ratio of observed and expected conservation.** The plot shows the cumulative distribution of the ratios between the observed number of conserved motif appearances and the number expected under a neutral model. Shown are the distributions for three different sets of motifs: target set (red), boundary (green) and background (blue). A full concordance with the neutral model would have resulted in a perfect lognormal distribution. The background distribution is the closest to lognormal, but still shows bias toward increased conservation due to the clustering of mutations in yeast promoters (see **Appendix B**). Target set k-mers are conserved above what is expected. This is evident from looking at the observed-expected ratio distribution, and also when comparing it to the observed-expected ratio distribution for background k-mers (KS (Kolmogorov-Smirnov test) = 0.42 for difference from background k-mers, $P < 3.8e - 43$) and the same is true for boundary k-mers, but to a lesser degree (KS=0.09 for difference from background k-mers, $P < 3.4e - 36$). **(C) Ratio of observed and expected substitutions.** The plot shows the cumulative distribution of the ratios between the observed number of substitutions among motifs and the number expected under a neutral model. Shown are plots for substitutions between target set and boundary k-mers (red), between boundary and background k-mers (green), and between background k-mers (blue). Substitutions between target set and boundary k-mers appear less than expected. Again, this is evident when looking both at the ratio distribution (> 0.6 of the data points have ratio < 1) and at its difference from the distribution for substitutions between background k-mers (KS=0.14, $P < 2.3e - 13$; p-value depends on KS and sample size). Substitutions between boundary k-mers and background k-mers are also occurring less than expected, but to a somewhat lesser extent (KS=0.1, $P < 3.8e - 169$). As with the conservation data, the background distribution here is not lognormal, due to the non uniform distribution of mutations.

the distribution expected under the assumption of a neutral model. Their conclusion was that about 30% of the promoter sequence in *S. cerevisiae* is under selection (see **Appendix B** for discussion of this estimation). Using our model, we can employ a local per-site test and count the percentage of the sequence that is influenced by characterized TFBSs (at each site we can simply check whether it is a part of a target set k-mer or a boundary k-mer, and hence under selection). By taking this approach, we wished to quantify how much of the observed selection on yeast promoters can be explained using what we know today on transcriptional regulation via high specificity binding sites.

We started by looking at direct functional selection, i.e., the fraction of promoter regions that is covered by binding sites. Using the known model (See 3.1), 1.77% of the promoter sequence is functional. Using the de-novo model (See 3.2), the fraction is 2.36%. These models may be extremely incomplete, but even using the entire repertoire of binding sites in the MacIsaac study (without conservation or LLR constraints), the fraction is only 3.24%. It is therefore reasonable to conclude that only a small fraction of the promoter sequences is under tight selection for losing high specificity binding sites.

As we mentioned in the previous section, according to our model assumptions, weaker selection affects boundary k-mers in addition to active binding sites. By including boundary k-mers, the fraction of sequence under selection increases significantly, to 27.1% in the literature based model and 29.2% in the de-novo model. As discussed above, the selection on boundary k-mers is very weak and unlikely to fully bridge the gap between the global estimation of on the fraction of yeast promoters under selection (30%) [10] and the estimation of selection due to binding sites ($< 4\%$). Part of this gap is probably a result of additional selection forces acting on promoters, reflecting the role of the DNA sequence in determining chromatin structure [38, 7] and involving weak-affinity transcriptional interactions [43].

4 Methods

4.1 Evolution under the effect of a regulatory code

Define a *regulatory code* to be a collection of sets $C_1 \ldots C_m$ where each C_t is a set of words of length k_t. C_t is called the *target set* of the t-th TF. Typically C_t will consist of highly similar words. We define an indicator function $\beta_t(s, i)$ whose value is 1 if the i'th position in sequence s falls inside a word from target set C_t (or formally, if the substring $s[i-j, \ldots, i-j+k_t-1] \in C_t$ for some $0 \leq j < k_t$), and 0 otherwise. Every occurrence of a word from C_t in the promoter sequence s is declared a binding site of the t-th TF. Our model therefore assumes that TFs recognize all loci bearing words from their target sets, and no other loci. It also assumes that all words from the same target set behave identically, and that the target sets (and therefore the DNA binding domains of the TFs) remain constant along the evolution between the studied species. See **Fig 4A** for an illustration of a regulatory code. A word of length k_t that does not belong to any target set is called a *boundary k-mer* for TF t if it is one substitution away from some k-mer in C_t (See **Fig 4B**).

To describe the evolutionary process given the simple model of regulation defined above, we shall assume that the promoter sequence is evolving with a neutral substitution rate, with the exception that substitutions that change the regulatory role of a nucleotide - either eliminating or introducing a TFBS - are selected against. This proposed principle is formalized by means of rates in a continuous time Markov model. Note that each TFBS affects several adjacent nucleotides, and multiple TFs may interact with a single nucleotide. Consequently, the substitution rate at each nucleotide is affected by the identity of several neighboring positions during evolution (in other words, the nucleotides that form a TFBS are *epistatically interacting*). It is therefore impossible to represent the evolutionary process as the combination of independent Markov processes at individual nucleotides, and one must use a Markov model over a larger state space.

We model neutral evolution using a standard instantaneous nucleotide substitution rate matrix P_b, defining $P_b(c_1 \rightarrow c_2)$ as the neutral rate of substitution from nucleotide c_1 to c_2. The effect of selection on TFBSs of the t-th TF is formalized using a *selection coefficient* $0 \leq \sigma_t \leq 1$. We assume that a substitution with neutral rate p has a reduced rate $\sigma_t p$ whenever it adds or removes a binding site for the TF t. The evolution of an entire promoter sequence (of length n) is modeled using a large continuous time Markov model with a state space that includes all sequences of length n. The rate of mutation at the i-th position of a sequence s is defined by:

$$a_{s,s'} = P_b(s[i] \rightarrow s'[i]) \prod_t \sigma_t^{|\beta_t(s,i) - \beta_t(s',i)|} \tag{1}$$

where s' is equal to s in all positions but i. A stochastic rate matrix A is defined using **Equation 1** on entries reflecting substitutions at one nucleotide and is 0 on entries reflecting a change in more than one nucleotide. The diagonal is defined by complementing row sums to 0. Based on the general theory of Markov models [23], the probability of arriving from arbitrary s_1 to s_2 in time t is obtained by exponentiation of the rate matrix A: $Pr(s_1 \rightarrow s_2, t) = exp(tA)_{s_1,s_2}$. Note that while the instantaneous rate of a direct transition from s_1 to s_2 may be zero, the exponentiation of the matrix is equivalent to allowing all paths of total time t between s_1 and s_2, through any set of intermediate sequences.

Since the state (sequence) space defined above is of size $N = 4^n$, any direct computation with the complete model is not practical. In particular, computing the likelihood of a model, which amounts to exponentiation of the rate matrix, must be approximated aggressively as discussed below. Still, the model represents what we believe to be a reasonable formalization of the evolutionary process, and is useful as a principled basis for approximations.

Generalizing the model to phylogenetic trees is straightforward, using the Markov process defined above to represent evolution on each branch. One should note that ancestral inference in our model is non-trivial (as it is for any context-aware model [22]), and consequently likelihood computation given missing data should be addressed carefully. Another easy extension to our model is to include additional context dependence in the neutral substitution rates [40], representing possible variations that are not a result of selection on TFBSs (e.g. nearest-neighbor effects).

4.2 The parsimonious Markov model

In order to reduce the complexity of computing the likelihood of a regulatory model given alignments, we shall use an approximate calculation. Instead of considering the space of all possible sequences of length n, we shall consider only the most parsimonious evolutionary paths between two extant sequences. By considering only parsimonious paths, we assume that conserved parts of the sequence were fixed over the entire evolutionary period under study. We can therefore confine the effect of selection on TFBSs to those parts of the sequence that include a target set or a boundary k-mer in the extant species. We ignore the evolutionary scenarios in which TFBSs affected other loci transiently, thereby simplifying computation drastically. Formally, we shall approximate the large Markov model using several independent processes over small parts of the sequence that contain overlapping binding sites (called below *(approximated) epistatic block*).

Given two aligned sequences s_1, s_2, we define the set S_{s_1,s_2} as the collection of sequences $\hat{s}$ such that for all positions i either $\hat{s}[i] = s_1[i]$ or $\hat{s}[i] = s_2[i]$. Sequences in S_{s_1,s_2} are called *parsimonious w.r.t* s_1, s_2. The parsimonious Markov model is the restriction of the original Markov model to the states in S_{s_1,s_2}, using the same transition probabilities, and an additional state (denoted OUT), that absorbs all transitions from states in S_{s_1,s_2} into states not in S_{s_1,s_2}, and has a self loop with transition probability of 1. We denote the rate matrix for the new model as $A^{(s_1,s_2)}$ and approximate the exact transition probabilities $P(s_1 \rightarrow s_2, t)$ by $P^{(s_1,s_2)}(s_1 \rightarrow s_2, t) = exp(tA^{(s_1,s_2)})_{s_1,s_2}$.

Given a regulatory code, we say that two positions i, j are *parsimoniously epistatic* if there exists a state $s' \in S_{s_1,s_2}$ and a TF t such that $s'[i]$ and $s'[j]$ are part of the same appearance of a k-mer from C_t, or a boundary k-mer for TF t. An *epistatic block* is defined to be a maximum interval in the alignment in which every two adjacent positions are parsimoniously epistatic. The simplest epistatic block is a single neutral nucleotide, which does not interact with any TF in the extant sequences or in any parsimonious trajectory between them. The next basic case is that of an interval including exactly one TFBS. Compare **Fig 4C**. In general, when there are several sites overlapping each other, the epistatic blocks define the smallest possible units for which we can compute the model likelihood independently.

Assume the epistatic blocks for the aligned sequences s_1, s_2 are $\{(a_1, b_1), (a_2, b_2), \ldots, (a_l, b_l)\}$, where a_i is the start index of the i-th block, and b_i is the corresponding end index. Define $s_1^j = s_1[a_j...b_j], s_2^j = s_2[a_j...b_j]$ $(1 \leq j \leq l)$. Then the probability $P^{(s_1,s_2)}(s_1 \rightarrow s_2, t)$ equals $\prod_j P^{(s_1^j,s_2^j)}(s_1^j \rightarrow s_2^j, t)$. Note that $P^{(s_1,s_2)}$ equals the integral of probabilities over all paths from s_1 to s_2 in the parsimonious model. The key observation is that the context-dependent transition rate of any substitution in a position i is affected only by the current state of nucleotides that are in the same epistatic block as i, and independent of the current state of any other nucleotide. We can therefore decompose any path from s_1 to s_2 into independent sub-paths in each epistatic block and compute the path probability as a product of independent contributions from the parsimonious models built on the sequences of each epistatic block.

Whenever the density of binding sites is not too extensive, one gets relatively small epistatic blocks. The probability for each such block can then be solved using exact exponentiation of the transition matrix or by any of the many available heuristics for continuous time Markov models. The approach we used in this work is to transform each continuous time model into a discrete time model on m time steps (m being large enough such that in each evolutionary time-step, at most one substitution occurs). Using this approach, we can approximate $P^{(s_1^j,s_2^j)}$ in each epistatic block using dynamic programming, spending $O(lm)$ time on a block with l parsimonious states.

In summary, to compute the likelihood of a model (regulatory code and selection coefficients) given a set of pairwise alignments we work in three phases (see **Fig 4D**). First we partition the alignment into epistatic blocks by searching for target set or boundary k-mers in the aligned sequences and their parsimonious combinations. We then compute the log likelihood of each block using the discrete time approximation to the parsimonious Markov model, and sum the contributions. Note that in a typical scenario, a substantial fraction of the sequence is neutral with respect to the model, which translates to epistatic blocks of length one (a single nucleotide). When computing the log likelihood ratio of some model vs. the null model, we can ignore all of these blocks.

One can estimate the error imposed by the parsimonious approximation, caused by losing probability absorbed into the OUT state. Unfortunately, the error increases linearly with the length of the sequence, yielding values very close to 1 even for relatively short sequences. Note however, that we apply the model to infer selection coefficient and learn regulatory models, so for our purposes, the parsimonious approximation can serve as an effective heuristic as long as it does not change the ranking of models likelihoods. Our empirical analysis suggests this is the case with the yeast data we analyzed here, but a more rigorous mathematical treatment of this problem is desirable.

4.3 Model Learning

We now turn to the problem of learning a maximum likelihood model given a set of pairwise alignments. Formally, given a set of pairwise alignments and assuming a background neutral substitution model P_b (which we compute directly from the alignments), we wish to find target sets $C_1 ... C_m$ and their selection coefficients $\sigma_1 ... \sigma_m$ such that the model likelihood (under the parsimonious model, and using the heuristic likelihood computation) is optimal.

Assume first that we are given the regulatory code $C_1 ... C_m$ and a set of pairwise alignments, and we wish to infer the maximum likelihood parameters $\sigma_1 ... \sigma_m$. The problem is difficult to solve analytically, given the different effects each σ_t has on each epistatic block (e.g., when there are overlapping binding sites). We therefore work in iterations, focusing each time on a single TF t and fixing σ_i for $i \neq t$. Optimizing σ for a single TF is still not completely straightforward. We currently use a standard binary-search algorithm to find σ_t that gives a maximum likelihood for the model. Empirically, this approach performs well since the optimization function appears to be nearly convex.

Learning the regulatory code from scratch is a more difficult problem, involving a vast solution space and relatively few restrictions on the number of TFs, the size of their target sets and the relations among them. We approach this problem heuristically, working in iterations and successively adding new target sets or updating the structure of existing ones. Since the number of possible target sets is extremely large, and since likelihood calculation and σ optimization are computationally demanding, we cannot search all possible changes per iteration. Instead, we use efficient pre-processing of all candidate k-mers (limiting k to a reasonable range) followed by a greedy search in the model space, guided by the results of the pre-processing.

The learning algorithm is organized as follows:

1. Compute statistics on all k-mers and obtain a small set of promising candidates.

2. Sort candidates by approximating their potential contribution to the model likelihood.

3. Iteratively try to add candidates to the model according to their approximated contribution. In each iteration, a new singleton target set containing the candidate is added to the model, and then expanded with highly similar k-mers. An addition to the model is accepted if and only if it increases the model likelihood.

We define n_m to be the number of times k-mer m appears in the promoter sequences of the first species, in the entire data. $n_{m,m'}$ is the number of times m appears in the first species aligned against m' in the second species. $n^p_{m,m'}$ is the number of times we expect m' to appear in the second species aligned against m in the first species, according to the neutral model. It is calculated as: $n^p_{m,m'} = n_m P_b(m \to m')$. Here, $P_M(m \to m')$ denotes the probability that k-mer m will be aligned against m' according to the model M (where b is the neutral model).

In the first stage we screen for conserved k-mers (with k ranging over some pre-selected range). Assuming $n_{m,m}$ is a binomial variable with success probability $P_b(m \to m)$, we compute a p-value for the conservation of each k-mer, conservatively correcting for multiple testing using the Bonferroni approach. k-mers with p-value ≤ 0.005 are considered as candidates in the subsequent stages. Additionally, we filter out candidates with less than a minimal number of hits (10 in this work). We call the set of remaining candidates F. For the yeast data, F contains a few hundred candidates.

In the second stage, we approximate the likelihood contribution of adding a k-mer to the model by roughly estimating the selection coefficient σ for it, and approximating the likelihood of the current model extended with the additional k-mer. We use M to denote the current model, and M' for the extended model. For a target set $C_t = \{m\}$, we set $\sigma_t^{appx} = \frac{\sum_{m' \neq m} n_{m,m'}}{\sum_{m' \neq m} n^p_{m,m'}}$. This approximation assumes that in the course of evolution, $\sum_{m' \neq m} n^p_{m,m'}$ appearances of m were affected by a substitution, but only $\sum_{m' \neq m} n_{m,m'}$ of them were fixated. The proportion of fixated mutations is an approximation to the selection coefficient. In reality, the situation is much more complex due to the effect of multiple substitutions, overlapping binding sites and selection against the emergence of TFBSs. The approximate likelihood ratio of extending the model with k-mer m is then calculated as: $\frac{\prod_{m'} P_{M'}(m \rightarrow m')^{n_{m,m'}}}{\prod_{m'} P_M(m \rightarrow m')^{n_{m,m'}}}$. The approximated likelihood contribution is computed as if all appearances of m form independent epistatic blocks and ignores the effects of the new model on existing motifs with occurrences overlapping m. Under this assumption, there is a small set of possible epistatic blocks, so the LLR can be quickly calculated. After calculating the estimated contribution for each $m \in F$, we sort F according to this score, and use the ordered list in the next stage.

In the third stage, we iteratively try to add and extend target sets, using the order computed in the previous stage. For each candidate k-mer m, we start by adding m as a new singleton target set in the model. Then, we perform a local search to add k-mers to the new target set. First we try to add k-mers that are one substitution away from m. An addition is accepted only if its LLR is above a pre-specified threshold. Then, we continue to expand with k-mers that are any possible combination of existing k-mers in the target set. Following the addition and expansion of a new target set, we return to Stage 2 and re-calculate the approximated contribution of each candidate. We continue iterating Stages 2 and 3 until no more candidates are left.

4.4 Data

We downloaded multiple alignments of *S. cerevisiae* with six other *Saccharomyces* species from the UCSC Genome Browser (version 1, Oct. 2003) [24]. Using gene annotations, we extracted pairwise alignments of *S. cerevisiae* promoters with each of *S. paradoxus*, *S. mikatae* and *S. bayanus*, using up to 300bp upstream of the TSS. Additionally, we extracted similar alignments with 100, 200, and so on up to 1000 bp upstream of the TSS, and with 300bp upstream and 100bp downstream of the TFF.

4.5 Computing expected values and z-scores for conservation and substitutions

The observed number of conserved appearances for each k-mer was derived directly from the pairwise alignment. The expected number (assuming neutrality) was calculated by multiplying the conservation probability, computed using the neutral model, by the total number of appearances in the first species. The observed and expected numbers for substitutions between k-mers were computed in the same way. The conditional z-score for negative selection between two target sets was computed as follows: As before, we obtained the observed number of substitutions between the two target sets from the pairwise alignment. However, the expected number of substitutions was then computed as follows: We calculated the posterior probability of the substitution given the conserved bases (e.g., for TTACCCG and TTACCCT we compute $P(TTACCCG \to TTACCCT | TTACCC \to TTACCC) = P(G \to T)$), and multiplied by the number of hits. These numbers were used to compute a conditional z-score.

5 Discussion

In this study we introduced a probabilistic model for the evolution of promoter regions that takes into account the combined effects of multiple TFs and their interactions. We developed an algorithm for calculating the likelihood of a model given pairwise alignments of promoters of orthologous genes from two species. Additionally, we developed algorithms for learning maximum likelihood model parameters. We applied our algorithms to *Saccharomyces* promoter regions, inferring a model using known TFBS sequences, and learning a full model from scratch. We analyzed the patterns of selection on promoter regions revealed by these models. Specifically, we used our models to study the intensity of selection on TFBSs and to estimate the amount of promoter region under selection due to high specificity TFBSs in these species.

Given our results, and based on evolutionary interpretation of all known transcriptional interactions in yeast, it is evident that even on very short evolutionary time scales, transcriptional regulation is highly dynamic. Indeed, the selection coefficients we computed for almost all TFBSs are higher (less tight) than what we might expect from functionally essential loci (averaging around 0.5). Selection seems to be weak, even if we restrict the analysis to functionally validated sites (ChIP targets). On the other hand, we observed a significant gap between the amount of selection we can account for using characterized TFBSs and the overall selection on yeast promoters. Taken together, it can be hypothesized that much of the functionality of transcriptional networks is encoded in ways other than strong TFBSs, and that due to high levels of redundancy, binding sites are under continuous remodeling [34]. Rather than being a deterministic and sparse network, transcriptional programs may be shaped as dense, noisy networks which are ever changing during evolution.

Much of the past research on comparative methods for non-coding regions has focused on the evolutionary dynamics of TFBSs, as they have relatively well defined features and a clear functional role. In addition to conservation-based methods for identifying TFBSs [14, 26, 50, 13, 6], several studies introduced methods for detecting TFBS motifs using phylogeny-based probabilistic models that distinguish between the evolution of TFBSs and of the neutral background [41, 39]. Other studies associated the evolutionary rate with the physical strength of TF-DNA interactions [44, 33, 27]. These studies strongly motivated the development of a general model for the evolution of regulatory regions in the presence of TFBSs.

The more general approaches for context-aware molecular evolution were so far limited to modeling of neutral evolutionary processes [37, 40, 5], or tailored to rigidly structured protein coding regions [36] or RNA coding genes [51]. The model we develop here is a small step toward overcoming the major computational difficulties in handling the evolution of large regions with heterogeneous function (many binding sites, sparsely and non-uniformly arranged). To make the model more realistic, additional effects will have to be considered, including binding sites with variable affinities, chromatin structure effects, combinatorial regulation and more. Computationally, the adaptation of our methods for computing likelihood and learning models to general phylogenies will require solution of the ancestral inference problem in our model. This problem, which is complex for any context aware model [22] is still an open challenge. We hope that continuous work on these challenges will open the way to faithful modeling of regulatory evolution in higher eukaryotes.

References

[1] Saccharomyces genome database, http://www.yeastgenome.org/.

[2] *Biochemical nomenclature and related documents*. the Biochemical Society, London, 1978.

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

[4] Pedersen AK., Wiuf C., and Christiansen FB. A codon-based model designed to describe lentiviral evolution. *Molecular Biology and Evolution*, 15(8):1069–81, 1998.

[5] P. F. Arndt and T. Burge, C. B. Hwa. DNA sequence evolution with neighbor-dependent mutation. *Journal of Computational Biology*, 10:313–322, 2003.

[6] D. Boffelli, J. McAuliffe, D. Ovcharenko, K. D. Lewis, I. Ovcharenko, L. Pachter, and E. M. Rubin. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299(5611):1391–4, 2003.

[7] M. J. Buck and J. D. Lieb. A chromatin-mediated mechanism for specification of conditional transcription factor targets. *Nat Genet.*, 38(12):1446–51, 2006.

[8] M. L. Bulyk. Computational prediction of transcription-factor binding site locations. *Genome Biology*, 5(1), 2003.

[9] M. L. Bulyk, A. M. McGuire, N. Masuda, and G. M. Church. A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in escherichia coli. *Genome Research*, 14:201–208, 2004.

[10] C. S. Chin, J. H. Chuang, and H. Li. Genome-wide regulatory complexity in yeast promoters: Separation of functionally conserved and neutral sequence. *Genome Research*, 15:205–213, 2005.

[11] S. S. Choi, W. Li, and B. T. Lahn. Robust signals of coevolution of interacting residues in mammalian proteomes identified by phylogeny-aided structural analysis. *Nature Genetics*, 37(12):1367–71, 2005.

[12] N. L. Clark and W. J. Swanson. Pervasive adaptive evolution in primate seminal proteins. *PLOS Genetics*, 1(3), 2005.

[13] P. Cliften, P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, J. Majors, R. Waterston, B. Cohen, and M. Johnston. Finding functional features in saccharomyces genomes by phylogenetic footprinting. *Science*, 301(5629):71–76, 2003.

[14] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(5):520–562, 2002.

[15] Tagle DA., Koop BF., Goodman M., Slightom JL., Hess DL., and Jones RT. Embryonic epsilon and gamma globin genes of a prosimian primate (galago crassicaudatus). nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *Journal of Molecular Biology*, 203(2):439–455, 1998.

[16] Tompa M. et. al. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol.*, 23(1):137–144, 2005.

[17] Pavesi G., Mauri G., and Pesole G. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, 17(Suppl 1):S207–14, 2001.

[18] A. P. Gasch, P. T. Spellman, C. M. Kao, O. Carmel-Harel, M. B. Eisen, G. Storz, D. Botstein, and P. O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, 11:4241–4257, 2000.

[19] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J. B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, 2004.

[20] Felsenstein J. and Churchill GA. A hidden markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, 13(1):93–104, 1996.

[21] Jensen JL. and Pedersen AK. Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Adv. in Appl. Probab.*, 32(2):499517, 2000.

[22] V. Jojic, N. Jojic, C. Meek, D. Geiger, A. Siepel, D. Haussler, and D. Heckerman. Efficient approximations for learning phylogenetic hmm models from data. *Bioinformatics*, 20(S1):I161–I168, 2004.

[23] S. Karlin and H. M. Taylor. *A second course in stochastic processes*. Academic press, 1981.

[24] D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, and W. J. Kent. The UCSC genome browser database. *Nucl. Acids Res*, 31(1):51–54, 2003.

[25] MacIsaac KD. and Fraenkel E. Practical strategies for discovering regulatory DNA sequence motifs. *PLOS Computational Biology*, 2(4), 2006.

[26] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E. Lander. Sequencing and comparison of yeast species to identify genes and regulatory motifs. *Nature*, 423:241–254, 2003.

[27] M. Lassig and V. Mustonen. Evolutionary population genetics of promoters: Predicting binding sites and functional phylogenies. *PNAS*, 102(44):15936–15941, 2005.

[28] Brenowitz M., Senear DF., Shea MA., and Ackers GK. Quantitative DNase footprint titration: a method for studying protein-DNA interactions. *Methods Enzymol.*, 130:132–81, 1986.

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

[30] Kimura M. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980.

[31] Schoniger M. and von Haeseler A. A stochastic model for the evolution of autocorrelated DNA sequences. *Mol Phylogenet Evol.*, 3(3):240–247, 1994.

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

[33] A. M. Moses, D. Y. Chiang, M. Kellis, E. S. Lander, and M. B. Eisen. Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evolutionary Biology*, 3(19), 2003.

[34] A. M. Moses, D. A. Pollard, D. A. Nix, V. N. Iyer, X. Y. Li, M. D. Biggin, and M. B. Eisen. Large-scale turnover of functional transcription factor binding sites in drosophila. *PLoS computational biology*, 10, 2006.

[35] Goldman N and Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution*, 11(5):725–36, 1994.

[36] A. K. Pedersen and J. L. Jensen. A dependent-rates model and an mcmc-based methodology for the maximum-likelihood analysis of sequences with overlapping reading frames. *Molecular Biology and Evolution*, 18:763–776, 2001.

[37] E. Schadt and K. Lange. Codon and rate variation models in molecular phylogeny. *Mol Biol Evol.*, 19(9):1534–49, 2002.

[38] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thastrom, Y. Field, I.K. Moore, J. Z. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 442:772–778, 2006.

[39] R. Siddharthan, E. D. Siggia, and E. van Nimwegen. PhyloGibbs: A gibbs sampling motif finder that incorporates phylogeny. *PLOS Computational Biology*, 1(7), 2005.

[40] A. Siepel and D. Haussler. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol*, 21:468–488, 2004.

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

[42] Muse SV. and Gaut BS. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution*, 11(5):715–724, 1994.

[43] A. Tanay. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Research*, 16(8):962–72, 2006.

[44] A. Tanay, I. Gat-Viks, and R. Shamir. A global view of the selection forces in the evolution of yeast cis-regulation. *Genome Research*, 14:829–834, 2004.

[45] Jukes TH. and Cantor CR. Evolution of protein molecules. *Mammalian protein metabolism*, 3:21–132, 1969.

[46] Bailey TL. and Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 28–36, 1994.

[47] R. Varshavsky, M. Linial, and D. Horn. COMPACT: A comparative package for clustering assessment. *Lecture Notes in Computer Science*, 3759:159–167, 2005.

[48] K. L. Wang and J. R. Warner. Positive and negative autoregulation of reb1 transcription in *saccharomyces cerevisiae*. *Mol. Cell. Biol.*, 18:4368–4376, 1998.

[49] E. Wingender, P. Dietze, H. Karas, and R. Knuppel. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Research*, 24(1):238–241, 1999.

[50] X. Xie, J. Lu, E. J. Kulbokas, T. Golub, V. Mootha, K. Lindblad-Toh, E. Lander, and M. Kellis. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–45, 2005.

[51] J. Yu and J. L. Thorne. Dependence among sites in RNA evolution. *Mol. Biol. Evol.*, 23(8):1525–1537, 2006.

[52] J. Zhu and M. Q. Zhang. SCPD: a promoter database of the yeast saccharomyces cerevisiae. *Bioinformatics*, 15(7):607–611, 1999.

A Appendix A: Simulations

A.1 The setup

We used simulations, varying model and data parameters, in order to characterize the problem's difficulty. The parameters that were explored were: (1) the size of the input (total length of all promoter sequences), (2) the neutral substitution rate, (3) the number of hits for each target set, (4) the length of the target set k-mers, (5) the size of each target set, and (6) the negative selection coefficient. To avoid assumptions on the steady state sequence distribution, we used actual *S. cerevisiae* promoters for the first species sequence set (s_1) and for the generation of the model as follows: Singleton target sets were generated by selecting a k-mer that appeared in s_1 with the required frequency. Target sets containing several k-mers were generated by selecting randomly a k-mer along with several of its neighbors (i.e. k-mers that differ from it by one substitution) that together have the required frequency. Given s_1 and the model we generated, we simulated the aligned sequences s_2 according to the model, by substituting each nucleotide independently with the neutral substitution probability, but rejecting a substitution with probability $1-\sigma_t$ is if disrupts or introduces a binding site for the TF t. The default parameters used for the simulation were 2M aligned bases, with a background mutation rate of 0.3, in which there are 250 hits of one target set comprised of 3 octamers, with selection coefficient 0.15.

A.2 Performance evaluation

To quantify the success of the model learning algorithm, we used three scores. The first two measure the accuracy of discovering the correct k-mers, regardless of their partition into target sets. Denote the correct solution by $A = \{A_1, ..., A_m\}$ where each A_i is a target set. Similarly, denote the learned solution by $B = \{B_i\}$. We define the *specificity* of B as $\frac{|(\cup A_i)\cap(\cup B_i)|}{|(\cup B_i)|}$ and the *sensitivity* as $\frac{|(\cup A_i)\cap(\cup B_i)|}{|(\cup A_i)|}$. The *Jaccard* score refers to the learned model as a clustering solution where each learned target set is a cluster, and applies a standard figure of merit used in cluster analysis, quantifying the extent to which pairs of k-mers are in the same cluster in the correct solution are also in the same cluster in the learned solution and vice versa (for the definition, see, e.g., [47]).

A.3 Experiments

We first examined a simple scenario, varying one parameter at a time. The results are shown in **Fig 10**, and represent an average of 50 simulations at each point. For the default parameters the sensitivity score is higher than the specificity score, indicating a higher susceptibility to false positives than to false negatives. However, for more extreme parameter values, the sensitivity score is usually affected more than the specificity. The effect of the different parameter values we examined is as expected for most parameters, e.g., improving results with more hits per target set, and with a smaller σ. The data size (ranging from 1Mb to 10Mb) and the motif length (ranging from 7 to 9) were found to have very little effect on the performance (results not shown).

Since we expect that many real TFs in yeast have less than 100 hits, we performed additional simulations to test the interplay between the number of hits and the selection coefficient. As seen in **Fig 11**, as the selection coefficient gets closer to zero, a smaller number of hits is sufficient.

We also tested two additional scenarios, which we expect to be particularly complex. In the first, two or more target sets are very similar to each other (k-mers in one set one substitution away from k-mers in the other set). In the second, there are two target sets whose k-mers are partially overlapping. These scenarios are known to occur in real biological data. Since most extant motif-finding algorithms cannot accurately handle such cases, we wanted to evaluate how well our algorithm deals with them.

In the first complex scenario, a pair of close target sets was generated by simulating a large target set and randomly partitioning it into two. When more than two close target sets are needed, we partitioned the large set into more groups. In reality, we would expect k-mers in the same target set to be more similar to the k-mers in the same target set than in other, similar target sets. The scenario we test here is therefore more stringent than what we expect to see in real biological data.

Fig 12 shows the results for the close target sets .These should be compared with results for the same number of arbitrary target sets (**Fig 10**). Overall the same trends are observed, but the scores are lower. As the number of target sets increases, the Jaccard score and the sensitivity decrease, while the specificity remains high. The Jaccard score deteriorates most when the target sets are close.

In the second complex scenario, there are two target sets with partially overlapping k-mers, and therefore some overlapping hits. We examined different fractions of overlapping hits, as well as different numbers of overlapping nucleotides per overlapping hit. Given the complex nature of the constraints, we could not use the simulation method described above to generate target sets, but rather planted in the original promoters hits of randomly generated target sets with the desired properties. Overlapping target sets were simulated by randomly generating the first target set, and then creating a second random target set, with the condition that its k-mers have the desired overlap with the first target set. The desired number of overlapping and non-overlapping k-mers are then planted in the data, taking into account appearances of the k-mers that existed in the original data. Performance results are shown in **Fig 13**. The size of the overlap was not found to have a significant effect. Overall, the effect of the overlapping motifs scenario on accuracy was more moderate than that of close target sets.

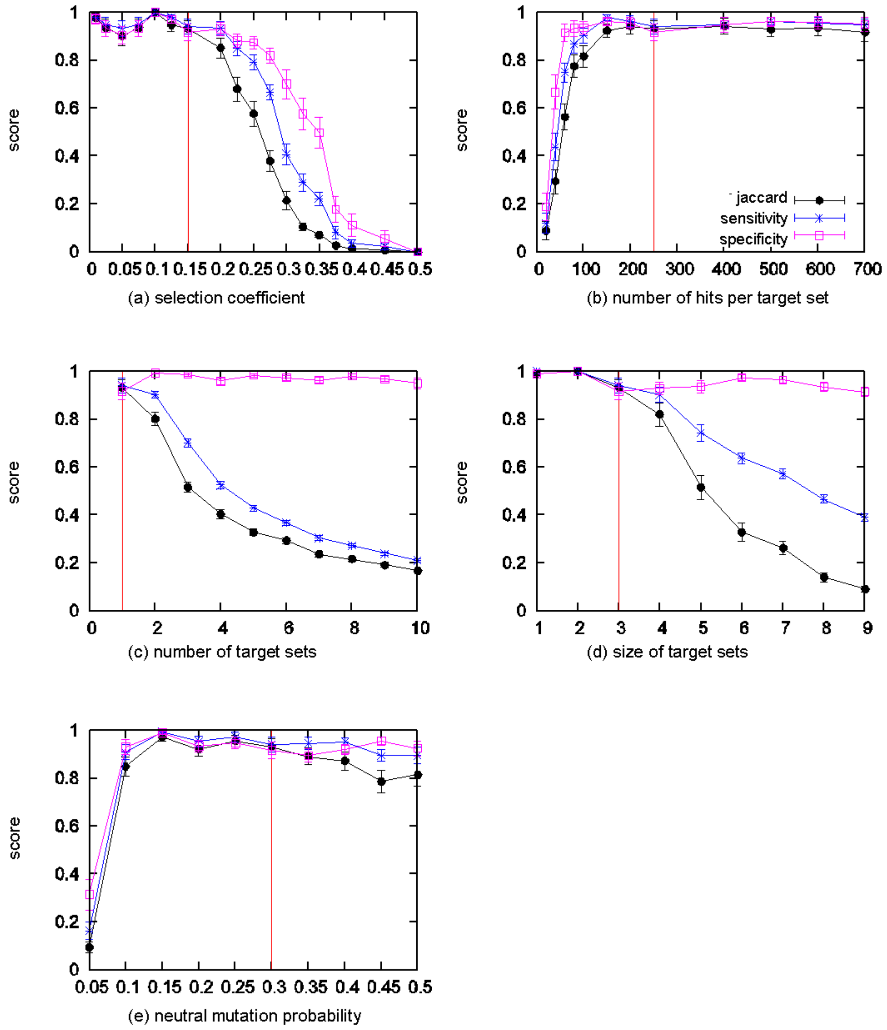


Figure 10: **Simulation results - Effect of parameters on model learning** Values are average scores (sensitivity, specificity and Jaccard) in 50 simulation trials. Error bars indicate 1 standard error of the mean. The vertical lines indicate the default value of the parameter. Default parameters were chosen to mimic the attributes of the yeast data. For the default parameters the average Jaccard score was 0.93, average sensitivity 0.94 and average specificity was 0.91. The specificity is lower than the sensitivity, suggesting a slightly higher tendency for false positives than false negatives. **(a) Effect of the selection coefficient:** As expected, performance improves with the strength of the negative selection. A sharp drop in the performance occurs above $\sigma = 0.25$. **(b) Effect of the number of hits per target set:** As expected, a larger hit number improves the performance. A reasonable accuracy is achieved around 100 hits, a fairly high hit-number. Above 150 hits increasing the hit number has a minor effect. **(c) Effect of the number of target sets:** While the parameter has almost no effect on specificity, increasing the number of target sets causes a significant decline in the sensitivity score (more false negatives ”missed” motifs) and the Jaccard score (more ”grouping” errors). **(d) Effect of target set size:** As expected, increasing the target set size leads to a decrease in all scores, with a stronger effect on the Jaccard score. **(e) Effect of neutral substitution rate:** For very conserved data (under 0.1 probability of substitutions), the score averages are quite low, since most of the sequence is conserved and signal detection is difficult. For very divergent data, there is a slight drop in performance.

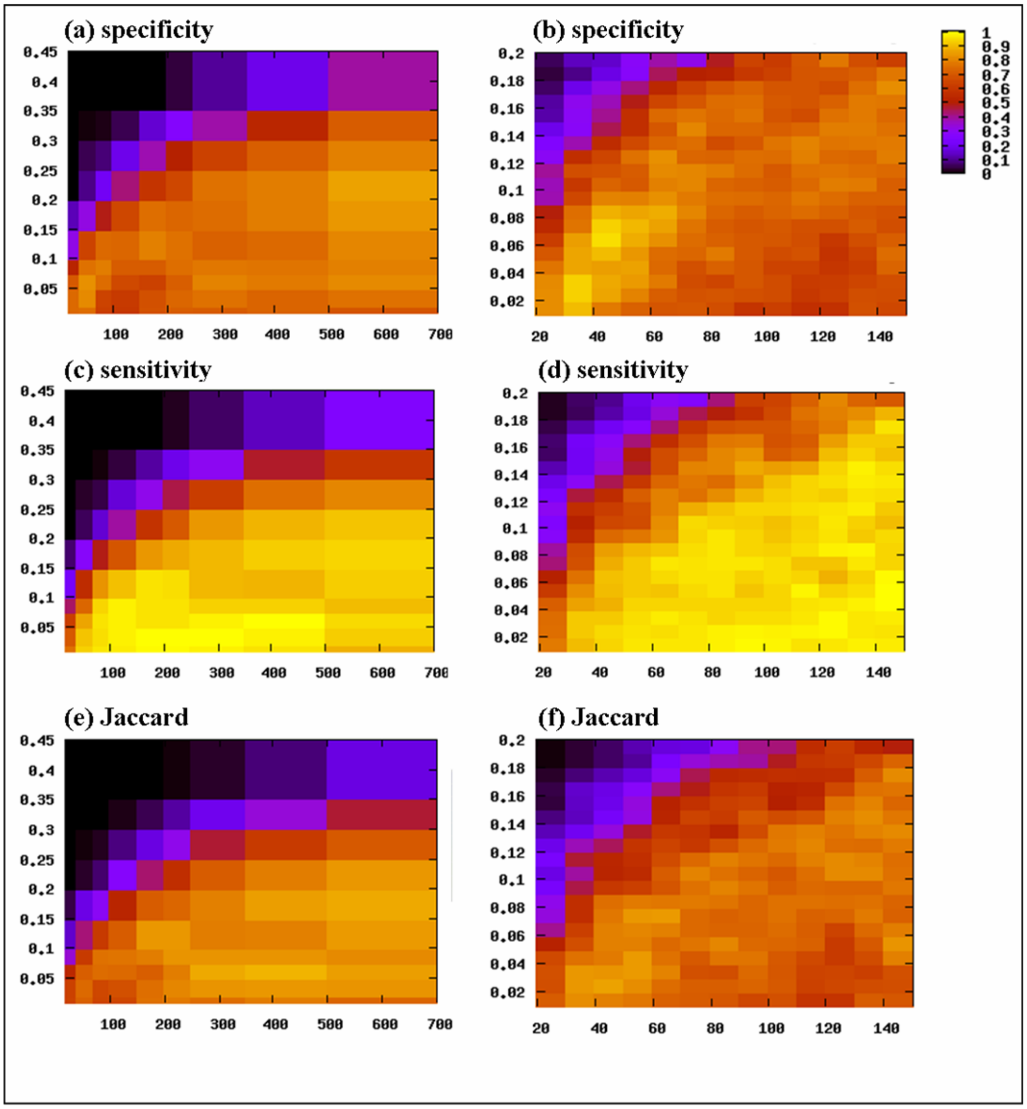


Figure 11: **Simulation results - Interplay of hits number and selection coefficient** Specificity, sensitivity and Jaccard score averages for different combinations of selection coefficient (y-axis) and hit number (x-axis). The plots on the left depict the results for a wide range for both parameters. The plots on the right show a higher resolution for a smaller range of parameters, in the 'twilight zone'. As the selection coefficient gets closer to zero, a smaller number of hits is sufficient.

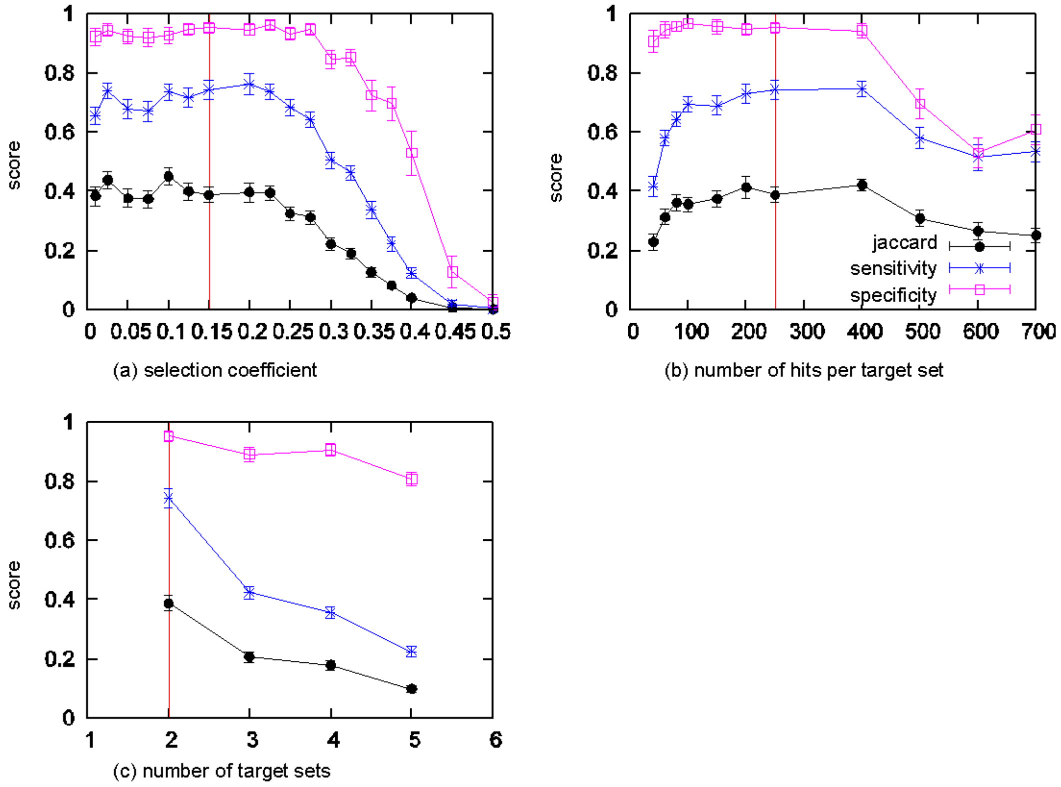


Figure 12: **Simulation results - Performance for close target sets** In the default scenario here (red vertical lines) there are two similar target sets, and the other parameters have the same default values as in the original setting. In this scenario, the specificity is near perfect, the sensitivity is slightly lower, suggesting a higher susceptibility to false negatives, and the Jaccard score is much lower, suggesting a difficulty in correctly partitioning into target sets. **(a) Effect of varying the selection coefficient ; (b) Effect of varying the hit number:** We observed the same trends as in the standard case, albeit at lower scores. Surprisingly, the performance drops when the target sets have over 400 hits. **(c) Effect of the number of target sets:** As expected, increasing the number of target sets causes a decrease in performance.

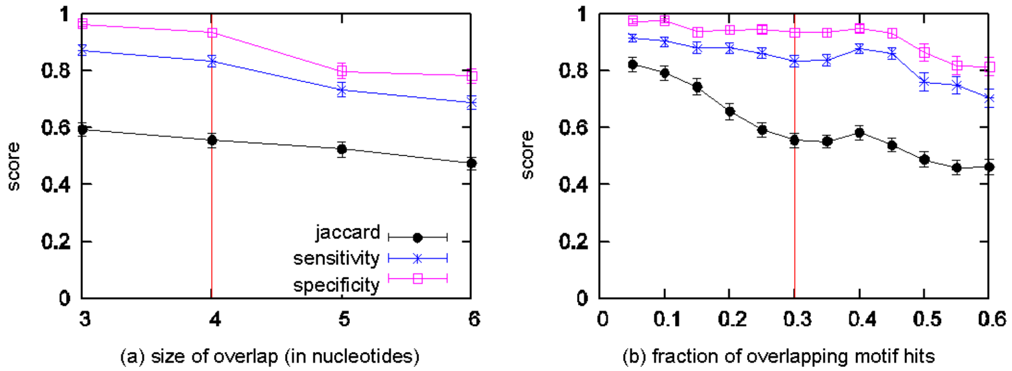


Figure 13: **Simulation results - Performance for overlapping motifs** The default setting in this scenario is two target sets consisting of one octamer each, overlapping in 4 nucleotides in 30% of their occurrences. The scores for the default setting are lower than the scores in the original simulations (See **Fig 10**), but higher than for the close motifs scenario. **(a) Effect of varying the size of the overlap in the motifs:** As the overlap includes more nucleotides, the performance slightly drops. **(b) Effect of varying the percentage of overlapping hits:** As the percentage of overlapping hits grows, we see a decline in the scores, as expected. The specificity is less affected, suggesting that a larger overlap has less effect on the amount of false positives.

B Appendix B: Non-uniform conservation rates along promoter regions

In previous studies, it has been established that the neutral mutation rate is more or less uniform across the yeast genome. In [10], no correlation was found between the mutation rates in neighboring genes. This and other similar results justify our whole-genome approach, and the assumption that the neutral mutation rates are uniform across the different promoters (and therefore one can use a single neutral model). However, we found that we cannot rule out correlations on a smaller scale. In analyzing substitutions in *S. cerevisiae-S. mikatae* promoter alignments, we found high correlation of mutation rates in neighboring windows of small sizes (10-100). Additionally, we found that mutation rates of close sites are correlated, and this correlation slowly decays as the distance between sites increases (0.08 for neighboring sites, 0.06 for sites one base apart, and so on). We also examined the distribution of the number of conserved bases in non-overlapping, gapless windows of size 10-200 in the yeast promoters, and found them to be different than expected if mutations were uniformly distributed across the genome. This may be due to several effects. One possibility is that it is an artifact of the alignment. Another is that TFBSs tend to cluster together, forming islands of high conservation inside the promoters. The latter hypothesis has been brought up in previous studies [8]. A third possibility is that promoters contain conserved islands that are not necessarily TFBS or even functional, but are conserved due to other reasons, such as chromatin positioning or base composition. At this point, we cannot rule out any of these options.

This finding has several practical implications for our study. When examining the conservation z-scores for different k-mers, we find that they are mostly positive. This is expected given the non-uniform distribution of mutations (Over-representation of mutations in certain regions imply under-representation in other regions, leaving large conserved stretches; Imagine an extreme case where all of the mutations are clustered at one end of the alignment). Therefore, it is not enough to examine the conservation of k-mers within a target set, but we need to compare these values to the background distribution for all k-mers. Another, stronger implication, is that there may be other selection forces acting on the promoter regions, and therefore not all conservation can be attributed to TFBS.