Probabilistic models of transcriptional regulation

John Reid

Department of Pure Mathematics and Mathematical Statistics
Peterhouse

October 2013

This dissertation is submitted for
the degree of Doctor of Philosophy

Declaration

I hereby declare that this dissertation is not substantially the same as any that I have submitted for a degree or diploma or other qualification at any other university.

I further state that no part of my dissertation has already been or is being concurrently submitted for any such degree or diploma or other qualification.

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

Probabilistic models of transcriptional regulation

John Reid

Summary

Regulation of transcription lies at the heart of many of the most critical cellular behaviours. This regulation is mediated by the binding of transcription factors to the genome. There is much uncertainty about many aspects of this process. Firstly, individual transcription factors preferentially bind to a variety of genomic sequences. Even when these preferences are known, predicting transcription factor binding sites has a high false positive rate. Secondly, the sequence preferences of many transcription factors are uncharacterised. Biological experiments only provide indirect evidence of transcription factor binding. The task of inferring the sequence preferences given these indirect data is difficult. Thirdly, transcription factors do not operate in isolation. Sets of interacting transcription factors are reused across distinct cellular programs. Many of these cellular programs and sets of transcription factors remain to be characterised. This thesis uses probabilistic modelling techniques to present three contributions that address the aforementioned uncertainty about transcriptional regulation.

I describe a novel method that uses conservation of genomic sequences between related species to improve predictions of transcription factor binding. Unlike other similar methods, a multiple alignment of the regions of interest is not required.

Some of the most popular and best methods for characterising the sequence binding preferences of transcription factors are not able to handle data sets of the size generated by recent biological techniques. I present a novel technique that uses suffix trees to speed up one of the most popular existing techniques so that it can be applied to much larger data sets.

I present a hierarchical non-parametric probabilistic model that captures interactions between transcription factors and their target genes. This model infers several well established interactions between transcription factors in an unsupervised setting.

Taken together these methods demonstrate how apposite probabilistic modelling techniques are for quantifying uncertainty in transcriptional regulation.

Contents

Chapter 1

Introduction

1.1 The biology of transcriptional regulation

1.1.1 Genomes and genes

An individual's genome is the sum of its inheritable traits. Most species, including all eukaryotes, encode a significant part of this information in deoxyribonucleic acid (DNA) macromolecules called chromosomes. A chromosome carries genetic information as a sequence of nucleotides. A nucleobase (or simply base) forms part of each nucleotide. In DNA there are four possible nucleobases: adenine; cytosine; guanine; and thymine. The genetic information stored in a chromosome is often summarised as a sequence of the four characters `A`, `C`, `G` and `T` representing the four possible bases. The chromosomes contain regions called genes that contain the information necessary for the cell to create molecules called gene products. Proteins are the archetypal gene products but some genes express ribonucleic acid (RNA) products. The process by which the information stored in a gene is converted into a gene product is called gene expression.

A single molecule of DNA exists as a single strand. Chromosomes are macromolecules that consist of two complementary molecules of DNA. The two strands are arranged in the well known double-helix configuration [Watson and Crick, 1953]. The complementary nature of the two strands is enforced by bonds between the nucleobases. Adenine binds to thymine and cytosine binds to guanine. If the sequence on one strand is known, the sequence on the complementary strand can be read by switching `A`s with `T`s and `C`s with `G`s. The backbone of a single-stranded DNA molecule is not symmetric and this imparts a directionality on the sequence associated with it. The two ends of the DNA molecule are labelled the 5' and 3' end. The 5' end has a terminal phosphate group and the 3' end has a terminal hydroxyl group. Conventionally the sequence is read from the

5' end to the 3' end. The two strands in double-stranded DNA are oriented in opposing directions. When considering the sequence on both strands of DNA we often talk of taking the reverse complement. The opposing sequences are complemented because of how the nucleobases bond and they are written in reverse order because of the directionality of the two strands. For example, if the sequence on one strand was `ATTCGG`, the reverse complement sequence on the other strand would be `CCGAAT`.

1.1.2 Regulatory networks

Gene expression levels affect many aspects of cellular function and the regulation of gene expression levels is crucial for a cell to function correctly. For example, correct spatio-temporal expression patterns in the *Drosophila* embryo are critical for the successful development of the organism's body plan [Johnston and Nüsslein-Volhard, 1992, Rivera-Pomar and Jãckle, 1996] The regulation of expression levels is crucial in yeast's cellular response to external stimuli such as heat shock [Boy-Marcotte et al., 1999].

Most aspects of cellular behaviour involve several gene products. Interactions between genes and their products can form complex gene regulatory networks that encode particular behaviours. Decoding (or reverse engineering) these networks is crucial to understand these behaviours at a molecular level. In gene regulatory networks the expression of each gene's products is regulated by the activity of other genes. Although regulation can occur at many points in the process of gene expression, in this thesis I focus on transcriptional regulation, one of the most important and pervasive methods of regulation in eukaryotes.

Evolution is a continuous and gradual process. Similar behaviours and phenotypes in related species are likely to derive from a common ancestor and thus share a common underlying mechanism. In particular regulatory networks are often shared across large evolutionary distances [Davidson, 2006] and the mechanisms that implement these networks are also commonly conserved [He et al., 2011]. The expectation that regulatory networks are shared across clades or greater evolutionary distances can be very helpful in the study of such systems. We can often use data from related species to improve or validate our inferences in the species of interest. These expectations should be tempered by evidence that some systems in closely related species have diverged significantly and implement the same behaviours using different mechanisms [Odom et al., 2007].

1.1.3 Transcriptional regulation

The first stage in gene expression is transcription. Transcription is the process of creating a complementary RNA copy of a sequence of DNA. As transcription is a necessary part of gene expression it is a convenient point at which a cell can regulate its gene expression levels. It is perhaps because of this convenience that so many cellular behaviours use transcriptional regulation as a control mechanism.

To understand transcriptional regulation, we should first understand transcription. In transcription, the RNA polymerase enzyme binds to a region upstream of the gene called the promoter and moves along the DNA sequence producing a complementary strand of RNA. The genomic location at which transcription starts is referred to as the transcription start site (TSS). The rate of transcription can be affected by proteins called transcription factors (TFs). That the expression of genes could be affected in this way was first discovered in studies of *Escherichia coli* [Gilbert and Maxam, 1973, Maizels, 1973, Dickson et al., 1975]. There are several types of TF: general TFs are necessary for the RNA polymerase to position at the promoter; activating TFs increase the rate at which the RNA polymerase associates with the promoter increasing the rate of gene expression; repressing TFs are similar to activating TFs but work oppositely, decreasing the interaction between the RNA polymerase and the promoter to reduce the rate of expression; specificity TFs alter the specificity of the RNA polymerase for a whole class of promoters, for example sigma factors in prokaryotic transcription [Gruber and Gross, 2003]; insulating TFs such as CTCF partition the genome into regulatory domains [Ohlsson et al., 2001].

1.1.4 Transcription factor binding

TFs act by binding to the DNA at transcription factor binding sites (TFBSs). TFBSs are also sometimes known as response elements (REs). Many TFs have sequence-specific binding preferences, that is, they prefer to bind to TFBSs that comprise a particular sequence of base pairs. These TFs can show variability in their binding sequences [Maniatis et al., 1975]. For example, the TATA-binding protein (TBP) which is a general TF, prefers to bind to the sequence `TATAAAA` but can also bind to sequences such as `TATATAT` or `TATATAA`. A molecular model of the structure of TBP-DNA binding is shown in Figure 1.1. The preferred binding sequence for a TF is termed its consensus sequence. In general a TF will have differing propensities for binding different DNA sequences. Sarai and Takeda demonstrated this by analysing the binding affinities of the *cro* and *repressor* TFs to mutated versions of the λ operator [Sarai and Takeda, 1989].

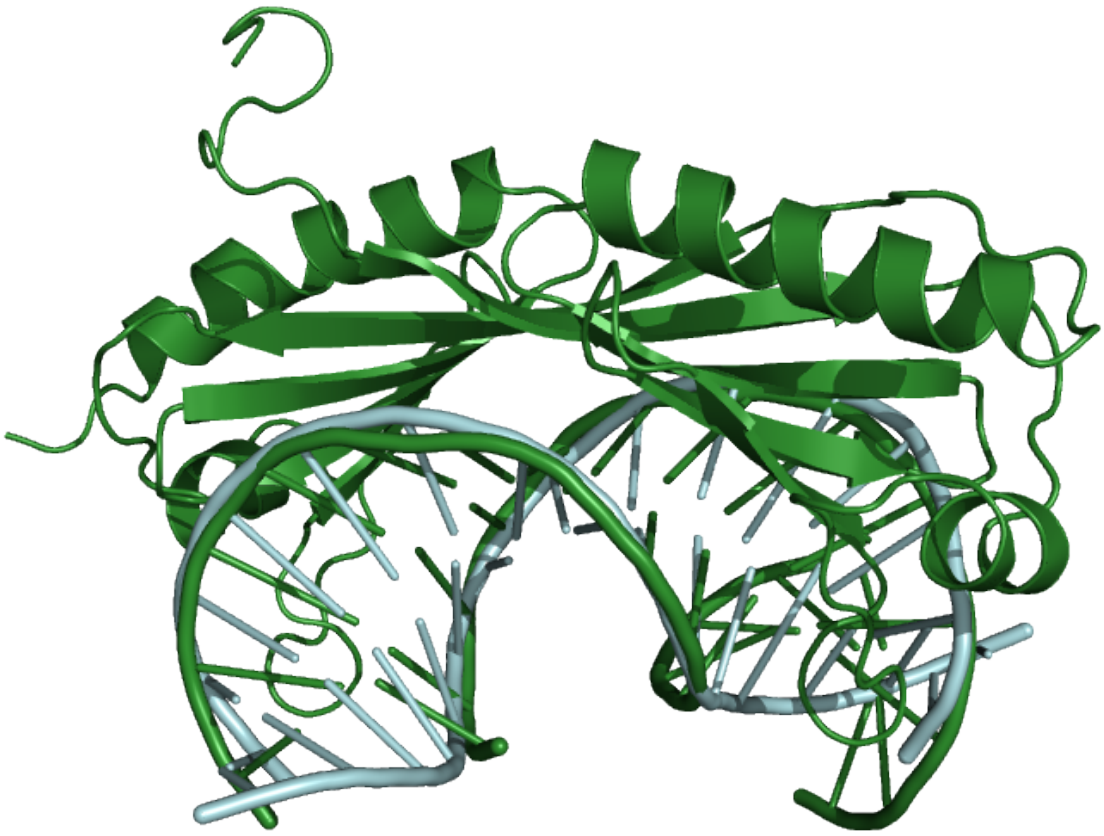


Figure 1.1: A molecular model of the structure of TBP-DNA binding. The model was generated by the 3D-DART web server [Dominguez et al., 2003].

TFs do not work in isolation from each other. TFBSs often appear in clusters or cis-regulatory modules (CRMs) (also known as enhancers), presumably to enable interactions between TFs binding there. As each TF is itself a gene product, CRMs allow the genome to encode responses to particular expression levels. This feedback is the basis of many well characterised gene regulatory networks, for example much of the gene expression in *Saccharomyces cerevisiae* can be explained in this way [Lee et al., 2002].

Most well characterised enhancers are close to the genes that they regulate but there are many examples of enhancer-gene interactions over large genomic distances. In the *Drosophila* Antennapedia complex the T1 enhancer acts on the distal Sex combs reduced gene, which is located on the other side of the nearby fushi tarazu gene [Gindhart et al., 1995, Calhoun et al., 2002, Calhoun and Levine, 2003]. The sonic hedgehog gene is regulated by an enhancer located nearly one megabase away [Amano et al., 2009]. Bateman et al. have shown that a *Drosophila melanogaster* enhancer can act on multiple genes on separate chromosomes [Bateman et al., 2012]. In general, TFBSs tend to be located close to transcription start sites (TSSs) although they can also be found in

intergenic regions and even gene deserts. There are also examples of intronic and exonic TFBSs. This distribution of sites around genomic features is not uniform and can be used to help in TFBS prediction.

Each TF has a particular protein structure and TFs can be organised into families on this basis. Members of the same family often exhibit similar binding preferences [Sandelin and Wasserman, 2004].

Sequence preferences are not the only way patterns of TF binding are determined. In the rest of this section I highlight some of the other mechanisms implicated in the process.

Competitive binding

Sterically two TFs cannot bind to the same section of DNA. If a TFBS can be bound by different TFs and both are present in the nucleus they will compete for that binding site [Teif and Rippe, 2010]. This is termed competitive binding. One competing TF might be an activator while another might be a repressor. As only one TF can bind the DNA at the TFBS this allows the genome to encode a response to relative expression levels of TFs. TFs of the same family are often implicated in competitive binding as their sequence preferences are typically similar.

Weak binding sites

The cell uses TFBSs of varying affinities to direct transcriptional responses. Tanay showed that weak binding sites in several *Saccharomyces* species are functional [Tanay, 2006]. Several computational models of gene expression support the idea that weak binding sites are important for correct expression patterns [Gertz et al., 2008, Segal et al., 2008, Koohy et al., 2010, McLeay et al., 2012]. These models predict that weak binding sites permit a sensitive response to TF concentration levels in the cell. Weak binding sites may also be relevant to competitive binding as weaker binding sites are more likely to match the binding preferences of more than one TF. There is some evidence to suggest that the strength of a binding site correlates with its use in a particular expression program. For example, it has been shown that medium or weak affinity FOXA2 binding sites are associated with liver-specific expression [Tuteja et al., 2008].

Cofactors

As well as binding competitively, TFs often bind synergistically [Hochschild and Ptashne, 1986]. TFs often bind to DNA as homodimers or heterodimers. For example, the TFs

POU5F1 and SOX2 are known to regulate pluripotency in mouse embryonic stem cells via synergistic binding [Loh et al., 2006]. POU5F1 binds to `ATGCAAAT` and SOX2 binds to a neighbouring SOX element. In this instance POU5F1 and SOX2 are termed cofactors. Weak binding sites can be relevant to synergistic binding as the presence of cofactors can be enough to overcome lack of binding due to low-affinity sites.

Multiple binding modes

There are several examples of TFs that exhibit more than one mode of binding. The TF ELK1 binds to high-affinity sites as a monomer. However, in the presence of the TF SRF, ELK1 can bind to low-affinity sites [Treisman et al., 1992]. Tanay et al. showed that REB1 has two distinct modes of binding associated with positive and negative autoregulation [Tanay et al., 2004a]. The TF Yin Yang 1 (YY1) is named after its ability to act as an activator, repressor, or initiator of transcription. YY1 has been shown to have distinct sequence preferences depending on whether it is acting as an activator or repressor [Whitfield et al., 2012]. The biophysical reasons for these two modes could be the presence of cofactors. Fordyce et al. have shown that the TF HAC1 binds DNA in two distinct modes [Fordyce et al., 2012].

A common theme amongst the few known examples of TFs with multiple modes of binding is that the modes are quite similar. That is, there is never too much divergence in the sequence preferences between the modes. Typically there is some evidence to suggest that the modes differentiate high-affinity sites from low-affinity sites.

Epigenetic effects

Accessibility to TFBSs can play a major role in patterns of TF binding. In eukaryotes DNA is wrapped around nucleosomes which package it into a structure called chromatin. Each nucleosome is an octamer of histone proteins. The positioning of the nucleosomes relative to a TFBS can affect its availability for binding. Some TFs such as Swi/Snf are known to affect how nucleosomes are positioned [Malik and Roeder, 2005], presumably to dynamically alter binding patterns.

In addition to nucleosome positioning, each histone protein can be chemically modified in various ways, each of which can affect how it interacts with the DNA and TFs [Kouzarides, 2007]. These modifications can be induced by TFs themselves, for example a TF might recruit CREB-binding protein (CBP) which can acetylate lysine amino acids on histone proteins. This acetylation can create binding sites for protein-protein interaction domains [Mujtaba et al., 2007]. TFs that act to remodel chromatin

are called pioneer TFs. There is increasing evidence that there is a combinatorial histone modification code [Wang et al., 2008]. This code partitions the genome on the basis of the combinations of histone modifications that are present. Different types of genomic regions have been found to be associated with particular combinations. For example, enhancers and promoters can be distinguished on the basis of these marks alone [Heintzman et al., 2007].

Deoxyribonuclease I (DNase I) is an endonuclease that cleaves DNA [Wu, 1980, Keene et al., 1981, McArthur et al., 2001]. It has been shown that regions that are hypersensitive to cleavage are associated with TF binding [Bernat et al., 2006, Hesselberth et al., 2009]

In addition to chemical modifications of the histone proteins, chemical modifications of the DNA itself can play a part in regulating TF binding. DNA methylation is such a modification [Choy et al., 2010]. For example, the insulating TF CTCF is known to have an enhancer blocking effect, that is, CTCF binding in between a gene and an enhancer can prevent the enhancer from affecting transcription of the gene. Bell and Felsenfeld showed that a particular CTCF TFBS is only occupied when the DNA is not methylated and that this affects expression levels [Bell and Felsenfeld, 2000].

All of the above modifications that do not affect the sequence of bases are termed epigenetic modifications despite evidence that they are heritable [Cavalli, 2002]. These epigenetic effects are dynamic in that they usually vary across cell types [Heintzman et al., 2009, Cui et al., 2009].

1.1.5 The combinatorics of transcriptional regulation

Particularly in higher organisms, combinatorial operations are often necessary for the response of a cell to external stimuli or developmental programs. Such a response is frequently implemented as a transcriptional switch where a combination of presence or absence of certain TFs regulates the expression of a certain gene. Several well characterised examples of the coordination of TFs are known. For instance, a set of well studied TFs in *Drosophila melanogaster* that govern spatial patterns of development in its embryo is described by Ingham [Ingham, 1988]; POU5F1, SOX2 and NANOG are known to interact to maintain pluripotency [Chen et al., 2008]; higher eukaryotes are known to use CRMs to integrate cellular signalling information [Arnosti and Kulkarni, 2005]; the development of the anterior pituitary gland is regulated by combinatorial actions of specific activating and restricting factors [Simmons et al., 1990] which determine cell type.

Conversely, cellular processes often involve the coordinated expression of sets of genes.

Hence there is reason to suppose that not only do particular sets of TFs regulate particular genes but that these sets are also reused across the genome: that is, co-regulated genes are often targets of the same TFs.

It is well known that the cell reuses overlapping sets of TFs as regulators at distinct developmental stages and in different tissue types. For example, the TF Krüppel is involved in the segmentation of the *Drosophila* embryo [Ingham, 1988] and is also implicated in the development of the embryo's central nervous system [Nakajima et al., 2010]. In the segmentation system Krüppel is part of a program that includes the other TFs giant, huckebein, hunchback, knirps and tailless. In the nervous system Krüppel works alongside the TFs hunchback, pdm, castor and sevenup. Commonly models of transcriptional regulation enforce mutual exclusivity on the sets of coordinated TFs. The methods in this thesis do not enforce this unrealistic assumption.

Cis-regulatory grammars

When TFBSs cluster together there is the question of grammar: is a precise spatial positioning of the TFBSs necessary for the correct interactions between the TFs to occur? There are several models of TFBS clusters related to this question. At one end of the scale is the enhanceosome model where the TFs are required to assemble on the DNA in a particular order and positioning. At the other end of the scale is the billboard model where the locations and orientations of the TFBSs are not important, it only matters that they are present. Arnosti and Kulkarni developed these models in a review paper [Arnosti and Kulkarni, 2005].

There are several known examples of enhanceosomes. The classical example of the enhanceosome is the interferon-beta enhanceosome [Thanos and Maniatis, 1995, Agalioti et al., 2000, Panne et al., 2007]. It has also become clear that the orientation of TFBSs can be important, that is, which DNA strand the TF binds to. In the presumptive neurogenic ectoderm of *Drosophila* TFBSs for the TFs twist and dorsal are required to have a particular orientation [Papatsenko and Levine, 2007]. Regulation of Pax2 in the *Drosophila* eye provides another example. There is an enhancer upstream of Pax2 that contains 12 TFBSs for the TFs Lozenge, Su and Ets. When the arrangement of TFBSs is altered, Pax2 is expressed in different cell types in the eye [Swanson et al., 2010]. That the arrangement of TFBSs can result in cell-type specific expression highlights the importance of the enhanceosome model.

It is believed that many developmental enhancers follow the billboard model albeit with a limited grammar [Levine, 2010]. Many such enhancers have been shown to be functional under limited rearrangements of their TFBSs.

Phylogenetic conservation

As discussed in Section 1.1.2 we can expect that TFBSs for homologous TFs in related species will be conserved. Furthermore we can expect regions that contain functional binding sites to show non-neutral rates of evolution [Moses et al., 2003]. However, there are well established examples of turnover of TFBSs between closely related species. In these cases evolution has destroyed and recreated the TFBSs necessary for regulatory networks to function.

TFBS binding patterns are not always conserved between related species. Odom et al. mapped the locations of four tissue-specific TFs in human and mouse hepatocytes [Odom et al., 2007,Schmidt et al., 2010]. They showed that between 41% and 89% of these binding events were species specific. Tsong et al. analysed a regulatory network controlling mating in a yeast lineage [Tsong et al., 2006]. They showed that although the outcome of the circuit remains identical the mechanism via which the outcome was attained has changed from using an activator to using a repressor. Borneman et al. showed that the TFBSs for the yeast TFs Ste12 and Tec1 diverged more quickly than the genes they regulated [Borneman et al., 2007]. They argue that this suggests evolution uses turnover of TFBSs as a niche specialisation mechanism. Kunarso et al. analysed the binding locations of the TFs POU5F1, NANOG and CTCF in murine and human embryonic stem cells [Kunarso et al., 2010]. They found that whilst CTCF binding sites were largely conserved, transposable elements had rewired the regulatory network controlling pluripotency and the TFBSs for POU5F1 and NANOG had diverged significantly. In a similar study of adipogenesis, Mikkelsen et al. showed that there was a significant turnover of TFBSs between human and mouse even when expression patterns were similar [Mikkelsen et al., 2010]. In contrast to the preceding examples, He et al. showed that the binding of the developmental TF twist is highly conserved across six *Drosophila* species [He et al., 2011].

1.1.6 Uncertainty in transcriptional regulation

It is clear from the preceding discussion of the biology of transcriptional regulation that there is much that we do not know about it. Modern high-throughput biological techniques allow us to learn much about individual TFs and genes. Recently there has also been a massive increase in the amount of data available about epigenetic effects. Both of these data sources have been useful for decoding combinatorial effects in regulatory networks. However, most of the work on combinatorial regulation has been performed on a few model systems. For example, the segmentation network in *Drosophila melanogaster*, the regulation of pluripotency in humans and mice and muscle and liver development

in humans have been particularly heavily studied. Many regulatory networks and many aspects of combinatorial transcriptional regulation remain to be characterised.

1.2 Experimental techniques

The discovery that gene expression could be regulated by TFs binding to DNA initiated an on-going effort in molecular biology to determine the binding preferences of these proteins. Here I give an overview of some of these techniques.

The electrophoretic mobility shift assay (EMSA) can determine if a TF binds to a particular DNA sequence [Fried and Crothers, 1981, Garner and Revzin, 1981]. It relies on the principle that a fragment of DNA will move more slowly through a gel when it is bound by a TF. However, it cannot determine the location of the TFBS in the DNA sequence so it is useful as a test for whether a TF binds to a particular promoter or enhancer but not for learning the TF's sequence preferences. DNase I footprinting is a technique that allows the location of TFBS in a particular sequence to be determined [Galas and Schmitz, 1978, Dynan and Tjian, 1983, Brenowitz et al., 1986]. DNase I footprinting relies on the fact the DNA bound by the TF is protected from cleavage when digested by DNase I. Both EMSA and DNase I footprinting are *in vitro* techniques and demonstrate that a TF can bind to a sequence outside the cellular environment. The *in vivo* footprinting assay extends the DNase I footprinting technique to allow the experimenter to determine if the TFBS is bound *in vivo*.

All the above techniques allow an experimenter to determine if a TF binds to one particular DNA sequence. To elucidate the sequence binding preferences of a TF, techniques that examine the TF's affinity for many sequences had to be developed. Systematic evolution of ligands by exponential enrichment (SELEX) works by repeatedly selecting sequences from a library of randomly generated sequences [Oliphant et al., 1989, Ellington and Szostak, 1990, Tuerk and Gold, 1990]. The sequences are selected based on their ability to bind the TF of interest. Yeast and bacterial one-hybrid systems are based on transforming yeast or bacterial cells with a TF and some potential binding sequences [Bulyk, 2005]. The cells are positively and negatively selected based on expression of reporter genes. Motif finding tools (see Section 1.4.5) can be used to derive the sequence binding specificities of the TF from the selected sequences.

Protein-binding microarrays (PBMs) are a high-throughput *in vitro* technique that can characterise the binding preferences of a TF [Mukherjee et al., 2004]. A typical configuration would place all 10-mer sequence variants on a microarray. A PBM experiment quantifies a TF's affinity for each of these 10-mers. PBMs do not reveal the locations of TFBSs themselves and they require an antibody for the TF under investigation.

Fordyce et al. have developed a micro-fluidic method mechanically induced trapping of molecular interactions (MITOMI) to determine TF-DNA affinities [Fordyce et al., 2010]. Whilst PBMs can discover both strong and weak TFBSs, they do not measure TF-DNA reactions at equilibrium as MITOMI does.

SELEX, the one-hybrid systems and PBMs are *in vitro* techniques. Recently high-throughput *in vivo* techniques have been developed. ChIP-chip combines chromatin immunoprecipitation (ChIP) with microarrays (chip) to resolve TFBS locations to within a few hundred base pairs [Blat and Kleckner, 1999, Aparicio et al., 2004, Buck and Lieb, 2004, Liu and Meyer, 2009]. ChIP-chip can be expensive and in recent years has been largely superseded by ChIP-seq. ChIP-seq combines chromatin immunoprecipitation with massively parallel DNA sequencing to achieve a similar effect [Johnson et al., 2007, Robertson et al., 2007, Park, 2009]. ChIP-seq has the potential to more accurately resolve TFBSs down to a few tens of base pairs [Jothi et al., 2008]. DNA adenine methyltransferase identification (DamID) is a method that does not require antibodies [van Steensel and Henikoff, 2000, Southall and Brand, 2007] for the TF under investigation. DamID has a resolution of about 200 base pairs.

In vivo techniques are important as many potential TFBSs are only utilised in particular tissues. Epigenetic modifications such as nucleosome positioning can render some parts of the genome inaccessible to TFs, effectively shutting down entire regulatory systems. The cell typically uses these modifications to achieve specific spatio-temporal expression patterns. Analysis of sequences for TFBSs alone cannot reveal which regions are available for binding in which tissue types. *In vivo* experimental techniques can be applied to particular tissue types at specific stages of development to determine the binding patterns that occur in that specific context.

There is also a growing number of experimental techniques that can provide data on epigenetic effects. FAIRE-seq [Giresi et al., 2007] can reveal which regions of a genome are depleted of nucleosomes. This depletion is associated with regulatory activity in any given cell sample independently of any given TF. ChIP-seq does not need to be performed with a TF specific antibody, histone modifications can also be determined via the appropriate antibodies. Methyl-seq is a technique to find methylated regions of DNA [Brunner et al., 2009]. DNase-chip is a technique to determine which genomic regions are hyper-sensitive to DNase I cleavage [Crawford et al., 2006].

1.3 Probabilistic models

Probabilistic models are a way of specifying a joint distribution over a set of random variables [Jordan, 2004]. They are used across a wide range of scientific disciplines for

a diverse set of tasks.

Probabilistic models have been used in state-of-the-art solutions for many problems. Also many *ad hoc* state-of-the-art methods have been reinterpreted as the application of standard inference techniques to particular probabilistic models. For example, the interpolated Kneser-Ney method of smoothing n-gram language models was a state-of-the-art method. Teh showed it to be equivalent to a particular form of approximate inference in a non-parametric Bayesian hierarchical model [Teh, 2006]. By using this model with a better inference technique Teh was able to improve performance.

1.3.1 Application

The application of probabilistic models requires three steps. Firstly, the model should be specified. The specification defines the relationship between all the random variables in the model. The relationships between the variables can be specified using one of several graphical formalisms that make the assumptions in the model explicit.

Secondly, given the model, inference will be performed. In any particular model, some of the variables will be observed and some will be latent (hidden). Inference produces a posterior distribution over the latent variables. Probabilistic models have been found to be a good fit for Bayesian inference techniques although they are not limited to them. In general, inference will produce a posterior distribution over all the latent variables. There are many different inference schemes to choose from. Some are designed for specific models, some are designed for efficiency and others for accuracy. The recent interest in machine learning techniques has resulted in a large literature of analysis of the properties of these inference schemes.

Thirdly, when inference is finished and there is a posterior distribution over the latent variables, it is used to solve the task under consideration.

1.3.2 Benefits

Probabilistic models have several advantages over *ad hoc* methods. Firstly the assumptions in the model are explicit. Conditional dependencies between the variables can be read from a representation of the model in graphical form.

Secondly, when compromises between accuracy and efficiency in the inference are necessary, the growing literature allows us to understand which ones might be acceptable.

Thirdly, uncertainty in the latent variables is quantified so we have a measure of confidence in the results. This quantification follows the laws of probability. These laws have

been shown to be the only reasonable way to reason in the face of uncertainty (see De Finetti's Dutch book arguments [De Finetti, 1937, De Finetti, 1964]). It is often useful to propagate this uncertainty through to downstream tasks.

Fourthly, the separation of model definition and inference is useful in a practical manner. The task of deciding on a suitable model or models is de-coupled from the task of inference. An expert in the system to be studied can build one or more models. An expert in inference can take these models and apply suitable inference techniques from the large number available in the literature.

1.3.3 Mixture models

Mixture models are widely used and have proved extremely flexible in modelling a variety of data. The methods in Chapters 3 and 4 use mixture models so I describe them here. An example of a simple mixture model serves to introduce graphical plate notation for probabilistic models.

A mixture model defines a distribution over a set of data, $x_1, \ldots, x_N$, using latent variables that assign each datum to a component. The number of components, K, is typically but not necessarily fixed.

The archetypal example of a mixture model is a mixture of Gaussians and we use this model to illustrate the idea. It is easiest to describe the model as a process that generates the N data, $x_1, \ldots, x_N$. Each datum, x_n, has a latent variable associated with it, z_n, that identifies the component it is generated by. The z_n are drawn at random from $\{1, \ldots, K\}$

$$z_n \sim \text{Uniform}(K) \tag{1.1}$$

Now for each component k, where $1 \leq k \leq K$ the model has latent variables representing the parameters of a Gaussian distribution: μ_k is the mean and the σ_k^2 is the variance. μ_k and σ_k^2 are drawn from some prior distribution. A typical choice is an inverse Gamma distribution

$$\sigma_k^2 \sim \text{inverse-Gamma}(\nu, \sigma_0^2) \tag{1.2}$$

and another Gaussian

$$\mu_k \sim \mathcal{N}(\mu_0, \lambda\sigma_k^2) \tag{1.3}$$

where ν, σ_0^2, μ_0 and λ are hyper-parameters of the model. Now we generate the data x_n using the parameters of the component it belongs to, z_n,

$$x_n \sim \mathcal{N}(\mu_{z_n}, \sigma_{z_n}^2) \tag{1.4}$$

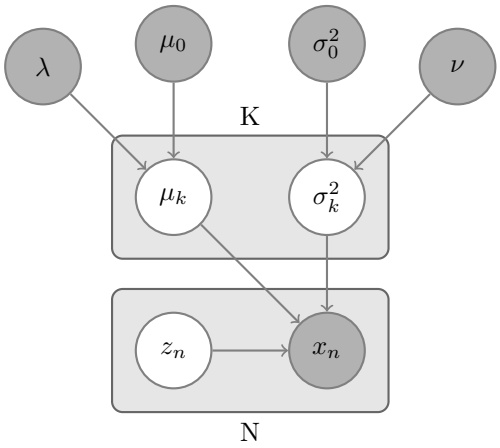


Figure 1.2: Our mixture of Gaussians represented in graphical plate notation. The plates represent multiplicities of variables. There are K copies of the μ_k and σ_k^2 nodes and N copies of the x_n and z_n nodes. I use the convention that observed nodes are shaded and latent (or hidden) nodes are left unshaded. The arrows represent conditional dependencies in the joint distribution of the variables. Arrows that cross plates also have that multiplicity. For example, this model represents KN dependencies between all N x_n and all K μ_k nodes but only one between any given pair of z_n and x_n nodes.

The dependencies between the variables in this model are easily seen using graphical plate notation. In this notation variables inside a plate (or box) have multiple instantiations. The plate is labelled with a number indicating how many instantiations of the variable are in the model. For example, in our mixture of Gaussians model, there would be a plate around the nodes labelled x_n and z_n and the plate would be labelled with N. Similarly the two nodes representing μ_k and σ_k^2 would be surrounded by a plate labelled with K. This is illustrated in Figure 1.2.

1.3.4 Likelihood functions and ratios

The likelihood function (or just likelihood) of a probabilistic model, $\mathcal{L}(\theta; D)$, is a function of the parameters of the model, θ. It is equivalent to the probability of the data, D, given those parameters

$$\mathcal{L}(\theta; D) = p(D|\theta) \tag{1.5}$$

This equivalence means that the likelihood is simply an alternative view of the model, one that is normally taken when considering a fixed set of data whilst varying the parameters. Given two parameter values, θ_1 and θ_2, the likelihood ratio, LR, is the ratio of the two likelihoods

$$\text{LR} = \frac{\mathcal{L}(\theta_1; D)}{\mathcal{L}(\theta_2; D)} = \frac{p(D|\theta_1)}{p(D|\theta_2)} \tag{1.6}$$

To avoid confusion it is worth noting that in classical statistics the test statistic, T, of the likelihood-ratio test is twice the logarithm of the LR

$$T = 2[\log p(D|\theta_1) - \log p(D|\theta_2)] = 2 \log \mathrm{LR} \tag{1.7}$$

When I use the term log-likelihood ratio in this thesis I always mean just the natural logarithm of LR.

1.3.5 Bayes factors

Jeffreys introduced the Bayes factor [Jeffreys, 1935, Jeffreys, 1998, Kass and Raftery, 1995], K, as a statistic for comparing a hypothesis, $\mathcal{H}$, to its alternative, $\overline{\mathcal{H}}$, given some data, D. The Bayes factor is defined as

$$K = \frac{p(D|\mathcal{H})}{p(D|\overline{\mathcal{H}})} \tag{1.8}$$

and it measures the change in odds when moving from the prior odds to the posterior odds [Lavine and Schervish, 1999]

$$\frac{p(\mathcal{H}|D)}{p(\overline{\mathcal{H}}|D)} = \frac{p(D|\mathcal{H})}{p(D|\overline{\mathcal{H}})} \frac{p(\mathcal{H})}{p(\overline{\mathcal{H}})} = K \frac{p(\mathcal{H})}{p(\overline{\mathcal{H}})} \tag{1.9}$$

In this sense it can be interpreted as a measure of the evidence that the data, D, provide for the hypothesis, $\mathcal{H}$. It is worth noting that the relationship between the prior odds, Bayes factor and posterior odds means that we can calculate any one of them given the other two.

Of course our models often have hidden variables, H, (or parameters, θ) and in a Bayesian framework we often integrate over these. In these cases we use the following substitution for $p(D|\mathcal{H})$

$$p(D|\mathcal{H}) = \int_H p(D, H|\mathcal{H}) \, \mathrm{d}H \tag{1.10}$$

in the equations above.

In practice Bayes factors are often converted to the logarithmic scale for two reasons. Firstly they are symmetrical in the hypotheses. That is, a log Bayes factor of $\log K$ in favour of $\mathcal{H}$ is a log Bayes factor of $-\log K$ in favour of $\overline{\mathcal{H}}$. Secondly it is convenient that log Bayes factors add rather than multiply for independent sets of data. This conversion to the logarithmic scale is also typical for likelihood ratios.

When several sets of data are available, $D_1, \ldots, D_N$, we can calculate a Bayes factor, $K_1, \ldots, K_N$, for each set

$$K_n = \frac{p(D_n|\mathcal{H})}{p(D_n|\overline{\mathcal{H}})} \tag{1.11}$$

If the data sets are independent, that is, $p(D_n|\mathcal{H}) = p(D_n|\mathcal{H}, D_{n'})$ for all $n \neq n'$, we can use the product of the individual Bayes factors to summarise the overall change in odds, K_{all}, when observing all the data sets

$$K_{\text{all}} = \prod_n K_n \tag{1.12}$$

The situation is more complicated when the data are not independent. Hypothetically, suppose we have the extreme case where the data are guaranteed to be equivalent, $D_n = D_{n'}$ for all n and n'. Hence we also have $K_n = K_{n'}$ for all n and n'. This is total dependence: knowing one datum allows us to predict the rest. If we know the data will be equivalent before we observe them, we gain no more evidence by observing all the data rather than just one. In this case the product of the K_n in Equation 1.12 would over-estimate the evidence provided by the data. Taking the geometric mean

$$K_{\text{dep}} = \sqrt[N]{\prod_n K_n} \tag{1.13}$$

gives a Bayes factor $K_{\text{dep}} = K_1$ as desired. K_{all} is theoretically justified and does indeed behave as we expect in cases where the data are independent. I have no theoretical justification for K_{dep} and it can only be seen as an *ad hoc* method of integrating the evidence from dependent data. However, K_{dep} does have some of the attributes we desire when integrating evidence from dependent data: it is symmetric in the K_n and it reduces to K_{all} when $N = 1$. When the dependencies between the data are not known or are hard to quantify, it is generally preferable to be cautious when integrating the evidence available. In this context, Equation 1.13 is a suitable *ad hoc* approach to summarise the overall change in odds given by dependent data.

1.3.6 Kullback-Leibler divergence

In Bayesian inference it is often necessary to compare how different two distributions are. The Kullback-Leibler divergence (KL-divergence), $KL(p||q)$, is such a measure of separation between two probability distributions or densities, $p(x)$ and $q(x)$ [Kullback, 1959, Kullback and Leibler, 1951]. It is not symmetric and does not satisfy the triangle inequality and therefore is not a metric. However, it does have the useful property that

$KL(p||q) = 0$ if and only if $p = q$. In the discrete case, KL-divergence is defined as

$$KL(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

and in the continuous case

$$KL(p||q) = \int_x p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x$$

Notationally perhaps it is simplest to regard it as the expectation of the log-likelihood ratio

$$KL(p||q) = \left\langle \log \frac{p(x)}{q(x)} \right\rangle_p$$

The KL-divergence is also known as the *relative entropy*.

The KL-divergence has an interpretation as the divergence of a prior from the truth. Suppose x is a random variable distributed as $p(x)$. Alan knows the distribution p but Brian does not, he believes x is distributed as $q(x)$. $KL(p||q)$ is the expected difference between Brian's surprisal and Alan's surprisal when the value of x is observed.

1.3.7 Expectation maximisation algorithm

The Expectation-Maximisation (EM) algorithm attempts to find the maximum likelihood estimate (MLE) of the parameters of a probabilistic model [Dempster et al., 1977]. It has two steps, the Expectation step (E-step) and Maximisation step (M-step), that are repeated until convergence of the expected value of the likelihood to a local maximum.

More exactly, suppose we have a model $p(D, H|\theta)$ that defines a joint distribution over some observed data, D, and some unobserved data, H, given some parameters, θ. We wish to find the parameters, $\hat{\theta}$, that maximise the expected value of the marginal likelihood of the observed data, $\langle p(D|\hat{\theta}) \rangle_{p(H|D,\hat{\theta})}$, under the posterior distribution of the hidden variables given our estimate $\hat{\theta}$, $p(H|D, \hat{\theta})$.

We start with any estimate of the parameters, θ_0. The E-step calculates the expected value of the marginal log-likelihood as a function of θ given our current estimate, θ_t,

$$Q(\theta|\theta_t) = \langle p(D|\theta) \rangle_{p(H|D,\theta_t)} \tag{1.14}$$

The M-step updates our estimate, θ_t, by maximising Q

$$\theta_{t+1} = \arg\max_\theta Q(\theta|\theta_t) \tag{1.15}$$

Q is bounded and the updates always increase Q so we are guaranteed to find a local maximum. Depending on the model, the EM algorithm can be sensitive to the initial estimate of the parameters. This is a particular problem with a multi-modal likelihood.

1.3.8 Variational inference

Variational inference is a Bayesian inference technique that approximates the posterior distribution over the hidden variables in a probabilistic model [Hinton and van Camp, 1993,Jaakkola, 1997,Jordan et al., 1999,Neal and Hinton, 1998,Winn, 2003]. Variational inference is used in instances where the full posterior of the model, $p(H|D)$ is intractable. A variational distribution, $q(H)$, is used to approximate the true posterior

$$q(H) \approx p(H|D) \tag{1.16}$$

Variational inference is the minimisation of the distance between $q(H)$ and $p(H|D)$. Any distance measure can be used however in practice the KL-divergence often simplifies the inference task.

Following the exposition in Winn's Ph.D. thesis [Winn, 2003] we show how using the KL-divergence simplifies the variational update equations. We want to minimise

$$KL(q||p) = \int_H q(H) \log \frac{q(H)}{p(H|D)} \, \mathrm{d}H = \left\langle \log \frac{q(H)}{p(H|D)} \right\rangle_{q(H)} \tag{1.17}$$

however we do not know $p(H|D)$ so we cannot do this directly. Making the substitution $p(H|D) = \frac{p(D,H)}{p(D)}$ we have

$$\begin{aligned} KL(q||p) &= \left\langle \log \frac{q(H)p(D)}{p(D,H)} \right\rangle_{q(H)} \\ &= \left\langle \log q(H) \right\rangle_{q(H)} + \log p(D) - \left\langle \log p(D,H) \right\rangle_{q(H)} \\ &= \log p(D) - \left[\mathbb{H}(q) + \left\langle \log p(D,H) \right\rangle_{q(H)} \right] \end{aligned} \tag{1.18}$$

where $\mathbb{H}(q) = -\langle \log q(H) \rangle_{q(H)}$ is the entropy of q. As $\log p(D)$ does not depend on $q(H)$ our task is to choose $q(H)$ such that $\mathbb{H}(q) + \left\langle \log p(D,H) \right\rangle_{q(H)}$ is maximised and hence $KL(q||p)$ is minimised. Following Winn we define $\mathcal{L}(q)$ to be the term to be maximised

$$\mathcal{L}(q) = \mathbb{H}(q) + \left\langle \log p(D,H) \right\rangle_{q(H)} \tag{1.19}$$

We can choose $q(H)$ to be of any form. Typically we will choose a factorised distri-

bution as this makes inference more straightforward. In general H is a collection of random variables, $H = \{H_1, \ldots, H_N\}$. A factorised distribution for q will be of the form $q(H) = \prod_n q_n(H_n)$. We can maximise $\mathcal{L}(q)$ iteratively by updating each of the $q_n(H_n)$ consecutively. When we are considering the update for $q_n(H_n)$, we factorise $q(H)$ into two distributions, $q(H) = q_n(H_n)q_{\neg n}(H_{\neg n})$, where we define $q_{\neg n}(H_{\neg n}) = \prod_{n' \neq n} q_{n'}(H_{n'})$ to be the factorised distribution over all the $H_{n'}$ except H_n. Due to the factorised nature of $q(H)$ we can take our expectations in any order, that is, $\langle . \rangle_q = \langle \langle . \rangle_{q_{\neg n}} \rangle_{q_n}$. Using this factorisation with Equation 1.19 and ignoring terms that do not depend on $q_n(H_n)$ gives

$$\begin{aligned}
\mathcal{L}(q) &= \mathbb{H}(q) + \big\langle \log p(D, H) \big\rangle_q \\
&= \sum_{n'} \mathbb{H}(q_{n'}) + \Big\langle \big\langle \log p(D, H) \big\rangle_{q_{\neg n}} \Big\rangle_{q_n} \\
&= \mathbb{H}(q_n) + \Big\langle \big\langle \log p(D, H) \big\rangle_{q_{\neg n}} \Big\rangle_{q_n} + \text{constant} \\
&= -KL(q_n || \langle \log p(D, H) \rangle_{q_{\neg n}}) + \text{constant} \qquad (1.20)
\end{aligned}$$

If we define the updated distribution $q_n^*(H_n)$ by

$$\log q_n^* = \langle \log p(D, H) \rangle_{q_{\neg n}} - \log Z \qquad (1.21)$$

where Z is the constant that makes q_n^* a proper distribution then we know that this will minimise the KL-divergence in Equation 1.20 (it will be 0). So we can maximise $\mathcal{L}(q)$ with respect to q_n by updating using Equation 1.21.

Each iteration of variational inference consists of applying the updates for each q_n consecutively. Each update will increase $\mathcal{L}(q)$ which we know to be bounded. We continue iterating until the rate of increase falls below some pre-defined small threshold. When this happens we should be close to a local maxima of $\mathcal{L}(q)$ and we can expect our variational distribution, $q(H)$, to be close to a mode of the posterior, $p(H|D)$.

The updates for the form of variational inference given above are very similar to those for the EM algorithm. In fact, the EM algorithm can be seen as a special case of variational inference. If H_{EM} are the hidden variables in our model and $\theta = \theta_1, \ldots, \theta_K$ are the model parameters then from our variational inference point of view these are all hidden variables giving $H_{\text{var}} = \{H_{\text{EM}}, \theta\}$. If we use a factorised variational distribution, $q(H_{\text{var}}) = q(H_{\text{EM}}) \prod_k q_k(\theta_k)$, where each $q_k(\theta_k)$ is restricted to being of the form of a point estimate then variational inference and the EM algorithm are equivalent.

1.3.9 Hypothesis testing

In this section I define some terms commonly used in statistical hypothesis testing that I will use later on. The null hypothesis typically describes a default position. The alternative hypothesis typically describes an effect conjectured to be present. A p-value is the probability of observing a test statistic at least as extreme as the test statistic generated by the data if the null hypothesis is true. The E-value is often used when multiple tests are involved. The E-value is the expected number of times a test statistic as extreme as the one generated by the data would be observed assuming the null hypothesis is true. The E-value is the number of tests multiplied by the p-value of the most extreme statistic.

1.3.10 Classification

In this section I introduce some statistics and graphics that are commonly used to compare classification and prediction methods. Suppose we have a binary classifier that predicts if some data possess some property. For example, we may have a classifier that predicts if putative TFBSs are bound *in vivo* by a TF. When the classifier predicts the property is present, the prediction is called a positive prediction; conversely a negative prediction is made when the classifier predicts the property is absent. A true prediction is one that the classifier gets correct; conversely a false prediction is one the classifier gets wrong. In this way, if the predicted and correct classifications are known then every prediction is either a true positive (TP), false positive (FP), true negative (TN), or false negative (FN). The false positive rate (FPR) of the classifier is the ratio of false positive predictions made by the classifier to the potential false positives (in other words the number of negatives), $\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}}$. Similarly the true positive rate (TPR) is defined as the ratio of the true positive predictions to the potential true positives (the number of positives), $\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$. Some authors use sensitivity and specificity instead of TPR and FPR. The sensitivity is the same as the TPR and the specificity is defined as $1 - \mathrm{FPR} = \frac{\mathrm{TN}}{\mathrm{FP+TN}}$. The false discovery rate (FDR) measures how many positive predictions were incorrect. It is the ratio of false positive to positive predictions, $\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{TP+FP}}$.

A binary classifier often works by thresholding a score associated with each datum. All those data that score above the threshold are predicted as positives and the remainder are predicted as negatives. Each threshold in the range of the scores defines such a classifier. When the threshold is set low the classifier always predicts positively; conversely a high threshold makes every prediction a negative. Between these extremes there will be a range of classifiers that represent a compromise between the number of FPs and FNs.

The performance of such a method can be investigated by examining this compromise graphically and statistically.

Graphically, this is typically achieved through plotting a receiver operating characteristic (ROC) curve. The ROC curve is a plot of the FPR against the TPR of the classifiers as the threshold varies. A perfect classifier has a ROC curve that passes through (0,0), (0,1) and (1,1). A ROC curve that is a line from (0,0) to (1,1) represents a method that classifies randomly. Two example ROC curves are plotted in Figure 1.3.

Note that a ROC curve represents a traversal of the data sorted by their score. As we move along the curve we are including more data at successively lower scores. An ambiguity arises when the method scores positive and negative data equally: should the positive or negative examples at this threshold be considered earlier on the ROC curve? There are three solutions to this ambiguity. Optimistically and unreasonably we can suppose that the method is better at predicting TPs than FPs. Agnostically we could suppose positive and negative data are treated equally. Pessimistically and cautiously we can place the negative data earlier on the curve. In this thesis, I choose the latter option following the example of [Håndstad et al., 2011].

Statistically, a method's performance can be measured by its area under curve (AUC) statistic. This statistic is defined as the area under the ROC curve. A perfect method will have an AUC statistic of 1 and a random method will have an AUC statistic of .5. An AUC statistic can be interpreted as the probability that a randomly chosen positive datum will be scored more highly than a randomly chosen negative example.

Sometimes we are not interested in the performance of a method over its entire range of thresholds. We are often interested in how the method performs at higher thresholds on the data for which it is more confident are positive. In this case statistics such as the AUC50 are commonly used [Håndstad et al., 2011]. The AUC50 is defined as the area under the ROC curve bounded by the FPR corresponding to having predicted 50 FPs. This is the region of the ROC curve that evaluates the most confident predictions. However the AUC50 statistic does have some shortcomings. Firstly it is sensitive to the number of data. Imagine two data sets, one contains two copies of each datum contained in the other. Any given method will have the same AUC on both data sets (if tied scores are handled appropriately). However, in general the AUC50s for the same method on the two data sets will vary by a factor of the order of two. Secondly 50 is a somewhat arbitrary choice. It is motivated by the idea that an experimenter might find no more than 50 FPs acceptable. However, for modern high-throughput data sets this represents a tiny FPR. I present a similar statistic to the AUC50 that is defined by a FPR threshold instead of a count of FPs later in this thesis. This statistic does not have these two shortcomings.

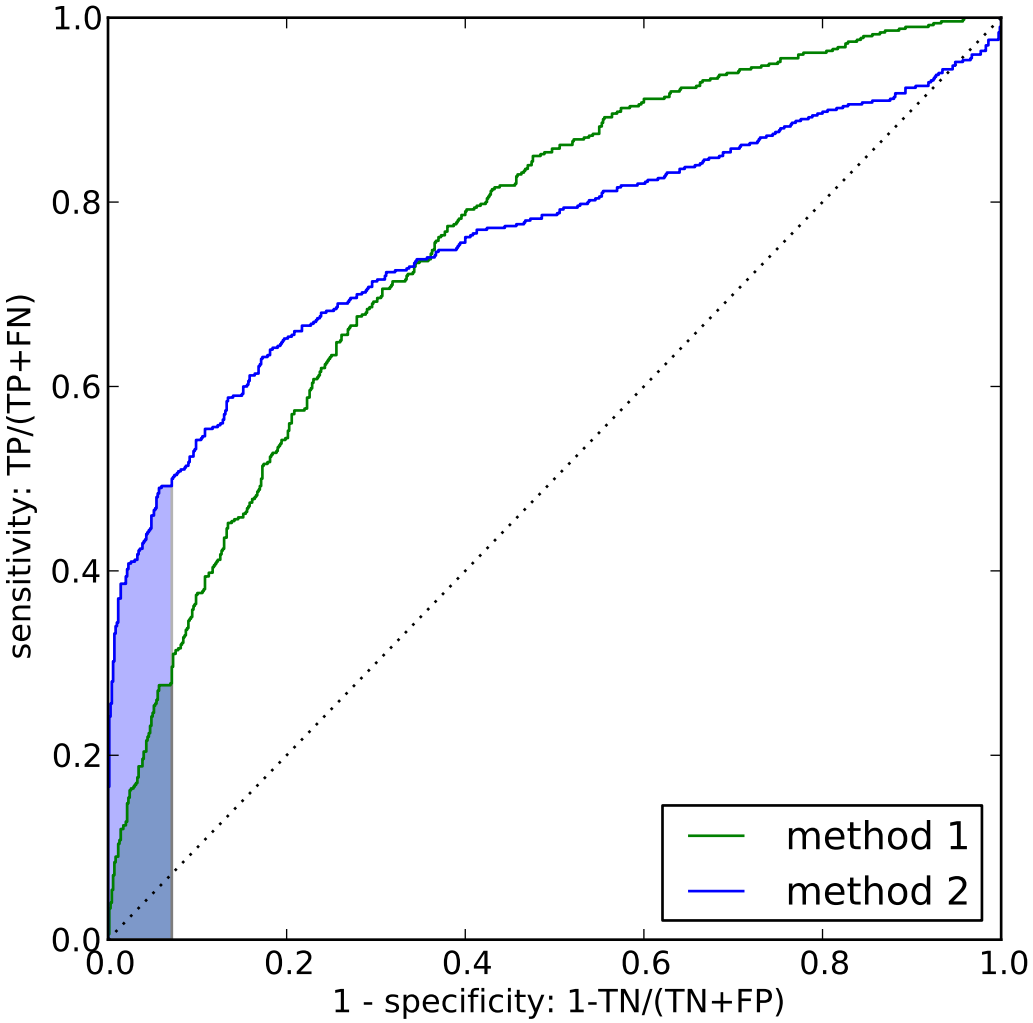


Figure 1.3: ROC curves for two fictional classification methods on simulated data. The AUC statistic for each method is the area under the whole curve. The AUC50 statistic is the area under the curve that corresponds to the first 50 FPs and is shaded in the plot. There were 700 negative data in the simulation so this area is bounded by a FPR of $\frac{50}{700}$. Method 1 has an AUC of 0.761 and an AUC50 of 0.013. Method 2 is better at low sensitivities and worse at high sensitivities. By chance method 2 has the same AUC statistic of 0.761. However method 2 has an AUC50 statistic of 0.03 which demonstrates superior performance at higher thresholds. The expected performance of a classification method based on random scores is shown as the dotted line (AUC=.5).

1.4 Models of transcriptional regulation

1.4.1 Motivation

Why do we model transcriptional regulation? Models have at least three useful aspects: they let us test how well data fit different hypotheses; they can help reduce the complexity of the system we are investigating and they allow us to predict behaviour in systems for which we do not have data. For the first aspect we can construct a model that encapsulates our hypothesis. The fit of the model to the data can be used as a proxy for our belief in the hypothesis. The second aspect is useful when models provide a summary or high-level view of complex systems by hiding or ignoring largely irrelevant low-level details. The third aspect is useful when we wish to extrapolate from what we know about a particular system to other related systems.

As well as being useful, computational models are necessary. We are compelled to use them due to the sheer volume of data generated by modern biological techniques. It is no longer possible to analyse these data by hand. For instance, a typical ChIP-seq experiment might report that a TF binds to regions spanning 10Mb of the genome. No biologist can inspect all these regions manually. To use these data effectively we need to build models with which we can infer the interaction between the TF and the DNA. In this thesis we focus on models of the sequence preferences of TFs and on models of cooperative effects between TFs.

1.4.2 Representations of binding sites

As described above, the sequences at TFBSs for a particular TF often exhibit significant variability. Models that characterise these TFBSs have to incorporate this variability [Stormo, 2000]. Knowledge of these preferences is useful in many ways. In the first place, it allows prediction of TFBSs from sequence data alone. These predictions can provide biologists with novel candidate genes that may be regulated by the TFs in the network they are studying. Conversely a biologist can scan CRMs in a network for instances of TFBSs not known to be associated with that network. In this way putative new TFs in the regulatory network are predicted. TFBS predictions can be used with other sequence based data such as single nucleotide polymorphisms (SNPs) to help discover the mechanisms by which genetic variants cause observed phenotypes. Homologous TFs in closely related species often have similar protein structures and share similar binding preferences. Knowledge of the binding preferences of a TF in one species can be applied across a clade or even greater evolutionary distances. In addi-

IUPAC code	Base
A	adenine
C	cytosine
G	guanine
T	thymine
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base

Table 1.1: The IUPAC code that incorporates ambiguity over bases.

tion, understanding the exact locations of TFBSs can help us understand combinatorial transcriptional regulation [Reid et al., 2009].

Consensus sequences

The most basic models of TF binding sequence preferences are consensus sequences. In their simplest form they describe the sequence that the TF is most likely to bind. The likelihood of the TF binding to alternate sequences is normally quantified by the number of mismatches to the consensus sequence.

Consensus sequences can be used with an alphabet representing the four DNA bases, `A`, `C`, `G` and `T`. They can also be used with an expanded alphabet which explicitly represents ambiguity in which base is preferred at each position. Typically the IUPAC alphabet is used (see Table 1.1). For example, the consensus sequence of TBP is `TATAAAA` in the DNA alphabet but variability in the TFBSs can be represented by the consensus sequence `TATAWAW` in the IUPAC alphabet. `W` represents `A` or `T`.

Position weight matrices

A position weight matrix (PWM) parameterises a probability distribution over words of a given length, say W. It is worth noting that some authors use the term PWM to refer to a scoring matrix (discussed in Section 2.1). PWMs are the archetypal method

for modelling the sequence-specific binding preferences of TFs. The distribution at each position is modelled as a discrete distribution over the four possible bases. PWMs treat each position in the TFBSs independently. That is, the base that occurs at position w_1 in a TFBS has no bearing on the base occurring at position w_2 if $w_1 \neq w_2$. When PWMs are used to model TF-DNA interactions this independence is equivalent to assuming that the binding energy of the TF-DNA contributes additively across positions.

A PWM is a probabilistic model and is parameterised by W discrete distributions, $\theta = \{\theta_1, \ldots, \theta_W\}$, over the four possible bases at each of its W positions. θ_{wb} is the probability of observing base b at position w in a TFBS. Out of all the models that assume position independence, PWMs are the most general. Consensus sequences can be seen as special cases of PWMs. PWMs exactly quantify the probability of the preferred base at each position and how unlikely any specific deviation is. The PWM models a W-mer, $X = x_1 \ldots x_W$ as

$$p(X|\theta) = \prod_{w=1}^{W} \theta_{wx_w} \tag{1.22}$$

PWMs have been shown to be more sensitive than consensus sequences when used to predict translational initiation sites [Stormo et al., 1982].

The information content of a PWM is a measure of how specific the TF's binding preferences are. There is more than one definition of the information content. I use the version where the information content, IC, is taken to be the KL-divergence between the PWM's distribution and a 0-order genomic background distribution, ϕ,

$$IC = KL(\theta||\phi) = \sum_{w} \sum_{b \in \{A,C,G,T\}} \theta_{wb} \log \frac{\theta_{wb}}{\phi_b} \tag{1.23}$$

This version is well founded and well established [Stormo, 1998]. The main alternative is due to Schneider and is equivalent when the genomic background distribution is uniform [Schneider, 1997].

PWMs are most often represented graphically by sequence logos [Schneider and Stephens, 1990]. The height of the sequence logo at each position gives a measure of the conservation at that position and the distribution over the four possible bases is represented by their relative heights. See Figure 1.4 for an example.

PWMs are motivated biophysically: Berg and von Hippel showed that the binding energy of the TF-DNA interaction is proportional to the logarithm of the frequencies of the bases [Berg and von Hippel, 1987]. Their analysis only applies to genomes with a uniform base composition but in practice the base compositions of most genomes are relatively close to uniform.

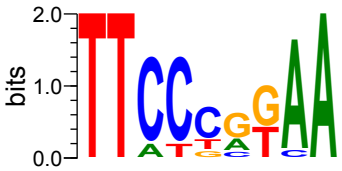


Figure 1.4: A sequence logo for the TF Stat5. We can see the consensus sequence is `TTCCCGGAA` by reading the bases on top at each position. The bases at each position are ordered top-down by their frequencies. The height of the logo at each position represents its information content, a measure of its conservation, which is measured in bits. The relative height of each base at each position represents their frequencies.

To give a feel for the molecular mechanics of TF-DNA binding I show an example with its associated sequence logo in Figure 1.5.

Only a relatively small fraction of TFs have had their sequence preferences characterised. For instance there are PWMs for a few hundred human TFs out of an estimated several thousand. The databases TRANSFAC [Matys et al., 2003], JASPAR [Sandelin, 2004, Portales-Casamar et al., 2009] and UniPROBE [Newburger and Bulyk, 2009] hold the most widely used collections of PWMs. The PWMs in these databases are mainly for TFs from *Homo sapiens* and such model organisms as *Saccharomyces cerevisiae*, *Drosophila melanogaster* and *Mus musculus*. JASPAR and UniPROBE are open-access databases, TRANSFAC has a commercial license.

Given a set of PWMs that represent the binding preferences of TFs under study, an investigator can scan genomic regions for matches to the binding preferences. There are several problems with this approach: the regions in which regulatory TFBSs are located are not normally known in advance; due to the size of the genome, algorithms that find putative binding sites are known to generate many false positives; and, unfortunately, JASPAR and TRANSFAC do not contain PWMs for all TFs of interest.

More complex models of TF specificity

The independence assumption inherent in PWMs has been shown to be realistic in some cases [Sarai and Takeda, 1989, Takeda et al., 1989, Desjarlais and Berg, 1994, Lustig and Jernigan, 1995] but there is evidence to suggest that it is not always justified [Man and Stormo, 2001, Bulyk et al., 2002]. Some models have been proposed for position interdependencies [Zhou and Liu, 2004, Barash et al., 2005, Ben-Gal et al., 2005, Naughton et al., 2006, Sharon et al., 2008]. On the other hand, it has been shown that the independence assumption is a close approximation for some cases where interdependencies exist [Benos et al., 2002, Man et al., 2004]. The advent of PBMs provided more data relevant to this

ongoing discussion. Badis et al. use it to argue that many TFs have multiple modes of binding [Badis et al., 2009]. Zhao and Stormo have argued that simple models are sufficient for most TFs [Zhao and Stormo, 2011]. For the time being the PWM remains the model of choice for TFBSs.

Several approaches model the idea of multiple modes of TF binding (Section 1.1.4). Hannenhalli and Wang use a mixture model approach [Hannenhalli and Wang, 2005]. Georgi and Schliep adapt this approach by allowing the model to share distributions at positions in the TFBS that do not vary across the mixture components [Georgi and Schliep, 2006]. Bais et al. present an approach that looks at spaced dyads [Bais et al., 2011]. One half of the dyad may have several modes depending on its cofactor.

The notion that TFBSs can vary in length has been relatively under-explored. In a publication outside of this thesis my coauthors and I examined some ChIP data for evidence of gapped motifs [Reid et al., 2010].

As mentioned in the discussion above on weak binding sites (Section 1.1.4), TFs belong to structural families. Sandelin and Wasserman were the first to model these families with familial binding profiles [Sandelin and Wasserman, 2004]. Piipari et al. have also developed a model for the binding preferences of TF families [Piipari et al., 2010]. We note that their metamotif model could also be used to model multiple modes of binding by the same TF.

1.4.3 Modelling genomic sequences

The genomic sequences of many species are known to contain dependencies between nearby base pairs [Chor et al., 2009, Zhou et al., 2008]. One of the most prevalent examples of this are CpG islands. CpG islands are regions of the genome with higher than expected `CG` dinucleotide content. Other regions of the genome are subject to CpG suppression where they have lower than expected `CG` dinucleotide content. CpG islands have been implicated in the determination of chromatin structure [Thomson et al., 2010] and may have an important role in transcriptional regulation.

Most algorithms for sequence analysis make the simplifying assumption that the genome can be modelled as a 0-order Markov model ignoring any higher order dependencies. Whilst this is a practical choice for many applications, several authors claim more complex background models improve their methods [Thijs et al., 2001, Aerts et al., 2003, Turatsinze et al., 2008, Thomas-Chollier et al., 2011]. In this thesis I will concentrate on 0-order Markov models but I will try to point out where more complex models could be applicable. More complex models are not restricted to higher order Markov models.

Hidden state space models such as hidden Markov models provide natural models of the mosaic structure present in many genomes [Down, 2005].

1.4.4 Associating regulatory regions with genes

Discovering evidence that a TF binds to the genome through sequence analysis or experimentation does not normally reveal which gene(s) it regulates. Models of transcriptional regulation commonly assume that CRMs regulate the expression of the closest gene. This may be a typical scenario but it is well known that CRMs can act over great distances and across chromosomes (see Section 1.1.4).

1.4.5 Algorithms to learn binding site representations

Modern high-throughput experimental methods such as ChIP-seq, ChIP-chip and DamID are able to pinpoint regions in the genome where particular TFs bind *in vivo*. However, the resolution of these techniques is still an order of magnitude or two larger than a typical TFBS [Gilchrist et al., 2009]. There remains a need to determine the binding sequence preferences of TFs and hence the exact locations of TFBSs from these data sets. This task of inferring the sequence preferences of a TF from such a set of regions is termed motif finding.

A typical high-throughput experiment might generate a data set of thousands of sequence fragments. Each fragment could be hundreds of base pairs long. The sequence preferences of a TF are relatively short, typically eight to twelve base pairs. Mismatches to the preferred bases are common in TFBSs. Determining these sequence preferences from the few binding sites in the fragments is a difficult problem. However, much effort has been dedicated to this motif finding problem and many algorithms and softwares exist for this purpose. The area has been reviewed several times [Hu et al., 2005, D'haeseleer, 2006a, MacIsaac and Fraenkel, 2006, Das and Dai, 2007, Håndstad et al., 2011].

Most motif finders can be broadly categorised as either combinatorial or probabilistic. Combinatorial motif finders search for consensus sequences. TFBSs are predicted on the basis of the number of mismatches with these consensus sequences. Probabilistic motif finders typically infer PWMs. Most of the probabilistic motif finders use either the expectation-maximisation (EM) algorithm [Dempster et al., 1977,D'haeseleer, 2006b] or a Gibbs sampling algorithm [Geman and Geman, 1984] for inference. Examples of motif finders that use the EM algorithm include [Lawrence and Reilly, 1990, Bailey et al.,

1994, Blekas et al., 2003, Moses et al., 2004a, Prakash et al., 2004, Sinha et al., 2004, Qi et al., 2005, MacIsaac and Fraenkel, 2006, Li, 2009].

The volume of available TF binding location data is rapidly increasing. Both the number and the size of data sets generated by techniques such as ChIP-chip, ChIP-seq, and DamID continue to grow. Unfortunately the run-time of most motif finders is at least linear in the size of the data. In our experience most motif finders are far too slow for such large data sets of sequences. Whilst it may be possible to let the motif finder run for several days, invariably the user would like to fine-tune parameters. This may involve several runs which makes motif finding impractical.

1.5 The rest of this thesis

The main contributions of this thesis are presented in the next three chapters.

Chapter 2 introduces a novel algorithm to scan DNA sequences for TFBSs. The algorithm incorporates phylogenetic information using an averaging of evidence technique. The core of the algorithm models billboard enhancers. In the maximal chain extension to the core algorithm, conservation of the order of TFBSs across species is explicitly modelled without the need for a multiple alignment of the sequences under consideration. This extension models enhanceosomes. A small example is given where the core algorithm was used to investigate a regulatory network in the mouse embryo. An empirical evaluation of the strengths and weaknesses of the core algorithm relative to three other comparable methods is presented.

Chapter 3 presents a novel and efficient approximation to one of the most popular motif finders, MEME. This approximation makes the application of the MEME algorithm possible to the large data sets generated by modern high-throughput experiments. The approximation is based on suffix trees which have been used in combinatorial motif finders. It is to the best of my knowledge the first application of suffix trees to probabilistic motif finders. Theoretical and empirical analyses of the approximation's properties are given.

Chapter 4 presents an application of a non-parametric hierarchical Bayesian model commonly used in document-topic modelling to model combinatorial effects in transcriptional regulation. The model discovers several well characterised sets of interacting TFs in an unsupervised fashion. The model also discovers some hitherto uncharacterised interactions.

Chapter 5 is a discussion chapter. It summarises the contributions made by the preceding three chapters. Some arguments are presented in favour of probabilistic models.

Possible ways to combine the application of the novel methods presented in the thesis are discussed. Types of data that might be used for this integration are identified. Some possible avenues for future research are presented.

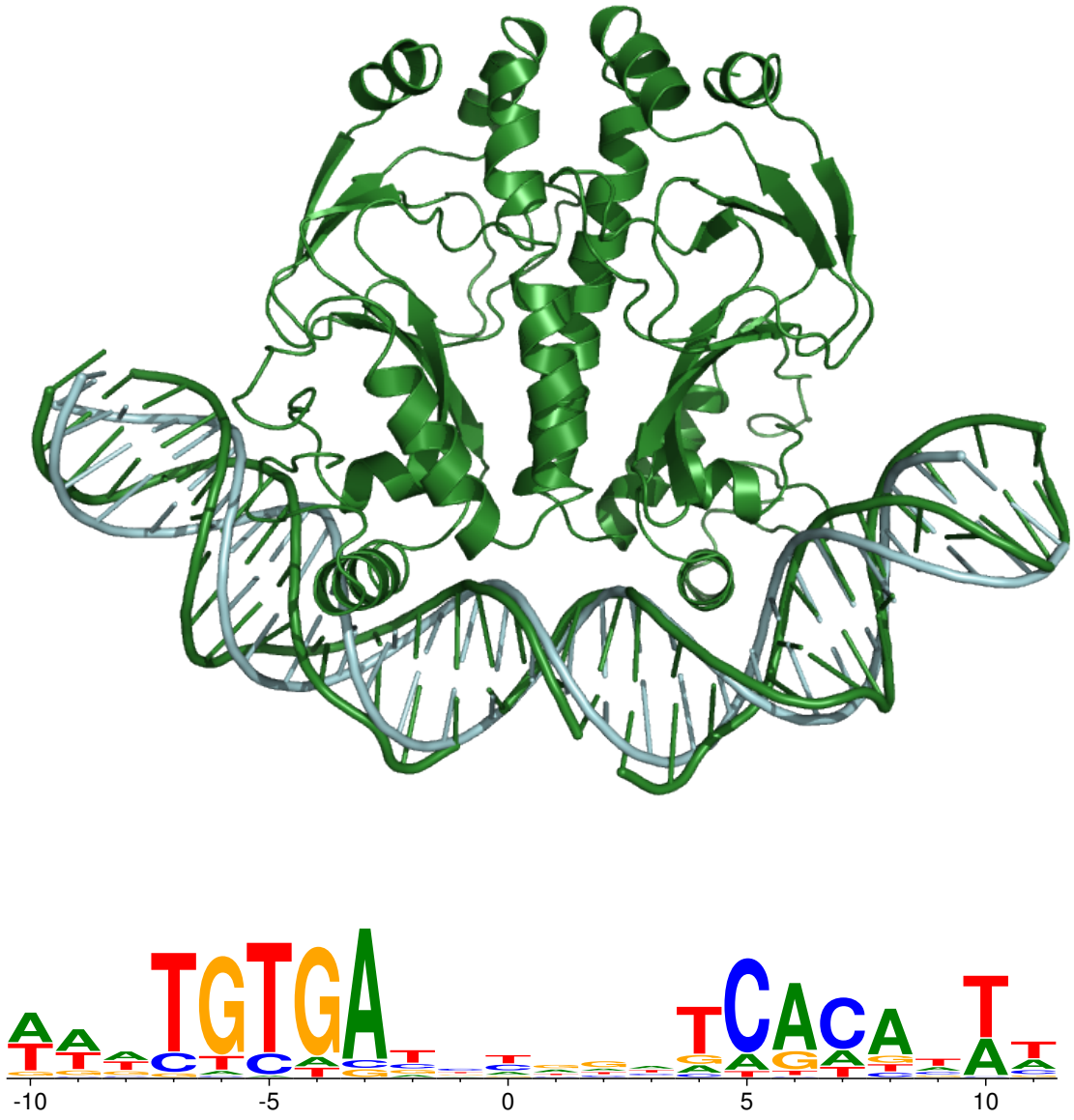


Figure 1.5: The catabolite activator protein (CAP) exists as a homodimer. It binds DNA with two DNA recognition helices which insert into consecutive turns of the DNA's major groove. *Top*: A molecular model of the structure of a CAP homodimer binding to DNA. The model was generated by the 3D-DART web server [Dominguez et al., 2003]. *Bottom*: A sequence logo representing 58 binding sites for the CAP homodimer [Robison et al., 1998]. Notice the two halves of the PWM, each one represents the binding preferences of one of the helices. They are spaced 11 base pairs apart, consistent with the distance between consecutive turns of the DNA's major groove. The two halves are approximate reverse complements of each other, reflecting the orientation of the helices in the homodimer.

Chapter 2

Predicting binding sites

Once we have characterised the binding preferences of a TF, a natural question to ask is: can we predict where in a DNA sequence the TF binds? This prediction task is called PWM scanning or motif scanning. Several prediction approaches have been developed.

2.1 Sequence based predictions

The simplest prediction methods require no more data than the sequence itself and of course the binding preferences of the TFs. There are a number of methods that tackle this problem from various angles. Here we summarise the most popular methodology. Later we will describe specific methods and how they relate to this basic framework. In the basic framework each position in the TFBS is treated independently. This is equivalent to the assumption that the binding energies are additive across the positions. In this formulation the binding preferences are summarised by a position specific scoring matrix (PSSM) , ψ. A PSSM is similar to a PWM but the values at each entry represent the score given to that base at that position in the TFBS rather than the frequency with which that base occurs. The total score, $S_\psi(x)$, for the putative TFBS, $x = x_1 \dots x_W$, is the sum of these scores.

$$S_\psi(x) = \sum_{w=1}^{W} \psi_{w,x_w} \tag{2.1}$$

Some methods normalise the score

$$S'_\psi(x) = \frac{S_\psi(x) - S_\psi^{\min}}{S_\psi^{\max} - S_\psi^{\min}} \tag{2.2}$$

where

$$S_\psi^{\max} = \arg\max_x S_\psi(x) \qquad \text{and} \qquad S_\psi^{\min} = \arg\min_x S_\psi(x)$$

so that $S' \in [0, 1]$.

A few methods stop here and report the score, S or S', however most calculate a p-value. The p-value summarises the chances of observing such a score, S, by chance. It is based on the distribution of the score, S, under some background genomic distribution. The background distribution is almost always taken to be a 0-order Markov model as this greatly simplifies the calculation of the p-values.

For many TFBS prediction tasks a binary output is required thus the score or p-value is thresholded. The threshold used may be PWM specific. PWM-specific thresholds are necessary if the score has been normalised. To see why, imagine two TFBSs, the first perfectly matches the very vague sequence preferences of one TF, the second TFBS perfectly matches the highly specific sequence preferences of another TF. Clearly the second TFBS is a much stronger candidate but they both score 1 under any normalised scoring scheme.

Popular p-value based prediction algorithms include FIMO [Grant et al., 2011], patser (implemented by Gerald Hertz, no reference available but can be downloaded from `http://stormo.wustl.edu/resources.html`) and matrix-scan-quick [Thomas-Chollier et al., 2011].

2.1.1 Pseudocounts

Most methods add pseudocounts to the PWMs before use. This ensures there are no forbidden bases and also makes sure PWMs built from an alignment of just a few sites are not overly confident. There has been some disagreement about the best value to use. Early authors used small values around 0.01. Later authors have used larger values of 0.25 or 1. Nishida et al. have made a study of the effect of pseudocounts on PWMs and predicting TFBSs [Nishida et al., 2008]. They recommend using a pseudocount of 0.8 in practice.

2.1.2 Log-likelihood scoring functions

The most prevalent scoring function is the log-likelihood ratio. To the best of my knowledge, the first example of its use is MATRIX SEARCH [Chen et al., 1995]. The log-likelihood ratio is the difference between the log-likelihood of the TFBS given the frequencies in the PWM and the background model. More exactly, suppose we have a PWM, θ, representing the frequencies of the bases in the binding sites and a 0-order Markov background model, ϕ. So $\theta_{w,b}$ is the probability of observing base b at position

w in a binding site and ϕ_b is the probability of observing base b in the genome. Our PSSM, ψ, is defined as

$$\psi_{w,b} = \log \frac{\theta_{w,b}}{\phi_b} \tag{2.3}$$

Here our score $S_\psi(x) = \sum_{w=1}^{W} \psi_{w,x_w}$ is the log-likelihood ratio.

The concept behind the likelihood ratio approach is clear: the two likelihoods are compared and the evidence in favour of one model over the other is the score. More generally treating the TFBS as a whole we can consider the log likelihood ratio of the PWM model to a background model of a higher order, Then the score, S_{LL}, is given by

$$S_{LL} = \log \frac{p(x|\theta)}{p(x|\phi)} \tag{2.4}$$

where $p(x|\theta)$ is the probability of x under the PWM model parameterised by θ and $p(x|\phi)$ is the probability of x under the background model parameterised by ϕ. When $p(x|\phi)$ is a 0-order model this reduces to the case above (Equation 2.3). As mentioned in Section 1.4.3 most genomes do not fit a 0-order model well. This ability to introduce more complex background models is an appealing aspect of this scoring method. These background models can include context-sensitive effects, for example whether the sequence is inside a CpG island. Neither does the model have to be a Markov model, any model can be used.

The log-likelihood ratio also has a pleasing biophysical interpretation. Berg and von Hippel used statistical mechanics theory to show that the log-likelihood ratio is proportional to the binding energy of the TF-DNA interaction under a set of simplifying assumptions [Berg and von Hippel, 1987]. The assumptions are that the background composition of the genome is uniform, that is that each base has an equal *a priori* probability and that the positions in the sequence contribute independently to the binding energy.

Stormo and Fields developed a separate justification for using the log-likelihood ratio as a measure of binding energy [Stormo and Fields, 1998]. Their justification applies to non-uniform genomes that are well represented by 0-order Markov models. Their starting point is a set of high-affinity sites for which the binding energies are unknown. Again assuming the positions contribute independently they show that the log-likelihood ratio is a maximum likelihood estimate of the binding energy between the TF and the sequence.

Posterior probability of binding

In a Bayesian framework we can treat the log-likelihood score, S_{LL}, as a log Bayes factor and estimate the probability that x is a TFBS. Denoting the event that x is a TFBS by B we have by Bayes' theorem

$$p(B|x) = \frac{p(x|B)p(B)}{p(x)} \tag{2.5}$$

and

$$p(\overline{B}|x) = \frac{p(x|\overline{B})p(\overline{B})}{p(x)} \tag{2.6}$$

where $\overline{B}$ is the event that x is not a TFBS. Dividing the two preceding equations to remove $p(x)$ and relating this to Equation 2.4 by $p(x|B) = p(x|\theta)$ (the likelihood of x under the PWM model) and $p(x|\overline{B}) = p(x|\phi)$ (the likelihood of x under the genomic background model) we have

$$\frac{p(B|x)}{p(\overline{B}|x)} = \frac{p(B)}{p(\overline{B})}e^{S_{LL}} \tag{2.7}$$

As B and $\overline{B}$ are mutually exclusive we have $p(\overline{B}) = 1 - p(B)$ and $p(\overline{B}|x) = 1 - p(B|x)$. Defining $T = \frac{p(B)}{1-p(B)}e^{S_{LL}}$ we have

$$p(B|x) = \frac{T}{1+T} \tag{2.8}$$

This is the posterior probability that x is a binding site under this model. It depends on S_{LL} and a prior, $p(B)$, that can be specified by the user of the method. This probabilistic approach is not as popular as using p-values although some recent methods do use it, for example MotEvo [Arnold et al., 2012]. Later I will develop this idea further.

A typical PWM scanning task might use hundreds of PWMs and sequences of hundreds of base pairs upwards. Just scanning 100 putative TFBSs with 100 PWMs will result in 10,000 tests. We are normally interested in ranking the putative TFBSs according to how strong our belief is that they are real TFBSs. Hence some method to compare the results of each test across different PWMs is generally required. p-value based scoring approaches use a threshold to decide which putative TFBSs are rejected as binding sites. This is a binary decision and whilst the p-value provides some measure of how strong the evidence, it is not clear that a direct comparison of p-values generated by different PWMs is justifiable. In contrast, this Bayesian framework gives us a posterior probability that the putative TFBS is actually a TFBS. Under the assumptions of the model, the posterior probabilities are comparable and we can use them to determine which TFBSs are stronger candidates. Additionally, whilst we can use a threshold with

p-value methods to make binary predictions, it is harder to carry forward any uncertainty about these predictions to later stages of an analysis. Quantification of the uncertainty about the predictions is a natural consequence of the Bayesian approach as this is exactly what the posterior represents. Admittedly this is subjective due to our specification of a prior but calibration of the method is possible given the right data.

2.1.3 The MatInspector and MATCH methods

Early PWM scanning algorithms such as MatInspector [Quandt et al., 1995, Cartharius et al., 2005] and MATCH [Kel et al., 2003] use *ad hoc* scoring functions where the sequence is matched against the PWM using heuristics that depend on the degree of conservation at each position in the PWM. We describe both in more detail mainly for their historical relevance.

Quandt et al. developed the MatInd and MatInspector programs to predict TFBSs in sequences based on PWMs in TRANSFAC [Quandt et al., 1995, Cartharius et al., 2005]. The MatInd program takes as input a PWM or a set of sequences representing the TFBSs and calculates a conservation score, $C_i(w)$, for every position w in the PWM or sequence alignment

$$C_i(w) = \frac{100}{\log 5}\left[\sum_{b\in A,C,G,T,\text{gap}} \theta_{wb}\log\theta_{wb} + \log 5\right] \tag{2.9}$$

so that $0 \leq C_i(w) \leq 100$. Note that $C_i(w) = 0$ when all $\theta_{wb} = \frac{1}{5}$ and $C_i(w) = 100$ when exactly one $\theta_{wb} = 1$. These $C_i(w)$ are used to weight the relative contributions of different positions in the TFBS to the overall score. Positions that are highly conserved contribute more to the score than vague positions. To assess a putative TFBS, the MatInspector program calculates a matrix similarity score, mat_sim, defined as

$$\text{mat\_sim} = \frac{\sum_{w=1}^{W} C_i(w)\theta_{wx_w}}{\sum_{w=1}^{W} C_i(w)\max_{b\in\{A,C,G,T,\text{gap}\}}\theta_{wb}} \tag{2.10}$$

The MATCH algorithm developed by Kel et al. [Kel et al., 2003] is similar to MatInspector in that it uses a heuristic scoring scheme that bears some resemblance to a likelihood ratio weighted by conservation information. Kel et al. define the matrix similarity score, mSS, as

$$mSS = \frac{\text{Current} - \text{Min}}{\text{Max} - \text{Min}} \tag{2.11}$$

where Current, Min and Max are defined as

$$\text{Current} = \sum_{w=1}^{W} I(w)\theta_{wx_w} \tag{2.12}$$

$$\text{Min} = \sum_{w=1}^{W} I(w) \min_{b\in\{A,C,G,T\}} \theta_{wb} \tag{2.13}$$

$$\text{Max} = \sum_{w=1}^{W} I(w) \max_{b\in\{A,C,G,T\}} \theta_{wb} \tag{2.14}$$

where $I(w)$ is the information content at position w in the PWM

$$I(w) = \sum_{b\in\{A,C,G,T\}} \theta_{wb} \log(4\theta_{wb}) \tag{2.15}$$

$I(w)$ in the MATCH algorithm is analogous to $C_i(W)$ in the MatInspector algorithm. Overall the MatInspector algorithm differs in that it can be used with gaps in the alignment but both implicitly assume a uniform 0-order background model. For many genomes this is not a good approximation. Otherwise the MATCH algorithm and the MatInspector algorithm are very similar: c.f. Equations 2.10 and 2.11. Kel et al. claim [Kel et al., 2003] that the MATCH algorithm is more discriminative than the similar MatInspector algorithm without showing supporting data. Kel et al. also claim [Kel et al., 1999] that methods such as MatInspector and MATCH that incorporate conservation information into the scoring system are to be preferred over scoring methods based on the likelihood ratio although they do not show data supporting this claim. The MATCH algorithm is associated with the TRANSFAC PWM database. Kel et al. have calculated different cut-off thresholds for each PWM in TRANSFAC. These are intended for users who wish to minimise the false positive rate, the false negative rate or a combination of both.

2.1.4 *p*-value calculations

As most methods report p-values there has been much research into algorithms to calculate these p-values efficiently. The p-value of a score is the probability of seeing this score or better in W-mers drawn from a background model. Again a 0-order background model is typically assumed. Staden devised the first method to tackle this problem [Staden, 1989]. He came up with an efficient numerical method for calculating the probabilities of finding motif matches in sequences. Claverie and Audic gave a method for calculating the distribution of scores for a given PSSM and 0-order genomic background frequen-

cies [Claverie and Audic, 1996]. Beckstette et al. developed a dynamic programming technique, ESAsearch, to calculate a suitable score threshold for a PSSM given a desired p-value or E-value threshold [Beckstette et al., 2006]. This threshold is based empirically on the matches to the PSSM in a large sequence database. Similarly Schones et al. present a method called STORM that uses gapped data structures to empirically estimate the p-values based on a large database of known promoter regions [Schones et al., 2007]. Touzet and Varré show that the problem of calculating p-values is NP-hard and present an improved calculator [Touzet and Varré, 2007]. The difficulty of calculating p-values is explained by the exponentially large number of potential TFBSs (4^W for TFBSs of size W). Each TFBS can have a different score and a naïve algorithm would enumerate all 4^W possible TFBSs. Pizzi et al. present a collection of algorithms to scan for TFBSs by p-value in large sequence sets [Pizzi et al., 2011]. Beckstette et al.'s PoSSuM software is also designed for large sequence sets and Pizzi et al. present a thorough comparison of speed. They are able to scan the human genome for the entire JASPAR motif set (123 PSSMs) in 18 minutes at a threshold of $p = 0.0001$.

2.2 Integrative approaches

Despite the work that has gone into the methodology of sequence based predictions it is well known that they suffer from high false positive rates which are difficult to control. After all most genomes are large and a lot of PWMs are reasonably degenerate. Thus one would expect to find a large number of spurious matches to any given PWM in a scan of sequences of any great length. All the above approaches make predictions using PWMs and sequence data alone. Many authors have attempted to improve the predictions by integrating other data. Plenty of other types of such data are available. For example: functional binding sites are likely to be conserved across species; some TFs are more likely to bind in the presence of cofactors; clusters of TFBSs in the genome can signify that the region is likely a regulatory region increasing the likelihood that other TFs bind there; epigenetic marks can be correlated with TF binding; TFBSs are more likely to be located near TSSs. In this section I review these concepts and highlight some of the methods that use them to predict TFBSs.

TFBSs are known to co-locate to enable interactions between the TFs. Several algorithms capitalise on this by looking for clusters of TFBSs. Cister uses a hidden Markov model to search for clusters of TFBSs [Frith et al., 2001]. Cluster-Buster [Frith et al., 2003] is the third generation of the Cister algorithm and uses a simplified model to improve its run-times. Rajewsky et al. present segmentation algorithms that locate enhancers in the *Drosophila* genome [Rajewsky et al., 2002].

Some TFBS prediction methods [Moses et al., 2004b, Siddharthan et al., 2005, Kheradpour et al., 2007, Hawkins et al., 2009, Xie et al., 2009] use alignments to related species to increase their predictive power. I describe these in more detail in the following Section 2.2.1 as they are relevant to the evaluation of the algorithm presented in this chapter.

TFBS prediction methods that treat each putative TFBS location independently can predict impossible overlapping binding sites. Dynamic programming approaches [Wasson and Hartemink, 2009] can overcome this deficiency. They model all potential DNA binding proteins and locations simultaneously. In this way they explicitly model limited access to DNA. These dynamic programming approaches can also take account of the effect of multiple weak binding sites [Rajewsky et al., 2002, Roider et al., 2007].

Recently experimental evidence for epigenetic effects has been integrated into prediction algorithms. CENTIPEDE [Pique-Regi et al., 2010] is an archetypal state-of-the-art TFBS predictor. It integrates information such as sequence, conservation, distance to TSS, activating and repressing histone modifications and DNase I cuts.

Cuellar-Partida et al. present a method to integrate epigenetic data into the motif scanning process as position-specific priors [Cuellar-Partida et al., 2012]. Their method is quite general in that any location data can be used to generate these priors. They tested their method using histone modification and DNase I hypersensitivity data. Their method is presented favourably in a comparison against the more complicated model of CENTIPEDE.

Ernst et al. provide another typical method [Ernst et al., 2010]: they use data such as sequence, conservation, estimated DNA melting temperature, GC-content, DNase I hypersensitivity, and histone modifications.

MotEvo [Arnold et al., 2012] is a method that combines the features of other tools that predict well. MotEvo incorporates an explicit evolutionary model; an unknown functional element concept; an enhancer predictor; and dynamic programming to model steric hindrance and weak binding sites.

2.2.1 Phylogenetic methods

Methods for TFBS prediction that use phylogenetic information can be broadly categorised four ways: alignment-free methods; simple alignment methods; phylogenetic motif model (PMM) methods and branch length score (BLS) methods. All methods use a set of related sequences. The sequence in the primary species under consideration is termed the central sequence. The other sequences are assumed to be from related species. These are termed the related sequences.

Alignment-free methods

Alignment-free TFBS prediction methods do not require a multiple alignment of the sequences being considered. Typically they predict TFBSs in the central sequence and modify these predictions based on an analysis of the related sequences. As they do not use an alignment, they are not subject to problems of mis-alignment and can handle situations where TFBSs have been lost and regained through evolution. They can also integrate evidence from multiple weak TFBSs in related sequences rather than relying on one strong TFBS to be aligned exactly with the site in the central sequence. On the other hand, a set of aligned TFBS predictions is strong evidence that a TFBS exists as non-neutral rates of evolution suggest the site is functional.

PhyloScan [Carmack et al., 2007,Palumbo and Newberg, 2010] is an algorithm that uses phylogenetic conservation and TFBS clustering to detect regulons in bacteria. PhyloScan uses p-value combination techniques to integrate the information from related sequences. It can work on aligned or unaligned sequences. It is primarily used for analysis of bacterial sequences.

Simple alignment methods

Simple alignment methods typically score a PWM in each sequence in an alignment. Scores in the central sequence are updated based upon scores in nearby regions of the alignment in the related sequences. For example Håndstad et al. introduced a weighted sum (WS) method [Håndstad et al., 2011]. In this method the scores in the central sequence were updated by adding half of the maximum score obtained in a window surrounding the aligned TFBS in each of the related sequences. The window was defined as the aligned TFBS extended by 15bp in both directions.

Phylogenetic motif models

A PMM is a generalisation of a PWM from a single sequence to a multiple alignment. Whereas a PWM models TFBSs in a single sequence, a PMM models TFBSs across all the sequences in an alignment. A PWM of width W is a matrix of size $W \times 4$, a PMM for a multiple alignment of N sequences is a tensor of size $W \times N \times 4$. The PMM provides the frequency of each possible base at each position of every sequence in the alignment. PMMs are normally applied in an analogous manner to PWMs. That is, the log-likelihood of a segment of a multiple alignment under the PMM is compared to the log-likelihood of the segment under a background model. Similarly to PWMs, positions

and sequences are treated independently and because of this the log-likelihood scores are additive.

PMMs are typically created from an existing PWM, a phylogenetic tree and an evolutionary model. The tree defines the evolutionary relationships and distances between the sequences. The evolutionary model defines how likely any base pair substitution is. To calculate a PMM accurately requires summing over all possible substitutions in every branch of the tree. Depending on the topology of the tree this process can be computationally intensive. This can restrict the application of PMM methods to multiple alignments with only a few sequences.

Examples of methods that use PMMs are MONKEY [Moses et al., 2004b], rMonkey [Moses et al., 2006], Motiph [Hawkins et al., 2009] and MotEvo [Arnold et al., 2012]. I will describe each in a little more detail.

MONKEY and rMonkey

To the best of my knowledge, MONKEY [Moses et al., 2004b] was the first TFBS predictor to use ideas of phylogenetic conservation. MONKEY was the top performer in an independent evaluation [Hawkins et al., 2009], performing better than the authors' own method, Motiph. MONKEY is a method for locating conserved TFBSs in multiple sequence alignments. It was originally developed and evaluated on *Saccharomyces* genomes.

MONKEY is based on a probabilistic model of TFBS conservation. It tests if a given segment of the multiple alignment is more likely to be part of the genomic background or part of a conserved TFBS. This is directly analogous to the likelihood ratio scoring functions described in Section 2.1.2 but generalised to the case with multiple sequences. The MONKEY model assumes that the site is present in all the aligned sequences.

MONKEY uses two separate evolutionary models: one for substitutions in TFBSs and the other for background substitutions. This captures the notion that substitutions in conserved binding sites should evolve more slowly than in background sequences. MONKEY allows the user to choose between models for substitutions of both types. For the motif substitutions, the default is a model from Halpern and Bruno (HB) [Halpern and Bruno, 1998], but a Jukes-Cantor model (JC) [Jukes and Cantor, 1969], a "simple" model and a model called "mk" are also available. The simple model is not actually an evolutionary model but one where MONKEY averages over the log-likelihoods in each sequence. In their original publication on MONKEY, [Moses et al., 2004b] note that the HB model out-performs this simple model but that even the simple model greatly outperforms scanning just one sequence. The mk model does not appear to be described

either in the MONKEY publication nor in the MONKEY software documentation. For the background substitutions, MONKEY provides the JC model by default and also the Hasegawa-Kishino-Yano model (HKY) [Hasegawa et al., 1985].

One drawback of the MONKEY method is that the run-time and memory requirements scale poorly in the number of aligned sequences. Each PMM position has 3^N free parameters where N is the number of aligned sequences. Calculation of these parameters from the motif substitution model and the single-species PWM is slow for more than a few sequences.

Moses et al. presented the rMonkey algorithm as part of a study highlighting the turnover of functional TFBSs between *Drosophila* species [Moses et al., 2006]. They studied the Zeste TF in four *Drosophila* species and estimated a turnover rate of around 5% of TFBSs between species. As part of this analysis they updated the MONKEY algorithm to create the rMonkey algorithm. rMonkey differs in allowing more flexibility in the alignment. rMonkey uses a greedy heuristic to update the alignment such that the highest-scoring sites align (so long as they align by at least one base pair in the original alignment). This heuristic is designed to overcome local mis-alignments that the authors expect to be common. Similarly to MONKEY, rMonkey's run-time and memory requirements scale poorly with the number of aligned sequences.

Motiph

The Motiph algorithm [Hawkins et al., 2009] is a variant of MONKEY that simply ignores regions with gaps. In contrast, MONKEY removes gaps when considering each TFBS. Motiph uses a star topology for the phylogenetic tree that simplifies the score calculations. In their evaluation on yeast data, the authors found that MONKEY outperformed Motiph.

MotEvo

It has been shown that TFBS prediction is improved by considering competition between TFs for nearby TFBSs [Rajewsky et al., 2002,Roider et al., 2007,Wasson and Hartemink, 2009]; clustering of TFBSs [Frith et al., 2001, Rajewsky et al., 2002]; and conservation of TFBSs in orthologous sequences [Kellis et al., 2003, Moses et al., 2004b, Siddharthan et al., 2005, Hawkins et al., 2009]. MotEvo [Arnold et al., 2012] was designed as the first TFBS predictor to incorporate all these features.

MotEvo incorporates an explicit evolutionary model in a similar way to MONKEY. However, the MotEvo model does not insist that TFBSs are conserved across all the

species in the given multiple alignment. MotEvo discards those sequences for which the the TFBS is more likely under the background model than the PWM model on a site-by-site basis. The authors of MotEvo argue that this improves the TPR of their method. MotEvo uses the Felsenstein evolutionary model (F81) [Felsenstein, 1981].

MotEvo also incorporates the concept of Unknown Function Elements (UFEs). Some conserved sequence elements may not be modelled by the PWMs available to MotEvo. UFEs model these elements explicitly to avoid inferences of TFBSs labelled with the incorrect TF. For instance, suppose we are searching for instances of a PWM but there exist TFBSs for a different TF with vaguely similar binding preferences. Even if we have no PWM for this different TF, its sites may be recognised because MotEvo's model rewards conservation across the sequences. However, its TFBSs will be associated with the closest matching PWM. UFEs model these TFBSs explicitly without prior knowledge of their PWMs.

MotEvo can use a maximum-likelihood approach to optimise the parameters of its model for the sequences presented to it. These include the background prior (prior probability of a TFBS), the prior probability of a UFE, and the PWMs themselves.

Branch length score methods

Like PMM methods, BLS methods also require an alignment and an evolutionary tree. BLS methods quantify their confidence in a TFBS prediction by summing the total length of branches in a phylogenetic tree connecting species that contain a TFBS at that point in the alignment. This method captures the concept that more TFBSs predictions in more distantly related species provide stronger evidence for a TFBS. The method explicitly allows binding site turnover, where TFBSs can be gained and lost through evolutionary substitutions. Multiple alignments are not perfect and a mis-alignment can be indistinguishable from the loss of a TFBS. BLS methods are more sympathetic to any such mis-alignments than PMM methods. BLS methods do not explicitly model the substitutions between the related species and hence do not have as many parameters as PMMs.

The first BLS method was presented by Kheradpour, Stark et al. in a study of twelve *Drosophila* genomes using PWMs for 83 TFs [Kheradpour et al., 2007,Stark et al., 2007]. This original method used a binary classifier at a given threshold for TFBS prediction. Xie et al. introduced the Bayesian Branch Length Score (BBLS) [Xie et al., 2009] which incorporates uncertainty into the TFBS predictions and reports the expectation of the sum of the branch lengths. Xie et al.'s comparison showed that the BBLS method significantly out-performed the BLS method. Håndstad et al. showed that BLS methods

could be used with scoring schemes other than log-likelihood ratios. They took scores from the MotifScan TFBS predictor [Naughton et al., 2006] and used them as input to the BBLS method [Håndstad et al., 2011].

2.3 The Binding Factor Analysis algorithm

The main contribution of this chapter is the Binding Factor Analysis (BiFA) algorithm. The BiFA algorithm is a log-likelihood ratio based method to predict binding sites. It works on a set of related sequences. This relationship would typically but not necessarily be phylogenetic. It is not necessary to specify the evolutionary distances involved or even a phylogenetic tree relating the species.

The BiFA algorithm is motivated in part by both the enhanceosome and the billboard models of enhancers (see Section 1.1.5). The core of the algorithm is based on a model where those TFs that bind *anywhere* in the related sequences are more likely to bind in the central sequence. This encapsulates the billboard model of enhancers. As well as this core algorithm there is a maximal chain extension. The maximal chain algorithm discovers the longest and most probable *conserved* sequence of TFBSs that occur in all of the sequences provided. This maximal chain algorithm encapsulates a relaxed version of the enhanceosome model of enhancers. A full encapsulation of the enhanceosome model would also consider the spacing and the orientation of the TFBSs across all the species.

2.3.1 Core algorithm

The basic concept of the core algorithm is that in a Bayesian framework we can use the log-likelihood ratio to probabilistically predict whether a TF binds to a TFBS (see Section 2.1.2). When considering a single sequence this is just another way of scoring a putative TFBS. However, the BiFA algorithm is designed to work on multiple sequences using a billboard model of enhancers. The billboard model suggests that if we believe the TF binds to the related sequences our belief that it binds to the central sequence should be stronger. Using the Bayesian framework described in Section 2.1.2, we can quantify our belief that the TF binds to each related sequence. Our belief that the TF binds to the given TFBS in the central sequence is then updated to reflect the evidence from the related sequences.

The input to the BiFA algorithm is a set of $M+1$ sequences $\{T_0, T_1, \ldots, T_M\}$ where T_0 is the central sequence and $T_1, \ldots, T_M$ are the related sequences. Suppose for the moment

that we are only interested in one PWM. Later for the maximal chain extension we will need to consider multiple PWMs. Each sequence, T_m, has a certain number, N_m, of potential TFBSs for the PWM. We label these TFBSs as $X_{m,1}, \ldots, X_{m,N_m}$. We represent the event that the TF binds to the nth TFBS in the mth sequence by $B_{m,n}$. Now using the method given in Section 2.1.2 we can calculate the probabilities $p(B_{m,n}|X_{m,n})$ for all m and n. Here we have to specify a prior probability of binding which we take to be constant across all sites, $p(B_{m,n}) = \alpha$. It would be possible to integrate other information into the BiFA algorithm by making this prior position-specific. For example, if we had DNase I hypersensitivity data we could use this to set a prior that varied across the possible TFBSs.

We would like to update the binding predictions for the central sequence $p(B_{0,n}|X_{0,n})$ using evidence of binding from the related sequences, $T_1, \ldots, T_M$. We represent the event that the TF binds to at least one potential TFBS in sequence T_m by B_m. We can predict whether the TF binds anywhere in each related sequence using

$$p(B_m|T_m) = 1 - \prod_n [1 - p(B_{m,n}|X_{m,n})] \tag{2.16}$$

that is one minus the probability that it does not bind anywhere in the sequence. In practice we only consider those $p(B_{m,n}|X_{m,n})$ that are above some small value that we call the *phylogenetic threshold*. This is a parameter of the BiFA algorithm.

Original updates

Now given a particular TFBS in the central sequence, $X_{0,n}$, and the related sequences, the original BiFA algorithm updates the probability that the TF binds to the TFBS

$$p(B_{0,n}|X_{0,n}, T_1, \ldots, T_M) = \sqrt[M+1]{p(B_{0,n}|X_{0,n}) \prod_m p(B_m|T_m)} \tag{2.17}$$

Arguably taking this geometric mean is an *ad hoc* solution to the problem of integrating evidence from the related sequences. When this method was first devised some time ago, I had in mind that this *ad hoc* approach could be justified via an argument that the sequences are not selected independently. Coming back to the method to write this thesis it is difficult to make this justification. The geometric mean in Equation 2.17 averages over probabilities but the intention was to average over the evidence that is contributed by each sequence. In our Bayesian setting, the probabilities in Equation 2.17 are a combination of this evidence with the prior. Averaging over terms that include the prior is counter-intuitive so later I will present a modified update method to Equation 2.17

that does not do this. However, as some of the work presented in this thesis is based on Equation 2.17 I present the method as it was used. Later I will compare results from both alternatives. The modified update method is not based on a full probabilistic model, but does average over the evidence and is perhaps easier to justify. I note that the original method was developed in collaboration with developmental biologists and computational biologists and as such has proved to be useful despite its *ad hoc* nature.

Modified updates

Here I present the modified update for the core BiFA algorithm that averages over the evidence provided by each sequence rather than the probabilities as in Equation 2.17. The modified update calculates the probability that the TF binds to the n'th binding site as follows. First, for each related sequence, T_m, a Bayes factor, K_m, representing the evidence in favour of the TF binding at least once to that sequence is calculated (see Equation 2.20 below). Because the sequences are related and therefore the evidence given by each is not independent, these Bayes factors are integrated via a geometric mean (or equivalently an arithmetic mean on the log probability scale) (see Section 1.3.5). So if $K_{0,n}$ is the Bayes factor in favour of the hypothesis $B_{0,n}$ as opposed to $\overline{B}_{0,n}$ we have

$$\frac{p(B_{0,n}|X_{0,n},T_1,\ldots,T_M)}{p(\overline{B}_{0,n}|X_{0,n},T_1,\ldots,T_M)} = \sqrt[M+1]{K_{0,n}\prod_{m=1}^{M}K_m}\frac{p(B_{0,n})}{p(\overline{B}_{0,n})} \tag{2.18}$$

Note that the geometric mean is only averaging the Bayes factors and that the prior is not included in this mean. This is in contrast to the original update method of Equation 2.17

To justify why we prefer this update method we note that it is difficult to model or quantify the degree of relatedness of the sequences the BiFA algorithm is presented with. In general, a user of the BiFA algorithm will have selected the sequences on the basis that they are conserved between species. How the user has measured that conservation is outside of the scope of the algorithm. However, we can certainly expect the presence of a TFBS in one sequence to correlate with the presence of a TFBS in another. For this reason we do not wish to over-estimate the strength of the evidence that the sequences provide. As Bayes factors are multiplicative on the probability scale (and additive on the log probability scale), this geometric (or equivalently arithmetic) mean seems a cautious way to incorporate information from the related sequences. Indeed, if we knew that all the related sequences were identical, the modified update method in Equation 2.18 would treat them as if we had only seen one of them. Admittedly this modified update is also an *ad hoc* method and I do not have a probabilistic model that justifies the above

equation. However, the fact that Bayes factors measure the evidence in favour of a hypothesis provides a certain justification.

It would be possible albeit difficult to construct a full probabilistic model that models the relatedness of the sequences and to do inference in this model as a way of integrating the information from the related sequences. This could lead to its own set of issues. The model would have to be parameterised and calibrated. This calibration would perhaps be suitable for some types of related sequences but I wanted the BiFA algorithm to be agnostic to the source of the sequences it analyses. The sequences presented to the BiFA algorithm may have been selected by any method. Any model of dependencies between them would impose assumptions. For this reason an algorithm with few parameters seems advisable.

I have not yet specified how to calculate the Bayes factors, K_m, for binding in the related sequences, T_m, in Equation 2.18. Using Equation 2.16, we are able to calculate the posterior probability that the TF binds at least once to the sequence, $p(B_m|T_m)$. It is straightforward to calculate the posterior odds, $\frac{p(B_m|T_m)}{1-p(B_m|T_m)}$, from the posterior probability. Now if we calculate the prior odds we can take the change from the prior odds to the posterior odds as the Bayes factor, K_m. To calculate the prior odds, we first calculate the prior probability that the TF binds at least once to the sequence

$$p(B_m) = 1 - \prod_n [1 - p(B_{m,n})] = 1 - (1-\alpha)^{N_m} \approx N_m \alpha \tag{2.19}$$

where the approximation is valid when $1 \gg \alpha N_m$. Again it is straightforward to convert the prior probability to the prior odds, $\frac{p(B_m)}{1-p(B_m)}$, and when combined with the posterior odds we have

$$K_m = \frac{1 - p(B_m)}{p(B_m)} \frac{p(B_m|T_m)}{1 - p(B_m|T_m)} \tag{2.20}$$

The difference between this modified update and the original update is that the averaging of evidence occurs in odds-space as opposed to probability-space. Both are *ad hoc* methods but I believe the odds-space version is justifiable and may be more powerful.

2.3.2 Maximal chain extension

First I give an overview of the intention and workings of the maximal chain algorithm. Later I will give the technical details.

Overview

The maximal chain extension to the BiFA algorithm models the TFBSs of an enhanceosome that are conserved across the BiFA input sequences. In order to do this it examines each sequence for the TFBSs of multiple PWMs. The algorithm looks for a sequence of TFBSs that is preserved across all the sequences. The PWMs for each TFBS must be the same for equivalent TFBSs in different sequences. Such a sequence is termed a chain. I will try to clarify this with an example. Suppose we are analysing four sequences for three PWMs, A, B and C. Suppose further that each sequence has strong binding sites for A, B and C in that order. The maximal chain algorithm would return A, B, C as the maximal chain. Now suppose that the binding sites for the second sequence are in the order A, C, B. In this configuration the longest chain that is conserved in order across the sequences is either A, C or A, B.

The algorithm does not just search for the longest chain irrespective of the strength of the TFBSs. Each TFBS is assigned a weight and the maximal chain algorithm searches for a chain conserved across all the sequences which has the largest weight. A chain's weight is defined as the sum of the weights of the TFBSs that comprise it. We weight each putative TFBS with a measure of our belief that it is a TFBS. The maximal chain algorithm also takes account of steric hindrance (see Section 1.1.4). Chains are not permitted to use overlapping TFBSs.

Definitions

In this section I define the terms needed for the maximal chain extension. As above, suppose the BiFA algorithm is given $M+1$ sequences, $T_0, \ldots, T_M$, only now suppose we are considering Q PWMs, $\theta_1, \ldots, \theta_Q$, each representing the sequence binding preferences of a TF. We need to extend our notation to accommodate the multiple PWMs and we add an extra subscript, after the sequence subscript, m, and before the TFBS subscript, n. So we have $X_{m,q,n}$ is the nth possible TFBS for the PWM, θ_q, in sequence T_m. Note that as the PWMs can have different widths, each PWM will have a different number of potential TFBSs, $N_{m,q}$, in each sequence, T_m. We use the log-likelihood ratio defined in Equation 2.4 as a score, $S_{m,q,n}$, for each potential TFBS, and only retain those TFBSs above some user-specified threshold, V.

Now we define a chain element, $\mathbf{u^q} = u_0^q, \ldots, u_M^q$, as a set of TFBSs for a PWM, θ_q, where each sequence is represented by exactly one TFBS,

$$\{X_{m,q,u_m^q} : 0 \leq m \leq M\}$$

A chain element represents the concept of a TFBS that has been conserved across all the sequences. Note that the number of chain elements for a particular PWM, θ_q, is $\prod_m |\{S_{m,q,n} \geq V\}|$ which grows exponentially as $O(|\{S_{m,q,n} \geq V\}|^{M+1})$ in the number of putative TFBSs above the threshold, V. We define a partial order on the set of chain elements for all PWMs by $\mathbf{u^q} < \mathbf{v^{q'}}$ if and only if $X_{m,q,u_m^q} < X_{m,q',v_m^{q'}}$ for all $0 \leq m \leq M$. Here we are using the natural partial ordering on TFBSs where $X_{m,q,n} < X_{m,q',n'}$ if and only if $X_{m,q,n}$ occurs strictly before (forbidding overlaps) $X_{m,q',n'}$ in the sequence, T_m. So one chain element precedes another if all of its TFBSs precede the corresponding TFBSs in the other chain. The weight (or score) of a chain element, $S_{\mathbf{u^q}}$, is defined as the sum of the weights of its constituent TFBSs,

$$S_{\mathbf{u^q}} = \sum_m S_{m,q,u_m^q}$$

Two chain elements, $\mathbf{u^q}$ and $\mathbf{v^{q'}}$ are said to be comparable if and only $\mathbf{u^q} < \mathbf{v^{q'}}$ or $\mathbf{v^{q'}} < \mathbf{u^q}$. A chain, U, with N_U elements is an ordered set of chain elements each pair of which are comparable,

$$U = \{\mathbf{u_i^{q_i}} : \mathbf{u_i^{q_i}} \text{ and } \mathbf{u_j^{q_j}} \text{ are comparable}, 1 \leq i \leq N_U, 1 \leq j \leq N_U, 1 \leq q_i \leq Q\}$$

A chain represents the concept of an enhanceosome, that is an ordered set of TFBSs that are conserved (in order) across all the sequences. The weight (or score) of a chain, is the sum of the scores of its elements,

$$S_U = \sum_i S_{\mathbf{u_i^{q_i}}}$$

A maximal chain is a chain such that no other chain has a higher score. A maximal chain can therefore be seen as the best sequence of PWMs for which there is a TFBS in each sequence. Furthermore the order of these TFBSs is conserved across all the sequences. This fits our enhanceosome model albeit ignoring the spacing and orientation of the TFBSs.

Implementation

It turns out that if maximal chains are defined in this way, there exist efficient algorithms to discover them due to Felsner, Müller and Wernisch [Felsner et al., 1997]. In the k-dimensional box representation terminology of Felsner et al. each chain element is a $(M+1)$-dimensional box and a maximal chain is a maximum weighted independent set in an interval graph. Felsner et al.'s algorithm to discover a maximal chain runs

in $O(n \log^M n)$ time where n is the number of chain elements. I have implemented the MAXCHAIN algorithm given by Felsner et al. in C^{++} and this forms the maximal chain extension of the BiFA algorithm.

TFBS orientation

The enhanceosome model is based upon specific protein-protein interactions. These interactions will almost certainly depend on the positioning and orientation of the TFs involved. Including the orientation of the TFBSs in the maximal chain algorithm would be straightforward. Each TFBS could be labelled with a '+' or a '-' depending on which strand binding is predicted. Only those TFBSs that matched their TF and their label would be considered as part of a chain element.

Modelling the spacing between TFBSs might also be worthwhile but could also be too restrictive. It is reasonable to suppose that the spacing between TFBSs could vary across species.

Alignment-free

The BiFA algorithm differs from most other algorithms that use phylogenetic models of TFBSs. Almost all other algorithms use an explicit model of the phylogenetic tree relating the species of the sequences under consideration. In addition most require a multiple alignment between the sequences. Whilst this has certain benefits in that species that are closely related can be treated differently from more distantly related species it also has drawbacks. Phylogenetic trees and multiple alignments are normally estimated using maximum likelihood procedures and there is no guarantee they are correct. In addition phylogenetic trees may not always be available for the relevant species. It can be time consuming for a user of a PWM scanning method to have to estimate a tree in order to make predictions. The use of a specific multiple alignment precludes the possibility of other alignments. The BiFA algorithm does not suffer from any of these drawbacks as neither a phylogenetic tree nor a multiple alignment is required. The BiFA algorithm models phylogeny implicitly by integrating evidence from each sequence and in a sense it integrates over all possible alignments by considering all the possibilities in the maximal chain algorithm.

Another point to note is that any method that uses multiple alignments may not work well when there has been high turnover of TFBSs. See Section 1.1.5 for a discussion of this in the context of phylogenetic conservation. This is perhaps most relevant for enhancers that follow the billboard model.

One drawback of BiFA's alignment-free model is that it can be sensitive to the lengths of the related sequences. The BiFA algorithm analyses the whole length of the related sequences for TFBSs. When the related sequences are long it is easier to find a match to the PWM. This can affect how the predictions in the central sequence are updated. Therefore, the length of the related sequences should not be much longer than the distance a TFBS might have moved between species. Unfortunately, we do not have good estimates for this distance. Enhancers are generally assumed to be of the order of a few hundred base pairs in length. The BiFA algorithm has been designed with this in mind. It has been evaluated on sequences of length within an order of magnitude of 300 base pairs. When the sequences are much longer than this it is reasonable to break them up into segments (or chunks) and perform separate runs of the algorithm. However, breaking up the sequences may require a multiple alignment to anchor the break points across the sequences.

2.3.3 An application

I present an example of the application of the BiFA algorithm. I used the algorithm to study an enhancer of Nodal, a factor important in cell-fate specification and patterning in the mouse embryo. This work was carried out for Jérôme Collignon's group at the Institut Jacques Monod and forms part of a publication in Developmental Biology [Granier et al., 2011]. The enhancer in question is known as the *PEE* cis-regulatory region.

Based on a sequence alignment technique [Baxter et al., 2012], regions upstream of the Nodal TSS conserved in mouse, human and cow were identified. The most 5' of these was the known *PEE* enhancer. I analysed the regions using the core BiFA algorithm and matrices from the commercial version of TRANSFAC. The algorithm predicted two strong LEF/TCF1 binding sites in the *PEE* enhancer (see Figure 2.1). LEF/TCF factors are known to be effectors of the canonical Wnt/β-catenin signalling pathway [Arce et al., 2006]. This pathway has been independently implicated in embryonic patterning providing further evidence that these TFBSs are likely to be functional.

2.4 A comparison of TFBS prediction methods

In contrast to the task of motif finding [Tompa et al., 2005, Sandve et al., 2007], TFBS prediction does not have a long-established benchmark. Until recently the lack of a gold standard of binding sites across a wide range of TFs hindered the evaluation of PWM scanning methods. The abundance of high-throughput high-quality ChIP data now allows better evaluations to be conducted. In this section I present an overview of

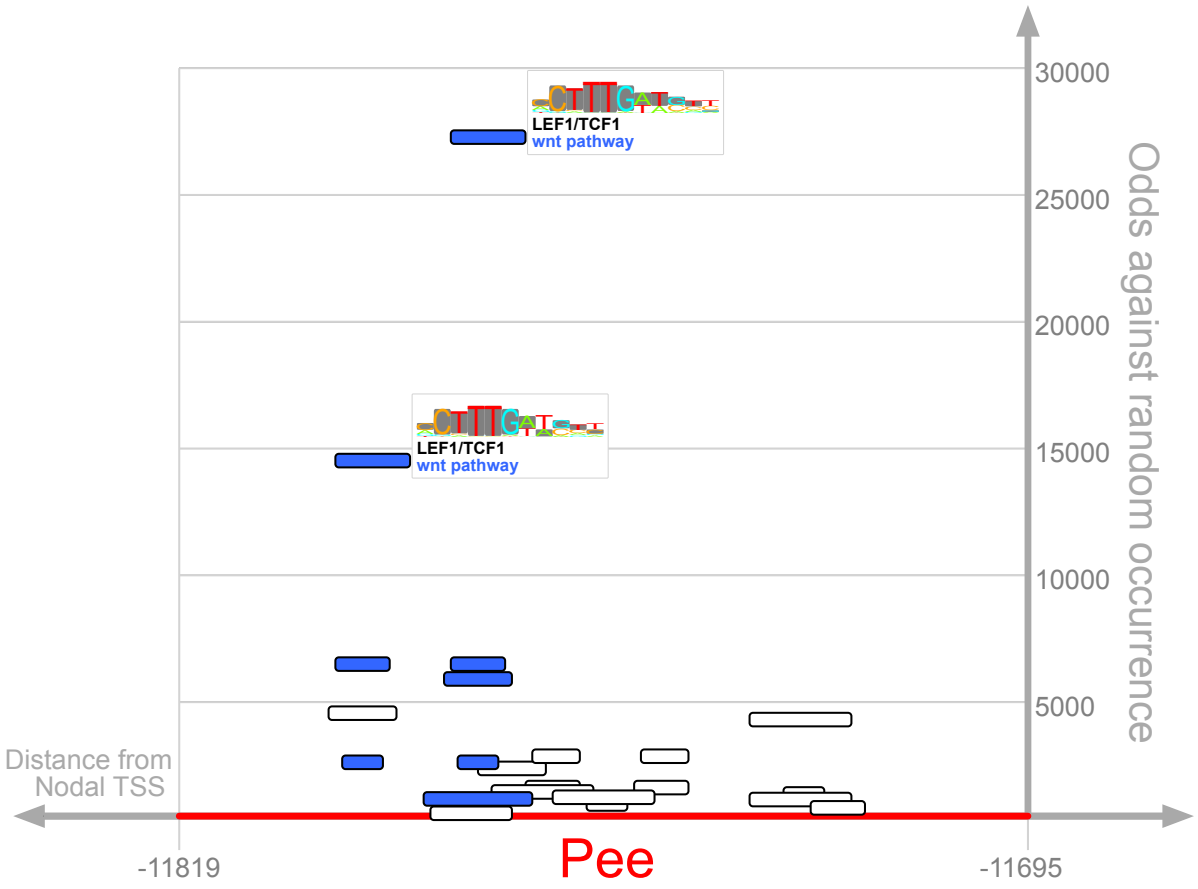


Figure 2.1: A graphical depiction of the posterior odds associated with putative TFBSs in the *PEE* enhancer upstream of the Nodal TSS. Each box represents a putative TFBS. TFBSs for TFs associated with the Wnt/β-catenin signalling pathway are coloured blue. The distance from the Nodal TSS in base pairs is given on the x-axis. The length of the boxes represent the region that the TFBS occupies on the DNA. The posterior odds for the TFBSs are given on the y-axis. The two most significant predictions are both for a PWM representing a LEF1/TCF1 heterodimer's binding preferences. Extra detail is provided for these TFBSs in the white region next to them. The sequence logo detailing the binding preferences of the PWM is given and for each binding site, the actual bases in the DNA sequence are shaded. This visualisation makes it possible to see that the more significant TFBS is a better match for the PWM as it has a `T` as opposed to an `A` in the fourth position from the end. This figure is produced by the implementation of the BiFA algorithm. It was edited slightly for publication.

other work that has compared phylogenetic TFBS predictors; I describe the data and methods I have used in the comparison in this thesis; and finally I present some results and a discussion highlighting the main points of the results which I relate to previous benchmark studies.

2.4.1 Previous comparisons of phylogenetic methods

When describing their BBLS method, Xie et al. compared its performance to that of MONKEY and BLS on ChIP data sets in human for the TFs: CTCF, NRSE, p53, Myc, STAT1, and NFκB. BBLS was found to be significantly better than both other methods on all the data sets.

An investigation into phylogenetic TFBS prediction methods [Hawkins et al., 2009] using a gold-standard set of TFBSs in *Saccharomyces cerevisiae* from SCPD [Zhu and Zhang, 1999] found that simple non-phylogenetic methods performed better than phylogenetic methods. The phylogenetic methods tested were: MONKEY, rMonkey and the authors' own method, Motiph. However, Hawkins et al. were concerned that their evaluation was biased by TFBSs that were missing from the gold standard. They used an approach that shuffles the columns of PWMs to estimate the distribution of scores on background sequence. Using this approach the phylogenetic motif scanners performed better than simple non-phylogenetic methods. The top-performing method was MONKEY. Hawkins et al. suggested that phylogenetic methods might be better at predicting weak TFBSs.

The MotEvo authors chose to evaluate their method on ChIP-seq data for five human TFs: CTCF; GABP; NRSF; SRF and STAT1 [Jothi et al., 2008, Valouev et al., 2008], using an alignment to six other mammals: mouse, dog, cow, monkey, horse and opossum. They compared MotEvo to the MONKEY and the PhyloScan algorithms. MotEvo was the top performer in this evaluation. The MotEvo authors did not choose to do a direct comparison of TFBS prediction performance using alignments of varying numbers of species. They chose to evaluate MotEvo's ability to predict enhancers as the number of species in the alignment varied. They investigated 76 experimentally validated blastoderm enhancers from *Drosophila* with alignments ranging from a single species to nine species. They found the coverage of predicted enhancers increased from 57% to 93% as the number of species in the alignment increased.

Recently, Håndstad et al. have proposed a benchmark framework for the TFBS prediction problem based on publicly available human ChIP-seq data [Håndstad et al., 2011]. They assessed five different TFBS prediction methods: PWM, MotifScan, WS, and two BBLS methods. They found that methods that use sequence conservation perform better in general than simpler methods. They found this effect was TF-dependent and

seemed to be strongest with PWMs of low information content and ChIP-seq peaks of high affinity. Indeed, they found that simpler methods can out-perform conservation based methods for TFs with high information content. Håndstad et al. came to the opposite conclusion about the utility of phylogenetic methods for detecting weak vs. strong TFBSs to Hawkins et al.: they suggest phylogenetic methods are better at detecting strong TFBSs.

One comparison that is lacking in all the evaluations above is a direct comparison of how increasing the number of sequences in an alignment correlates with a method's performance. The MotEvo evaluation comes closest to an evaluation of the benefits associated with including more species in an alignment. However, it measures this in an indirect manner, by evaluating the method's ability to predict enhancers. The study presented in this chapter addresses this question directly.

2.4.2 The benchmarks

The benchmarks presented in this study are based on the evaluations by Håndstad et al. Their evaluations only used binding data from experiments in human cell lines. In this study I have extended their proposed framework with additional benchmarks derived from *Drosophila* binding data. I have also extended the framework to study the effects of varying the number of aligned sequences presented to a single method. In contrast, Håndstad et al. compared distinct methods that used different numbers of aligned sequences. I have also used a slightly different statistic for measuring the performance of methods at high specificity.

To clarify the structure of the benchmarks I define some terms: a *benchmark* refers to a collection of genomic binding data for one species that has typically come from one experimental source but may contain data for more than one TF. Each benchmark has a multiple alignment associated with it over a given set of genomes, including the genome of the benchmark's species. Each TF has a PWM associated with it that describes its binding preferences. A benchmark has multiple *test cases* for each TF. Each test case is a genomic region usually with one positive *sub-region* (where the TF is supposed to bind), surrounded by several negative sub-regions (where the TF is not supposed to bind).

The Håndstad site benchmark

The Håndstad site benchmark is designed to address the problem of locating a TFBS in a sequence bound by a known TF. Håndstad et al. took publicly available ChIP-seq data from the ENCODE project [Birney et al., 2007]. The ENCODE project is a

large consortium whose aim is "to identify all functional elements in the human genome sequence". Most of the binding data used in this benchmark is from the K562 cell line although some is from the HeLa-S3 cell line. These HeLa cell line data are treated separately in my analysis and their TFs are labelled as "hela". I followed the protocol described in [Håndstad et al., 2011] for creating test cases from the called ChIP-seq peaks which I describe briefly here. Each peak was considered a positive sub-region and surrounded by a 20kb region that was partitioned into sub-regions of 200bp each of which were considered negative sub-regions. This resulted in 69,267 test cases, each containing one positive sub-region and 100 negative sub-regions. The benchmark includes data for the eight TFs: c-Fos; c-Jun; c-Myc; E2F4; GATA1; Max; NFKB; NRSF. The TFs c-Fos, c-Myc, E2F4 and Max have data for both cell lines, the other TFs only have data for the K562 cell line. The number of cases per TF can be seen in Figure 2.2. The benchmark uses an alignment of 18 placental mammals extracted from the 28-way alignment of the hg18 assembly downloaded from the UCSC Genome Browser [Kent et al., 2002]. The phylogenetic tree is shown in Figure 2.3. I used the same PWMs as Håndstad et al. These are shown in Table 2.1.

The turnover benchmark

Bradley et al. examined experimental evidence for the binding of six TFs in two closely related *Drosophila* species: *melanogaster* and *yakuba* [Bradley et al., 2010]. They found that almost all the bound regions were also bound in the orthologous sequence. However, there was considerable variation in the levels of binding between the species. This variation was highly correlated across all six TFs suggesting a factor-independent reason for this variation such as chromatin accessibility. However, the correlation was not perfect and they found numerous instances where the variation was apparently driven by the gain or loss of TFBSs in one of the sequences.

I downloaded the *Drosophila melanogaster* peaks from the Gene Expression Omnibus (accession GSE20369). The TFs investigated were bicoid (bcd); caudal (cad); giant (gt); hunchback (hb); knirps (kni); Krüppel (kr). Following the protocol defined for the Håndstad site benchmark, I converted the peaks into 29,343 test cases. The benchmark uses a multiple alignment of the two *Drosophila* species extracted from the 15-way alignment of the dm3 assembly downloaded from the UCSC Genome Browser. The TFs and their PWMs are shown in Table 2.2.

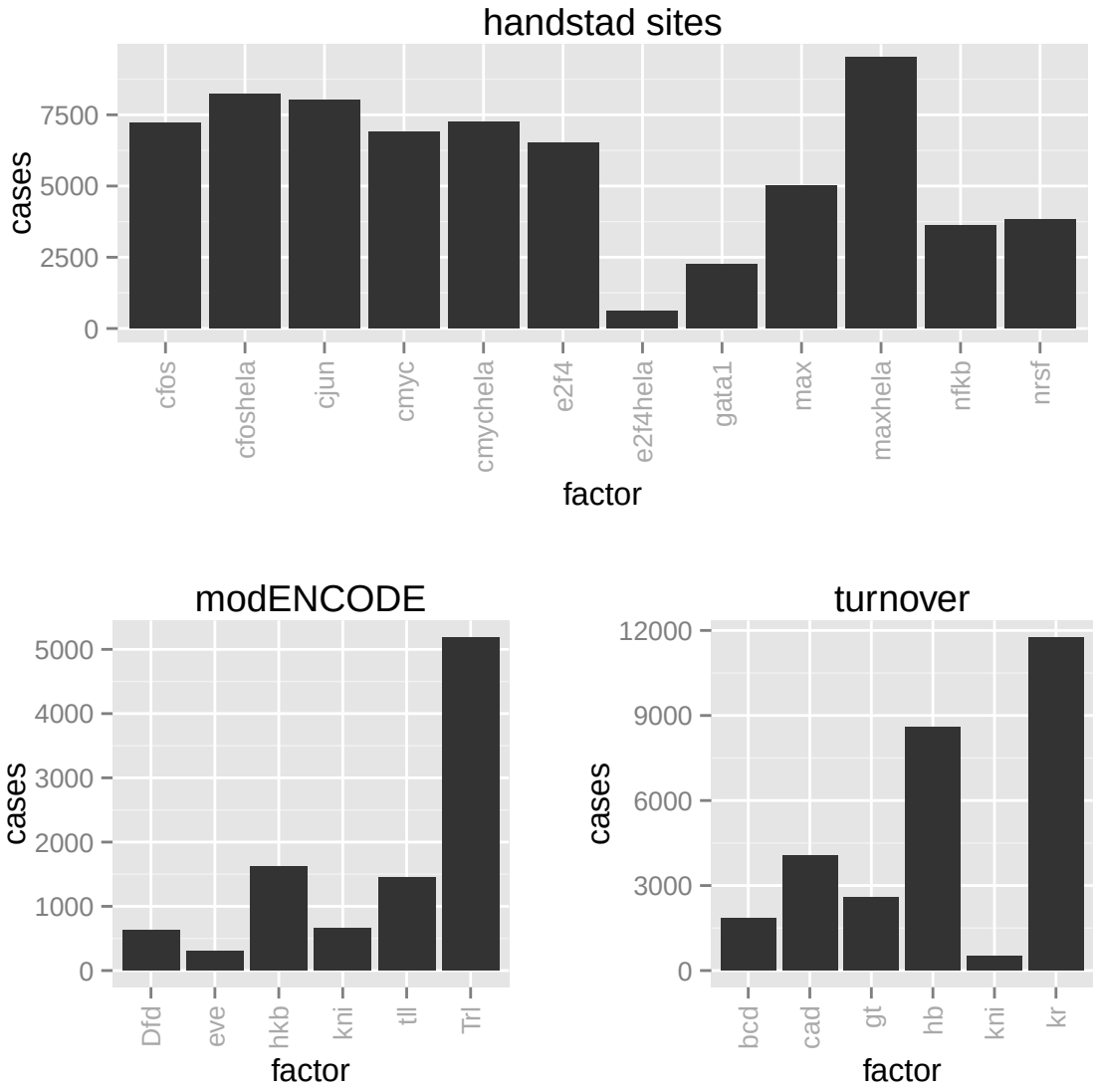


Figure 2.2: The number of test cases in the Håndstad sites, modENCODE and turnover benchmarks.

The modENCODE benchmark

The modENCODE project aims to "provide the biological research community with a comprehensive encyclopedia of genomic functional elements in the model organisms *C. elegans* and *D. melanogaster*" [Celniker et al., 2009]. To this end, members of the project have conducted several high quality TF binding experiments for *Drosophila melanogaster*.

I downloaded peaks from the modENCODE experiment "Chromatin Binding Site Mapping of Transcription Factors in D. melanogaster by ChIP-seq" from modMine [Contrino et al., 2012] for the TFs Deformed (Dfd), even skipped (eve), huckebein (hkb), knirps

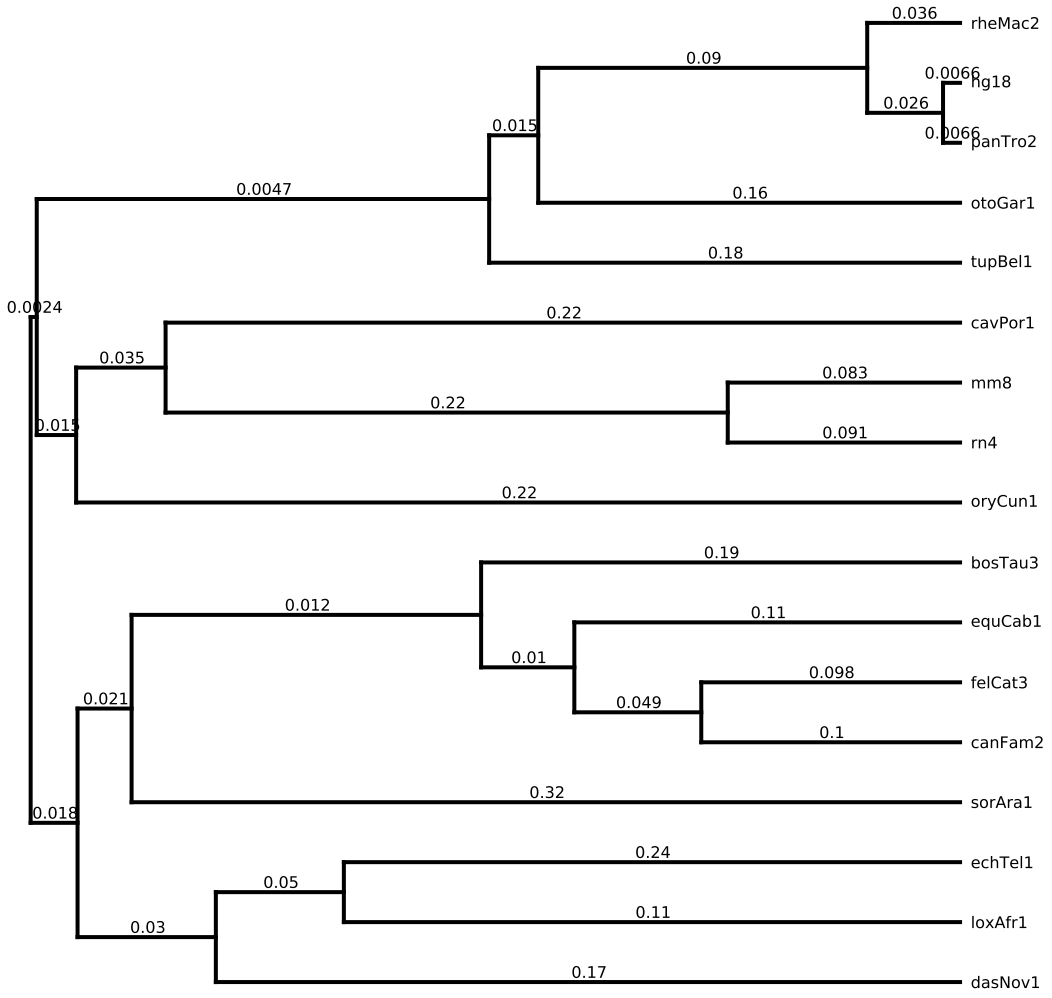


Figure 2.3: The phylogenetic tree used for the Håndstad site benchmark. The tips are labelled with their UCSC Genome Browser assembly identifiers. The branch labels represent relative evolutionary time.

(kni), tailless (tll), and translucent (Trl). Peaks in the mitochondrion genome were discarded. Following the protocol defined for the Håndstad site benchmark, I converted the remaining peaks into 9,861 test cases. The number of cases per TF is shown in Figure 2.2. The benchmark uses a multiple alignment of ten *Drosophila* species extracted from the 15-way alignment of the dm3 assembly downloaded from the UCSC Genome Browser. The ten species are: *melanogaster*, *simulans*, *yakuba*, *erecta*, *ananassae*, *pseudoobscura*, *williston*, *virilis*, *mojavensis* and *grimshawi*. The phylogenetic tree for the benchmark is shown in Figure 2.4. The TFs and their PWMs are shown in Table 2.3.

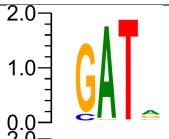
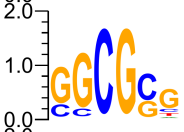
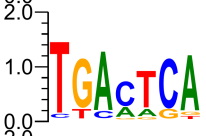
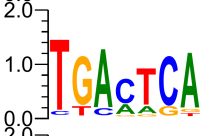
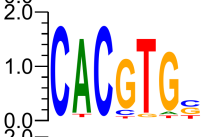
TF	PWM	Width	IC	IC/base	Logo
GATA1	MA0036.1	5	6.85	1.37	
E2F4	M00803	6	8.87	1.48	
c-Fos	MA0099.2	7	9.19	1.31	
c-Jun	MA0099.2	7	9.19	1.31	
c-Myc	M00799	7	10.62	1.52	
Max	MA0058.1	10	14.14	1.41	
NFKB	MA0105.1	11	16.54	1.5	
NRSF	M01028	19	25.23	1.32	

Table 2.1: The TFs and PWMs in the Håndstad site benchmark, sorted by information content (IC) given in bits. PWM identifiers that start MA (respectively M) refer to the JASPAR (respectively TRANSFAC) database. Note that c-Fos and c-Jun share the same PWM.

2.4.3 Framework

In this section I present how the benchmark framework evaluates the methods. When a TFBS prediction method is applied to a benchmark, the method is presented with each test case individually. The method's predictions for each test case are mapped to the positive and negative sub-regions in the test case. Each such sub-region is scored by the strongest prediction in that sub-region. In this way the number of TPs, FPs, TNs and FNs for each TF in a benchmark can be calculated for each scoring threshold.

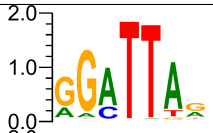
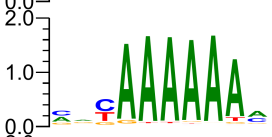
TF	PWM	Width	IC	IC/base	Logo
bicoid	BCD.MTF	7	9.91	1.42	
caudal	CAD.MTF	10	10.99	1.1	
giant	GT.MTF	12	13.08	1.09	
hunchback	HB.MTF	10	13.5	1.35	
knirps	KNI.MTF	13	14.04	1.08	
Krüppel	KR.MTF	11	13.35	1.21	

Table 2.2: The TFs and PWMs in the turnover benchmark, sorted by information content (IC) given in bits. PWM identifiers refer to the iDMMPMM library [Kulakovskiy and Makeev, 2010].

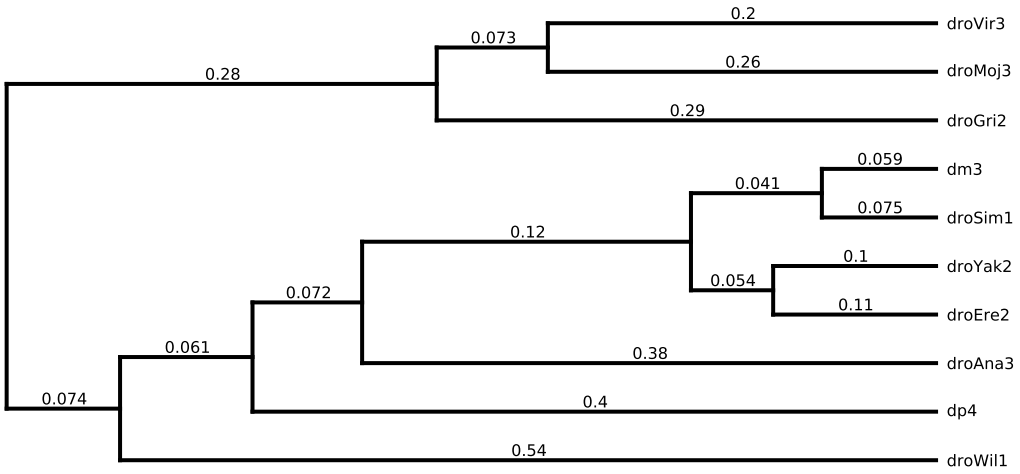


Figure 2.4: The phylogenetic tree used for the modENCODE benchmark. The tips are labelled with their UCSC Genome Browser assembly identifiers. The branch labels represent relative evolutionary time.

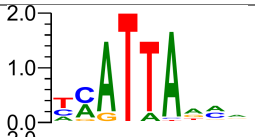
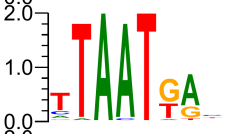
TF	PWM	Width	IC	IC/base	Logo
even skipped	EVE.MTF	9	10.14	1.13	
Deformed	DFD.MTF	8	11.02	1.38	
translucent	MA0205.1	10	11.17	1.12	
tailless	TLL.MTF	10	12.68	1.27	
huckebein	HKB.MTF	10	13.65	1.36	
knirps	KNI.MTF	13	14.04	1.08	

Table 2.3: The TFs and PWMs in the modENCODE benchmark, sorted by information content (IC) given in bits. PWM identifiers refer to the iDMMPMM library [Kulakovskiy and Makeev, 2010] except for MA0205.1 which is from the JASPAR library.

Parallelisation

The benchmarks consist of over 100,000 test cases. Each test case has around 20kb of central sequence, making around 2Gb in total. Depending on the benchmark, this is aligned to almost as much sequence in each of 17, 9 or 1 other species. Running several methods with varying parameters on these benchmarks is a technical challenge. I built a test harness that allows parallelisation of this task across several Linux servers.

Alignments

In contrast to the work done by Håndstad et al., the number of species used by each method in this framework is not fixed by the prediction method. In their work the WS method used three species and the BBLS method used 18 placental mammals. I have extended the framework so that each prediction method can work with multiple alignments over varying numbers of species. When a method is applied to the benchmarks, one parameter is the maximum number of species available to it. This is applied on a test case-by-test case basis as each test case may not be aligned to all the species in

the benchmark-wide alignment. That is, each test case may use aligned sequences from different sets of species, but each such set is limited in number. This has allowed me to investigate more closely the benefits of using data from varying numbers of related species. The framework calculates phylogenetic sub-trees and sub-alignments as needed for the different restricted subsets of species available to each test case.

Statistics

I plotted ROC curves and calculated AUC statistics (see Section 1.3.10) for each combination of benchmark and method considered. However, I have not used the AUC statistic as the primary measure of the methods' performance. TFBS prediction has a high FDR and we typically test many regions at a time, leading to large numbers of positive predictions. Usually there are only limited resources to follow up on the predictions, so there is particular interest in the performance of methods at thresholds that correspond to high confidence predictions. A standard statistic to measure the performance at high thresholds is the AUC50. However, as discussed earlier (Section 1.3.10), this statistic has two shortcomings: firstly, it is sensitive to the size of the benchmark; and secondly, 50 is an arbitrary choice that may not make sense on large (or small) benchmarks as it corresponds to an extremely low (respectively high) FPR. [Håndstad et al., 2011] use both the AUC and AUC50 statistics in their comparison of TFBS prediction methods. I prefer to use a statistic that represents the area under the ROC curve bounded above by a given FPR. I call this statistic the AUCFPR. Given similar ratios of positive to negative examples, this AUCFPR statistic is comparable across benchmarks of varying sizes. The FPR threshold can be defined as the problem demands or as the researcher prefers. I have used AUCFPR with a FPR of 5% as the primary measure of the performance of the methods in this study.

The classification of regions into positive and negative examples will not be perfect. Some of the peaks may not contain TFBSs for the TF in question as the cross-linking step in ChIP-seq can detect indirect binding of the target TF via an intermediary TF. Conversely, some of the negative examples may contain TFBSs that are not occupied under the experimental conditions tested. For these reasons I do not expect any of the methods to achieve high AUC statistics.

The Wilcoxon paired signed-rank test is a non-parametric hypothesis test [Wilcoxon, 1945] applicable to matched samples. The null hypothesis for the test is that the difference between the matched pairs for the two samples are symmetric about 0. The alternative hypothesis in the one-sided version of the test I use is that the distribution of one of the samples has a positive location shift. I use the test in this chapter to compare samples of the AUCFPR statistic for pairs of methods.

The box plots in this chapter follow the normal box plot conventions: the hinges at the extremes of the box represent the first and third quartiles of the data and the band inside the box represents the median (second quartile). The ends of the whiskers represent the highest (respectively lowest) datum within 1.5 times of the inter-quartile range (IQR) of the third (respectively first) quartile.

Prediction methods

I chose to compare the BiFA algorithm to two other phylogenetic TFBS prediction methods: MONKEY and MotEvo and a non-phylogenetic TFBS prediction method, FIMO.

2.4.4 Results

The MONKEY, MotEvo, FIMO and BiFA methods have several parameters. Before evaluating them I endeavoured to determine which parameter settings had the best performance on the benchmarks. For these tests, I limited the number of species in each test case to four. Note that the turnover benchmark only has two species and the FIMO method only uses the central sequence in each test case.

I prefer to optimise the parameters rather than use the default parameters as these are not always well chosen. One potential drawback is overfitting the parameters to the benchmarks. However, I expect that the large size of the benchmarks and the relatively small number of parameters should mitigate against this. I optimised most methods using evaluations that repeatedly sub-sample different test cases. This should also help to prevent overfitting.

Optimisation of MONKEY parameters

MONKEY implements four motif substitution models and two background substitution models. I compared all combinations to see which performed best on the benchmarks. For the HKY background substitution model, I did not vary the transition transversion rate ratio parameter but always left it at the default setting of 2.

The choice of MONKEY parameters made little absolute difference to either the AUC and AUCFPR statistics across all benchmark/TF combinations (see Figures 2.5 and 2.6). The changes were of the order of 0.001 and 0.0001 respectively which represent a change of about 0.1% and 1% respectively. Hence the effect of changing the parameters was small in both cases but greater on higher compared to lower confidence predictions.

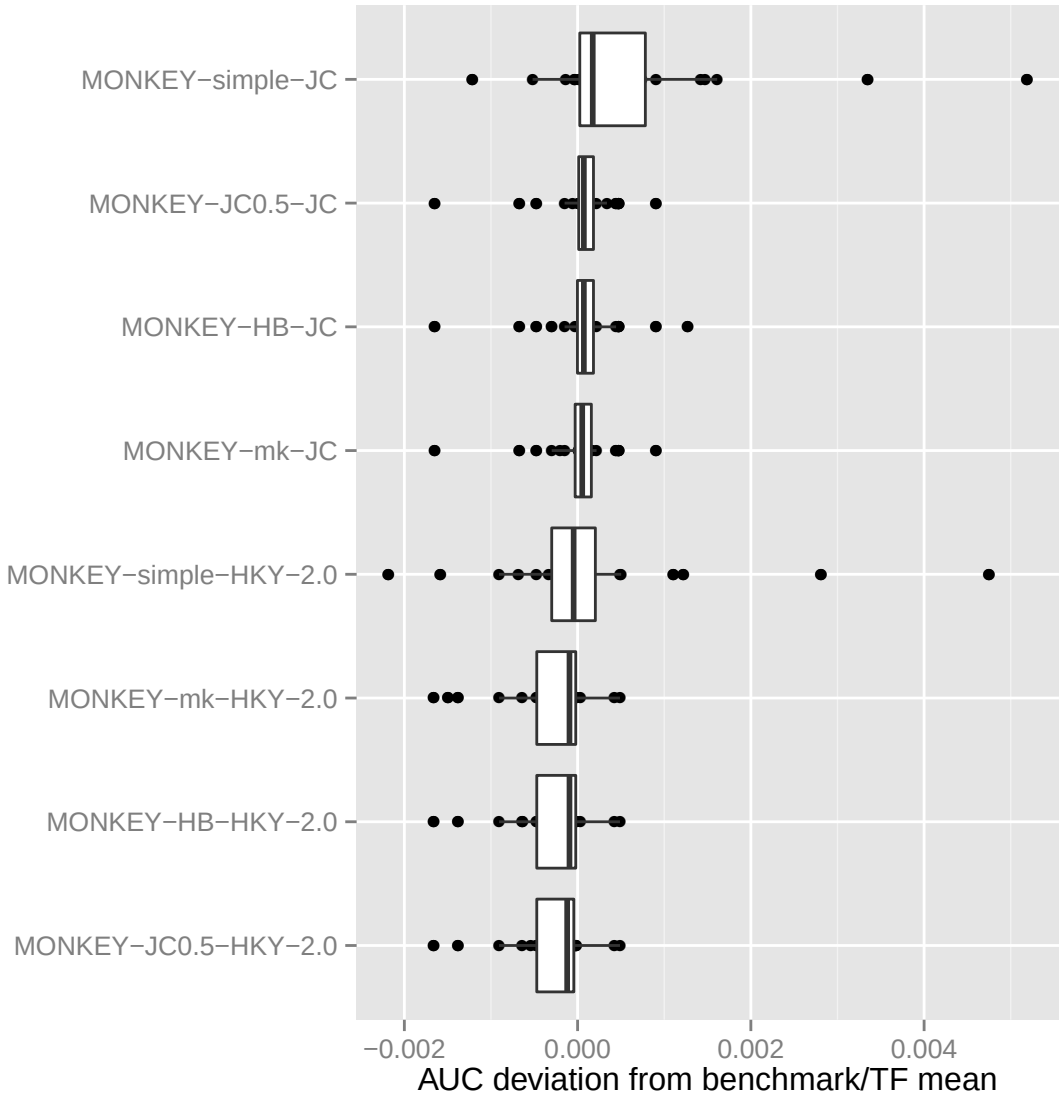


Figure 2.5: The performance of MONKEY as measured by the AUC statistic for different parameter settings. An AUC statistic was calculated for each benchmark/TF/parameter combination. The mean AUC statistic for each of the 33 benchmark/TF combinations was calculated over all methods. The box plots represent the deviations from these means grouped by the labelled parameters. The parameters are labelled by the motif substitution model followed by the background substitution model. MONKEY allows the motif model to be one of HB, mk, JC (with a substitution rate of 0.5), and simple. The background model is one of JC or HKY (with a transition transversion rate ratio of 2). The methods are sorted by descending median deviation.

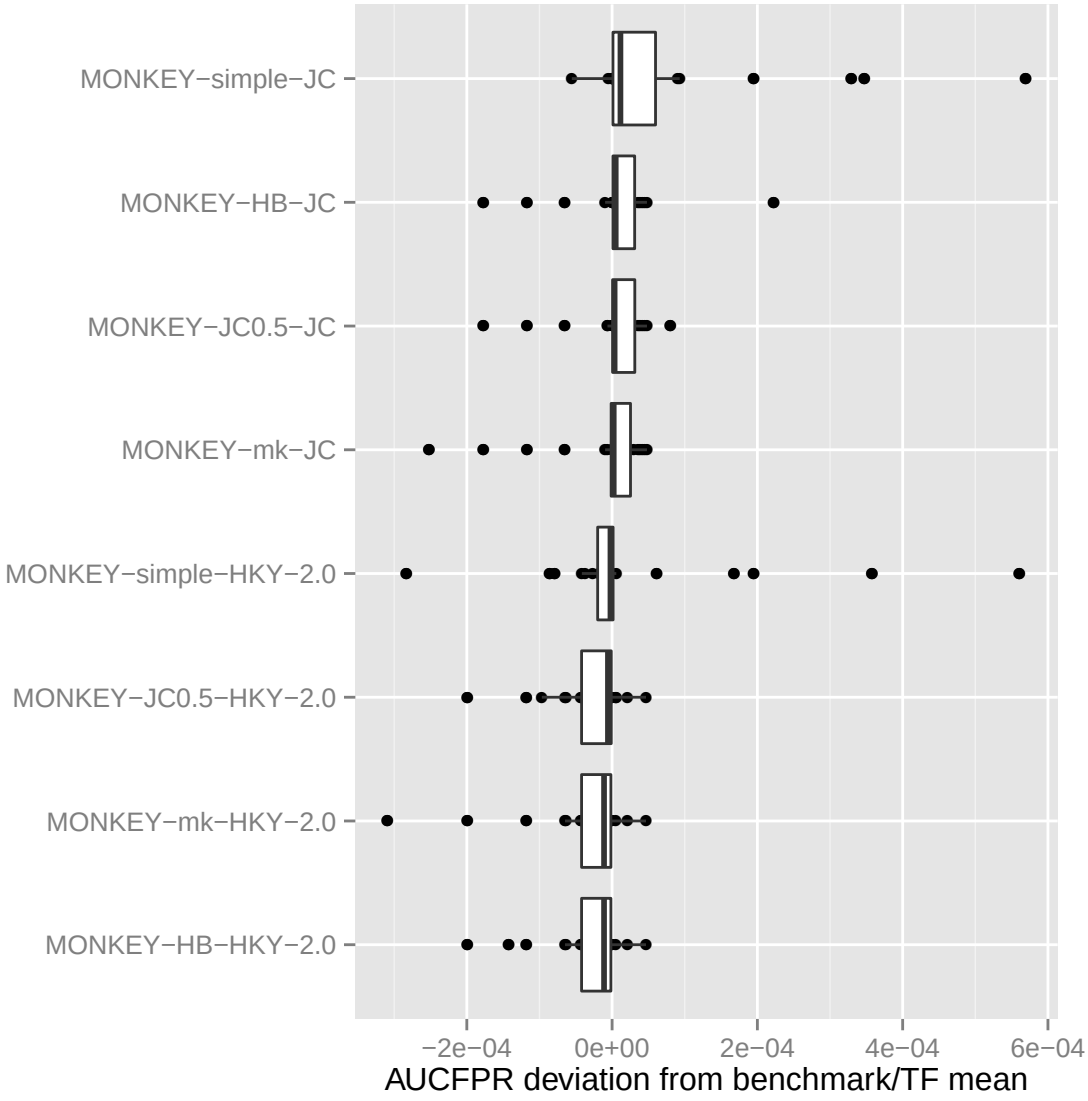


Figure 2.6: The performance of MONKEY as measured by the AUCFPR statistic for different parameter settings. See the caption of Figure 2.5 for a description of how the plot was generated.

However, despite these small differences, the JC background substitution model appeared to perform consistently better under both statistics when compared to the HKY model. Surprisingly the simple motif substitution model also appeared to consistently outperform the explicit evolutionary models on both statistics. This contrasts directly with the results published by the MONKEY authors [Moses et al., 2004b]. I tested if these results were significant using a one-sided Wilcoxon paired signed-rank test. Using this test, both the AUC and AUCFPR statistics of the MONKEY-simple-JC method were significantly greater at the 0.05 level than those of the MONKEY-simple-HKY-2.0 method (p-values of 1.2e-5 and 5.2e-6 respectively). This suggests using the JC back-

ground substitution model improves the performance of MONKEY. I also tested if the effect of changing the motif substitution model was significant. I compared the AUCFPR statistics of the top performing method MONKEY-simple-JC against the MONKEY-JC0.5-JC, MONKEY-mk-JC and MONKEY-HB-JC methods. The simple method was significantly better than both the JC0.5 and mk methods but not the HB method at the 0.05 level (p-values of 0.041, 0.041 and 0.064 respectively).

In Moses et al.'s original publication on their MONKEY algorithm they compared the simple scoring scheme to the HB motif substitution model and found that the HB model was significantly better. I have used a different data set for my evaluation. Moses et al. used data from yeast, I have data from *Drosophila* and human. Further work is needed to identify the reasons for the discrepancy between their results and these results. In particular, I plan to include a yeast benchmark in a later version of this study.

Optimisation of MotEvo parameters

To determine the best parameter settings for MotEvo, I considered the UFE and background priors and the option that allows the priors to be learnt. I did not test if optimising the PWMs improved performance. As the priors can take any value in a continuous range I choose to use the spearmint method [Snoek et al., 2013] to find the best parameter values. Spearmint is an algorithm that uses Gaussian processes to model how an objective function varies as a function of the parameters of a method. Spearmint uses the predictive mean and uncertainty of a Gaussian process over the parameter-space to choose candidate parameter settings for evaluation.

Here I used the average AUCFPR statistic across all the benchmark/TF combinations as the objective function. For each evaluation of the objective function the number of test cases per combination was down-sampled to 500 so that the spearmint algorithm could complete in a reasonable time. This down-sampling adds noise to the objective function. However, spearmint's model of the objective function incorporates a noise term and should adapt to this noise. After evaluating 215 different parameter settings, spearmint chose as optimal a background prior of 0.983 and a UFE prior of 301 without automatic learning of the prior values. These are the settings I have used in the following unless otherwise stated.

Optimisation of BiFA parameters

Each test case is around 20Kb in length and hence is unsuitable for a direct application of the BiFA algorithm (see Section 2.3.2). I partitioned each test case up into chunks

using the multiple alignment and applied the BiFA algorithm to each chunk. I treated the size of the chunks as a parameter of the BiFA method.

I used the spearmint tool to optimise four parameters for the BiFA method: the prior odds; the phylogenetic threshold; the chunk size; and the choice of update method (see Section 2.3.1). After 202 evaluations of the objective function spearmint chose the following optimal parameters: prior odds of $10^{-5.5}$; a phylogenetic threshold of 10^{-6}; a chunk size of 155bp and to use the original rather than the modified update method. I use these parameters in all that follows unless otherwise stated.

It was disappointing that spearmint chose not to use the modified update method; however the second best parameters evaluated did use the modified update method. Their average AUCFPR was only 1.1×10^{-5} lower than the optimal parameters. Later on I directly compare the modified and original update methods.

Optimisation of FIMO parameters

FIMO only has one relevant parameter to optimise: a pseudo-count that is added to the PWM. I chose several different pseudo-counts and evaluated FIMO using each value. I used a range of values above and below the default value of 0.1 including 0.8, the value recommended by Nishida et al. in their study of the effect of pseudo-counts on the performance of TFBS prediction methods [Nishida et al., 2008]. The results are scatter-plotted in Figure 2.7. I found the most variation in AUCFPR occurred in the Håndstad sites benchmark. Both the modENCODE and turnover benchmarks were less affected by the pseudo-counts. Nishida et al. highlighted a possible relationship between the optimal pseudo-count and the information content of the PWM. The scatter plot suggests that low (respectively high) pseudo-counts may be more suitable for PWMs with low (respectively high) information contents. As the linear models fitted in the plot suggested higher pseudo-counts (especially 5) might perform best, I tested if the AUCFPR for the pseudo-count of 5 were significantly greater at the 0.05 level than those of pseudo-counts of 1.5, 1 and 0.8 using a one-sided Wilcoxon paired signed-rank test. They were not (p-values of 0.14, 0.16 and 0.10 respectively). Hence I followed Nishida et al.'s recommendation: all further FIMO results in this study use a pseudo-count of 0.8.

Method comparison

Having established optimal parameter settings for the four algorithms in the comparison, I plotted ROC curves and made a comparison of each using AUCFPR statistics across all benchmark/TF/method combinations. All the ROC curves are given in Appendix A.

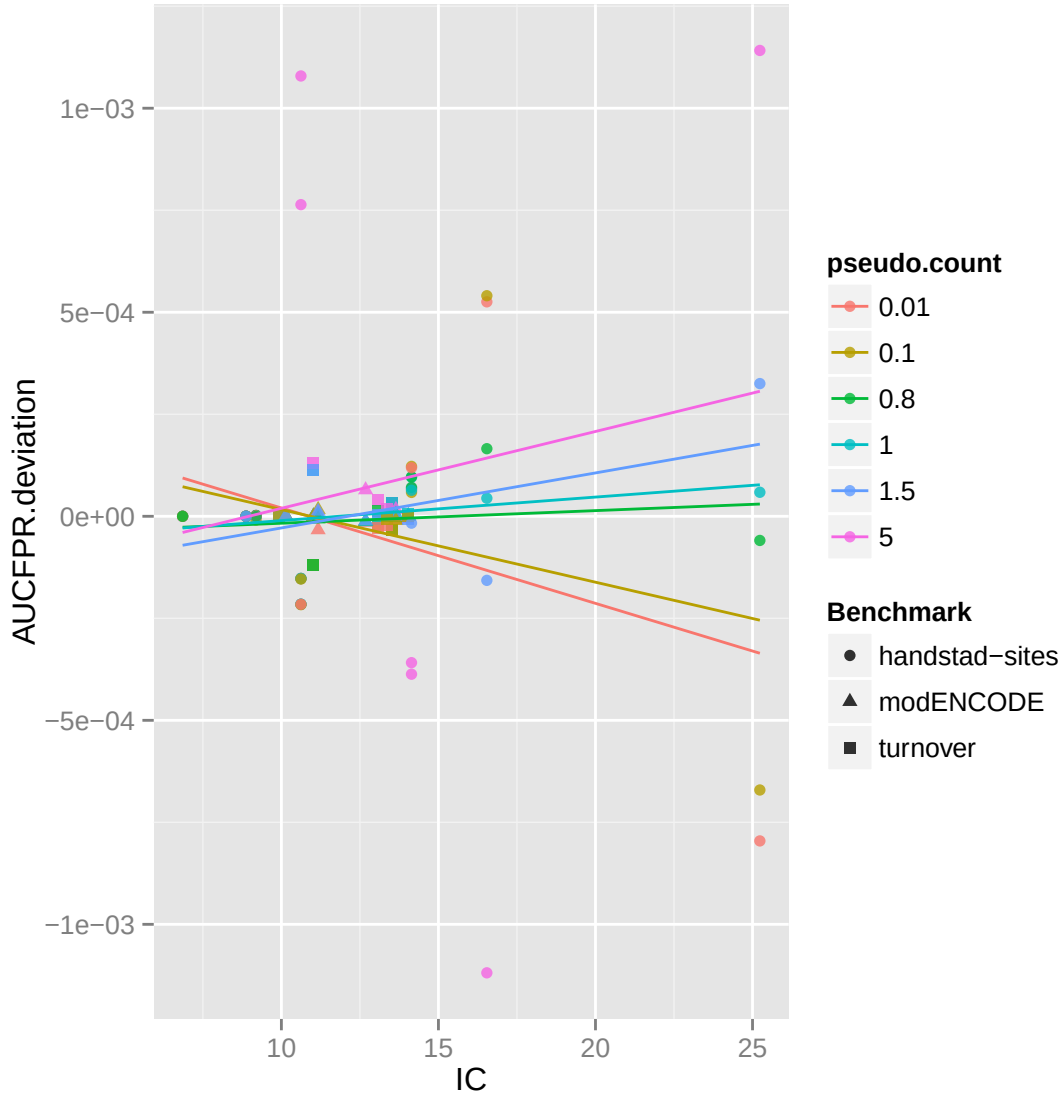


Figure 2.7: The effect of choosing different pseudo-counts for the FIMO algorithm. A scatter plot of information content of the PWM against deviation from the mean AUCFPR for that benchmark/TF combination. Each pseudo-count/benchmark/TF combination is a separate point coloured by pseudo-count and shaped according to the benchmark. The lines show linear models fitted for each pseudo-count.

The AUCFPR statistics are plotted in Figure 2.8. I used a one-sided Wilcoxon paired signed-rank test to investigate which methods performed better at the 0.05 significance level and give the results in Table 2.4. FIMO is dominated by all three other methods and MotEvo dominates all three other methods. The significance test does not discriminate between the performance of MONKEY and BiFA.

Further to these pair-wise comparisons of methods, I investigated what correlation there was between method performance and the information content of the PWMs. The

	FIMO	BiFA	MONKEY
BiFA	**0.018**		0.9
MONKEY	**0.0063**	0.11	
MotEvo	**0.001**	**0.0069**	**0.028**

Table 2.4: One-sided Wilcoxon paired signed-rank test p-values to compare method AUCFPR performance. The p-value is the result of a test whether the method of the row has better AUCFPR statistics than the method of the column. Empty cells represent tests that were not performed. A bold font indicates the test is significant at the 0.05 level.

results are plotted in Figure 2.9. The most striking features of the plot are that the phylogenetic methods, BiFA, MONKEY and MotEvo perform relatively well on PWMs of lower information content. In contrast the non-phylogenetic method FIMO does better on PWMs of higher information content. This is in agreement with the findings of Håndstad et al. Most of the variation in the plot is from the Håndstad sites benchmark: the benchmarks from *Drosophila* are more tightly clustered. The fitted linear models do not show the dominance of MotEvo as clearly as the significance tests as they are heavily influenced by the statistics for the outlying NRSF PWM that has an information content above 25 bits.

I investigated the outliers from the comparison more closely. I define outliers as in the box plots in this chapter: any AUCFPR more than 1.5 times the IQR away from the nearest of the first or third quartiles. FIMO had three outliers which were all low AUCFPRs: both cell lines of E2F4 data and the c-Fos data from the HeLa-S3 cell line. E2F4 has a palindromic short CG-rich PWM with the second lowest information content of all the PWMs in the study (see Figure 2.1). The ROC curves in Figures 2.10 and 2.11 show that FIMO is disadvantaged by my (and Håndstad et al.'s) cautious policy of handling ties of scores (see Section 1.3.10). To see why, note that the FIMO ROC curves for E2F4 and E2F4-hela start as step functions. The first step represents those positive and negative sub-regions of the test cases where there was a perfect match for the consensus sequence of E2F4. However, these steps cover the FPR=5% threshold for the AUCFPR statistic. The cautious tie-handling policy assigns an AUCFPR of 0 to FIMO. An agnostic tie-handling policy would result in a diagonal line instead of a step. Under this policy FIMO would have approximately the same AUCFPR as MONKEY in both cases. To understand why this affects FIMO more than the other methods, consider that the shorter a PWM is, the more perfect matches there will be to its consensus sequence in the central sequences. However, the phylogenetic methods have the related sequences to analyse. Variation in these related sequences means that their scores will take more distinct values than those of FIMO. This leads to the steps in their ROC curves being much thinner giving a more diagonal shape and higher AUCFPR.

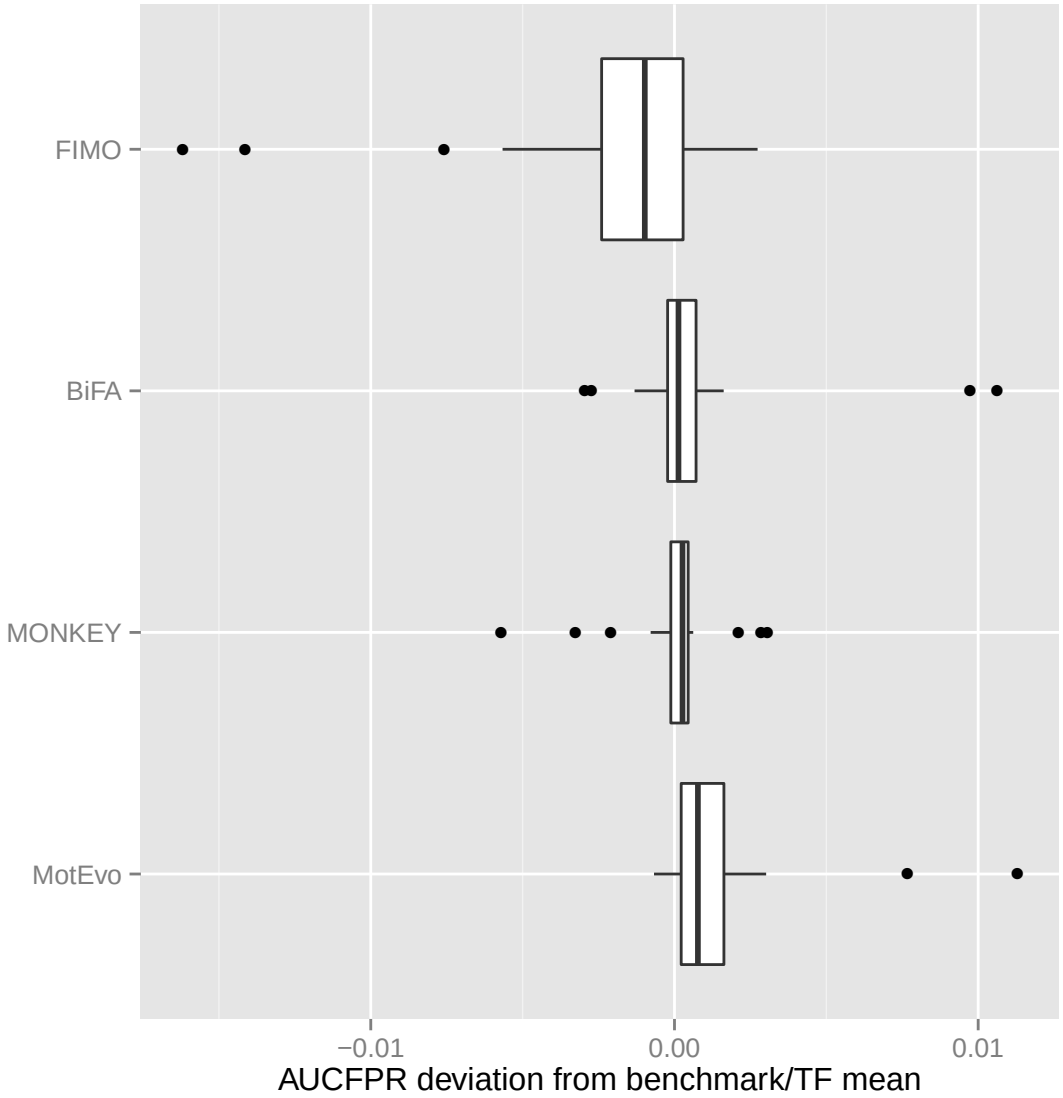


Figure 2.8: The AUCFPR statistics for the four methods. The box plots represent the deviation from the AUCFPR mean for each TF/benchmark combination. The methods are sorted by their median deviation. If sorted by their mean deviation, the BiFA and MONKEY algorithms would swap positions.

Hence there is a bias against FIMO in the AUCFPR and AUC statistics.

The BiFA algorithm had four outliers: two positive outliers for E2F4 in both cell lines (see Figure 2.10) and two negative outliers for MAX and NRSF (see Figure 2.12). MONKEY's outliers were E2F4 (both cell lines) and NRSF in the negative direction and c-Fos (both cell lines) and c-Jun in the positive direction. MotEvo's only outliers were E2F4 (both cell lines) in the positive direction.

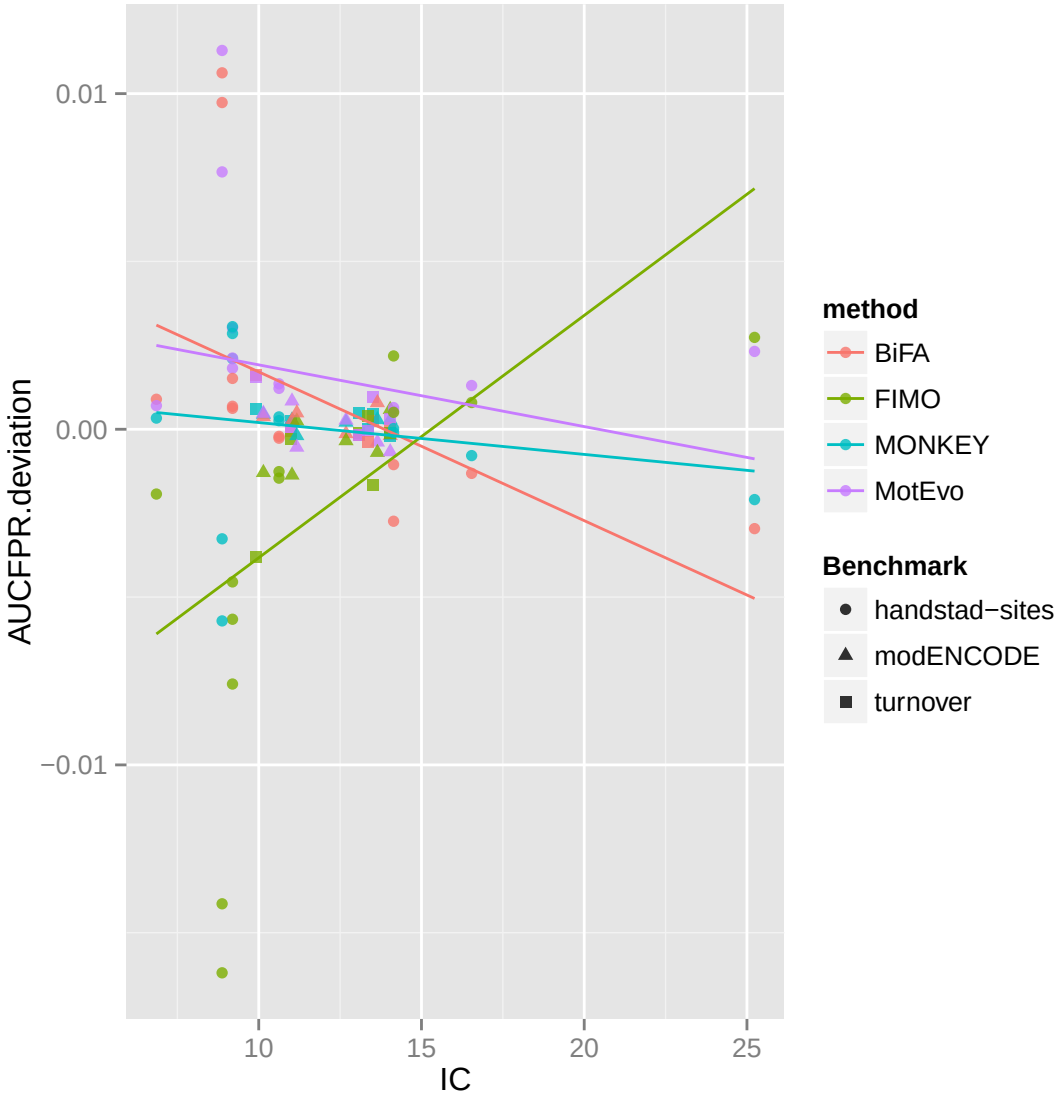


Figure 2.9: Relationship of AUCFPR statistics for the four methods with information content. The deviation from the AUCFPR mean for each TF/benchmark combination is plotted against the information content in bits of the PWM. Each method is identified by colour and has a fitted linear model shown. The shapes identify which benchmark the deviation is from.

BiFA update method

In the description of the BiFA algorithm, I presented two different methods for updating BiFA's belief that there is a TFBS in the central sequence using evidence from the related sequences (see Section 2.3.1). When optimising the BiFA parameters, spearmint chose to use the original updates. I wanted to test how much difference the choice of update method makes to the AUCFPR over all the test cases. I ran BiFA once with each method and calculated the AUCFPR. The differences between the AUCFPR statistics for each

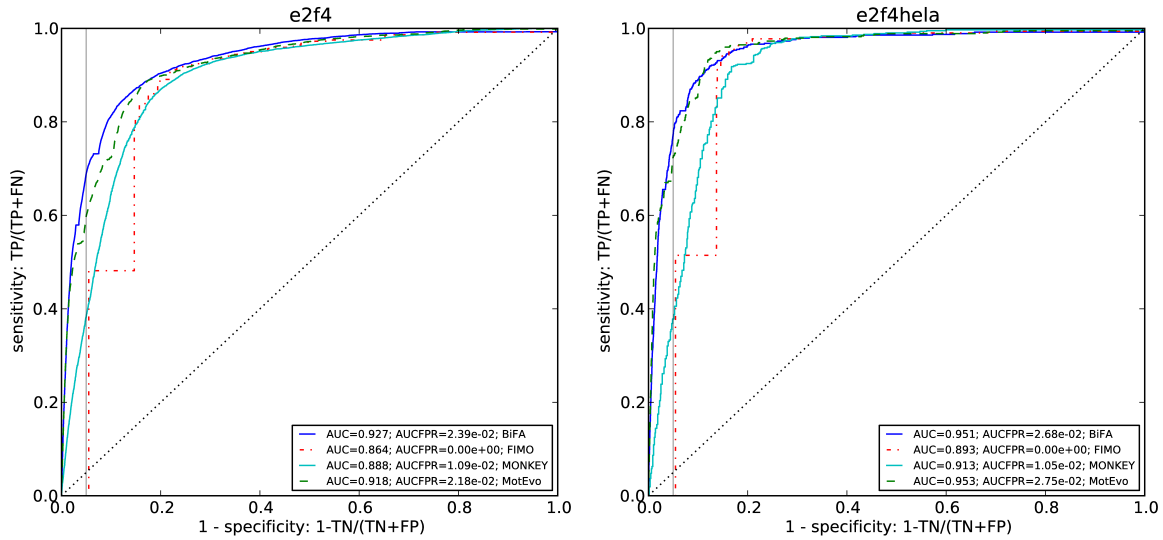


Figure 2.10: ROC curves for the TF E2F4 from the Håndstad sites benchmark. *Left*: the ROC curve for the K562 cell line. *Right*: the ROC curve for the HeLa-S3 cell line. The feint vertical lines show the FPR=5% threshold. The black dotted lines show the expected performance of a random classifier.

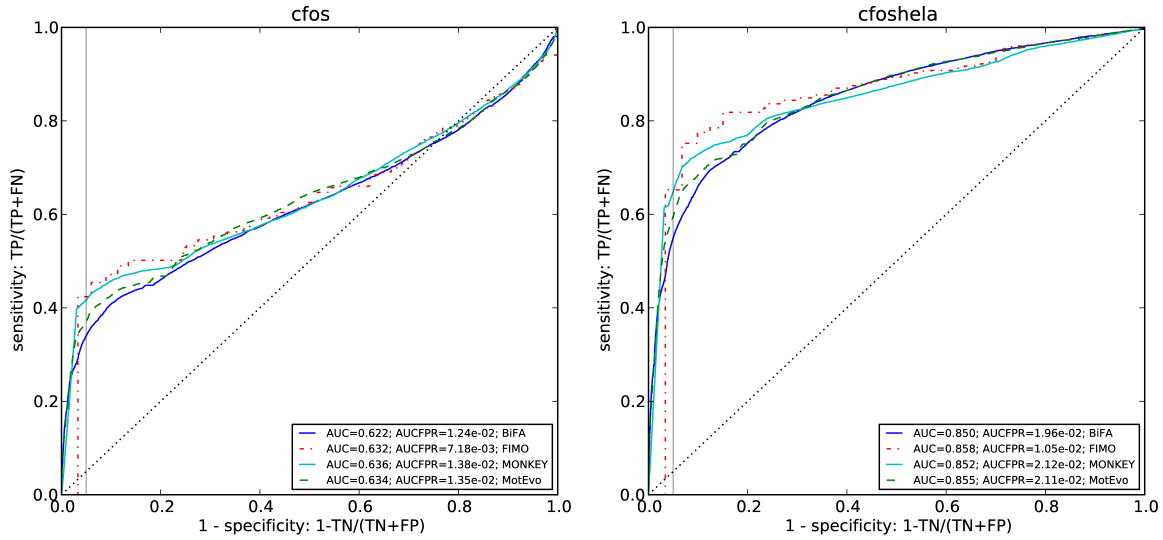


Figure 2.11: ROC curves for the TF c-Fos from the Håndstad sites benchmark. *Left*: the ROC curve for the K562 cell line. *Right*: the ROC curve for the HeLa-S3 cell line. The feint vertical lines show the FPR=5% threshold. The black dotted lines show the expected performance of a random classifier.

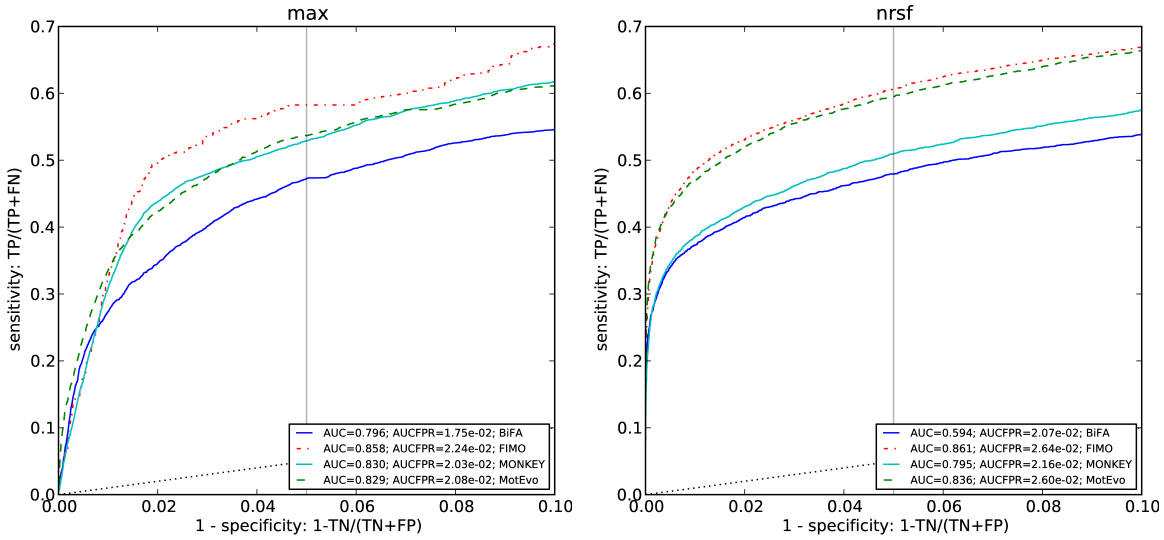


Figure 2.12: ROC curves for the TFs MAX and NRSF from the Håndstad sites benchmark. *Left*: the ROC curve for MAX. *Right*: the ROC curve for NRSF. The x-axis is scaled to show FPRs between 0% and 10%. The feint vertical lines show the FPR=5% threshold. The black dotted lines show the expected performance of a random classifier.

benchmark/TF combination are quite small (of the order of 0.1%). Nevertheless, I evaluated both update methods against each other using two one-sided Wilcoxon paired signed-rank tests. Neither method was significantly better at the 0.05 level: p-values of 0.3 (respectively 0.7) for the test of whether modified update method was better than the original (respectively vice versa). The AUCFPR deviations are plotted in Figure 2.13 in relation to the information content of the PWMs.

Effect of alignment size

I ran the BiFA, MONKEY and MotEvo methods on alignments of two, three and five species on the Håndstad sites and modENCODE benchmarks. I did not use the turnover benchmark for this comparison as it only has two species. Box plots of the AUCFPR statistics are shown in Figure 2.14. There are two trends evident: firstly MotEvo outperforms MONKEY which outperforms BiFA and secondly all three methods perform better with fewer sequences in the alignment.

I examined the difference in AUCFPR between alignments of two and five species for each benchmark/TF/method combination in relation to the information content of the TF's PWM. The results are scatter-plotted in Figure 2.15. For most benchmark/TF/method combinations, performance was degraded when adding species to the alignment. However, there is a negative correlation with information content and the BiFA method

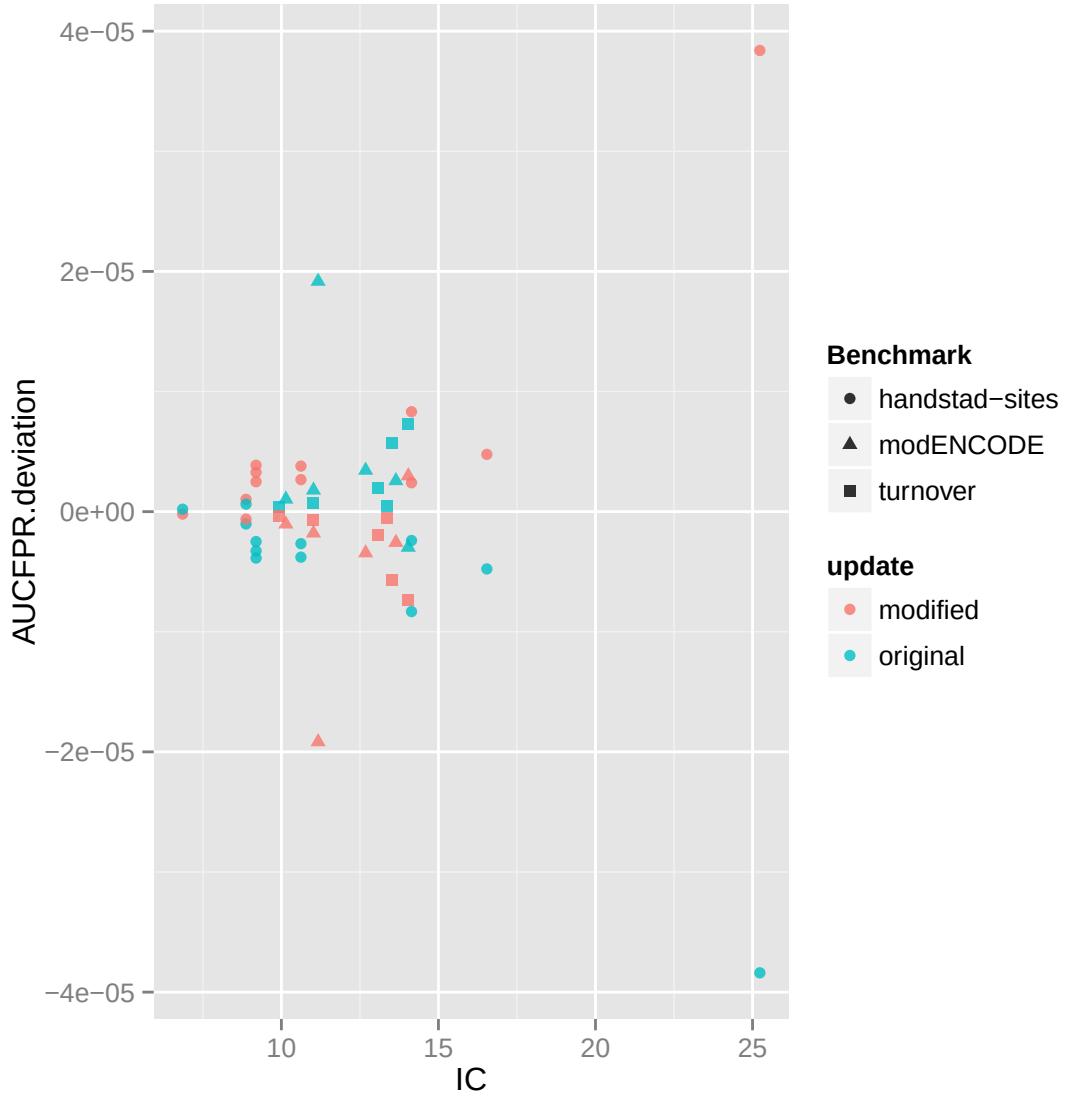


Figure 2.13: Relationship of AUCFPR statistics for the two different BiFA update methods with information content. The deviation from the AUCFPR mean for each TF/benchmark combination is plotted against the information content in bits of the PWM. Each update method is identified by colour. The shapes identify which benchmark the deviation is from.

shows this effect most strongly.

I tested whether each of the BiFA, MONKEY and MotEvo methods performed better with two species than five using a one-sided Wilcoxon paired signed-rank test. At the 0.05 significance level, both MONKEY and MotEvo do perform better with fewer species (p-values of 0.00021 and 0.027 respectively) but BiFA does not (p-value of 0.35).

I decided to investigate a particular test case for which this difference was most marked. I looked for the test case which had the largest drop in AUCFPR between the MotEvo

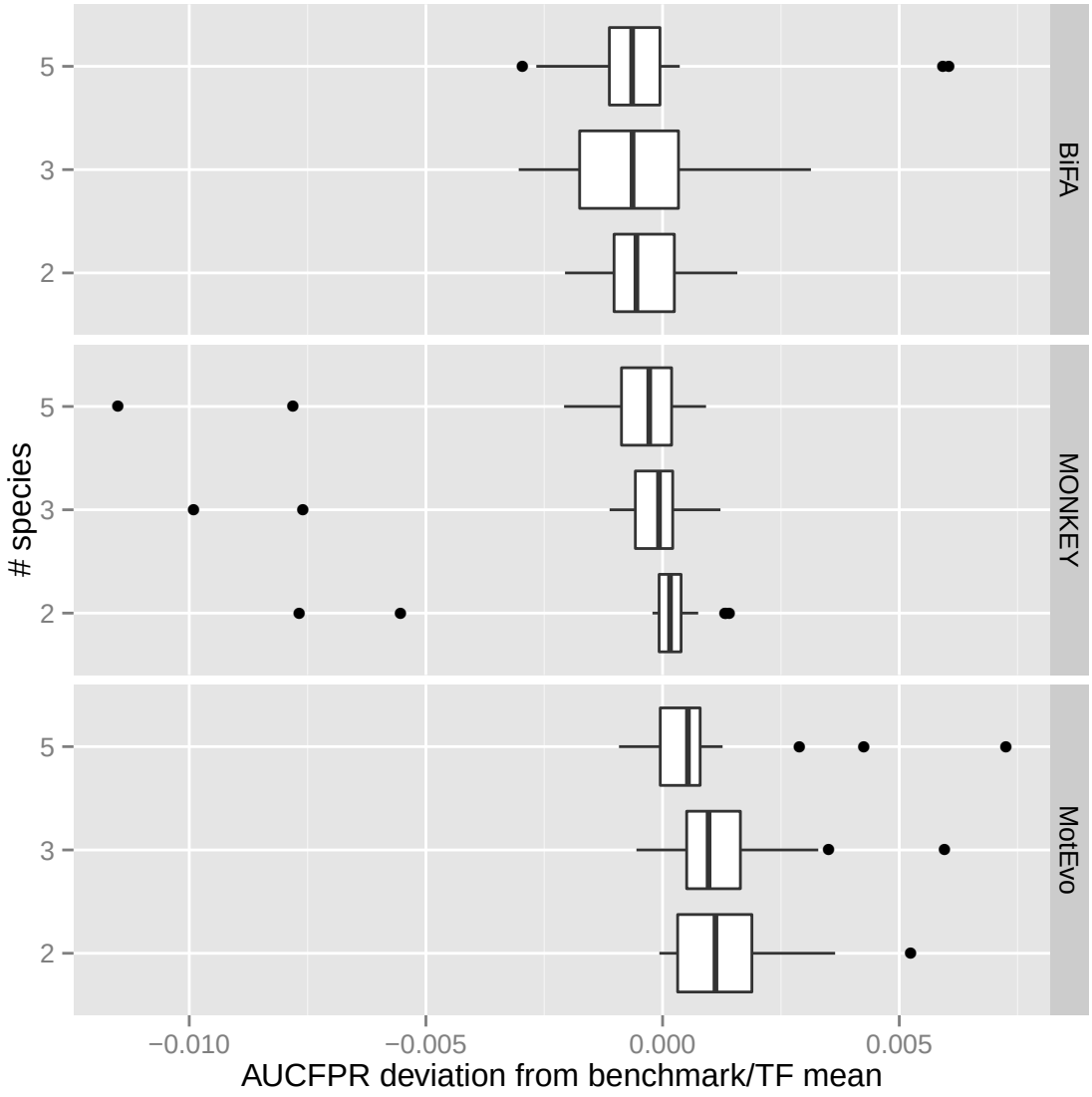


Figure 2.14: AUCFPRs for the BiFA, MONKEY and MotEvo methods on alignments of two, three and five species. The mean of all the AUCFPRs for each benchmark/TF combination is calculated. The box plots show the deviations from this mean for each combination of method with number of species.

method with two species and with five species. Whilst this extreme example may not be typical of the differences when the number of species is varied, I hoped it would provide some insight. This test case is for the knirps TF from the modENCODE benchmark and starts at position 11.535,473 of chromosome chr2R. On closer inspection of this locus and MotEvo's predictions on the two-way and five-way multiple alignments of it, I determined that a strong putative TFBS for knirps was being ignored in the five-way alignment and that this was responsible for the loss of AUCFPR. The alignment and TFBS are shown in Figure 2.16. The alignment of the TFBS in the three extra species

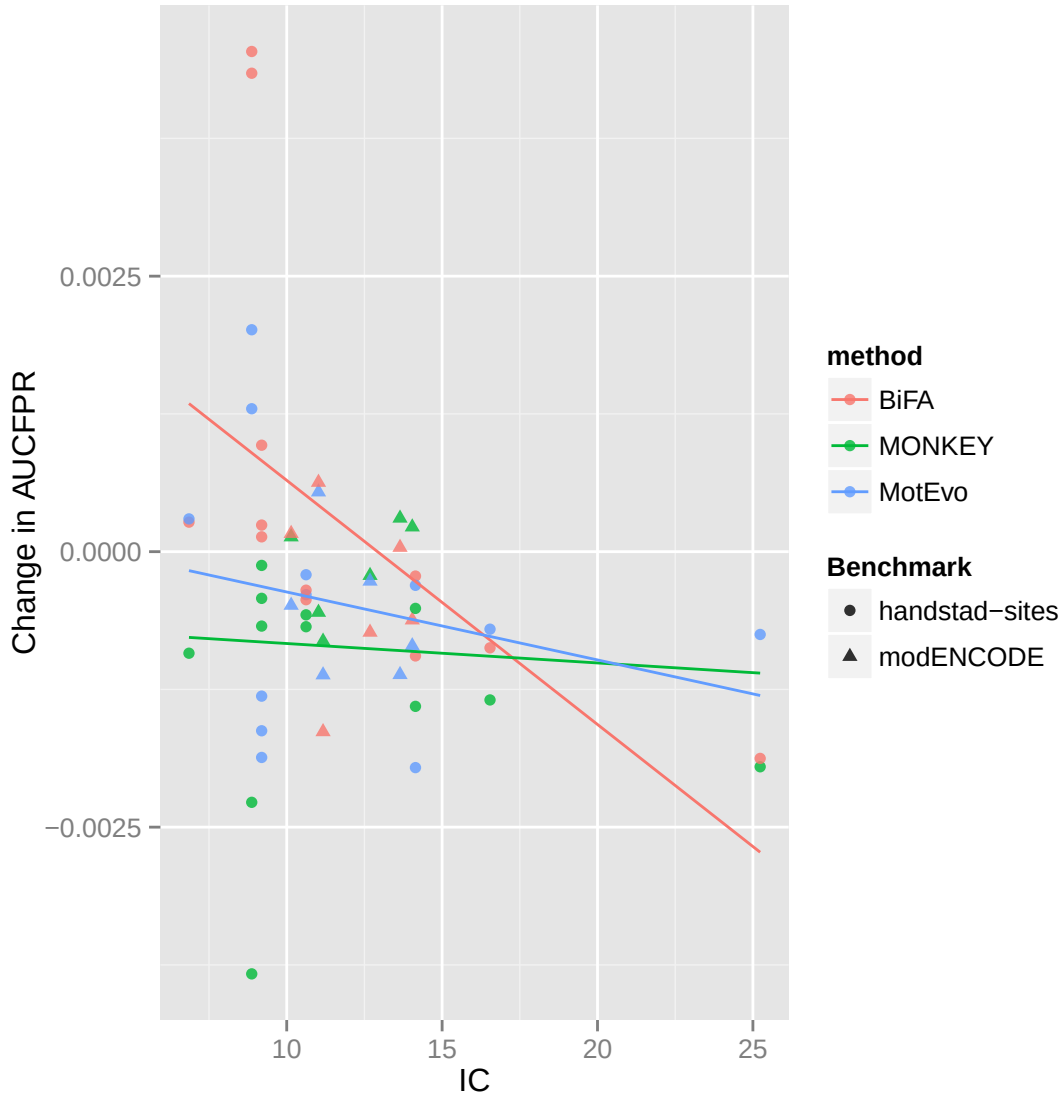


Figure 2.15: Relationship between change in AUCFPR when increasing alignment size from two species to five and the information content of the TF's PWM in bits. Each point represents a benchmark/TF/method combination. The x-axis shows the information content of the PWM for the TF. The y-axis shows the improvement in AUCFPR when moving from an alignment of two species to an alignment of five species. That is positive values represent combinations for which including more species in the alignment improved the performance of the method. The lines show linear models fit to each method.

contains a gap at the penultimate base. MotEvo will not consider TFBSs in species for which the alignment contains a gap. It is easy to see that despite these gaps, a strong putative TFBS exists in all five sequences. Even if the extra three sequences had significant substitutions and the TFBS was almost erased in them, I would argue there is hardly much less evidence for the TFBS than in the two-species alignment. MotEvo was designed to allow for site loss between species and this is presented as one of its features in its publication. The drop in posterior probability for this one TFBS does not seem consistent with this design. This TFBS seems to highlight two potential problems with the MotEvo algorithm: firstly, it ignores aligned TFBSs which have almost insignificant gaps in them and secondly, the reported posterior probability of a TFBS can drop more quickly than seems reasonable when TFBSs are missing in related sequences.

```
    dm GGGG---CG-TGT-GGTTCTAAACTAGACTAACAGATAGGTTTCTTCG-ATT
droWil -------AA-TTG-GGTTCTAAACTAAACT-ACAGACAAGTTTCTTCGTTTT
droVir TTATCTATA-GTT-GGTTCTAAACTAGACT-AC-TACAAGTTTCTTCG-ATT
droYak GTGT---CG-GGT-GGTTCTAAACTAGACTAACAGATAGGTTTCTTCG-ATT
    dp -------GA-TTT-GGTTCTAAGCTAGACT-ACAGACATGTTTCTTCG-AAT
```

Figure 2.16: A putative knirps TFBS inside a positive sub-region of a test case for which MotEvo weakens its prediction when more species are added to the alignment. *Drosophila melanogaster* (dm) and *Drosophila yakuba* (droYak) are the species in the two-way alignment with which MotEvo scores the TFBS relatively highly (posterior probability=0.033). When the other three species are included in the alignment, MotEvo scores the TFBS relatively lowly (posterior probability=0.00068). *Top*: a five-way multiple alignment, those species that are only in the five-way alignment are coloured red, those species that are in both the two-way and five-way alignment are coloured blue. 13 bases from the aligned `T` are coloured in each sequence. Note the gap in the putative TFBS in the three extra species. *Bottom*: PWM for the TF knirps (13bp).

2.4.5 Discussion

Decoding transcriptional regulatory networks is difficult and has only been attempted for a small fraction of networks. One source of relevant data is the conservation of TFBSs across related species. The evolutionary pressures on this conservation are not well understood. Phenomenon such as TFBS turnover through site gain, loss or movement have been identified but we do not know how prevalent these events are. One way we can investigate these effects is through the evaluation of models of this phenomenon,

using predictive performance of the models on experimental data as a proxy for model accuracy. The idea is that models that perform well capture an inherent facet of the evolutionary pressures involved. The implementation of a benchmark framework including three large data sets for two model organisms enabled the investigation of many issues related to phylogenetic TFBS prediction. I believe using a framework of this size and variety is essential to determine which models of phylogenetic conservation of TFBSs perform best in a prediction setting.

If evolutionary pressures on transcriptional regulatory networks and TFBSs varied between genera, some methods might be more suited to the analysis of data from particular genera. No such systematic variation in performance that correlated with the two species or the three benchmarks was apparent. This supports the hypothesis that the evolutionary pressures on transcriptional regulatory networks that our models can detect are similar across genera.

Phylogenetic vs. non-phylogenetic

The first point to note from the results of the comparison is that phylogenetic methods perform better than FIMO, the non-phylogenetic method studied. This comes with a proviso that the method of calculating the AUCFPR statistics is biased against methods such as FIMO whose distribution of possible scores has narrower support. Visual inspection of the ROC curves in Appendix A reveals that the picture is not as clear as the Wilcoxon tests of the AUCFPR statistics suggest. I used the same policy as Håndstad et al. to break ties and calculate AUC statistics. Håndstad et al. also noted that this bias might exist although they did not publish all the ROC curves from which their statistics are generated. They showed curves for E2F4 in which the bias is evident and curves for NRSF which do not display the bias as expected for a longer PWM. Their claim that phylogenetic methods perform better on short and information-poor PWMs covers exactly the cases where the bias is strongest. Hence further work is needed to either confirm the effect or establish if it is an artefact of this bias.

Supposing that the bias is not too strong, the results of this study broadly agree with Håndstad et al.'s claim that phylogenetic methods are to be preferred when the PWM is short or information-poor. However, we found that MotEvo appeared to perform as well as FIMO on information-rich PWMs (for example see the ROC curves for NRSF, NFκB and Max in Appendix A.1). MotEvo allows for site loss between species and this is consistent with its comparable performance to FIMO on these TFs. Both MONKEY and the BiFA algorithm heavily penalise TFBSs that have been lost in related sequences. This is a likely cause of their poor performance on information-rich PWMs.

Aligned vs. alignment-free

Of the phylogenetic methods, BiFA is the only alignment-free method. MONKEY and MotEvo score putative TFBSs by directly examining the aligned segments of the related sequences for matches to the PWM. Conservation of a putative TFBS across many species is certainly strong evidence that it is functional. However, there is a growing body of evidence that points to frequent site turnover (loss, gain and movement) between closely related species. Both MONKEY and MotEvo penalise such turnover (MONKEY more than MotEvo). Their advantage lies in identifying lower substitution rates at putative TFBSs than at surrounding neutrally evolving sequence. The BiFA algorithm's ability to model site turnover does not seem to compensate for the disadvantage of not using the alignment directly. This study has not been able to quantify how important these two effects are. This would be easiest using a single model that incorporated both concepts.

Alignment-free methods do have a practical advantage over those that require alignments. The BiFA algorithm just needs the central and related sequences as input. Both MONKEY and MotEvo require the specification of a phylogenetic tree with branch lengths and a multiple alignment. Multiple alignments for model organisms are readily available from the UCSC browser. Phylogenetic trees with branch lengths can be more difficult to come by. Both tools have fairly strict requirements on how these are presented to the algorithm (MONKEY especially). Conversion of the tree and alignment into the correct format can be error-prone and time-consuming.

Simple models of TFBS conservation

There is some evidence that methods based on simple models of TFBS conservation perform well. In my tests, MONKEY's simple model surpassed the more complicated evolutionary models. MONKEY's simple model is equivalent to averaging log Bayes factors across the TFBS alignment (c.f. the BiFA algorithm). All information about evolutionary substitution rates and phylogenetic distances is ignored by this model. The WS method in Håndstad et al.'s work is very similar: a weighted average of the log-likelihood scores where the weights are a function of the branch lengths. The WS method uses the highest score in a window around the aligned TFBS in the related sequences. This window technique is presumably less sensitive to problems of mis-alignments. The WS method was better than BBLS when evaluated at all thresholds using the AUC statistic. It may not have performed so well with the AUC50 statistic due to the bias highlighted earlier. In Xie et al.'s evaluation of their BBLS method, they compare it to a method that simply reports TFBSs that are conserved across human, mouse and

rat at some PWM-specific threshold. On some of their data sets this score performs comparably to BBLS. To summarise, most evaluations of complex phylogenetic TFBS predictors have not proven their superiority to simpler methods, either in general or on a subset of TFs.

Effect of alignment size

Interestingly, I did not discover that increasing the number of species in the alignment significantly improved the performance of any of the phylogenetic methods. Indeed, I found the opposite for MONKEY and MotEvo: their performance significantly decreased. The authors of neither method studied this effect directly in their publications, nor did Håndstad et al. in their comparison. The authors of MotEvo presented some evidence that their method improves when given more aligned sequences. MotEvo can be used as an enhancer predictor by looking for cluster of TFBSs. The authors showed that MotEvo's performance at this task improved on alignments of more species.

The performance decrease is disturbing in that the availability of more information to an algorithm should not degrade its performance. This effect suggests the PMM models in MONKEY and MotEvo can be improved upon. Perhaps the simplest explanation is that the performance decrease could be caused by over-penalising the loss of sites. As more species are added to an alignment, the chance that a site is lost in one or more species (or that there is a mis-alignment) grows. If a method is overly sensitive to this loss it may perform worse on these larger alignments.

Consistency with previous evaluations

There have been two dedicated evaluations of phylogenetic TFBS predictors [Hawkins et al., 2009, Håndstad et al., 2011]. Also, the publications of the methods described earlier all present some comparison of their methods with other TFBS prediction methods. In this section I discuss how my results relate to the results in these evaluations. I have already discussed how the results in this thesis relate to Håndstad et al.'s results in Section 2.4.5 above.

Hawkins et al.'s work was on yeast data. They found that phylogenetic methods were worse than non-phylogenetic methods at predicting known sites. They argued this could be because of missing weak TFBSs. They used a column-shuffling technique to estimate a minimum FDR for different methods at different thresholds. They found that MONKEY outperformed a non-phylogenetic method, consistent with the results from my study.

MotEvo's authors evaluated their method against MONKEY and PhyloScan on data for five human TFs. As I did, they found MotEvo performed significantly better than MONKEY. They showed that MotEvo's ability to predict enhancers increased as more sequences were added to the alignment. As noted above, my results suggest MotEvo's performance decreases as more sequences are aligned which is in contrast to their enhancer predicting results. This certainly bears further investigation. MotEvo's authors chose to use seven species in their evaluations. Considering the methods' models of site loss, this may have disadvantaged MONKEY more than MotEvo.

The disparity between the conclusions drawn from separate evaluations of phylogenetic TFBS predictors highlights how much is unknown about TF binding. On the one hand, complex aligned methods are reported as out-performing other simpler methods. On the other hand, a closer look at the statistics reveals a more complicated picture where simpler methods can surpass more complex models. Ultimately I expect continued iterations of model checking via predictive performance and model updating will help us understand how TFBSs evolve.

2.4.6 Further work

There is much work that could be done to further investigate models of TFBS conservation. I describe some ideas in this section.

Improving the benchmarks

In terms of this benchmark framework, I think the priority must be to re-analyse the methods' performances using a more equitable policy to break tied scores. Whilst a bias exists against non-phylogenetic or simpler methods, no firm conclusions can be drawn.

The BBLS method performed very well in Xie et al.'s evaluation. It is probably the strongest candidate for the next method to include in the framework although there are others such as rMonkey, PhyloScan and some of the simpler methods such as WS.

Integration of the data sets that have already been used for comparison of methods would allow my results to be directly compared to those evaluations. In addition, using data from other model organisms such as *Arabidopsis*, bacteria, yeast, *Caenorhabditis elegans* and mouse would add more weight to the results. There is no longer a shortage of good experimental binding data in these organisms.

Some of the features of MotEvo, especially its model of competitive binding, require PWMs for more than one TF for proper evaluation. Similarly the maximal chain extension to the BiFA algorithm only makes sense when predicting TFBSs for more than

one TF. The benchmark framework as presented does not implement any tests of this nature. This is also the context in which most experimentalists will use these methods: typically, an experiment will suggest a locus in the genome is of interest and an experimentalist will want to scan for TFBSs of any of a number of TFs. Tests where methods are asked to report which of a number of TFs might bind to a locus would enable better comparisons of these models.

These multi-TF tests would also help evaluate if method scores are comparable across PWMs. As mentioned earlier, some PWM scanning approaches use per-PWM thresholds to assess significance. Methods that report comparable scores across PWMs without the need for calibration are far more useful. Multi-TF tests would help investigate how methods perform in this regard. For instance p-values of predictions decrease monotonically as log-likelihood ratios increase but only on a per-PWM basis: the same log-likelihood ratio for different PWMs will have a different p-value. The introduction of more than one PWM into a single test case would allow us to investigate if the posterior probability of binding methods described in Section 2.1.2 are better calibrated for comparison than p-value methods.

Hawkins et al. and Håndstad et al. drew contradictory conclusions regarding whether phylogenetic TFBS predictors are better at identifying strong or weak TFBSs. Integration of some measure of TFBS strength into the benchmark framework would allow this question to be addressed. An obvious choice is the peak height in ChIP-seq data as Håndstad et al. have already done.

Improving the methods

The question of how many species to include in an alignment for the purposes of TFBS prediction has not been answered in any work to date. This could be because performance appears to suffer as the number of species increases. Understanding which models perform best with more sequences is key to identifying models that capture how TFBSs evolve. Similarly, the question of which evolutionary distances are optimal for use in TFBS prediction has not been addressed. Some recent work has identified liver-specific and heart-specific enhancers that are poorly conserved [Blow et al., 2010, Schmidt et al., 2010, May et al., 2011]. The ideal distances could depend on the type of transcriptional regulatory network under investigation. Some networks may be shared across large evolutionary distances, others may be specific to individual species.

To evaluate how well the core BiFA algorithm models billboards and the maximal chain extension models enhanceosomes, it would be worthwhile to hand-curate a collection of known billboards and enhanceosomes.

0-order Markov models do not fit genomic sequences particularly well. Recent work has shown that models of order up to 7 can be optimal for certain sequence analysis problems [Narlikar et al., 2013]. I discuss this in more detail in Section 5.3.2. In any case, although many of the available methods will work with higher-order models, little work has been done evaluating the methods' performance when they are used.

Finally, whilst there are clearly some advantages to be had from considering an explicit alignment and evolutionary model, the PMM methods that take this approach (MONKEY and MotEvo) seem too strict in some cases where TFBSs have been lost. On the one hand, simple log-likelihood based averaging approaches that do not use an alignment may not penalise these cases so harshly. On the other hand, they are not directly using information from the alignment about TFBSs that are under evolutionary constraint. An approach that tries to combine the best of both worlds would be interesting to investigate. One idea is a consensus meta-method that integrates the output of two or more methods (perhaps MotEvo and BBLS). Alternatively, perhaps the MotEvo model could be extended to handle site loss better than it currently does.

Chapter 3

The STEME motif search algorithm

The work in this chapter has been published in Nucleic Acids Research [Reid and Wernisch, 2011] under the authorship of myself and my supervisor, Lorenz Wernisch. My contributions to the paper were the main technical idea for the efficient approximation, the implementation of this idea, the accuracy and efficiency tests and the majority of the text of the paper.

3.1 Introduction

MEME [Bailey et al., 1994] is one of the most popular motif finders. It has a long pedigree: the original version was published in 1994. MEME was one of the best performing motif finders in a comparative benchmark review [Tompa et al., 2005]. MEME has a large user-base that understand its parameters and trust its results: the primary paper describing its algorithm is cited over 300 times on PubMed. Unfortunately MEME takes a prohibitively long time to run on large data sets. The MEME authors acknowledge this and recommend discarding data from large data sets in order to make run-times practical. They suggest a limit of 200,000 base pairs on the size of input data set. In our experience, the users of MEME are not always aware of this advice and can be frustrated when using MEME on large data sets. In any event, discarding data is a far from ideal work-around as it necessarily detracts from the power of the method. Hence there is a need to make MEME and other motif finders more efficient. This chapter details an efficient approximation to the EM algorithm that is a core component of MEME and many other motif finders.

Various attempts have been made to speed up MEME in recognition of its poor performance on large data sets. The authors of MEME have implemented a parallel version of MEME, ParaMEME [Grundy et al., 1996]. Other approaches use specialised hardware

such as parallel pattern matching chips on PCI cards [Sandve et al., 2006] or off-loading the computations onto powerful GPUs [Chen et al., 2008]. All of these techniques require hardware that is not commonly available to the typical researcher.

In this thesis we propose an alternative route to accelerate MEME by using suffix trees. A suffix tree [Gusfield, 1997] is a data structure that represents a sequence or set of sequences. Suffix trees are well suited to algorithms that require efficient access to subsequences by content rather than by position. They have been used in several areas of bioinformatics: sequence alignment [Schatz et al., 2007], indexing genome-scale sequences [Phoophakdee and Zaki, 2008] and short read mapping [Mäkinen et al., 2010]. They have also been used for combinatorial motif finding [Federico and Pisanti, 2009, Marsan and Sagot, 2000, Pavesi et al., 2001] and scanning for PWMs [Beckstette et al., 2006]. To the best of our knowledge the work presented here is the first application of suffix trees to probabilistic motif finding and the EM algorithm in particular.

MEME is not the only motif finding algorithm. Other motif finders have been proposed specifically for large data sets. One such motif finder is DREME [Bailey, 2011], developed by an author of MEME. DREME does not use a full probabilistic model: it searches in the space of IUPAC words for those that discriminate best between an input sequence set and a background sequence set. It is designed to find short motifs and does not scale well to find longer motifs.

In Section 3.2 we describe MEME's algorithm, MEME's probabilistic model and how MEME uses the EM algorithm to optimise its parameters. We describe an approximation to EM and show how suffix trees can be used to implement this approximation. We call this approximation the Suffix Tree Expectation-Maximisation Motif Elicitation (STEME) algorithm We analyse the efficiency gains we expect to achieve with this approximation. We describe our open source implementation of the STEME algorithm. In Section 3.3 we describe the tests we have done to establish the accuracy and efficiency of STEME in practice. We examine the effect of varying the motif width and the main parameter in our algorithm on the accuracy and efficiency. In Section 3.4 we look at the implications of the results and suggest how our algorithm can be best used. We conclude with an outlook for future work.

3.2 Materials and methods

3.2.1 MEME

The MEME algorithm is based on a probabilistic model. W-mers in the input sequences are modelled as a mixture of draws from a background model and a model that represents

the motif being searched for. MEME uses the EM algorithm to improve the motif model iteratively. In each iteration, the locations of the binding sites are estimated using the current model of the motif and the motif is updated using the predicted sites weighted by their likelihoods. The EM algorithm is guaranteed to converge to a local maximum of the likelihood function but is very sensitive to initial conditions. To mitigate this sensitivity, MEME runs the EM algorithm many times from different starting points (also known as *seeds*). MEME tests every W-mer in the input sequences as a potential seed and runs the EM algorithm on the most promising seeds. The work in this chapter provides an efficient replacement for MEME's EM algorithm. Efficiently finding the most promising seeds on large data sets is outside the scope of this work.

MEME's model

For a particular motif width, W, MEME treats every subsequence of length W (henceforth W-mer) in the data independently. Given a motif width, W, MEME models each W-mer in the sequences as an independent draw from a two-component mixture. One mixture component models the background sequence composition, the other models binding sites. The binary latent variables, $Z = \{Z_1, \ldots, Z_N\}$, indicate whether each W-mer, X_n, is drawn from the background component or the binding site component. MEME has several different variants of its model which the user can choose between. They vary in how the sites are distributed amongst the sequences. The *oops* variant insists that there is exactly One Occurrence Per Sequence. For most experimental data this is not a realistic assumption and those sequences that do not contain a site can reduce MEME's ability to find the motif. The *zoops* variant allows Zero or One Occurrences Per Sequence. This is more plausible for most experimental data sets but will not take statistical strength from more than one site in a sequence. The *anr* variant allows any number of binding sites in each sequence. This variant is the most flexible and is the most suitable for most applications. However, it is also the most computationally demanding: care must be taken in the algorithm when sites overlap otherwise MEME will tend to converge on self-overlapping motifs. This is because MEME's assumption that the W-mers are independent breaks down as each W-mer will overlap with up to $2(W-1)$ other W-mers. Nevertheless, as homotypic clusters of binding sites are common in transcriptional networks, we will focus on this *anr* variant in the rest of this chapter.

In the *anr* variant the background component is modelled using a Markov model parameterised by θ_{BG}, the binding site component is modelled by a position-weight matrix parameterised by θ_{BS}, and λ parameterises the probability that any given W-mer is

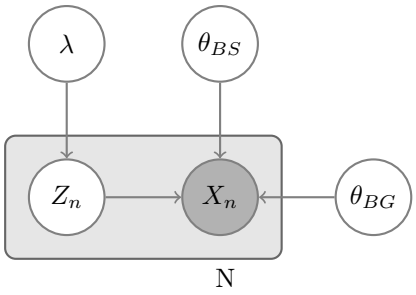


Figure 3.1: MEME's model: λ, the prior probability of a binding site; Z_n, the hidden variable representing whether the n'th W-mer is an instance of the motif; X_n, the n'th W-mer; θ_{BS}, the parameters of the motif; θ_{BG}, the parameters of the background distribution.

drawn from the binding site component. Thus the model is

$$p(Z_n = 1|\lambda) = \lambda \tag{3.1}$$

$$p(X_n|Z_n, \theta_{\mathrm{BG}}, \theta_{\mathrm{BS}}) = p(X_n|\theta_{\mathrm{BS}})^{Z_n} p(X_n|\theta_{\mathrm{BG}})^{1-Z_n} \tag{3.2}$$

where $\{X_1, \ldots, X_N\}$ are the W-mers and $\{Z_1, \ldots, Z_N\}$ are latent variables indicating whether the W-mers are drawn from the background or binding site model. This gives the joint distribution

$$\begin{aligned} & p(X, Z|\lambda, \theta_{\mathrm{BG}}, \theta_{\mathrm{BS}}) \\ = & \prod_{n=1}^{N} p(Z_n|\lambda) p(X_n|Z_n, \theta_{\mathrm{BG}}, \theta_{\mathrm{BS}}) \\ = & \prod_{n=1}^{N} [\lambda p(X_n|\theta_{\mathrm{BS}})]^{Z_n} [(1-\lambda) p(X_n|\theta_{\mathrm{BG}})]^{1-Z_n} \end{aligned} \tag{3.3}$$

The model is depicted in plate notation in Figure 3.1.

Expectation maximisation

In the E-step of expectation maximisation MEME derives the expected value of the log-likelihood, LL, w.r.t. the latent variables, Z, given the current parameter estimates, $\theta = \{\theta_{\mathrm{BS}}, \theta_{\mathrm{BG}}, \lambda\}$. All expectations, $\langle . \rangle_{Z|\theta}$, are w.r.t. $Z|\theta$ unless specified.

$$\langle LL \rangle = \langle \log p(X, Z|\lambda, \theta_{\mathrm{BG}}, \theta_{\mathrm{BS}}) \rangle \tag{3.4}$$

$$\begin{aligned} = & \sum_{n=1}^{N} \langle Z_n \rangle \log[\lambda p(X_n|\theta_{\mathrm{BS}})] \\ & + (1 - \langle Z_n \rangle) \log[(1-\lambda) p(X_n|\theta_{\mathrm{BG}})] \end{aligned} \tag{3.5}$$

From Equation (3.3) and an application of Bayes' theorem

$$\langle Z_n \rangle = \frac{\lambda p(X_n|\theta_{\text{BS}})}{\lambda p(X_n|\theta_{\text{BS}}) + (1-\lambda)p(X_n|\theta_{\text{BG}})} \tag{3.6}$$

The M-step maximises the expected log-likelihood w.r.t. each parameter in turn to calculate their new estimates. On inspection of (3.5) we can see

$$\begin{aligned} \lambda &\mapsto \arg\max_{\lambda} \sum_n \langle Z_n \rangle \log \lambda + (1 - \langle Z_n \rangle) \log(1-\lambda) \\ &= \frac{\sum_n \langle Z_n \rangle}{N} && (3.7) \\ \theta_{\text{BS}} &\mapsto \arg\max_{\theta_{\text{BS}}} \sum_n \langle Z_n \rangle \log\, p(X_n|\theta_{\text{BS}}) && (3.8) \\ \theta_{\text{BG}} &\mapsto \arg\max_{\theta_{\text{BG}}} \sum_n (1 - \langle Z_n \rangle) \log\, p(X_n|\theta_{\text{BG}}) && (3.9) \end{aligned}$$

MEME uses a PWM model for binding sites where $\theta_{\text{BS}} = \{\theta_{wb}\}$. θ_{wb} parameterises the probability of seeing base b at position w in a TFBS.

$$p(X_n|\theta_{\text{BS}}) = \prod_w \theta_{wX_{nw}} \tag{3.10}$$

Here X_{nw} is the wth base of the n'th W-mer. The update equations are

$$\theta_{wb} \mapsto \frac{\sum_n \langle Z_n \rangle \mathbb{I}(X_{nw} = b)}{\sum_n \langle Z_n \rangle} = \frac{c_{wb}}{S} \tag{3.11}$$

where $c_{wb} = \sum_n \langle Z_n \rangle \mathbb{I}(X_{nw} = b)$ is the expected number of times we see base b at position w in a binding site and $S = \sum_n \langle Z_n \rangle$ is the expected number of binding sites

MEME uses a 0-order Markov model for θ_{BG}. This is updated by the expected counts of the bases which are not in binding sites.

If MEME just used the EM algorithm as described above to update its estimates of the $\langle Z_n \rangle$ it would run into problems when the estimate of the motif allows for overlapping instances. For instance, suppose that there are 12 consecutive `A`s in the data and the current estimate of the motif models binding sites of 8 consecutive `A`s. MEME would assign $\langle Z_n \rangle \approx 1$ to the five 8-mers in the consecutive `A`s. The sum of the window. To avoid this situation MEME's algorithm leaves the highest $\langle Z_n \rangle$ unchanged and scales the others down so that they sum to at most 1. Without this adjustment, repetitive sections in the input sequences can cause MEME to converge on motifs of low complexity that have frequently overlapping binding sites.

Expected running time

Each iteration of MEME's EM algorithm evaluates the current estimate of the motif on each W-mer taking $O(NW)$. The algorithm to adjust for overlaps also runs in $O(NW)$ time hence an iteration of EM completes in $O(NW)$ time. However, it should be noted MEME's algorithm as a whole is quadratic in N as the number of seeds is proportional to N.

3.2.2 Approximation to EM

The updates in the M-step of the EM algorithm all involve sums of the form $\sum_n \langle Z_n \rangle \ldots$ where n ranges over all W-mers in the data set. In any given iteration of EM, depending largely on the current θ_{BS}, most of these $\langle Z_n \rangle$ will be negligible. We can make an approximate M-step by ignoring those n for which $\langle Z_n \rangle$ is small. We formalise this by defining a subset of the n thresholded by $\langle Z_n \rangle$

$$T_\delta = \{n : \langle Z_n \rangle \geq \delta, 1 \leq n \leq N\} \tag{3.12}$$

Intuitively, T_δ indexes those W-mers that match our current motif estimate. As an approximation to Equation (3.11) we define $\hat{\theta}_{wb}, \hat{c}_{wb}, \hat{S}$

$$\hat{\theta}_{wb} = \frac{\sum_{n \in T_\delta} \langle Z_n \rangle \mathbb{I}(X_{nw} = b)}{\sum_{n \in T_\delta} \langle Z_n \rangle} = \frac{\hat{c}_{wb}}{\hat{S}} \tag{3.13}$$

For convenience of notation we define $\bar{c}_{wb} = \sum_{n \notin T_\delta} \langle Z_n \rangle \mathbb{I}(X_{nw} = b)$, $\bar{S} = \sum_{n \notin T_\delta} \langle Z_n \rangle$, and $\bar{N} = N - |T_\delta|$ so that

$$\begin{aligned} c_{wb} &= \hat{c}_{wb} + \bar{c}_{wb} \\ S &= \hat{S} + \bar{S} \\ N &= |T_\delta| + \bar{N} \end{aligned}$$

The relative error, ϵ_δ, in our approximation $\hat{\theta}_{wb}$ of θ_{wb} is

$$\begin{aligned} \epsilon_\delta &= \frac{\theta_{wb} - \hat{\theta}_{wb}}{\theta_{wb}} = \frac{c_{wb} - \frac{S}{\hat{S}} \hat{c}_{wb}}{c_{wb}} \\ c_{wb} \epsilon_\delta &= \bar{c}_{wb} - \left(\frac{S}{\hat{S}} - 1 \right) \hat{c}_{wb} = \bar{c}_{wb} - \frac{\bar{S}}{\hat{S}} \hat{c}_{wb} \end{aligned}$$

Noting that $\frac{\bar{S}}{\hat{S}}$ and all the counts, c, are positive

$$|\epsilon_\delta| \leq \max\left\{\frac{\bar{c}_{wb}}{c_{wb}}, \frac{\bar{S}\hat{c}_{wb}}{\hat{S}c_{wb}}\right\} \tag{3.14}$$

We know from the definition of T_δ that $\bar{S} \leq \delta\bar{N}$ and that $\bar{c}_{wb} \leq \delta\bar{N}$

$$|\epsilon_\delta| \leq \max\left\{\frac{\delta\bar{N}}{\hat{c}_{wb}}, \frac{\delta\bar{N}}{\hat{S}}\right\} \leq \frac{\delta N}{\hat{c}_{wb}} \tag{3.15}$$

So $\delta \geq \frac{|\epsilon_\delta|\hat{c}_{wb}}{N}$. If we knew $\hat{c}_{wb}$ we would know which δ would guarantee our desired bound in the motif estimation relative error. Unfortunately at the beginning of an EM iteration when we want to choose δ we do not know $\hat{c}_{wb}$. In practice this is not a problem. Given λ and the current motif parameters we can estimate $\hat{c}_{wb}$ fairly accurately. When the EM iteration has completed, $\hat{c}_{wb}$ is available. We can check the above equations to ensure that the relative error is less than $|\epsilon_\delta|$. If it is not, we can calculate a new δ for which the relative error is guaranteed to be small enough and re-run the iteration. In our tests this was never necessary. Also $\hat{c}_{wb}$ tends to change slowly over iterations, this makes its estimation straightforward in all but the first iteration.

3.2.3 Suffix trees

A suffix tree is a data structure that stores a sequence or a set of sequences. Typically, sequences are stored as contiguous buffers. This permits fast access to subsequences indexed by their position. Suffix trees are alternative data structures that allow efficient access to subsequences by their content. Re-writing algorithms to use suffix trees can often achieve significant efficiencies.

Suppose we have a sequence, $Y = y_1 \ldots y_T$. A suffix of Y is any subsequence, $y_t \ldots y_T$, that ends at y_T. A suffix tree stores every suffix of the given sequence(s) in a tree structure. An example of a suffix tree is shown in Figure 3.2. Now we show how to iterate over all the subsequences of length W in Y (the W-mers). Each such W-mer is the start of a suffix. Hence descending the tree to depth W iterates over all the W-mers. If two W-mers have the same content, they will be represented once by the same path in the tree. Contrast this with the random-access of a typical contiguous buffer data structure for sequence storage. A contiguous buffer permits fast random access to a W-mer at a given position but takes no account of identical or similar W-mers. If we have an application where we are not interested in the position of the W-mers, a suffix tree can be a more efficient data structure to iterate over them.

Another attractive property of suffix trees is that they can be constructed in linear time

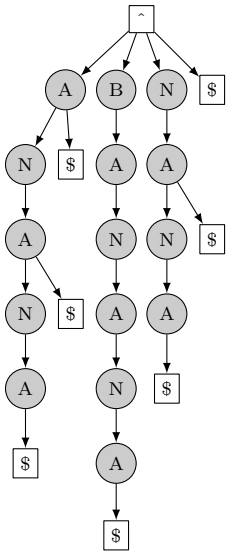


Figure 3.2: A suffix tree that represents the sequence "BANANA". The beginning of a suffix is represented by the ^ symbol and $ is a termination symbol. Note that the subsequence "ANA" occurs twice in the sequence but is represented once in the tree.

and space [McCreight, 1976, Ukkonen, 1995, Giegerich and Kurtz, 1997]. One might suppose that when the sequence has T distinct characters, constructing a suffix tree might take $O(T^2)$ time as the distinct suffixes have sizes $T, T-1, \ldots, 1, 0$. However, each edge in the suffix tree need not be labelled with the entire part of the suffix it represents. It can be labelled with a start and end index into the sequence in $O(1)$ time and $O(1)$ space. This leads to construction algorithms that are $O(T)$ in time and space. Despite the simplicity of this idea, efficient suffix tree construction is not straightforward and is the subject of much ongoing research. In particular, techniques to efficiently index genome sized sequences have recently been developed [Phoophakdee and Zaki, 2008, Barsky et al., 2008, Barsky et al., 2009, Mansour et al., 2011].

Suffix tree EM efficiencies

The EM algorithm must visit every W-mer in the sequences once on each iteration. By using a suffix tree to enumerate all the W-mers we immediately achieve two efficiency improvements. Firstly, if any two W-mers are identical, the $\langle Z_n \rangle$ calculations are equivalent and we do not repeat them. Secondly as we descend the suffix tree to enumerate the W-mers, we make partial evaluations of our current motif on what we have seen of the W-mer so far. These partial evaluations are shared across all the W-mers below the current node in the tree. On the other hand MEME evaluates every base in each W-mer once.

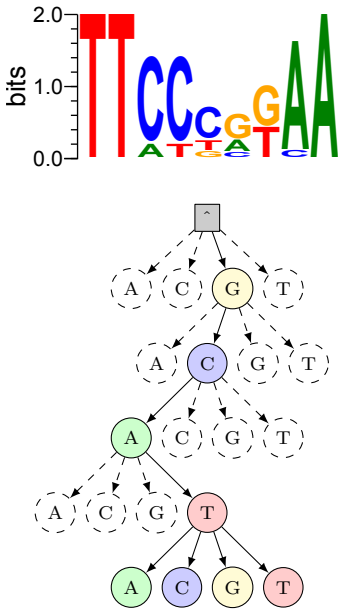


Figure 3.3: An illustration of how the STEME branch-and-bound algorithm works. *Top*: The current estimate of the motif in the EM algorithm. This is actually the motif for Stat5 from the TRANSFAC database (M00223). *Bottom*: Part of the suffix tree representing the sequences. We can see that if we have descended the tree to the node that represents the prefix, `GCAT`, our match to the motif is poor. If the bound for the $\langle Z_n \rangle$ of all the nodes below this is small enough, we can stop our descent here.

3.2.4 Branch-and-bound

Recall from Equation (3.12) that we need to identify all n with $\langle Z_n \rangle \geq \delta$ for a given δ. We iterate over the W-mers by descending the suffix tree. Suppose we have an upper bound on the $\langle Z_n \rangle$ of all the W-mers below any node. If this bound is below δ, then we can ignore the entire branch of the suffix tree below the node. In this way we avoid evaluating large parts of the tree that do not fit the current estimate of the motif well. We illustrate the idea in Figure 3.3.

We define X_n^{w-} as the prefix of X_n of length w and X_n^{w+} as the suffix of length $W - w$ (so that $X_n = X_n^{w-} X_n^{w+}$). We can write the likelihoods of the X_n in terms of their prefixes and suffixes

$$\begin{aligned} p(X_n|\theta_{\mathrm{BS}}) &= p(X_n^{w-}|\theta_{\mathrm{BS}}) p(X_n^{w+}|\theta_{\mathrm{BS}}) \\ p(X_n|\theta_{\mathrm{BG}}) &= p(X_n^{w-}|\theta_{\mathrm{BG}}) p(X_n^{w+}|\theta_{\mathrm{BG}}) \end{aligned}$$

We can enumerate the W-mers in the data by descending a suffix tree. Each node we visit represents the prefix of all of the W-mers below it. Given our binding site and background models we can calculate the $p(X_n^{w-}|\theta_{\mathrm{BS}})$ and $p(X_n^{w-}|\theta_{\mathrm{BG}})$ exactly. Sup-

pose we can also bound $p(X_n^{w+}|\theta_{\mathrm{BS}})$ from above and $p(X_n^{w+}|\theta_{\mathrm{BG}})$ from below. Recalling (3.6) we can use these bounds to bound $\langle Z_n \rangle$ above. In more detail, suppose $p(X_n^{w+}|\theta_{\mathrm{BS}}) \leq \overline{p(X_n^{w+}|\theta_{\mathrm{BS}})}$ and $p(X_n^{w+}|\theta_{\mathrm{BG}}) \geq \underline{p(X_n^{w+}|\theta_{\mathrm{BG}})}$ then using 0 as a lower bound for $p(X_n^{w+}|\theta_{\mathrm{BS}})$ we have

$$\langle Z_n \rangle \leq \overline{\langle Z_n \rangle} = \frac{\lambda p(X_n^{w-}|\theta_{\mathrm{BS}})\overline{p(X_n^{w+}|\theta_{\mathrm{BS}})}}{(1-\lambda)p(X_n^{w-}|\theta_{\mathrm{BG}})\underline{p(X_n^{w+}|\theta_{\mathrm{BG}})}} \tag{3.16}$$

The upper bounds, $\overline{p(X_n^{w+}|\theta_{\mathrm{BS}})}$ are easy to calculate from θ_{BS}. In practice the background model does not change very much over the course of the EM algorithm as only a small fraction of the base pairs are explained as binding sites. Therefore we keep the background model fixed and precompute the lower bounds, $\underline{p(X_n^{w+}|\theta_{\mathrm{BG}})}$, in an initialisation step.

3.2.5 Expected efficiencies

In order to understand the computational savings this approximation achieves we give an analysis of a simplified example. We investigate the expected fraction of nodes we ignore at each depth in our descent of the suffix tree.

Suppose our current estimate of the PWM has a preferred base at each position. Each preferred base has probability a and the other three bases at each position are equally likely with probability $\frac{1-a}{3}$. When $a = 1$ our PWM is equivalent to a consensus sequence, when $a = \frac{1}{4}$ our PWM has a uniform distribution. As $\hat{c}_{wb} \approx \lambda N a$ we set $\delta = \epsilon\lambda a$ where ϵ is the maximum relative error we will tolerate. Suppose also our background model is a uniform 0-order Markov model, then $p(X_n|\theta_{\mathrm{BG}}) = 4^{-W}$. As $1 \approx 1 - \lambda$ and recalling (3.16), we want to know when the following holds

$$\langle Z_n \rangle \leq \overline{\langle Z_n \rangle} = \frac{\lambda p(X_n^{w-}|\theta_{\mathrm{BS}})a^{W-w}}{4^{-W}} \leq \delta = \epsilon\lambda a$$

Let Y be the number of preferred bases in X_n^{w-}. Assuming that X_n^{w-} is drawn from our background distribution we have $Y \sim \mathrm{Binomial}(w, \frac{1}{4})$. Now $\log p(X_n^{w-}|\theta_{\mathrm{BS}}) = Y \log a + (w - Y) \log \frac{1-a}{3}$. Hence whenever

$$Y \leq \frac{\log \epsilon - W \log 4 - (W-1)\log a}{\log a - \log(1-a) + \log 3} + w \tag{3.17}$$

we can ignore all the nodes with prefix X_n^{w-}. For any given values of ϵ, W, and a, the expected fraction of nodes ignored at depth w is the probability that Equation (3.17) holds. As Y is distributed according to a binomial distribution, these values can be read

directly from the binomial cumulative distribution function. We plot these expected fractions for some parameter values in Figure 3.4.

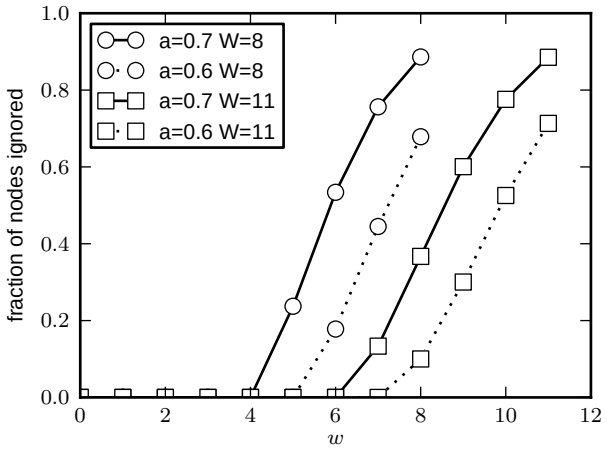


Figure 3.4: The probability of discarding a W-mer drawn from a uniform 0-order Markov background at different depths, w, in the suffix tree. Here we used $\epsilon = .4$. As explained in the text, a represents how sharp the current estimate of the motif is. The higher a is, the sharper the motif. Examining the graph reveals that with a moderately sharp motif ($a = .7$) of width 8 we can expect to discard over half the nodes in the tree at depth $w = 6$.

3.2.6 Open source implementation

We have implemented the STEME algorithm in C^{++} as an open source library. For the suffix tree implementation we used the SeqAn library [Döring et al., 2008]. In addition to the C^{++} interface we have implemented a Python scripting interface to make it more accessible. The codes are tested on Linux with GCC 4.4 and Python 2.6 but should work with any modern C^{++} compiler and version of Python 2 newer than 2.5. Our implementation is available at `http://sysbio.mrc-bsu.cam.ac.uk/johns/STEME/` Our implementation requires 500Mb of memory to work with data sets of up to 13Mb, which is well within the range of modern desktop or laptop machines. Building the suffix tree for such a data set takes 19 seconds on my laptop. These space and time requirements scale linearly in the size of the input.

3.2.7 Test data sets

We used data from two sources for our tests (see Table 3.1): a set of six smaller ChIP-chip and ChIP-seq data sets we had previously worked with [Reid et al., 2010]; and five larger data sets from the ENCODE project [Consortium, 2004].

The six data sets we had previously worked with were prepared as follows. The data for Sp1 were extracted from TRANSFAC professional 11.4 and the flanking bases added by TRANSFAC were removed. The data sets for GABP, Stat1, Stat5a and Stat5b were processed to extract the binding site sequences using the cisGenome software suite v1.0 [Ji et al., 2008]. In every case both sequences and controls were used. Binding region boundary refinement was used and then the region extended on each side by 30bp. GABP peaks were selected if there were more than 18 reads in a rolling 100 bp sequence window compared to the control. This higher figure was selected to remove visually noisy peaks and 10,767 peaks were detected. Cutoffs of 30 and 20 reads were used for the Stat5a and Stat5b data respectively yielding 814 and 154 peaks. RepeatMasker was used on all the test data sets to mask repetitive elements using the genomic context for each sequence.

The five larger data sets from the ENCODE project were produced by the Myers Lab at the HudsonAlpha Institute for Biotechnology. We downloaded the data for SRF, ZBTB33, RXRA, TCF12 and CTCF from the ENCODE Data Coordination Center at UCSC.

3.2.8 Tests

In order to test the accuracy and efficiency of the STEME approximation, we ran our STEME implementation and MEME's EM implementation to completion on the data sets.

We wanted to try a range of typical parameters so we ran MEME's seed searching algorithm with the default arguments. We used motif widths of 8, 11, 15 and 20. The number of sites parameter took values of 2, 4, 8, 16, 32, 64, 128, 256 and 500. MEME

TF	Sequences	Base pairs	Publication
Stat5b	144	19,379	[Liao et al., 2008]
Stat5a	737	94,250	[Liao et al., 2008]
Sp1	296	207,325	[Cawley et al., 2004]
GABP	2,275	500,203	[Valouev et al., 2008]
Stat1	2,360	500,409	[Robertson et al., 2007]
SRF	2,155	674,443	[Consortium, 2004]
ZBTB33	3,342	1,589,893	[Consortium, 2004]
RXRA	19,126	8,118,061	[Consortium, 2004]
TCF12	35,714	12,540,202	[Consortium, 2004]
CTCF	41,069	13,214,001	[Consortium, 2004]

Table 3.1: The test data sets.

uses the number of sites parameter to initialise λ and also to look for the best seed (consensus sequence) for the motif. This gave us 6 data sets, 4 motif widths and 6 different number of sites parameters for a total of 144 separate test cases. Additionally we wanted to test the effect of varying the permitted relative error so when we ran STEME, we used ϵs of 0, 0.2, 0.4, 0.6, and 0.8.

Once we had run the test cases we needed some way of comparing the results of the different implementations and the different settings for the permitted relative error, ϵ. Comparison of the resulting PWMs would have been possible but we chose to perform a simplified analysis by converting the resulting PWMs into consensus sequences and using the Hamming distance as a distance metric. To test the accuracy of the STEME approximation, we calculated two statistics: the *mismatch rate*, that is, how often the resulting consensus sequences from the same starting point differed in any base; and the *mismatch fraction*, that is, what proportion of the bases of the resulting consensus sequences differed.

We ran the tests using version 4.5.0 of MEME which was released October 8, 2010. We modified the MEME source code in order to obtain precise timing information for its EM algorithm. The modifications are available as a patch included with the STEME source code.

3.3 Results

3.3.1 How ϵ affects STEME's accuracy

We compared the accuracy of STEME when using different bounds on the relative error, ϵ. When $\epsilon = 0$ no approximations are made and we used this as a baseline for comparison. The average mismatch rate and mismatch fraction statistics are plotted in Figure 3.5. Even when a very large relative error of 0.8 is permitted, only 1 in 6 of the resulting consensus sequences differ and less than 1 in 20 bases differ. When using a reasonable value of $\epsilon = .4$, only around 1 in 8 of the test cases differed from the baseline and only 1 in 30 of the resulting bases differed.

3.3.2 STEME's accuracy relative to MEME

We also analysed the accuracy of STEME relative to MEME. We had hoped that the STEME algorithm with the approximation turned off ($\epsilon = 0$) would produce identical results to MEME. For reasons we present in Section 3.4, this is not the case. These

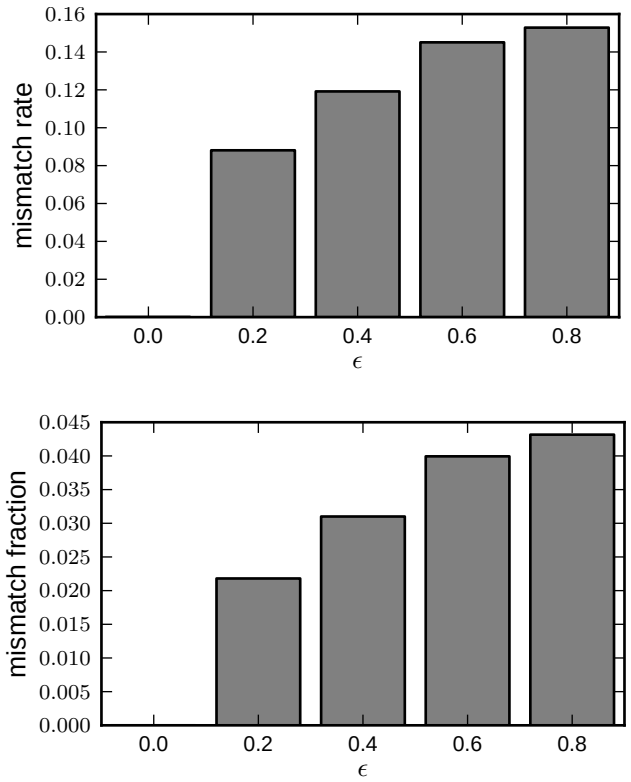


Figure 3.5: An analysis of how increasing the permitted relative error, ϵ, affects the outcome of STEME. The STEME algorithm was run from from the initialisations described in the text for various values of ϵ. *Top*: The mismatch rate: The fraction of resulting consensus sequences that differed from those when $\epsilon = 0$. *Bottom*: The mismatch fraction: The fraction of bases in the resulting consensus sequences that differed from those when $\epsilon = 0$.

results are presented in Figure 3.6. When $\epsilon = 0$, less than 1 in 4 of the test cases had a different outcome but only around 1 in 20 of the bases in the resulting consensus sequences differed. As an example, when the seed `ATCCTGTTCTC` is used with 16 sites on the Sp1 data set, MEME converges to `CTTCCTTCTCT` and STEME converges to `CTCCCTTCTCT`.

3.3.3 Efficiency

We compared the running time for an iteration of STEME to an iteration of the MEME EM algorithm. The relative speeds are dependent on the value of ϵ chosen and on the width of the motif, as shown in Figure 3.7. STEME is significantly quicker than MEME for reasonable values of ϵ and typical motif widths.

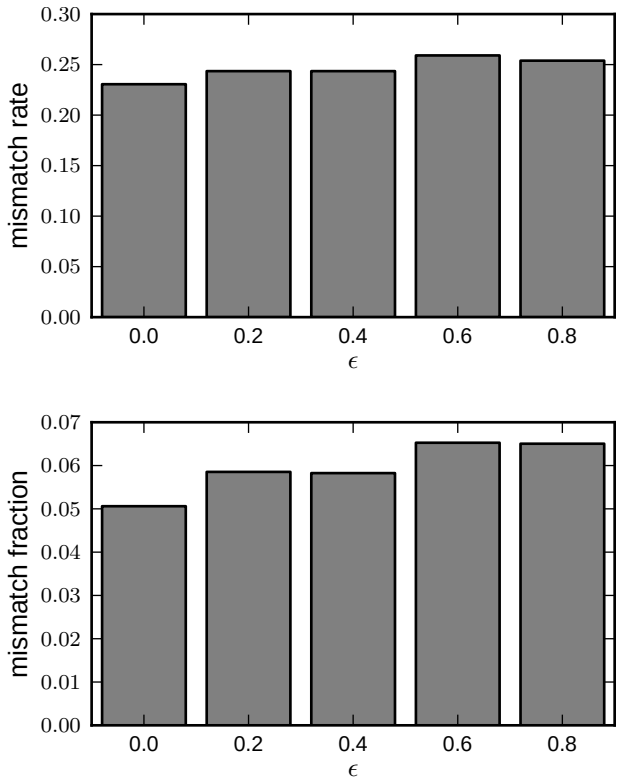


Figure 3.6: An analysis of the accuracy of STEME for various values of ϵ relative to MEME. The STEME algorithm was run from from the initialisations described in the text for various values of ϵ. *Top*: The mismatch rate: the fraction of resulting consensus sequences that differed from the results of MEME. *Bottom*: The mismatch fraction: the fraction of bases in the resulting consensus sequences that differed from the results of MEME.

3.4 Discussion

3.4.1 Accuracy

Examining Figure 3.5 we can see that when $\epsilon = .4$ about one in eight of our applications of EM had some discrepancy with the exact algorithm and about one base in thirty differed overall. In our experience this represents a satisfactory compromise of speed and accuracy. In any case it is not clear if all the differences introduced by the approximation have a negative effect. Our approximation ignores those putative binding sites that are not a good match to the motif rather than discounting them. It could be that by only examining the higher quality binding sites, our algorithm builds a better model of the motif. We hope to investigate this possibility in further work integrating our STEME algorithm in a motif finder.

We also compared STEME without any approximation to MEME's EM implementation, see Figure 3.6. We had hoped the implementations would agree. Unfortunately there

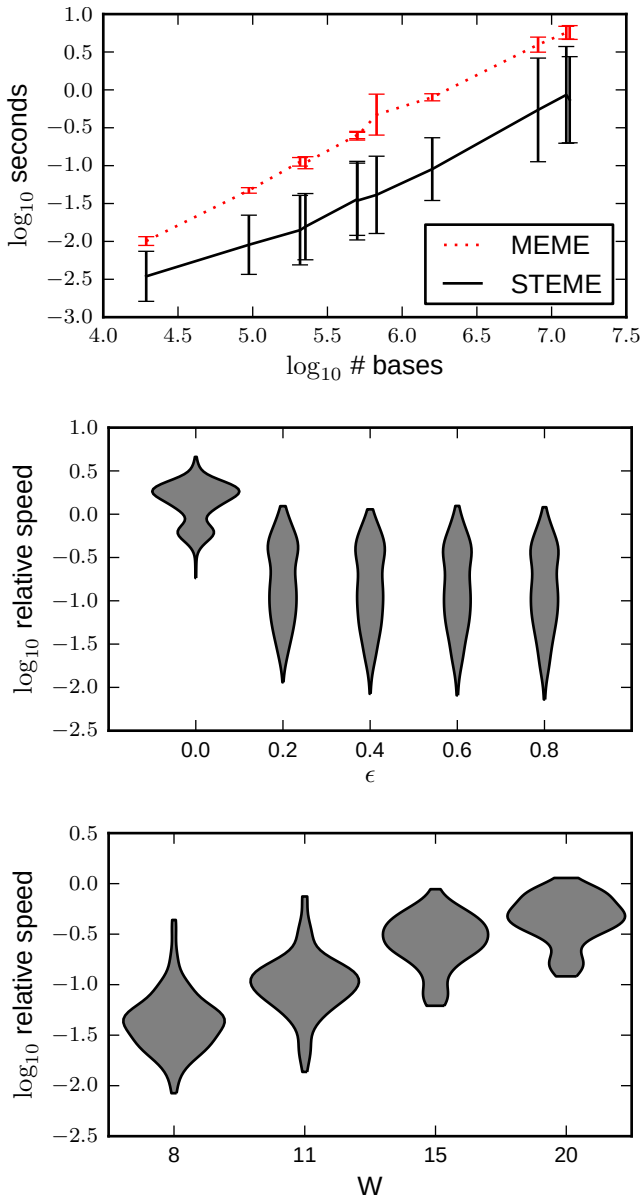


Figure 3.7: A comparison of the speed of STEME and MEME on one iteration of the EM algorithm. *Top*: Using $\epsilon = .4$ as a typical setting, the iteration speeds across all the data sets are plotted on a log 10 scale. The error bars represent the standard deviations. *Middle*: A violin plot of the relative speeds of MEME and STEME grouped by ϵ. With $\epsilon = 0$, STEME can be slower than MEME although we would expect this to reverse on larger data sets. As ϵ grows, STEME's advantage grows. The contours of the violin plots are kernel density estimates that are truncated at the minimum and maximum values. The y-axes are on a log 10 scale. *Bottom*: Using $\epsilon = .4$ as a typical setting, the relative speeds grouped by motif width. For motifs of width 8, STEME is between $10^{.3} \approx 2$ and $10^{2.1} \approx 125$ times quicker than MEME.

were some discrepancies. We spent some time reverse-engineering the MEME source code and discovered some inconsistencies between the published MEME algorithm [Bai-

ley et al., 1995] and the latest implementation. In particular, the handling of reverse complements is not discussed in the published algorithm. STEME treats each draw as a 50-50 mixture between a binding site on the positive strand and a binding site on the negative strand. The MEME implementation essentially doubles the size of the data by adding a reverse-complemented copy of the data. Despite this, STEME and MEME converge on essentially the same motifs. On average, only 1 base in 20 differs.

Interestingly it appears that there is significant overlap between the test cases for which STEME without any approximation differs from MEME and those test cases for which the result of the STEME changes as the permitted error is allowed to grow. This can be seen in Figure 3.6 as the difference between $\epsilon = 0$ and $\epsilon = .4$ is smaller than the analogous difference in Figure 3.5.

3.4.2 Efficiency

Figure 3.7 shows that the speed-up possible through the STEME approximation is dependent both on the width of the motif considered and the relative error tolerated in the estimation of the motif. For motifs of reasonable size ($W = 8$ or 11) an order of magnitude increase in speed over MEME can be expected when using a relative error of $\epsilon = .4$. Our approximation is consistently quicker than MEME's implementation of EM which is already highly optimised. STEME achieves an order of magnitude increase in speed on data sets of moderate size for a wide range of reasonable parameters. In the coming years we expect the average size of data sets to continue increasing.

3.4.3 Applicability

We have not presented a complete motif finder but we have shown how any motif finder that uses the EM algorithm on a compatible model can be adapted to handle larger data sets. We would have liked to have presented an efficient drop-in replacement for MEME but were prevented from doing so for some technical reasons that we elaborate on here.

The EM algorithm is not a motif finder on its own. The result of EM is dependent on how the parameters are seeded. Hence to find motifs, suitable seeds must be found. MEME's search for seeds is inefficient on large data sets. Integrating our fast EM algorithm with MEME's slow search for seeds would offer little benefit as run-times would be dominated by the seed search. We are working on using suffix trees to re-implement MEME's search for seeds more efficiently, however this is a major undertaking in its own right. We have included an implementation of our work-in-progress with the source code for STEME.

It is of practical value for motifs up to width 8 on large data sets (> 500Kb) however the efficiencies tail off quickly as the motif width increases (see Table 3.2). For example, on the 674Kb SRF data set, MEME took over four hours to find a motif of width 8, in contrast our implementation with STEME finished in 13 minutes, 18 times quicker.

In addition, the way that MEME calculates the significance of the motifs involves a preprocessing step that does not scale well to large numbers of sites. Typically a user will want to choose the number of sites proportionally to the number of sequences in the data set. Hence for large data sets, the significance calculation needs to be reimplemented more efficiently. We are working on this using approximations to the LLR p-value calculations.

3.5 Conclusion

Reverse-engineering transcriptional networks remains an important *in silico* challenge. Modern biology continues to generate ever larger data sets and this trend can be expected to continue. Hence there exists a need for good motif finders that can handle large data sets. MEME is well trusted but does not handle these data sets well. We have presented an approximation to EM for models of the type used in the MEME algorithm. We have demonstrated that this approximation has a minor effect on the outcome on the algorithm and is an order of magnitude faster. We have supplied an implementation of this algorithm and hope that it will be incorporated into existing and novel motif finders.

Motif finding is a popular and competitive research area. Perhaps this is due to the simplicity of the problem statement combined with the difficulty of the problem. Many motif finders have been developed which can handle data sets of the size STEME can cope with, for example DREME [Bailey, 2011], Trawler [Ettwiller et al., 2007], ChIP-

TF	Base pairs (Kb)	W	STEME (secs)	MEME (secs)	Speed-up
SRF	674	8	792	14,760	18
ZBTB33	1,590	8	933	78,339	84
TCF12	12,540	8	2,122	4,928,532	2,322
TCF12	12,540	10	27,424	5,176,744	189
TCF12	12,540	12	379,891	4,597,053	12

Table 3.2: Timings for STEME with search for seeds and complete MEME algorithm. The times to run MEME on the TCF12 data set are estimated from partial runs as otherwise they would have taken months to complete.

Munk [Kulakovskiy et al., 2010], Amadeus [Linhart et al., 2008]. However, we hope that the familiarity, success and popularity of the MEME algorithm ensures STEME will find an audience amongst genomic researchers.

Chapter 4

Transcriptional programs

The work in this chapter has been published in BMC Bioinformatics [Reid et al., 2009] under the authorship of myself, Sascha Ott and my supervisor, Lorenz Wernisch. My contributions to the paper were the main technical idea for the application of the model to combinatorial aspects of transcriptional regulation, the implementation of this idea, the analysis of the results and the majority of the text of the paper.

4.1 Background

4.1.1 Combinatorics of transcriptional regulation

As discussed in section 1.1.5 TFs often work in coordinated sets. The idea is that in specific cellular contexts a particular TF might coordinate its activities with distinct sets of TFs. For example, the TFs SOX2, POU5F1 and NANOG are known to form part of a regulatory network maintaining pluripotency in human embryonic stem cells [Boyer et al., 2005]. However eye development is regulated in part by the interaction of SOX2 with another TF, PAX6 [Kondoh et al., 2004]. This is believed to be independent of the interaction of SOX2 with POU5F1 and NANOG in pluripotent embryonic stem cells.

Biophysically these interactions manifest themselves as many different molecular mechanisms, for example: cooperative binding of TFs [He et al., 2010], competitive binding of TFs [Wasson and Hartemink, 2009], mutual transcriptional regulation [Stathopoulos and Levine, 2005], formation of enhanceosomes [Merika and Thanos, 2001]. Current biological experimental techniques cannot directly detect these interactions. Genomic data commonly available today, such as expression data or TF localisation data, typically inform us about individual TFs and/or genes. These data link TFs to genes but

provide little direct evidence of coordination amongst TFs. The higher order structure of interactions between TFs must be inferred from the data.

So on the one hand, inference of biophysical interactions between specific TFs is a goal of this work. On the other hand, cellular behaviour and regulatory networks are frequently studied at the systems level. Due to the number of regulatory connections between TFs and genes, these systems can be difficult to study as a whole. In these cases a summary or view of higher order structure in transcriptional regulation is more suitable. This work also provides such a summary.

4.1.2 Our model

Our model aims to discover cooperative effects between TFs in noisy sequence analysis data. We use a model that has had success in the field of document modelling where the task is to infer the latent topics that best summarise a corpus of documents. Each document is modelled as a mixture of several topics drawn from a shared pool of unknown topics and each topic is modelled as a collection of words. Only the documents are given as input to the model: both the mixtures of topics that characterise the documents and the distributions over words that characterise the topics are inferred from the data.

To explain the use of this model in the context of transcriptional regulation we draw an analogy: in our model a document is analogous to a gene; a word is analogous to a TF and the occurrence of a word in a document is analogous to a binding site in a gene's CRM. To complete the picture, a topic is analogous to what we term a transcriptional program (TP). A TP captures the notion of a set of TFs that act in a coordinated manner across a set of target genes. So in the same way that a document's topics define its context, a gene's transcriptional programs summarise its transcriptional regulation. Figure 4.1 shows how transcriptional programs can summarise regulatory relationships. To provide a concrete example for the analogy, consider that the word 'apple' might belong to both a topic about fruit and to an almost entirely non-overlapping topic about computers. In the same way the TF SOX2 may be reused across distinct regulatory programs in the cell.

Hierarchical Dirichlet processes (HDPs) are a natural framework to use in document-topic modelling and hence for our work in transcriptional regulation. In our framework, transcriptional programs are modelled as distributions over TFs. Each gene's transcriptional regulation is modelled as a mixture of these programs. Dirichlet process mixtures (DPMs) are a non-parametric technique for modelling mixtures where the number of components is unknown. We use DPMs to model the mixture associated with each gene's transcriptional regulation. In order to share transcriptional programs between

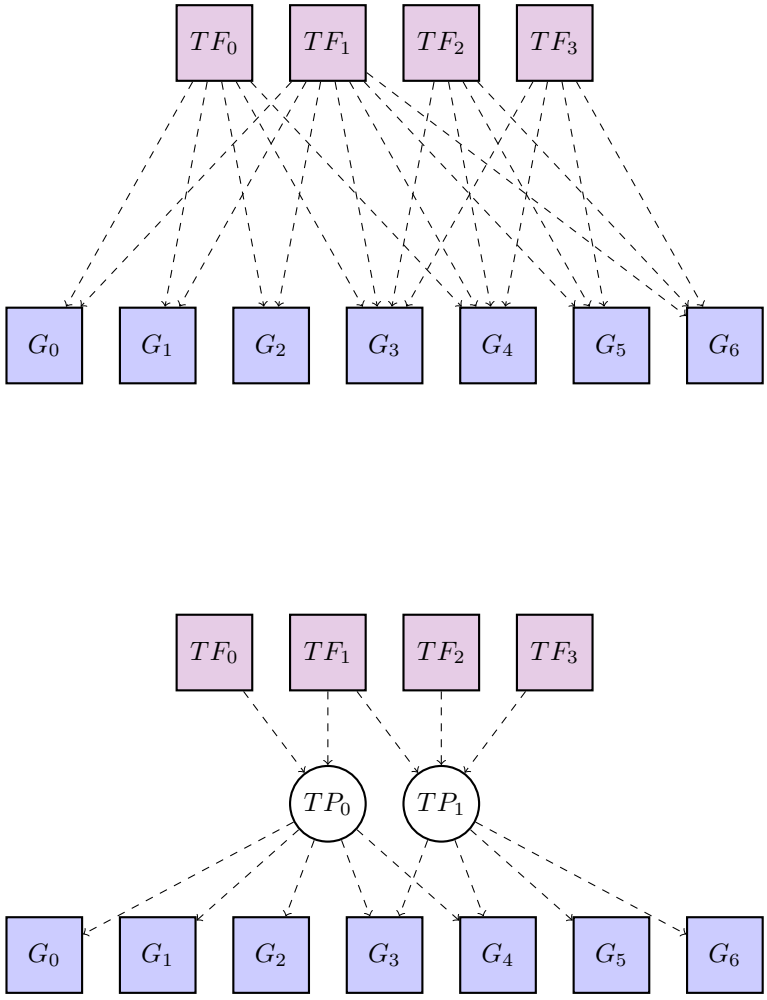


Figure 4.1: Two schematics of the same regulatory network. Both representations have four TFs at the top and seven genes at the bottom. The lower network uses latent transcriptional programs as intermediaries to reduce its complexity. Note that the transcriptional programs can overlap, for example TF_1 is in both programs and that the same gene can be targeted by multiple programs, for example G_3 and G_4.

genes we use a common base distribution for the DPMs which is itself a DPM. This step makes our model hierarchical. An extensive review of HDPs is given in [Teh et al., 2006].

4.1.3 Previous work

Quite a few approaches have been suggested in the literature to identify groups of TFs that co-regulate genes, often called transcriptional modules (TM). They all differ from our approach in several respects. One major difference is that our concept of transcriptional programs is slightly more abstract than TMs. A TM is often defined as a set of

TFs that physically bind next to each other in the vicinity of the regulated gene. Many approaches either enumerate all possible combinations of TFs up to a certain number (fewer than half a dozen or so) as potential TMs and search for over-representation of TMs in various groups of genes [Sharan et al., 2003, Kreiman, 2004, Ho Sui et al., 2007, Singh et al., 2007]. There is usually a computationally intensive post-processing step involved in clustering TMs according to *ad hoc* rules in these combinatorial approaches to reduce the number of highly similar TMs. Alternatively, an incidence matrix (or bipartite graph) is calculated linking each TF to the genes it regulates [Lemmens et al., 2006, Chen et al., 2007, Jensen et al., 2007] (as in Figure 4.1 top).

In contrast, a transcriptional program, as we define it, comprises TFs as well as genes (see Figure 4.1 bottom) and does not necessarily require a physical vicinity of binding sites for all the TFs in the program. For example, if two transcriptional modules have some common TFs, not necessarily sharing all of them, they might be merged into one transcriptional program by our algorithm. Whether this happens depends on the amount of overlap and the number of co-occurrences of their TFs. In a way, transcriptional programs generalise both transcriptional modules and TF-gene incidence matrices and provide a higher-level summary of these structures. To our knowledge, the only other work defining transcriptional programs in a similar way is by Tanay et al. [Tanay et al., 2004b]. In contrast to their work, where such programs are found by enumeration, scoring and filtering, we model transcriptional programs explicitly within a comprehensive probabilistic model.

Some work, as discussed below, insists on clusters of co-regulated genes or groups of co-regulating TFs to be disjoint. Our approach is open to the possibility that genes as well as TFs can be members of several transcriptional programs simultaneously. Indeed this fits our biological understanding well: TFs are well known to be reused in different combinations in different cellular contexts. A further difference is that many approaches require a positive gene set, for example, by co-expression, as well as a background set to detect TMs that characterise one set against the other. Our approach is essentially an unsupervised one, where transcriptional programs are discovered from one sequence set. This is a more challenging problem but it requires less input from the user and avoids problems of mis-identification of the positive set.

To our knowledge, our approach is the first application of a document topic model to transcriptional regulation. Such models have the distinct advantage of using very few free parameters that need to be specified.

Being more specific about previous work, CREME [Sharan et al., 2003] uses a sliding window to look for combinations of TFBSs that are over-represented in promoters of the genes of interest. Only combinations whose sites are physically close to one another

can be detected in this way. The user must specify the maximum number of factors in a promoter. oPOSSUM2 [Ho Sui et al., 2007] looks for pairs and triplets of TFs that are over-represented in the promoters of the genes. TREMOR [Singh et al., 2007] is similar but uses the Mahalanobis distance to distinguish between similar PWMs that represent different members of the same family of TFs. It also removes some dependence on arbitrary p-value thresholds. All of these methods discriminate between a positive user-specified set of genes and a negative (background) set. Our method differs in that it fits a model of the entire set of genes at once.

Kreiman [Kreiman, 2004] looks for over-representation of combinations of up to four TFs in co-expressed genes. Blüthgen et al. [Blüthgen et al., 2005] use Cluster-Buster [Frith et al., 2003] to identify groups of potentially co-regulating TFs which are then further filtered by statistical enrichment of classes of regulated genes in the Gene Ontology (GO) catalogue [Ashburner et al., 2000].

There is some work that integrates more than one data source. Some combination of ChIP-chip, binding site analysis (either *de novo* or PWM-based) and expression data are commonly used. Heuristics or probabilistic models are used to search for consistent structure amongst these data sources. Almost all this work has been carried out in *Saccharomyces cerevisiae*. ReMoDiscovery [Lemmens et al., 2006] builds on the Apriori framework in a two-step procedure which examines expression profiles and ChIP-chip data. MOFA [Wu et al., 2006] combines binding data with time-series microarray data to build transcriptional modules and explicitly models which TFs up or down-regulate which genes. SAMBA [Tanay et al., 2004b] is a biclustering framework that analyses gene expression, protein interaction, growth phenotype, and TF binding data. In COGRIM, Chen et al. [Chen et al., 2007] use Gibbs sampling in a Bayesian hierarchical model to integrate expression data, PWM analyses and ChIP-chip data. They model uncertainty in each data source independently but each module is associated with exactly one TF. As discussed above most of this work is reconstructs pair relationships of TFs and regulated genes.

Segal et al. [Segal et al., 2003b] have integrated a motif search algorithm and gene expression data to find motif profiles (analogous to transcriptional programs) in *Saccharomyces cerevisiae*. Their model partitions the genes into a fixed number of mutually exclusive sets which have the same expression pattern across experiments. Each gene is the target of exactly one motif profile, hence their model does not allow so much structure in the latent profiles/programs. Also, the number of partitions must be fixed somewhat arbitrarily in advance by the user. They focus on *Saccharomyces cerevisiae* which has a simpler transcriptional code than *Mus musculus*, the focus of our study.

Various other probabilistic models that require specification of the number of modules

by the user have been implemented. Xu et al. [Xu et al., 2004] build on the module networks of Segal et al. [Segal et al., 2003a]. These models also partition the gene set to find transcriptional modules. Our model allows genes to be the target of more than one transcriptional program.

Other algorithms also use non-parametric probabilistic models to obviate the need to specify the number of modules. Gerber et al. [Gerber et al., 2007] use hierarchical Dirichlet processes to discover expression programs in human microarrays. They use a similar model to ours, except their data are discretised expression levels rather than putative TFBSs. They use a Markov chain Monte Carlo (MCMC) method for inference which takes an order of magnitude longer than our variational approach. The MCMC method produces a posterior distribution over the unknowns in their model. One of the latent variables in their model is the structure of the gene hierarchy. Identifiability issues force them to use a complex set of heuristics to summarise this hierarchy. Liu et al. [Liu et al., 2007] use a Bayesian hierarchical model to examine yeast gene expression and ChIP-chip data. Their extension of an infinite mixture model limits each program to represent binding data for at most one TF. It is difficult to see how cooperative effects are estimated by the model.

4.2 Methods

4.2.1 Binding site analyses

We extracted 1,000 repeat-masked base pairs upstream of the mouse transcriptional start sites (assembly July 2007) as defined in the UCSC Genome browser [Kent et al., 2002]. After removing strongly repeat-masked sequences we were left with 18,445 sequences for analysis.

We extracted a set of PWMs from TRANSFAC version 11.4 for which we could map the factors they represent onto Ensembl gene identifiers [Hubbard et al., 2007]. From each promoter we need an estimate of the number of times there is a binding site for that PWM in the CRM as input to our HDPM.

We scored each putative TFBS in the promoters with the log-likelihood scoring scheme (Section 2.1.2). We used a threshold of $S_{LL} > 7.8$ to predict TFBSs. Our experience working with biologists has shown us that this is a reasonable threshold to use. Our model does allow for noisy data and should accommodate false positives in the large vague transcriptional programs that do not model cooperative effects. We were constrained by our computational resources from lowering this threshold significantly.

Up to this point we have been dealing with each PWM independently. Unfortunately they are not independent as, for instance, there are many factors for which TRANSFAC has more than one PWM. Two PWMs for the same factor are very likely to represent TFBSs at the same location in a promoter. We do not wish our model to learn this strong correlation instead of true transcriptional programs. We therefore reduce our set of TFBSs by taking the highest scoring set of non-overlapping TFBSs for each promoter. To do this we used the BiFA maximal chain algorithm applied to just one sequence.

4.2.2 Topic document model

In the field of information retrieval HDPs [Teh et al., 2006] are often used to model latent topics in documents. We apply them to TFBSs in promoters to infer latent transcriptional programs.

Our model is best described generatively, that is, we describe how to sample a suitable TF from our model given a target gene. We follow the description in [Teh et al., 2008]. A gene g is linked to a distribution over transcriptional programs, which is represented by a (possibly infinite) vector $\theta_g = (\theta_{g1}, \theta_{g2}, \ldots, \theta_{gk}, \ldots)$, where θ_{gk} is the contribution of program k to gene g. All θ_{gk} sum up to one for each g. A program k in turn is linked to a similar distribution over TFs, that is, program k is represented by a vector $\phi_k = (\phi_{k1}, \ldots, \phi_{kJ})$, where ϕ_{kj} is the contribution of TF j to program k assuming there are J TFs in total. All ϕ_{kj} sum up to one for each k.

To sample a random TF for binding site i upstream of gene g, we first sample a multinomial random variable variable $z_{ig} \sim \text{Mult}(\theta_g)$ which indicates the transcriptional program the factor is drawn from. Next, we sample a second multinomial random variable $x_{ig} \sim \text{Mult}(\phi_{z_{ig}})$ taking the selected transcriptional program z_{ig} into account. Sample x_{ig} specifies which TF binds at binding site i upstream of gene g.

When calibrating the model using data, the task is to infer posterior distributions for parameters θ_g and ϕ_k. In order to do this, we place conjugate Dirichlet priors on the parameter vectors and use a variational approach to approximate their posterior distribution (for details see [Teh et al., 2008]). More specifically, we set $\theta_g \sim \text{Dir}(\alpha\pi)$ and $\phi_k \sim \text{Dir}(\beta\tau)$. α and β are scalar strength parameters that control the variances of the θ_g and ϕ_k respectively. π and τ are vectors and represent their respective means.

We do not wish to constrain our model to use a fixed number of transcriptional programs. Instead, we use a non-parametric approach where we allow a countably infinite number of transcriptional programs. Now θ_g and π are infinite dimensional vectors. π is modelled using an explicit stick-breaking construction [Sethuraman, 1994] where γ controls how

many transcriptional programs are used. Formally, the stick-breaking model is defined by

$$\pi_k = \tilde{\pi}_k \prod_{l=1}^{k-1}(1-\tilde{\pi}_l) \qquad \tilde{\pi_k} \sim \text{Beta}(1,\gamma) \qquad \text{for } k = 1, 2, \ldots$$

Intuitively, probabilities π_k are obtained by starting with a stick of length 1, and continuing to break pieces off it, their lengths representing the probabilities π_k. Even if continued indefinitely the pieces all sum up to one, the total length of the stick, forming a proper probability distribution over the natural numbers. The size of piece k is determined as a fraction $\tilde{\pi}_k$ of the remaining stick, whose length is $\prod_{l=1}^{k-1}(1-\tilde{\pi}_l)$, where $\tilde{\pi}_k$ is a random number from the interval $[0, 1]$ distributed according to a Beta distribution.

We also place priors on all the other hyperparameters of the model,

$$\alpha \sim \text{Gamma}(a_\alpha, b_\alpha) \quad \beta \sim \text{Gamma}(a_\beta, b_\beta) \quad \gamma \sim \text{Gamma}(a_\gamma, b_\gamma) \quad \tau \sim \text{Dir}(a_\tau)$$

Our model is presented graphically in Figure 4.2.

4.2.3 Inference

We implemented the collapsed variational inference technique described by Teh et al. [Teh et al., 2008] complete with the Gaussian approximation for non-zero counts.

4.2.4 Thresholding the posterior

In our model each transcriptional program is represented as a distribution over factors, ϕ_k, and each gene can be summarised as a distribution over programs, θ_g. In order to examine the programs we have learnt, we thresholded these distributions to discover which programs are over-represented in which genes and which factors are over-represented in which programs. However, due to the collapsed nature of the inference algorithm we do not directly obtain a posterior over them as they have been integrated out. The inference algorithm does infer which factors have binding sites in which genes due to which transcriptional programs. These inferences are summarised as the expectations of various counts and these allow us to estimate the θ_g and ϕ_k and hence associate transcriptional programs with genes and with TFs.

More formally, in an analogous notation to [Teh et al., 2008], we define n_{gkf} as the number of binding sites for factor f drawn from transcriptional program k in the promoter of gene g. A '.' in the subscript indicates summation over that index. For example, $n_{.kf}$

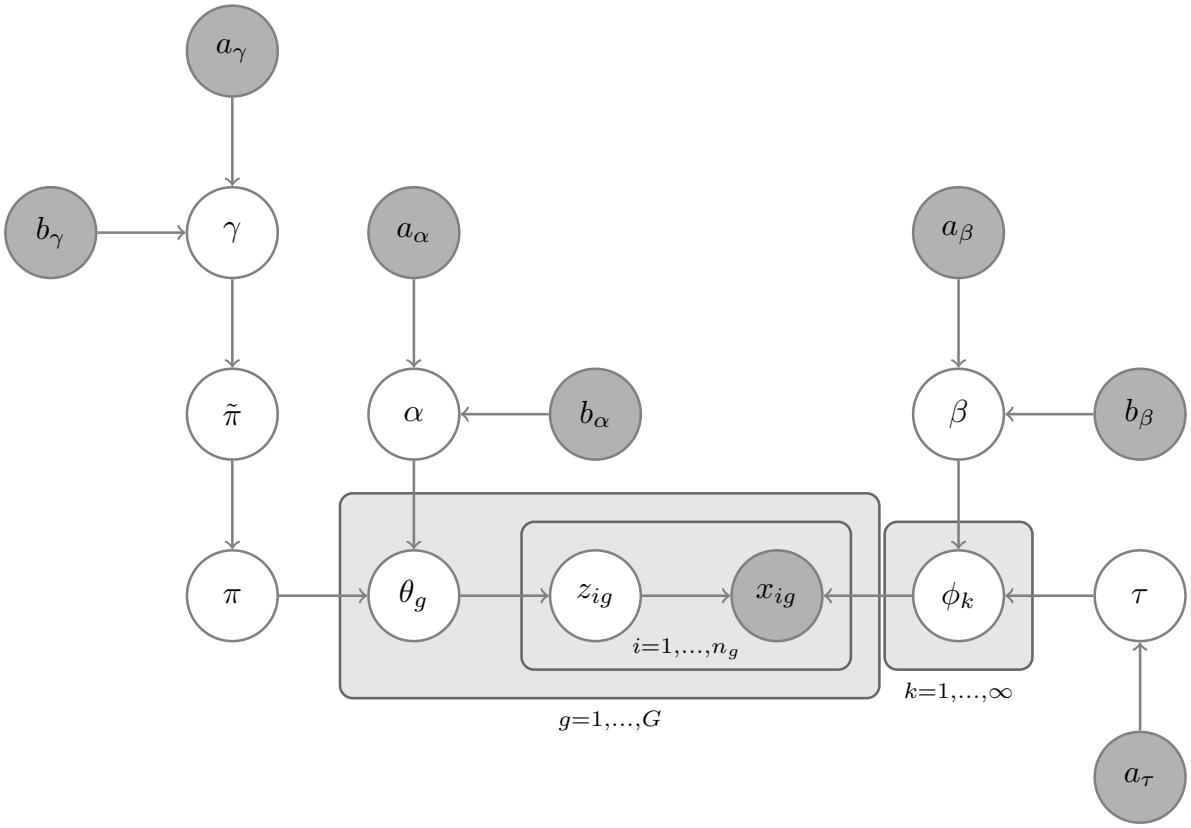


Figure 4.2: We present our model graphically. The shaded nodes represent observed variables (or equivalently from the model's perspective, fixed hyper-parameters). The clear nodes are the latent variables in the model. The boxes are called *plates*. If a node is inside a plate, its corresponding variable has a multiplicity equal to the size of the plate. For example, there are G instances of the θ_g variable as its node is inside the $g = 1, \ldots, G$ plate. See the text for a description of the variables.

is the number of binding sites of factor f drawn across all genes from program k. Now we make point estimates:

$$\widehat{\theta}_{gk} = \frac{\mathbb{E}(n_{gk.})}{n_{g..}} \qquad \widehat{\phi}_{kf} = \frac{\mathbb{E}(n_{.kf})}{\mathbb{E}(n_{.k.})}$$

We define $\Phi_f = \frac{n_{..f}}{n_{...}}$ to be the empirical distribution of factors. Now we associate with transcriptional program k all those factors, f, for which $\frac{\widehat{\phi}_{kf}}{\Phi_f} > 2$. Likewise we define $\widehat{\Theta}_k = \frac{\mathbb{E}(n_{.k.})}{n_{...}}$ and associate those genes, g, with transcriptional programs, k, for which $\frac{\widehat{\theta}_{gk}}{\widehat{\Theta}_k} > 2$. We found our method was insensitive to the actual choice of threshold: when we varied it between $1\frac{1}{2}$ and 10 the results were not affected significantly. Of course it is possible for a program to have no factors nor genes associated with it if its distributions are close to the empirical distributions.

4.2.5 Validation

Correcting for multiple testing in a GO ontology is difficult due to its hierarchical nature. To validate the strength of our results we generated random samples from the same populations of genes and TFs to test for enrichment. We choose sample sizes to cover the range of sizes in the discovered transcriptional programs. For each size we sampled 100 independent sets and calculated the exponent of the best p-value found in a GO term enrichment analysis (see Section 4.3.3).

Significance of p-values

We used a bootstrap approach to assess the significance of the results from the GO enrichment analysis. We generated random samples of factors and random samples of genes without replacement. We analysed each sample for GO enrichment using the same procedure as for the transcriptional programs that our model predicts. For each sample we took the best uncorrected p-value and refer to its base 10 logarithm as its p-score. In Figure 4.3 we show box plots of the p-scores for the randomly sampled factors and target genes. We sampled 100 times at each of 50 different sample sizes for the factors and the targets. The sizes were chosen to reflect the range of sizes of the actual transcriptional programs. Hence each plot represents $50*100 = 5000$ independent samples. The sample size does not appear to affect the extreme value distribution of the best p-scores' exponents. From 10,000 independent samples, the lowest p-score is around -6.

Each p-score represents the smallest p-value obtained when a random sample of factors or targets was tested against every term in the GO hierarchy. We wanted to correct the p-values for these multiple tests. Plotting the sorted p-scores against the base 10 logarithm of the proportion that are equal or better gave us a good linear fit (Figure 4.3). This fit has an intercept very close to -2 which, together with the linear relationship, suggest adding 2 to the p-score to obtain a multiple testing corrected p-value exponent. That is, we would expect to have to generate 10^2 random samples in order to achieve a p-score of -4. Hence we would deem a p-score of -4 significant at the 0.01 level.

4.3 Results and Discussion

We analysed the promoter regions of 18,445 *Mus musculus* genes using PWMs from TRANSFAC. This generated 78,085 putative TFBSs of 149 TFs which scored above a stringent threshold (see Section 4.2.1). We ran our model on these putative TFBSs and it discovered 68 latent transcriptional programs.

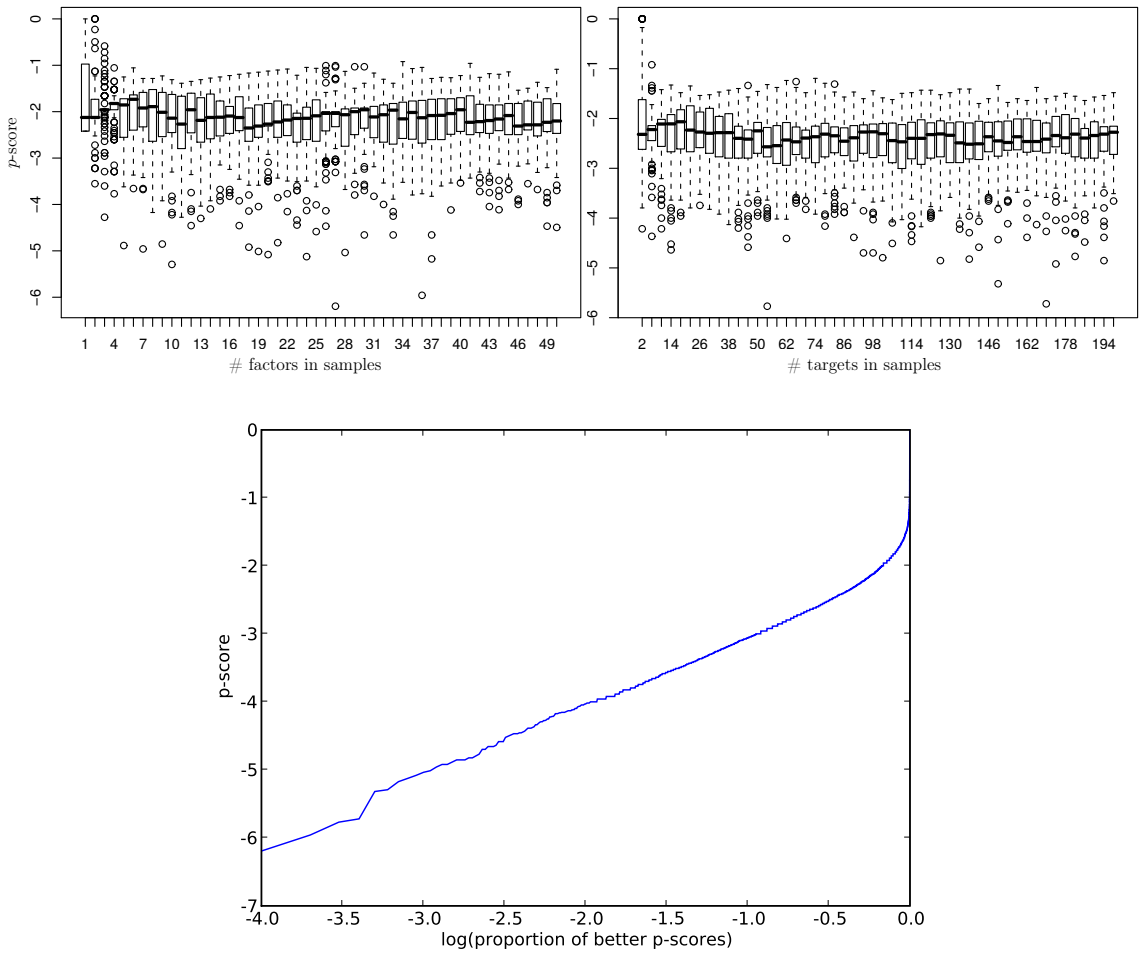


Figure 4.3: Assessment of the significance of the results from the GO enrichment analysis by random samples of factors and genes. We show a boxplot of the p-scores for the randomly sampled factors on the top left and the targets on the top right. The x-axes are the sample sizes and the y-axes are the p-scores. We sampled 100 times at each of 50 different sample sizes for the factors and the targets. The lower plot shows the sorted p-scores plotted against the base 10 logarithm of the proportion that are equal or less.

4.3.1 Inference

As variational inference only converges to a local maximum, we ran the algorithm 24 times from different initialisations. Each initialisation differed only in the variational distribution over the assignment of TFBSs to programs, $q(z)$. These were drawn randomly from a Dirichlet distribution. Figure 4.4 shows how the converged expected log-likelihoods varied between the 24 restarts.

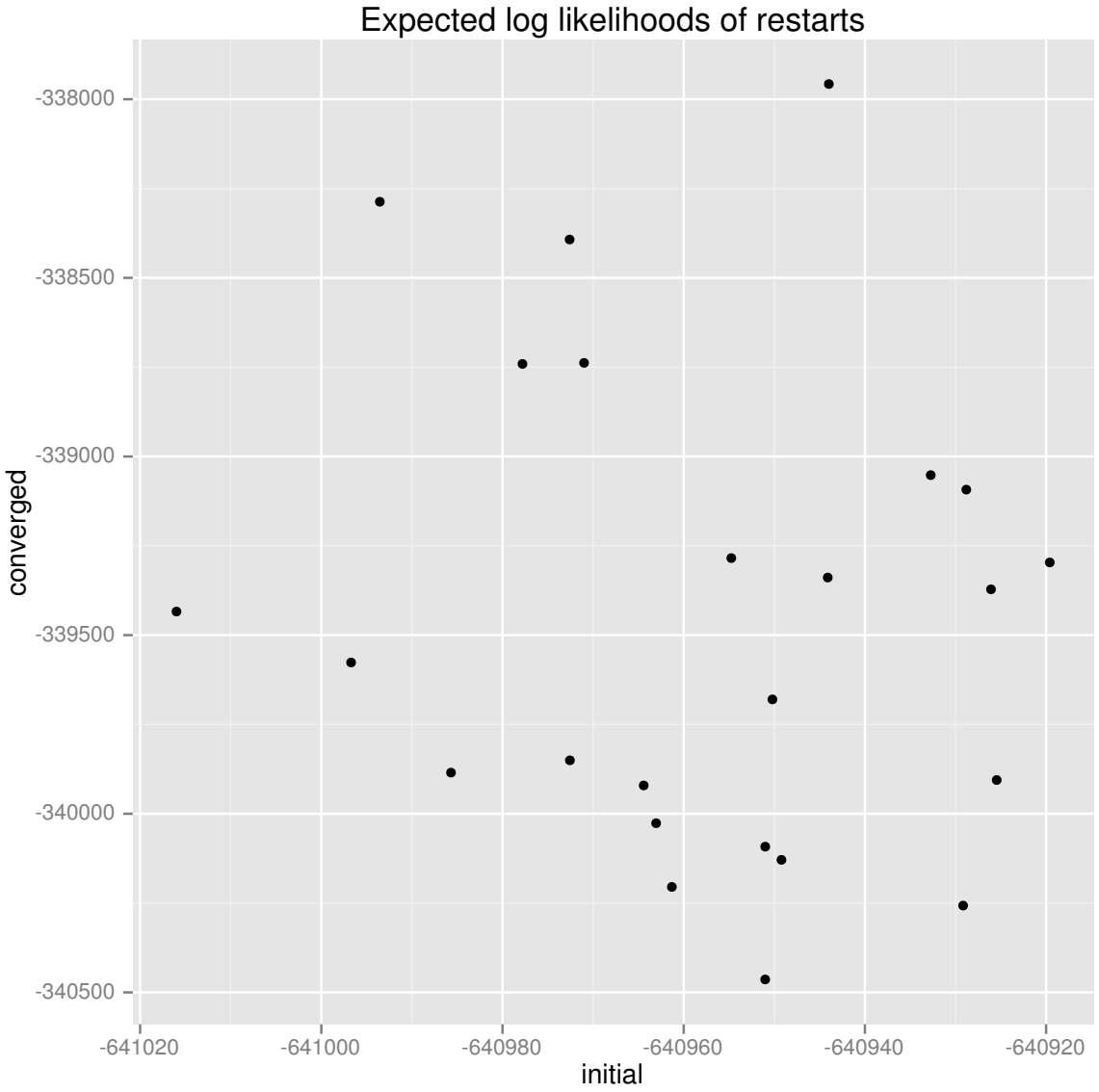


Figure 4.4: Each point represents one run of the variational inference algorithm. The expected log-likelihood at the initial conditions is given on the x-axis and the expected log-likelihood after convergence is given on the y-axis.

4.3.2 Structure of the programs

The number of TFBSs explained by each of the 68 programs varied considerably. Most of the TFBSs were explained by the largest 10 programs (Figure 4.5). We examined how many target genes and transcription factors were associated with each program and how this related to the number of TFBSs explained by each program. In general we found a good separation of these statistics across the programs (Figure 4.6) showing that our model had inferred structure across large sets of genes as well as structure across small

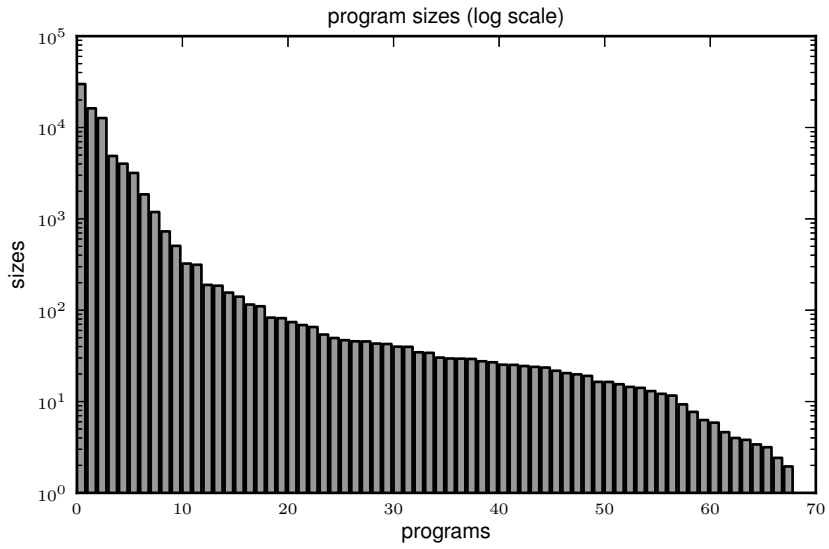


Figure 4.5: We show how many TFBSs are generated by each transcriptional program in our model. The number of binding sites that each program in our model explains is shown on a log-scale. A by-product of our algorithm is that the programs are sorted by the number of TFBSs they are responsible for. The most frequently used transcriptional programs accounted for almost 30,000 and 15,000 binding sites respectively and the smallest just for a handful. The largest programs are composed predominantly of common TFs and in general the smaller programs explain occurrences of rarer TFs.

gene sets.

Coverage of the genome

As the model associates each TFBS with a program, even those TFBSs for which cooperative effects cannot be found must be associated with a program. The model uses the largest two programs (programs 0 and 1) for these TFBSs: their distribution over factors is vague and they target many genes. To some extent, the programs that explain more binding sites are less likely to represent true cooperative effects. They can be seen as a 'catch-all' for those TFBS that the model could not find higher-order structure for. We looked at the number of target genes of the programs in this context. That is, we analysed the total number of target genes of all programs smaller than a given size (Figure 4.7). The size of a program is measured by the number of binding sites it explains. Including the first two vague programs, a total of over 10,000 genes are associated with our programs. Most of the binding sites are explained by the first ten programs and using this as a cut-off we can see that the remainder of the programs still target over 4,000 genes. Thus a sizeable proportion of the genome can be associated with the cooperative combinations of factors defined by our programs.

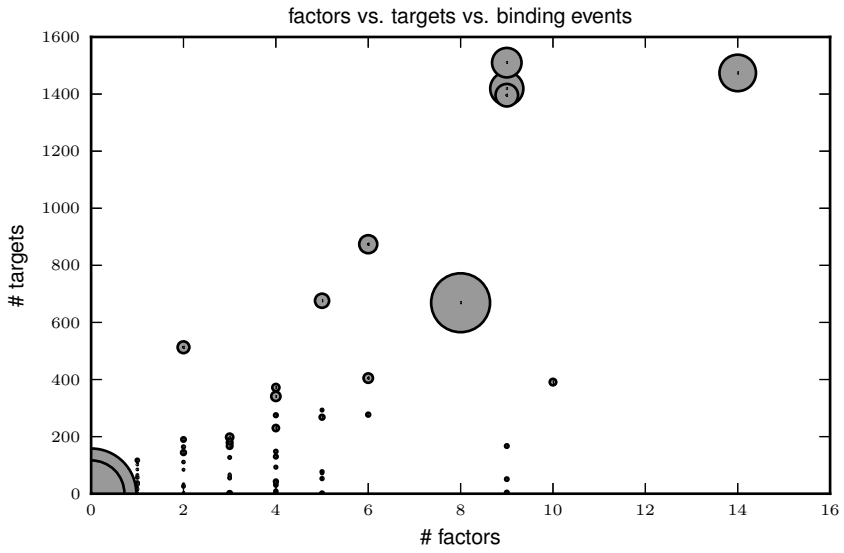


Figure 4.6: A scatter plot of the programs showing the number of TFs against target genes. The area of each scatter point is proportional to the number of binding sites the program is responsible for. Note the first two programs do not have any genes or targets associated with them, their distributions over TFs are very similar to the genomic distribution and they are ubiquitous.

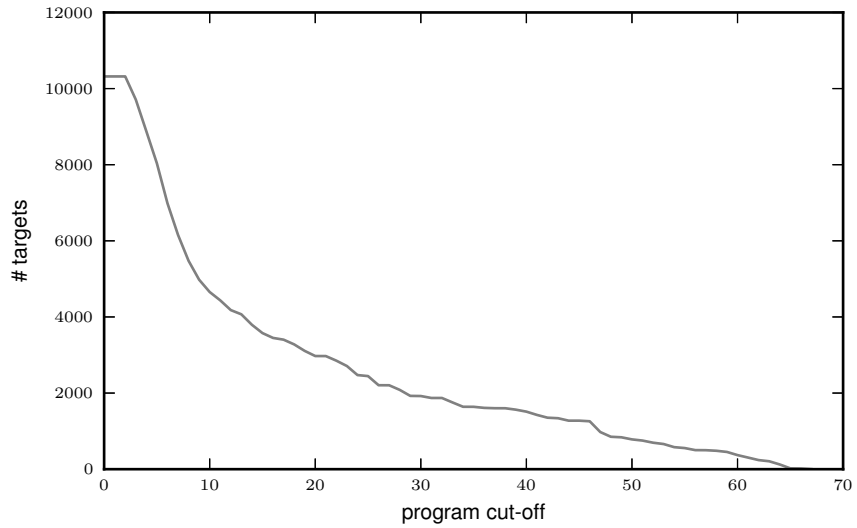


Figure 4.7: We plot how many genes are targeted by the programs smaller than a given size. The programs that account for more binding sites are less interesting in terms of cooperative effects, so we plot the size of the set of all targets of all programs smaller than a given size. The size cut-off varies along the x-axis (indexed by program) and the y-axis represents the total number of genes targeted by those programs. For example, excluding the first 10 programs, just over 4,000 distinct genes are targeted by the remainder of the programs.

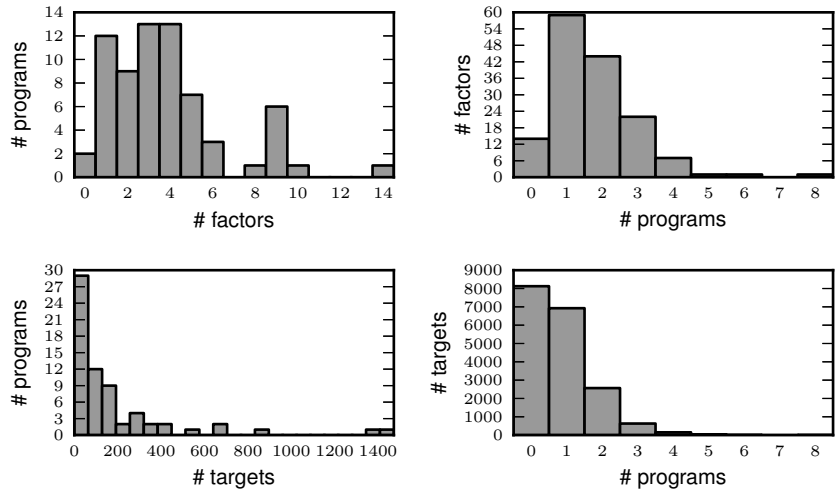


Figure 4.8: Each transcriptional program is associated with a set of TFs and a set of target genes. We examined the relationships between the programs and their targets and factors. The top left figure shows that most programs have fewer than seven factors associated with them. The top right illustrates that most factors are in fewer than five programs. The bottom left shows that a few programs target many genes but most programs have fewer than 200 targets and the bottom right demonstrates most genes are targeted by two or fewer programs.

Separation between the programs

In general, we found a good separation between the programs, in that any given TF or gene is unlikely to be associated with many programs and conversely that most programs were associated with a small number of TFs and genes (Figure 4.8). This was confirmed by our analysis of the intersection between pairs of programs' TFs and the overlap between their target genes (Figure 4.9).

4.3.3 Validation

In order to test whether the transcriptional programs capture real biological structure we validated the transcriptional programs using an analysis of enrichment for GO terms [Ashburner et al., 2000], signalling pathways from the KEGG database [Kanehisa, 2006], tissue specific co-expressed genes from SymAtlas [Su et al., 2002], and groups of known interacting TFs from the literature. We present those transcriptional programs that were noteworthy in the validation in Table 4.1.

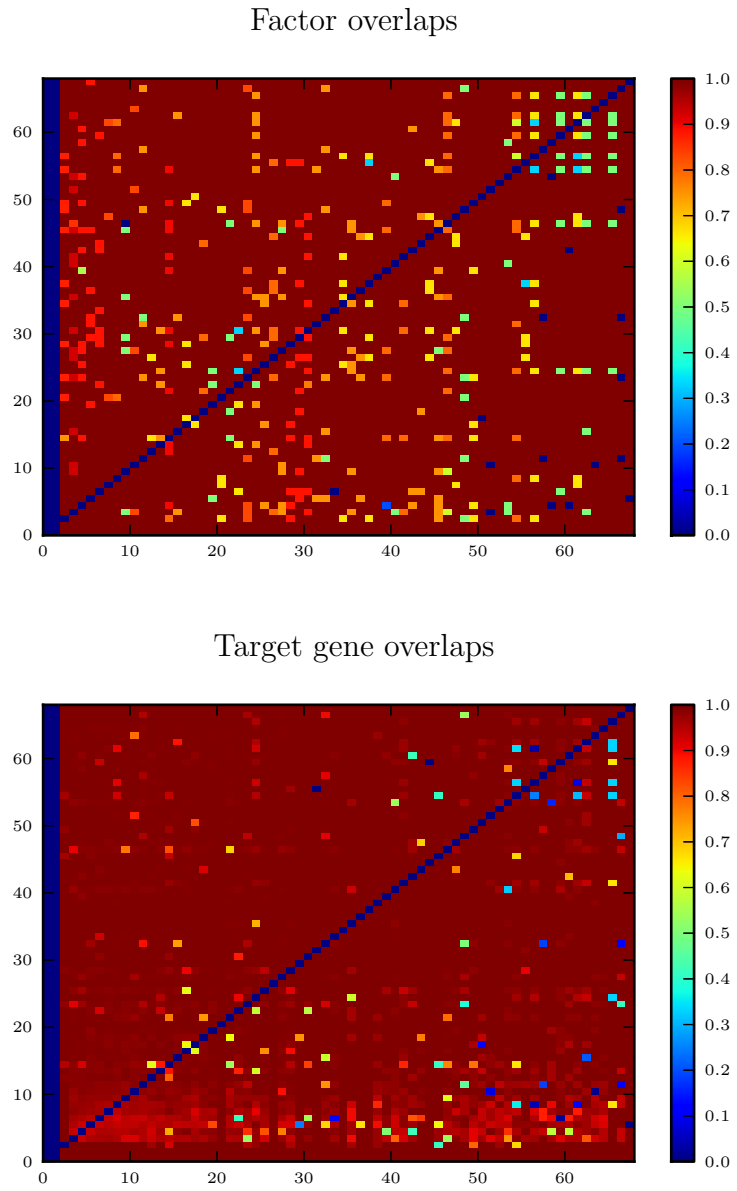


Figure 4.9: The intersection of all the programs' factors and targets. On top, we show how much the factors of the programs overlap. This is represented as the ratio of the size of the intersection between the factors to the size of one of the sets of factors. The bottom is the same analysis of the programs' targets. The overlap between targets and factors is negligible in almost all cases. The sets of factors that do overlap to some extent are those that are not responsible for many TFBSs. The first two programs do not have any factors or targets associated with them.

TP	# Targets	Factors
2	669	Nkx2-1 Dbp Ahr Srebf1 Egr2 Tcfap2a Sp1 Egr1
3	1474	Rest Pparg Pax6 Creb1 Vdr Ets1 Hivep1 Pbx1 Dmrt2 Hand1 Dmrta1 Irf8 Atf2 Ar
4	1419	Gabpa Gzf1 Ppara Stat3 Hoxa5 Ikzf1 Hnf4a Srf Pax5
5	1510	Atf4 Dmrt1 Lhx3 Nkx6-1 Stat5a Runx2 Irf2 Pax4 Pax1
12	198	Nfya E2f1 Mtf1
13	372	Foxo3 Foxj2 Dmrt3 Nr2f1
15	230	Pbx1 Nr5a1 Sry Rora
18	268	Pou1f1 Pax2 Ets2 Cux1 Tbp
25	275	Srf T CT025657.12-201 Pou5f1
28	167	Cebpa Gabpa Cebpg Dbp Tgif1 Atf3 Rela Hes1 CT025657.12-201
53	111	Gzf1 Atf2
55	54	Klf4 Prdm1 Atf3

Table 4.1: Some of the more interesting transcriptional programs.

GO term enrichment

Each program is associated with a set of TFs and a set of target genes. We tested the genes and the factors in each program for enrichment of terms in the biological process GO ontology. We used a standard hypergeometric test in conjunction with the weight method implemented in the topGO R-package [Alexa et al., 2006] as a significance test. Table 4.2 shows the result of the GO enrichment analysis.

KEGG pathway enrichment

We tested the genes and the factors in each program for enrichment in signalling pathways defined in the KEGG database. After Bonferroni correction for multiple testing, we found no significant results. However, we did find a significant result in conjunction with our analysis of known interacting TFs from the literature.

SymAtlas enrichment

We tested the target genes in each program for enrichment in tissue-specific co-expressed genes from the SymAtlas dataset. Genes over-expressed in embryonic tissues were significantly enriched in the targets of transcriptional program 53. Program 53 accounts

TP	Factors	GO term		GO description	annotated	p-value
28	9	GO:0001889	BP	liver development	5/7	4.7e-06
TP	Targets	GO term		GO description	annotated	p-value
2	669	GO:0004842	MF	ubiquitin-protein ligase activity	16/121	8.7e-06
2	669	GO:0051216	BP	cartilage development	12/71	9.0e-06
3	1474	GO:0005550	MF	pheromone binding	11/27	2.8e-06
4	1419	GO:0005132	MF	interferon-alpha/beta receptor binding	7/8	1.2e-07
5	1510	GO:0005132	MF	interferon-alpha/beta receptor binding	6/8	7.3e-06
12	198	GO:0000786	CC	nucleosome	14/136	4.2e-10
12	198	GO:0005634	CC	nucleus	84/4115	8.2e-08
12	198	GO:0003677	MF	DNA binding	54/1993	4.0e-07
12	198	GO:0003697	MF	single-stranded DNA binding	6/37	4.2e-06
12	198	GO:0006334	BP	nucleosome assembly	14/147	1.9e-09
12	198	GO:0006260	BP	DNA replication	16/151	2.4e-06
13	372	GO:0004984	MF	olfactory receptor activity	50/1001	5.4e-09
13	372	GO:0007186	BP	G-protein coupled receptor protein signa...	81/1968	1.4e-09
15	230	GO:0034097	BP	response to cytokine stimulus	5/15	1.0e-06
18	268	GO:0004984	MF	olfactory receptor activity	43/1001	6.0e-11
18	268	GO:0007166	BP	cell surface receptor linked signal tran...	74/2606	2.4e-08
18	268	GO:0007608	BP	sensory perception of smell	10/90	7.9e-07
25	275	GO:0004556	MF	alpha-amylase activity	4/5	2.1e-07
28	167	GO:0007186	BP	G-protein coupled receptor protein signa3...	38/1968	4.1e-06
55	54	GO:0032183	MF	SUMO binding	2/2	1.0e-05

Table 4.2: We tested each of the factors and target genes associated with each of the 68 transcriptional programs for enrichment of GO terms across the three GO ontologies: biological process (BP), molecular function (MF) and cellular component (CC). The number of factors (resp. targets) associated with the program is displayed followed by information about the GO term. 'Annotated' shows the number of factors (resp. targets) annotated with the term in the program compared to the total numbers of factors (resp. targets) annotated with the term. The p-values are not corrected for multiple testing. Based on a bootstrap analysis described in Section 4.2.5 any p-value below 10^{-4} might be deemed significant at the 0.01 level.

for fewer than 100 binding sites out of the 78,085 sites, yet was strongly predictive of membership of the group of over-expressed genes. This demonstrates the ability of our method to find small signals in large datasets.

Literature

We took well known sets of interacting TFs from the literature and looked for programs that contained them. We looked for sets of TFs associated with the liver, muscle development, and the cell cycle. The three factors in transcriptional program 12 (E2F, NFY, MTF1) contain two of the three TFs in our analysis that are known to regulate the cell cycle (E2F, CREB, NFY [Elkon, 2003]). When we tested the targets of program 12 for enrichment in the KEGG cell cycle pathway (without correcting for multiple testing) we obtained a p-value of 9e-4. The extra TF in program 12 that is not in our literature derived set, MTF1, has been implicated in the cell cycle [Lichtlen et al., 2001] and as a co-regulator with E2F [Joshi et al., 2005].

4.3.4 Structure at many scales

Our model found programs over a wide range of sizes. However a DPM would be expected to have components of varying sizes. This expectation is made explicit in the DPM's representation using the stick breaking construction. As demonstrated in the GO enrichment validation (Table 4.2) our model was able to find significant signals in both large and small programs. This does not prove that the DPM's prior over program sizes fits the data well but it does demonstrate that the found programs are not just an artefact of the DPM's prior.

4.3.5 Biological interpretation

Several of the discovered programs have well defined biological meanings. Not many of the factors of the transcriptional programs were significantly enriched for GO terms. However, program 28 did contain five of seven TFs that are annotated with the term 'liver development' in its nine factors.

Several of the target sets of the programs were strongly associated with different GO terms. In particular, program 12 was particularly enriched for genes with nuclear products and those that are involved in nucleosome assembly. Program 18 appears to be associated with the sense of smell as it has strong enrichment for 'olfactory receptor activity' and 'sensory perception of smell'.

4.3.6 Potential improvements

There are a number of ways the model and analysis as presented could be improved.

PWM scanning has a high FDR

Predicting TFBSs by scanning for PWMs is a notoriously difficult task. PWMs are typically reasonably degenerate motifs and will match many putative TFBSs. At any reasonable threshold the number of false positives is high. In our method we have used a simple scan of the promoter sequence to predict TFBSs. The high false discovery rate could be managed in part by integrating other sources of data that are indicative of TF binding such as *in vivo* location data for particular TFs from ChIP-seq experiments, epigenetic data regarding histone modifications or DNA methylation, and phylogenetic conservation across related species. Methods to reduce the FDR have been discussed in Chapter 2.

Weak binding sites

Our method must discard those TFBSs that are not amongst the strongest in the regulatory region. It must do this to avoid introducing a bias when there is an overlap between PWMs (see Section 4.2.1). Several recent studies have shown that in many instances, weak TFBSs are crucial for regulatory networks to generate proper expression patterns. In our analysis we are selecting against these sites. It is difficult to assess how much this affects our analysis. It is also difficult to modify the analysis to include these weak binding sites particularly because we already have the problem of the high false positive rate when scanning for PWMs.

Poor knowledge of PWMs

It is believed that the mouse genome contains a much greater number of sequence-specific transcription factors than we have PWMs for. Our analysis uses 149 PWMs but the mouse genome is estimated to contain of the order of 2,500 TFs. The 149 PWMs are biased towards the most highly studied and perhaps most important TFs but nevertheless our model is only looking at a small fraction of the TFs. Recently much progress has been made elucidating the binding preferences of more TFs. In particular many TFs' binding preferences have been characterised using protein-binding microarrays. Additionally many TFs' binding preferences have been learnt from motif search in large ChIP data sets. Despite this progress it is believed the majority remain to be discovered.

As well as a lack of knowledge of some PWMs, there is the problem that many TFs are related and have similar binding preferences. It is easy to mis-classify TFBSs for separate members of these families of TFs. Some PWMs represent a lowest common denominator of binding for such families, for example a core binding region that is the same across all members of the family. However the activity of different members of the family may be quite distinct: they may interact with different partner TFs and have distinct biological roles. In this case an analysis using the familial binding preferences is not ideal. Whilst in some cases TFBSs for members of the same family can be differentiated on the basis of their PWMs, in general this is a difficult issue and one that would require a more sophisticated treatment. Perhaps familial PWMs could be removed from the set of PWMs when the binding preferences of a sufficient number of family members was known. In other cases it might be beneficial to use the familial binding PWMs in the analysis in place of more specific PWMs for family members.

Paralogous sequences

Gene duplication events can also duplicate regulatory regions. After such an event, the two regions would be likely to diverge under evolutionary pressures. However recent duplication events will not have diverged far and could bias the results of the analysis. If two such paralogous regulatory regions were included in the model, they would likely contain very similar sets of TFBSs. This is precisely the situation our model is built to detect. However in this case, spurious non-functional TFBSs would stand a good chance of being detected as part of a transcriptional program. Furthermore, paralogous genes are commonly related and these transcriptional programs might well score highly in any gene set significance tests such as those performed above. We have not controlled for this effect but a more rigorous analysis would investigate this issue. We are not aware of any useful information about paralogy between specific genomic regions. Information on pairs of paralogous genes is available that could be used as a proxy for whether particular regulatory regions are paralogous. Any sets of paralogous regions could be replaced with just one typical member to avoid this issue.

Distal regulatory regions

In this study we have examined the promoter regions of known mouse genes. Due to recent advances in experimental techniques it has become clear that much regulation of gene expression occurs at some distances from promoters (at least when the distance is measured along the chromosome). Our method and model do not attempt to use any regions distal from target genes. This is mainly due to the size of the mouse genome

and the absence of reliable information about which regions might be regulatory when this study was carried out. More recently more data concerning those epigenetic modifications that accompany regulatory activity have become available in mouse and human (for example, the ENCODE project). In some models of transcriptional regulation distal regulatory regions encode expression patterns and promoters are more concerned with assembling the transcriptional machinery rather than encoding regulatory logic. In these cases the interesting programs would be active at the distal regions rather than the promoters. Thus modelling the make-up of distal regulatory regions could increase the power of the method significantly. However this modelling comes with its own problems. It is straightforward and reasonable to assume that a promoter regulates the gene it sits next to on the chromosome. Distal regions may or may not regulate the gene that is closest to them. There are well known examples where a region may regulate a more distant gene or even a gene on another chromosome. This uncertainty could be modelled explicitly or simplifying assumptions about which gene might be regulated by a region could be made. In either case we would expect that the integration of distal regions into the model would be beneficial.

Cell type specific activity

We have ignored the issue of cell type. It is well established that different regulatory programs are active in different cell types. The activity of regulatory regions is associated with certain epigenetic modifications. These modifications vary across different cell types. One popular model suggests that a gene's activity is regulated by different regulatory regions in different cell types. In this model the gene's expression patterns in different cell types would be different under the same inputs (TF concentrations) because the accessibility and activity of each regulatory region is altered in a cell-type specific manner by epigenetic modifications. This aspect of transcriptional regulation has typically been ignored by models that predict expression or look for transcriptional programs or modules. One would expect that particular transcriptional programs would be associated with regulatory regions that are active in the same cell types. The recent explosion in availability of epigenetic data means that models incorporating this information should be useful.

TF expression

In addition to cell-type specific epigenetic modifications, different genes are expressed in different cell types. In particular the genes encoding TFs in any given transcriptional program would be expected to be expressed in similar cell types. The wealth of expres-

sion data across multiple cell types could provide yet more information from which to infer transcriptional programs.

Restrictive DPM priors

We have not thoroughly examined the suitability of the model for the data. HDPMs are flexible models in that they do not constrain the data to fit in a particular number of programs. On the other hand, the prior that they place on the number of programs does not necessarily match our expectations about our data. A DPM prior suggests that the number of components should grow as the logarithm of the number of data points. Our model is not a simple DPM but a HDPM, however the sizes of the programs found by our model seem to decay exponentially in line with this behaviour. We have not investigated if our model prior enforces this behaviour or if it is inherent in the data. Pitman-Yor processes are more flexible generalisations of Dirichlet processes that can escape this logarithmic behaviour. Replacing our HDPM with a hierarchical Pitman-Yor process mixture model would be possible but inferences are more difficult in these models and it is not clear if there are significant gains to be made using them.

4.4 Conclusions

Discovering structure in sequence analyses is a difficult task. We are limited by the set of PWMs available, our inability to predict regulatory genomic regions and the high false positive rate of PWM scanning. Out of the three sets of interacting TFs that are most cited in the literature, we only recovered one of them. However, our method is looking for structure in a much larger dataset than other methods and does not have a positive set and a negative set of genes with which to discriminate.

Our model does find significant structure in these analyses and it is reasonable to suppose that this structure underlies some mechanisms of transcriptional regulation. This is to be expected given our understanding of the underlying biology. A valuable property of our method is that it finds structure at both large and small scales.

We have shown that non-parametric probabilistic models are useful tools for unsupervised learning in this context. Techniques for genomic data integration are just starting to be applied with success to higher eukaryotes and we believe HDPM models are useful non-parametric tools for this task. We believe that probabilistic models are natural and principled tools for integrating diverse types of data. We expect their popularity to increase as more experimental techniques and data become available.

Chapter 5

Discussion

5.1 Contributions

In this thesis I have applied probabilistic methods to three different problems each of which involves a different aspect of transcriptional regulation.

In Chapter 2 I presented an algorithm to predict binding sites. The algorithm is not a rigorous application of a probabilistic model and an inference technique. However, it does use Bayes factors as principled measures of evidence despite the *ad hoc* technique used to combine them. Without a more complex approach that explicitly models the dependencies between the sequences, a Bayes factor approach is perhaps the closest it is possible to get to a justifiable probabilistic model. I have evaluated this BiFA algorithm against three other methods in an attempt to assess its strengths and weaknesses. In doing so I discovered some of the strengths and weaknesses of the competing algorithms. I was also able to assess some of the statistics and methods used to evaluate these methods in other work. Whilst the BiFA algorithm did not out-perform the other algorithms, it is an alignment-free algorithm and my evaluation together with previous work, suggests that alignment-free algorithms have a place in phylogenetic TFBS prediction.

In Chapter 3 I presented STEME, an efficient approximation to MEME which is an established algorithm for motif finding. STEME allows the application of the algorithm to data sets of the size generated by modern biological techniques. The EM algorithm is one of the most popular inference algorithms and STEME demonstrates how significant efficiencies can be achieved at a small cost in accuracy, albeit only when applied to a very specific probabilistic model. It would be interesting to investigate if suffix trees could be used to improve the efficiency of inference in more complex models. For example, efficient inference in hidden Markov models on genome-scale data would be a useful tool for computational biologists. It could be that the location dependent hidden state of

such models prevents a data structure that ignores location such as a suffix tree from achieving such efficiencies.

In Chapter 4 I presented a more complex non-parametric probabilistic model inspired by document-topic models from the machine learning literature. I applied the BiFA algorithm from Chapter 2 to *Mus musculus* promoters to provide the data for this model. I applied the collapsed variational inference technique to show that the abstract concept of transcriptional programs can capture biological relationships between sets of TFs and genes. I demonstrated that the model captures relationships on many scales. That is, the model learns relationships that affect many genes and TFs and also relationships that are specific to limited sets of genes and TFs.

5.2 Probabilistic models

For a system to be worthy of study, be it transcriptional regulation or something entirely different, some uncertainty about it must exist. The act of learning is the reduction of this uncertainty. Probabilistic models are suitable tools for research because they explicitly quantify uncertainty. Complementary to this quantification of uncertainty is the ability to apply the laws of probability to make inferences. Jaynes describes these laws as "The Logic of Science" [Jaynes, 2003]. I hope that the methods in this thesis have shown that in combination, the quantification of uncertainty and the application of the laws of probability are powerful tools.

As an example consider the task of looking for transcriptional programs. Given a data set of TFBSs there are many ways one could go about searching for transcriptional programs. A natural approach that suggests itself is a combinatorial one that looks for over-represented pairs, triplets or higher order combinations of TFs and/or genes. Once these pairs or triplets are identified a subsequent step in the approach could be to amalgamate them into transcriptional programs that link TFs to genes. Typically in this sort of approach we would be compelled to define one or more significance thresholds. Combinations that passed this threshold would be considered equivalently and those combinations that just failed to pass would be ignored. As uncertainty is represented explicitly in probabilistic models they have a softer character when inferring such combinations. That is, the evidence in favour of a particular transcriptional program might be weak from the standpoint of the TFs or the genes but in combination it could be enough to justify our belief that the transcriptional program represents a biological program. Admittedly in my method I use a thresholding approach to generate a concrete set of transcriptional programs from the variational approximation to the posterior. However,

this is a post-processing step. I believe that methods that propagate uncertainty further through the inference process are more powerful in general.

5.3 Future work

The three preceding chapters include discussions of ideas that might stimulate further work on the models presented therein. In this section, I try to identify common ideas and research directions for the methods and their applications. I also try to highlight what relevant data has or is about to become available.

5.3.1 Integration

The three methods presented in this thesis have all been developed fairly independently but there is a natural order in which all three could be combined. The utility of models of combinatorial interactions would be increased by predictions for more TFs. Increasing the numbers of TFs for which we can computationally predict TFBSs relies upon learning sequence binding preferences for those TFs. Hence, given suitable data, STEME could add to the databases of PWMs; the BiFA algorithm could make predictions based on these PWMs; and the transcriptional programs model could be used to discover combinatorial structure in those predictions.

Relevant data

What types of data might be interesting to study in the context of integrating the three methods? Enhancers are believed to implement most of the logic of transcriptional regulatory networks. Any data that could narrow our search for TFBSs and interactions between those TFBSs to enhancers would be ideal. Recently, there has been much interest in the field of epigenetics. Several epigenetic marks are strongly associated with enhancer activity, typically in a tissue-specific manner: the acetylation of histone H3 at lysine 27 (H3K27ac) [Creyghton et al., 2010]; the monomethylation of histone H3 at lysine 4 (H3K4me1) [Heintzman et al., 2007]; DNase I hypersensitive sites [Dorschner et al., 2004]; and DNA methlyation [Xie et al., 2013, Schlesinger et al., 2013]. Also, the p300 protein is associated with enhancer activity [Visel et al., 2009]. ChIP-seq technology together with suitable antibodies has made genome-wide data on these marks available in multiple cell-lines of several model organisms. For example, this data has been used to estimate that hundreds of thousands of enhancers exist in the human genome [Shen et al., 2012, Dunham et al., 2012, Zhu et al., 2013]. The ENCODE project

has such data in many human and mouse cell lines and the modENCODE project has similar data for several *Caenorhabditis* and *Drosophila* species.

To summarise, it is now possible to study tissue-specific TF binding by analysing tissue-specific enhancers from different cell lines. These enhancers have been identified on a genomic scale. The STEME algorithm could identify motifs that are relatively prevalent in the enhancers of particular tissues as compared to other tissues. The BiFA algorithm could apply these motifs to provide TFBS predictions as data for the transcriptional programs model. Recalling that the transcriptional programs model is a hierarchical Dirichlet process, it is easy to imagine extending the hierarchy to include one or more tissue-specific levels. These extra levels could capture tissue-specific TF interactions.

Another technology that will help our understanding of transcriptional regulatory networks is Chromosome Conformation Capture (3C) [Dekker et al., 2002] and its extensions: Circularized Chromosome Conformation Capture (4C) [Simonis et al., 2006, Zhao et al., 2006] and Carbon-Copy Chromosome Conformation Capture (5C) [Dostie and Dekker, 2007]. These techniques analyse the 3-dimensional spatial organisation of chromosomes in the nucleus. They are relevant to transcriptional regulatory networks because they can determine which enhancers and genes are in close physical proximity. This appears to be a frequent phenomenon even when they are well separated by genomic distance. Decoding transcriptional regulatory networks relies upon associating enhancer activity with changes in gene expression. Historically it has been difficult to reliably infer these associations. Until recently models designed to integrate expression data with binding data have typically associated genomic binding locations with the nearest gene. Whilst this is a practical solution given the lack of information, long-range regulation is known to occur. However, it is not known how prevalent it is. 3C, 4C and 5C technology will help to investigate this. It would be possible to extend the transcriptional programs model to incorporate expression data, perhaps in the same way that Gerber et al. did [Gerber et al., 2007]. In this case, I would hope to include chromosome conformation data in the model.

5.3.2 Technical extensions

Higher order background models

Both tasks of TFBS prediction and motif search rely on detecting TFBSs against a background of genomic sequence. It is well known that 0-order Markov models do not fit genomic sequences particularly well. Common conceptions exist that second or third order models are suitable when a better fit is desired (see Section 1.4.3). However, recent work has demonstrated that such common conceptions may be misplaced [Narlikar et al.,

2013]. Narlikar et al. applied the Akaike information criterion (AIC) [Akaike, 1974] and Bayesian information criterion (BIC) [Schwarz, 1978] model selection criteria to Markov models of varying order. They studied the tasks of motif discovery and phylogeny reconstruction and found that order 7 models were optimal for human genomes. They suggest that the performance of many sequence analysis algorithms could be improved easily through the determination of the best Markov order to use.

STEME as a full motif finder

The work described in Chapter 3 is an efficient approximation to the EM algorithm at the core of the MEME algorithm. However, replacing this core with the STEME algorithm results in a motif finder whose run-time is dominated by two other parts of the MEME algorithm: the search for seeds (starting points for the EM algorithm) and the significance calculations used to assess the discovered motifs. I have implemented efficient solutions to both these problems. To search for seeds I have again used suffix trees to implement an efficient algorithm. For the significance calculations I have made a small generalisation to an existing efficient algorithm [Nagarajan et al., 2005] to adapt it to the MEME algorithm. I have left this work outside the scope of this thesis but I hope to publish it soon. A motif search web-server that uses this implementation is available at `http://sysbio.mrc-bsu.cam.ac.uk/STEME/`.

5.4 Acknowledgements

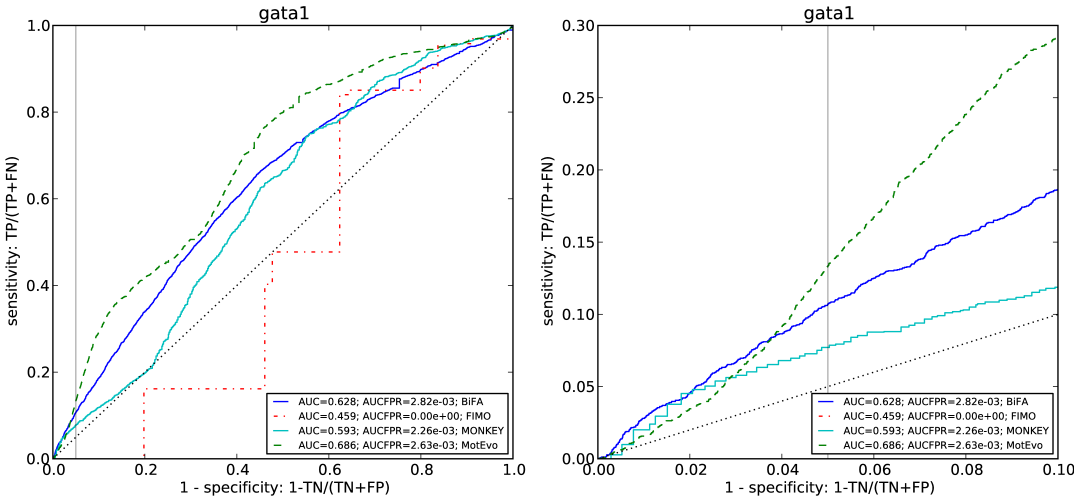
I would like to thank Georgy Koentges and Sascha Ott for many stimulating discussions about the cooperative effects of TFs. Of course I would not have been able to complete this thesis without the support and advice of my supervisor Lorenz Wernisch. Discussions with Sascha Ott were invaluable when designing the BiFA algorithm. Also vital was the support of my wife Kathryn who proofread my papers and this thesis. The quality of English in this thesis would have been much lower without her help. She also sent me interesting links about *Drosophila* from time to time. Many thanks also go to my parents and sisters who knew when it was best not to ask me how the thesis was progressing.

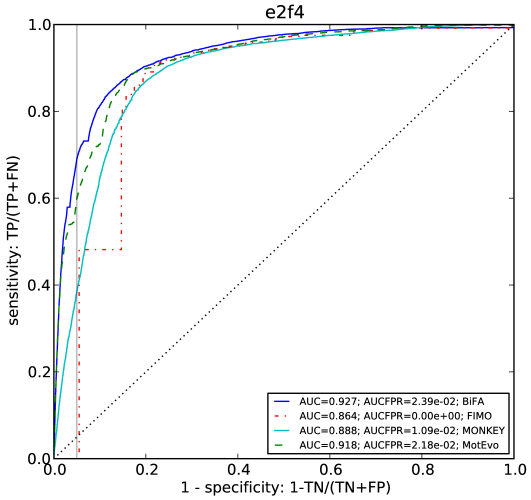
Appendix A

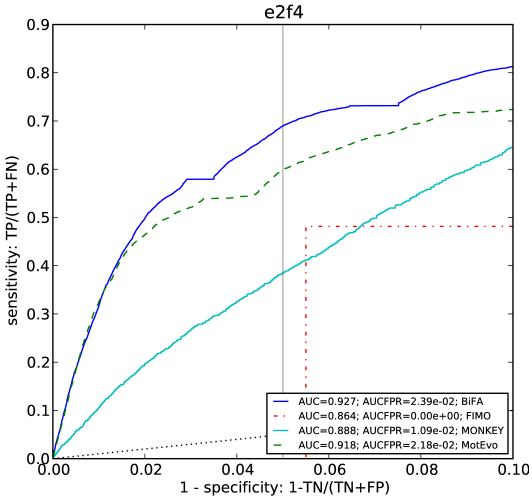
TFBS predictor ROC curves

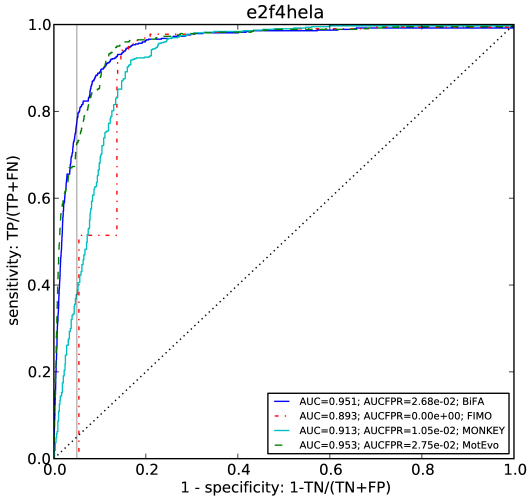
In this appendix I provide a full set of ROC curves from the evaluation of phylogenetic TFBS predictors in Chapter 2. The feint vertical lines show the FPR=5% threshold. The black dotted lines show the expected performance of a random classifier. Note that the plots on the right are the same as the left but the x-axis is scaled to show FPRs between 0% and 10% and the y-axis is scaled to fit the data. The TFs are presented in order of increasing information content of their PWMs as in Tables 2.1, 2.2 and 2.3.

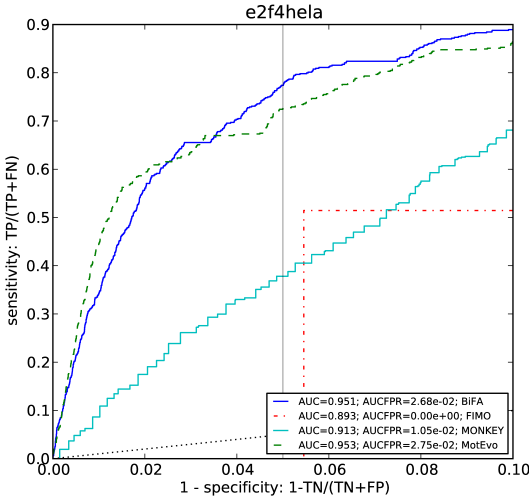
A.1 Håndstad sites benchmark

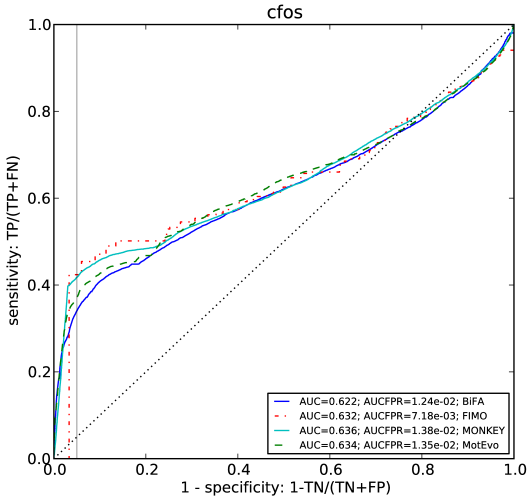


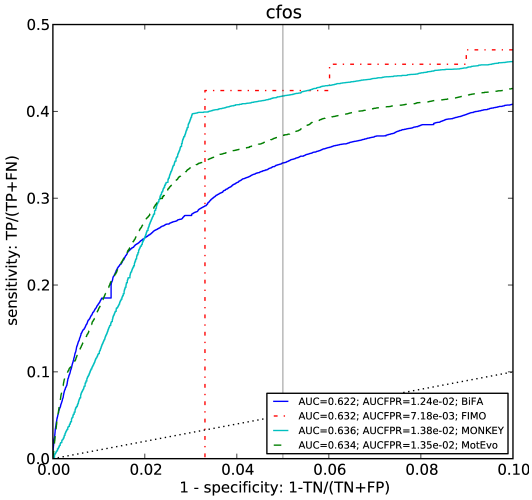

e2f4
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.927; AUCFPR=2.39e-02; BiFA
AUC=0.864; AUCFPR=0.00e+00; FIMO
AUC=0.888; AUCFPR=1.09e-02; MONKEY
AUC=0.918; AUCFPR=2.18e-02; MotEvo

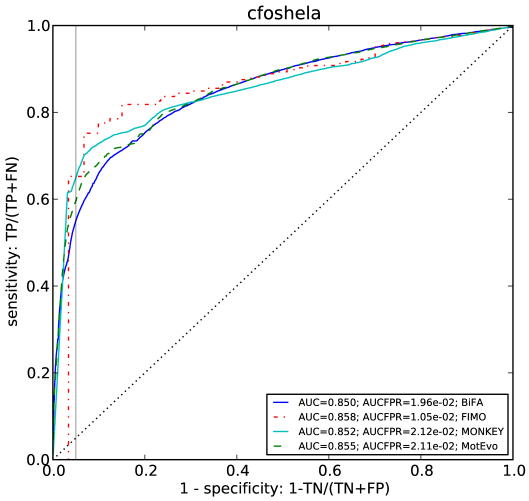

e2f4
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.927; AUCFPR=2.39e-02; BiFA
AUC=0.864; AUCFPR=0.00e+00; FIMO
AUC=0.888; AUCFPR=1.09e-02; MONKEY
AUC=0.918; AUCFPR=2.18e-02; MotEvo

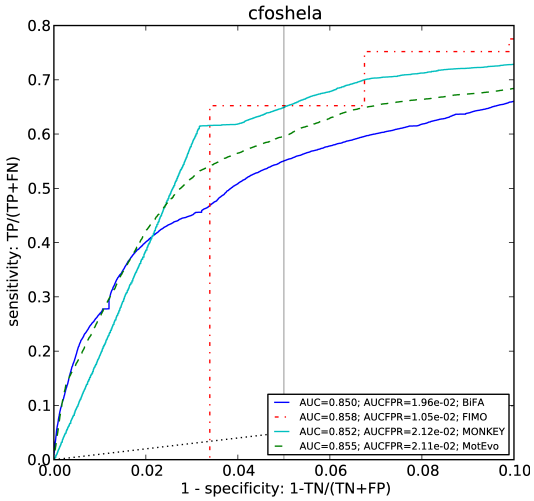

e2f4hela
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.951; AUCFPR=2.68e-02; BiFA
AUC=0.893; AUCFPR=0.00e+00; FIMO
AUC=0.913; AUCFPR=1.05e-02; MONKEY
AUC=0.953; AUCFPR=2.75e-02; MotEvo

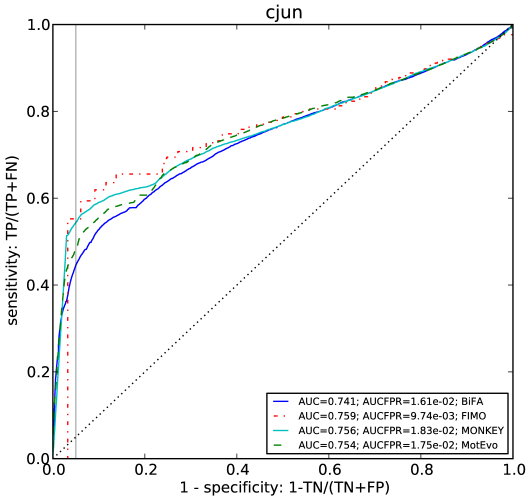

e2f4hela
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.951; AUCFPR=2.68e-02; BiFA
AUC=0.893; AUCFPR=0.00e+00; FIMO
AUC=0.913; AUCFPR=1.05e-02; MONKEY
AUC=0.953; AUCFPR=2.75e-02; MotEvo

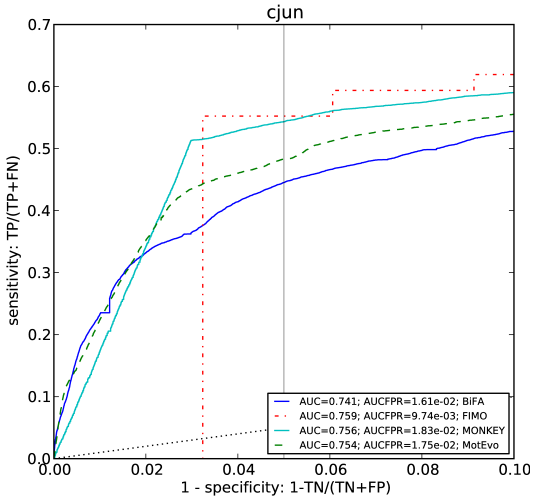

cfos
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.622; AUCFPR=1.24e-02; BiFA
AUC=0.632; AUCFPR=7.18e-03; FIMO
AUC=0.636; AUCFPR=1.38e-02; MONKEY
AUC=0.634; AUCFPR=1.35e-02; MotEvo

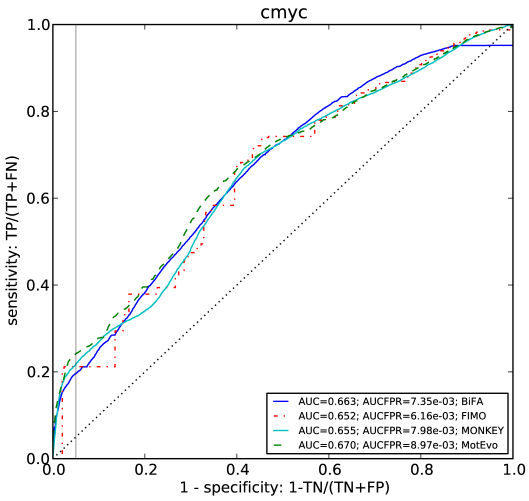

cfos
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.622; AUCFPR=1.24e-02; BiFA
AUC=0.632; AUCFPR=7.18e-03; FIMO
AUC=0.636; AUCFPR=1.38e-02; MONKEY
AUC=0.634; AUCFPR=1.35e-02; MotEvo

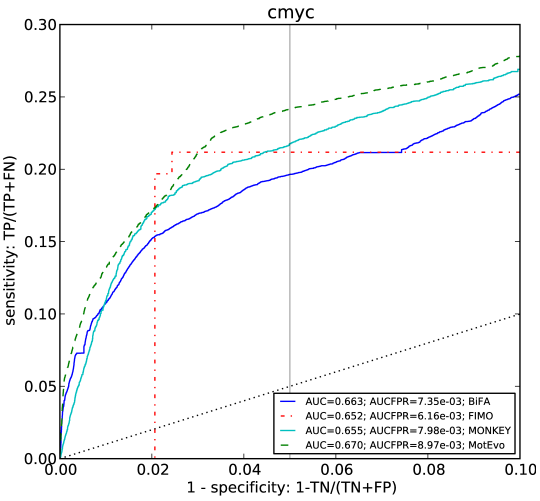

cfoshela
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.850; AUCFPR=1.96e-02; BIFA
AUC=0.858; AUCFPR=1.05e-02; FIMO
AUC=0.852; AUCFPR=2.12e-02; MONKEY
AUC=0.855; AUCFPR=2.11e-02; MotEvo

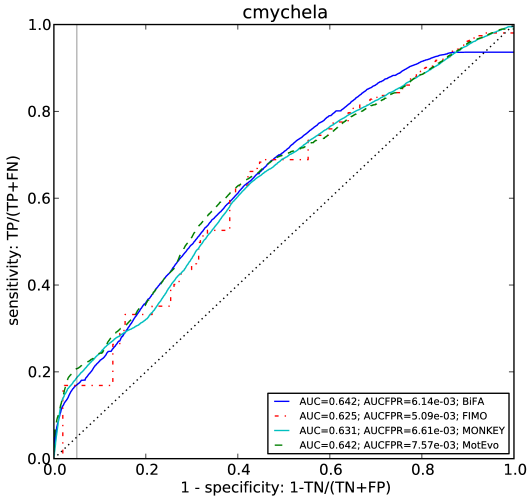

cfoshela
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.850; AUCFPR=1.96e-02; BIFA
AUC=0.858; AUCFPR=1.05e-02; FIMO
AUC=0.852; AUCFPR=2.12e-02; MONKEY
AUC=0.855; AUCFPR=2.11e-02; MotEvo

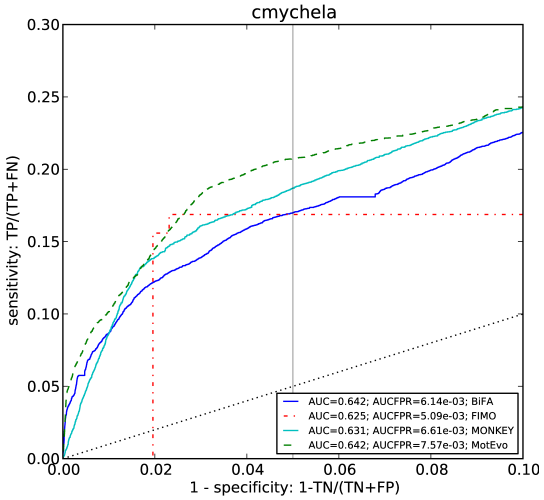

cjun
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.741; AUCFPR=1.61e-02; BIFA
AUC=0.759; AUCFPR=9.74e-03; FIMO
AUC=0.756; AUCFPR=1.83e-02; MONKEY
AUC=0.754; AUCFPR=1.75e-02; MotEvo

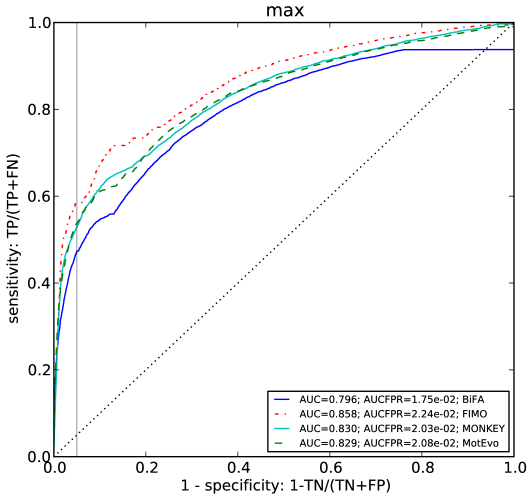

cjun
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.741; AUCFPR=1.61e-02; BIFA
AUC=0.759; AUCFPR=9.74e-03; FIMO
AUC=0.756; AUCFPR=1.83e-02; MONKEY
AUC=0.754; AUCFPR=1.75e-02; MotEvo

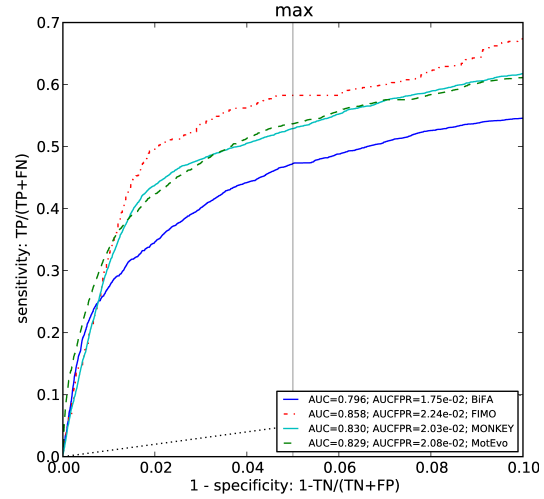

cmyc
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.663; AUCFPR=7.35e-03; BIFA
AUC=0.652; AUCFPR=6.16e-03; FIMO
AUC=0.655; AUCFPR=7.98e-03; MONKEY
AUC=0.670; AUCFPR=8.97e-03; MotEvo

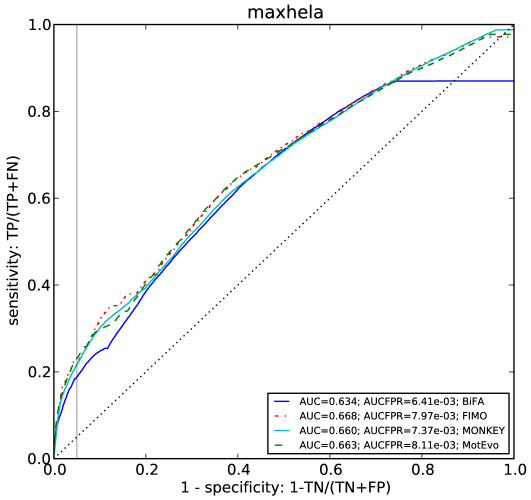

cmyc
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.663; AUCFPR=7.35e-03; BIFA
AUC=0.652; AUCFPR=6.16e-03; FIMO
AUC=0.655; AUCFPR=7.98e-03; MONKEY
AUC=0.670; AUCFPR=8.97e-03; MotEvo

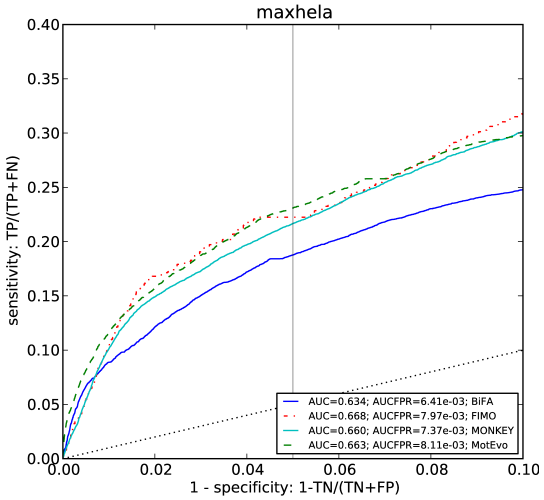

cmychela
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.642; AUCFPR=6.14e-03; BiFA
AUC=0.625; AUCFPR=5.09e-03; FIMO
AUC=0.631; AUCFPR=6.61e-03; MONKEY
AUC=0.642; AUCFPR=7.57e-03; MotEvo

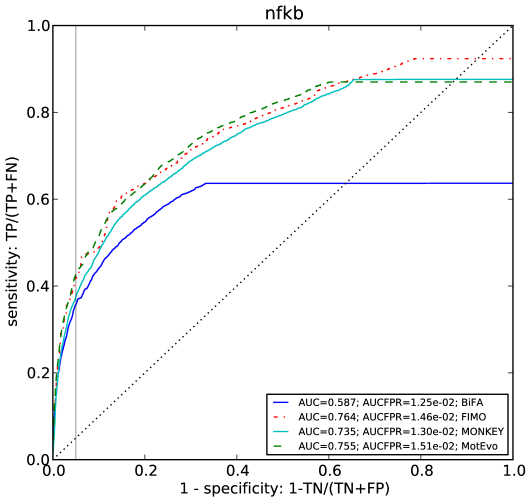

cmychela
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.642; AUCFPR=6.14e-03; BiFA
AUC=0.625; AUCFPR=5.09e-03; FIMO
AUC=0.631; AUCFPR=6.61e-03; MONKEY
AUC=0.642; AUCFPR=7.57e-03; MotEvo

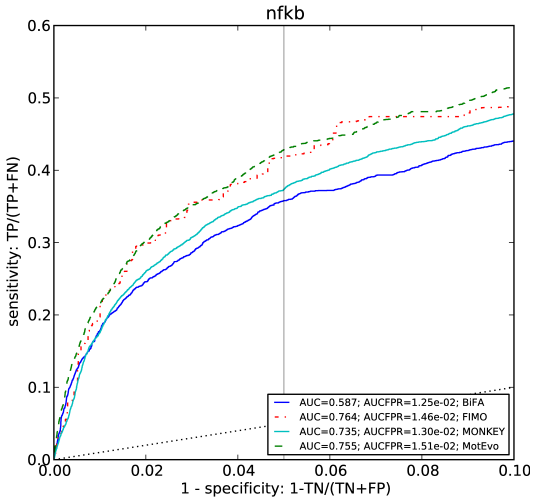

max
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.796; AUCFPR=1.75e-02; BiFA
AUC=0.858; AUCFPR=2.24e-02; FIMO
AUC=0.830; AUCFPR=2.03e-02; MONKEY
AUC=0.829; AUCFPR=2.08e-02; MotEvo

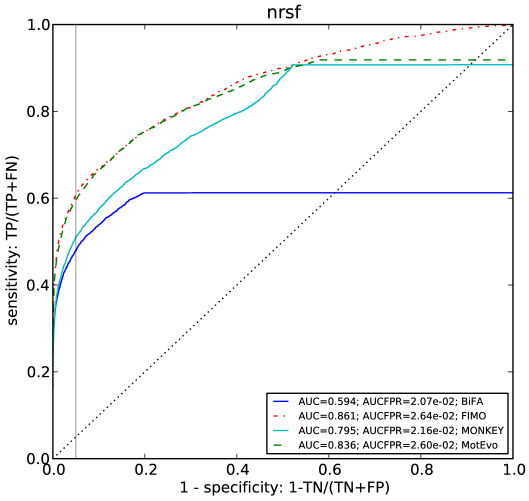

max
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.796; AUCFPR=1.75e-02; BiFA
AUC=0.858; AUCFPR=2.24e-02; FIMO
AUC=0.830; AUCFPR=2.03e-02; MONKEY
AUC=0.829; AUCFPR=2.08e-02; MotEvo

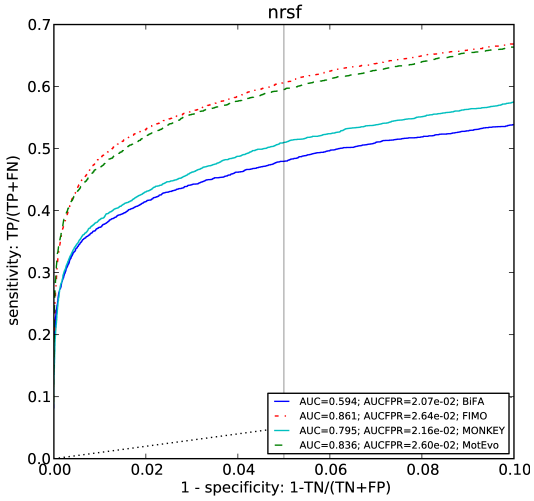

maxhela
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.634; AUCFPR=6.41e-03; BiFA
AUC=0.668; AUCFPR=7.97e-03; FIMO
AUC=0.660; AUCFPR=7.37e-03; MONKEY
AUC=0.663; AUCFPR=8.11e-03; MotEvo


maxhela
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.634; AUCFPR=6.41e-03; BiFA
AUC=0.668; AUCFPR=7.97e-03; FIMO
AUC=0.660; AUCFPR=7.37e-03; MONKEY
AUC=0.663; AUCFPR=8.11e-03; MotEvo

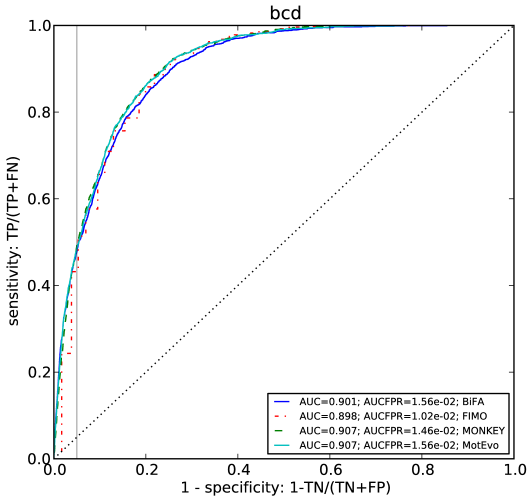

nfkb
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.587; AUCFPR=1.25e-02; BIFA
AUC=0.764; AUCFPR=1.46e-02; FIMO
AUC=0.735; AUCFPR=1.30e-02; MONKEY
AUC=0.755; AUCFPR=1.51e-02; MotEvo

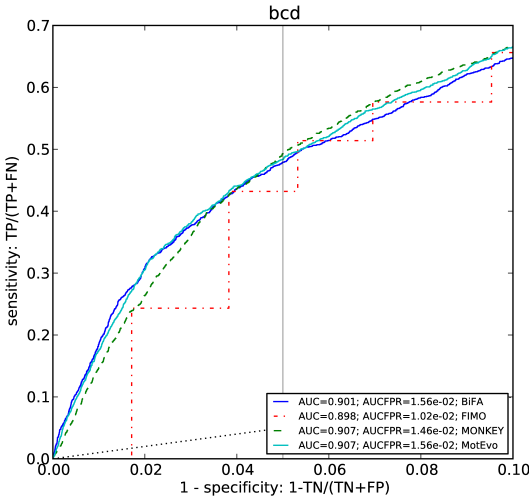

nfkb
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.587; AUCFPR=1.25e-02; BIFA
AUC=0.764; AUCFPR=1.46e-02; FIMO
AUC=0.735; AUCFPR=1.30e-02; MONKEY
AUC=0.755; AUCFPR=1.51e-02; MotEvo

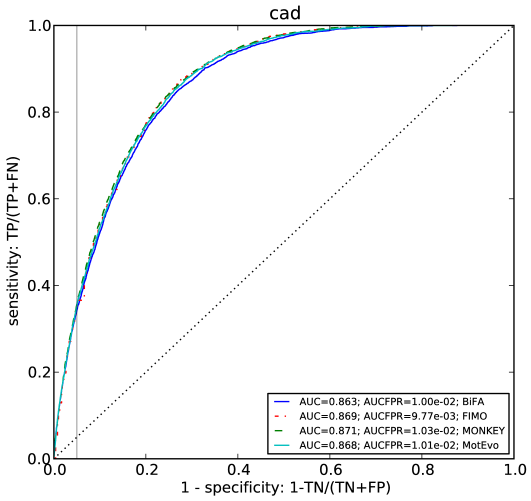

nrsf
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.594; AUCFPR=2.07e-02; BIFA
AUC=0.861; AUCFPR=2.64e-02; FIMO
AUC=0.795; AUCFPR=2.16e-02; MONKEY
AUC=0.836; AUCFPR=2.60e-02; MotEvo

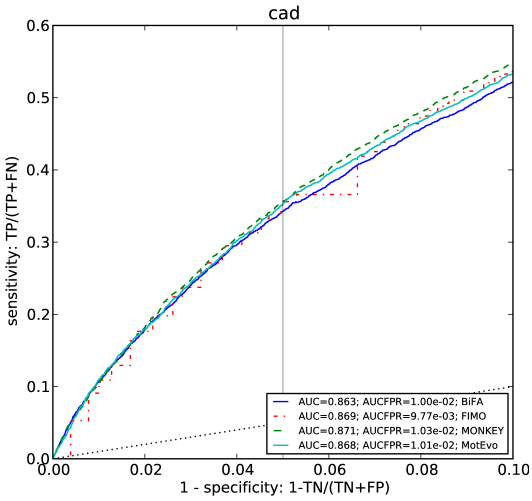

nrsf
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.594; AUCFPR=2.07e-02; BIFA
AUC=0.861; AUCFPR=2.64e-02; FIMO
AUC=0.795; AUCFPR=2.16e-02; MONKEY
AUC=0.836; AUCFPR=2.60e-02; MotEvo

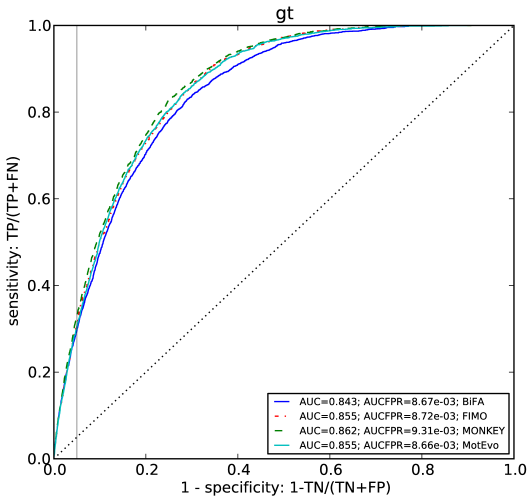
A.2 Turnover benchmark

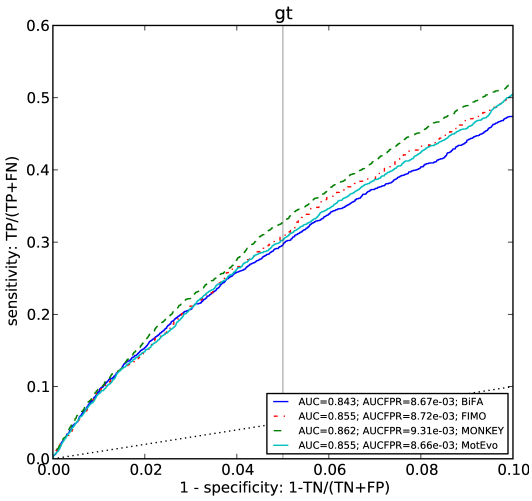


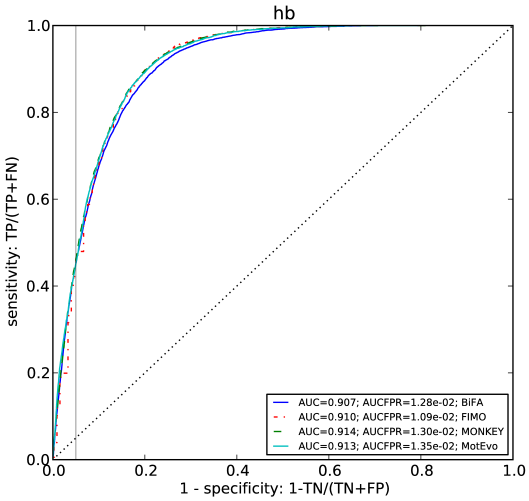


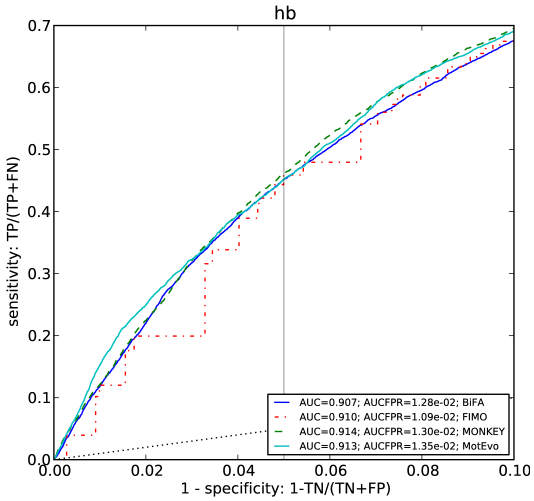


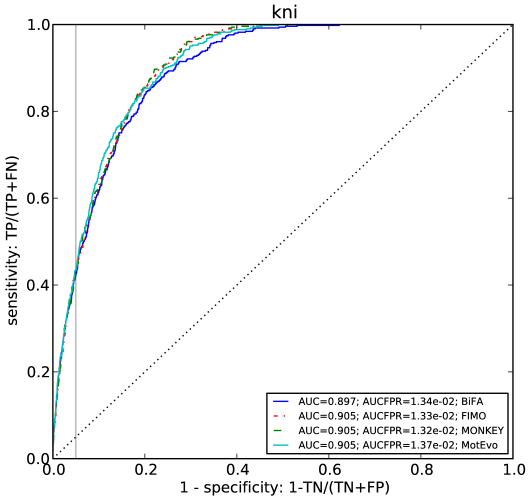


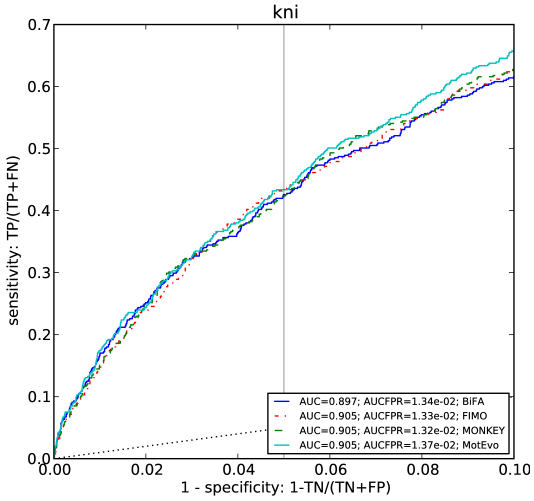


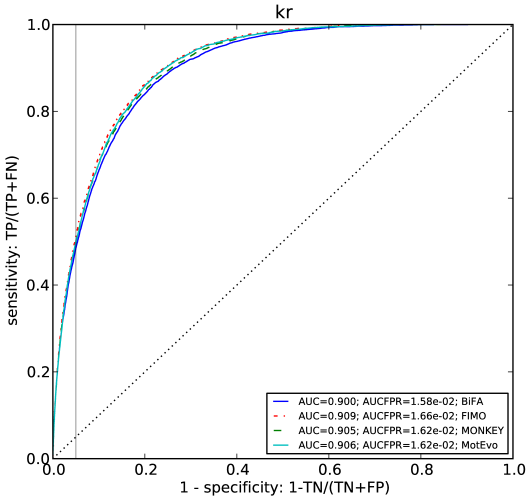


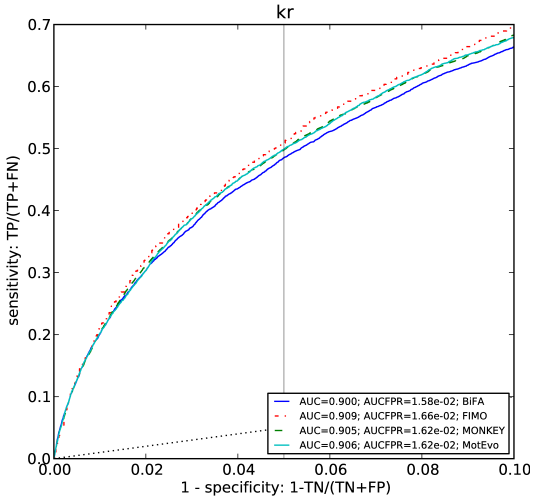

hb
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.907; AUCFPR=1.28e-02; BiFA
AUC=0.910; AUCFPR=1.09e-02; FIMO
AUC=0.914; AUCFPR=1.30e-02; MONKEY
AUC=0.913; AUCFPR=1.35e-02; MotEvo


hb
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.907; AUCFPR=1.28e-02; BiFA
AUC=0.910; AUCFPR=1.09e-02; FIMO
AUC=0.914; AUCFPR=1.30e-02; MONKEY
AUC=0.913; AUCFPR=1.35e-02; MotEvo

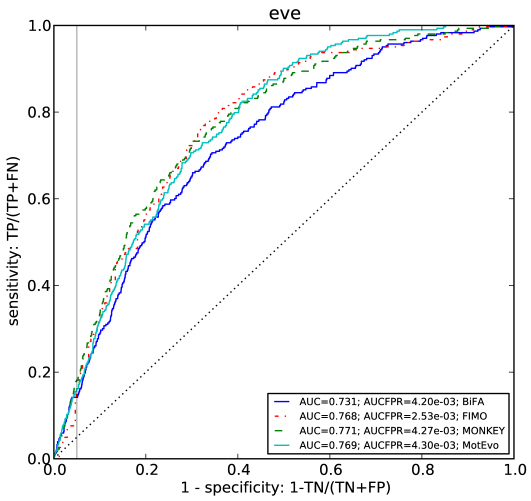

kni
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.897; AUCFPR=1.34e-02; BiFA
AUC=0.905; AUCFPR=1.33e-02; FIMO
AUC=0.905; AUCFPR=1.32e-02; MONKEY
AUC=0.905; AUCFPR=1.37e-02; MotEvo

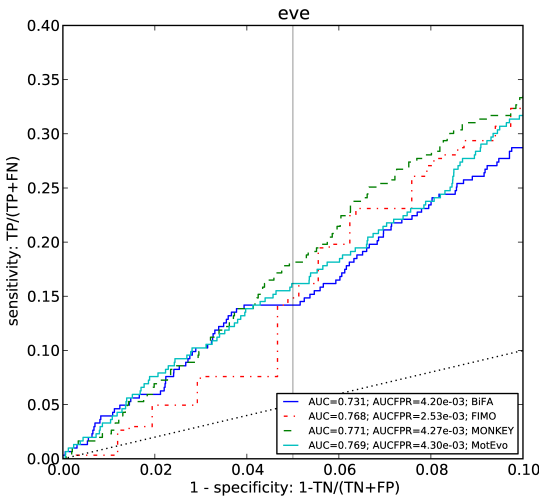

kni
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.897; AUCFPR=1.34e-02; BiFA
AUC=0.905; AUCFPR=1.33e-02; FIMO
AUC=0.905; AUCFPR=1.32e-02; MONKEY
AUC=0.905; AUCFPR=1.37e-02; MotEvo

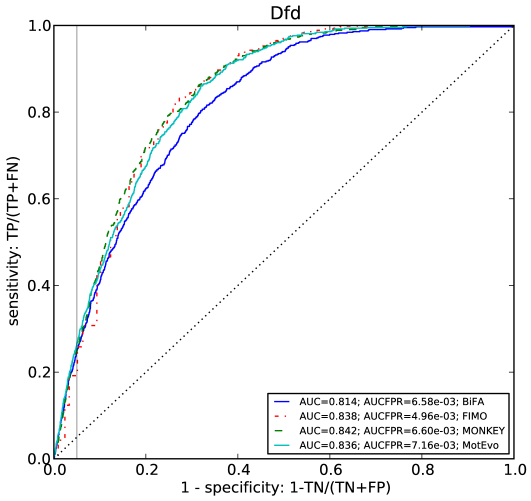

kr
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.900; AUCFPR=1.58e-02; BiFA
AUC=0.909; AUCFPR=1.66e-02; FIMO
AUC=0.905; AUCFPR=1.62e-02; MONKEY
AUC=0.906; AUCFPR=1.62e-02; MotEvo

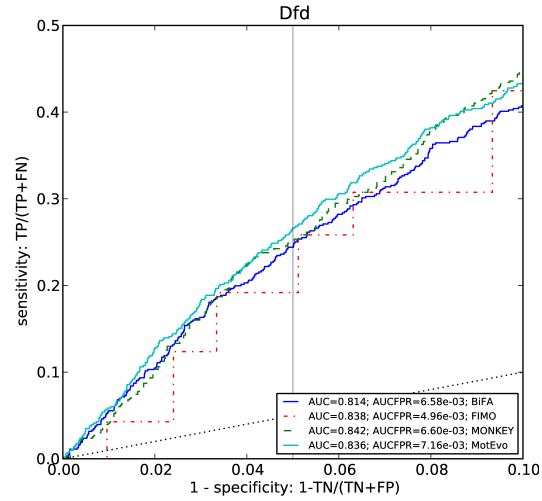

kr
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.900; AUCFPR=1.58e-02; BiFA
AUC=0.909; AUCFPR=1.66e-02; FIMO
AUC=0.905; AUCFPR=1.62e-02; MONKEY
AUC=0.906; AUCFPR=1.62e-02; MotEvo

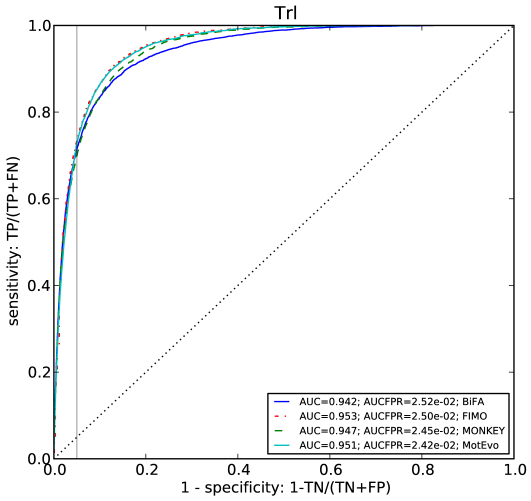
A.3 modENCODE benchmark

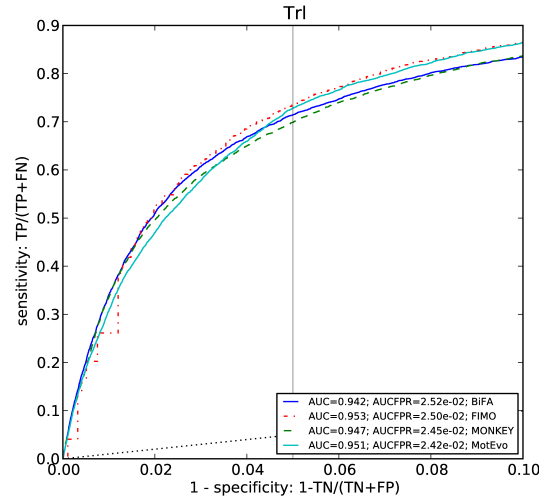


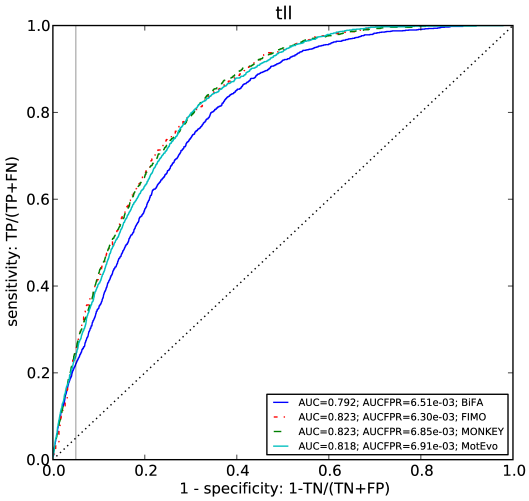


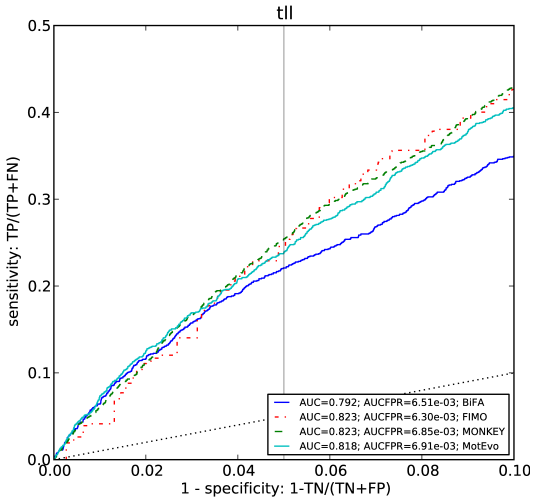


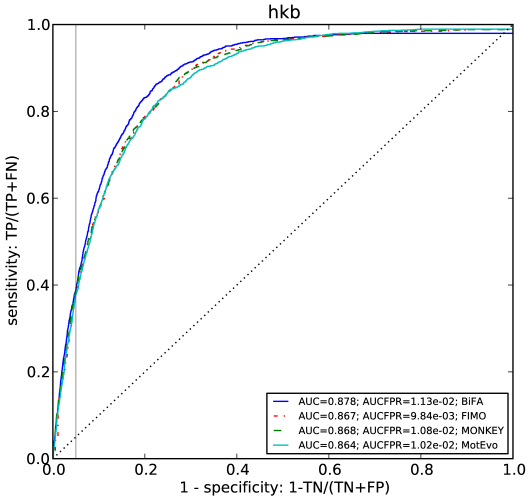


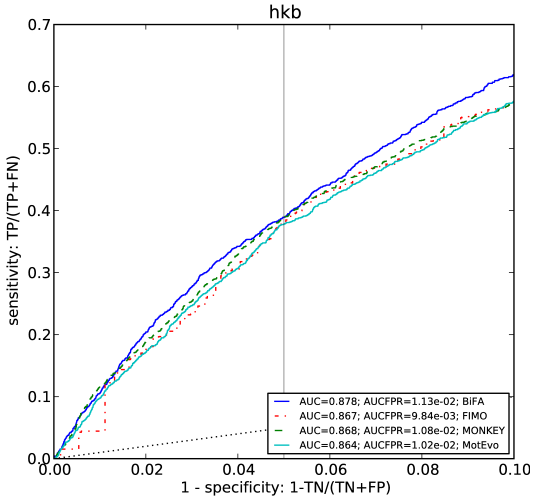


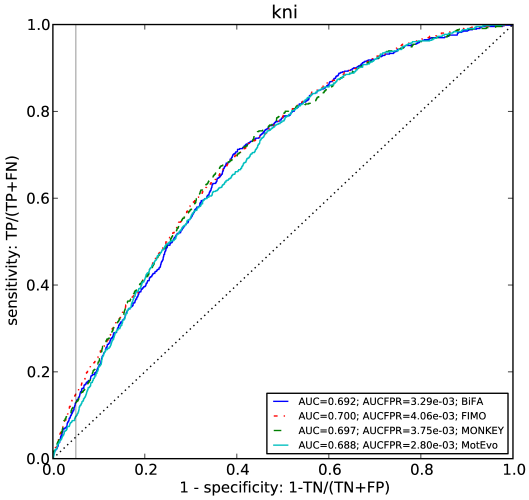


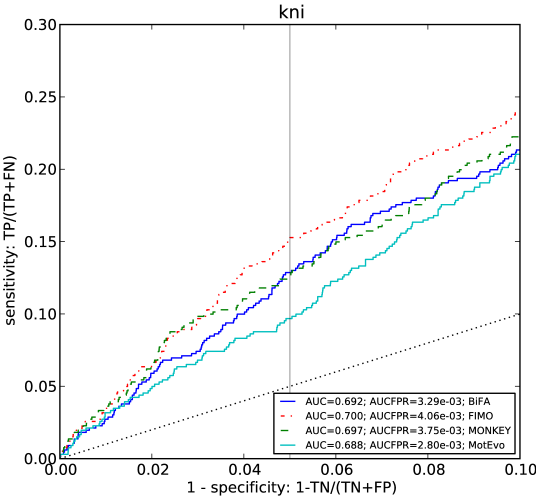

tll
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.792; AUCFPR=6.51e-03; BIFA
AUC=0.823; AUCFPR=6.30e-03; FIMO
AUC=0.823; AUCFPR=6.85e-03; MONKEY
AUC=0.818; AUCFPR=6.91e-03; MotEvo


tll
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.792; AUCFPR=6.51e-03; BIFA
AUC=0.823; AUCFPR=6.30e-03; FIMO
AUC=0.823; AUCFPR=6.85e-03; MONKEY
AUC=0.818; AUCFPR=6.91e-03; MotEvo


hkb
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.878; AUCFPR=1.13e-02; BIFA
AUC=0.867; AUCFPR=9.84e-03; FIMO
AUC=0.868; AUCFPR=1.08e-02; MONKEY
AUC=0.864; AUCFPR=1.02e-02; MotEvo


hkb
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.878; AUCFPR=1.13e-02; BIFA
AUC=0.867; AUCFPR=9.84e-03; FIMO
AUC=0.868; AUCFPR=1.08e-02; MONKEY
AUC=0.864; AUCFPR=1.02e-02; MotEvo


kni
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.692; AUCFPR=3.29e-03; BIFA
AUC=0.700; AUCFPR=4.06e-03; FIMO
AUC=0.697; AUCFPR=3.75e-03; MONKEY
AUC=0.688; AUCFPR=2.80e-03; MotEvo


kni
sensitivity: TP/(TP+FN)
1 - specificity: 1-TN/(TN+FP)
AUC=0.692; AUCFPR=3.29e-03; BIFA
AUC=0.700; AUCFPR=4.06e-03; FIMO
AUC=0.697; AUCFPR=3.75e-03; MONKEY
AUC=0.688; AUCFPR=2.80e-03; MotEvo

Glossary

adenine a nucleobase, complementary to thymine. 1

alternative hypothesis the hypothesis that supposes there is an effect. It is tested against the null hypothesis. 18

AUC area under curve: a statistic that measures the performance of a classification method defined as the area under the ROC curve. 20, 57

AUC50 a statistic that measures the performance of a classification method on those examples scored higher than the first 50 FP. 20, 57

AUCFPR a statistic that measures the performance of a classification method up until a certain FPR. 57

BBLS Bayesian branch length score: a scoring scheme that sums the expected total branch lengths in a phylogenetic tree between all leaf nodes that represent a TFBS (see Section 2.2.1). 40, 48, 50, 56, 73, 75

BiFA Binding Factor Analysis: an algorithm to predict TFBSs given phylogenetically related sequences. 41

billboard a model of combinatorial TF binding where the positioning and orientation of the TFBSs is unimportant. 7

BLS branch length score: a scoring scheme that sums the total branch lengths in a phylogenetic tree between all leaf nodes that represent a TFBS (see Section 2.2.1). 37, 40

central sequence the primary sequence under consideration in a phylogenetic analysis. 37, 41, 42, 47, 54, 58, 67

chain a sequence of TFBSs that are conserved across multiple sequences. 44

chain element a set of TFBSs for the same PWM, one in each sequence. That is, an element of the maximal chain. 45

ChIP chromatin immunoprecipitation: a technique that determines which genomic regions are associated with a protein. 10

chromatin the combination of DNA and proteins that make up the contents of the nucleus of a cell. 6

chromosome a DNA macromolecule that carries genetic information. 1

cofactor a TF that synergistically encourages another TF to bind. 5

competitive binding when two or more TFs compete for a TFBS. 5

consensus sequence the sequence that a TF prefers to bind to. 3

CpG island a region of the genome with high CG dinucleotide count. 26

CpG suppression the effect whereby regions of the genome have a low CG dinucleotide count. 26

CRM cis-regulatory module: a regulatory region of the genome containing multiple TFBSs. 3

cytosine a nucleobase, complementary to guanine. 1

DamID DNA adenine methyltransferase identification: a technique that locates TFBSs by expressing the proposed TF as a fusion protein with DNA methyltransferase. 10

DNA deoxyribonucleic acid: a nucleic acid. One of the three major macromolecules essential for all known forms of life. 1

DNase I deoxyribonuclease I: a nuclease that cleaves DNA. 6, 9

DNase I footprinting an assay to detect protein-DNA binding using DNase I. 9

effector a molecule, chemical, or structure that regulates a pathway by increasing or decreasing the pathway's reaction rate. 48

EM algorithm Expectation-Maximisation algorithm: an algorithm that makes point estimates of parameters in a probabilistic model that maximise the expected likelihood of the data. 16

EMSA electrophoretic mobility shift assay: a common affinity electrophoresis technique used to study protein–DNA or protein–RNA interactions. 9

enhanceosome a model of combinatorial TF binding where the positioning and orientation of the TFBSs is critical. 7

enhancer a regulatory region of the genome some distance from its target gene. 3

entropy measures the expected amount of information (or surprisal) given by a draw from a distribution. The entropy can be seen as a measure of the uncertainty associated with a distribution. 17

epigenetic modification a heritable change caused by mechanisms other than changes to DNA sequence. 7

***E*-value** the number of tests multiplied by the p-value. Suppose you performed (or could have performed) N tests and the most extreme p-value you found was p. The E-value is Np. 18

FDR false discovery rate: the percentage of positive predictions a classifier incorrectly makes. 19, 57, 74

FN false negative: an incorrect negative prediction. 19, 54

FP false positive: an incorrect positive prediction. 19–21, 53, 130

FPR false positive rate: the number of false positives as a proportion of the total negatives. 19–21, 57, 130

gene unit of genetic information. 1

gene expression process by which information from a gene is used to make a functional gene product. 1

gene product product of gene expression, commonly a protein. 1

gene regulatory network a set of genes that interact with each other to control the rates at which their gene products are expressed. 2

genome the sum of an individual's inheritable traits. 1

guanine a nucleobase, complementary to cytosine. 1

in vivo footprinting assay an *in vivo* extension of the DNase I footprinting assay to detect protein-DNA binding. 9

IQR inter-quartile range: the difference between the first and third quartiles. 57, 64

KL-divergence Kullback-Leibler divergence: a measure of separation between two probability distributions or densities, also known as the relative entropy. 15

likelihood function a function of the parameters of a probabilistic model that is equivalent to the probability of the data given those parameters. 13

maximal chain the longest and most probable sequence of TFBSs that appear in all the sequences in a BiFA analysis. 41

microarray a high-throughput screening method that typically measures expression levels. 10

MITOMI mechanically induced trapping of molecular interactions: uses micro-fluidics to determine the energy landscape of TF-DNA interactions. 10

MLE maximum likelihood estimate: an estimate of some parameters of a probabilistic model that maximises the likelihood function. 16

motif finding the task of searching for a TF's binding preferences in a set of sequences. 26

motif scanning the task of searching for TFBSs in a sequence given the TF's sequence preferences. 30

nucleobase part of each nucleotide. 1

nucleosome the basic unit of DNA packaging in eukaryotes. 6

null hypothesis the hypothesis that the alternative hypothesis is tested against. It typically describes a default position. 18

one-hybrid system assays to detect protein-DNA binding by transforming cells with a TF of interest and potential binding sequences. 10

PBM protein-binding microarray: a microarray that measures the affinity of TF-DNA interactions. 10

pioneer TF a TF that remodels chromatin allowing other TFs to access the DNA. 6

PMM phylogenetic motif model: a model that describes how likely each base pair is in a multiple alignment. Used to scan multiple alignments for instances of motifs (see Section 2.2.1). 37, 38, 40, 74, 76

position-specific prior a prior in a probabilistic model that depends on the location in the genome. It is a flexible method of integrating diverse location data into motif scanning and motif finding models. 36

probabilistic model a way of specifying a joint distribution over a set of random variables. 11

promoter a genomic region near the TSS where the transcriptional machinery assembles. 3

PSSM position specific scoring matrix: a matrix of scores for different bases at each position in a TFBS. 30

***p*-value** the probability of observing a test statistic at least as extreme as the test statistic generated by the data if the null hypothesis is true. 18

PWM position weight matrix: a matrix of frequencies identifying how often each base occurs at each position in a TFBS. 22, 38, 50

PWM scanning the task of searching for TFBSs in a sequence given the TF's sequence preferences. 30

RE response element: another term for a TFBS. 3

regulon a collection of genes or operons under regulation by the same TF. 37

related sequence a sequence that is phylogenetically related to the central sequence. 37, 41, 47, 67

reporter gene a gene that is attached to a regulatory region of interest. The expression of the gene normally results in a phenotype that can be measured (or reported). 10

RNA ribonucleic acid: a nucleic acid. One of the three major macromolecules essential for all known forms of life. 1

ROC receiver operating characteristic: a comparison of the performance of a binary classifier as the discrimination threshold varies. 19, 20, 57

SELEX systematic evolution of ligands by exponential enrichment: a method to determine which sequences from a random library a TF binds to. 10

STEME algorithm an approximation to the EM algorithm that uses suffix trees. 78

surprisal a measure of the information content associated with the outcome of a random variable. Less likely outcomes have higher surprisals. 16

TF transcription factor: a protein that affects the rate of transcription of a gene, usually by binding to DNA at a TFBS. 3

TFBS transcription factor binding site: the location on the DNA that a TF binds to. 3

thymine a nucleobase, complementary to adenine. 1

TN true negative: a correct negative prediction. 19, 53

TP true positive: a correct positive prediction. 19, 53

TPR true positive rate: the number of true positives as a proportion of the total positives. 19, 40

transcription the first stage in gene expression in which DNA is copied into RNA. 2

transcriptional program transcriptional program: a set of TFs that act in a coordinated fashion to regulate a set of target genes. 95

TSS transcription start site: the genomic location where transcription of a gene starts. 3

variational inference a Bayesian inference technique that approximates the posterior distribution over the hidden variables in a probabilistic model. 17

WS weighted sum: a scoring scheme that sums weighted motif scores across species in a phylogenetic tree (see Section 2.2.1). 37, 50, 56, 73, 75

Bibliography

[Aerts et al., 2003] Aerts, S., Thijs, G., Coessens, B., Staes, M., Moreau, Y. and Moor, B. D. (2003). Toucan: deciphering the cis-regulatory logic of coregulated genes. Nucleic Acids Research *31*, 1753–1764.

[Agalioti et al., 2000] Agalioti, T., Lomvardas, S., Parekh, B., Yie, J., Maniatis, T. and Thanos, D. (2000). Ordered recruitment of chromatin modifying and general transcription factors to the IFN-beta promoter. Cell *103*, 667–678.

[Akaike, 1974] Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control *19*, 716–723.

[Alexa et al., 2006] Alexa, A., Rahnenführer, J. and Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics *22*, 1600–1607.

[Amano et al., 2009] Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H. and Shiroishi, T. (2009). Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription. Developmental Cell *16*, 47–57.

[Aparicio et al., 2004] Aparicio, O., Geisberg, J. V. and Struhl, K. (2004). Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. Current protocols in cell biology / editorial board, Juan S. Bonifacino ... [et al.] *Chapter 17*, Unit 17.7.

[Arce et al., 2006] Arce, L., Yokoyama, N. N. and Waterman, M. L. (2006). Diversity of LEF/TCF action in development and disease. Oncogene *25*, 7492–7504.

[Arnold et al., 2012] Arnold, P., Erb, I., Pachkov, M., Molina, N. and van Nimwegen, E. (2012). MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. Bioinformatics *28*, 487–494.

[Arnosti and Kulkarni, 2005] Arnosti, D. N. and Kulkarni, M. M. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? Journal of Cellular Biochemistry *94*, 890–898.

[Ashburner et al., 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics *25*, 25–29.

[Badis et al., 2009] Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., Kuznetsov, H., Wang, C.-F., Coburn, D., Newburger, D. E., Morris, Q., Hughes, T. R. and Bulyk, M. L. (2009). Diversity and Complexity in DNA Recognition by Transcription Factors. Science *324*, 1720–1723.

[Bailey, 2011] Bailey, T. L. (2011). DREME: Motif discovery in transcription factor ChIP-seq data. Bioinformatics *27*, 1653–1659.

[Bailey et al., 1994] Bailey, T. L., Elkan, C., University of California, S. D. D. o. C. S. and Engineering (1994). Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Bipolymers. Citeseer.

[Bailey et al., 1995] Bailey, T. L., Elkan, C. et al. (1995). The value of prior knowledge in discovering motifs with MEME. In Proc Int Conf Intell Syst Mol Biol vol. 3, p. 21–9, Menlo Park, Calif. : AAAI Press, c1993-.

[Bais et al., 2011] Bais, A. S., Kaminski, N. and Benos, P. V. (2011). Finding subtypes of transcription factor motif pairs with distinct regulatory roles. Nucleic Acids Research *39*, e76–e76.

[Barash et al., 2005] Barash, Y., Elidan, G., Kaplan, T. and Friedman, N. (2005). CIS: compound importance sampling method for protein–DNA binding site p-value estimation. Bioinformatics *21*, 596–600.

[Barsky et al., 2008] Barsky, M., Stege, U., Thomo, A. and Upton, C. (2008). A new method for indexing genomes using on-disk suffix trees. In Proceedings of the 17th ACM conference on Information and knowledge management p. 649, ACM Press.

[Barsky et al., 2009] Barsky, M., Stege, U., Thomo, A. and Upton, C. (2009). Suffix trees for very large genomic sequences. In Proceedings of the 18th ACM conference on Information and knowledge management p. 1417, ACM Press.

[Bateman et al., 2012] Bateman, J. R., Johnson, J. E. and Locke, M. N. (2012). Comparing Enhancer Action in cis and in trans. Genetics *191*, 1143–1155.

[Baxter et al., 2012] Baxter, L., Jironkin, A., Hickman, R., Moore, J., Barrington, C., Krusche, P., Dyer, N. P., Buchanan-Wollaston, V., Tiskin, A., Beynon, J., Denby, K. and Ott, S. (2012). Conserved Noncoding Sequences Highlight Shared Components of Regulatory Networks in Dicotyledonous Plants. The Plant Cell Online *online before print*.

[Beckstette et al., 2006] Beckstette, M., Homann, R., Giegerich, R. and Kurtz, S. (2006). Fast index based algorithms and software for matching position specific scoring matrices. BMC bioinformatics *7*, 389.

[Bell and Felsenfeld, 2000] Bell, A. C. and Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature *405*, 482–485.

[Ben-Gal et al., 2005] Ben-Gal, I., Shani, A., Gohr, A., Grau, J., Arviv, S., Shmilovici, A., Posch, S. and Grosse, I. (2005). Identification of transcription factor binding sites with variable-order Bayesian networks. Bioinformatics *21*, 2657–2666.

[Benos et al., 2002] Benos, P. V., Bulyk, M. L. and Stormo, G. D. (2002). Additivity in protein–DNA interactions: how good an approximation is it? Nucleic Acids Research *30*, 4442–4451.

[Berg and von Hippel, 1987] Berg, O. G. and von Hippel, P. H. (1987). Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters. Journal of molecular biology *193*, 723–743.

[Bernat et al., 2006] Bernat, J. A., Crawford, G. E., Ogurtsov, A. Y., Collins, F. S., Ginsburg, D. and Kondrashov, A. S. (2006). Distant conserved sequences flanking endothelial-specific promoters contain tissue-specific DNase-hypersensitive sites and over-represented motifs. Human Molecular Genetics *15*, 2098–2105.

[Birney et al., 2007] Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Stamatoyannopoulos, J. A., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas,

P. A., Neri, F., Parker, S. C. J., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu1, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Dutta, A., Guigó, R., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Flicek, P., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermüller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K., Sung, W.-K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Dermitzakis, E. T., Margulies, E. H., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C.-L., Ruan, Y., Snyder, M., Birney, E., Struhl, K., Gerstein, M., Antonarakis, S. E., Gingeras, T. R., Brown, J. B., Flicek, P., Fu, Y., Keefe, D., Birney, E., Denoeud, F., Gerstein, M., Green, E. D., Kapranov, P., Karaöz, U., Myers, R. M., Noble, W. S., Reymond, A., Rozowsky, J., Struhl, K., Siepel, A., Stamatoyannopoulos, J. A., Taylor, C. M., Taylor, J., Thurman, R. E., Tullius, T. D., Washietl, S., Zheng, D., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Collins, F. S., Margulies, E. H., Cooper, G. M., Asimenos, G., Thomas, D. J., Dewey, C. N., Siepel, A., Birney, E., Keefe, D., Hou, M., Taylor, J., Nikolaev, S., Montoya-Burgos, J. I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Brown, J. B., Huang, H., Zhang, N. R., Bickel, P., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., Sequencing Program*, N. C., Human Genome Sequencing Center*, B. C. o. M., Genome Sequencing Center*, W. U., Institute*, B., Oakland Research Institute*, C. H., Gerstein, M., Antonarakis, S. E., Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Pachter, L., Green, E. D., Sidow, A., Weng, Z., Trinklein, N. D., Fu, Y., Zhang, Z. D., Karaöz, U., Barrera, L., Stuart, R., Zheng, D., Ghosh, S., Flicek, P., King, D. C., Taylor, J., Ameur, A., Enroth, S., Bieda, M. C., Koch, C. M., Hirsch, H. A., Wei, C.-L., Cheng, J., Kim, J., Bhinge, A. A., Giresi, P. G., Jiang, N., Liu, J., Yao, F., Sung, W.-K., Chiu, K. P., Vega, V. B., Lee, C. W., Ng, P., Shahab, A., Sekinger, E. A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Clelland, G. K., Wilcox, S., Dil-

lon, S. C., Andrews, R. M., Fowler, J. C., Couttet, P., James, K. D., Lefebvre, G. C., Bruce, A. W., Dovey, O. M., Ellis, P. D., Dhami, P., Langford, C. F., Carter, N. P., Vetrie, D., Kapranov, P., Nix, D. A., Bell, I., Patel, S., Rozowsky, J., Euskirchen, G., Hartman, S., Lian, J., Wu, J., Urban, A. E., Kraus, P., Van Calcar, S., Heintzman, N., Hoon Kim, T., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G., Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X., Xu, M., Haidar, J. N. S., Yu, Y., Birney*, E., Weissman, S., Ruan, Y., Lieb, J. D., Iyer, V. R., Green, R. D., Gingeras, T. R., Wadelius, C., Dunham, I., Struhl, K., Hardison, R. C., Gerstein, M., Farnham, P. J., Myers, R. M., Ren, B., Snyder, M., Thomas, D. J., Rosenbloom, K., Harte, R. A., Hinrichs, A. S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A. S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R. M., Karolchik, D., Haussler, D., Kent, W. J., Dermitzakis, E. T., Armengol, L., Bird, C. P., Clark, T. G., Cooper, G. M., de Bakker, P. I. W., Kern, A. D., Lopez-Bigas, N., Martin, J. D., Stranger, B. E., Thomas, D. J., Woodroffe, A., Batzoglou, S., Davydov, E., Dimas, A., Eyras, E., Hallgrímsdóttir, I. B., Hardison, R. C., Huppert, J., Sidow, A., Taylor, J., Trumbower, H., Zody, M. C., Guigó, R., Mullikin, J. C., Abecasis, G. R., Estivill, X., Birney, E., Bouffard, G. G., Guan, X., Hansen, N. F., Idol, J. R., Maduro, V. V., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., Young, A. C., Blakesley, R. W., Muzny, D. M., Sodergren, E., Wheeler, D. A., Worley, K. C., Jiang, H., Weinstock, G. M., Gibbs, R. A., Graves, T., Fulton, R., Mardis, E. R., Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh, K., Lander, E. S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B. and de Jong, P. J. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799–816.

[Blat and Kleckner, 1999] Blat, Y. and Kleckner, N. (1999). Cohesins Bind to Preferential Sites along Yeast Chromosome III, with Differential Regulation along Arms versus the Centric Region. Cell *98*, 249–259.

[Blekas et al., 2003] Blekas, K., Fotiadis, D. I. and Likas, A. (2003). Greedy mixture learning for multiple motif discovery in biological sequences. Bioinformatics *19*, 607–617.

[Blow et al., 2010] Blow, M. J., McCulley, D. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Bristow, J., Ren, B., Black, B. L., Rubin, E. M., Visel, A. and Pennacchio, L. A. (2010). ChIP-Seq identification of weakly conserved heart enhancers. Nature Genetics *42*, 806–810.

[Blüthgen et al., 2005] Blüthgen, N., Kiełbasa, S. M. and Herzel, H. (2005). Inferring combinatorial regulation of transcription in silico. Nucleic Acids Research *33*, 272–279.

[Borneman et al., 2007] Borneman, A. R., Gianoulis, T. A., Zhang, Z. D., Yu, H., Rozowsky, J., Seringhaus, M. R., Wang, L. Y., Gerstein, M. and Snyder, M. (2007). Divergence of Transcription Factor Binding Sites Across Related Yeast Species. Science *317*, 815–819.

[Boy-Marcotte et al., 1999] Boy-Marcotte, E., Lagniel, G., Perrot, M., Bussereau, F., Boudsocq, A., Jacquet, M. and Labarre, J. (1999). The heat shock response in yeast: differential regulations and contributions of the Msn2p/Msn4p and Hsf1p regulons. Molecular Microbiology *33*, 274–283.

[Boyer et al., 2005] Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., Gifford, D. K., Melton, D. A., Jaenisch, R. and Young, R. A. (2005). Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells. Cell *122*, 947–956.

[Bradley et al., 2010] Bradley, R. K., Li, X.-Y., Trapnell, C., Davidson, S., Pachter, L., Chu, H. C., Tonkin, L. A., Biggin, M. D. and Eisen, M. B. (2010). Binding Site Turnover Produces Pervasive Quantitative Changes in Transcription Factor Binding between Closely Related Drosophila Species. PLoS Biol *8*, e1000343.

[Brenowitz et al., 1986] Brenowitz, M., Senear, D. F., Shea, M. A. and Ackers, G. K. (1986). Quantitative DNase footprint titration: a method for studying protein-DNA interactions. Methods in enzymology *130*, 132–181.

[Brunner et al., 2009] Brunner, A. L., Johnson, D. S., Kim, S. W., Valouev, A., Reddy, T. E., Neff, N. F., Anton, E., Medina, C., Nguyen, L., Chiao, E., Oyolu, C. B., Schroth, G. P., Absher, D. M., Baker, J. C. and Myers, R. M. (2009). Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. Genome Research *19*, 1044–1056.

[Buck and Lieb, 2004] Buck, M. J. and Lieb, J. D. (2004). ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics *83*, 349–360.

[Bulyk et al., 2002] Bulyk, M., Johnson, P. and Church, G. (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic acids research *30*, 1255–1261.

[Bulyk, 2005] Bulyk, M. L. (2005). Discovering DNA regulatory elements with bacteria. Nature Biotechnology *23*, 942–944.

[Calhoun and Levine, 2003] Calhoun, V. C. and Levine, M. (2003). Long-range enhancer–promoter interactions in the Scr-Antp interval of the Drosophila Antennapedia complex. Proceedings of the National Academy of Sciences *100*, 9878–9883.

[Calhoun et al., 2002] Calhoun, V. C., Stathopoulos, A. and Levine, M. (2002). Promoter-proximal tethering elements regulate enhancer-promoter specificity in the Drosophila Antennapedia complex. Proceedings of the National Academy of Sciences *99*, 9243–9247.

[Carmack et al., 2007] Carmack, C. S., McCue, L. A., Newberg, L. A. and Lawrence, C. E. (2007). PhyloScan: identification of transcription factor binding sites using cross-species evidence. Algorithms for Molecular Biology *2*, 1.

[Cartharius et al., 2005] Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M. and Werner, T. (2005). MatInspector and Beyond: Promoter Analysis Based on Transcription Factor Binding Sites. Bioinformatics *21*, 2933–2942.

[Cavalli, 2002] Cavalli, G. (2002). Chromatin as a eukaryotic template of genetic information. Current Opinion in Cell Biology *14*, 269–278.

[Cawley et al., 2004] Cawley, S., Bekiranov, S., Ng, H. H., Kapranov, P., Sekinger, E. A., Kampa, D., Piccolboni, A., Sementchenko, V., Cheng, J., Williams, A. J. et al. (2004). Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell *116*, 499–509.

[Celniker et al., 2009] Celniker, S. E., Dillon, L. A. L., Gerstein, M. B., Gunsalus, K. C., Henikoff, S., Karpen, G. H., Kellis, M., Lai, E. C., Lieb, J. D., MacAlpine, D. M., Micklem, G., Piano, F., Snyder, M., Stein, L., White, K. P. and Waterston, R. H. (2009). Unlocking the secrets of the genome. Nature *459*, 927–930.

[Chen et al., 2008] Chen, C., Schmidt, B., Weiguo, L. and Müller-Wittig, W. (2008). GPU-MEME: Using graphics hardware to accelerate motif finding in DNA sequences. Pattern Recognition in Bioinformatics *5265*, 448–459.

[Chen et al., 2007] Chen, G., Jensen, S. and Stoeckert, C. (2007). Clustering of genes into regulons using integrated modeling-COGRIM. Genome biology *8*, R4.

[Chen et al., 1995] Chen, Q. K., Hertz, G. Z. and Stormo, G. D. (1995). MATRIX SEARCH 1.0: A Computer Program That Scans DNA Sequences for Transcriptional Elements Using a Database of Weight Matrices. Computer applications in the biosciences : CABIOS *11*, 563–566.

[Chen et al., 2008] Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., Loh, Y.-H., Yeo, H. C., Yeo, Z. X., Narang, V., Govindarajan, K. R., Leong, B., Shahab, A., Ruan, Y., Bourque, G., Sung, W.-K., Clarke, N. D., Wei, C.-L. and Ng, H.-H. (2008). Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. Cell *133*, 1106–1117.

[Chor et al., 2009] Chor, B., Horn, D., Goldman, N., Levy, Y. and Massingham, T. (2009). Genomic DNA k-mer spectra: models and modalities. Genome Biology *10*, R108.

[Choy et al., 2010] Choy, M.-K., Movassagh, M., Goh, H.-G., Bennett, M., Down, T. and Foo, R. (2010). Genome-wide conserved consensus transcription factor binding motifs are hyper-methylated. BMC Genomics *11*, 519.

[Claverie and Audic, 1996] Claverie, J.-M. and Audic, S. (1996). The Statistical Significance of Nucleotide Position-Weight Matrix Matches. Computer applications in the biosciences : CABIOS *12*, 431–439.

[Consortium, 2004] Consortium, T. E. P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636–640.

[Contrino et al., 2012] Contrino, S., Smith, R. N., Butano, D., Carr, A., Hu, F., Lyne, R., Rutherford, K., Kalderimis, A., Sullivan, J., Carbon, S., Kephart, E. T., Lloyd, P., Stinson, E. O., Washington, N. L., Perry, M. D., Ruzanov, P., Zha, Z., Lewis, S. E., Stein, L. D. and Micklem, G. (2012). modMine: flexible access to modENCODE data. Nucleic Acids Research *40*, D1082–D1088.

[Crawford et al., 2006] Crawford, G. E., Davis, S., Scacheri, P. C., Renaud, G., Halawi, M. J., Erdos, M. R., Green, R., Meltzer, P. S., Wolfsberg, T. G. and Collins, F. S. (2006). DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. Nature methods *3*, 503–509.

[Creyghton et al., 2010] Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A., Boyer, L. A., Young, R. A. and Jaenisch, R. (2010). Histone H3K27ac separates

active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences of the United States of America *107*, 21931–21936.

[Cuellar-Partida et al., 2012] Cuellar-Partida, G., Buske, F., McLeay, R., Whitington, T., Noble, W. and Bailey, T. (2012). Epigenetic priors for identifying active transcription factor binding sites. Bioinformatics *28*, 56–62.

[Cui et al., 2009] Cui, K., Zang, C., Roh, T.-Y., Schones, D. E., Childs, R. W., Peng, W. and Zhao, K. (2009). Chromatin Signatures in Multipotent Human Hematopoietic Stem Cells Indicate the Fate of Bivalent Genes during Differentiation. Cell Stem Cell *4*, 80–93.

[Das and Dai, 2007] Das, M. and Dai, H. (2007). A survey of DNA motif finding algorithms. BMC bioinformatics *8*, S21.

[Davidson, 2006] Davidson, E. H. (2006). Gene Regulatory Networks and the Evolution of Animal Body Plans. Science *311*, 796–800.

[De Finetti, 1937] De Finetti, B. (1937). La Prévision: Ses Lois Logiques, Ses Sources Subjectives. Annales de l'Institut Henri Poincaré *17*.

[De Finetti, 1964] De Finetti, B. (1964). Foresight: Its Logical Laws in Subjective Sources. In Studies in Subjective Probability, (Kyburg, H. and Smokler, H., eds), pp. 93–158. Wiley.

[Dekker et al., 2002] Dekker, J., Rippe, K., Dekker, M. and Kleckner, N. (2002). Capturing Chromosome Conformation. Science *295*, 1306–1311.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B *39*, 1–38.

[Desjarlais and Berg, 1994] Desjarlais, J. R. and Berg, J. M. (1994). Length-encoded multiplex binding site determination: application to zinc finger proteins. Proceedings of the National Academy of Sciences *91*, 11099–11103.

[D'haeseleer, 2006a] D'haeseleer, P. (2006a). How does DNA sequence motif discovery work? Nature biotechnology *24*, 959–961.

[D'haeseleer, 2006b] D'haeseleer, P. (2006b). What are DNA sequence motifs? Nature biotechnology *24*, 423–425.

[Dickson et al., 1975] Dickson, R. C., Abelson, J., Barnes, W. M. and Reznikoff, W. S. (1975). Genetic regulation: the Lac control region. Science (New York, N.Y.) *187*, 27–35.

[Dominguez et al., 2003] Dominguez, C., Boelens, R. and Bonvin, A. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical data. NMR-based docking of protein-protein complexes *125*, 51.

[Döring et al., 2008] Döring, A., Weese, D., Rausch, T. and Reinert, K. (2008). SeqAn An efficient, generic C++ library for sequence analysis. BMC Bioinformatics *9*, 11.

[Dorschner et al., 2004] Dorschner, M. O., Hawrylycz, M., Humbert, R., Wallace, J. C., Shafer, A., Kawamoto, J., Mack, J., Hall, R., Goldy, J., Sabo, P. J., Kohli, A., Li, Q., McArthur, M. and Stamatoyannopoulos, J. A. (2004). High-throughput localization of functional elements by quantitative chromatin profiling. Nature Methods *1*, 219–225.

[Dostie and Dekker, 2007] Dostie, J. and Dekker, J. (2007). Mapping networks of physical interactions between genomic elements using 5C technology. Nature Protocols *2*, 988–1002.

[Down, 2005] Down, T. A. (2005). NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. Nucleic Acids Research *33*, 1445–1453.

[Dunham et al., 2012] Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B.-K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., Shoresh, N., Simon, J. M., Song, L., Trinklein, N. D., Altshuler, R. C., Brown, J. B., Cheng, C., Djebali, S., Dong, X., Ernst, J., Furey, T. S., Giardine, B., Greven, M., Harris, R. S., Herrero, J., Hoffman, M. M., Iyer, S., Kheradpour, P., Lassmann, T., Li, Q., Lin, X., Marinov, G. K., Merkel, A., Mortazavi, A., Parker, S. C. J., Reddy, T. E., Rozowsky, J., Schlesinger, F., Thurman, R. E., Wang, J., Ward, L. D., Whitfield, T. W., Wilder, S. P., Wu, W., Xi, H. S., Yip, K. Y., Zhuang, J., Gunter, C., Pazin, M. J., Lowdon, R. F., Dillon, L. A. L., Adams, L. B., Kelly, C. J., Zhang, J., Wexler, J. R., Good, P. J., Feingold, E. A., Yu, Y., Wrobel, J., Risk, B. A., Gunawardena, H. P., Kuiper, H. C., Maier, C. W., Xie, L., Chen, X., Mikkelsen, T. S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R., Coyne, M. J., Durham, T., Ku, M., Truong, T., Eaton, M. L., Dobin, A., Tanzer, A., Lagarde, J., Lin, W., Xue, C., Williams, B. A., Zaleski, C., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Batut, P., Bell, I., Bell, K., Chakrabortty, S., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S.,

Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Li, G., Luo, O. J., Park, E., Preall, J. B., Presaud, K., Ribeca, P., Robyr, D., Ruan, X., Sammeth, M., Singh Sandhu, K., Schaeffer, L., See, L.-H., Shahab, A., Skancke, J., Maria Suzuki, A., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Hayashizaki, Y., Reymond, A., Antonarakis, S. E., Hannon, G. J., Ruan, Y., Carninci, P., Sloan, C. A., Learned, K., Malladi, V. S., Wong, M. C., Barber, G. P., Cline, M. S., Dreszer, T. R., Heitner, S. G., Karolchik, D., Kirkup, V. M., Meyer, L. R., Long, J. C., Maddren, M., Raney, B. J., Grasfeder, L. L., Giresi, P. G., Battenhouse, A., Sheffield, N. C., Showers, K. A., London, D., Bhinge, A. A., Shestak, C., Schaner, M. R., Ki Kim, S., Zhang, Z. Z., Mieczkowski, P. A., Mieczkowska, J. O., Liu, Z., McDaniell, R. M., Ni, Y., Rashid, N. U., Jae Kim, M., Adar, S., Zhang, Z., Wang, T., Winter, D., Keefe, D., Iyer, V. R., Zheng, M., Wang, P., Gertz, J., Vielmetter, J., Christopher Partridge, E., Varley, K. E., Gasper, C., Bansal, A., Pepke, S., Jain, P., Amrhein, H., Bowling, K. M., Anaya, M., Cross, M. K., Muratet, M. A., Davis, N. S., McCue, K., Eggleston, T., Fisher-Aylor, K. I., DeSalvo, G., Meadows, S. K., Balasubramanian, S., Nesmith, A. S., Scott Newberry, J., Newberry, K. M., Parker, S. L., Pusey, B., Levy, S. E., Absher, D. M., Wong, W. H., Blow, M. J., Visel, A., Pennachio, L. A., Petrykowska, H. M., Abyzov, A., Aken, B., Barrell, D., Barson, G., Berry, A., Bignell, A., Boychenko, V., Bussotti, G., Davidson, C., Despacio-Reyes, G., Diekhans, M., Ezkurdia, I., Frankish, A., Gilbert, J., Manuel Gonzalez, J., Griffiths, E., Harte, R., Hendrix, D. A., Hunt, T., Jungreis, I., Kay, M., Khurana, E., Leng, J., Lin, M. F., Loveland, J., Lu, Z., Manthravadi, D., Mariotti, M., Mudge, J., Mukherjee, G., Notredame, C., Pei, B., Manuel Rodriguez, J., Saunders, G., Sboner, A., Searle, S., Sisu, C., Snow, C., Steward, C., Tapanari, E., Tress, M. L., van Baren, M. J., Washietl, S., Wilming, L., Zadissa, A., Zhang, Z., Brent, M., Haussler, D., Valencia, A., Addleman, N., Alexander, R. P., Auerbach, R. K., Balasubramanian, S., Bettinger, K., Bhardwaj, N., Boyle, A. P., Cao, A. R., Cayting, P., Charos, A., Cheng, Y., Eastman, C., Euskirchen, G., Fleming, J. D., Grubert, F., Habegger, L., Hariharan, M., Harmanci, A., Iyengar, S., Jin, V. X., Karczewski, K. J., Kasowski, M., Lacroute, P., Lam, H., Lamarre-Vincent, N., Lian, J., Lindahl-Allen, M., Min, R., Miotto, B., Monahan, H., Moqtaderi, Z., Mu, X. J., O'Geen, H., Ouyang, Z., Patacsil, D., Raha, D., Ramirez, L., Reed, B., Shi, M., Slifer, T., Witt, H., Wu, L., Xu, X., Yan, K.-K., Yang, X., Struhl, K., Weissman, S. M., Penalva, L. O., Karmakar, S., Bhanvadia, R. R., Choudhury, A., Domanus, M., Ma, L., Moran, J., Victorsen, A., Auer, T., Centanin, L., Eichenlaub, M., Gruhl, F., Heermann, S., Hoeckendorf, B., Inoue, D., Kellner, T., Kirchmaier, S., Mueller, C., Reinhardt, R., Schertel, L., Schneider, S., Sinn, R., Wittbrodt, J., Wittbrodt, B., Jain, G., Balasundaram, G., Bates, D. L., Byron, R., Canfield, T. K., Diegel,

M. J., Dunn, D., Ebersol, A. K., Frum, T., Garg, K., Gist, E., Scott Hansen, R., Boatman, L., Haugen, E., Humbert, R., Johnson, A. K., Johnson, E. M., Kutyavin, T. V., Lee, K., Lotakis, D., Maurano, M. T., Neph, S. J., Neri, F. V., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Rynes, E., Sanchez, M. E., Sandstrom, R. S., Shafer, A. O., Stergachis, A. B., Thomas, S., Vernot, B., Vierstra, J., Vong, S., Wang, H., Weaver, M. A., Yan, Y., Zhang, M., Akey, J. M., Bender, M., Dorschner, M. O., Groudine, M., MacCoss, M. J., Navas, P., Stamatoyannopoulos, G., Beal, K., Brazma, A., Flicek, P., Johnson, N., Lukk, M., Luscombe, N. M., Sobral, D., Vaquerizas, J. M., Batzoglou, S., Sidow, A., Hussami, N., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M. W., Schaub, M. A., Miller, W., Bickel, P. J., Banfai, B., Boley, N. P., Huang, H., Jessica Li, J., Stafford Noble, W., Bilmes, J. A., Buske, O. J., Sahu, A. D., Kharchenko, P. V., Park, P. J., Baker, D., Taylor, J., Lochovsky, L., Bernstein, B. E., Crawford, G. E., Dekker, J., Elnitski, L., Farnham, P. J., Gerstein, M., Giddings, M. C., Gingeras, T. R., Green, E. D., Guigó, R., Hardison, R. C., Hubbard, T. J., Kellis, M., James Kent, W., Lieb, J. D., Margulies, E. H., Myers, R. M., Snyder, M., Stamatoyannopoulos, J. A., Tenenbaum, S. A., Weng, Z., White, K. P., Wold, B. and Birney, E. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

[Dynan and Tjian, 1983] Dynan, W. S. and Tjian, R. (1983). The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. Cell *35*, 79–87.

[Elkon, 2003] Elkon, R. (2003). Genome-Wide In Silico Identification of Transcriptional Regulators Controlling the Cell Cycle in Human Cells. Genome Research *13*, 773–780.

[Ellington and Szostak, 1990] Ellington, A. D. and Szostak, J. W. (1990). In vitro selection of RNA molecules that bind specific ligands. Nature *346*, 818–822.

[Ernst et al., 2010] Ernst, J., Plasterer, H. L., Simon, I. and Bar-Joseph, Z. (2010). Integrating multiple evidence sources to predict transcription factor binding in the human genome. Genome Research *20*, 526–536.

[Ettwiller et al., 2007] Ettwiller, L., Paten, B., Ramialison, M., Birney, E. and Wittbrodt, J. (2007). Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. Nature Methods *4*, 563–565.

[Federico and Pisanti, 2009] Federico, M. and Pisanti, N. (2009). Suffix tree characterization of maximal motifs in biological sequences. Theoretical Computer Science *410*, 4391–4401.

[Felsenstein, 1981] Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of molecular evolution *17*, 368–376.

[Felsner et al., 1997] Felsner, S., Müller, R. and Wernisch, L. (1997). Trapezoid graphs and generalizations, geometry and algorithms. Discrete Applied Mathematics *74*, 13–32.

[Fordyce et al., 2010] Fordyce, P. M., Gerber, D., Tran, D., Zheng, J., Li, H., DeRisi, J. L. and Quake, S. R. (2010). De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. Nature Biotechnology *28*, 970–975.

[Fordyce et al., 2012] Fordyce, P. M., Pincus, D., Kimmig, P., Nelson, C. S., El-Samad, H., Walter, P. and DeRisi, J. L. (2012). Basic leucine zipper transcription factor Hac1 binds DNA in two distinct modes as revealed by microfluidic analyses. Proceedings of the National Academy of Sciences of the United States of America *109*, E3084–E3093.

[Fried and Crothers, 1981] Fried, M. and Crothers, D. M. (1981). Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. Nucleic Acids Research *9*, 6505–6525.

[Frith et al., 2001] Frith, M. C., Hansen, U. and Weng, Z. (2001). Detection of cis -element clusters in higher eukaryotic DNA. Bioinformatics *17*, 878–889.

[Frith et al., 2003] Frith, M. C., Li, M. C. and Weng, Z. (2003). Cluster-Buster: finding dense clusters of motifs in DNA sequences. Nucleic Acids Research *31*, 3666–3668.

[Galas and Schmitz, 1978] Galas, D. J. and Schmitz, A. (1978). DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Research *5*, 3157–3170.

[Garner and Revzin, 1981] Garner, M. M. and Revzin, A. (1981). A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. Nucleic Acids Research *9*, 3047–3060.

[Geman and Geman, 1984] Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Trans. Pattern Anal. Mach. Intell. *6*, 721–741.

[Georgi and Schliep, 2006] Georgi, B. and Schliep, A. (2006). Context-specific independence mixture modeling for positional weight matrices. Bioinformatics *22*, e166–e173.

[Gerber et al., 2007] Gerber, G. K., Dowell, R. D., Jaakkola, T. S. and Gifford, D. K. (2007). Automated Discovery of Functional Generality of Human Gene Expression Programs. PLoS Computational Biology *3*, e148.

[Gertz et al., 2008] Gertz, J., Siggia, E. D. and Cohen, B. A. (2008). Analysis of combinatorial cis-regulation in synthetic and genomic promoters. Nature *457*, 215–218.

[Giegerich and Kurtz, 1997] Giegerich, R. and Kurtz, S. (1997). From Ukkonen to McCreight and Weiner: A unifying view of linear-time suffix tree construction. Algorithmica *19*, 331–353.

[Gilbert and Maxam, 1973] Gilbert, W. and Maxam, A. (1973). The Nucleotide Sequence of the lac Operator. Proceedings of the National Academy of Sciences of the United States of America *70*, 3581–3584.

[Gilchrist et al., 2009] Gilchrist, D. A., Fargo, D. C. and Adelman, K. (2009). Using ChIP-chip and ChIP-seq to study the regulation of gene expression: Genome-wide localization studies reveal widespread regulation of transcription elongation. Methods *48*, 398–408.

[Gindhart et al., 1995] Gindhart, J. G., King, A. N. and Kaufman, T. C. (1995). Characterization of the cis-regulatory region of the Drosophila homeotic gene Sex combs reduced. Genetics *139*, 781–795.

[Giresi et al., 2007] Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R. and Lieb, J. D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Research *17*, 877–885.

[Granier et al., 2011] Granier, C., Gurchenkov, V., Perea-Gomez, A., Camus, A., Ott, S., Papanayotou, C., Iranzo, J., Moreau, A., Reid, J., Koentges, G., Sabéran-Djoneidi, D. and Collignon, J. (2011). Nodal cis-regulatory elements reveal epiblast and primitive endoderm heterogeneity in the peri-implantation mouse embryo. Developmental Biology *349*, 350–362.

[Grant et al., 2011] Grant, C. E., Bailey, T. L. and Noble, W. S. (2011). FIMO: Scanning for Occurrences of a Given Motif. Bioinformatics *27*, 1017–1018.

[Gruber and Gross, 2003] Gruber, T. M. and Gross, C. A. (2003). Multiple sigma subunits and the partitioning of bacterial transcription space. Annual Review of Microbiology *57*, 441–466.

[Grundy et al., 1996] Grundy, W. N., Bailey, T. L. and Elkan, C. P. (1996). ParaMEME: A parallel implementation and a web interface for a DNA and protein motif discovery tool. Computer applications in the biosciences: CABIOS *12*, 303–310.

[Gusfield, 1997] Gusfield, D. (1997). Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology. Cambridge University Press.

[Halpern and Bruno, 1998] Halpern, A. L. and Bruno, W. J. (1998). Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Molecular biology and evolution *15*, 910–917.

[Håndstad et al., 2011] Håndstad, T., Rye, M. B., Drabløs, F. and Sætrom, P. (2011). A ChIP-Seq Benchmark Shows That Sequence Conservation Mainly Improves Detection of Strong Transcription Factor Binding Sites. PLoS ONE *6*, e18430.

[Hannenhalli and Wang, 2005] Hannenhalli, S. and Wang, L.-S. (2005). Enhanced position weight matrices using mixture models. Bioinformatics *21*, i204–i212.

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

[Hawkins et al., 2009] Hawkins, J., Grant, C., Noble, W. S. and Bailey, T. L. (2009). Assessing phylogenetic motif models for predicting transcription factor binding sites. Bioinformatics *25*, i339–i347.

[He et al., 2011] He, Q., Bardet, A. F., Patton, B., Purvis, J., Johnston, J., Paulson, A., Gogol, M., Stark, A. and Zeitlinger, J. (2011). High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species. Nature Genetics *43*, 414–420.

[He et al., 2010] He, X., Samee, M. A. H., Blatti, C. and Sinha, S. (2010). Thermodynamics-Based Models of Transcriptional Regulation by Enhancers: The Roles of Synergistic Activation, Cooperative Binding and Short-Range Repression. PLoS Computational Biology *6*, e1000935.

[Heintzman et al., 2009] Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., Ching, K. A., Antosiewicz-Bourget, J. E., Liu, H., Zhang, X., Green, R. D., Lobanenkov, V. V., Stewart, R., Thomson, J. A., Crawford, G. E., Kellis, M. and Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112.

[Heintzman et al., 2007] Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A., Wang, W., Weng, Z., Green, R. D., Crawford, G. E. and Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature Genetics *39*, 311–318.

[Hesselberth et al., 2009] Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., Thurman, R. E., Neph, S., Kuehn, M. S., Noble, W. S., Fields, S. and Stamatoyannopoulos, J. A. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nature Methods *6*, 283–289.

[Hinton and van Camp, 1993] Hinton, G. E. and van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In Proceedings of the sixth annual conference on Computational learning theory COLT '93 p. 5–13, ACM, New York, NY, USA.

[Ho Sui et al., 2007] Ho Sui, S. J., Fulton, D. L., Arenillas, D. J., Kwon, A. T. and Wasserman, W. W. (2007). oPOSSUM: integrated tools for analysis of regulatory motif over-representation. Nucleic Acids Research *35*, W245–W252.

[Hochschild and Ptashne, 1986] Hochschild, A. and Ptashne, M. (1986). Cooperative binding of lambda repressors to sites separated by integral turns of the DNA helix. Cell *44*, 681–687.

[Hu et al., 2005] Hu, J., Li, B. and Kihara, D. (2005). Limitations and potentials of current motif discovery algorithms. Nucleic Acids Research *33*, 4899–4913.

[Hubbard et al., 2007] Hubbard, T. J. P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S. C., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Herrero, J., Holland, R., Howe, K., Howe, K., Johnson, N., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Melsopp, C., Megy, K., Meidl, P., Ouverdin, B., Parker, A., Prlic, A., Rice, S., Rios, D., Schuster, M., Sealy, I., Severin, J., Slater, G., Smedley, D., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wood, M., Cox, T., Curwen, V., Durbin, R., Fernandez-Suarez, X. M., Flicek, P., Kasprzyk, A., Proctor, G., Searle, S., Smith, J., Ureta-Vidal, A. and Birney, E. (2007). Ensembl 2007. Nucleic Acids Research *35*, D610–D617.

[Ingham, 1988] Ingham, P. W. (1988). The molecular genetics of embryonic pattern formation in Drosophila. Nature *335*, 25–34.

[Jaakkola, 1997] Jaakkola, T. S. (1997). Variational methods for inference and estimation in graphical models. Thesis Massachusetts Institute of Technology. Thesis (Ph. D.)–Massachusetts Institute of Technology, Dept. of Brain and Cognitive Sciences, 1997.

[Jaynes, 2003] Jaynes, E. T. (2003). Probability Theory: The Logic of Science. Cambridge University Press.

[Jeffreys, 1935] Jeffreys, H. (1935). Some Tests of Significance, Treated by the Theory of Probability. Mathematical Proceedings of the Cambridge Philosophical Society *31*, 203–222.

[Jeffreys, 1998] Jeffreys, H. (1998). Theory of Probability, Third Edition. Oxford University Press.

[Jensen et al., 2007] Jensen, S. T., Chen, G. and Stoeckert, Jr., C. J. (2007). Bayesian variable selection and data integration for biological regulatory networks. Annals of Applied Statistics *1*, 612–633.

[Ji et al., 2008] Ji, H., Jiang, H., Ma, W., Johnson, D. S., Myers, R. M. and Wong, W. H. (2008). An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nature Biotechnology *26*, 1293–1300.

[Johnson et al., 2007] Johnson, D. S., Mortazavi, A., Myers, R. M. and Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science *316*, 1497–1502.

[Johnston and Nüsslein-Volhard, 1992] Johnston, D. S. and Nüsslein-Volhard, C. (1992). The origin of pattern and polarity in the Drosophila embryo. Cell *68*, 201–219.

[Jordan, 2004] Jordan, M. I. (2004). Graphical models. Statistical Science *19*, 140–155.

[Jordan et al., 1999] Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. and Saul, L. K. (1999). An introduction to variational methods for graphical models. Machine learning *37*, 183–233.

[Joshi et al., 2005] Joshi, B., Ordonez-Ercan, D., Dasgupta, P. and Chellappan, S. (2005). Induction of human metallothionein 1G promoter by VEGF and heavy metals: differential involvement of E2F and metal transcription factors. Oncogene *24*, 2204–2217.

[Jothi et al., 2008] Jothi, R., Cuddapah, S., Barski, A., Cui, K. and Zhao, K. (2008). Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Research *36*, 5221–5231.

[Jukes and Cantor, 1969] Jukes, T. and Cantor, C. (1969). Evolution of Protein Molecules. In Mammalian Protein Metabolism, (Munro, H., ed.), pp. 121–132. Academic Press New York.

[Kanehisa, 2006] Kanehisa, M. (2006). From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Research *34*, D354–D357.

[Kass and Raftery, 1995] Kass, R. E. and Raftery, A. E. (1995). Bayes Factors. Journal of the American Statistical Association *90*, 773.

[Keene et al., 1981] Keene, M. A., Corces, V., Lowenhaupt, K. and Elgin, S. C. (1981). DNase I hypersensitive sites in Drosophila chromatin occur at the 5' ends of regions of transcription. Proceedings of the National Academy of Sciences of the United States of America *78*, 143–146.

[Kel et al., 2003] Kel, A., Gößling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. and Wingender, E. (2003). MATCHTM: a tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Research *31*, 3576–3579.

[Kel et al., 1999] Kel, A., Kel-Margoulis, O., Babenko, V. and Wingender, E. (1999). Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells. Journal of Molecular Biology *288*, 353–376.

[Kellis et al., 2003] Kellis, M., Patterson, N., Endrizzi, M., Birren, B., Lander, E. et al. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature *423*, 241–254.

[Kent et al., 2002] Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. and Haussler, D. (2002). The Human Genome Browser at UCSC. Genome Research *12*, 996–1006.

[Kheradpour et al., 2007] Kheradpour, P., Stark, A., Roy, S. and Kellis, M. (2007). Reliable prediction of regulator targets using 12 Drosophila genomes. Genome research *17*, 1919–1931.

[Kondoh et al., 2004] Kondoh, H., Uchikawa, M. and Kamachi, Y. (2004). Interplay of Pax6 and SOX2 in lens development as a paradigm of genetic switch mechanisms for cell differentiation. The International Journal of Developmental Biology *48*, 819–827.

[Koohy et al., 2010] Koohy, H., Dyer, N., Reid, J., Koentges, G. and Ott, S. (2010). An alignment-free model for comparison of regulatory sequences. Bioinformatics *26*, 2391.

[Kouzarides, 2007] Kouzarides, T. (2007). Chromatin modifications and their function. Cell *128*, 693.

[Kreiman, 2004] Kreiman, G. (2004). Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes. Nucleic Acids Research *32*, 2889–2900.

[Kulakovskiy et al., 2010] Kulakovskiy, I. V., Boeva, V. A., Favorov, A. V. and Makeev, V. J. (2010). Deep and wide digging for binding motifs in ChIP-Seq data. Bioinformatics *26*, 2622–2623.

[Kulakovskiy and Makeev, 2010] Kulakovskiy, I. V. and Makeev, V. J. (2010). Discovery of DNA motifs recognized by transcription factors through integration of different experimental sources. Biophysics *54*, 667–674.

[Kullback, 1959] Kullback, S. (1959). Information theory and statistics. Wiley.

[Kullback and Leibler, 1951] Kullback, S. and Leibler, R. A. (1951). On Information and Sufficiency. The Annals of Mathematical Statistics *22*, 79–86.

[Kunarso et al., 2010] Kunarso, G., Chia, N. Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y. S., Ng, H. H. and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nature genetics *42*, 631–634.

[Lavine and Schervish, 1999] Lavine, M. and Schervish, M. J. (1999). Bayes Factors: What They Are and What They Are Not. The American Statistician *53*, 119.

[Lawrence and Reilly, 1990] Lawrence, C. E. and Reilly, A. A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins *7*, 41–51.

[Lee et al., 2002] Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J.-B., Volkert, T. L., Fraenkel, E., Gifford, D. K. and Young, R. A. (2002). Transcriptional Regulatory Networks in Saccharomyces cerevisiae. Science *298*, 799–804.

[Lemmens et al., 2006] Lemmens, K., Dhollander, T., De Bie, T., Monsieurs, P., Engelen, K., Smets, B., Winderickx, J., De Moor, B. and Marchal, K. (2006). Inferring transcriptional modules from ChIP-chip, motif and microarray data. Genome biology *7*, R37.

[Levine, 2010] Levine, M. (2010). Transcriptional enhancers in animal development and evolution. Current biology *20*, R754–R763.

[Li, 2009] Li, L. (2009). gadem: A Genetic Algorithm Guided Formation of Spaced Dyads Coupled with an EM Algorithm for Motif Discovery. Journal of Computational Biology *16*, 317–329.

[Liao et al., 2008] Liao, W., Schones, D. E., Oh, J., Cui, Y., Cui, K., Tae-Young, R., Zhao, K. and Leonard, W. J. (2008). Priming for T helper type 2 differentiation by interleukin 2-mediated induction of IL-4 receptor α chain expression. Nature immunology *9*, 1288–1296.

[Lichtlen et al., 2001] Lichtlen, P., Wang, Y., Belser, T., Georgiev, O., Certa, U., Sack, R. and Schaffner, W. (2001). Target gene search for the metal-responsive transcription factor MTF-1. Nucleic Acids Research *29*, 1514–1523.

[Linhart et al., 2008] Linhart, C., Halperin, Y. and Shamir, R. (2008). Transcription factor and microRNA motif discovery: The Amadeus platform and a compendium of metazoan target sets. Genome Research *18*, 1180–1189.

[Liu et al., 2007] Liu, X., Jessen, W. J., Sivaganesan, S., Aronow, B. J. and Medvedovic, M. (2007). Bayesian hierarchical model for transcriptional module discovery by jointly modeling gene expression and ChIP-chip data. BMC Bioinformatics *8*, 283.

[Liu and Meyer, 2009] Liu, X. S. and Meyer, C. A. (2009). ChIP-Chip: Algorithms for Calling Binding Sites. In Microarray Analysis of the Physical Genome, (Pollack, J. R., ed.), number 556 in Methods in Molecular Biology$^{\text{TM}}$ pp. 165–175. Humana Press.

[Loh et al., 2006] Loh, Y.-H., Wu, Q., Chew, J.-L., Vega, V. B., Zhang, W., Chen, X., Bourque, G., George, J., Leong, B., Liu, J., Wong, K.-Y., Sung, K. W., Lee, C. W. H., Zhao, X.-D., Chiu, K.-P., Lipovich, L., Kuznetsov, V. A., Robson, P., Stanton, L. W., Wei, C.-L., Ruan, Y., Lim, B. and Ng, H.-H. (2006). The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. Nature Genetics *38*, 431–440.

[Lustig and Jernigan, 1995] Lustig, B. and Jernigan, R. L. (1995). Consistencies of individual DNA base-amino acid interactions in structures and sequences. Nucleic Acids Research *23*, 4707–4711.

[MacIsaac and Fraenkel, 2006] MacIsaac, K. D. and Fraenkel, E. (2006). Practical strategies for discovering regulatory DNA sequence motifs. PLoS computational biology *2*, e36.

[Maizels, 1973] Maizels, N. M. (1973). The Nucleotide Sequence of the Lactose Messenger Ribonucleic Acid Transcribed from the UV5 Promoter Mutant of Escherichia coli. Proceedings of the National Academy of Sciences of the United States of America *70*, 3585–3589.

[Mäkinen et al., 2010] Mäkinen, V., Välimäki, N., Laaksonen, A. and Katainen, R. (2010). Unified View of Backward Backtracking in Short Read Mapping. In Algorithms and Applications, (Elomaa, T., Mannila, H. and Orponen, P., eds), number 6060 in Lecture Notes in Computer Science pp. 182–195. Springer Berlin Heidelberg.

[Malik and Roeder, 2005] Malik, S. and Roeder, R. G. (2005). Dynamic regulation of pol II transcription by the mammalian Mediator complex. Trends in Biochemical Sciences *30*, 256–263.

[Man and Stormo, 2001] Man, T.-K. and Stormo, G. D. (2001). Non-independence of Mnt repressor–operator interaction determined by a new quantitative multiple fluorescence relative affinity (QuMFRA) assay. Nucleic Acids Research *29*, 2471–2478.

[Man et al., 2004] Man, T.-K., Yang, J. S. and Stormo, G. D. (2004). Quantitative modeling of DNA–protein interactions: effects of amino acid substitutions on binding specificity of the Mnt repressor. Nucleic Acids Research *32*, 4026–4032.

[Maniatis et al., 1975] Maniatis, T., Jeffrey, A. and Kleid, D. G. (1975). Nucleotide sequence of the rightward operator of phage lambda. Proceedings of the National Academy of Sciences of the United States of America *72*, 1184–1188.

[Mansour et al., 2011] Mansour, E., Allam, A., Skiadopoulos, S. and Kalnis, P. (2011). ERA: efficient serial and parallel suffix tree construction for very long strings. Proceedings of the VLDB Endowment *5*, 49–60.

[Marsan and Sagot, 2000] Marsan, L. and Sagot, M. (2000). Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. Journal of Computational Biology *7*, 345–362.

[Matys et al., 2003] Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A. E., Kel-Margoulis, O. V., Kloos, D.-U., Land, S., Lewicki-Potapov, B., Michael, H., Münch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. and Wingender, E. (2003). TRANSFAC®: transcriptional regulation, from patterns to profiles. Nucleic Acids Research *31*, 374–378.

[May et al., 2011] May, D., Blow, M. J., Kaplan, T., McCulley, D. J., Jensen, B. C., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Afzal, V., Simpson, P. C., Rubin, E. M., Black, B. L., Bristow, J., Pennacchio, L. A. and Visel, A. (2011). Large-scale discovery of enhancers from human heart tissue. Nature Genetics *44*, 89–93.

[McArthur et al., 2001] McArthur, M., Gerum, S. and Stamatoyannopoulos, G. (2001). Quantification of DNaseI-sensitivity by Real-time PCR: Quantitative Analysis of DNaseI-hypersensitivity of the Mouse β-Globin LCR. Journal of molecular biology *313*, 27.

[McCreight, 1976] McCreight, E. M. (1976). A space-economical suffix tree construction algorithm. Journal of the ACM (JACM) *23*, 262–272.

[McLeay et al., 2012] McLeay, R. C., Lesluyes, T., Partida, G. C. and Bailey, T. L. (2012). Genome-wide in silico prediction of gene expression. Bioinformatics *28*, 2789–2796.

[Merika and Thanos, 2001] Merika, M. and Thanos, D. (2001). Enhanceosomes. Current opinion in genetics & development *11*, 205–208.

[Mikkelsen et al., 2010] Mikkelsen, T. S., Xu, Z., Zhang, X., Wang, L., Gimble, J. M., Lander, E. S. and Rosen, E. D. (2010). Comparative Epigenomic Analysis of Murine and Human Adipogenesis. Cell *143*, 156–169.

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

[Moses et al., 2004a] Moses, A. M., Chiang, D. Y. and Eisen, M. B. (2004a). Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing *vol*, 324–335.

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

[Moses et al., 2006] Moses, A. M., Pollard, D. A., Nix, D. A., Iyer, V. N., Li, X.-Y., Biggin, M. D. and Eisen, M. B. (2006). Large-Scale Turnover of Functional Transcription Factor Binding Sites in Drosophila. PLoS Computational Biology *2*, e130.

[Mujtaba et al., 2007] Mujtaba, S., Zeng, L. and Zhou, M.-M. (2007). Structure and acetyl-lysine recognition of the bromodomain. Oncogene *26*, 5521–5527.

[Mukherjee et al., 2004] Mukherjee, S., Berger, M. F., Jona, G., Wang, X. S., Muzzey, D., Snyder, M., Young, R. A. and Bulyk, M. L. (2004). Rapid Analysis of the DNA Binding Specificities of Transcription Factors with DNA Microarrays. Nature genetics *36*, 1331–1339.

[Nagarajan et al., 2005] Nagarajan, N., Jones, N. and Keich, U. (2005). Computing the P-value of the information content from an alignment of multiple sequences. Bioinformatics *21*, i311–i318.

[Nakajima et al., 2010] Nakajima, A., Isshiki, T., Kaneko, K. and Ishihara, S. (2010). Robustness under Functional Constraint: The Genetic Network for Temporal Expression in Drosophila Neurogenesis. PLoS Computational Biology *6*, e1000760.

[Narlikar et al., 2013] Narlikar, L., Mehta, N., Galande, S. and Arjunwadkar, M. (2013). One size does not fit all: On how Markov model order dictates performance of genomic sequence analyses. Nucleic Acids Research *41*, 1416–1424.

[Naughton et al., 2006] Naughton, B. T., Fratkin, E., Batzoglou, S. and Brutlag, D. L. (2006). A graph-based motif detection algorithm models complex nucleotide dependencies in transcription factor binding sites. Nucleic acids research *34*, 5730–5739.

[Neal and Hinton, 1998] Neal, R. and Hinton, G. E. (1998). A View Of The Em Algorithm That Justifies Incremental, Sparse, And Other Variants. In Learning in Graphical Models p. 355–368, Kluwer Academic Publishers.

[Newburger and Bulyk, 2009] Newburger, D. E. and Bulyk, M. L. (2009). UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. Nucleic Acids Research *37*, D77–D82.

[Nishida et al., 2008] Nishida, K., Frith, M. C. and Nakai, K. (2008). Pseudocounts for transcription factor binding sites. Nucleic Acids Research *37*, 939–944.

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

[Ohlsson et al., 2001] Ohlsson, R., Renkawitz, R. and Lobanenkov, V. (2001). CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. Trends in Genetics *17*, 520–527.

[Oliphant et al., 1989] Oliphant, A. R., Brandl, C. J. and Struhl, K. (1989). Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. Molecular and Cellular Biology *9*, 2944–2949.

[Palumbo and Newberg, 2010] Palumbo, M. J. and Newberg, L. A. (2010). Phyloscan: locating transcription-regulating binding sites in mixed aligned and unaligned sequence data. Nucleic Acids Research *38*, W268–W274.

[Panne et al., 2007] Panne, D., Maniatis, T. and Harrison, S. C. (2007). An Atomic Model of the Interferon-β Enhanceosome. Cell *129*, 1111–1123.

[Papatsenko and Levine, 2007] Papatsenko, D. and Levine, M. (2007). A rationale for the enhanceosome and other evolutionarily constrained enhancers. Current biology: CB *17*, R955–957.

[Park, 2009] Park, P. J. (2009). ChIP–seq: advantages and challenges of a maturing technology. Nature Reviews Genetics *10*, 669–680.

[Pavesi et al., 2001] Pavesi, G., Mauri, G. and Pesole, G. (2001). An algorithm for finding signals of unknown length in DNA sequences. Bioinformatics *17*, S207–S214.

[Phoophakdee and Zaki, 2008] Phoophakdee, B. and Zaki, M. J. (2008). TRELLIS+: an effective approach for indexing genome-scale sequences using suffix trees. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing *vol*, 90–101.

[Piipari et al., 2010] Piipari, M., Down, T. A. and Hubbard, T. J. P. (2010). Metamotifs-a generative model for building families of nucleotide position weight matrices. BMC bioinformatics *11*, 348.

[Pique-Regi et al., 2010] Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y. and Pritchard, J. K. (2010). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Research *21*, 447–455.

[Pizzi et al., 2011] Pizzi, C., Rastas, P. and Ukkonen, E. (2011). Finding Significant Matches of Position Weight Matrices in Linear Time. IEEE/ACM Transactions on Computational Biology and Bioinformatics *8*, 69–79.

[Portales-Casamar et al., 2009] Portales-Casamar, E., Thongjuea, S., Kwon, A. T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W. W. and Sandelin, A. (2009). JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic Acids Research *38*, D105–D110.

[Prakash et al., 2004] Prakash, A., Blanchette, M., Sinha, S. and Tompa, M. (2004). Motif discovery in heterogeneous sequence data. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing *vol*, 348–359.

[Qi et al., 2005] Qi, Y., Ye, P. and Bader, J. S. (2005). Genetic interaction motif finding by expectation maximization – a novel statistical model for inferring gene modules from synthetic lethality. BMC Bioinformatics *6*, 288.

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

[Rajewsky et al., 2002] Rajewsky, N., Vergassola, M., Gaul, U. and Siggia, E. D. (2002). Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. BMC Bioinformatics *3*, 30.

[Reid et al., 2010] Reid, J., Evans, K., Dyer, N., Wernisch, L. and Ott, S. (2010). Variable structure motifs for transcription factor binding sites. BMC genomics *11*, 30.

[Reid et al., 2009] Reid, J., Ott, S. and Wernisch, L. (2009). Transcriptional programs: Modelling higher order structure in transcriptional control. BMC bioinformatics *10*, 218.

[Reid and Wernisch, 2011] Reid, J. and Wernisch, L. (2011). STEME: efficient EM to find motifs in large data sets. Nucleic acids research *39*, e126–e126.

[Rivera-Pomar and Jãckle, 1996] Rivera-Pomar, R. and Jãckle, H. (1996). From gradients to stripes in Drosophila embryogenesis: filling in the gaps. Trends in Genetics *12*, 478–483.

[Robertson et al., 2007] Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., Thiessen, N., Griffith, O. L., He, A., Marra, M., Snyder, M. and Jones, S. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nature Methods *4*, 651–657.

[Robison et al., 1998] Robison, K., McGuire, A. M. and Church, G. M. (1998). A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. Journal of Molecular Biology *284*, 241–254.

[Roider et al., 2007] Roider, H. G., Kanhere, A., Manke, T. and Vingron, M. (2007). Predicting Transcription Factor Affinities to DNA from a Biophysical Model. Bioinformatics *23*, 134–141.

[Sandelin, 2004] Sandelin, A. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Research *32*, 91D–94.

[Sandelin and Wasserman, 2004] Sandelin, A. and Wasserman, W. (2004). Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. Journal of molecular biology *338*, 207–215.

[Sandve et al., 2007] Sandve, G., Abul, O., Walseng, V. and Drabløs, F. (2007). Improved benchmarks for computational motif discovery. BMC bioinformatics *8*, 193.

[Sandve et al., 2006] Sandve, G., Nedland, M., Syrstad, Ø., Eidsheim, L., Abul, O. and Drabløs, F. (2006). Accelerating motif discovery: Motif matching on parallel hardware. Algorithms in Bioinformatics *4175*, 197–206.

[Sarai and Takeda, 1989] Sarai, A. and Takeda, Y. (1989). Lambda Repressor Recognizes the Approximately 2-Fold Symmetric Half-Operator Sequences Asymmetrically. Proceedings of the National Academy of Sciences *86*, 6513–6517.

[Schatz et al., 2007] Schatz, M., Trapnell, C., Delcher, A. and Varshney, A. (2007). High-throughput sequence alignment using Graphics Processing Units. BMC Bioinformatics *8*, 474.

[Schlesinger et al., 2013] Schlesinger, F., Smith, A. D., Gingeras, T. R., Hannon, G. J. and Hodges, E. (2013). De novo DNA demethylation and noncoding transcription define active intergenic regulatory elements. Genome Research *23*, 1601–1614.

[Schmidt et al., 2010] Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S., Talianidis, I., Flicek, P. and Odom, D. T. (2010). Five-Vertebrate ChIP-Seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. Science *328*, 1036–1040.

[Schneider and Stephens, 1990] Schneider, T. and Stephens, R. (1990). Sequence logos: a new way to display consensus sequences. Nucleic acids research *18*, 6097.

[Schneider, 1997] Schneider, T. D. (1997). Information content of individual genetic sequences. Journal of theoretical biology *189*, 427–441.

[Schones et al., 2007] Schones, D., Smith, A. and Zhang, M. (2007). Statistical significance of cis-regulatory modules. BMC Bioinformatics *8*, 19.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the Dimension of a Model. The Annals of Statistics *6*, 461–464.

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

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

[Segal et al., 2003b] Segal, E., Yelensky, R. and Koller, D. (2003b). Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. Bioinformatics *19*, i273–i282.

[Sethuraman, 1994] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. Statistica Sinica *4*, 639–650.

[Sharan et al., 2003] Sharan, R., Ovcharenko, I., Ben-Hur, A. and Karp, R. (2003). CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments. Bioinformatics *19*, i283–i291.

[Sharon et al., 2008] Sharon, E., Lubliner, S. and Segal, E. (2008). A feature-based approach to modeling protein–DNA interactions. PLoS computational biology *4*, e1000154.

[Shen et al., 2012] Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. V. and Ren, B. (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116–120.

[Siddharthan et al., 2005] Siddharthan, R., Siggia, E. D. and van Nimwegen, E. (2005). PhyloGibbs: A Gibbs Sampling Motif Finder That Incorporates Phylogeny. PLoS Computational Biology *1*, e67.

[Simmons et al., 1990] Simmons, D. M., Voss, J. W., Ingraham, H. A., Holloway, J. M., Broide, R. S., Rosenfeld, M. G. and Swanson, L. W. (1990). Pituitary cell phenotypes involve cell-specific Pit-1 mRNA translation and synergistic interactions with other classes of transcription factors. Genes & Development *4*, 695–711.

[Simonis et al., 2006] Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B. and de Laat, W. (2006). Nuclear organization of active

and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C). Nature Genetics *38*, 1348–1354.

[Singh et al., 2007] Singh, L., Wang, L. and Hannenhalli, S. (2007). TREMOR—a tool for retrieving transcriptional modules by incorporating motif covariance. Nucleic acids research *35*, 7360–7371.

[Sinha et al., 2004] Sinha, S., Blanchette, M. and Tompa, M. (2004). PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences. BMC Bioinformatics *5*, 170.

[Snoek et al., 2013] Snoek, J., Larochelle, H. and Adams, R. P. (2013). Practical Bayesian Optimization of Machine Learning Algorithms. In Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012, (P Bartlett, ed.), Advances in Neural Information Processing Systems 25. Curran Associates, Inc.

[Southall and Brand, 2007] Southall, T. D. and Brand, A. H. (2007). Chromatin profiling in model organisms. Briefings in Functional Genomics & Proteomics *6*, 133–140.

[Staden, 1989] Staden, R. (1989). Methods for calculating the probabilities of finding patterns in sequences. Computer applications in the biosciences : CABIOS *5*, 89–96.

[Stark et al., 2007] Stark, A., Lin, M. F., Kheradpour, P., Pedersen, J. S., Parts, L., Carlson, J. W., Crosby, M. A., Rasmussen, M. D., Roy, S., Deoras, A. N., Ruby, J. G., Brennecke, J., Matthews, B. B., Schroeder, A. J., Gramates, L. S., Pierre, S. E. S., Roark, M., Jr, K. L. W., Kulathinal, R. J., Zhang, P., Myrick, K. V., Antone, J. V., Gelbart, W. M., Yu, C., Park, S., Wan, K. H., Celniker, S. E., Hodges, E., Hinrichs, A. S., Caspi, A., Paten, B., Park, S.-W., Han, M. V., Maeder, M. L., Polansky, B. J., Robson, B. E., Aerts, S., Helden, J. v., Hassan, B., Gilbert, D. G., Eastman, D. A., Rice, M., Weir, M., Hahn, M. W., Park, Y., Dewey, C. N., Pachter, L., Kent, W. J., Haussler, D., Lai, E. C., Bartel, D. P., Hannon, G. J., Kaufman, T. C., Eisen, M. B., Clark, A. G., Smith, D. and Kellis, M. (2007). Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. Nature *450*, 219–232.

[Stathopoulos and Levine, 2005] Stathopoulos, A. and Levine, M. (2005). Genomic Regulatory Networks and Animal Development. Developmental Cell *9*, 449–462.

[Stormo, 1998] Stormo, G. D. (1998). Information Content and Free Energy in DNA–Protein Interactions. Journal of Theoretical Biology *195*, 135–137.

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

[Stormo and Fields, 1998] Stormo, G. D. and Fields, D. S. (1998). Specificity, free energy and information content in protein-DNA interactions. Trends in biochemical sciences *23*, 109–113.

[Stormo et al., 1982] Stormo, G. D., Schneider, T. D., Gold, L. and Ehrenfeucht, A. (1982). Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. Nucleic Acids Research *10*, 2997–3011.

[Su et al., 2002] Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., Wiltshire, T., Orth, A. P., Vega, R. G., Sapinoso, L. M., Moqrich, A., Patapoutian, A., Hampton, G. M., Schultz, P. G. and Hogenesch, J. B. (2002). Large-scale analysis of the human and mouse transcriptomes. Proceedings of the National Academy of Sciences *99*, 4465–4470.

[Swanson et al., 2010] Swanson, C. I., Evans, N. C. and Barolo, S. (2010). Structural Rules and Complex Regulatory Circuitry Constrain Expression of a Notch- and EGFR-Regulated Eye Enhancer. Developmental cell *18*, 359–370.

[Takeda et al., 1989] Takeda, Y., Sarai, A. and Rivera, V. M. (1989). Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments. Proceedings of the National Academy of Sciences of the United States of America *86*, 439–443.

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

[Tanay et al., 2004a] Tanay, A., Gat-Viks, I. and Shamir, R. (2004a). A Global View of the Selection Forces in the Evolution of Yeast Cis-Regulation. Genome Research *14*, 829–834.

[Tanay et al., 2004b] Tanay, A., Sharan, R., Kupiec, M. and Shamir, R. (2004b). Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. Proceedings of the National Academy of Sciences of the United States of America *101*, 2981.

[Teh, 2006] Teh, Y. W. (2006). A Bayesian interpretation of interpolated Kneser-Ney. Technical Report TRA2/06 National University of Singapore School of Computing.

[Teh et al., 2006] Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006). Hierarchical dirichlet processes. Journal of the American Statistical Association *101*, 1566–1581.

[Teh et al., 2008] Teh, Y. W., Kurihara, K. and Welling, M. (2008). Collapsed variational inference for HDP. Advances in neural information processing systems *20*, 1481–1488.

[Teif and Rippe, 2010] Teif, V. B. and Rippe, K. (2010). Statistical–mechanical lattice models for protein–DNA binding in chromatin. Journal of Physics: Condensed Matter *22*, 414105.

[Thanos and Maniatis, 1995] Thanos, D. and Maniatis, T. (1995). Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. Cell *83*, 1091–1100.

[Thijs et al., 2001] Thijs, G., Lescot, M., Marchal, K., Rombauts, S., Moor, B. D., Rouzé, P. and Moreau, Y. (2001). A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics *17*, 1113–1122.

[Thomas-Chollier et al., 2011] Thomas-Chollier, M., Defrance, M., Medina-Rivera, A., Sand, O., Herrmann, C., Thieffry, D. and van Helden, J. (2011). RSAT 2011: regulatory sequence analysis tools. Nucleic Acids Research *39*, W86–W91.

[Thomson et al., 2010] Thomson, J. P., Skene, P. J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A. R. W., Deaton, A., Andrews, R., James, K. D., Turner, D. J., Illingworth, R. and Bird, A. (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature *464*, 1082–1086.

[Tompa et al., 2005] Tompa, M., Li, N., Bailey, T. L., Church, G. M., Moor, B. D., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., Régnier, M., Simonis, N., Sinha, S., Thijs, G., Helden, J. v., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C. and Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. Nature Biotechnology *23*, 137–144.

[Touzet and Varré, 2007] Touzet, H. and Varré, J.-S. (2007). Efficient and accurate P-value computation for Position Weight Matrices. Algorithms for molecular biology : AMB *2*, 15.

[Treisman et al., 1992] Treisman, R., Marais, R. and Wynne, J. (1992). Spatial flexibility in ternary complexes between SRF and its accessory proteins. The EMBO Journal *11*, 4631–4640.

[Tsong et al., 2006] Tsong, A. E., Tuch, B. B., Li, H. and Johnson, A. D. (2006). Evolution of alternative transcriptional circuits with identical logic. Nature *443*, 415–420.

[Tuerk and Gold, 1990] Tuerk, C. and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science (New York, N.Y.) *249*, 505–510.

[Turatsinze et al., 2008] Turatsinze, J.-V., Thomas-Chollier, M., Defrance, M. and van Helden, J. (2008). Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. Nature Protocols *3*, 1578–1588.

[Tuteja et al., 2008] Tuteja, G., Jensen, S. T., White, P. and Kaestner, K. H. (2008). Cis-regulatory modules in the mammalian liver: composition depends on strength of Foxa2 consensus site. Nucleic Acids Research *36*, 4149–4157.

[Ukkonen, 1995] Ukkonen, E. (1995). On-Line Construction of Suffix Trees. Springer-Verlag.

[Valouev et al., 2008] Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. M. and Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nature Methods *5*, 829–834.

[van Steensel and Henikoff, 2000] van Steensel, B. and Henikoff, S. (2000). Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. Nature biotechnology *18*, 424–428.

[Visel et al., 2009] Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M. and Pennacchio, L. A. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature *457*, 854–858.

[Wang et al., 2008] Wang, Z., Zang, C., Rosenfeld, J. A., Schones, D. E., Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Peng, W., Zhang, M. Q. and Zhao, K. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. Nature Genetics *40*, 897–903.

[Wasson and Hartemink, 2009] Wasson, T. and Hartemink, A. J. (2009). An ensemble model of competitive multi-factor binding of the genome. Genome Research *19*, 2101–2112.

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

[Whitfield et al., 2012] Whitfield, T., Wang, J., Collins, P., Partridge, E. C., Aldred, S., Trinklein, N., Myers, R. and Weng, Z. (2012). Functional analysis of transcription factor binding sites in human promoters. Genome Biology *13*, R50.

[Wilcoxon, 1945] Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. Biometrics Bulletin *1*, 80.

[Winn, 2003] Winn, J. (2003). Variational message passing and its applications. PhD thesis, Cambridge University.

[Wu, 1980] Wu, C. (1980). The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. Nature *286*, 854–860.

[Wu et al., 2006] Wu, W.-S., Li, W.-H. and Chen, B.-S. (2006). Computational reconstruction of transcriptional regulatory modules of the yeast cell cycle. BMC Bioinformatics *7*, 421.

[Xie et al., 2013] Xie, M., Hong, C., Zhang, B., Lowdon, R. F., Xing, X., Li, D., Zhou, X., Lee, H. J., Maire, C. L., Ligon, K. L., Gascard, P., Sigaroudinia, M., Tlsty, T. D., Kadlecek, T., Weiss, A., O'Geen, H., Farnham, P. J., Madden, P. A. F., Mungall, A. J., Tam, A., Kamoh, B., Cho, S., Moore, R., Hirst, M., Marra, M. A., Costello, J. F. and Wang, T. (2013). DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. Nature Genetics *45*, 836–841.

[Xie et al., 2009] Xie, X., Rigor, P. and Baldi, P. (2009). MotifMap: a human genome-wide map of candidate regulatory motif sites. Bioinformatics *25*, 167.

[Xu et al., 2004] Xu, X., Wang, L. and Ding, D. (2004). Learning module networks from genome-wide location and expression data. FEBS Letters *578*, 297–304.

[Zhao and Stormo, 2011] Zhao, Y. and Stormo, G. D. (2011). Quantitative analysis demonstrates most transcription factors require only simple models of specificity. Nature biotechnology *29*, 480–483.

[Zhao et al., 2006] Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Singh Sandhu, K., Singh, U., Pant, V., Tiwari, V., Kurukuti, S. and Ohlsson, R. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nature Genetics *38*, 1341–1347.

[Zhou et al., 2008] Zhou, F., Olman, V. and Xu, Y. (2008). Barcodes for genomes and applications. BMC Bioinformatics *9*, 546.

[Zhou and Liu, 2004] Zhou, Q. and Liu, J. S. (2004). Modeling within-motif dependence for transcription factor binding site predictions. Bioinformatics *20*, 909–916.

[Zhu et al., 2013] Zhu, J., Adli, M., Zou, J. Y., Verstappen, G., Coyne, M., Zhang, X., Durham, T., Miri, M., Deshpande, V., De Jager, P. L., Bennett, D. A., Houmard, J. A., Muoio, D. M., Onder, T. T., Camahort, R., Cowan, C. A., Meissner, A., Epstein, C. B., Shoresh, N. and Bernstein, B. E. (2013). Genome-wide Chromatin State Transitions Associated with Developmental and Environmental Cues. Cell *152*, 642–654.

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