Co-occurrence of Transcription Factor Binding Sites

Holger Klein

Dezember 2009

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Gutachter:
Prof. Dr. Martin Vingron
Prof. Dr. Hanspeter Herzel

1. Referent: Prof. Dr. Martin Vingron
2. Referent: Prof. Dr. Hanspeter Herzel

Tag der Promotion: 12. Mai 2010

Contents

Chapter 1

Overview

1.1 Motivation and Thesis Structure

Transcriptional Regulation Gene regulation deals with the processes, that enable an organism to create a large variety of cells and cell states from the same genome. A cell uses different genes under divergent conditions, for example in two phases of the cell cycle. In metazoan organisms completely different cell types, like a liver and a brain cell, are encoded in the same genome. However, the set of genes needed in both cases differs.
A crucial step at which a cell regulates the production of proteins and other gene products is transcriptional regulation. The molecular machinery that transcribes the gene assembles at the transcriptional start site, the point from where an RNA copy of the gene is transcribed. The RNA is subsequently processed and often later on used as a blueprint for protein translation. The assembly of the transcriptional apparatus is governed by transcription factors: proteins that recognize and bind to the DNA and support tethering of the transcriptional apparatus to the transcriptional start site. We present a concise overview about the molecular biology of transcription in Section 2.1.
To date the known target spectrum of transcription factors ranges from very specific factors with only a tiny number of targets under distinct conditions to ubiquituous factors, which influence the transcription of a large fraction of genes in a large variety of cell types and conditions. The estimated fraction of different transcription factors encoded in the human genome varies between 6 and 8%, leading to more than 2,500 different transcription factors [13]. The number of transcription factors expressed per human tissue is between 150 and 350 [342].

Transcription Factor Interactions and Regulatory Regions Aggregates of several transcription factors can act synergistically or antagonistically. Using combinations of transcription factors is beneficial to an organism, not only because of the many possible interactions which allow for fine-grained regulation, but also because a partial redundance of transription factors makes the transcriptional response more stable. The transcriptional network is dynamic, and the number of possible combinations of factor-factor and factor-DNA interactions in a complete organism is enormous due to the size of the genome, the number of different transcription factors, and the large number of different cell states and cell types. We summarize details about transcription factor interactions in Section 2.2.
Transcription factors bind to two major classes of regulatory regions — promoters and enhancers. On the genomic sequence the promoters are located close to the regulated gene. Enhancers can be far away; in this case the interaction with the transcriptional machinery is

possible because of bending and looping of the DNA. The promoters harbor binding sites for general transcription factors which drive a low level of transcription, as well as for specific factors, that modulate the expression strength dependent on various factors. Enhancers usually influence specific expression. For properties of regulatory regions we refer the reader to Section 3.3.1. We present an overview of the large variety of experimental methods for the detection of transcription factor binding sites, transcription factor interactions and the detection of regulatory regions in Sections 2.3.1 and 2.3.3.

Computational Methods Already the knowledge of the first binding motifs of transcription factors led researchers to search for more potential binding sites in other parts of the genome with *in silico* methods. Over time this lead to the development of copious methods for the prediction of transcription factor binding sites. A general problem of these methods lies in the nature of transcription factor binding sites. They are short and degenerate, which means that a high number of occurrences of a matching DNA motif is present by chance, rather than for a functional reason. We explain the general ideas and the problems that arise in Section 3.1.

The prediction of transcription factor interactions is a difficult problem. Available methods make use of expression data, experimental binding data, overrepresented sequence motifs, or predicted transcription factor binding sites and apply a large variety of statistical methods. We summarize these methods in Section 3.2.2.

Also for the prediction of regulatory regions one has the choice between many different tools. Partly they are sequence based only and deploy low level features like GC content or higher level features such as binding motifs. Other tools additionally utilise experimental data. We review various approaches in Section 3.3.

Working Hypothesis The underlying assumption of the present work is, that *in regulatory regions, binding sites of interacting transcription factors co-occur more often than expected by chance.*
The prediction of individual transcription factor binding sites is hampered by a large number of false positive results. Nevertheless we expect our assumption to be true also for *predicted* binding sites, even if the signal that emanates is weakened.

To that end we develop two new methods, one for the prediction of transcription factor interactions, and one for the prediction of regulatory regions based on commonness of transcription factor binding site combinations.

A Co-occurrence Score for Transcription Factor Binding Sites For the prediction of functional transcription factor interactions, we develop a counting method, which applies a sliding window over annotated transcription factor binding sites. The counting procedure is able to deal with overlapping windows, homotypic clusters, and overlapping binding sites. For the detection of overrepresented TFBS pairs, we calculate as a co-occurrence score the log odds score of observed over expected number TFBS pairs. We estimate the number

of expected pairs using a label permutation procedure with subsequent recountings. We describe our method in Section 4.
We assess the counting procedure and the co-occurrence score on artificially generated datasets with defined number of co-occurrences in Section 6.1 and show, that the method is able to detect low TFBS pair enrichments. We apply the method to yeast regulatory regions in Section 6.2, and find that a large number of overrepresented TFBS pairs in fact belong to transcription factors known to interact. Moreover we examine the similarity of binding sites of interacting pairs and reasons for underrepresentation of TFBS pairs. In Section 6.5 we compare the co-occurrence score with the *costat* method by Pape et al. [253]. Subsequently we apply the method to vertebrate data in Chapter 6. Despite the much higher complexity of the regulatory network of vertebrates compared to yeast, we still find many known interactions among the top scoring pairs. This is the case for a genome wide study in human in Section 6.3, as well as for genes expressed in human embryonic kidney cells (Section 6.4.2), and tissue specific gene sets from mouse (Section 6.4.3).

Binding Site Graphs for the Prediction of Regulatory Regions Many tools for the prediction of regulatory regions explicitly or implicitly apply sequence properties like the GC content or the presence of CpG islands for the detection of regulatory regions. We aim to design a method, which is less dependent on low level features and hence makes use of the knowledge about over- and underrepresented TFBS pairs in known regulatory regions to measure the regulatory potential. We represent the predicted binding sites in a sequence to be characterised as the vertices in a graph. Subsequent assignment of co-occurrence scores as edge weights for all vertex pairs leads to a binding site graph. The co-occurrence scores originate from known regulatory regions and are calculated with the method from Section 4. Using this graph, we now can calculate various edge-weight based scores for the input sequence, which we call regulatory potentials and which represent the level of abundance of transcription factor combinations typical for regulatory regions. We present our approach in Chapter 5.
In Chapter 7 we apply the methods to known regulatory regions. We show the performance of one of the scores on the well examined regulatory regions of the murine *PAX6* gene and known human enhancer regions from the VISTA set. In Section 7.2 we assess the reliability of our method for genome-wide prediction of regulatory regions based on test sets, consisting of promoter and enhancer regions as positive sets and an artificial, shuffled sequence set and an intergenic set as negative sets. We find, that although the biggest factor playing a role in the prediction of regulatory function again is the GC content of a candidate sequence, our method should be used to filter out false positive predictions of regulatory function based on GC content.

In Chapter 8 we summarize and discuss our findings.

1.1.1 Publications

Some of the work presented in this thesis has been published in the paper "Using Transcription Factor Binding Site Co-Occurrence to Predict Regulatory Regions", Klein & Vingron, Genome Informatics, 2007. The *costat* method, to which we compare our co-occurrence score was published in the paper "Statistical detection of co-operative transcription factors with similarity adjustment.", Pape, Klein, Vingron, Bioinformatics, 2009.

1.1.2 Acknowledgements

During my time in the gene regulation group in the department for computational molecular biology at the Max Planck Institute for Molecular Genetics, I enjoyed an intellectually stimulating and friendly environment. First of all I want to thank Martin Vingron for the opportunity to work on an interesting topic in his group, and for the ideas and continuous support that I received from him in the time that I spent at the MPI. Moreover, I would like to acknowledge the International Research and Training Group (IRTG), the PIs and fellows of which provided an additional inspiring forum to discuss ideas related to this thesis. I spent two months in Zhiping Weng's lab, during which I received a lot of input from Zhiping and other group members.
Apart from thanking all members of the department, I would like to individually mention a number of people for longer discussions related to my work, providing data, and/or proofreading of the present manuscript: Hannes Luz, Hugues Richard, Sarah Behrens, Utz Pape, Marcel Schulz, Sean O'Keefe, Abha Singh Bais, Steffen Grossmann, Stefan Röpcke, Hans-Jörg Warnatz, Helge Roider, Ho-Ryun Chung, Christoph Dieterich, Szymon Kiełbasa. I enjoyed sharing my office with Christoph and Szymon.
The scientific and professional interactions within the group and the institute often evolved to friendship. I would not want to miss that.
I would like to acknowledge my parents Birgid and Hans-Josef Klein for their ongoing support and encouragement (not only during the time of my thesis). Lastly, I want to thank Vesna for her patience and support.

Part I

Background

Chapter 2

Transcriptional Regulation

The focus of the present work lies on the prediction of transcription factor interactions and their use for the prediction of regulatory regions. We will begin this chapter with a short summary of the biological context of eukaryotic gene regulation. We will then shift the focus towards transcription factors and their co-operation. Subsequently we will provide an overview of the most important experimental methods for the detection of transcription factor binding, transcription factor co-operation, and regulatory regions. Here we can only give a short introduction into the topics. For more details we refer to textbooks like Alberts et al. [8].

2.1 Molecular Biology of Gene Regulation

2.1.1 From DNA to Proteins

The Structure of DNA The genetic information of a living organism is stored in the *deoxyribonucleic acid* (DNA). The DNA consists of a chain of *nucleotides*; sugar molecules combined with one of the bases adenine (A), cytosine (C), guanine (G), and thymine (T). Combined via phosphodiester bonds, the sugar molecules form the backbone of the DNA. Chargaff [57] found that the bases C and G as well as A and T form pairs and hence occur in equal fractions in the DNA. The base pairing leads to the well-known double-helix structure of the DNA, published by Watson and Crick [353]. It consists of two anti-parallel, complementary strands, the direction of which is defined by the position of the phosphodiester bonds in the sugar molecules: the bonds connect the 5' carbon of one deoxyribose to the 3' carbon of the next. The synthesis and processing of DNA only takes place in the direction from the 5' end to 3' end. This directionality also defines the terms *upstream* (5') and *downstream* (3'). Figure 2.1 shows the chemical structure of the four bases and the pairing of bases.

Cells of eukaryotic organisms, unlike bacteria, contain nuclei which harbour the *chromosomes*. The chromosomes are structures that consist of the double stranded DNA and accessory structural proteins. The DNA is wound around the *nucleosomes*, each of which consists of a *histone octamer*. One loop of the DNA around a nucleosome has the length of 146 base pairs (bp).

Figure 2.2 shows different packing levels of the DNA. The DNA is only in the fully condensed state when entering the process of cell division. Otherwise the condensation state can either be active *euchromatin*, or silent *heterochromatin*, meaning that the DNA is easily

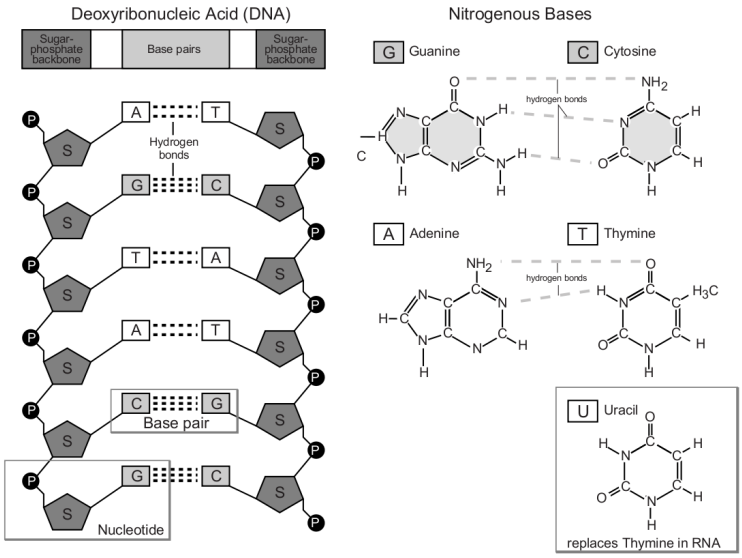


Figure 2.1: Building blocks if the DNA and base pairing. Guanine / Cytosine pairs are connected by three, Adenine / Thymine pairs by two hydrogen bonds. Image reproduced with permission from `http://www.accessexcellence.org/`, National Health Museum, USA

accessible for other proteins or not.

Genes, Transcription and Translation The classical definition of a gene is a stretch of DNA, that encodes functional cellular components like proteins. In a process called *transcription*, the enzyme *RNA polymerase* transcribes the DNA of a gene into *ribonucleic acid* (RNA). The RNA is a linear bio-polymer, similar to the DNA, but consists only of a single strand, and with *uracil* (U) instead of the thymine in the DNA (Figure 2.1). The RNA plays a role in direct catalysis of metabolic processes and as a structural component in RNA protein complexes, and on the other hand as *messenger RNA* (mRNA) as a template for the production of proteins.

For the production of proteins the first step is the splicing of the primary transcript. The *introns* are removed, resulting in the *mature* mRNA, which only consists of the *exons*, combined to the *open reading frame* (ORF), and 5' and 3' untranslated regions (UTR). By variation of exon usage, splicing can lead to different splice forms or *alternative transcripts*, finally resulting in different protein products from the same gene. Subsequently the mature mRNA is transported out of the cell nucleus to the *ribosomes*, in which translation to a chain of amino acids based on the genetic code takes place. In human, roughly 1.5% of the complete genome are exonic sequences. Figure 2.3 exemplifies a eukaryotic gene with the transcriptional unit containing exons, introns, and UTR. The flow of information from DNA to RNA to protein is commonly known as *the central dogma of molecular biology*, a term coined by Crick [70].

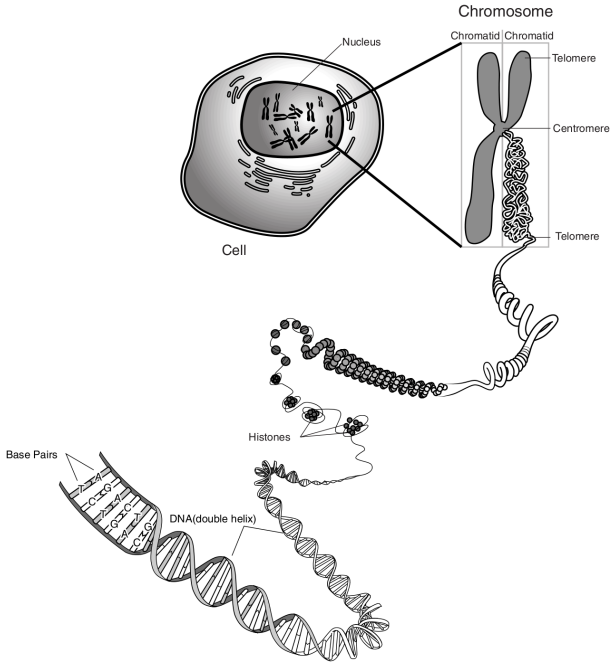


Figure 2.2: Structural organization of the genome. The nucleus in the cell contains the chromosomes. The chromosomes consist of identical sister chromatids. The DNA is wound around the nucleosomes, which are built from eight histone molecules (octamer). Image reproduced with permission from `http://www.genome.gov/`, National Human Genome Research Institute, USA

2.1.2 Transcriptional Regulation

An organism requires the various genes in different amounts and under different conditions. Thus the creation of gene products is controlled at all levels, from chromatin modifications over transcriptional regulation to post-translational modifications and protein degradation. In the following, we will focus on transcriptional regulation, since this is the most important aspect with respect to this thesis.

Regulatory Regions and the Transcriptional Machinery A part of a gene not mentioned before is the *promoter*. It contains the information needed for the activation or repression of gene [256, 309] . In eukaryotes, the transcription of a genes starts with the assembly of the *pre-initiation complex* on the promoter. The pre-initiation complex contains 10 to 12 proteins, among them the general *transcription factors* [194, 246] and one of three different RNA polymerases, in case of protein coding genes *RNA polymerase II*.
A common definition for the promoter is the region upstream of the *transcriptional start site* (TSS). The promoter is divided into two parts: the *core* or *basal* promoter at roughly -35bp relative to the TSS, which the aforementioned transcriptional apparatus binds to. For a part of the genes, it contains the *TATA box*, a sequence motif which the *TATA-binding protein*

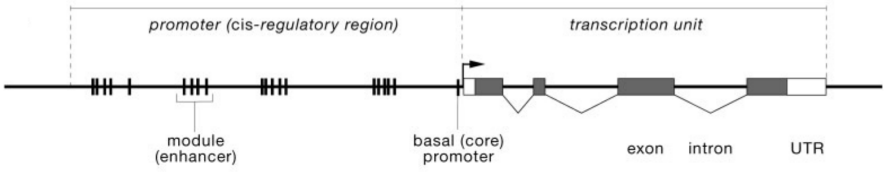


Figure 2.3: Structural organization of a eukaryotic gene. The promoter contains several cis-regulatory modules. The transcriptional unit consists of exons, introns, and untranslated regions. Image reproduced from Wray et al. [361].

(TBP) binds to [202]. The *proximal promoter* ranges up to a few hundred bp upstream of the TSS. It contains binding sites for for other, mostly sequence-specific TFs, needed for the activation or repression of the regulated gene in various conditions [196]. Another type of genomic region that influences the transcription is the *enhancer* [177]. It also harbours binding sites for sequence-specific TFs, but can be far away from the TSS, up to 100,000bps [150] or even 1,000,000bps away [259, 179]. Nevertheless the TFs that bind to enhancers interact with the transcriptional apparatus at the promoter by looping of the DNA.
The TFs, that are part of the transcriptional apparatus activate unspecific expression on a low level. For higher expression levels as well as specific spatial or temporal regulation of gene expression, TFs that bind to the proximal promoter and enhancers play a role. Figure 2.4 illustrates the binding of the transcriptional machinery to the promoter and the interaction with enhancer modules.

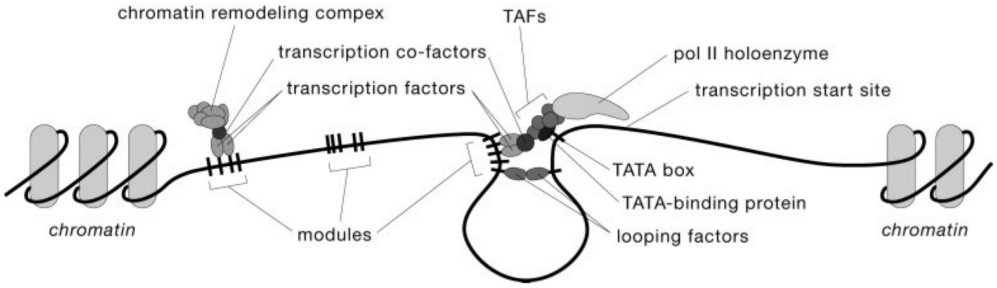


Figure 2.4: Transcriptional regulation. Several general transcription factors bind to the basal promoter, among them the TATA binding protein TBP and other TATA associated factors (TAF), to form the preinitiation complex. The regulatory regions contain several cis-regulatory modules, to which combinations of TFs bind. Image reproduced Wray et al. [361]

For an extensive review of regulatory elements see Maston et al. [217].

Number of Transcription Factors The human body consists of more than 200 different cell types [337] and an estimated number of 20,000-25,000 different genes spread over a genome of more than 3,000,000,000bps [65], and a part of the tissues requires tight spatial

and temporal regulation of gene expression. Only a subset of the genes in a eukaryotic cell are expressed at any given time, and the proportion and amount changes during life cycle, among cell types, in response to external conditions [161, 356, 11]. Hence it is not a surprise that eukaryotic organisms have a many different TFs. Nimwegen [236] finds that large genomes tend to have more TFs. The number of genes with known DNA-binding domains in baker's yeast *Saccharomyces cerevisiae* is 245, which corresponds to 3.9% of the genes. In human there are 2604 genes with known DNA-binding domains, amounting to roughly 8.1% of all genes [13]. Vaquerizas et al. [342] estimate that roughly 6% of human genes are TFs, with an upper bound of 8.2%. In human, between 150 and 350 different TFs are expressed per tissue. On average 6% of expressed genes in a tissue are TFs [342]. The expression patterns of TFs themselves are highly complex and expression takes place in distinct domains [75, 76].

Transcription Factor Binding Up to now, eight different structural groups of transcription factors are known. They belong to 54 different structural families [207] and can draw on 12-15 different DNA-binding domains [143]. The *transcription factor binding sites* (TFBSs) on which the TFs interact with the DNA are typically short and degenerate. The length of the DNA region with physical contact with a TF is usually between 10 and 20bps with a core of five to eight bps which are required to contain specific bases. The distance between TFBSs is constrained for sterical reasons [12]. A typical promoter has an occupancy of TF binding of 10 to 20% of its sequence.

Modulation of Transcription Factor Activity An organism has a number of possibilities to influence the activity of a TF. An obvious way is to change the expression of the factor itself on the level of transcription as well as on the level of translation [1]. Formation of dimers, presence of cofactors, as well as post-translational modifications are another option.

On the other hand modifications of the DNA allow the modulation of TF binding to the DNA: *Histone modifications* like (de-)acetylation or (de-)methylation of the side chains lead to a higher or lower compaction of the DNA-histone complexes, hence increasing or decreasing the accessibility of regulatory regions for the transcriptional machinery [113]. The most common modification of the DNA molecule itself is the *DNA methylation*, specifically of the cytosine in `CG` dinucleotides (often referred to as `CpG`, with p representing the phosphodiester bond). Cytosines in CpG dinucleotides are methylated by DNA methyltransferases to form 5-methyl-cytosine. Methyl-cytosines are prone to transition mutation to a thymine [288], hence the `CG` dinucleotide is rare in mammalian genomes. Evolutionary constraints lead to higher amounts of `CG` dinucleotides in regulatory regions, often occurring in clusters called *CpG islands*. The organism can selectively methylate and demethylate `CG` dinucleotides. A methylated `CG` dinucleotide in a regulatory region prevents binding of TFs, which results in inhibition of expression. The promoters of roughly 40% of mammalian genes contain `CpG` islands [101]. The original definition for a `CpG` island requires a size of over 200bps, a `GC` percentage of more than 50%, and an observed over expected `CpG` ratio above 60% [120]. Saxonov et al. [287] identified two classes of human promoters, based on their `CpG` observed over expected ratio, with a tendency of the high `CpG` promoters belonging to "house keeping"

genes, that are ubiquitously expressed and the low CpG promoters belonging to specifically expressed genes.

2.2 Cooperation of Transcription Factors

Proteins fulfill their function in the organism in cooperation with other proteins. Estimated numbers of interactions for *E. Coli* proteins are within a range of two to ten [213], in yeast the average number of interactions is five [129]. This is also true for TFs: although a variety of factors is available in a typical eukaryotic cell, the complex expression patterns mentioned in the previous section require combinations of TFs to achieve specific regulation.
In eukaryotes transcription factors are usually part of bigger complexes consisting of several factors and co-factors. Normally it is not a single transcription factor which regulates the spatial and temporal expression patterns of a gene, but combinations of transcription factors. A common transcription factor is organized in a modular fashion and consists of at least one DNA-binding domain and a trans-activating or interaction domain, which enables the factor to interact with other transcription factors or co-factors. Some transcription factors carry other domains, like ligand-binding domains in the case of hormone-receptors, which upon binding of a ligand modulate the activity of a factor. The TFs involved in such a complex bind the DNA in sterical proximity, hence the respective transcription factor binding sites lie close to each other on the DNA sequence. The footprint of a transcription factor complex is a cluster of TFBSs commonly called a *cis-regulatory module* (CRM). This concept goes back to the group of Eric Davidson [180, 372, 24]. For metazoans a typical CRM consists of 10 to 50 binding sites for at least three to 15 different sequence-specific transcription factors spread over up to 500bp [361, 197, 12]. Balmer and Blomhoff [24] estimate the mean length of a CRM to 600bps, with a mean of 24.5 TFBSs [24]. Sometimes multiple similar binding sites increase the sensitivity for a TF [150], lead to a more robust transcriptional response [304], play a role in the activation of morphogen TFs in response to low local TF concentration [133], or simply lead to the binding of a homo-oligomer of the TF (e.g. *p53*, or *NF-κB*). These homotypic clusters exist in various organisms, for example yeast [348], *Drosophila melanogaster* [201, 12], and human [375]. Some transcription factors have well known interaction partners. We call the type of interaction *homotypic* if the TF interacts with a second factor of the same type. The factor *GATA-1* is a well known example displaying homotypic interactions [72]. The interaction is called *heterotypic* if the factor interacts with a factor of another type. Moreover, DNA-binding transcription factors can interact with other DNA-binding factors in an indirect way by mediation of co-factors. Often a given transcription factor can interact with a variety of other factors. Replacement of a TF in a complex can but does not need to change the function of the whole complex. A variety of known cis-regulatory modules can be accessed from the TRANSCompel database [173].

TFs can act together in synergistic or antagonistic fashion [12, 104, 372]. A well studied example of a complex of transcription factors acting together synergistically is the combination of *nuclear factor of activated T cells* (NFAT) belonging to the *REL* family of factors and the *AP-1* hetero-dimer, which itself consists of the two factors *Fos* and *Jun*. The complex activates the transcription of many genes playing a role in immune response. The crystal structure of the complex is shown in Figure 2.5. NFAT and AP-1 directly interact with

each other, hence the respective binding sites lie directly next to each other on the DNA sequence [59].

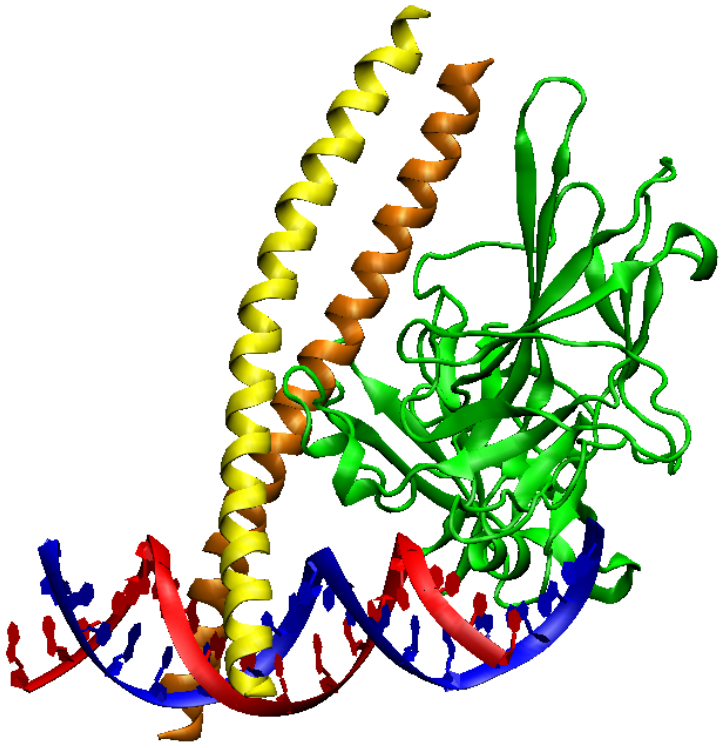


Figure 2.5: Crystal structure of the *NFAT* – *AP-1* complex binding to a DNA fragment from the *interleukin-2* (IL-2) promoter (PDB id 1A02 [59], image created with VMD [155]). *NFAT* (green) belongs to the Rel-family transcription factors and contains a Rel homology region (RHR). *AP-1* is a heterodimer *leucine zipper* consisting of the factors *JUN* (orange) and *FOS* (yellow). The DNA sequence (strands in blue and red) which is bound contains a binding site for *NFAT* (`GGAAA`) and a binding site for *AP-1* (`TGTTTCA`) divided by a two base long spacer sequence (`TT`). *NFAT* and *AP-1* have a large contact surface with mostly polar interactions. Chen et al. [59] expect that the complex shown here is part of a larger complex containing more partners.

Another immune-response related example for interacting transcription factors are *ATF3* from the *CREB/ATF* family of transcription factors and the nuclear factor *NF-κB* [124]. The binding of *ATF3* and its interaction partners was shown to repress transcription of cytokine genes. Like *NFAT*, *NF-κB* contains a *REL* domain needed for interaction with other factors. Gilchrist et al. [124] found the TFBSs for the combination of *ATF3* and *NF-κB* to lie close to each other in respective targets. Moreover they discovered a direct interaction of the factors - as well as interactions of *ATF3* with the aforementioned *Jun* and *Fos*.

Implications of Combinatorial Regulation There are several salient advantages of combinatorial regulation. Due to interactions between two or more factors, an organism can realize a wide variety of transcriptional responses already with a small number of different transcription factors. Combinatorial regulation allows for a much higher number of distinct expression states than the number of transcription factors present in an organism [169]. Moreover simple exchange of single components can modulate the function that a complex fulfills: this way the expression of target genes can vary from cell type to cell type or between different conditions. If a TF in a complex can be replaced without a change in the transcriptional response, the regulatory system gains robustness [335]. Also TFs which themselves do not have a clear DNA binding preference might "inherit" specificity from a more specific interaction partner. For the factors from the bZIP family including for example *Jun*, *Fos*, and *CREB/ATF*, Ryseck and Bravo [283] showed differences in the binding specificity depending on the interaction partners [283]. The homeodomain TF pair *MATa1* / $MAT_{\alpha}2$ plays a role in the yeast cell cycle and its combined DNA affinity is much higher than that of the individual TFs [199]. Other homeodomain TFs also change their binding specificity depending on co-factors [344]. This is in accordance with the finding of Bilu and Barkai [36], that regions which bind many transcription factors tend to have shorter and fuzzier TFBSs than sites which bind only few TFs.

2.3 Experimental Methods

2.3.1 Transcription Factor Binding Sites

Several experimental methods for the detection of transcription factor binding sites have been developed over the years. The methods can be divided into time consuming low throughput methods, which can precisely localize a transcription factor binding site and high throughput methods which require a subsequent computational analysis [96]. Moreover the methods differ depending on whether the factor for which binding sites are sought is known or not.
Regulatory regions are the parts of the genome to which transcription factors bind to influence transcription of genes. Thus the experimental methods used for the detection of regulatory regions largely overlap with the ones used for the detection of transcription factor binding sites.
In the following we give an overview of the most important techniques.

DNase I Hypersensitivity In regions with an open chromatin structure the DNA is accessible for other proteins like transcription factors or nucleases such as *DNase I*. This can be exploited to detect regions which are sensitive for cleavage by *DNase I* and thus are also potentially accessible for transcription factors [123].
The state of the chromatin depends on factors such as the cell type, the developmental stage or the environment normally leading to different DNase I hypersensitive regions. The genome contains large scale DNase hypersensitive regions with sizes between 10 and 100 kilobases [193] but also local hypersensitive regions with sizes between 100 and 400bp [130]. Hypersensitivity in non coding portions of the genome can be used as a marker for transcription factor binding [56].
While the above mentioned methods are considered to be *low throughput* there are *high*

throughput methods, which in their results differ in the resolution. Indirect end-labelling provides a resolution of ca. 500bp [363], while quantitative PCR methods provide a resolution on nucleotide level [366, 220]. Other high throughput methods include quantitative chromatin profiling [86], massive parallel signature sequencing [69] and using tiling arrays for the determination of nucleosome positions [371]. The method can be applied for genome wide detection of potential regulatory regions.

Promoter Analysis Gene expression assays measure the amount of a reporter protein in response to changes in cis-regulatory elements in the respective upstream promoter. The most common reporter constructs use *luciferase*, a gene whose product causes bioluminiscence [79], and *green fluorescent protein* (GFP) from jellyfish, which displays fluorescence when exposed to ultraviolet light [336]. The bioluminescence and the fluorescence allow for simple methods to measure the expression levels of the reporter genes. After incorporation of the reporter into a plasmid, the plasmid is used to transfect a cell. The expression level of the reporter depends on the promoter sequence upstream of the reporter gene. Transcription factors in the cell bind to the promoter and influence the expression of the reporter. Mutations of sequence positions which play a role in the binding of a factor influence the expression level of the reporter gene, thus allowing the identification of transcription factor binding sites. Different groups adapted the method for high throughput usage by alternative transfection methods such as lipofection [323], coinjection [232], and nucleofection [303].

Mobility Shift Assays In this type of assay one applies an electrical potential to a polyacrylamide gel which contains DNA and transcription factors. The voltage causes a movement of the molecules in the gel, which depends on the size of the molecules. DNA bound by a transcription factor moves slower than unbound DNA.
Because the DNA is radio-nucleotide labelled, bands in the gel show up and it is possible to read out the distance that a molecule has travelled.

The *electric mobility shift assay* (EMSA) was one of the first methods to investigate protein-DNA interactions [110, 121, 145]. The assay uses a transcription factor and fragments of DNA of roughly 25bp. If the DNA is not bound by the transcription factor, the gel contains two bands, one for the unbound DNA and one for the transcription factor. If the transcription factor binds the DNA, one expects a third band for a molecule of higher weight, the complex of the DNA bound by the transcription factor. Using different DNA sequences allows for scrutiny of the specificity of the transcription factor [167]. A disadvantage of the method is the potential detection of non-specific DNA-protein interactions [183].

An alternative method combines the DNA-protein binding reaction of EMSA with the cleavage reaction of DNase I. *DNase I footprinting* uses the fact that DNase I can not cleave transcription factor-bound DNA. Visualization of the bands for the radio-nucleotide labelled DNA fragments shows regions devoid of bands representing binding sites in a semi-continuous ladder of bands. DNA sequences used in a footprinting experiment have a length of up to 500bp, resulting in the possibility to localize several transcription factor binding sites at once [116]. Newer methods make use of fluorescent labels [243] instead of radio-nucleotide labelling. Chemical cleavage is an alternative to the usage of nucleases. This solves enzyme

specific problems [89], but does not tackle the issue of unspecific DNA-protein interactions [183].

Nitrocellulose Binding Assay An early method to measure binding of a transcription factor to DNA is the nitrocellulose binding assay [360]. The basis of the assay is negatively charged nitrocellulose paper. Proteins with a net-positive charge bind to the nitrocellulose while negatively charged DNA does not. Washing removes the non-bound DNA from the nitrocellulose filter. The DNA still present is considered to be bound to the transcription factor. Afterwards the bound DNA is eluted by a denaturing enzyme and analyzed in a subsequent gel electrophoresis. Using this assay it is not possible to detect the binding site on the DNA itself. Nowadays the assay is rarely used and has been replaced by a variety of other methods.

NMR and X-ray structures *X-ray* and *NMR* spectroscopy resolves the structure of a transcription factor bound to DNA. This time-consuming procedure is carried out for a small number of factors. An example is the X-ray structure of the heterodimer consisting of *NFAT* and *AP-1* together with a piece of DNA [59]. Using structures it is possible to not only investigate the sequence of a transcription factor binding site but also if the factor bends or torts the DNA. Apart from requiring a lot of time to obtain structures it is impossible to crystallize some proteins. Another drawback is that one usually can only observe a single specific binding site, which might not be representative. Luscombe et al. [207] present an overview of available structures of transcription factors in the PDB database.

SELEX and CAST SELEX (*Selective Evolution of Ligands by Exponential Enrichment* [242, 339, 109]) and CAST (*Cyclic Amplification and Selection of Targets* [362]) are both *in vitro* approaches for the detection of the binding specificity for a known transcription factor. Both screen large pools of short, random oligonucleotides and amplify sequences that bind to the transcription factor in several rounds.
Recent *in vitro* approaches include *DIP-ChIP* (DNA immunoprecipitation with microarray detection) [204] and double-stranded DNA microarray chips [51, 16, 230].

Chromatin Immunoprecipitation Assays *Chromatin Immunoprecipitation* (ChIP) is a method to experimentally determine regions of the genome to which a specific transcription factor binds *in vivo* [313, 45, 245]. Formaldehyde or another cross-linking agent induce the formation of covalent bonds between the transcription factor and the DNA. Subsequently sonication causes cleavage of the DNA into pieces of a size of 100 to 500bp. A specific antibody containing an anchor tags the transcription factor. The antibody allows for retrieval by precipitation of the parts of the DNA cross-linked to the transcription factor. After reversal of the cross-linking process the analysis of the precipitated DNA results in the regions of the genome to which the factor in question bound. The main bottleneck for the ChIP method is the limited availability of transcription factor specific antibodies. A problem of the method is the detection of indirect contacts due to protein-protein interactions.

ChIP Amplification The analysis for the DNA fragments retrieved from the chromatin immunoprecipitation are determined by PCR and followed by sequencing. The preparation of primers for the PCR requires previous knowledge about the genomic region where the transcription factor is supposed to bind. The precise binding location within the genomic fragments can not be determined by sequencing only. Combining the chromatin immunoprecipitation with DNase protection and cleavage allows for to identify the specific binding site [168].

ChIP-Chip and ChIP-Seq ChIP-chip and ChIP-seq both combine chromatin immunoprecipitation with a high throughput technology which permit genome wide analyses of ChIP results. In ChIP-chip one identifies the precipitated DNA by microarray analysis [275, 161, 139]. Hybridization of the DNA fragments to microarray probes and subsequent readout delivers information about enrichment of sequences bound to the transcription factor. Early applications of the method used cDNA microarrays [275], later followed by 50mer oligonucleotide tiling arrays [178]. ChIP-chip methods were applied to a wide variety of transcription factors, genomic regions, cell types and organisms, for example intergenic regions in yeast [275], putative promoter regions in human [200, 239], human CpG island associated promoters [354], genome wide promoters in fibroblasts [178]. Subsequently the statistical analysis of spot intensities is crucial for the determination of most probable binding regions of the transcription factor in question. Buck and Lieb [49] summarize common methods.
A more recent alternative to the identification of precipitated DNA using microarray chips is next-generation sequencing. Here the DNA is *sequenced-by-synthesis* in parallel after being attached to a surface. Pyrosequencing [215], usage of fluorescent reversible terminator deoxyribonucleotides [30, 31], or sequential ligations [300] permit fast and accurate sequencing of large amounts of DNA at the same time. A successive mapping of the TF bound DNA sequences to the genome is necessary. For a review of next-generation sequencing technologies see [299]. Recent examples for ChIP-seq applications are the determination of STAT1 binding sites [277] and the investigation of nucleosome positioning in human promoters [291]. While requiring a high quality reference genome sequence, the advantage of ChIP-seq methods over ChIP-chip is that the accuracy of the result is not limited by the location of probes in the genome. The resolution of binding site localization is in the order of tens of base pairs. The procedures require less material and less replicate experiments, and the results are highly similar to the ones of ChIP chip [214].

Detection of Transcriptional Start Sites The transcriptional start site (TSS) can be used to define putative promoter regions of genes. Cap Analysis of Gene Expression (CAGE) [301] is the most common method for the determination of TSSs. The CAGE method consists of capturing, sequencing, and mapping the 5' end of mRNAs in a biological sample to a reference genomic sequence, leading to TSS locations.

The TSS then defines putative promoter regions in the vicinity. Depending on the organism, various regions around transcriptional start sites are considered to comprise promoter regions. While there are reports of regulatory interactions between regulatory elements as far away from the TSS as 1Mbp [259, 179] or 100kbp [150], the majority of interactions with the basal transcription machinery seems to happen in a short range on the core promoter. For *Homo sapiens* Qian et al. [268] use a region of -1000 to 0 bp relative to the TSS, since

it was claimed to contain the highest density of known binding sites [369]. Xie et al. [365] find the the highest density of evolutionarily conserved motifs in the region between -500 and +200bp relative to the TSS. The ENCODE project found that 67% of real TFBSs are located within 2.5kb of TSS [63]. Tabach et al. [328] show that the region with the highest abundance of location-specific TFBSs in human and mouse extends from -200 to +100bp relative to the TSS[328]. The concentration of functional and conserved TFBSs close to the TSS suggests a limitation of the extracted sequence region to a few hundred base pairs to 1kbp to limit the fraction of false positive TFBS predictions.

Summary A plethora of experimental methods are available for the examination of transcription factor binding specificities and the localization of the respective sites. For a known transcription factor, technologies like SELEX or CAST can elucidate the binding specificity, while ChIP combined with a consecutive analysis of bound DNA produces the location of the binding sites. Elaborate methods involving the resolution of transcription factor structures are time-consuming, but carried out in cases of specific interest. If the transcription factor in question is not yet known, methods like DNase I hypersensitivity generate possible genome wide candidate regions for regulatory activity. The combination of EMSA and DNase I also resolves binding site locations. Reporter constructs make the examination of binding sites for known or unknown factors possible for smaller numbers of candidate sequences. In recent time large scale approaches using ChIP combined with modern sequencing technologies became feasible and prevalent. Detection of the transcriptional start site supports the definition of putative promoters in the vicinity of the TSS.

2.3.2 Collections of Binding Sites and Regulatory Regions

Transcription Factor Binding Sites Two bigger collections of transcription factor binding sites exist. At this time the TRANSFAC database [219] contains 885 different binding site profiles (version 2009.1) of different quality, many of which stem from the same factor. JASPAR [346] comprises a smaller, but non-redundant set of profiles (123 matrices in version 3.0). Various other projects collect profiles for specific organisms, like yeast [374, 142, 208], or different bacteria [210, 284]. The recently started PAZAR project [264] aims to unify various collections of experimentally determined transcription factor binding sites in an open-source and open-access fashion.

Collections of Regulatory Regions The *Saccharomyces cerevisiae promoter database* (SCPD) [374] annotates regulatory regions in yeast. New experimental data from genome-wide detection of DNase I hypersensitive regions in yeast provide a more complete and detailed picture of regulatory regions in yeast [35, 98].
One of the first collections of regulatory regions for higher organisms was the eukaryotic promoter database EPD. It started as a manually curated compilation of published promoter sequences [48] in 1986. The most recent version 100 [292] contains 4809 promoter sequences from various species, the majority from mammalia.
The *database of transcriptional start sites* (DBTSS) [349] is a collection of experimentally determined 5'-end sequences of full-length cDNAs for *Homo sapiens* and *Mus musculus*. The current release 6.0.1 contains approximately 19 million 5'-end sequences derived from

next-generation sequencing mapped to the human genome.
Another source of transcriptional start sites is the EnsEMBL database [153], which annotates transcripts for a large variety of organisms based on evidence from the EMBL nucleotide sequence database [62], the protein sequence collection UniProtKB [66], and the manually annotated mRNAs and proteins from NCBI RefSeq [355]. The FANTOM project provides collections of full-length cDNA sequences based on the CAGE technology [170, 64].

2.3.3 Experimental Methods Protein-Protein and TF-TF Interactions

Most experimental methods which provide information about transcription factor interactions are not specific for transcription factors, but are designed to detect general protein-protein interactions (PPI). There are approaches to screen PPIs and others to verify PPIs. Some methods detect interactions *in vivo*, others *in vitro*. Because of the normally low expression levels of transcription factors themselves and of methodological reasons the elucidation of interactons between transcription factors is generally more difficult than between other proteins. In the following we present the main approaches.

Screening Methods In the *yeast two-hybrid* (*Y2H*) experiment [106] the coding sequences of the DNA-binding domain and the activation domain of a transcription factor, often *GAL4*, are separately fused with a *bait* protein and potential interaction partners as *prey* proteins. If the bait and a prey protein interact, the separated domains come together and activate the transcription of a reporter gene which is expressed in case of an interaction. The method functions in *quasi in vivo* circumstances. The method has been fully automated, is highly sensitive but has many false positive predictions. Using Y2H for the elucidation of TF interactions is problematic, because the domains fused to bait and prey proteins possibly interact with other domains then their original partner, leading to less reliable results. Yeast two-hybrid screens exist for several organisms, for example for yeast [341, 159] or for human [320, 282].

Protein cross-linking works by forming covalent cross-links between lysine residues of proteins and can detect weak interactions [258]. Neighbouring proteins from a complex can lead to false-positive predictions. Cross-linking can be performed *in vitro* and *in vivo*.

Tandem Affinity Purification (TAP) [276] is a method to rapidly purify protein complexes under natural conditions such that one can identify the components of a complex using mass spectrometry. It comprises of a two-step purification after fusing a TAP tag with a target protein. TAP is not limited to binary interactions, but in some setups the false-negatives caused by low abundance and transient interactions are problematic.

High-Throughput Mass Spectrometric Protein Complex Identification (HMS-PCI) directly identifies protein complexes by mass-spectrometry [149]. A one-step immuno-affinity purification based on epitope tags allows to capture bait proteins, which are subsequently used for the immunoprecipitation of multi-protein complexes. One resolves the complexes using gel

electrophoresis with subsequent staining and cutting them out. Following a tryptic digestion, the proteins are identified using mass spectrometry and a comparison with MS-spectra from databases. Like TAP, HMS-PCI provides information about multi-protein complexes.

Phage display [311] is a purely *in vitro* high-throughput method. In phage display one presents proteins on the outside of a bacteriophage by fusing them to the respective coat proteins. One identifies a protein's binding partners from large recombinant phage display libraries. The displayed proteins bind to the protein of interest. Subsequent washing removes phages with non-binding proteins on the surface. The system imposes size limits on the proteins which have to be checked, and only works for proteins which can be secreted to the surface of the bacteriophage.

Verification Methods In *Co-immuno-precipitation* (coIP) [261] a specific antibody binds to the protein of interest in a cell lysate, followed by a precipitation using affinity beads. After washing, one analyzes the protein and its interaction partners by *western blotting* or immuno-detection. CoIP functions *in vivo* or *in vitro*. Problems can arise due to indirect interactions and sometimes due to low sensitivity.

In *pull-down assays* [261] one attaches a bait protein to a matrix (usually as a fusion protein). One puts a cell-lysate containing possible interaction partners for the bait onto the matrix. After washing, interaction partner enrichment takes place due to their direct binding to the bait, and thus indirect binding to the matrix as well. The method shows direct physical interactions, and requires pure proteins and large amounts of the bait.

Sub-cellular immuno-fluorescence co-localization [46] works by fixation and permeabilization of cells expressing two proteins of interest. Specific primary antibodies bind to the respective poteins in the cells. Secondary antibodies coupled to fluorescent dyes bind to the primary antibodies. One determines cellular localization of the proteins by fluorescence microscopy.

Summary While the methods to screen and verify interacting proteins do not account for their potential transcriptional activity, some of the experimental methods for the detection of binding sites for individual transcription factors are also useful to obtain information about potentially interacting transcription factors. For example, promoter analysis can deliver information about various binding sites in the upstream region of the reporter gene. Further, for a small set of transcription factor complexes structures derived using X-ray and NMR methods exist.

2.3.4 Collections of Protein and Transcription Factor Interactions

The availability of high-throughput data amended the knowledge about general protein-protein interactions (yeast [341, 159], human [320, 282, 218], comparison of multiple species [118]). *Interactome*-Databases containing protein-protein interaction information have the advantage of putting transcription factors into a bigger context of a network of interacting proteins. Examples for manually curated databases with interaction data derived from

scientific literature are MPact (yeast) [134], MPPI (mammalia) [249, 226], or DIP (various organisms, hand-curated and automatic annotation) [285]. Early large-scale databases for protein-protein interactions rely on automated literature parsing [40, 244, 73]. Later homology-based methods added to the information available for protein-protein interactions [257, 213, 156].

Note that the results from different large-scale experiments expose a small overlap only (yeast [14], human [114]). Futschik et al. [114] assume that the reason for the small overlap is to be sought in selection and detection biases and the inability of the yeast two-hybrid method to detect protein modifications. Databases like STRING [347] or UniHI [58] condense the methods and data sources mentioned before.

2.3.5 Summary

Although large collections of protein and transcription factor interactions exist, they do not cover the complete interactome. The reasons lie in dynamic changes of protein interactions over the time, experiments specific to cell types, or experimental bias. Large scale protein interaction studies often cover a subset of the transcription factors present in an organism. An overview of methods for the computational prediction of protein- and specially TF interactions is presented in Section 3.2.2.

Chapter 3

Computational Prerequisites

3.1 Computational Prediction of Transcription Factor Binding Sites

3.1.1 Models

Transcription factors commonly show a binding preference to certain DNA sequences. Experimental evidence for binding sites allows to build a model that represents the binding specificity of a transcription factor.
In this section we describe how to get from a set of experimentally derived binding sequences to consensus sequences and to position-specific score matrices, and how to use them for the detection of new binding sites.
The concept to use position weight matrices to represent transcription factor binding sites was introduced by Stormo [321]. We use the methods of Rahmann et al. [270] to search for binding sites.

Multiple Alignment of Binding Sites For building any model of a transcription factor binding site, we require a multiple alignment of experimentally determined binding sites. TFBSs are usually short and the number of binding sites normally is relatively low. Hence standard methods for multiple alignment, such as `Clustal W` [332] are applicable. For an overview of more recent methods see Notredame [237].

Transcription factors usually show some variability of their binding preferences. In a set of experimentally determined binding sites for the same factor, one can normally observe positions with high conservation as well as positions with a certain variability.

For an example of aligned binding site sequences see Figure 3.1 a). The figure shows an alignment of seven (of 31 in the respective TRANSFAC [219] entry) experimentally verified binding sites for the transcription factor NF-AT (M00935 / V$NFAT_Q4_01). The sites originate from different publications and experimental methods, for example crystallization [59], DNase I footprinting [80, 281], gel shift, functional assay [164, 90].

Profiles Observing a nucleotide in one position of a binding site usually does not influence the probability of observing a nucleotide in another position. Only insignificant dependencies between positions could be shown [294, 322]; thus it is appropriate to assume independence of sequences positions in binding sites. The models here are therefore position-independent. A profile is a probabilistic description of a sequence set. Assume a finite alphabet $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. Let $\pi = (\pi_j)_{j\in\Sigma}$ be a probability distribution over the letters of the alphabet Σ. We consider π as a vector with $|\Sigma| = 4$ non negative components such that $\sum_{j\in\Sigma} \pi_j = 1$. It is a probability distribution for the letters of Σ for each sequence position. Given a sequence of length L and an alphabet Σ we present a profile P as an $L \times P$ matrix $(P_{ij})\,(i = 1, \ldots, L; j \in \Sigma)$, such that $P_{ij} \geq 0$ for all positions i and letters j and $\sum_{j\in\Sigma} P_{ij} = 1$ for all positions i.

Position-specific Count Matrix / Position-specific Frequency Matrix The next step, common to the two models presented here, is the building of the position-specific count matrix (PSCM) of dimensions $L \times \Sigma$. It contains the counts $\kappa_{i,j}$ for each nucleotide j in each position or row i of the multiple alignment:

$$C = \begin{pmatrix} \kappa_{1,1} & \cdots & \kappa_{L,1} \\ \vdots & \ddots & \vdots \\ \kappa_{1,4} & \cdots & \kappa_{L,4} \end{pmatrix}$$

Each row of the PSCM C sums up to the total number of sequences N in the multiple alignment.

The profile or position-specific frequency matrix (PSFM) has the same dimensions as the PSCM. It contains a maximum likelihood profile of the multiple alignment or frequency of each nucleotide j at each position i. We calculate the frequency ϕ_{ij} as the fraction of the counts of nucleotide j at position i, and the sum of nucleotide counts at position i:

$$\phi_{ij} = \frac{\kappa_{ij}}{\sum_{j\in\Sigma} \kappa_{ij}} \tag{3.1}$$

The PSFM F is then defined as

$$F = \begin{pmatrix} \phi_{1,1} & \cdots & \phi_{L,1} \\ \vdots & \ddots & \vdots \\ \phi_{1,4} & \cdots & \phi_{L,4} \end{pmatrix}$$

Figure 3.1 b) shows an example for a position-specific count matrix, derived from the 31 sequences of TRANSFAC entry V$NFAT_Q4_01.

Consensus Sequences Two concepts exist for the definition of a consensus sequence. The *biochemical consensus* is defined as the single sequence variant with the highest affinity for the transcription factor *in vitro*. The *informatic consensus* is defined as a representative sequence where the nucleotide at each position is the most abundant nucleotide with the largest $\kappa_{i,j}$. For example, the informatic consensus for the PSCM in Figure 3.1 b) is `GTGGAAAATC`.

For the informatic consensus, sequences based on IUPAC degeneracy nucleotide symbols [67, 78] are a more suitable representation, because they can represent ambiguities to a certain degree. IUPAC symbols in Table 3.1 represent alternative choices of nucleotides for a given position in the alignment. The representation of a profile as a consensus sequence implies a loss of information due to discretization.

Symbol	Meaning
A	A
C	C
G	G
T	T
U	U
M	A or C
R	A or G
W	A or T
S	C or G
Y	C or T
K	G or T
V	A or C or G
H	A or C or T
D	A or G or T
B	C or G or T
X	G or A or T or C
N	G or A or T or C

Table 3.1: IUPAC-IUB symbols for nucleotide nomenclature

The IUPAC consensus for the count matrix for V$NFAT_Q4_01 is `NWGGAAANWB`.

To search for binding sites using a consensus sequence simply relies on a search for the described pattern. If one assumes the known binding sites to be a subset of the real binding specificity of the transcription factor, one sometimes allows for one or more mismatches [269].

Position Weight Matrix To decide whether a sequence $T = (t_1, \ldots, t_L)$ of length L, one calculates a score describing the similarity of T to P in contrast to the background probability distribution on Σ, given by the vector $\pi_b = \pi_b(j), j \in \Sigma$. The background describes medium- or large-scale properties of the genomic sequences under scrutiny, for example the GC-content. The most simple background distribution is the uniform distribution $\tau = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$, where one assumes that the nucleotides in the background are i.i.d.

with τ.

Consider the frequencies $\phi_{i,j}$ of the PSFM F as probabilities to observe letter j at position i, under the assumption that $T = (t_1, \ldots, t_L)$ was generated by P.

The probability that a profile P generates a fixed sequence $T = (t_1, \ldots, t_L)$ then is:

$$Prob[T|P] = \prod_i \phi_{i,s_i} \tag{3.2}$$

We use the likelihood-ratio of the probability of T being generated from the profile or the background model to decide, whether a sequence is a binding site or not. Consider the binding site profile P and a background profile matrix Π_b of the same length with each row consisting of the same probability vector π. The log-odds score is the log-ratio of the probabilities that a sequence T is generated from foreground profile P and background profile Π_b:

$$S(T) := \log \frac{Prob[T|P]}{Prob[T|\Pi_b]} = \sum_{i=1}^{L} \log \left(\frac{P_i(T_i)}{\pi_b(T_i)} \right) \tag{3.3}$$

The score $S(T)$ is > 0 if it is more likely that T was generated from the profile, and < 0 if it is more likely that T was generated from the background.

With a fixed background distribution π, one calculates the position weight matrix PWM, by setting position and nucleotide specific scores $S_{i,j} = \dfrac{P_{i,j}}{\Pi_{bj}}$ with $(i = 1, \ldots, L)$ and $(j \in \Sigma)$:

$$PWM = \begin{pmatrix} S_{1,1} & \ldots & S_{L,1} \\ \vdots & \ddots & \vdots \\ S_{1,4} & \ldots & S_{L,4} \end{pmatrix}$$

To build the PWM, instead of using the raw nucleotide counts from the PSCM we apply position-specific *regularization* by adding pseudo-counts to the counts as described in Rahmann et al. [270]. The regularization prevents the rejection of a previously unobserved nucleotide in a given position, resulting in a higher generalization ability of the profile.
The described score sums the individual contributions of every position i, consisting of $\log \left(\frac{P_i(S_i)}{\pi(S_i)} \right)$. Using biophysical models of DNA-protein interactions, it was shown that this term correlates with the contribution to the total binding free energy of the individual position [32, 105].

The position weight matrix for TRANSFAC entry V$NFAT_Q4_01 in Figure 3.1 c) shows the log-odds scores calculated using regularization and assuming the uniform background distribution.

Sequence Logo Sequence logos are a common way to illustrate DNA profiles. A sequence logo has the same length as the respective profile and stacked nucleotide symbols of different heights proportional to the Shannon information content at the respective position [293]. Assuming a uniform background distribution the information content for a position i is calculated using the probability $P_j(i)$ of observing nucleotide j at this position:

$$D(i) = \log_2 |\Sigma| + \sum_{j \in \Sigma} P_j(i) \cdot \log_2 P_j(i) \tag{3.4}$$

The higher the preference of a transcription factor for a nucleotide at a given position is, the higher is the information content at that position. The lower the conservation, the less specific the factor is at that position, leading to small heights at the position in the logo. For the nucleotide alphabet consisting of four letters, the maximum information content of a position is 2bits.

The logo representation of the NF-AT transcription factor is shown in Figure 3.1 d). The positions which have an exclusive preference for one nucleotide in the count matrix have an information content of 2 bits (position 3 and 5). On the other hand positions with a non-specific distribution of nucleotides like position 1 have a small information content.

Searching for Binding Sites The score describing the similarity of a piece of sequence to our profile enables us to search for new binding sites by calculating the score for uncharacterized sequences.
Let $X = S(W)$ be the score of a sequence W. Furthermore we define $\mathbb{P}_P$ and $\mathbb{P}_{\Pi b}$ as the two probability distributions for the signal model P and the background model Π_b associated with W. We represent the probability of W being generated by the signal profile by $\mathbb{P}_P$ and the probability of W being generated by the background by $\mathbb{P}_{\Pi b}$. To distinguish between the two cases we employ a statistical test.
Under the *null hypothesis*, one assumes that the sequence W was generated from the background distribution (the score X is distributed according to Π).

The *alternative hypothesis* H_1 is generation of W by the signal profile (X is distributed according to P). As a *test statistic* we use the log-odds score $S(W)$ described in Equation 3.3. We reject H_0 if $S(W)$ is greater or equal a given threshold t.

Figure 3.2 shows the signal distribution in red and the background distribution in blue. If the score $S(W)$ for a sequence W is greater or equal the threshold (vertical black line), one rejects the null hypothesis and regards the sequence as a binding site.

Except for the correct predictions two kinds of errors (marked in red) can occur:

	prediction: true	prediction: false
positive	true positive (TP)	false positive (FP)
negative	false negative (FN)	true negative (TN)

a) alignment of binding sites

site	experimental binding sequence
1	GAGGAAAAAC
2	GAGGAAAATT
3	GAGGAAAAAC
4	GAGGAAATGA
5	CTGGAATTTC
6	TTGAAAATAT
7	GCGGAAACTT
	...

b) position-specific count matrix (PSCM)

pos	A	C	G	T
1	6	7	12	6
2	9	5	1	16
3	0	0	31	0
4	1	0	30	0
5	31	0	0	0
6	29	0	2	0
7	27	2	0	2
8	12	8	4	7
9	9	4	2	16
10	3	11	5	12

c) position weight matrix (PWM)

pos	A	C	G	T
1	-0.153	-0.149	0.410	-0.258
2	0.161	-0.445	-1.857	0.708
3	-4.940	-6.173	1.383	-5.706
4	-1.973	-5.758	1.349	-5.292
5	1.383	-5.885	-5.030	-5.418
6	1.314	-5.374	-1.323	-4.907
7	1.242	-1.345	-4.319	-1.332
8	0.444	-0.030	-0.539	-0.119
9	0.164	-0.664	-1.240	0.702
10	-0.865	0.332	-0.418	0.423

d) sequence logo

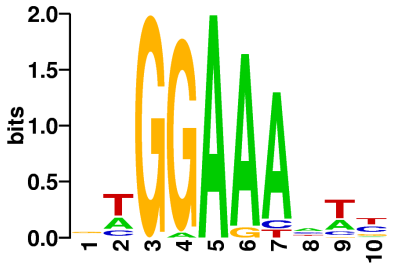


Figure 3.1: From nucleotide counts to the score matrix - An example binding site description from TRANSFAC for the transcription factor NF-AT (V$NFAT_Q4_01 / M00935) in different representations: **a**) 7 out of 31 aligned experimentally verified binding sites for NF-AT from the TRANSFAC database. The different sequences show some degree of variability. Note that for example site 1 and site 5 have only 50% of the nucleotides in common. **b**) The position-specific count matrix (PSCM). It contains the numbers of different nucleotides found at each position of the alignment of the experimentally determined sites. The consensus sequence is GTGGAAAATC. The IUPAC consensus sequence is NWGGAAANWB. **c**) The score matrix / position weight matrix (PWM) for NF-AT using a uniform nucleotide distribution ($\pi_A = \pi_C = \pi_G = \pi_T = \frac{1}{4}$) as background model. The probabilities of nucleotides have been regularized using the methods of Rahmann et al. [270] to prevent zero entries. **d)** Sequence logo of the motif for V$NFAT_Q4_01. The horizontal axis represents the position in the motif. The height at each position is proportional to the total bits of information at the position in the sequence. Positions with a high information content deviate strongly from the background, that is, the position contains mostly one specific nucleotide. We created the logo using the tool WebLogo [71].

The *type I error* or *false positive* (FP) means that H_0 is rejected although it is true. This happens when a sequence W generated from background appears to be generated from the signal profile due to a log-likelihood $S(W) \geq t$. This situation corresponds to the blue area under the blue curve in Figure 3.2. The *type II error* or *false negative* (FN) occurs in case of acceptance of H_0 although it is false. A sequence W generated from the signal appears to be generated from the background distribution due to a log-odds score $S(W) < t$. This situation corresponds to the red area under the red curve in Figure 3.2.

Once the score distributions are known, the calculation of fixed thresholds corresponding to error levels becomes possible. The three main variations are a fixed type I, a fixed type II, and a balanced error level, in which the type I and the type II error are equal.

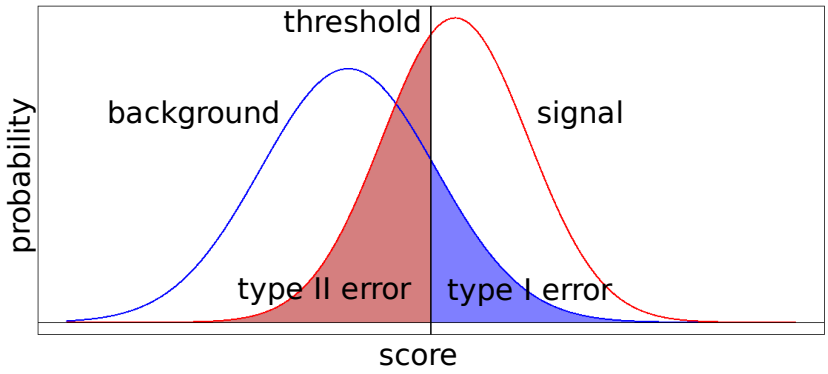


Figure 3.2: An example for score distributions of the background (blue) and signal profile (red). The threshold determines the score above which a a sequence W is regarded as positive. The type I or false positive error occurs if a score generated from the background profile is bigger than the threshold (light blue area). The type I or false negative error happens in case a score generated from the signal profile is smaller than the threshold.

Searching for binding sites usually involves calculating the scores for overlapping windows which are not independent. In this case the calculation of the exact type I and type II errors is complex and requires approximations.
Using the easier to calculate window level errors, Rahmann et al. [270] calculate the sequence level errors under an independence assumption for the windows.
Various approaches for the efficient calculation of score distribution have been developed (e.g. McLachlan [221], Staden [319], Tatusov et al. [329], Wu et al. [364], Rahmann et al. [270], Beckstette et al. [27]). For our project we use the methods developed by Rahmann et al. [270], and scan for putative binding sites at fixed false positive error rates of 0.05, 0.01, and 0.005, and the balanced cutoff, at which the false positive and the false negative error rates are equal to each other.
Different alternative threshold-based approaches exist. Kel et al. [172] use predefined thresholds, while Hertz and Stormo [146] and Turatsinze et al. [340] apply variable background models for the calculation of appropriate score thresholds.

We illustrate the actual search for putative binding sites in Figure 3.3. The log-likelihood score described before is calculated for every window of a sequence. When the score exceeds a predefined threshold representing a fixed error rate, we regard the window as a putative binding site.

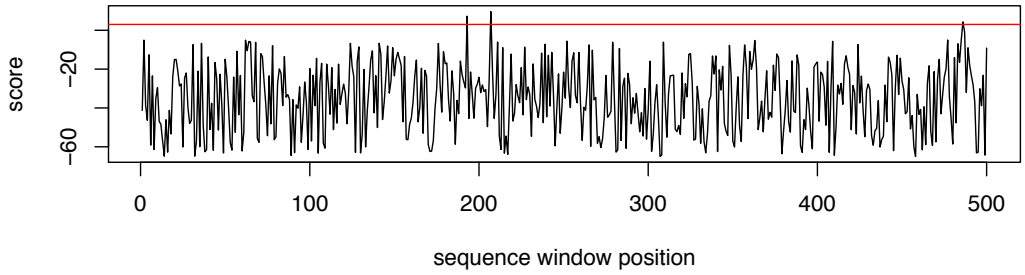


Figure 3.3: To determine potential transcription factor binding sites given a sequence and profile, one calculates a log-likelihood score (y-axis) for every sequence window (sequence window positions on the x-axis). The threshold (red line) represents a fixed error level. One considers windows with a score above the threshold as potential transcription factor binding sites. In the given example we show three binding sites in locations close to 200bp and shortly before 500bp.

3.1.2 Application and Problems

Typical transcription factor binding sites are short and degenerate. Hence in a genome wide prediction one expects a plethora of predicted binding sites including many false positives. For example, Wasserman and Sandelin [352] coined the term *futility theorem*: for the muscle-specific transcription factor *myoD* they find 1000 false-positive binding site prediction for every true-positive if carried out for the whole human genome.
However, the problems leading to the high number of false positive predicted binding sites do not primarily lie in insufficient prediction methods: For the transcription factor *HNF-1* Tronche et al. [334] tested the *in vitro* binding of predicted binding sites, and could show that the majority of predicted TFBSs could be bound by the transcription factor. Regulation of transcription is dependent on a plethora of other levels which determine, whether a transcription factor can bind to its preferred motif on the DNA: competition of factors [128], chromatin structure [50, 295], DNA-, chromatin-, histone modifications, alternative splicing of RNA, mRNA stability, translational controls, post-translational modifications, or presence of interaction partners [322]. Transcriptional regulation is context-dependent. Given a basically functional TFBS, various circumstances can prevent the binding site to be used: The factor itself is not present in proximity of the binding site; an adjacent binding site is occupied with another factor causing sterical hindrance; the transcription factor is present but inactive; a different factor has a higher affinity for the binding site; cofactors are missing or altering binding specificity of the transcription factor.

To improve prediction quality, several properties of regulatory regions are used: Promoters are located close to transcriptional start sites (TSSs). If these are known by experiments or a reliable prediction is available, one can restrict the search for TFBSs to regions surrounding the TSSs. Moreover regulatory regions have been found to be under evolutionary constraints. In plenty of examples, regulatory regions are conserved over several related species. The usage of conservation information in the context of regulatory regions and transcription factor binding sites is called *phylogenetic footprinting / shadowing* [91, 10, 82, 175, 41]. For a set of skeletal-muscle specific transcription factors, Wasserman et al. [351] found that 98% of experimentally defined binding sites are located in the 19% of non-coding sequence most conserved between human and mouse. In the same study, for the factor *SP1*, which has a general preference for `GC` rich sequences and a low specificity, still 75% of verified binding sites lie in the conserved blocks. Other properties of regulatory regions exploited to overcome the false-positive problem include sequence composition, like the GC content, or the presence of CpG islands or experimental data about histone modifications and DNase hypersensitivity.

Usually the modelling and prediction of transcription factor binding sites is just a first step in series of computational experiments to explain regulation of individual or sets of genes. A common strategy is to apply some sort of enrichment analysis to the detected TFBSs. An early example for the construction of position weight matrices can be found in Bucher [47]. Examples for the successful application of TFBS prediction include the analysis and prediction of muscle- [350] and liver-specific [190] regulatory regions or the detection of transcription factors with importance for the cell-cycle [83]. For a review of several methods for the prediction of transcription factor binding sites and applications of position weight matrix based methods see Bulyk [52].

3.2 Computational Prediction of Transcription Factor Interactions

3.2.1 Prediction of Protein-Protein Interactions

In this section we present methods for the prediction of protein-protein or domain-domain interactions. The presented methods here are not specific to the domain of TFs. For a broader coverage of methods for the prediction of protein-protein interactions see the reviews of Shoemaker and Panchenko [302] and Skrabanek et al. [308]

- **gene neighbor / cluster methods** harness that in bacteria genes with related functions often lie on the same operon, are transcribed together, and are candidates for physical interaction partners. Despite the evolutionary distance the respective homologues in eukaryotes they are often co-regulated. The works of Ermolaeva et al. [99] and Moreno-Hagelsieb and Collado-Vides [227] are examples for operon-prediction in bacteria. Teichmann and Babu [330] found that gene order is not only useful in bacteria but also in the eukaryotes yeast and and worm.

- **phylogenetic profile methods** assume that that potentially interacting proteins co-evolve and have orthologs in the same set of organisms [257, 93, 312]. A phylogenetic profile consists of a presence-absence indicator for each protein in each of a set of organisms. Protein profiles close to each other after clustering are putative interaction partners [43, 44]. Pagel et al. [248] used phylogenetic profiles for domain-domain interaction prediction.

- The **Rosetta Stone method** also uses sequence information from different genomes. The method is based on the fact that two proteins which were fused into one protein in another genome often interact [213, 97, 212].

- **Sequence-based co-evolution methods** use the correlation coefficient of the distance matrices of two protein families as a measure of similarity between the phylogenetic trees. A high similarity is likely due to co-evolution of the two families and thus mark potential interaction of the proteins [125, 254, 166].

- **Classification methods** use various data sources, such as InterPro-based sequence signatures [316, 317], Pfam domain assignments [60], amino acid triplet occurrences [298], or sequence signature products [216] to train classifiers like support vector machines [187], bayesian networks, [162] or random forests [60] for the separation of interacting and non-interacting protein-pairs. Fang et al. [100] and Neduva et al. [235] developed methods to mine for motifs in sets of proteins that share common interaction partners. Qi et al. [267] provide an overview of classification methods.

General methods for protein-protein interaction prediction usually result in information only for a small set of transcription factors. Moreover the biological features used for protein-protein interactions in general as presented above are not necessarily as pronounced as needed for reliable predictions.

3.2.2 Prediction of Transcription Factor Interactions

In the following we present methods that were developed specifically for the prediction of transcription factor interactions.
Methods for the prediction of protein interactions which are based on the amino acid sequence of a protein sometimes are problematic when applied to transcription factors. The repertoire of TFs in eukaryotes likely originates from a small set of conserved super families [13]; Harrison estimates the amount of DNA-binding domains to 12-15 in eukaryotes [143]. The different members of transcription factor families arose from duplications [331].
Due to the resulting similarity in the amino acid sequence of TFs, most of the approaches for the prediction of TF interactions do not rely on the amino acid sequence but on joint binding of two transcription factors instead. As source of binding information the methods use either experimental data (see Section 2.3.1) or predictions of transcription factor binding (see Section 3.1).

Interaction Prediction Using Experimental Transcription Factor Binding Data

The most common source of TF binding data used for prediction of TF interactions is chromatin immuno-precipitation. Other experimental datal like expression patterns, usually can improve the prediction of interactions. Banerjee and Zhang [25] use cell-cycle expression data [61] together with genome-wide location data for transcription factors in yeast, derived from ChIP chip experiments [195] to calculate cooperativity of transcription factors. They devise an expression coherence score, which describes the correlation of expression patterns of genes in sets bound by two different factors, and the expression patterns of genes in a set bound by both factors.
Nagamine et al. [233] analyze the same TF binding data for yeast [195] but complement it with protein-protein interaction data [367]. The underlying assumption of Nagamine's method is that proteins close to each other in the interaction network are more likely to be co-regulated by the same set of TFs. Thus for a set of genes co-regulated by two TFs the distances in the PPI network are expected to be significantly lower than the distances of the set of genes regulated by the individual factors.

Interaction Prediction Based on Predicted Transcription Factor Binding Data

A wide variety of methods predicts TF interactions without large-scale experimental binding data from ChIP experiments. Often the methods use expression data in combination with predicted TFBSs, either based on known TF motifs, on overrepresented motifs, or based on a combination of the two.

GuhaThakurta and Stormo [131] developed the method Co-Bind. It identifies motifs for transcription factors which cooperatively bind close to each other in the promoters of small co-expressed gene sets. Co-Bind applies a Gibbs sampling strategy and maximizes the joint likelihood of occurrence of two TFBS motifs. GuhaThakurta and Stormo applied the method on small sets of co-regulated yeast and E. coli genes and show that the detected motifs belong to the TFs which are known to regulate the respective gene sets.

Pilpel et al. [262] annotate yeast upstream regions for genes in different functional categories from the MIPS [224] database with known TF motifs and additionally derive motifs using AlignAce [154]. Their method identifies motif-pairs that co-occur in large numbers of promoters from a set of genes with correlated expression patterns to calculate the probability of obtaining the observed or a higher co-occurrence rate for two motifs. Sudarsanam et al. [325] apply the method to identify interactions of rRNA related transcription factors.

Beer and Tavazoie [28] use a Bayesian approach to map regulatory sequence elements to expression patterns. To start, they cluster microarray expression data to obtain gene sets with distinct expression patterns. Afterwards they annotate overrepresented motifs also using AlignAce [154] and known motifs [194]. The bayesian network inferred subsequently describes the sequence elements and their combinatorial constraints needed for a gene to be expressed with a specific pattern.

The tool CisModule developed by Zhou and Wong [373] uses a hierarchical mixture-model, that simultaneously infers PWMs and cis-regulatory modules (CRMs). It consists of two mixture levels — the first mixing CRMs and background and the second mixing motifs within the CRM and a CRM-internal background. The authors apply the method on homotypic clusters in *D. melanogaster* and on human muscle-expressed genes.

Das et al. [74] apply non-linear models on cell-cycle expression data and annotated TFBS motifs from several other publications [262, 174]. The authors use multivariate adaptive regression splines to correlate binding site occurrences and pairs thereof to the ratio of gene expression levels.
Kato et al. [169] combine experimental data from ChIP experiments and gene expression with predicted TFBS motifs to identify combinatorial regulation. The statistical test applied is based on a contingency table with the number of sequences from a foreground- and a background set containing the respective motif and combinations. The authors define the respective sets using the ChIP data.

Elkon et al. [94] identify potential synergistic partners for a specific transcription factor by searching for overrepresented binding sites of other factors and their combinations in known targets. The program PRIMA determines co-occurring pairs of PWMs calculating a hyper-geometric p-value for observing a number f_{ab} or more promoters containing hits for two PWMs, with f_a and f_b as number of promoters with hits for PWMs P_a and P_b in m promoters in total:

$$p = \sum_{h=f_{ab}}^{min(f_a,f_b)} \frac{\binom{f_a}{h}\binom{m-f_a}{f_b-h}}{\binom{m}{f_b}} \tag{3.5}$$

The authors apply PRIMA to find transcription factor interactions in the human cell cycle.

Zhu et al. [375] combine known TF motifs, *de novo* motif finding, phylogenetic footprinting and microarray data to predict interactions. First they define anchor TFBSs based on known motifs and subsequently use AlignACE [154] for pattern finding in regions around the anchors. The method calculates a *neighbor specificity* score based on the binomial distribution. The probability of getting the observed number of promoters or more containing the anchor and neighboring motif is

$$P = \sum_{i=x}^{Na} \binom{N_a}{i} p^i (1-p)^{N_a - i} \tag{3.6}$$

with N_a as the number of promoters with the anchor motif, and x the number of promoters with the anchor and the neighbor motif. The probability p of a random anchor-containing promoter to also contain the neighbor motif is approximated as $p \approx \frac{N_n}{N_t} \times \frac{W}{L} \times C_a \times C_n$. Here N_n is the number of promoters with the neighbor motif, N_t the total number of promoters, W the window size, L the promoter length, C_a the average copy number of anchor motifs in an anchor containing promoter, and C_n the respective copy number for neighbor containing promoters. Afterwards the authors test significant combinations for functionality by means of the expression-coherence score of Banerjee and Zhang [25] (see above). Zhu et al. [375] apply the method on human cell cycle data and detect various TFs whose binding sites occur in homotypic pairs and a set of heterotypic combinations with functional importance

in cell-cycle.

Yu et al. [368] present a method which detects overrepresented motifs that are important for subsets of genes. Their program Motif-PIE first identifies the most overrepresented single motifs and subsequently enumerates all possible pair combinations between the top motifs. The authors calculate a combined p-value consisting of an over-representation term P_{occ} and a distance constraint P_d. P_{occ} is a hyper-geometric p-value with g as the number of motif pairs occurring in the input, G the genome-wide occurrence of a pair, n the number of input promoters and N number of genome-wide promoters:

$$P_{occ} = \sum_{k=g}^{min(n,G)} \frac{\binom{n}{k}\binom{N-n}{G-k}}{\binom{N}{G}} \tag{3.7}$$

The authors derive the distance p-value from a Kolmogorov-Smirnov test that compares the observed distance distribution of the two motifs with a background distribution from motif pairs that do not interact. Applying Motif-PIE on yeast [368] and human data [370], the authors demonstrate that some factors have different interaction partners depending on the conditions or cell-types.

Other attempts also take into account distance preferences of predicted binding sites towards each other: FitzGerald et al. [108] determine the position of all possible DNA 8-mers relative to human transcription start sites and find known TF motifs overrepresented in specific positions in distance histograms. For a subset of motifs they also identify specific distances between their occurrence, which might be due to cooperativity of the factors. Based on known motifs Vardhanabhuti et al. [343] exploit position specificity of predicted TFBS towards the transcription start site and each other to identify cooperating transcription factors. Assuming binomial distributions the authors apply Z-scores to quantify motif conservation, positional specificity and motif-pair distance specificity. Apart from finding 25% of motifs with a distance specificity relative to the TSS, Vardhanabhuti et al. [343] show that a part of the TFBS pairs with distance preferences between each other show significant functional associations.

Several other methods count TFBS pairs present on a sequence within a maximum distance. To assess the significance of a number of co-occurrence events these methods compare the observed pair count with an expected pair count, derived in different ways.
Hannenhalli and Levy [141], Levy et al. [198] define a co-localization index for TFBS pairs. They first annotate TFBS genome-wide and then count TFBS pairs with a maximum allowed distance. The co-localization index CI for two PWMs i and j is

$$CI = \frac{N_{ij}}{R_{ij}} \tag{3.8}$$

with N_{ij} as number of co-localizations for two TFBS i and j within w bp in sequence and R_{ij} number of co-localizations within w bp in sequence set after perturbation of the binding site labels (note that in Levy et al. [198] the score is defined as a log-odds score). Applying about 250 PWMs from TRANSFAC to the whole human genome they find some TFs tending to occur in homotypic pairs, a correlation of the CI and TF synergism, as well as

a correlation of PWM similarity and the CI. The counting procedure does not explicitly take into account problems which can arise due to homotypic clusters of TFBS. A window containing multiple TFBSs of the same type also results in multiple TFBS pairs of that type. The background model preserves the density of TFBSs and amount of TFBSs for each type.

Rateitschak et al. [272] annotate TFBS in conserved regions of human promoters. They count the number of promoters in which a hit for two PWMs i and j occurs within a maximum distance at least once. The authors calculate a log-odds score S_{ij} for co-occurrence as

$$S_{ij} = \log \frac{m_{ij}}{\pi_i \pi_j} \tag{3.9}$$

with the pair probability $m_{ij} = \frac{c_{i,j}}{\sum_{i,j} c_{i,j}}$ and the associated marginals of the co-occurrence count matrix $\pi_i = \frac{\sum_j c_{i,j}}{\sum_{i,j} c_{i,j}}$. The expected number of promoters arises from the number of promoters containing the individual TFBSs. It does not preserve the binding site density from the original data and implicitly models the binding site locations as uniformly distributed.

Hu et al. [152] predict TF interactions by identifying enriched TFBS combinations in the same promoter. They calculate a log-odds pair enrichment score to compare the frequency of a pair in the promoter set with the frequency in a background set derived by shuffling the bases of the original sequences and subsequent re-annotation with transcription factor binding sites. The authors combine the pair enrichment score with a hyper-geometric distribution based gene enrichment score to provide a link to biological function of TFBS pair targets. The authors count all combinations of TFBSs within a iteratively growing maximum distance and do not account for multiple counting of pair types caused by homotypic clusters. The background model does neither preserve binding site density nor the number of individual binding sites, since the permutation procedure operates on the bases of the DNA and potentially destroys basepair combinations typical for regulatory DNA.

Haiminen et al. [135] assess different counting methods for co-occurring TFBS and various background models that arose from some of the publications presented above and the method that we present in Section 4 (uniform, shuffling of TFBS labels). They develop an extended background model (shuffling of TFBS labels, while keeping the positions of one TFBS type constant). Using simulated and real data, the authors show that the simple uniform background model yields a higher false-positive rate than the more complex density-preserving models.

Pape and Vingron [251] and Pape et al. [253] proposed a method for the significance calculation of TFBS co-occurrence events based solely on the contents of PWMs and GC content of the sequence. The method calculates the probability of co-occurrence events of two PWM hits in a random sequence of a given length. The method considers similarity of the PWMs and does not need to assume position independence of the PWM occurrences. The authors derive a count distribution for co-occurrence events and thus compute the significance of TF cooperativity. The cooperativity is computed as a p-value for the number of observed

co-occurrence events using the binomial distribution. Moreover they present a way to calculate the window size given the co-occurrence probability of two PWMs. Pape et al. [253] show an approximation based on the Chen-Stein method for the calculation of co-operativity p-values in overlapping windows.

3.2.3 Summary

A variety of methods for the prediction of protein-protein and TF-TF interactions exists. The general PPI prediction methods often exhibit a moderate performance when applied to TF interactions [60, 316, 317, 213, 125, 254, 43]. Most methods for the prediction of TF interactions make use of predicted TFBS (based on known motifs [74, 169, 94, 272, 152, 135, 251, 141], pattern finding [131, 373, 375, 108], or both [262, 325, 28]) or experimentally determined [25, 61, 233, 367, 169] binding locations of transcription factors. Some of the methods amend the TF binding data with other sources like expression data [233, 25, 131, 262, 28, 375, 368, 370, 74, 94], sequence conservation [375, 272], or PPI networks [233]. The approaches comprise methods like Gibbs sampling [131], Bayesian inference, hierarchical mixture models [373], and statistical tests, for example based on the hypergeometric [25, 262, 94, 368, 152] or binomial distribution [375, 343, 251]. Methods counting the co-occurrence of predicted TFBSs differ in their way of counting and in the proposed background models. Some count TFBS pairs once per sequence [272], some ignore multiple occurrence of the same binding site type and homotypic clusters [141, 152], some take care for overlapping TFBSs [152]. The methods apply either empirical background models based on permutations [141, 135, 152] or estimate expected number of co-occurrences from the individual TFBS counts [272] or from complex statistical models [251, 253].

3.3 Computational Prediction of Regulatory Regions

In this section we first present an overview of properties of regulatory regions that can be exploited for their prediction. Subsequently we review the main prediction approaches. Some of these are specific to promoters only, others are applicable to enhancers as well.

3.3.1 Properties of Regulatory Regions

In silico prediction of promoter regions is a difficult task [3]. Although it is known that regulatory sequences differ from other regions of the genome [255, 6], they lack universal structural features that are present in coding sequences with open reading frames or a codon bias [361]. The evolutionary aspects of regulatory regions are not yet completely understood [361], nevertheless besides conservation other evolutionary aspects appeared useful [91].

One of the main properties of regulatory regions is that transcription factors bind to it. When no experimental data about transcription factor binding is available, a common strategy is to predict TFBSs. As described already in Section 3.1.1, TFBSs are short and degenerate, which complicates the application of predicted sites for the detection of regulatory regions. Still, the presence of specific motifs like the `TATA` box, the downstream

promoter element [53], or a high density of predicted motifs [361, 12] helps in the recognition of regulatory regions. It has been known for a long time that a big part of, but by far not all eukaryotic promoters contain a TATA box roughly 25bp upstream of the transcriptional start site [202]. The fraction of promoters with a TATA box varies between species. This motif is abundant in yeast [324] and less abundant in mammals [327]. Another characteristic feature of regulatory regions is the increased GC content compared to other intergenic regions. CpG islands are an additional feature used to recognize regulatory regions. Both properties are more pronounced in promoters than in enhancers. In human about 70% of the promoters contain CpG islands [287].

3.3.2 Prediction of Regulatory Regions

First, we present methods specialized on the prediction of promoters. Subsequently we shift the focus to more general methods for the prediction of cis-regulatory modules, which include promoters and enhancers.

Promoter Prediction

Promoter prediction tools have classically been used for *in silico* gene finding when no experimental data about transcriptional start sites were available. Given the availability of vast amounts of TSS data for many organisms, the main focus nowadays lies on the analysis of gene regulation and in genome annotation [22].
Tools for promoter prediction use various sources of information, which are used individually or in different combinations: the presence of CpG islands in vicinity to transcriptional start site (*TSS*) [81, 21, 140, 77, 263, 157], detection of specific transcription factor binding sites like the TATA box or the downstream promoter element DPE [186, 87, 273, 241, 314], density of predicted transcription factor binding sites [103, 266], statistical properties of the promoter sequence opposed to intergenic sequences [21, 20, 77, 87, 273, 186], and using transcript information [203]. Some other methods use sequence conservation to other organisms for phylogenetic footprinting or phylogenetic shadowing [91, 38, 41, 95, 82, 314] or analyse large-scale structural properties of the DNA like base-stacking [2].
The tools used to discriminate promoters from non-promoters based on the properties mentioned above have its origins in different areas of machine learning and statistics: They include support [126, 119] and relevance [87] vector machines (SVMs and RVMs), neural networks [20, 21, 186, 273, 241], discriminant analysis [77, 157, 314], Hidden Markov Models (HMMs) [240], and various statistical analyses of promoters [20, 21, 77, 87, 241, 263, 290]. Commonly these tools require some amount of training data to be able to distinguish between promoters and non-promoters.

Cis-Regulatory Modules

Detection methods for cis-regulatory modules (CRMs) are a more general approach, which in principle allow for the prediction of promoters as well as enhancers. In this section we adopt the division of methods into CRM scanners, CRM builders and CRM genome screeners proposed in the review by Loo et al. [206] and presented in Figure 3.4. The scanners

detect CRMs based on predefined models, for example pairs of PWMs within a specific distance. The builders detect regulatory patterns that are specific to subsets of genes which are assumed to be co-regulated. Some builders use *a priori* knowledge from PWM libraries, possibly subsets of transcription factors important for a set of genes under scrutiny. Other builders apply methods from pattern detection [333] and extend them to include cooperativity of TFs or TFBSs. The CRM genome screeners do not make assumptions about co-operation of specific TFs, use PWM databases and genomic sequences as input, and assess regulatory regions based on homotypic or heterotypic clustering of TFBSs.
Some of the methods presented here overlap with the methods for the prediction of TF interactions presented in Section 3.2.2. Moreover some methods show up in more than one of the categories stated above.

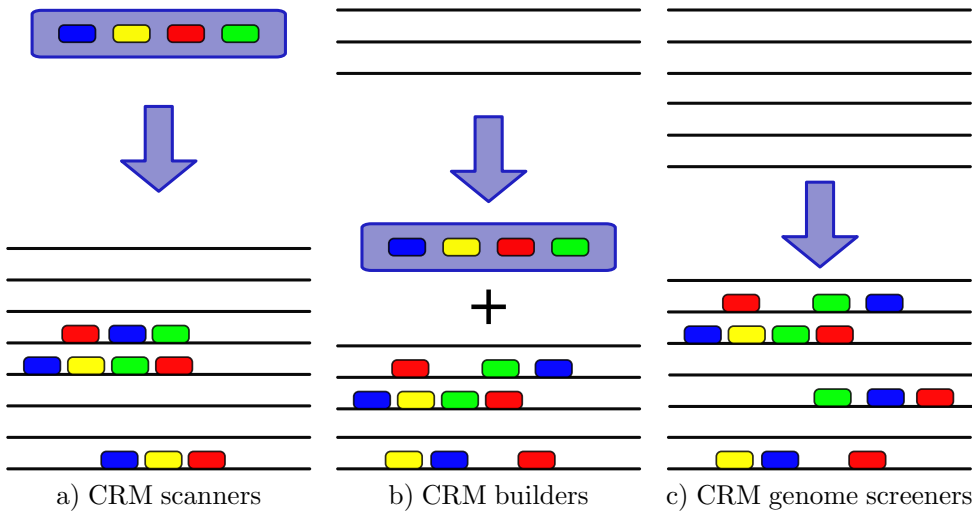


Figure 3.4: Different categories of methods for the detection of cis-regulatory modules: a) CRM scanners require user-defined motif combinations as input to search for putative regulatory regions. b) CRM builders analyse a set of co-regulated genes as input and produce candidate motif combinations, as well as similar target regions. c) CRM genome screeners search for homotypic or heterotypic motif clusters without making assumptions about the involved TFs. Schema adapted from Loo et al. [206]

CRM Scanners CRM scanners search for sequence regions, which fulfil the requirements of a predefined model, such as the occurrence of two specific TFBSs close to each other. Klingenhoff et al. [182] and Kel et al. [171] search for the joint occurrence of two known TFBS motifs or *composite elements* for the prediction of genes with similar expression patterns. Wagner [348] detects homotypic or heterotypic clustering of one or two PWMs and evaluates the significance by modeling the occurrences using independent Poisson processes. CIS-ANALYST [34, 136] counts the number of TFBSs derived from a small set of PWMs in a given sequence window and predicts a CRM if the TFBS count exceeds a predefined threshold. Wasserman and Fickett [350] and Krivan and Wasserman [190] perform logistic regression analysis using a set of known cis-regulatory modules and a negative set, as well as a set of PWMs known to be important for the sequences under research. They apply

the method on liver and muscle specific genes in human and mouse. MSCAN, developed by Johansson et al. [165] and Alkema et al. [9] needs a set of PWMs as input and calculates a CRM score for hit combinations based on the p-values of individual TFBS hits. The authors calculate a p-value by fitting the CRM score to a statistical distribution. In their tool AHAB, Rajewsky et al. [271] compute the probability that a piece of sequence is derived from a set of input PWMs as opposed to a background using a maximum-likelihood approach. Frith et al. [111, 112] developed the three tools Cister, COMET, and Cluster-Buster. All three need a set of PWMs as input. Cister relies on Hidden Markov Models where each position in the sequence can either be "motif", "intra-CRM-background", or "inter-CRM-background". COMET also learns an HMM, but applies Viterbi decoding instead of posterior-decoding in the case of Cister. Cluster-Buster is similar to Cister, but improves run-time and can learn weights of PWMs if a training set is available. Stubb from Sinha et al. [306, 307] also learns an HMM, but explicitly models the spatial relationships of TFBSs and uses comparative genomics. Based on a set of PWMs MCAST by Bailey and Noble [17] learns an HMM similar to COMET, with slight differences in modeling of the background. ModuleFinder by Philippakis et al. [260] uses the number of TFBSs in a sequence window, their homo- or heterotypic clustering and the evolutionary conservation to calculate the likelihood of the presence of a CRM. Grad et al. [127] use Markov Chain discrimination (log-odds score of CRM- and background HMM) in their tool PFRSearcher. The tool uses phylogenetically foot-printed non-coding regions (PFRs). ModuleScanner by Aerts et al. [5] finds the best scoring combination of PWMs for a given sequence set by a tree search. ModuleFinder by Philippakis et al. [260] calculates the likelihood that a piece of sequence is CRM for a set of PWMs, and takes into account homo- and heterotypic site clustering and evolutionary conservation. EEL by Hallikas et al. [137] first predicts TFBSs in two species and then creates alignments of TFBSs instead of sequences, relying on the order of TFBSs. It calculates a combined score including an affinity term, a clustering term, and a conservation term and punishes shifts of TFBSs between species. EvoPromoter by Wong and Nielsen [359] applies phylogenetic HMMs. Besides aligned sequences and a PWM set it requires a phylogenetic tree as input. SimAnn by Bais et al. [18, 19] performs a simultaneous alignment of sequences and the annotation of a small PWM set to detect regulatory regions.

CRM Builders Using PWM Libraries The methods in this class predict CRMs based on a PWM set and regulatory regions of co-regulated or co-expressed genes. Creme by Sharan et al. [297] detects co-occurring TFBSs from a library of PWMs in conserved upstream regions of co-expressed genes. Using a hashing-algorithm Creme enumerates all possible PWM combinations and calculates the combined significance. Prediction of novel CRMs is carried out for sequence windows based on significant combinations. Aerts et al. [5] presented ModuleSearcher alongside with the tool ModuleScanner presented above. The tool uses conserved non-coding regions and finds CRMs similar to a given training set. MOPAT, developed by Hu et al. [151], is a graph-based method to predict recurrent CRMs from known motifs. The method builds a motif pair tree and calculates the significance of a motif combination using Poisson clump heuristics. MARSMOTIF by Das et al. [74] takes a set of candidate motifs and microarray expression data and models the expression values as a function of motif content of a sequence. The core of the method are multivariate adaptive regression splines. In their tool ModuleMiner Loo et al. [206] identify CRMs that are most discriminative for a given set of genes with respect to the genomic background.

CRM Builders Using PWMs from Pattern Detection The CRM builders, which perform pattern recognition on their own, often just need a set of co-regulated or co-expressed genes as input. GuhaThakurta and Stormo [131] developed the method Co-Bind. It identifies motifs for transcription factors which cooperatively bind close to each other in the promoters of small co-expressed gene sets. Co-Bind applies a Gibbs sampling strategy and maximizes the joint likelihood of occurrence of two TFBS motifs. GuhaThakurta and Stormo applied the method on small sets of co-regulated yeast and E. coli genes and show that the motifs detected belong to the transcription factors known to regulate the respective gene sets. Kreimann [189] uses a set of co-regulated genes and a background set of genome-wide putative regulatory sequences to exhaustively try all possible combinations of up to four detected PWMs. The PFRSampler, also presented by Grad et al. [127] alongside with the PFRSearcher, uses simulated annealing on the sum of PFRSearcher scores and needs a set of co-regulated genes. CisModule by Zhou and Wong [373] applies hierarchical mixture models in two steps: CRM vs. background and within the CRM TFBS vs. CRM-background. It consists of an iterative algorithm, which estimates parameters given TFBS positions and CRMs, and subsequently estimates new occurrences of CRMs and TFBS positions.

CRM Genome Screeners As opposed to the CRM builders that use small sets of PWMs assumed to be important for a set of genes, CRM genome screeners use big libraries of PWMs and usually work on large sets of regulatory regions. These methods detect homotypic or heterotypic clusters of TFBSs. Trafac, developed by Jegga et al. [163], looks for sequence windows with high densities of predicted TFBSs in evolutionarily conserved regulatory regions. Elnitski et al. [95] and Kolbe et al. [188] measure the similarity of patterns in aligned regions with the patterns in known regulator regions using a simplified alignment alphabet and higher order Hidden Markov Chains. PreMod by Blanchette et al. [39] bases on the detection of cluster of phylogenetically conserved TFBSs. It identifies up to five different tag-TFBSs per CRM and homotypic clustering and calculates a CRM score using only the tag TFBSs.

3.3.3 Summary

A plethora of approaches for the prediction of regulatory regions is available. The variety of methods applied to the problem is huge and encompasses methods from statistics and machine learning. Also the required input data varies, some methods just work on sequences, some require PWM libraries, some use conservation between species or other data. There are tools specialized on promoters which use features typical for promoters, like the presence of specific motifs or low-level properties of DNA. Other tools apply more generalized approaches to detect CRMs, which also allows for the prediction of enhancers.
The more is known about a system under scrutiny *a priori*, the more successful the prediction of regulatory regions is. General methods are often only successful in subclasses of regulatory regions, like GC or CpG rich regions, or regions with homotypic clusters of TFBSs. For methods using complex models it can be hard to discern the influence of DNA low-level features like GC content from the high-level model components like combinations of certain TFBSs. In general, even with complex models it is easier to find regulatory regions which are GC rich and contain CpG islands.
For reviews and performance assessment of different methods for promoter prediction see

the publications of Fickett and Hatzigeorgiou [103], Bajic et al. [22], Abnizova and Gilks [3], and Abeel et al. [2]. For a review of methods for the detection of cis-regulatory modules see Loo and Marynen [205] and Narlikar and Ovcharenko [234].

3.4 Similarity and Clustering of Position Weight Matrices

As already mentioned in Section 2.3.2, collections of TF binding profiles contain redundancy. The underlying cause of the redundancy is partly technical, partly biological. On the technical side, the PWM collections often contain more than one binding site description per factor because the same factor might have been topic of multiple different investigations [219, 346]. Moreover the application of sets of *de novo* motif discovery tools leads to redundant sets of PWMs [333]. On the biological side, we have the limited number of DNA binding domains already described in Section 2, leading to different TFs recognizing similar motifs [143, 207, 286]. Itzkovitz et al. [160] find that similar motifs often correspond to TFs with similar biological functions, so that problems caused by mis-recognition are minimized. Mahony et al. [209] provide an overview of the various methods for the creation of representative sets of PWMs.

In this work we use the methods by Pape et al. [252] to generate clustered PWM sets. Pape et al. [252] first define a *natural similarity measure*, which regards two PWMs as similar, if they describe similar binding sites, implying a high number of overlapping occurrences on a random sequence. Thus the number of hits on the sequence is correlated. Pape et al. [252] represent the correlation using the asymptotic covariance of the number of hits for two PWMs. For the application in clustering of binding sites, one also needs to calculate the relative shift yielding the highest overlap probability. The covariance summarizes the overlap probabilities for all relative shifts. Hence, for clustering it is more convenient to substitute the covariance by the maximum overlap probability for all relative shifts. For comparison with other pairs of PWMs, the maximum overlap probability is normalized by dividing by the probability of joint occurrences under the independence assumption. Applying the logarithm function, this retrieves a log-odd score and the best relative position for each pair of PWMs.
The resulting gapless alignment of two PWMs enables Pape et al. [252] to merge PWMs in a three step procedure. A selection step chooses the pair of PWMs with the highest similarity. In a merging step they create a new familial binding profile, based on the gapless alignment resulting from the optimal shift of the two PWMs. The new familial binding profile consists of the sum of the respective count matrices, with non-overlapping positions filled with background frequencies. A verification step ensures that the resulting familial PWMs are similar to the original PWMs that lead to the familial profile. If the similarity is too low, the merging step is discarded. Taking a GC content as background, a scanning threshold method, and a set of PWMs, the method provides a set of PWMs representative for the input set.

3.5 Graph Theory and Graph Matching

In this section we will introduce the basics of graph theory. We will define the problem of matching on a graph and lay out a way from alternating paths to an algorithm for maximum-cardinality matching. We will shortly sketch the extension of the method to maximum-weight matching. The concepts presented in this section are the foundation of binding site graphs and the calculation of the regulatory potential, which we present in Chapter 7.

3.5.1 Graph Theory Definitions

A graph $G = (V, E)$ is a representation of objects and the links between them. The objects are embodied by a set V of vertices (or nodes) and the links by a multiset E of edges. Two vertices connected by an edge are called *adjacent* and an edge connected to a vertex is *incident* to the vertex. The *degree* of a vertex is the number of edges connected to it.
The number of vertices in a graph G is the order $|V|$. The total number of edges is the size of the graph $|E|$. The number of edges connected to an individual vertex is the *degree* of the respective vertex. In the most simple case, edges are undirected and symmetric (Figure 3.5). In *directed* graphs, edges contain information about their direction and can be represented as ordered pairs.
A *weighted graph* $G = (V, E, w)$ is a graph where edges have weights. The sum of all edge weights in a graph G is the *weight* of the graph.

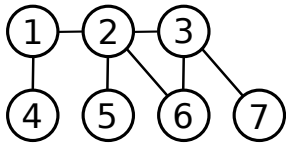


Figure 3.5: Example of a simple, undirected graph $G = (V, E)$ with 7 vertices $V = \{1, 2, 3, 4, 5, 6, 7\}$ and 7 edges $E = \{\{1, 4\}, \{1, 2\}, \{2, 5\}, \{2, 6\}, \{2, 3\}, \{3, 6\}, \{3, 7\}\}$. Vertex 2 has a degree of 4, since it is connected to the vertices 1, 3, 5, and 6.

A bipartite graph contains two disjoint partitions of vertices U and V. Edges in bipartite graphs can only connect the two sets, having one endpoint in U and one in V. A bipartite graph can not contain three or more vertices, all of which are connected to all others. For example, the graph shown in Figure 3.5 is not a bipartite graph, because vertices 2, 3, and 6 are all connected to each other. For that at least one disallowed edge within one partition would be needed. Moreover a graph is bipartite if and only if it is 2-color-able, that is after assigning colors to vertices such that no edge has equally-colored endpoints not more than two colors are needed (Figure 3.6).

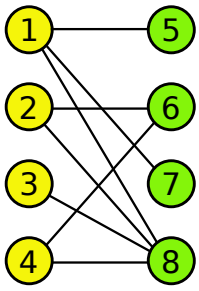


Figure 3.6: Example of a bipartite graph $G = (U, V, E)$ consisting of the vertex set $U = \{1, 2, 3, 4\}$ with 4 vertices (colored yellow), vertex set $V = \{5, 6, 7, 8\}$ also consisting of 4 vertices (colored green) and the edge set $E = \{\{1, 5\}, \{1, 7\}, \{1, 8\}, \{2, 6\}, \{2, 8\}, \{3, 8\}, \{4, 6\}, \{4, 8\}\}$. Every edge has one endpoint in the green and one endpoint in the yellow partition of the graph.

3.5.2 Graph Matching

Given a graph $G(V, E)$, a matching $M \in E$ is a set of pairwise non-adjacent edges, that is a set of edges without common vertices (Figure 3.7). Edges that belong to M are *matched*, all other edges $E \notin M$ are *unmatched* or *free*. Likewise, a vertex with an incident matched edge is matched while all other vertices are unmatched or free. A matching is *maximal* if no edges ending in free vertices exist in G. Adding another edge to a maximal matching results in a set of edges that is not a matching. A *maximum-cardinality* matching is the matching with the largest possible number of edges for a given graph G. A matching is said to be *perfect* if all vertices in the graph are matched, resulting in an order $|M| = \frac{|V|}{2}$.

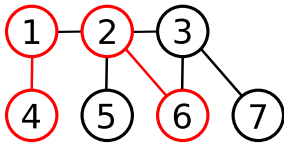


Figure 3.7: Example of a matching. Here, the graph $G = (V, E)$ from Figure 3.5 contains a matching $M = \{\{1, 4\}, \{2, 6\}\}$, marked in red.

Maximum-Weight Matching

In case of a weighted graph $G(V, E, w)$, a *maximum-weight matching* (MWM) the matching with the maximal weight $w(M) = \sum_{e \in M} w(e)$. (Figure 3.8). A *maximum-weight perfect matching* is a perfect matching with the highest possible weight. The two problems are closely related and can be reduced into each other as described in Schaefer [289].

Finding the maximum-weight matching in a graph $G(V, E)$ with n vertices and m edges is a non-trivial task. Figure 3.8 illustrates some of the problems that arise, for example that the edge $\{3, 4\}$ with the highest weight of 11 is not part of the matching, or that only two of nine edges in the graph belong to the maximum-weight matching. Other matchings for

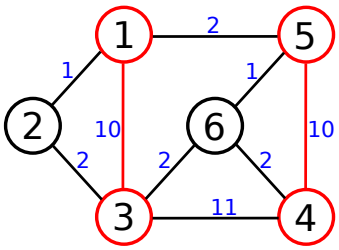


Figure 3.8: The graph G with vertex set $V = \{1,2,3,4,5,6\}$ and edge set $E = \{\{1,2\},\{1,3\},\{2,3\},\{1,5\},\{3,4\},\{3,6\},\{5,6\},\{6,4\},\{4,5\}\}$ has a maximum-weight matching $M = \{\{1,3\},\{4,5\}\}$ with the weight 20. No other combination of non-adjacent edges has a higher edge weight. Vertices 2 and 6 are unmatched. The edge weights are written in blue.

the given graph with a higher order and a smaller weight exist.

In 1965, Edmonds [92] found the first polynomial time algorithm with a run-time complexity of $O(n^4)$ for the maximum-weight matching problem. Gabow [115] improved the algorithm to a run-time complexity of $O(n^3)$. In 1986, Galil [117] enhanced the performance to a run-time complexity of $O(n \cdot m \cdot log(n))$.
Maximum-weight matching in a bipartite graph is a special case of weighted matching in a general graph. The first polynomial time algorithm for maximum-weight matching in a bipartite graph $G(U,V,E)$ with $n := |U| + |V|$ and $m := |E|$ was the *Hungarian method* developed by Kuhn [191] . Galil [117] found an algorithm improving the a run-time complexity to $O(n \cdot (m + n \cdot log(n)))$.

Because the description of an algorithm for maximum-weight matching is complex, we can not cover it in detail in this thesis. Here, we only present a short overview and refer the interested reader to the publications of Mehlhorn and Schaefer [223] and Schaefer [289], which contain an in-depth presentation of the field.

Schaefer [289] first shows an algorithm for maximum-cardinality matching. It is based on the search for an *augmenting path* with respect to a matching M. An augmenting path is an *alternating path* that starts from and ends in free vertices. The edges of an alternating path belong alternatively to the matching and not to the matching (Figure 3.9 a)). If an augmenting path exists, the matching is not of maximum cardinality, and one can use the augmenting path to improve the matching (Figure 3.9 b)). The simple algorithm for a maximum-cardinality matching repeatedly searches for augmenting paths and in turn improves the respective matching. When no augmenting path can be found any more, the matching is of maximum cardinality.

The search for an augmenting path described in Schaefer [289] relies on an alternating tree, consisting of alternating paths with respect to the given matching. If the graphs contain cycles of odd length, a simple version of the algorithm is not guaranteed to find all augmenting paths, because the algorithm relies on even or odd distances of vertices from

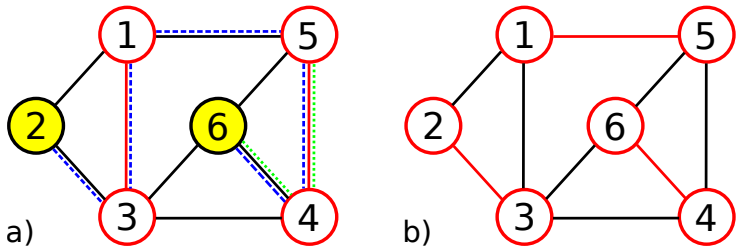


Figure 3.9: a) The graph G with vertex set $V = \{1,2,3,4,5,6\}$ and edge set $E = \{\{1,2\},\{1,3\},\{2,3\},\{1,5\},\{3,4\},\{3,6\},\{5,6\},\{6,4\},\{4,5\}\}$ has a matching $M = \{\{1,3\},\{4,5\}\}$ with order 2 marked in red. $p_{alt} = (6,4,5)$ is an example for an *alternating* path relative to M - $\{4,5\}$ is part of the matching, $\{4,6\}$ is not part of the matching (marked with green small-dotted line). The alternating path $p_{aug} = (6,4,5,1,3,2)$ starting and ending with an unmatched vertex (2 and 6, marked yellow) is *augmenting* with respect to the matching M (marked with dashed blue line). b) Using the augmenting path p_{aug}, one can construct a matching $M' = \{\{6,4\},\{5,1\},\{3,2\}\}$ with the order 3 using all edges on the augmenting path *unmatched* in M. This way the order of the resulting matching M' is one edge bigger than the order of the original matching M.

the root of the tree. Odd-length cycles can lead to multiple possible paths from the root to a given vertex, which can be of even and odd length (Figure 3.10).

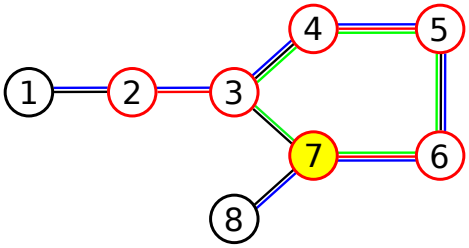


Figure 3.10: In the example graph the path $p = (1,2,3,4,5,6,7,8)$ (marked blue) is augmenting with respect to the matching marked in red. Using the simple algorithm described by Schaefer [289] it may occur, that the path is not detected. Vertex 7 (yellow) is even if the tree-building procedure enters the odd-length cycle in a clockwise fashion at its base (vertex 3), or odd if it enters counterclockwise. In the second case we do not detect the augmenting path starting in vertex 8, since we only search for a free edge starting in an even vertex. This problem is present whenever the graph contains odd-length cycles (marked green in the example).

The solution for this problem was introduced by Edmonds [92]. He showed that odd-length cycles can be shrunk to *blossoms*. Blossoms are vertices representing all vertices in an odd-length cycle. After replacement of all odd-length cycles in a graph the simple approach leads to augmenting paths, possibly including blossoms. Subsequently the blossoms are expanded again, leading to an augmenting path with respect to the matching on the complete graph (Figure 3.11).

The extension of the blossom shrinking approach for maximum-cardinality matching to maximum-weight matching involves a formulation of the problem in the language of combinatorial optimization. Similar to Edmonds [92], Schaefer [289] first states the problem in

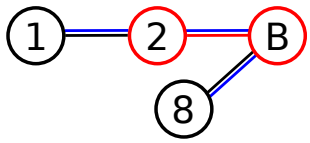


Figure 3.11: We shrink the odd-length cycle from Figure 3.10 to blossom $\mathcal{B}$. If there is an augmenting path $p' = (1, 2, B, 8)$ (marked in blue) containing the shrunken blossom $\mathcal{B}$, there is also an augmenting path in the original graph containing the respective expanded blossom [92].

an integer programming fashion, relax it to a linear programming formulation. Using the primal-dual method, the linear program is then transformed into a dual formulation, finally leading to a solution to the original integer program based on an iterative method.

3.5.3 Summary

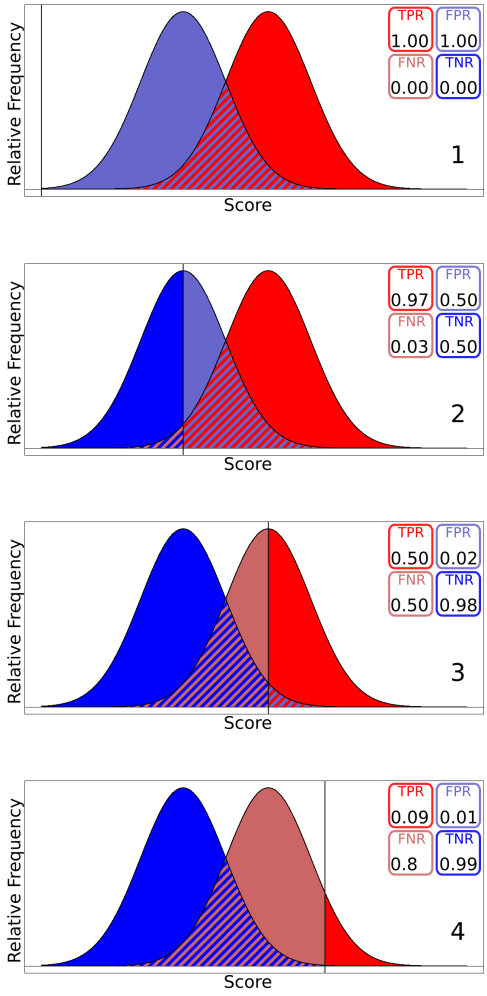
In this section we presented the basics of graph theory and some aspects of the problem of graph matching. We sketched methods for maximum-cardinality matching and maximum-weight matching. We will use maximum-weight matching in Section 5.1 to calculate a regulatory potential on binding site graphs.

3.6 Assessment of Results

3.6.1 Receiver Operator Characteristics

The receiver operating characteristic (ROC) is a graphical means for the quality-assessment of a binary classifier. To generate a ROC curve, one calculates error rates at a variable discrimination threshold for the binary classifier. Given a distribution of scores from a positive and a negative set, the ROC curve displays the *true positive rate* (TPR or *sensitivity*) on the y-axis and the *false positive rate* (FPR or (1-*specificity*), *type I error*) on the x-axis. The *true positive rate* is defined as the number of positives, which are correctly classified divided by the total number of positives. The *false positive rate* is the number of negatives that are incorrectly classified as positive divided by the total number of negatives.

Figure 3.12 shows a ROC curve for two slightly overlapping distributions of scores for a negative and a positive set. Relative to each other the score distribution of the negative set is shifted to the left of the score distribution of the positive set. For a small value of the discrimination threshold points at the top right of the ROC curve are drawn (high TPR, high FPR). Shifting the threshold to higher values, the points in the ROC curve move to the bottom left (low TPR, low FPR). The *area under the curve* (AUC) is a measure to assess the discrimination quality of the score. A random classification results in an AUC of roughly 0.5, while a perfect classification would be reached with an AUC of 1 (the point with a 100% true positive predictions and no false positives in the top left corner of the plot exists). The AUC can be regarded as representing the likelihood that a classifier assigns a



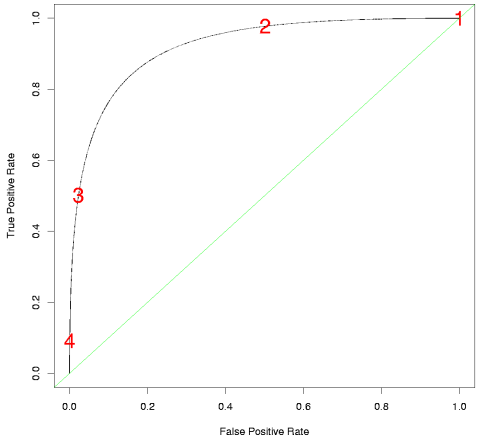


Figure 3.12: Example ROC curve: the images 1-4 on the left show two score distributions for a negative and a positive set. The score is used as a binary classifier. A variable discrimination threshold is used to calculate the true positive rate (TPR, sensitivity) and the false positive rate (FPR, (1-specificity), which are plotted against each other for every threshold. The resulting ROC curve is shown on the right. The red crosses correspond to the TPR and FPR at the example thresholds displayed on the left.

higher value to a randomly chosen positive item than to a randomly chosen negative item.

For an overview about ROC curves, AUC and related methods, implementations and interpretations see Hanley and McNeil [138] and Fawcett [102]. In the present work we use the `ROCR` package to create ROC plots [305].

The standard approach for the calculation of ROC curves requires a positive and a negative set. In our application, only reliable positive sets are available, while the negative have to be constructed and are less trustworthy. In Section 6 we predict TF interactions and assess the results with ROC curves using known interactions as a positive set. As a negative set we use all interactions from the dataset which are not known. This set includes non-interacting TFs as well as yet unknown TF interactions. We assume that the fraction of *real* negative TF pairs in that set is large enough to work as a negative set for the calculation of the ROC curve. In Section 7 the goal is to detect regulatory regions. Also here we apply the ROC method and calculate the area under the curve to evaluate the results. Here we use known regulatory regions as a positive set. As a negative set we take regions which are less likely to contain real regulatory elements, like random intergenic sequences. Again, we can not rule out that the negative set also contains sequences with regulatory function, but we assume that their fraction is small enough.
Using negative sets in the ROC method which still contain positive examples results in smaller AUCs compared to using clean negative sets. Due to the formulation of our problems an AUC of 1 can not be reached.

Part II

Methods

Chapter 4

A Co-Occurrence Score for the Prediction of Transcription Factor Interactions

Transcription factors fulfill their role in complexes of multiple proteins, some of which bind the DNA themselves (Section 2.2). Except for long-range interactions with enhancers for example, the binding sites for the involved TFs lie in relative proximity to each other. Although TFs are usually able to interact with multiple factors, one does not observe arbitrary functional combinations of TFs: specific combinations of TFs play specific roles in an organism, resulting in frequently occurring combinations of TFs.

Our motivation for the development of a co-occurrence score is the assumption, that predicted binding sites of functionally interacting TFs co-localize more often than expected by chance. Despite the high numbers of false positive predictions of *individual* TFBSs, we hypothesize that the signal of co-occurrence patterns is strong enough to identify interacting transcription factors.

4.1 Predicting TF interactions Based on TFBS Co-Occurrence

4.1.1 Synopsis

In this section we present the calculation of the co-occurrence score for TFBSs, used to identify over- or underrepresented TFBS combinations. The TFBS co-occurrence score is a log-odds score of observed and expected counts of TFBS pairs. The section is divided into three parts: First we discuss the problems which arise when counting TFBS combinations naively and solutions for the problems. Then we explain how to obtain the expected number of TFBS pairs in set of sequences annotated with TFBSs using a permutation procedure. Finally we present the co-occurrence score itself.

4.1.2 Counting Co-Occurring TFBSs

Preliminaries Assume a set of regulatory sequences, on which we annotate binding sites with a set of m position weight matrices $\{p_1, p_2, ..., p_m\}$ using the method described in Section 3.1. The resulting annotation is a set of n predicted TFBSs $\{t_1, \ldots, t_n\}$. Each TFBS t_i has genomic coordinates $t_{i \to start}$ and $t_{i \to end}$, a strand $t_{i \to strand} \in [+, -]$, and a PWM type

$t_{i\to type}$, with which it was detected. We assume sorted TFBSs with $t_{i\to start} \leq t_{i+1\to start}$. The length of a binding site t_i is $|t_{i\to end} - t_{i\to start}|$ and is equal to the length of the respective PWM. In our counting procedure we ignore the predicted strand of a TFBS.

4.1.3 Counting Pairs in a Single Window

We start with a pair counting procedure for TFBSs in a single window. Take a sequence window $\mathcal{W} = \{t_i, t_{i+1}, \ldots, t_j\}$ of size w containing the TFBSs t_i to t_j, with $t_{j\to end} - t_{i\to start} \leq w$. The window size limits the maximal distance, within which we consider two TFBSs as co-occurring.
The most simple approach to count co-occurring TFBSs now is to count all combinations $(t_{i\to type}, t_{j\to type}), t_i, t_j \in \mathcal{W}$. The problem with this simple approach is that some TFBSs occur in homotypic clusters. Thus we expect to find sequence windows with many predicted binding sites for one factor. In this case the simple counting method described above leads to a bias towards high numbers of TFBS combinations involving TFBSs that occur in homotypic clusters. In the context of the prediction of transcription factor interactions based on overrepresented combinations of TFBSs this can confound the results.
To overcome this problem, we reduce the the influence of homotypic TFBS clusters by not counting repeated occurrences of a pair of TFBSs $(t_{i\to type}, t_{j\to type})$. We implement this by counting the combination of *TFBS types* present at least once in a window instead of counting all combinations. Like this, we would loose the information about homotypic pairs in the window. Thus we count one homotypic pair (t_i, t_i), while at the same time we count heterotypic pairs (t_i, t_j) for all $t_j \in \mathcal{W} \setminus \{t_i\}$ as if t_i would only occur once in the window. We illustrate the problem and the solution in Figure 4.1.

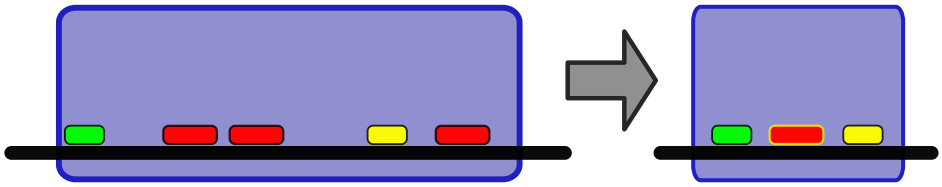


Figure 4.1: Counting of pairs within a window. The red TFBSs form a homotypic cluster. We count the repeated occurrence of TFBS pairs involving red TFBSs only once in a window. The pairs that are actually counted are shown in the window with the "effective" content on the right. The yellow border of the red TFBS signifies the counting of the homotypic pair (red, red). In the example we count each of the pairs $(\text{red}, \text{green})$, $(\text{red}, \text{yellow})$, and (red, red) once.

The example window on the left contains five TFBSs: a green, three red, and a yellow one. Using the simple approach, we would count three times $(\text{red}, \text{green})$, three times $(\text{red}, \text{yellow})$, three times (red, red), and once $(\text{green}, \text{yellow})$. We count the pairs of TFBS types, or in our example on the right, TFBS colors. This leads to counting once $(\text{red}, \text{green})$, once $(\text{red}, \text{yellow})$, and once (red, red). As a sign for the special treatment of the homotypic combination, we mark the red TFBS with a yellow frame.

Large PWM data sets are often redundant and contain PWMs which are similar. This can happen due to multiple binding site descriptions for the same factor or similar binding sites for different factors. Due to that, overlapping predicted TFBSs and even stacks of predicted TFBSs can appear. On the other hand, for many TFs it is known that they interact with other TFs, which have sites that overlap with each other. Moreover due to experimental artefacts, the PWM describing the binding site of a TF can be larger than the real binding site of a TF.
The magnitude of the described problem depends highly on the redundancy that is present in the PWM set used to annotate the TFBSs. Hence we want to be able to assess the impact of overlapping TFBSs on our co-occurrence score. Thus, we either count all possible TFBS type combinations, or we count conservatively and ignore overlapping TFBS pairs.

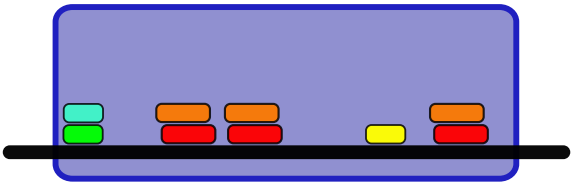


Figure 4.2: Overlapping TFBSs potentially lead to overestimating the pair counts for similar PWM pairs. The conservative way of counting TFBS pairs ignores overlapping TFBS hits and does not count combinations of TFBSs on the same stack.

Figure 4.2 contains a window with overlapping TFBS predictions. Ignoring overlapping TFBSs leads to not counting (blue, green) and (orange, red).
The counting procedure above enables us to count TFBS combinations with a maximum distance to each other in non-overlapping windows. This leads to the problem, that depending on the boundaries of the windows we disregard possible functional TFBS combinations.

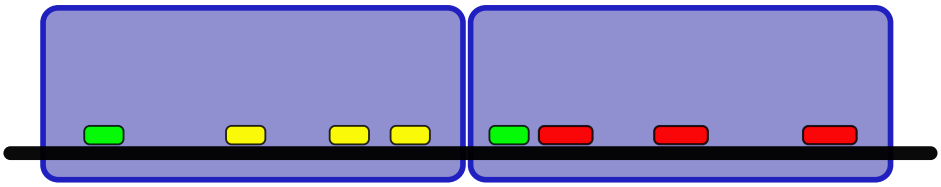


Figure 4.3: When we use non-overlapping windows in the counting procedure, we can not count the transcription factor binding site combinations which are close enough but lie in neighbouring windows. In the left window of the present example we count the combinations green - yellow and yellow - yellow. We count the combinations green - red and red - red in the right window. We disregard the combination of yellow and green although the binding site distance is small enough.

We illustrate this in Figure 4.3. The window boundary in the middle of the sequence regions impedes counting of TFBS pairs, which are close enough to be considered co-occurring given the window size.

4.1.4 Counting Pairs in a Sliding Window

We solve the problem of window boundaries by extending the procedure to use a sliding window. Opposed to non-overlapping windows, this has the advantage that we can take into account every TFBS pair (Figure 4.4).

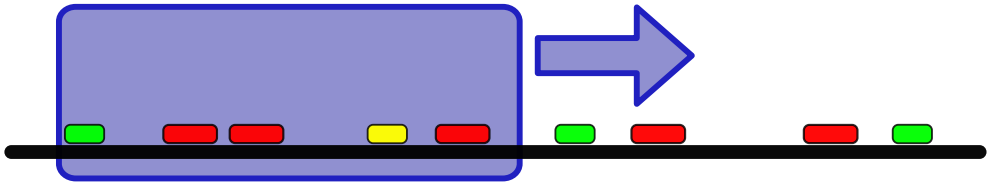


Figure 4.4: Sliding window of size w. Coloured boxes represent predicted TFBSs. First we count the TFBS pairs in the current window. Subsequently we shift the window to the next TFBS position.

We define a starting sequence window $\mathcal{W}$ of length w. The window starts at TFBS t_1 on the left. Furthermore the window $\mathcal{W}$ contains all TFBSs t_j for which $t_{j\to end} - t_{1\to start} \leq w$. Next we count the TFBS pairs as in the non-overlapping window case described above. Subsequently we shift the window such that it starts in TFBS t_2 and contains all TFBSs t_j for which $t_{j\to end} - t_{2\to start} \leq w$. We continue to shift the window in the same fashion, resulting in $\mathcal{W}$ containing $t_{j\to end} - t_{i\to start} \leq w$, until t_j is the last binding site in the actual sequence.
The problem with overlapping windows becomes apparent: If we just apply the procedure for the non-overlapping windows, we count some of the pairs more than once. A pair of TFBSs can be a part of several subsequent windows. To account for this we use a blacklist $\mathbb{B}$ containing all pairs counted before in windows, that overlap with the current window. We illustrate the blacklisting procedure in Figure 4.5.

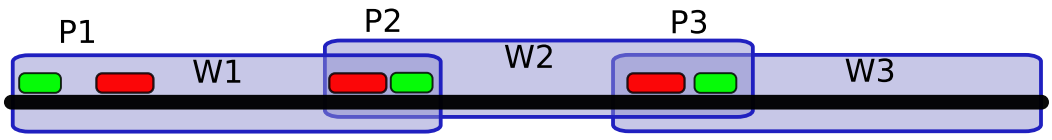


Figure 4.5: Blacklisted TFBS pairs are not counted. We count the $(\text{red}, \text{green})$ pair once in window W1. We ignore P2, since P1 is of the same type. The combination is in the blacklist also in window W2, leading to an ignorance even of P3 in window W2, although P1 and P3 have a greater distance than the window size. This is not a problem, as the counting of P3 is simply postponed to window W3.

If the blacklist $\mathbb{B}$ contains a TFBS pair, we ignore it every subsequent time we find it in a sequence window. Moreover we need to extend the procedure for the treatment of homotypic combinations: We also ignore combinations of transcription factor binding sites if a pair of the same binding site types is present in the blacklist, even if the respective binding sites are not identical.

4.1.5 Algorithm

Before we presented a method to count co-occurring TFBSs, that accounts for homotypic clusters of TFBSs, potential problems of similar PWMs, and overlapping sequence windows. In this section we explain the implementation of the method in an algorithm.

As input we require a set of predicted TFBSs ordered by sequence id and start position of the TFBS. The output is a set of pair counts $\mathcal{C}_{i,j}$ for all possible combinations for the input PWM set.
The algorithm uses a sliding window and considers several problems arising from typical properties of the input and from the usage of overlapping windows that we described above in section 4.1.3:

1 PWM sets can contain similar PWMs and result in overlapping predicted TFBSs.

2 Predicted and experimentally verified TFBSs sometimes arise in homotypic clusters of TFBSs leading to a high pair count of the same type.

3 We decided to count the TFBS pairs using a sliding window to not miss any combinations. This in turn means that in a given window TFBS pairs were potentially counted beforehand.

The algorithm iterates over all TFBS t_i in the set of predicted binding sites. Each t_i defines a new window $\mathcal{W}$ of size w. The window contains all TFBSs between t_i and t_e with $t_{e \to end} \leq t_{i \to start} + w$. which is the last TFBS, which is completely inside the window. Within each window we consider all combinations (t_i, t_j), and check, if they overlap (in the case of conservative counting), or if they are blacklisted. Blacklist updates are carried out with every window update. Whenever the distance from the start of the first TFBS of a blacklisted pair is bigger than the difference between the start of the current window and the window size, we remove the pair from the blacklist. We also use the blacklist to make sure not to count a pair of similar types a second time within a window. Finally, we update the pair count matrix with pairs that are neither blacklisted nor overlapping.

The procedure can then be formulated as shown in Algorithm 1.

4.1.6 Log-Odds Score and Expected Number of Pairs

The observed pair counts $\mathcal{C}_{i,j}^{obs}$ depend on the individual occurrences of the binding sites t_i and t_j. Thus, to obtain an informative value, the raw pair counts have to be set in relation with an expectation derived from a null model. We calculate a co-occurrence score $\mathcal{S}_{i,j}$ defined as a log-odds score of observed vs. expected pair counts.

$$\mathcal{S}_{i,j} := \log \frac{\mathcal{C}_{i,j}^{obs}}{\mathcal{C}_{i,j}^{exp}}. \tag{4.1}$$

We determine the observed number of TFBS pairs $\mathcal{C}_{i,j}^{obs}$ on the original TFBS annotation as described in Section 4.1.2.

Algorithm 1: algorithm to count co-occurring TFBS pairs

```
// Blacklist for TFBS pairs:  𝔹 ← ∅
// Set of TFBS: (t_1, ..., t_n)
// Window size:  w
// iterate over all TFBSs
for t_i ← t_1 to t_n do
   // t_j is the rightmost TFBS, which the window 𝒲 of size w completely
      contains.  𝒲 contains all TFBSs between t_i and t_j.
   𝒲 ← (t_i, ..., t_j), with t_{j→end} − t_{i→start} ≤ w
   // update blacklist:  remove all pairs (t_k, t_l) from blacklist, which only
      are element of windows, that do not overlap with the current.
   foreach (t_k, t_l) ∈ 𝔹 do
      if (t_{i→end} − t_{k→start}) < w then
         remove (t_k, t_l) from 𝔹
      end
   end
   forall t_j ∈ 𝒲 \ t_i do
      // overlap check for conservative counting method
      if t_i and t_j overlap then
         // ignore combination
         next t_j
      end
      // t_i and t_j do not overlap and are close enough
      // check if combination of the same types is in 𝔹
      if types of (t_i, t_j) ∈ 𝔹 then
         next t_j
      end
      // increment counter for pair (t_i, t_j)
      𝒞_{i,j} ← 𝒞_{i,j} + 1
      add pair (t_i, t_j) to 𝔹
   end
end
```

Calculating the Expected Number of Pairs We approximate the expected number of TFBS pairs using a permutation procedure: we keep the positions from the original TFBS annotation fixed, while we shuffle the labels of the TFBSs and count the co-occurring pairs again. We repeat this procedure a number of p times and take the average value of TFBS pair occurrences $\mathcal{C}_{i,j,k}^{perm}$ as the expected count $\mathcal{C}_{i,j}^{exp}$.

This method of calculating the null model has the advantage of preserving the local binding site density in the original annotation. It also preserves the number of individual binding site occurrences, implicitly taking into account the tendency of TFBSs to occur in clumps.

$$\mathcal{C}_{i,j}^{exp} := \frac{\sum_{k=1}^{p} \mathcal{C}_{i,j,k}^{perm}}{p}. \tag{4.2}$$

The resulting log-odds score $\mathcal{S}_{i,j}$ is large and positive if the corresponding pair of TFBSs (i, j) occurs more often than expected by chance. It becomes negative if (i, j) occurs less often than expected and equals zero if observed $\mathcal{C}_{i,j}^{obs}$ and expected pair counts $\mathcal{C}_{i,j}^{exp}$ are the same.
For TFBSs which occur individually in small numbers only, it happens that the number of observed or expected combinations with other TFBSs are zero. Since we cannot rule out that this is an effect of a small sample size, we add a pseudo count $\Pi = 1$ to the denominator and the numerator, so that the probability of a pair occurrence is now negligibly small instead of zero.

We define the complete co-occurrence score by eq. 4.3.

$$\mathcal{S}_{i,j} := \log \frac{\mathcal{C}_{i,j}^{obs} + \Pi}{\frac{\sum_{k=1}^{p} \mathcal{C}_{i,j,k}^{perm}}{p} + \Pi}. \tag{4.3}$$

4.1.7 Summary

We presented a new method to count co-occurring TFBSs that uses a sliding window. It deals with homotypic clusters and overlapping TFBSs. Based on the co-occurrence counts, we developed a co-occurrence log-odds score, that employs a permutation procedure to obtain a background distribution of co-occurrence counts. In the results in Chapter 6 we refer to the score simply as co-occurrence score. When we distinguish between the co-occurrence score calculated taking into account or ignoring overlapping TFBSs, we refer to it as *COOC/count OL* or *COOC/ignore OL* score, respectively.

4.2 An Empirical PWM Similarity Measure

Collections of known transcription factor binding sites usually contain partly redundant data. One reason is redundancy on the biological level. Some factors recognize the same or similar sequences of the DNA. The cause for similarity lies in the limited set of different DNA-binding domains [207, 13]. The similarity of binding preferences allows for competition of transcription factors or the possibility of a factor to take over parts of the functional spectrum of another factor. Other factors do not have a pronounced binding preference. Another reason for redundancy found in binding site databases are of technical nature. Descriptions of a factor's binding motif in multiple different publications leads to multiple entries in binding site collections for the same factor.

Thus it is of interest to evaluate the similarity of two position weight matrices.

Assume a set of TFBSs annotated with two PWMs i and j. Let $\mathcal{N}_i$ be the number of predicted TFBSs for PWM i and $\mathcal{N}_j$ be the number of predicted TFBSs for PWM j. The number of overlapping hits for the two different PWMs is called $\mathcal{N}_{i,j}$.

The fraction of TFBSs predicted with PWM i overlapping with TFBSs predicted with PWM j relative to $\mathcal{N}_i$ is then defined as

$$\mathcal{F}_{i,j} := \frac{\mathcal{N}_{i,j}}{\mathcal{N}_i}. \tag{4.4}$$

The fraction of TFBSs predicted with PWM j overlapping with TFBSs predicted with PWM i relative to $\mathcal{N}_j$ is then defined as

$$\mathcal{F}_{j,i} := \frac{\mathcal{N}_{i,j}}{\mathcal{N}_j}. \tag{4.5}$$

$\mathcal{N}_{i,j}$ is the same in both cases, but $\mathcal{N}_i$ and $\mathcal{N}_j$ can be different, which in turn leads to different frequencies. In Equation 4.6 we define the similarity $Sim_{i,j}^{emp}$ as the maximum of the two fractions $F_{i,j}$ and $F_{j,i}$.

$$Sim_{i,j}^{emp} := \max \left\{ \begin{array}{l} F_{i,j} \\ F_{j,i} \end{array} \right. \tag{4.6}$$

Using the larger of the two values leads to a high similarity value also in cases of two PWMs of different specificity. Otherwise PWMs with a low specificity always obtain low similarities due to their high number of ocurrences in the sequences. In case of homotypic combinations, $Sim_{i,i}^{emp}$ can take values between 0 and 1. In this case the value is to be interpreted as a tendency for self-overlapping hits.

This procedure implicitly takes into account several properties of the sequence set and the method of scanning for TFBSs. Specific patterns or GC content of the sequence set (for example a set of regulatory regions) are accounted for. Moreover different score thresholds in the scanning procedure result in different similarity values for two PWMs. The calculation of PWM similarity needs to be carried out anew for each data set at the scanning parameters used for binding site annotation.

4.3 Methods for Assessment of Results

4.3.1 Synopsis

In Section 3.6.1 we presented the receiver operator characteristics and the area under the curve as a method to test the quality of a score to separate a positive from a negative set. Here we introduce the relative rank sum (RRS) as an alternative method, which takes into account the rank of known interactions relative to other interactions of a single factor only. Moreover we describe the common neighborhood score as a measure to be used for the definition of a positive set for TF interactions, provided protein-protein interaction data are available for the organism under scrutiny.

4.3.2 Relative Rank Sum of Interactions in Positive Set

An alternative method to evaluate TF interaction prediction results is the calculation of a relative rank sum RS_{rel} of interactions from the set of positive combinations (Equation 4.7). While the ROC curve with the AUC presents a global measure, which assesses the ranks of known interactions within all possible interactions, the RRS assesses the ranks of known interactions within all possible interactions of *a single factor* only. Let D_i be the set of co-occurrence scores for combinations of PWM i with all other PWMs, and P the set of interactions in the positive set. For each PWM i part of a pair from the positive set P we calculate the rank of the combination (i, j) in D_i. The total number of combinations possible at a given parameter set can vary, so the rank is normalized with the number of scores in set D_i. We sum the relative ranks for all known interactions. A lower relative rank sum RRS signifies a better detection of known interactions at a given parameter set.

$$RRS := \frac{\sum_{i,j \in P} \frac{\text{rank}(i,j,D_i)}{|D_i|}}{\mid P \mid}. \tag{4.7}$$

The relative rank sum is only comparable within the same sequence set, PWM set, and positive interaction set. We use it to assess parameter combinations on the same data set.

4.3.3 Common Neighborhood Score

The set of known direct TF interactions is relatively small compared to all possible TF combinations, thus we often have the problem of small positive sets for the assessments of results. If a protein-protein interaction network is available for the organism in question, we can make use of indirect interactions. We expect that using TF pairs with a shortest path length > 2 as a positive set adds too much noise for our application. Thus to extend the positive set while keeping the noise level low, we define a score which represents the common neighborhood of two TFs with a shortest path length of 2.

This neighborhood overlap score is the fraction of the number of intersecting neighbors and total number of neighbors of two TFs

$$S_{no} =: \frac{N_{intersect}}{N_{union}}. \tag{4.8}$$

Chapter 5

Prediction of Regulatory Regions with Binding Site Graphs

5.1 Transcription Factor Binding Site Graphs

5.1.1 Synopsis

In this section we present the concept of *binding site graphs* and their use for the calculation of the *regulatory potential* of a genomic sequence.
Binding site graphs represent the putative transcription factor binding sites and interactions in a piece of genomic sequence. The prediction of individual TFBSs is error-prone. Since transcription factors often act in specific combinations, the goal is to exploit the combinations of their binding sites to make prediction of regulatory regions more reliable.
The concepts which we present here are an amalgamation of individual binding site prediction (Section 3.1), the co-occurrence score (Chapter 4), and graph theory (Section 3.5).

A TFBS graph represents the predicted transcription factor binding sites in a piece of DNA sequence and putative interactions between the corresponding transcription factors. We annotate a piece of sequence with a set of position weight matrices $\{p_1, \ldots, p_n\}$ using the methods described in Section 3.1 and obtain a set of predicted TFBSs $\mathcal{T}$. Furthermore we need a set of co-occurrence scores S containing all combinations $S_{i,j}$ of the set of PWMs. We use co-occurrence scores computed on a suitable set of sequences (that is a large set of known regulatory regions) using the procedure described in Section 4.

5.1.2 Building Binding Site Graphs

Complete TFBS Graph

The TFBS graph contains one vertex per predicted TFBS $t \in \mathcal{T}$. We extend the TFBS graph to be complete by adding edges from each vertex to all other vertices. We set the weight of an edge $w(e)$ connecting two vertices representing TFBSs t_i and t_j to $S_{i,j}$. This way, combinations of TFBSs common in known regulatory regions get large positive weights while combinations of TFBSs atypical for regulatory regions get weights smaller than zero. For a graph with $n = |\mathcal{T}|$ vertices representing the TFBS the number of edges in the general TFBS graph is $m = \frac{n \cdot (n-1)}{2} = \frac{n^2 - n}{2}$.

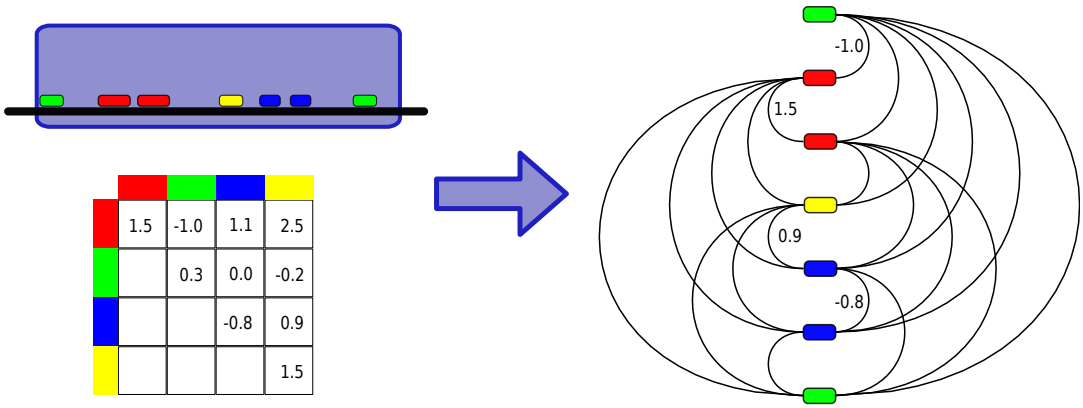


Figure 5.1: Construction of a general TFBS graph. Each vertex in the complete weighted graph on the right represents a predicted TFBS in the DNA sequence window (blue box). We set the edge weights to the co-occurrence score $S_{i,j}$ for the respective PWMs i and j. Four examplary weights are annotated with co-occurrence scores from the matrix on the left.

Bipartite TFBS Graph

As a second possibility to build the TFBS graph we realize a bipartite TFBS graph (see Section 3.5.1). Here each TFBS $t \in \mathcal{T}$ is represented by one vertex in the first partition and another vertex in the second partition. We put edges between each vertex from one partition to all vertices in the other partition except for the vertex representing the same TFBS. In the bipartite case the graph contains $n = 2 \cdot |\mathcal{T}|$ vertices, one per TFBS in each partition of the graph. The number of edges in the bipartite TFBS graph is $m = n \cdot (n-1) = n^2 - n$.

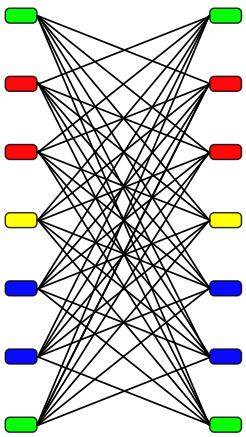


Figure 5.2: Construction of a bipartite TFBS graph. Each of the predicted TFBSs in the sequence is represented by one vertex in the left and one vertex in the right partition of the bipartite graph. We set the edge weights to the co-occurrence score $S_{i,j}$ for the respective PWMs i and j.

5.1.3 Calculation of Regulatory Potential from TFBS Graphs

The TFBS graphs contain information about the predicted transcription factor binding sites in DNA sequence encoded in the vertices. Moreover the edge weights encode the commonness of TFBS pairs in the training set (normally known regulatory regions).

In the following we show different ways to calculate a regulatory potential $\mathcal{R}$ from a given TFBS graph. Four are based on the complete TFBS graph presented before and one on the bipartite TFBS graph.

All the regulatory potentials presented depend on the number of TFBSs in a piece of DNA sequence. Thus we show different ways to normalize the scores, leading to regulatory potential scores which are independent from raw counts of TFBSs.

Maximum-Weight Matching

We build a complete weighted TFBS graph $G = (V, E, w)$ using the TFBS annotation of a DNA sequence and the set of co-occurrence scores S for all possible PWM pairs for that set. Afterwards we calculate a maximum-weight matching $M \subseteq E$ on G. Then we define the regulatory potential $\mathcal{R}_{\mathrm{MWM}}$ as the weight of the resulting matching:

$$\mathcal{R}_{\mathrm{MWM}} := \sum_{e \in M} w(e). \tag{5.1}$$

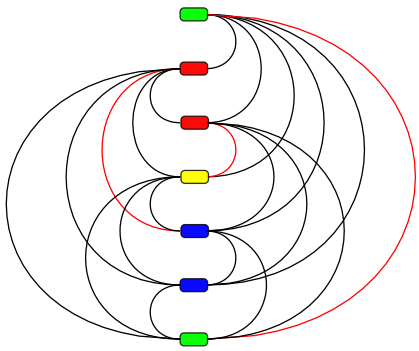


Figure 5.3: Example for regulatory potential $\mathcal{R}_{\mathrm{MWM}}$ on the example from Figure 5.1: the edges which are part of the matching are marked in red. We use the weight of the matching as the score $\mathcal{R}_{\mathrm{MWM}}$. Using the edge weights from Figure 5.1, the matching contains a green - green edge with weight 0.3, a red - blue edge with weight 1.1, and a red - yellow edge with weight 2.5. The resulting edge weight of the matching is $\mathcal{R}_{\mathrm{MWM}} = 2.3$.

The matching allows for one edge per vertex. The co-occurrence score represents the importance of an interaction between two TFBSs. $\mathcal{R}_{\mathrm{MWM}}$ then takes into account the combination of most important interactions of each site on a given piece of DNA sequence, while only allowing for a *single* interaction per site.

We normalize $\mathcal{R}_{\mathrm{MWM}}$ based on the total number of edges in the graph $|E| = \frac{|V|^2-|V|}{2}$ and the maximum number of edges of a matching of a given graph $|M| = \lfloor \frac{|V|}{2} \rfloor$. This leads to

$$\mathcal{R}_{\mathrm{MWM}}^{\mathrm{norm.e}} := \frac{\mathcal{R}_{\mathrm{MWM}}}{|E|}. \tag{5.2}$$

and

$$\mathcal{R}_{\mathrm{MWM}}^{\mathrm{norm.m}} := \frac{\mathcal{R}_{\mathrm{MWM}}}{|M|}. \tag{5.3}$$

Summation of Unique Edges

Let $G = (V, E, w)$ be a complete weighted TFBS graph built from the TFBS annotation of a DNA sequence, and S the co-occurrence scores for the corresponding PWM set. Given two PWMs i and j with co-occurrence score $S_{i,j}$ and their types t_i and t_j the type of an edge is defined as $b = (t_i, t_j)$. The respective edge weight is $w(b) = S_{i,j}$. The set of present positive edge types $\mathcal{E}^+$ contains every edge type, that is present once or more in the graph G and has a positive edge weight.

We define the regulatory potential $\mathcal{R}_{\mathrm{SUE}}$ as

$$\mathcal{R}_{\mathrm{SUE}} := \sum_{b \in \mathcal{E}^+} w(b). \tag{5.4}$$

The design of $\mathcal{R}_{\mathrm{SUE}}$ is motivated by the common occurrence of homotypic clusters of predicted TFBS for some PWMs. This can be either due to real binding of multiple transcription factors of the same type close to each other or to a low specificity of a PWM leading to many predictions in a given region. $\mathcal{R}_{\mathrm{SUE}}$ evaluates the combination of TFBSs in a piece of sequence, but at the same time makes sure that homotypic clusters of one or more TFBS types do not influence the result too strongly.

The normalization for $\mathcal{R}_{\mathrm{SUE}}$ is based on the total number of edges in the graph $|E| = \frac{|V|^2-|V|}{2}$. This leads to

$$\mathcal{R}_{\mathrm{SUE}}^{\mathrm{norm.e}} := \frac{\mathcal{R}_{\mathrm{SUE}}}{|E|}. \tag{5.5}$$

The set of present positive edge types $\mathcal{E}^+$ for the example graph in Figure 5.1 contains the edge types red:blue (1.1), red:yellow (2.5), red:red (1.5), and yellow:blue (0.9), resulting in a regulatory potential $\mathcal{R}_{\mathrm{SUE}} = 6.0$.

Summation of All Edge Weights

A complete weighted TFBS graph $G = (V, E, w)$ is built using the TFBS annotation of a piece of sequence and the set of co-occurrence scores S for all possible PWM pairs i, j for that set. Afterwards the weights of all edges in the graph G are summed up. The regulatory potential $\mathcal{R}_{\mathrm{SAE}}$ is then defined as the weight of the complete graph:

$$\mathcal{R}_{\mathrm{SAE}} := \sum_{e \in G} w(e) \tag{5.6}$$

This way, the score accounts for all potential interactions of a transcription factor. Because $\mathcal{R}_{\mathrm{SAE}}$ contains negative edge weigths, TFBS combinations underrepresented in the training set are penalized.

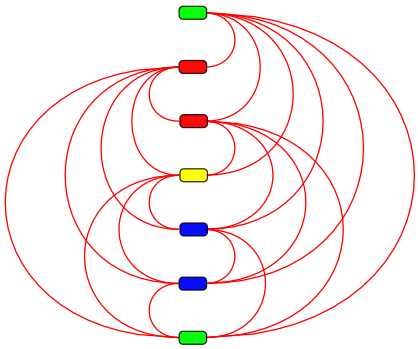


Figure 5.4: Example for regulatory potential $\mathcal{R}_{\mathrm{SAE}}$ on the example from Figure 5.1: all edges are taken into account for the calculation of $\mathcal{R}_{\mathrm{SAE}}$. Summing up all edge weights from Figure 5.1 weight is $\mathcal{R}_{\mathrm{SAE}} = 1.9$.

The normalization for $\mathcal{R}_{\mathrm{SAE}}$ is based on the total number of edges in the graph $|E| = \frac{|V|^2-|V|}{2}$.

The normalized regulatory potential is then

$$\mathcal{R}_{\mathrm{SAE}}^{\mathrm{norm.e}} := \frac{\mathcal{R}_{\mathrm{SAE}}}{|E|}. \tag{5.7}$$

Summation of All Positive Edge Weights

A complete weighted TFBS graph $G = (V, E, w)$ is built using the TFBS annotation of a DNA sequence and the co-occurrence scores S for the corresponding PWM set. Afterwards the weights of all edges in the graph G are summed up. The regulatory potential $\mathcal{R}_{\mathrm{SAPE}}$ is than defined as the sum of all positive edge weights:

$$\mathcal{R}_{\mathrm{SAPE}} := \sum_{e \in G} |w(e)|. \tag{5.8}$$

This way all potential interactions of a transcription factor that are overrepresented in the training set, are accounted for.

The normalization for $\mathcal{R}_{\mathrm{SAPE}}$ is based on the total number of edges in the graph $|E| = \frac{|V|^2-|V|}{2}$.

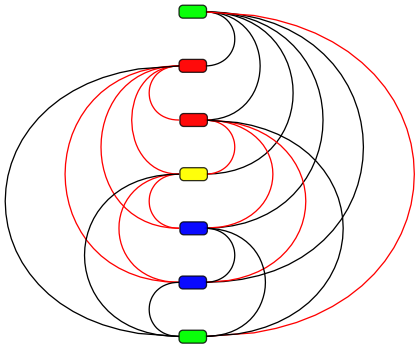


Figure 5.5: Example for regulatory potential $\mathcal{R}_{\mathrm{SAPE}}$ on the example from figure 5.1: all positive weight edges are taken into account for the calculation of $\mathcal{R}_{\mathrm{SAPE}}$. Summing up all positive edge weights from Figure 5.1 weight is $\mathcal{R}_{\mathrm{SAPE}} = 7.1$.

The normalized regulatory potential is then

$$\mathcal{R}_{\mathrm{SAPE}}^{\mathrm{norm.e}} := \frac{\mathcal{R}_{\mathrm{SAPE}}}{|E|}. \tag{5.9}$$

Maximum-Weight Bipartite Matching

A bipartite weighted graph $G = (U, V, e, w)$ is built based on the set of predicted TFBS $\mathcal{T}$ for a piece of DNA sequence and the set of co-occurrence scores S for all possible PWM pairs i, j. Each $t \in \mathcal{T}$ is represented by a vertex u in partition U and by a vertex v in partition V. In the following a maximum-weight bipartite matching is calculated on G. The regulatory potential $\mathcal{R}_{\mathrm{MBPM}}$ is the weight of the resulting matching M:

$$\mathcal{R}_{\mathrm{MBPM}} := \sum_{e \in M} w(e). \tag{5.10}$$

Like in the maximum-weight matching case, one edge per vertex is allowed. The co-occurrence score represents the importance of an interaction between two TFBSs. $\mathcal{R}_{\mathrm{MBPM}}$ then takes into account the combination of most important interactions of each site on a given piece of DNA sequence, while in principle allowing for a *two* interactions per site. Below, in Section 5.1.4 we will show, that in practice only one interaction is taken into account.

For the normalization two measures are possible, the number of edges in the bipartite graph

$$|E| = 4 \cdot |\mathcal{T}|^2 - 2 \cdot |\mathcal{T}| \tag{5.11}$$

and the number of possible edges in the matching

$$|M| = \begin{cases} |T| & if\ |T| \text{ is even} \\ |T| - 1 & if\ |T| \text{ is odd} \end{cases} \tag{5.12}$$

.

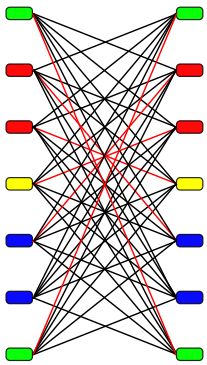


Figure 5.6: Example for regulatory potential $\mathcal{R}_{\mathrm{MBPM}}$ on the example from Figure 5.2: the edges part of the matching, the weight of which is used, are marked in red. Note that the same edge types like in the $\mathcal{R}_{\mathrm{MWM}}$ are used twice each, resulting in a weight of the matching $\mathcal{R}_{\mathrm{MBPM}} = 4.6$.

The normalized regulatory potentials follow as

$$\mathcal{R}_{\mathrm{MBPM}}^{\mathrm{norm.e}} = \frac{\mathcal{R}_{\mathrm{mbpm}}}{|E|} \tag{5.13}$$

and

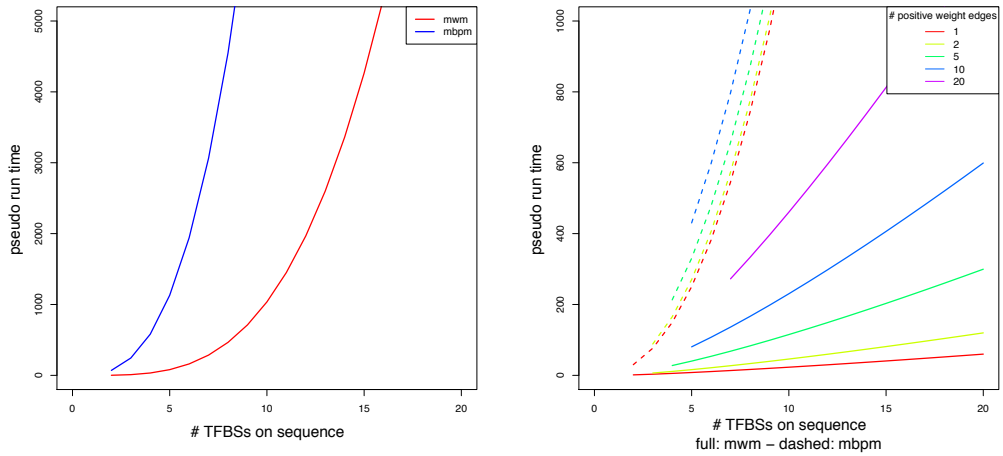
$$\mathcal{R}_{\mathrm{MBPM}}^{\mathrm{norm.m}} = \frac{\mathcal{R}_{\mathrm{mbpm}}}{|M|} \tag{5.14}$$

5.1.4 Equivalence and Run-time Comparison of $\mathcal{R}_{\mathrm{MWM}}$ and $\mathcal{R}_{\mathrm{MBPM}}$

On the same sequence and the same set of annotated TFBSs the results for $\mathcal{R}_{\mathrm{MWM}}$ and $\mathcal{R}_{\mathrm{MBPM}}$ are equivalent. While for $\mathcal{R}_{\mathrm{MWM}}$, one vertex per TFBS exists, in $\mathcal{R}_{\mathrm{MBPM}}$ one TFBS has assigned two vertices in the different partitions. In both cases the combination of edges which are part of the matching has the highest possible sum of edge weights, when only one edge per vertex is allowed. For the maximum-weight bipartite matching on the TFBS graph in the case of $\mathcal{R}_{\mathrm{MBPM}}$ the edges are symmetric: if an edge connects a vertex $u_i \in U$ (of TFBS type i) and a vertex $v_j \in V$ (of TFBS type j), there also exists an edge connecting a vertex $u_j \in U$ with vertex $v_i \in V$. Both edges have the same weight, since the co-occurrence score is symmetric ($S_{i,j} = S_{j,i}$). Due to its design, the bipartite weighted TFBS graph can be viewed as a duplication of the general weighted TFBS graph. This leads to a perfect correlation of the regulatory potentials $\mathcal{R}_{\mathrm{MWM}}$ and $\mathcal{R}_{\mathrm{MBPM}}$. In our examples in Figures 5.3 and 5.6 the regulatory potential $\mathcal{R}_{\mathrm{MBPM}}$ is exactly twice as big as the regulatory potential $\mathcal{R}_{\mathrm{MWM}}$.

As stated above, the worst-case time complexity for the maximum-weight matching algorithm is $O(n \cdot m \cdot log(n))$ while in the bipartite case the worst-case time complexity is $O(n \cdot (m + n \cdot log(n)))$ (see Section 3.5.2). For the bipartite version of the binding site graphs every TFBS is assigned to two vertices in the graph, naturally resulting in more edges in the bipartite graph as well. For a given TFBS set the weights of the resulting

matchings just differ by a scaling factor and are perfectly correlated. This is easily understood, when considering that in the bipartite version the most favourable edge for a vertex in a matching will correspond to the same combination of TFBSs in the simple graph. To assess the performance of the algorithms on typical problem sizes, we use the $O()$ notations of the complexity as formulas to compare pseudo run time. In Figure 5.7 a) we plot the time consumption calculated from the algorithm complexity for maximum-weight matching and maximum-weight bipartite matching on complete graphs of different sizes. The time consumption for maximum-weight matching is always lower. Calculating time consumption for example graphs with only a small amount of positive weight edges (Figure 5.7 b)) produces similar results.
When we apply the regulatory potentials in Chapter 7, we do not use $\mathcal{R}_{\text{MBPM}}$, because it produces the same results as $\mathcal{R}_{\text{MWM}}$, but has a less favorable run-time.



a) *b)*

Figure 5.7: Comparison of the time complexity for MWM and MBPM algorithm on TFBS graphs. a) Comparison for equal number of TFBSs $\mathcal{T}$. The pseudo run time for the maximum-weight matching is drawn in red, the pseudo run time for the maximum bipartite matching is drawn in blue. For the general graph the number of vertices is $n = |\mathcal{T}|$ and the total number of edges is $m = \frac{n^2-n}{2}$. In the bipartite graph the number of vertices is $n = 2 \cdot |\mathcal{T}|$ and the number of vertices is $m = n \cdot (n-1)$. The pseudo run time for the MWM algorithm is always lower than the one for the MBPM on the same problem. b) Comparison for different number of edges ($\widehat{=}$ number of TFBS combinations with a co-occurrence score > 0). Pseudo run time for maximum-weight matching (full lines) and maximum-weight bipartite matching (dashed lines). Numbers are not plotted for combinations of n and m which correspond to impossible graphs. In all cases the maximum-weight matching has a more favourable run time.

5.1.5 Implementation

We implemented the binding site graph routines in Perl (version 5.8.4) and C++, using the GNU compiler collection version 3.2.1. We used the implementation of the maximum weight matching and the maxmimum weight bipartite matching from the LEDA C++ library [222]

version 4.4.
Compared to Edmonds' implementation presented in Section 3.5 with a complexity of $O(n^3)$ the implementation present in LEDA has a complexity of $O(n \cdot m \cdot log(n))$. The gain of performance is primarily due to the heavy use of concatenable priority queues, especially for the representation of surface blossoms and trees.

5.1.6 Summary

In this section we presented the concept of transcription factor binding site graphs. We described the construction of a binding site graph from predicted TFBSs and co-occurrence scores, as presented in Chapter 4. Subsequently we presented a variety of regulatory potentials, which describe the abundance of TFBS combinations typical for regulatory regions based on binding site graphs. We will apply the various regulatory potentials for the prediction of regulatory regions in Chapter 7.

Part III

Applications

Chapter 6

Prediction of Transcription Factor Interactions

In Chapter 4, we presented a new method for predicting TF interactions. Here we assess the performance of our method and predict yet unknown transcription factor interactions. Moreover we examine the dependencies between TF interactions and similarity of the respective binding sites. In Section 6.1 we perform simulations and evaluate the sensitivity of our method. Section 6.2 contains the application of the method on yeast regulatory sequences. In Section 6.3 we apply the method on genome wide human sequences, in Section 6.4.2 on upstreams of genes expressed in human embryonic kidney cells, and in Section 6.4.3 on tissue specific sequence sets from mouse. Section 6.5 contains a performance comparison with the theoretical measure presented in Pape et al. [253].

6.1 Detection of Overrepresented PWM Pairs in Simulated Datasets

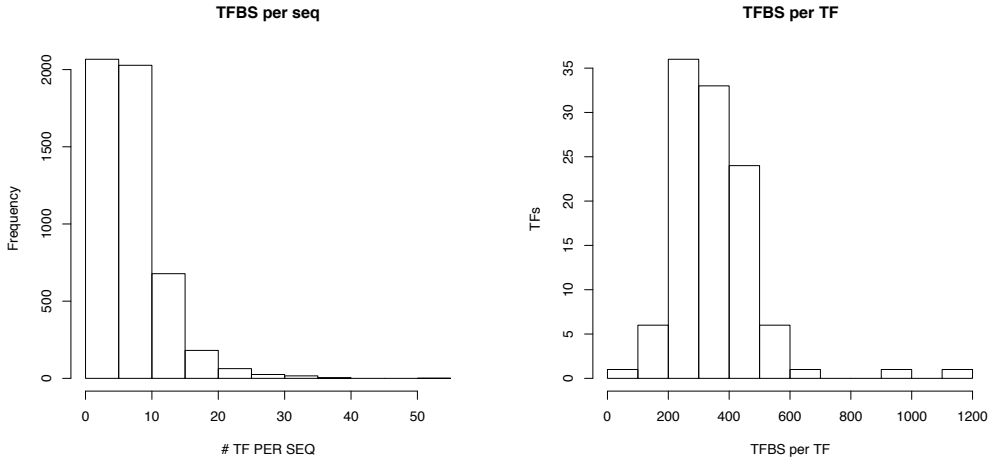
6.1.1 Synopsis

We perform a simulation study to assess the performance of the co-occurrence score described in Section 3.2.2. We generate random annotation sets that preserve the positions of binding sites and the overall abundance of each TFBS type. We implant interacting pairs of TFs at varying proportions of the individual TFBSs. Afterwards we evaluate the rank of the co-occurrence score achieved by the overrepresented TFBS pair. Varying the fraction of the overrepresented TFBS pair we are able to estimate the sensitivity of our co-occurrence score.

6.1.2 Simulation of a PWM Annotation Set

To preserve properties typical for real datasets, we simulate an annotation set based on the TFBS positions from the annotation of a yeast promoter set. This scheme maintains the number of binding sites per individual TF, the distribution of binding site counts per promoter sequence, and their clustering properties. The yeast promoter set and consequently the simulated dataset contains 37,955 TFBSs for 109 different TFs in 5,062 sequences of length 300bp. Figure 6.1 summarizes the distribution of TFBS on the set of sequences. The majority of sequences contains less than 15 predicted binding sites. The median count of

binding sites predicted per TF is 337, with two outliers possessing more than 700 predicted binding sites. Keeping the binding site positions fixed, we permute the binding site labels. Thus we obtain a random annotation set with the same number of each TFBS as well as binding site clustering properties from the yeast annotation.



a) *b*)

Figure 6.1: Distribution of TFBSs in the artificial dataset. a) histogram of TFBSs per sequence: most sequences harbour less than 15 TFBSs b) histogram of TFBSs per TF: the median number of TFBSs predicted per TF is 337. Only two outlier TFs have more than 700 predicted TFBSs.

We generate the artificial set by drawing TFBS positions without replacement and assigning them to TFBS. We proceed in two steps:

1 positions of pairs to be enriched: for each pair of TFBSs to be enriched, we draw a single position and implant a second position for the interaction partner exactly 20bps apart. We remove a random position in the same sequence to compensate for the newly created binding site.

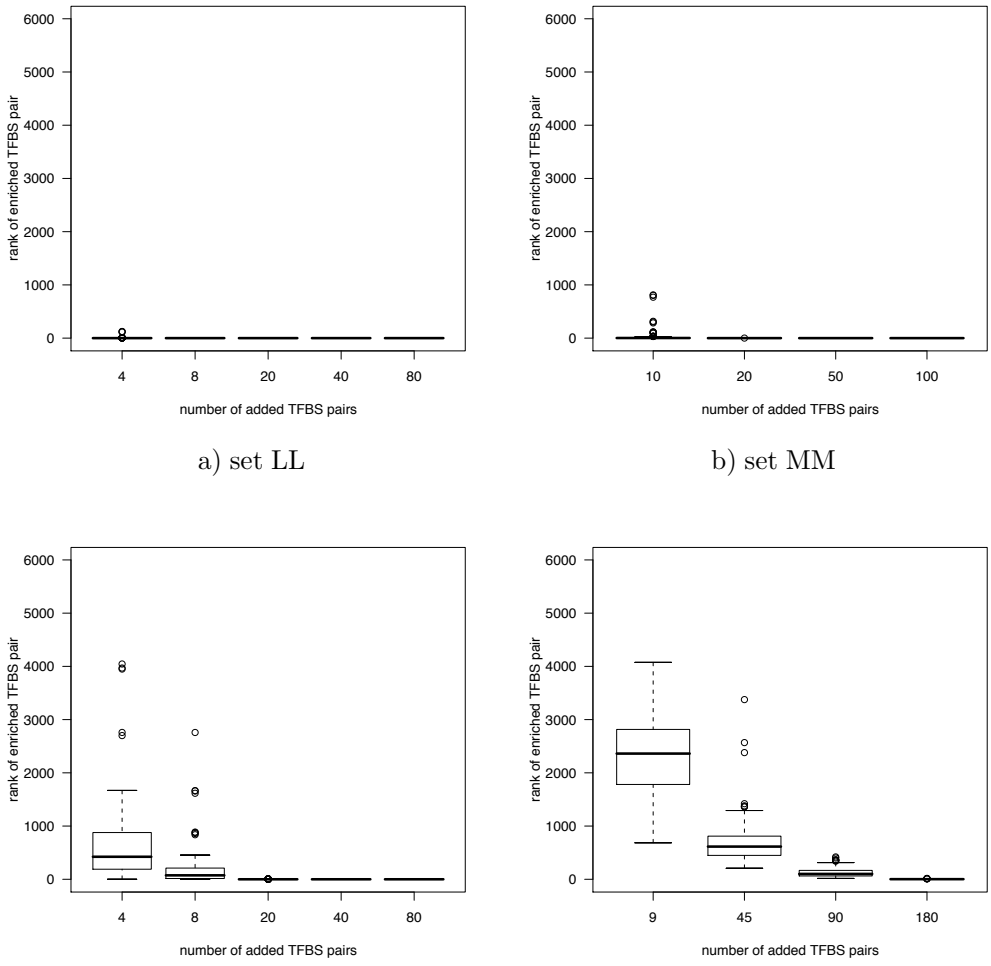
2 all other TFBSs: we assign all remaining positions randomly according to the remaining TF counts in the set.

The co-occurrence score for a pair will be influenced by the proportion of the individual TFBSs in the data set. We select four scenarios of TFBS pairs:

- *set LL*: both TFs occur rarely; TF 108 (82 TFBSs) and TF 41 (108 TFBSs).
- *set MM*: both TFs occur in medium numbers; TF 74 (224 TFBSs) and TF 65 (225 TFBSs).
- *set LH*: one TF occurs rarely, one is abundant; TF 61 (906 TFBSs) and TF 108 (82 TFBSs).
- *set HH*: both TFs are abundant; TF 61 (906 TFBSs) and TF 22 (1154 TFBSs).

Furthermore we set the fraction of TFBS pairs to various levels relative to the individual occurrence of the less abundant TF.

6.1.3 Co-occurrence Scores for Artificially Enriched PWM Pairs



a) set LL

b) set MM

c) set LH

d) set HH

Figure 6.2: Boxplots containing the ranks of co-occurence scores, that the enriched pairs obtain in a random background across 100 repetitions. Each subfigure contains boxplots for variable enrichment of TFBS pairs, dependent on the individual TFBS counts of the TFs part of a pair. A low rank represents a high co-occurrence score. The higher the co-occurrence score, the easier is the detection of the enriched pair based on the score. a) LL set: combination of two TFs with low numbers of TFBSs. b) MM set: combination of two TFs with medium numbers of TFBSs c) LH set: combination of a TF with low number of TFBSs with a TF with a high number of TFBSs d) HH set: combination of a TFs with high numbers of TFBSs.

We generate 100 artificial datasets for each of the scenarios from Section 6.1.2 at various enrichment levels and subsequently calculate co-occurrence scores. To test if we can detect an artificially enriched TFBS pair using the co-occurrence score, we look at its rank distribution across the simulations.
Figure 6.2 contains boxplots for the four scenarios at different enrichment levels.

- *set LL*, Figure 6.2 a): The numbers of enriched TFBS pairs which we tested are 4, 8, 20, 40, and 80, which corresponds to roughly 5%, 10%, 25%, 50%, and 100% of the less abundant TFBS. The enriched pair obtains best ranks in 95% of all cases.
- *set MM*, Figure 6.2 b): The numbers of enriched TFBS pairs are 10, 20, 50, and 100, which corresponds to roughly 4.5%, 9%, 22%, and 45%, of the less abundant TFBS. In all simulations except for a 8 at enrichment level 10 the enriched pair obtains best ranks.
- *set LH*, Figure 6.2 c): We test enrichment levels for the TFBS pairs of 4, 8, 20, 40, and 80, corresponding to roughly 5%, 10%, 25%, 50%, and 100% of the less abundant TFBS and roughly 0.45%, 0.9%, 2%, 4.5%, and 9% of the more abundant TFBS. Above 2% of two percent of the more abundant TFBS the enriched pair obtains the best ranks again. Lower enrichment leads to worse results.
- *set HH* Figure 6.2 d): The numbers of enriched TFBS pairs are 9, 45, 90, and 180, corresponding to 1%, 5%, 10%, and 20%. Here the enriched pair obtains best ranks only at the level of 20%. At the 10% level the enriched pair obtains acceptable ranks, while at the 5% and 1% levels the enriched pair can not be detected reliably any more.

Applying the method to count co-occurring TFBS pairs and the subsequent calculation of a co-occurrence score presented in Chapter 4, we are able to detect artifical TFBS pairs planted in a random background under multiple scenarios. The level of enrichment needed to reliably detect co-occurrence of TFBSs depends on the individual abundance of the TFBSs. While for very abundant TFBSs the method requires an enrichment at a level of 10 to 20% of total occurences (*set HH* and *set HL*), the method performs well for combinations of TFs with small or average amounts of predicted TFBSs (*set LL* and *set MM*). Here we can detect co-occurring pairs even if the fraction of co-occurring versus randomly distributed pairs is below 5%.

6.1.4 Summary

In this this section we demonstrated the performance of our co-occurrence score on simulated data. Apart from problematic TF combinations where both TFBSs are very abundant in the annotated sequences, we are able to detect the overrepresented TF pair among the highest ranks even at low enrichment levels. Based on the present results we expect that our method performs well, especially on specific TF interactions.

6.2 Predicting Transcription Factor Interactions in Yeast

6.2.1 Synopsis

In this section we use the co-occurrence score to predict TF interactions for the baker's yeast *Saccharomyces cerevisiae*. Compared to other eukaryotes, this unicellular organism has a small genome, which consists of roughly 13,000,000 base pairs (bp) and contains 6,275 genes. Its small intergenic regions as well as the extensive research carried out on yeast and its regulation make it an interesting candidate to apply our method.

6.2.2 Yeast TFBSs and Positive Interaction Set

Sequences As a putative promoter set for yeast we extract sequence regions ranging from -250 to +100 relative to the annotated transcriptional start site (TSS) of all yeast genes in EnsEMBL version 46. If the candidate region overlaps with a neighboring gene, it is trimmed at the TSS of the neighbor. We merge overlapping regions in the sequence set, such that no region occurs more than once in our input set. Subsequently we mask repetitive regions from the sequences [310]. The resulting set contains 5,408 different sequences, out of which 866 stem from merging of overlapping sequences and consists of roughly 1.5 Mbps of unmasked sequence in total. The average GC content is 37%

Binding Sites We annotate the yeast promoter sequences with putative TFBSs using two sets of motifs. The first set consists of the 124 PWMs from MacIsaac et al. [208]. To generate the second set, we apply the clustering procedure by Pape et al. [252] described in Section 3.4 at a GC content of 37%, which results in 37 different clustered PWMs.
As explained in Section 3.1, the prediction method by Rahmann et al. [270] requires a background model, which we set to the average GC content of 37%.
To be able to assess the influence of the scanning threshold on the detection of co-occurrence of predicted TFBSs, we scan with a balanced cutoff (159,363 TFBSs), and fixed type I error thresholds at false positive error rates of 0.05 (63,485 TFBSs), 0.01 (34,633 TFBSs), and 0.005 (29,131 TFBSs).

6.2.3 Known TF Interactions

To evaluate the performance of the interaction prediction, we require known TF interactions. We use the MIPS database [225] as a source for protein-protein interactions, as well as known functional interactions from TRANSFAC [219] and the interactions collected from the literature by Bar-Joseph et al. [26]. We combine these sources to 42 homotypic and 13 heterotypic unique interactions for the MacIsaac PWM set. Moreover we calculate the *common neighborhood score* described in Section 4.3.3 for the TFs based on the MIPS interactions. As a positive set we use all heterotypic TF pairs with a common neighborhood score > 0. Homotypic TF combinations naturally get a neighborhood score of 1. Of the TF combinations with a shortest path of 2 the majority have a common neighborhood score below 0.1 (Figure 6.3) The resulting positive set contains 140 heterotypic interactions with

a neighborhood score > 0.

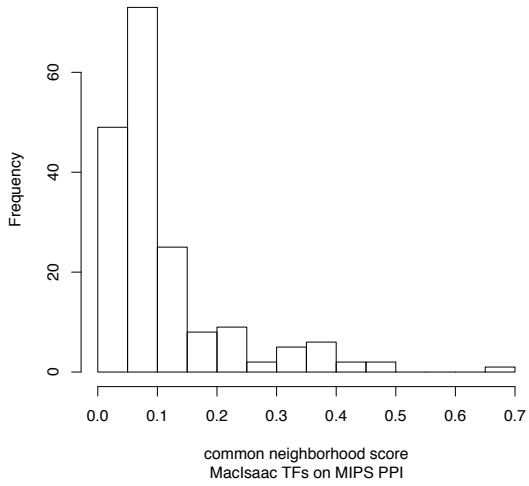


Figure 6.3: Histogram of common neighborhood scores for TF pairs from MacIsaac set. The majority of pairs have a distance > 2 in the MIPS interaction graph and do not share any neighbors. Thus they have a common neighborhood score of 0. In the histogram we omit these pairs as well as homotypic pairs with a common neighborhood score of 1, so that the histogram contains the common neighborhood scores for 140 TF pairs.

To obtain a set of positive interactions for the clustered PWMs, we map the interactions of the individual transcription factors to the clusters that represent them. The resulting positive set contains 18 known direct interactions, out of which 9 are homotyic.

6.2.4 PWM Similarity for Yeast TFs

As mentioned in Chapter 2, TFBSs are often similar to each other. This could lead to false positive predictions of heterotypic interactions in our method.

To asses the similarity of the PWMs in the MacIsaac set we calculate the empirical similarity presented in Section 4.2 and the difference in GC content ΔGC for each PWM pair for the MacIsaac PWM set. We use the TFBS annotation with a fixed false positive rate of 0.05. The majority of PWM pairs is dissimilar to each other. 97.7% of the pairs have a similarity score smaller or equal than 0.1 (Figure 6.4 b)). The set of known heterotypic interactions consists of only 15 pairs, the set of known homotypic interactions is even smaller and contains six pairs. For homotypic combinations the similarity score represents the tendency of a binding site to overlap with a second binding site of the same type. The small size of the positive set makes it hard to derive general statements about real interactions. Nevertheless we see that there is no particular tendenciy to be similar or dissimilar for the binding sites of TFs known to interact (Figure 6.4 a)).

We calculate ΔGC for a PWM pair by taking the absolute difference of the individual GC contents. As expected when comparing ΔGC to PWM similarity (Figure 6.4 b)), similar TFBS pairs have a small ΔGC. A ΔGC of 1, which would represent a pair of PWMs one of which only consists of {A,T} and the other only of {G,C}, does not exist. On the other hand a huge number of pairs have the same GC content, while not being similar. For this reason just using the GC content difference is not enough. There are no directly interacting pairs of TFs in the positive set, the PWMs of which have a GC content difference > 0.25. The PWMs representing known interactions tend to show a similar GC content while *not* recognizing the same sequences. Assuming TFBSs in the direct neighborhood to each other, this makes sense since it simplifies co-operative binding in sequence regions with a homogeneous GC content.

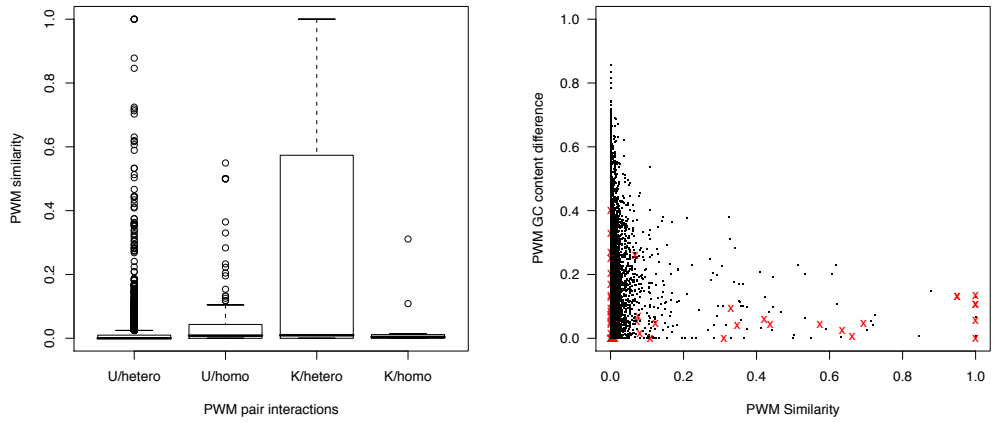
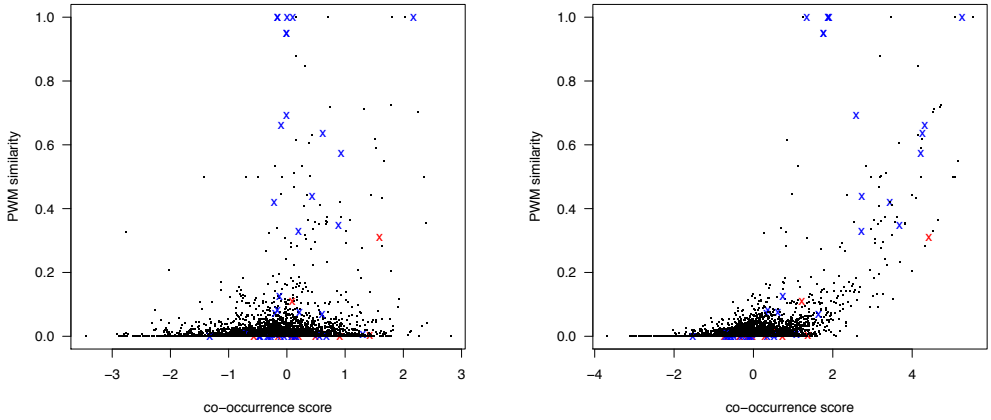


Figure 6.4: a) Boxplot of the PWM similarity for all PWM pairs in MacIsaac set. U: pairs unknown to interact; K: pairs known to interact; homo: homotypic combinations; hetero: heterotypic combinations. b) Scatterplot of PWM similarity vs. GC content difference for a pair of PWMs. Values for pairs known to interact marked in red.

Similar pairs of PWMs tend to have higher co-occurrence scores more often (Figure 6.5). The high similarity of the corresponding binding sites makes it impossible to distinguish the combination from an overrepresentation of a homotypic combination of one of the PWMs. The set of known interactions marked in red also contains PWM pairs which are overrepresented and similar, which forbids to treat highly similar overrepresented PWM pairs as false positives.

6.2.5 Influence of Window Size, Scanning Threshold, and TFBS Overlap

Here we will select the optimal set of parameters by monitoring the AUC for each parameter combination. The size of the sliding window and the scanning threshold for predicted TFBSs influences the results. We expect that small window sizes will lead to the specific detection of TF combinations, which bind to the DNA directly next to each other and possibly directly



a) *b)*

Figure 6.5: Relation beween co-occurrence scores and PWM similarity. We mark TFBS pairs known to interact with red (heterotypic) and blue (homotypic) X. The co-occurrence scores shown in this plot stem from a calculation at a window size of 100bp and a scanning threshold for a fixed false positive rate of 0.05. Figure a) contains co-occurrence scores calculated ignoring overlapping TFBS pairs. Figure b) contains co-occurrence scores calculated counting overlapping TFBS pairs

interact. Moreover, we expect that counting overlapping TFBSs during the calculation of the co-occurrence score leads to a slightly higher rate of false positive predictions due to self-similar PWMs. Ignoring overlapping TFBSs will solve this problem but on the other hand might lead to a higher false negative rate, since several classes of TFs have overlapping binding sites [176].

ROC/AUC We calculate ROC curves for the different parameter combinations as described in Section 3.6.1. Figure 6.6 a) shows the ROC curves calculated at a TFBS scanning threshold for a fixed false positive rate of 0.01 at different window sizes. Figure 6.6 b) contains the areas under the curve (AUCs) calculated for all window size and threshold combinations with co-occurrence scores calculated with overlapping TFBSs. Figure 6.6 c) contains AUCs for the same window and threshold combinations on co-occurrence scores calculated ignoring overlapping TFBSs. The AUC is bigger when using co-occurrence scores calculated using overlapping TFBSs. Some of the known interacting pairs have similar PWMs, which leads to lower scores if we ignore overlapping TFBSs in the counting process (Figure 6.4). We obtain the best AUC for a combination of a window size of 50bp and the scanning cutoff for a fixed false positive rate of 0.01. Neighboring window sizes, as well as a more stringent or a more relaxed scanning cutoff perform slightly worse. The extremely stringent cutoff of 0.005 performs worst. At the cutoff of 0.05 we also find a relatively good AUC at bigger window sizes.

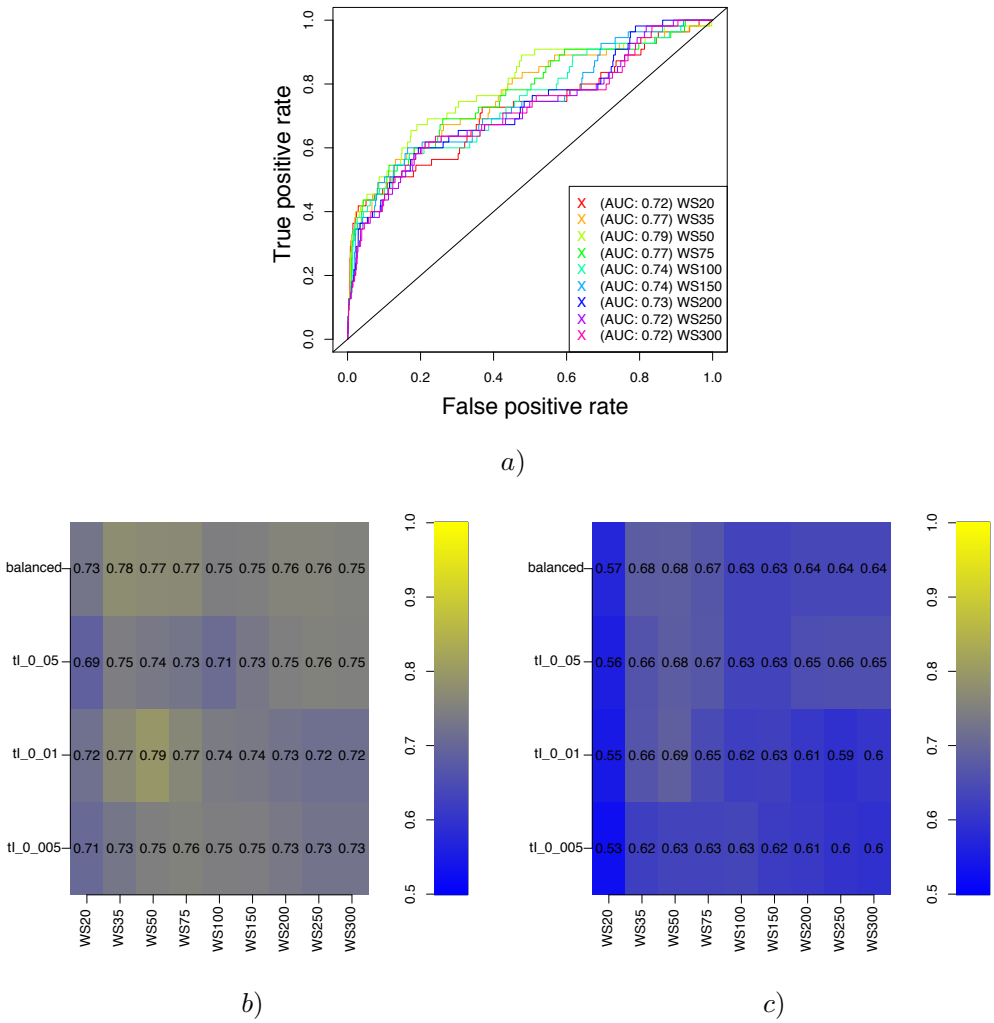


Figure 6.6: a) ROC curves for co-occurrence scores in yeast calculated at a threshold for a fixed false positive rate of 0.01 and different window sizes. As a positive set we take the co-occurrences scores for direct interactions from the MIPS database and known interactions from TRANSFAC. The negative set consists of scores for all other TF combinations. The window size for the biggest area under the curve (AUC) is 50bp. b) AUC matrix for co-occurrence scores derived counting overlapping TFBSs. We calculate ROC curves for combinations of different window sizes and scanning thresholds and color a matrix depending on the AUC of the ROC curve for the parameter combination. Blue means an AUC of 0.5, meaning that the co-occurrence score does not separate the positive set from the negative set. Yellow represents an AUC of 1. An AUC of 1 implies a perfect separation. As we expect to find real, but yet unknown interactions in our negative set, the AUC will always be lower than 1. For details see Section 3.6.1. We obtain the biggest AUC of 0.79 for the stringent scanning treshold for a fixed false-positive rate of 0.01 at a window size of 50bps. c) AUC matrix for co-occurrence scores derived ignoring overlapping TFBSs. We calculate ROC curves for combinations of different window sizes and scanning thresholds and color a matrix depending on the AUC of the ROC curve for the parameter combination. Blue means an AUC of 0.5, meaning that the co-occurrence score does not separate the positive set from the negative set. We obtain the biggest AUC of 0.69 for the stringent scanning treshold for a fixed false-positive rate of 0.01 at a window size of 50bps.

Relative Rank Sum of Known Interactions Here we apply the relative rank sum method presented in Section 4.3.2 to find optimal parameters for the calculation of the co-occurrence score. Figure 6.7 contains the relative rank sum matrix for the same data set. The parameter combination for the smallest relative rank sum is a scanning threshold of 0.01 at a window size of 50bp, counting overlapping TFBSs. The result confirms the parameter combination identified using ROC and AUC in the previous section.

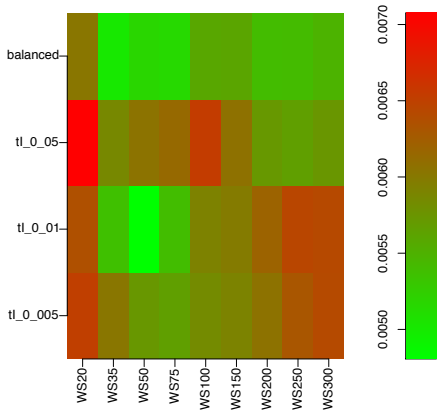


Figure 6.7: Relative rank sums for co-occurrence scores in yeast using known direct interactions from MIPS as a positive set. The relative rank sum for the positive set is best at a fixed false positive rate of 0.01 for TFBS scanning. The preferred window size is 50bp.

ROC/AUC with Common Neighborhood Score In this paragraph we test the influence of using the common neighborhood score to define the positive set of potentially interacting TFs. In contrast to the direct interaction set, it does not contain homotypic pairs. In addition to the positive set consisting only of direct interactions, it contains pairs of TFs connected with a path length of 2 in the MIPS PPI network, which share common neighbors.

The resulting AUCs in Figure 6.8 tend to be slightly smaller than the AUCs based on the direct interactions. One can speculate that a longer path between two TFs also implies more potential sources of error, thus leading to a higher noise level in the positive set. In general the results based on the common neighborhood score support the optimal parameters found using the AUC based on direct interactions and the ones found using the relative rank sum.

6.2.6 Differences Between Homotypic and Heterotypic TF pairs

Some TFBSs are known to occur in homotypic clusters. To assess whether our method is able to detect homotypic as well as heterotypic combinations, we repeat the calculation of AUCs for positive interaction sets consisting only of homotypic or heterotypic PWM pairs.

Heterotypic combinations (Figure 6.9 a)) obtain the best AUC at different parameter combinations that homotypic ones (Figure 6.9 b)). Both positive sets obtain good AUCs at small

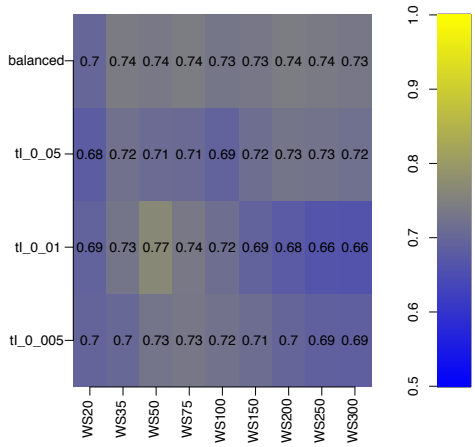
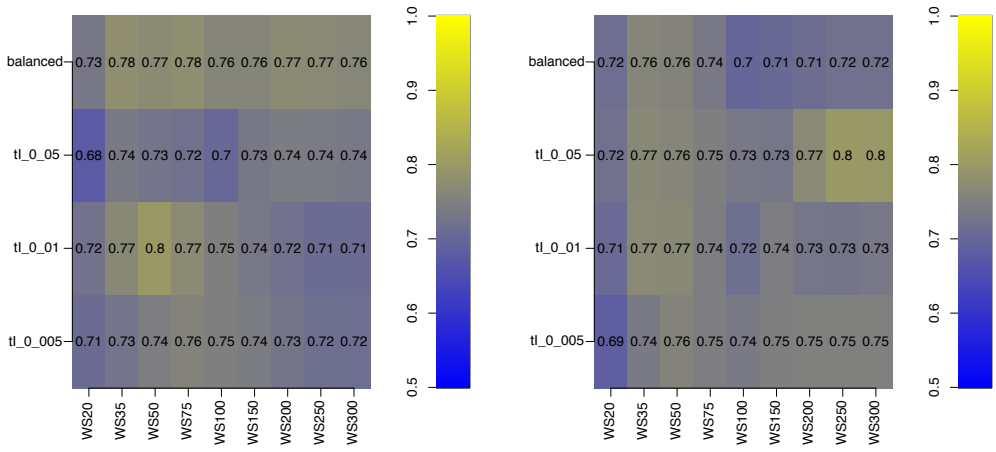


Figure 6.8: AUC matrix using common neighborhood score. We calculate ROC curves for combinations of different window sizes and scanning thresholds and color a matrix depending on the AUC of the ROC curve for the parameter combination. As a positive set we take TF pairs with a common neighborhood score > 0.25. Blue means an AUC of 0.5, meaning that the co-occurrence score does not separate the positive set from the negative set. Yellow represents an AUC of 1.



a) *b)*

Figure 6.9: AUC matrices for homotypic and heterotypic positive set: a) AUC matrix for positive set consisting only of known heterotypic TF pairs b) AUC matrix for positive set consisting only of known homotypic TF pairs

window sizes. The AUC usually does not drop below 0.7. Unexpectedly, the homotypic combinations obtain good AUCs also at window sizes of 250 and 300bp at a scanning cutoff for a fixed false positive rate of 0.05. The good results using only heterotypic pairs as a positive set show that the we do not bias the selection of parameters towards classifying only the known homotypic interactions in our positive set.

6.2.7 Over- and Underrepresented TF combinations

In the previous section we find the best AUC with a fixed false positive rate of 0.01 and a window size of 50bp. The relative rank sums supports the usage of the same values. We find that counting overlapping TFBSs leads to a better detection of known interacting TFs, no matter if the positive set consists of homotypic or heterotypic combinations only.

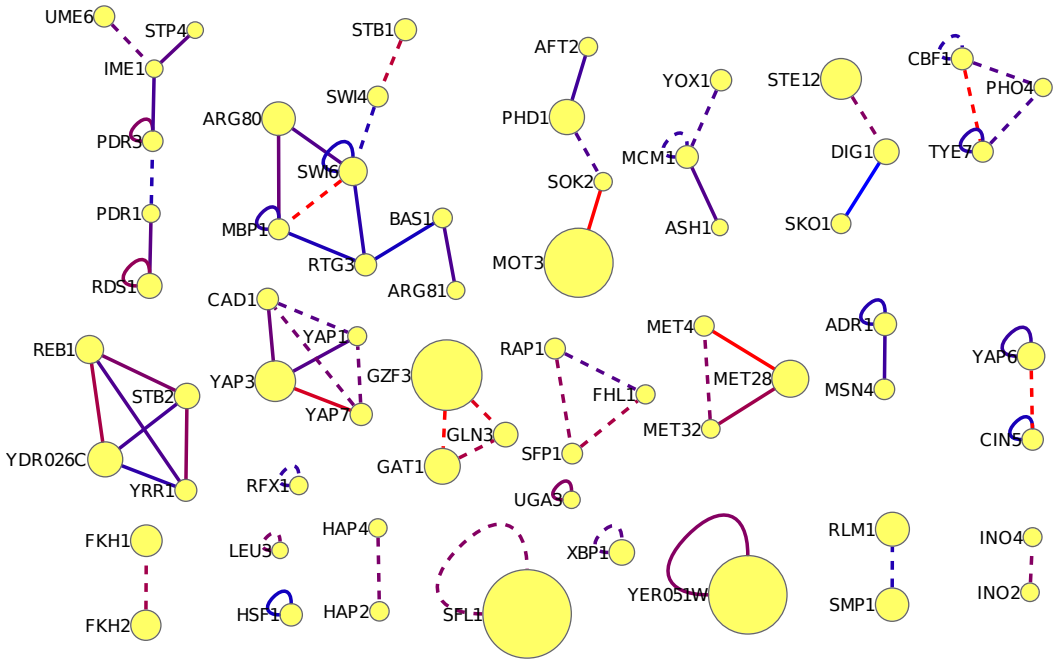


Figure 6.10: Predicted interactions of yeast TFs. TFBS co-occurrences are overrepresented for a window size of 50bp at a scanning threshold for a fixed false positive rate of 0.01. For each factor, the vertex size is proportional to the binding site counts for, ranging from 19 TFBSs for LEU3 to 1216 TFBSs for SFL1. The edges are colored according to the empirical similarity between the binding sites (from blue: no similarity to red: completely similar). Dashed edges represent interactions, for which there is experimental evidence.

Potentially Interacting TFs Figure 6.10 shows the network with the top scoring predicted TF interactions for yeast, calculated using a window size of 50bp and a TFBS scanning threshold for a fixed false positive rate of 0.01. Dashed edges represent interactions for which there is experimental evidence. Apart from the direct interactions annotated in the MIPS set and known interactions from TRANSFAC we use to calculate the ROC curves, we mark edges as dashed, if there is frequent common binding of the two TFs in Harbison et al. [142],

or if we find other experimental data in the literature (*UME6:IME1* [357], *PDR1:PDR3* [238], *MCM1:MCM1* [23], *MCM1:YOX1* [265], *PHO4:TYE7* [278], *GZF3:GLN3:GAT1* [315], *LEU3:LEU3* [107], *RFX1:RFX1* [274], *SFL1:SFL1* [250], *XBP1:XBP1* [250]).

The edge color represents the similarity of the TFBSs involved. Some PWMs belonging to overrepresented TF combinations are very similar. If at least one of the involved factors has an overrepresented homotypic interaction as well, the homotypic and the heterotypic interactions will be indistinguishable, as the binding site pairs leading to the overrepresentation might be occupied by a homotypic as well as a heterotypic combination. Examples are the pair *CBF1:TYE7* and *MBP1:SWI6*. We do not regard those cases as false positives though, because of examples from the positive set, which are TFs that can self-interact and interact with a partner possessing a similar binding site. Among the top scoring interactions are several combinations whose binding sites overlap frequently (edges in red). Partly the corresponding factors are known to interact. Lowering the co-occurrence score threshold further results in a network with many more edges.

Co-avoiding PWM pairs We also find PWM pairs that are underrepresented with respect to the expected numbers. The most underrepresented combinations with a co-occurrence score of -2 or smaller usually stem from combinations of PWMs with a small number of binding sites, for which the expected pair count is below 15, and the number of pairs found is 0. Since we can not rule out that this happens by chance, in Table 6.1 we only report pairs which are underrepresented but are present in the dataset with at least 10 TFBS pairs.
At this point we can only speculate about functional reasons for underrepresentation of TFBS pairs. We can rule out an underrepresentation due a high ΔGC of the two PWMs (maximum is 0.3). A literature survey for the underrepresented combinations did not result in explanations for the underrepresentation of the PWM combinations above. A possible explanation might be a specific function of a TF pair, that is only present in rare conditions and thus underrepresented.

TF1	TF2	Co-occurrence score
YER051W	YHP1	-1.18
DAL80	REB1	-1.17
DAL80	YER051W	-1.16
YER051W	GZF3	-1.15
MATA1	SFL1	-1.09
GZF3	SFL1	-1.09
YHP1	SFL1	-1.07
HAP5	SFL1	-1.04
HAP3	SFL1	-1.03
MOT3	SFL1	-0.98

Table 6.1: Table of co-avoiding PWM pairs. The table contains PWM combinations with the lowest co-occurrence scores while at the same time having at least ten co-occurrences in the dataset. The parameter combination is again scanning threshold for a fixed false positive rate of 0.01 and a window size of 50bp.

Co-occurrence Scores of Known Interactions Some of the interactions in our positive set obtain negative co-occurrence scores or scores around zero (Figure 6.11). While for TFs with a large number of TFBSs the results from the simulations show (Section 6.1) that for abundant TFs it is hard to obtain high scores for an individual combination, there are

some TFs like *SWI4* or *ACE2*, which are not very abundant, but obtain a negative score for a known combination. At the same time both factors have overrepresented combinations with other TFs. A technical explanation of an underrepresentation of a known interacting combination might be a too stringent cutoff for the binding site prediction, leading to a non-detection of functional TFBSs and a subsequent underrepresentation of respective TFBS combinations. Moreover, as we showed in the simulation study in Section 6.1, it is hard to detect individual interactions of factors which are abundant and promiscuous. On the other hand, a pronounced specificity of a TFBS combination can cause an underrepresentation, implying that the regulatory function of the TF pair is rarely needed and possibly even selected against in other circumstances.

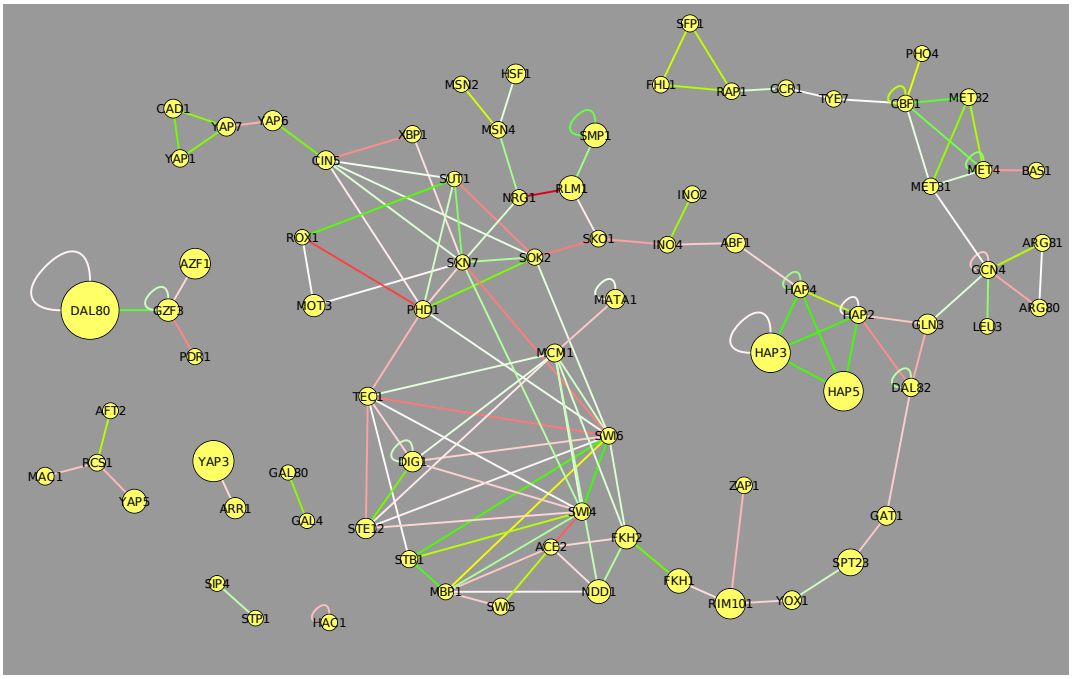


Figure 6.11: Co-occurrence score network of known interactions in yeast. For each factor the vertex size is proportional to the binding site counts. Green and yellow edges signify a high co-occurrence score and thus an overrepresentation of the respective TFBS pair. Red edges mean low co-occurrence scores and an underrepresentation of the respective pair. White means that the observed and expected number of pairs are in the same order of magnitude.

6.2.8 Co-occurence Scores for a Clustered PWM set

Figure 6.12 shows the AUC matrix calculated for the 37 clustered PWMs. We obtain the biggest AUC of 0.73 for a window size of 50bp and a fixed false positive rate of 0.05 well as for 50bp and a fixed false positive rate of 0.01 and 0.005, but the overall performance for the detection of known interactions is worse than in the unclustered case. The creation of a positive set for the clustered PWMs probably contributes to the smaller AUC here, since we map interactions of individual TFs to the clustered PWMs. The positive set is

even smaller than the original interaction set. Another factor which could play a role are different properties of the PWMs in the set. While the information content distribution of the two PWM sets is similar, the average PWM length of the original MacIsaac PWM set is 9.8, and the average PWM length for the clustered set is 13.6. This could lead to more overlapping predicted TFBSs, resulting in a weaker co-occurrence signal.

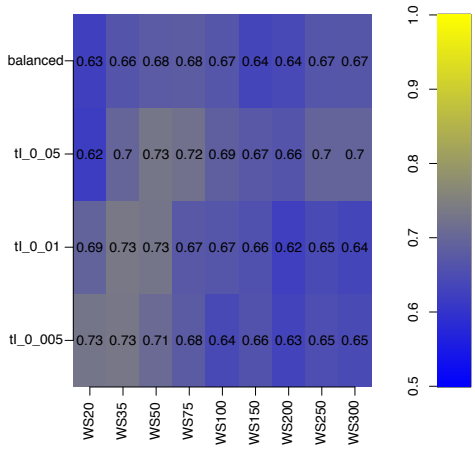


Figure 6.12: AUC matrix for ROC curves the clustered PWM set. The best AUC under the curve is 0.73. We obtain the value for window sizes smaller than 75bps and at a scanning threshold for a fixed false positive rate of 0.05 and more stringent cutoffs.

6.2.9 Summary

In this section we predicted interaction TFs for yeast. We evaluated different ways of finding optimal parameters for the calculation of the co-occurrence score: Using the common neighborhood score to extend the positive sets turned out to have no advantages over using only direct functional interactions. The relative rank sum proposed as an alternative to the calculation of the AUC from ROC curves supports the same parameter combination.
We assessed the results using known interaction sets. Clustering of PWMs resulted only in average results. Using the complete yeast PWM set, we obtained good results: The best AUC found is around 0.8. PWM similarity and the difference of the GC content of PWM pairs play a role, but do not dominate the calculation of co-occurrence scores. Known interactions of TFs with similar binding sites exist. On the other hand we find known interactions with a similar GC content but dissimilar binding sites, implying a preference of the TF pair for sequence regions with similar GC content while still recognizing specific sites. Counting overlapping binding sites during the calculation of the co-occurrence score leads to vastly improved results compared to the co-occurrence score calculated without counting overlapping pairs. Counting overlapping TFBSs allows for the detection of TF combinations, which are known to use overlapping sites to increase co-operativity and specificity, like for example homeodomain factors and nuclear receptors [176]. Using the co-occurrence score we are able to detect homotypic combinations of PWMs as well as heterotypic combinations with a good sensitivity. Using top scoring PWM pairs, we derived a predicted interaction net-

work for yeast. For the majority of predicted interactions we find support in the literature. Interestingly, we also detected underrepresented pairs in yeast promoters. Also for some of the known interactions we obtain negative co-occurrence scores. Possible explanations are specific and not widely used functions of the respective combination, large numbers of possible other interaction partners for the involved TFs, or a too stringent scanning cutoff for one of the TFs involved.

6.3 Predicting Genome-wide TF Interactions in Human

6.3.1 Synopsis

In this section we use the co-occurrence score to predict TF interactions in human. As other vertebrates, *Homo sapiens* has a much larger genome than simple eukaryotes like yeast and features a much more complex regulatory system. The genome consists of roughly 3.2×10^9 bp. The number of genes is estimated to lie between 20,000 and 25,000.

6.3.2 Human TFBSs and Positive Interaction Set

Candidate Regulatory Sequences We define putative regulatory sequences in human using size and conservation to mouse as main criteria. The suggestions for suitable promoter regions vary. While the ENCODE project finds that 67% of functional binding sites lie within 2,500bp upstream of the TSS [63], Qian et al. [268] identify the -1,000 to 0 region as the one with the highest density of known binding sites. Tabach et al. [328] detect the -200 to 0 region as the one with the highest abundance of binding sites.

We create large and a short upstream sequence set: one ranging from -1,000 to +200 relative to the most 5' TSS annotated for each EnsEMBL v. 46 gene, and the second one covering the -250 to +50 region. To prevent the inclusion of coding sequences into the set, we trim upstream regions at the start of neighboring genes. Moreover, to prevent redundancy, we remove overlapping pieces of sequence in the set. We mask repetitive regions in the resulting sequence set [310]. Conservation of regulatory regions is assumed to lower the number of false positive binding site predictions (for example see Dieterich et al. [84]), while at the same time losing binding sites which are species specific. To be able to test the influence of this parameter, we also create a conserved sequence set which contains the part of the human sequences, which are conserved to upstream regions of orthologuous mouse genes. We extract the mouse regions using EnsEMBL and identify conserved regions using blastz. The sequence sets contain between 2.5×10^6 and 20×10^6 unmasked base pairs (Table 6.2).

	short	long
number of sequences	24,419	23,622
bp complete	5,830,914	19,607,508
bp conserved	2,577,591	8,318,251

Table 6.2: Number of base pairs in human sequence sets after repeat masking.

PWM Set and Predicted TFBSs We annotate the sequence sets using the method described in Section 3.1 with the local GC content of the respective sequence as the background. The TRANSFAC database version 11.3 [219] provides a non-redundant vertebrate PWM set containing 214 matrices. Since it sometimes includes more than one PWM for a single TF, in those cases we remove PWMs in such a way that each TF has one corresponding PWM in the set. This leaves a set of 142 vertebrate PWMs which we use to annotate the sequence sets with.

In the case of yeast, the different scanning thresholds influenced the detection of interacting TFs. For that reason we annotate binding sites at various thresholds again (Table 6.3).

	balanced	type I - 0.05	type I - 0.01	type I - 0.005
short/conserved	498,381	142,260	40,014	25,011
short/complete	1,098,572	303,953	82,148	49,973
long/conserved	1,558,976	421,474	111,987	67,878
long/complete	3,649,115	975,637	252,098	149,558

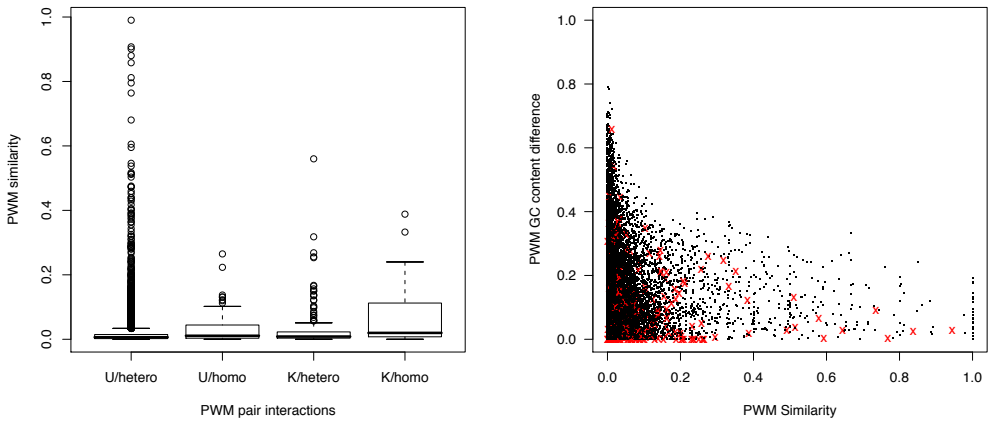
Table 6.3: Number of predicted binding sites at different scanning cutoffs in the various sequence sets.

We generate a second set of PWMs, by clustering the complete PWM set using the method from Pape et al. [252] at a GC content of 50% as described in Section 3.4. This way we obtain 45 different clustered PWMs, resulting in 226,060 (balanced), 45,061 (type I, 0.05), 12,417 (type I, 0.01), and 7,192 (type I, 0.005) predicted binding sites.

Positive Interaction Set To assess the influence of the various parameters we use known interactions annotated in the TRANSFAC database. The set contains 235 known functional interactions, consisting of 188 heterotypic and 47 homotypic PWM pairs. Some of TFs in the positive set like *SP1* or *TBP* have large numbers of known interactions. To be able to assess the prediction quality for specific interactions we generate a smaller positive set, only consisting of heterotypic interactions of factors which have at most five known interactions. We refer to the second positive set as the *non-promiscuous* positive set.

6.3.3 PWM Similarity for Vertebrate TFs

Here we examine the PWM similarity of the vertebrate PWM set described above. We calculate the empirical overlap similarity described in Section 4.2 on a TFBS set created using a fixed type I cutoff for a false positive rate of 0.05. The set of known TF interactions contains PWM pairs which are similar to each other (Figure 6.13 a). Compared to yeast (Section 6.2.4) a larger proportion of known homotypic combinations has self-overlapping TFBSs, while less similar heterotypic combinations exist. Due to potential bias in the respective positive sets, we can not draw a general conclusion out of this observation. As in yeast, the majority of PWM pairs is dissimilar to each other (Figure 6.13 b). The PWM pairs that belong to known interactions can be dissimilar as well as similar. As expected, PWM pairs which are similar to each other usually also have a similar GC content. Similar



a) *b)*

Figure 6.13: a) Box plot of the PWM similarity for all PWM pairs in the vertebrate PWM set. U: pairs unknown to interact; K: pairs known to interact; homo: homotypic combinations; hetero: heterotypic combinations. b) Scatter plot of PWM similarity vs. GC content difference for a pair of PWMs. Values for pairs known to interact marked in red.

PWM pairs normally have a low ΔGC. Pairs with a big difference in GC content are also dissimilar.

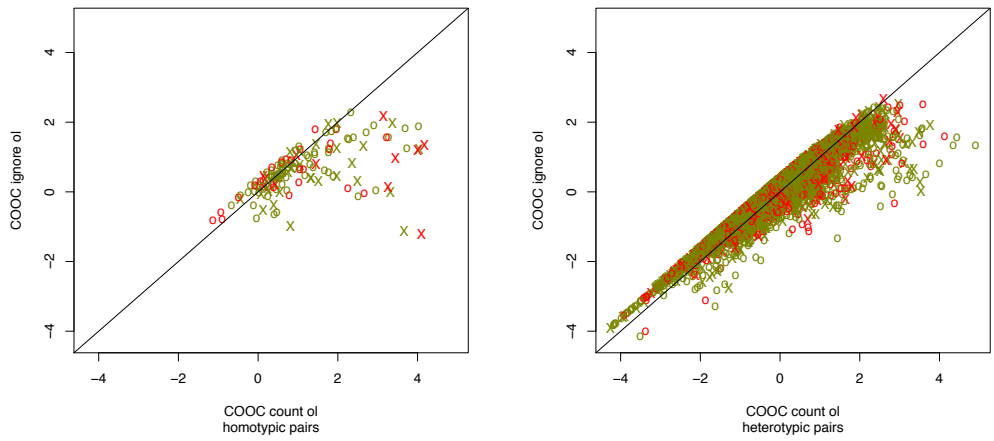
6.3.4 Counting or Ignoring Overlapping TFBSs

The co-occurrence scores calculated while counting overlapping TFBSs ("COOC/count OL") and the scores calculated ignoring overlapping TFBSs ("COOC ignore OL") are strongly correlated. While the majority of PWM pairs get the same or similar scores, some combinations get a higher COOC/count OL score (Figure 6.14). Among the pairs that obtain higher scores is no preference for similar ($SIM > 0.2$) or dissimilar ($SIM \leq 0.2$) PWM pairs. Counting overlapping TFBSs affects a bigger fraction of homotypic than heterotypic pairs, but in absolute numbers more heterotypic pairs obtain higher scores due to counting overlapping TFBSs.

The comparison of the co-occurrence vs. similarity plots for COOC/count OL and COOC/ignore OL (Figure 6.15) shows that a big part of known pairs are shifted to higher values when we count overlapping TFBSs. While there are PWM combinations that are not known to interact which also have an increased COOC/count OL score, a large part of PWM combinations does not obtain a higher score (Figure 6.14). The increase COOC/count OL score implies that overlapping TFBSs for co-operating TFs are a common situation.

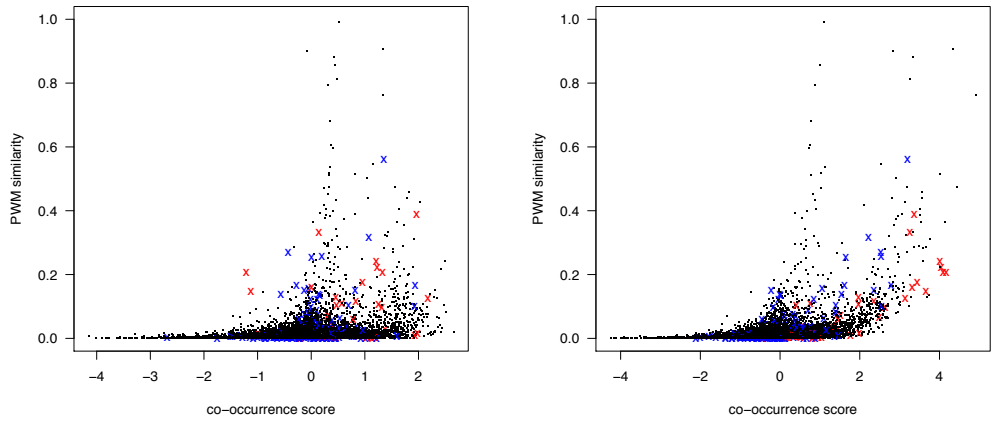
Influence of Parameters on the Detection of Known Interactions

In this section we assess the influence of various parameters on the co-occurrence score and the possibility to find TF interactions already known. As in yeast (Section 6.2.5) we test



$a)$ $b)$

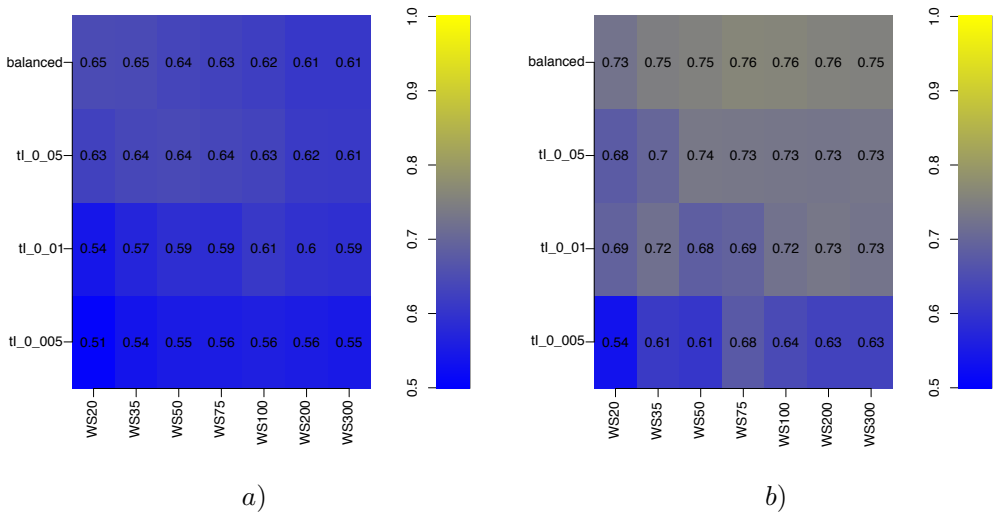
Figure 6.14: Relation between co-occurrences scores calculated counting and ignoring overlapping TFBSs. Green points represent dissimilar TFBS pairs. Red points present similar TFBS pairs. Known combinations are shown as X, while unknown combinations are shown as o. Panel a) contains homotypic and panel b) heterotypic combinations only.



$a)$ $b)$

Figure 6.15: Relation between co-occurrences scores and PWM similarity. We mark TFBS pairs known to interact with red (heterotypic) and blue (homotypic) X. The co-occurrence scores shown in this plot stem from a calculation at a window size of 100bp and a scanning threshold for a fixed false positive rate of 0.05. Figure a) contains co-occurrence scores calculated ignoring overlapping TFBS pairs. Figure b) contains co-occurrence scores calculated counting overlapping TFBS pairs

the influence of TFBS scanning threshold and window size. Moreover, for human we check the complete, as well as the non-promiscuous positive set, and the influence of sequence conservation and upstream length on the results.



a) *b)*

Figure 6.16: Choice of positive set. a) AUC matrix for all known interactions b) AUC matrix for the non-promiscuous positive set (known interactions of TFs with maximally five known partners)

The use of the complete set of known interactions as a positive set leads to mediocre results for almost all parameter combinations. The best AUC achieved is 0.65 in case of conserved, short upstreams and the COOC/count OL score (Figure 6.16 a)). For the non-promiscuous set we reach a better separation of positive and negative sets. A combination of the balanced threshold with window sizes 75bp, 100bp, and 200bp achieve an AUC of 0.76. Thus, for the further assessment of parameters we only present results for non-promiscuous positive set.

As expected from Section 6.3.4, the resulting AUCs for the four different sequence sets (complete and conserved regions of short and long upstream sizes) are always better for the COOC/count OL score (Figure 6.18) than for the COOC/ignore OL score (Figure 6.17). Surprisingly, the influence of conservation and size of the upstream regions is relatively small. The best AUC reached is around 0.75 for all sequence sets, usually also for window sizes between 75bps and 200bps for the relaxed balanced cutoff. The influence of window size and scanning threshold is similar on all of the sequence sets. In most of the cases a combination of parameters performs similar in one sequence set to the other sequence sets. The best AUC reached is 0.78 for the most stringent scanning cutoff at a window size of 75bps for short and complete upstream regions. We choose the conserved, short upstreams at a balanced cutoff and a window size of 100bps for subsequent analyses though, since we find good AUCs for this window size/threshold combination in all the sequence sets.

The relative rank sum matrix (Figure 6.19) for the conserved, short upstream set supports similar parameter combinations as the AUC, although the differences between the various

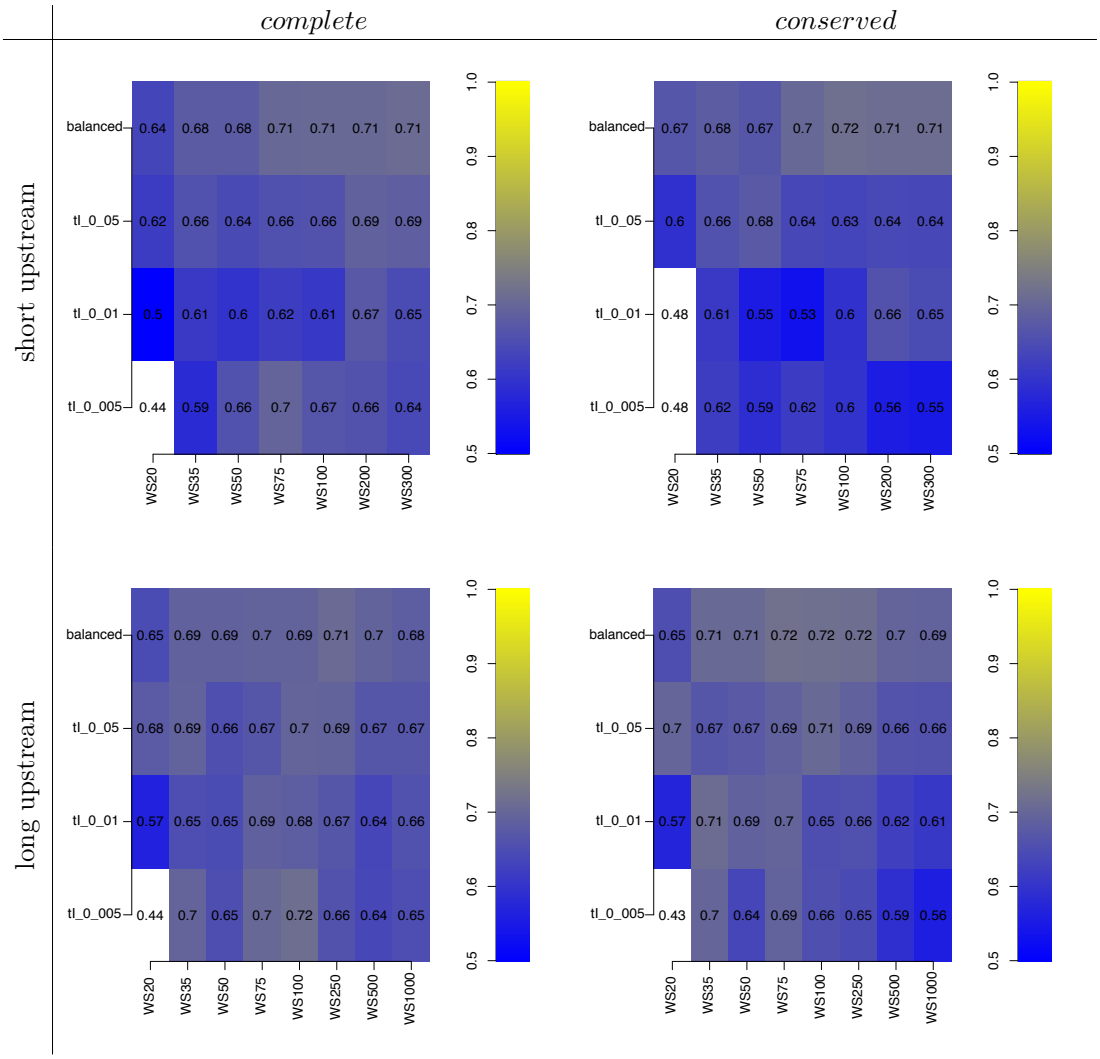


Figure 6.17: Detection of known TF interactions in human. AUC for threshold and window size combinations for four sequence sets. *Short upstream* covers the region from -250 to +50 relative to most 5' EnsEMBL annotated TSS, *long upstream* the region from -1,000 to +50. *Complete* refers to the the whole non-repeat masked sequence, *conserved* to regions in human sequence conserved to mouse. Co-occurrences scores are calculated *ignoring* overlapping binding sites.

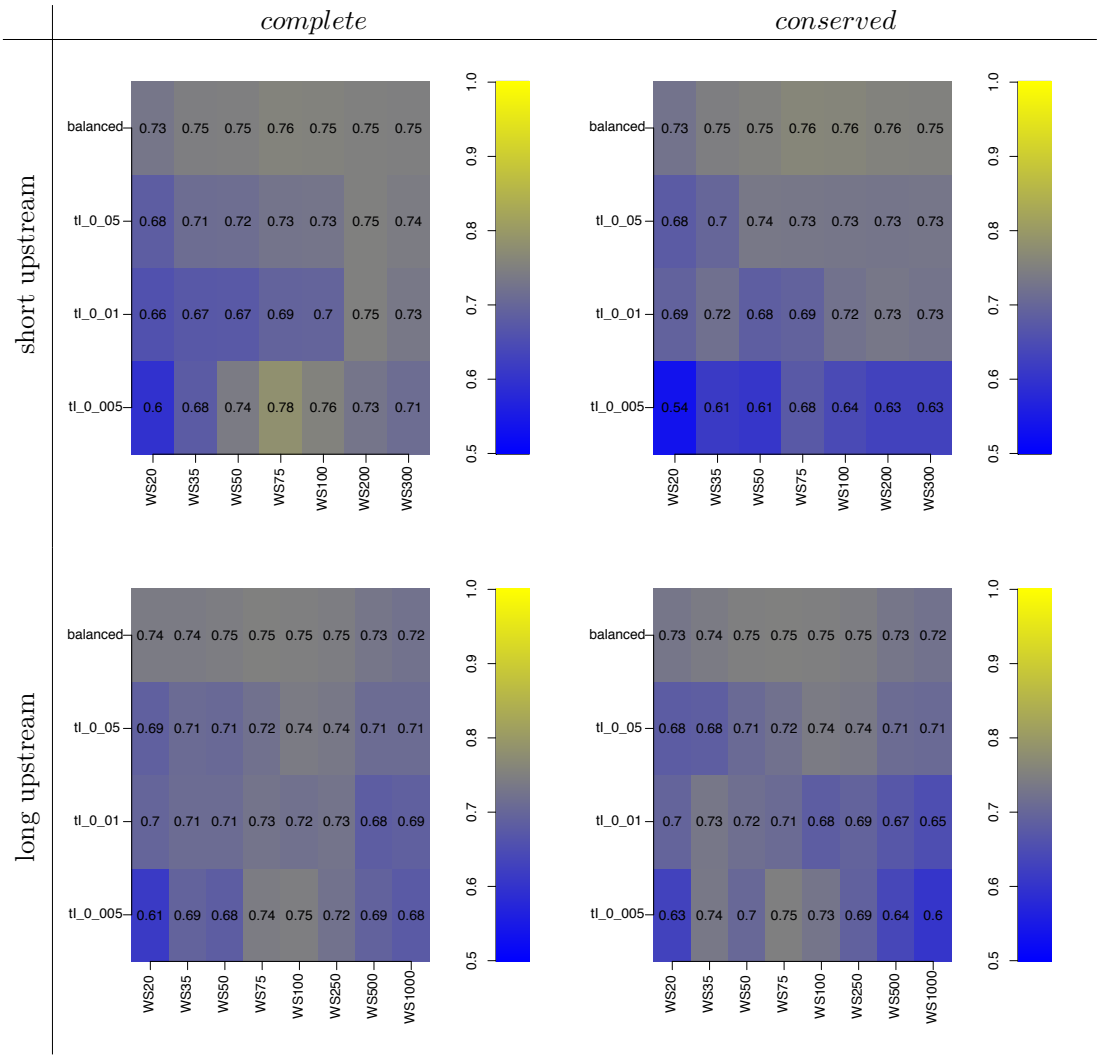


Figure 6.18: Detection of known TF interactions in human. AUC for threshold and window size combinations for four sequence sets. *Short upstream* covers the region from -250 to +50 relative to most 5' EnsEMBL annotated TSS, *long upstream* the region from -1,000 to +50. *Complete* refers to the the whole non-repeat masked sequence, *conserved* to regions in human sequence conserved to mouse. Co-occurrences scores are calculated *counting* overlapping binding sites.

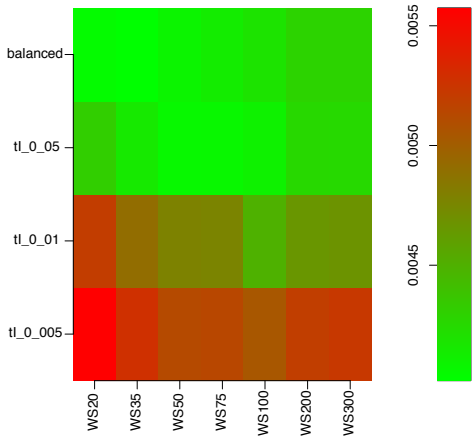


Figure 6.19: Detection of known TF interactions in human. Relative rank sum matrix for parameter combinations on short conserved upstreams.

window sizes are even smaller than for the assessment using the AUCs.

Using a window size of 100bps also captures most of the binding site distances from composite elements of the TransCOMPEL database [219]. For TransCOMPEL 10.3, of 375 known entries, 98.16% have a distance between the experimentally determined TFBS below 100bps (Figure 6.20).

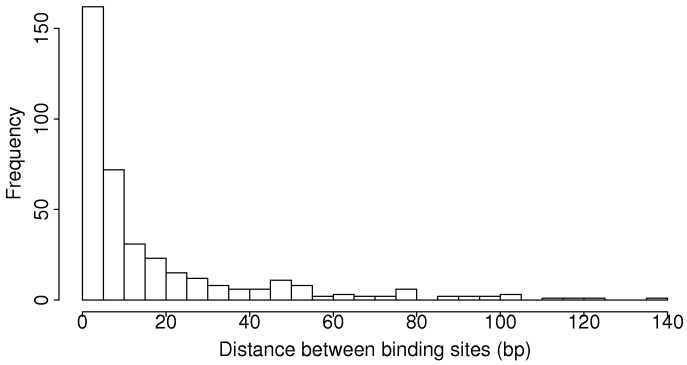


Figure 6.20: Distance distribution of transcription factor binding sites from composite elements from TRANSCompel 10.3.

Clustered PWM set Mapping the positive set for the complete PWM set onto the clustered PWM set, we calculate AUCs for the short conserved upstream regions using the COOC/count OL score (Figure 6.21). The best parameter combination obtains an AUC of 0.74 at a window size of 100bp and a balanced scanning threshold. The results are slightly worse than for the non-clustered case in Figure 6.18.

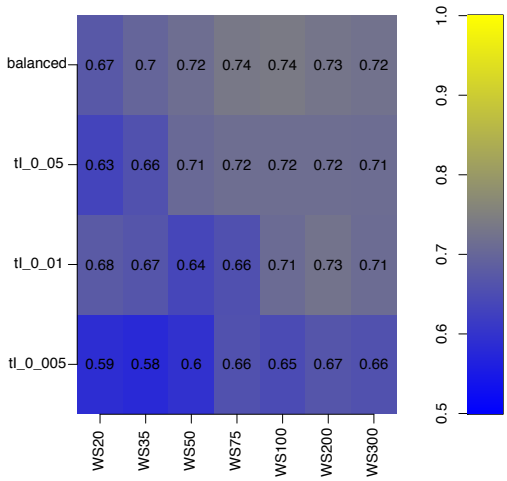


Figure 6.21: AUC for parameter combinations using the clustered PWM set for short conserved upstream regions.

6.3.5 Potentially Interacting TFs

In this section we present the TF combinations which get the highest co-occurrence score on a genome wide upstream sequence set. Figure 6.22 shows the network of top scoring pairs calculated using a balanced cutoff and a window size of 100bps on the short and conserved upstream set.

We mark known interactions with a dashed line. Apart from the interactions in the positive set, we find evidence for interactions for other top scoring pairs (CEBP:HNF3 [88]; CEBP:OCT [29]; GATA6:TBP [7]; TBP:CDXA [122]; NFAT:HMGIY [181]; PPAR:HNF4 [279]). The set of overrepresented known combinations contains some pairs which consist of similar PWMs, like *TBP:CDXA* or *DEC:EBOX*. Moreover we detect combinations that are known to be competing or repressing each others function (edges marked with *A*), like *NFAT:HMGIY* or *PPAR_DR1:HNF4*. Moreover we see a number of interactions involving homeobox TFs (e.g. *CDXA, HNF1, OCT, NKX2-5*). Although some of the involved motifs are similar to each other and the individual combinations are hard to tell apart from each other, homeodomain TFs usually form homo- or heterodimers and have binding sites found to overlap *in vivo* [176, 358].

To check whether PWM similarity results in false positive predictions of TF combinations especially with respect to the homeodomain TFs, we look at the top TF pairs of clustered PWMs (Figure 6.23). Also here the PWMs representing various homeodomain PWMs show up in overrepresented homotypic as well as heterotypic combinations. In the clustered PWM set the similarity between the different homeodomain PWMs is low. This, as well as the occurrence of overrepresented homeodomain combinations among the top COOC/ignore OL pairs (data not shown) demonstrates, that the homeodomain combinations we detect are not only artefacts due to similar, overlapping and potentially palindromic motifs.

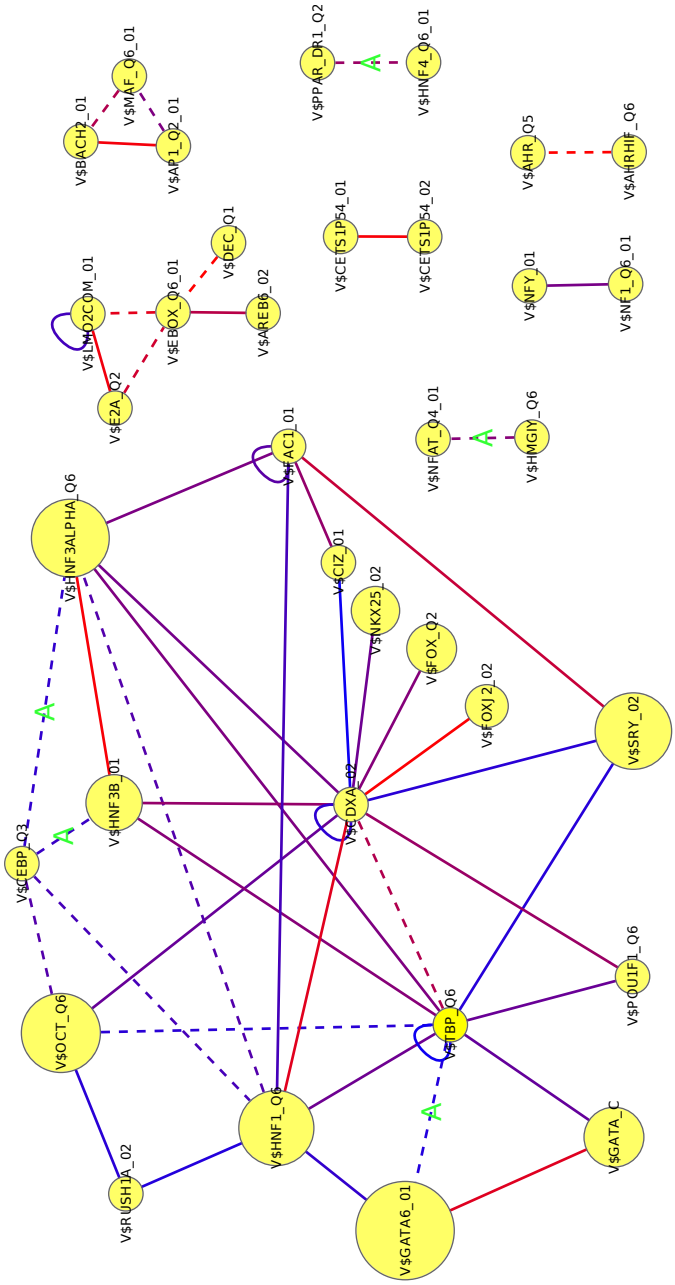


Figure 6.22: Predicted interactions in human. Additional to the top co-occurrence scores we exclude all edges with less than 1000 co-occurrences in the data set. Predicted interactions of human TFs. TFBS co-occurrences are overrepresented for a window size of 100bp at a scanning threshold for a balanced cutoff. Each edge represents at least 1,000 TFBS co-occurrences in the data set. We exclude edges with high co-occurrence scores but less than 1,000 co-occurring TFBSs in the data set. The vertex size is proportional to the individual TFBS counts for the respective factor. The edge colors come from a gradient from blue to red; blue, if the two TFBSs are not similar ($Sim_{emp}(p_i, p_j) = 0$), and red if similar ($Sim_{emp}(p_i, p_j) = 1$). Dashed Edges represent interactions, for which there is experimental evidence. Edges labelled with a green **A** represent antagonism of the involved factors and are either found to be competing for repressing each other's activity.

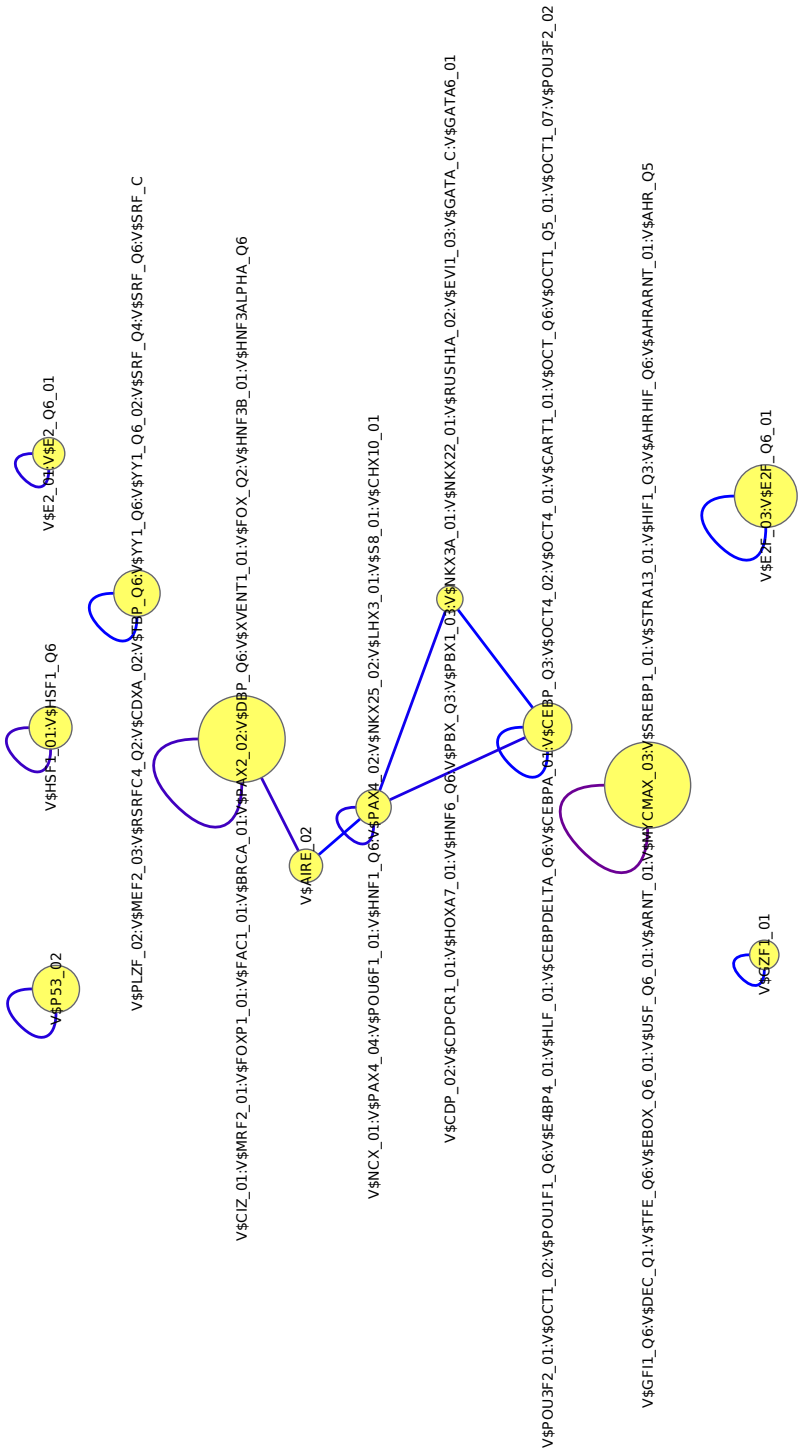


Figure 6.23: Predicted interactions in human. Predicted interactions of human TFs. TFBS co-occurrences are overrepresented for a window size of 100bp at a scanning threshold for a balanced cutoff. The vertex size is proportional to the individual TFBS counts for the respective factor. The edge colors come from a gradient from blue to red; blue, if the two TFBSs are not similar ($Sim_{emp}(p_i, p_j) = 0$), and red if similar ($Sim_{emp}(p_i, p_j) = 1$).

6.3.6 Discussion

In this section we predicted genome-wide TF interactions in human. Here we had several parameters to choose: short or long upstream regions, using conserved or complete regions, TFBS scanning threshold, window size, and counting or ignoring overlapping TFBSs in the co-occurrence calculation procedure. We selected optimal parameters based on a set of known interactions. The option to count overlapping TFBSs results in a prediction performance that is 10 percentage points better in the AUC than ignoring the overlapping TFBSs. Keeping in mind that some interacting TFs have overlapping binding sites (see for example Khorasanizadeh and Rastinejad [176] and Figure 6.13) this makes sense. Window size and scanning threshold influenced the results as well; keeping all other parameters fixed, the difference between the best and the worst value can be up to 15 percentage points in the AUC. Often the maximum AUC is achieved for window sizes below 100bps, which matches the expectations (Figure 6.20). Surprisingly, when looking at the best window size / threshold combination the influence of short or long, as well as conserved or complete upstream region only has a minor impact on the results, with a very small advantage for conserved, short upstream regions (Section 6.3.4). Confirming the results from the simulation study in Section 6.1, we find that more specific interactions are easier to detect (Figure 6.16).

The top predicted TF pairs contain many interactions from the positive set and others, for which we found evidence in the literature. Among them are interactions which have similar binding sites, some of which are known to be antagonistic. We find putative interactions for the TATA binding protein *TBP* as well as for TFs with more specific roles. 15 of the TFs involved in top scoring pairs have one predicted interaction partner, 20 have more than one predicted interactions with a maximum of 11. In the predicted interaction graph we also find cliques. But having a clique of size three for example, does not necessarily imply that the individual overrepresented TF combinations concern the same gene targets. Since we choose an arbitrary cutoff to present the top interactions, many more known or potentially interesting combinations are present but not shown.
We detect an abundance of interactions of homeodomain proteins which is surprising on the first glance. The homeodomain proteins are the second largest group of TFs [342]. They have a similar structure but distinct binding preferences [33]. While originally assumed to be mostly important in development, Morgan [229] propose a larger spectrum of function also in adulthood. Moreover, Mannervik [211] suggests that some homeodomain TFs regulate the majority of genes in *Drosophila*. Homeodomain TFs form homo- and heterodimers with overlapping binding sites [338], which achieve a higher specificity than the individual TFs [42]. This taken together could explain the high number of overrepresented homeodomain TF combinations, also involving partly similar PWMs. To make sure that we do not see artefacts caused by similar PWMs, we repeat the analysis using clustered PWMs and still find overrepresented, now almost completely dissimilar overrepresented hetero- and homotypic homeodomain TF combinations.

6.4 Prediction of TF Interactions in Specific Sequence Sets

6.4.1 Synopsis

In this section we predict TF interactions in subsets of regulatory sequences. We start with upstream sequences of genes expressed in human embryonic kidney (HEK) cells, and afterwards present results for small upstream sequence sets of genes, specifically expressed in mouse lung and testis.

6.4.2 Human Embryonic Kidney Cells

In this section we analyze upstream sequences of genes expressed in human embryonic kidney (HEK293T) cells. Sultan et al. [326] performed deep sequencing on RNAs transcribed in this cell line and find in total 12,567 genes expressed.

Sequences, TFBS Annotation, and Co-occurrence Scores

We extract the region around the most 5' TSS of every expressed gene in the set, starting at -300 and ending at +100bp relative to the TSS. Subsequently we calculate the normalized CpG content of each sequence [280] and divide the sequences into a CpG depleted set with a normalized CpG content < 0.5 and a CpG enriched set with a normalized CpG content ≥ 0.5. After trimming at neighboring TSSs, merging of overlapping sequences and repeat masking, the sequence set contains 3,259,739 non-repeat masked bases in the CpG enriched and 584,813 non-repeat masked bases in the CpG depleted set. Subsequently we annotate the sequence set using the non-redundant/unique vertebrate PWM set from TRANSFAC (v. 2008/4) consisting of 147 PWMs at a threshold for a fixed false-positive rate of 0.05. We obtain 204,219 predicted TFBSs for the CpG enriched and 32005 predicted TFBSs for the CpG depleted set. We calculate co-occurrence scores (COOC/count OK) for the annotation using a window size of 100bps. We were able to map 41 out of the 147 TFs in the PWM set to factors expressed in HEK cells.

Top Scoring TF Combinations

In the top interactions for the CpG depleted (Figure 6.25) as well as for the CpG enriched (Figure 6.26) we find TF combinations known to interact (dashed edges). We find 14 out of 45 (CpG depleted) and 11 out of 40 (CpG enriched) to be expressed in the cell line (marked in green). Finding non-expressed TFs taking part in overrepresented TF combinations can have different reasons: Since the genes with some evidence for expression in the publication by Sultan et al. [326] amount to roughly 40% of the genes annotated in EnsEMBL, a big part of the current gene set is not specific to HEK cells. Thus overrepresented TF combinations are not necessarily functional in HEK cells as well. Another cause might be an imprecise mapping of PWMs to TFs, since we only map a PWM to one TF instead of more than one. A third reason might lie in the deep sequencing data: Sultan et al. [326] estimate that between 83% and 92% of expressed genes are detectable at the given reads coverage. TFs have a tendency to be lower expressed than other genes [342], which is also true for their

expression levels in HEK cells (Figure 6.24). Thus we expect that the deep sequencing data misses a higher percentage of expressed TFs than average genes. For the PWM set we use, we find 41 out of 147 TFs with a link to a motif to be expressed, corresponding to 27.9%. If we assume a detection rate of 83% of expressed TFs, we expect at least 9TFs from our set to be expressed but not being detected by Sultan et al. [326].

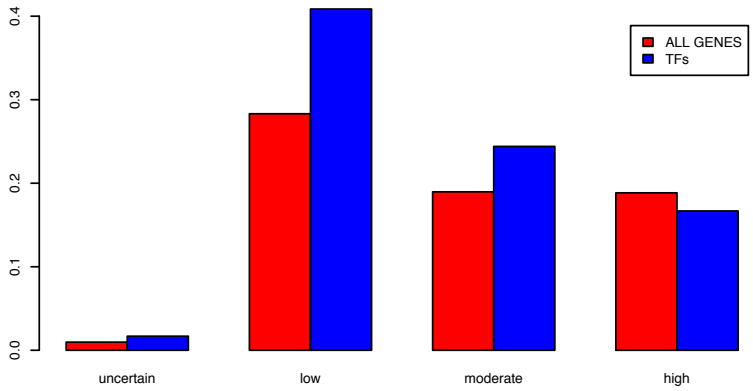


Figure 6.24: Fraction of all genes (red) / TFs (blue) that are detected as highly, moderately, lowly, and uncertainly expressed. Compared to all genes, a bigger fraction of TFs is in *low* category. The TF set consists of 1318 EnsEMBL genes with the GO term annotation GO:0003700 ("transcription factor activity"). We use the categories low, high, moderate, and uncertain from Sultan et al. [326].

Apart from interactions in our positive set from Section 6.3.2 we find experimental evidence for functional interactions for the following TF combinations (not necessarily in HEK cells) in the literature: *TGIF:MEIS1* [184]; *E4BP4:HLF* [158] (antagonism); *HMGIY:NFAT* [158]; *HNF4:PPAR* [85]; *PPAR alpha/RXR:COUP* (antagonism) [15]; *Brachyury:Brachyury* [231]; *v-JUN, ATF* self interactions [4]; *MEF2:MEF2* [37]; *SRY:SOX9* [185, 296]; *EVI1:GATA* [192]; *GATA:RUSH* [147]. Independent of the normalized CpG frequency of the set, we find comparable numbers of HEK expressed TFs among the top scoring interactions, as well as comparable numbers of known interactions. In both cases we find interactions for *HNF* factors, which play important roles in kidney [148] or *NKX3a*, known to regulate expression in the urogenital tract [318]. In both sets we also detect several interactions of homeodomain TFs, known to be play important roles in development as well as during adulthood [229].

To find out possible differences in the regulation of the CpG enriched and the CpG depleted genes, we look at the TF combinations which obtain high co-occurrence scores in one set and low in the other (Tables 6.4 and 6.5).

While there are only two PWM combinations that obtain a co-occurrence score > 2 in the CpG depleted set and a score much different ($COOC(CpG_{depleted}) - COOC(CpG_{enriched}) > 1$), there are 146 combinations which obtain higher scores in the CpG enriched set than in the CpG depleted set with a score difference of at least 1. From the PWMs part of the TF pairs in Tables 6.4 and 6.4 only V$E2_01, V$HLF_01, and V$E4BP4_01 contain

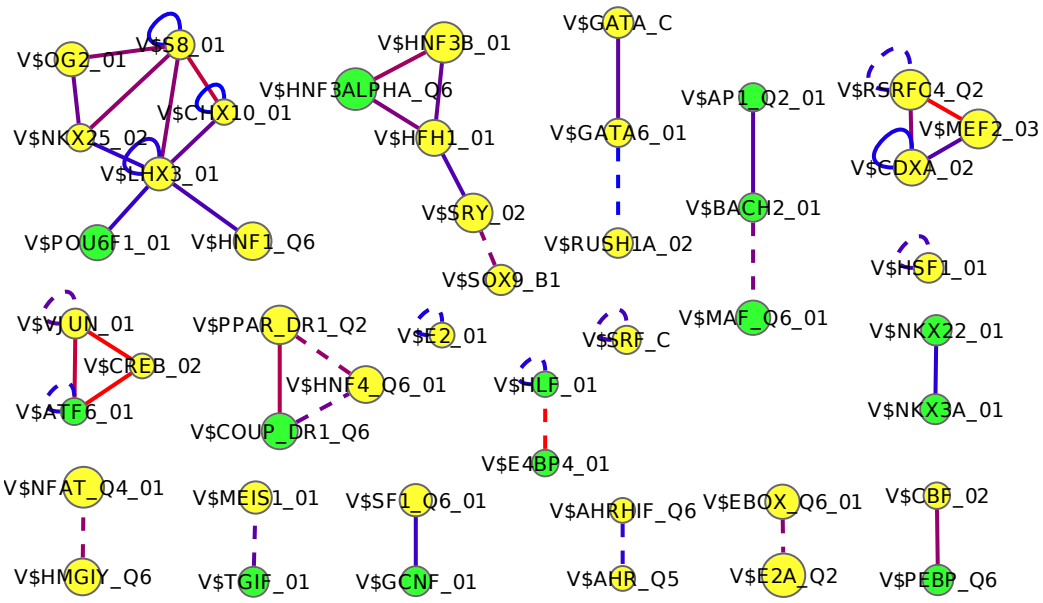


Figure 6.25: Predicted regulatory network in HEK cells, CpG-depleted upstreams. TFBS co-occurrences are overrepresented for a window size of 100bp at a scanning threshold for a fixed false positive rate of 0.05. Edge color represents PWM similarity as a gradient from blue (dissimilar) to red (similar).

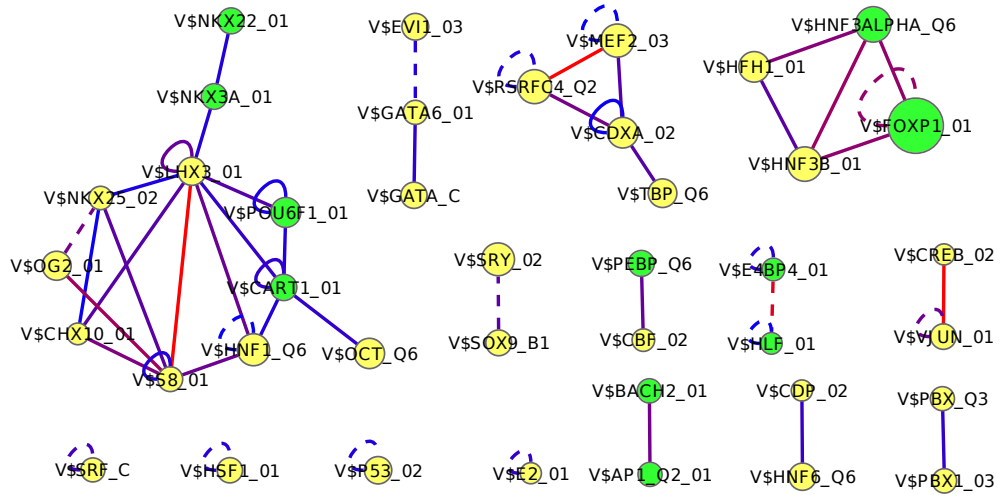


Figure 6.26: Predicted regulatory network in HEK cells, CpG-enriched upstreams. TFBS co-occurrences are overrepresented for a window size of 100bp at a scanning threshold for a fixed false positive rate of 0.05. Edge color represents PWM similarity as a gradient from blue (dissimilar) to red (similar).

PWM 1	PWM 2	COOC CpG depleted	COOC CpG enriched	Δ_{COOC}
V$POU1F1_Q6	V$RUSH1A_02	2.1	0	2.1
V$CDP_02	V$SOX9_B1	2	0.85	1.15
V$AIRE_02	V$TBP_Q6	2.1	1.15	0.95

Table 6.4: top differences of co-occurrence scores between CpG depleted and CpG enriched sequence set

PWM 1	PWM 2	COOC CpG enriched	COOC CpG depleted	Δ_{COOC}
V$E2_01	V$E2_01	3.9	0.85	3.05
V$HNF1_Q6	V$LHX3_01	3.3	0.65	2.65
V$E4BP4_01	V$HLF_01	4.9	2.2	2.7
V$FOXJ2_02	V$POU1F1_Q6	2.7	0.2	2.5
V$CDXA_02	V$TBP_Q6	4.2	1.7	2.5
V$E4BP4_01	V$E4BP4_01	3.9	1.5	2.4
V$CART1_01	V$HNF1_Q6	3.4	1.1	2.3
V$CDXA_02	V$HNF3ALPHA_Q6	2.3	0.05	2.25
V$CART1_01	V$LHX3_01	3.6	1.4	2.2
V$HNF6_Q6	V$S8_01	2.7	0.6	2.1

Table 6.5: top differences of co-occurrence scores between CpG enriched and CpG depleted sequence set

CpG dinucleotides. For both sequence sets we predict interactions for factors known to be involved in kidney regulation, like the *HNF* family or *NKX3a*. The top scoring pairs for the CpG enriched set contain a number of putative interactions for *HNF* factors, which are not overrepresented in the CpG depleted set, implying some TF-TF interactions specific for the CpG enriched set.

Discussion

In this section we calculated co-occurrence scores for upstream sequences of genes expressed in human embryonic kidney cells. Among the top scoring pairs are many known interactions, some of which play a role in kidney or embryonic development. Not all of the factors part of the predicted top interactions themselves are found expressed by Sultan et al. [326], this might be due to failure to detect transcripts of lowly expressed genes, or due to functions of a TF combinations outside of the cell line analysed. In the HEK sequence sets we also find predicted and known interactions, which are not kidney or development specific. This is due to the big number of genes expressed in the cell line, most of which are not specifically expressed. Surveying CpG depleted enriched upstreams independently, we find many top scoring combinations overrepresented in both sequence sets. Looking at overrepresented combinations specific for one of the sets only, we find only few interactions which are specifically overrepresented in the CpG depleted set, while we find almost 150 combinations, which are highly overrepresented in the CpG enriched set while obtaining much lower co-occurrence scores in the depleted set. As in the genome wide sequence set, we predict interactions between various homeodomain transcription factors. Due to their role in development this is not unexpected. Their abundance in the genome wide data suggests that those combinations are not specific for the HEK cells.

6.4.3 Tissue-specific Genes in Mouse

Roider et al. [280] analyse tissue-specific TF binding in sets of mouse expressed genes. Based on EST clusters from the GeneNest [132], they extract upstream regions of tissue-specific gene sets and further divide them into CpG depleted and CpG enriched sequence, using normalized CpG content of 0.5 as a threshold.
Here we use the sequence sets from Roider et al. [280] ranging from -200 to 0 relative to the most 5' annotated EnsEMBL TSS and annotate them with the non-redundant PWM set described in Section 6.3.2 and calculate co-occurrence scores using a balanced threshold and a window size of 100bps.

Lung

For the 156 lung specific genes in the CpG depleted category, after repeat masking the set contains 29,393bps and 5,994 predicted TFBSs. The CpG enriched set contains 345 genes, 65,603bps and 13,430 predicted TFBSs.

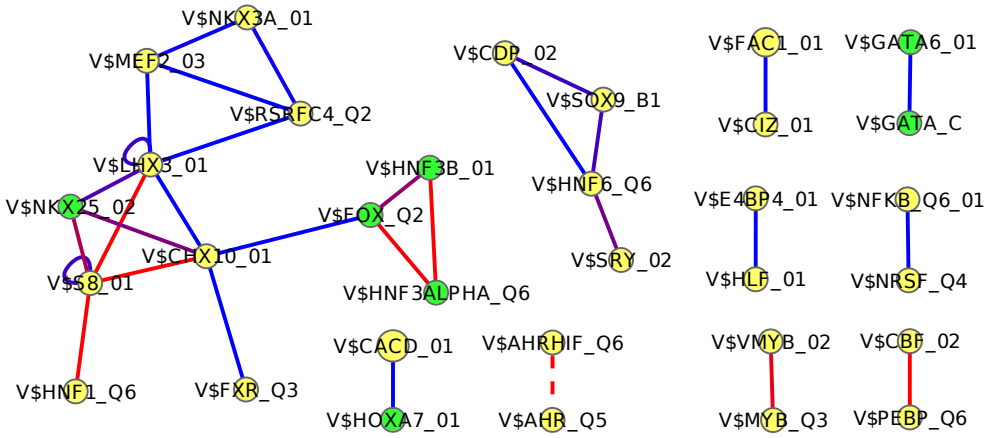


Figure 6.27: Predicted regulatory network in mouse lung expressed genes, CpG depleted upstreams. TFBS co-occurrences are overrepresented for a window size of 100bp at a scanning threshold for a balanced threshold. Edge color represents PWM similarity as a gradient from blue (dissimilar) to red (similar).

Among the top predicted interactions in the CpG depleted (Figure 6.27) as well as for the CpG enriched (Figure 6.28) set we find combinations of many TFs known to be involved in lung development, like diverse *HNF*, *GATA*, *Forkhead*, and *HOX* TFs [68]. We mark the nodes with a known role in lung in green. Some have the same predicted interaction partners independent of the sequence set, like the *HNF3*:*Forkhead* combination, others have different top interaction partners, for example *HOXA7*:*CACD* in the CpG depleted case and *HOXA7*:*NFY* in the CpG enriched case.

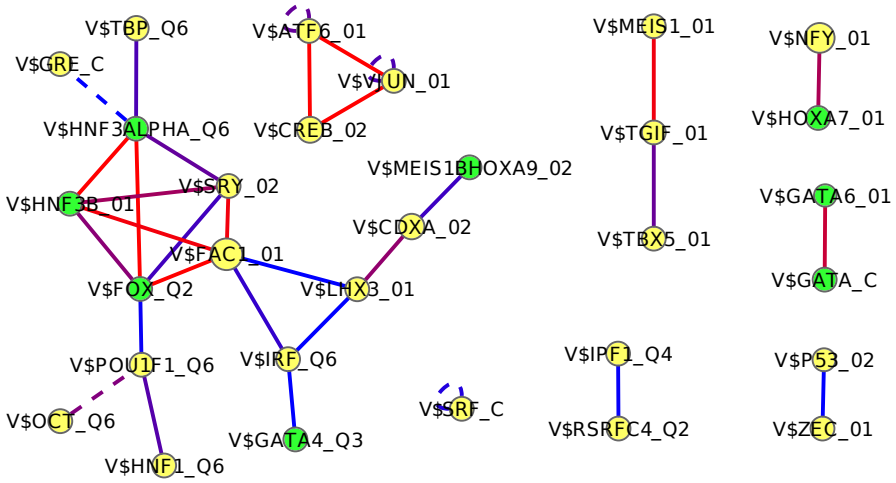


Figure 6.28: Predicted regulatory network in mouse lung expressed genes, CpG enriched upstreams. TFBS co-occurrences are overrepresented for a window size of 100bp at a scanning threshold for a balanced threshold. Edge color represents PWM similarity as a gradient from blue (dissimilar) to red (similar).

Testis

For the 274 testis specific genes in the CpG depleted category, after repeat masking the set contains 52,397bps and 9,065 predicted TFBSs. The CpG enriched set contains 226 genes, 43,126bps and 8,252 predicted TFBSs.

The known interactions we find are not necessarily specific for testis. As in liver, some of the factors that are known to play a role in testis show up in both sets, some are specific to only the CpG enriched or depleted set. For the factors that have overrepresented interactions in both sets, usually the top interaction partners are different, for example *SRY:HNF6* in CpG depleted vs. four interaction partners for *SRY* in the CpG enriched set, which do not contain *HNF6*.

Discussion

The analysis on the remaining tissue specific upstream sets results results in a similar picture like for lung and testis. Among the top 10 interactions for the respective CpG enriched and depleted sets we usually find TFs known to play a role in the respective tissue. While sometimes the TFs themselves differ between the two sets, sometimes just the predicted interaction partners differ.

We do not expect to find all TFs known for a tissue to be part of interactions with high scores due to the "intra-tissue" background. Since some of the TFs are overrepresented individually [280], they occur in high numbers also in our permutation based background model, leading to higher expected numbers for their TFBS combinations as well.

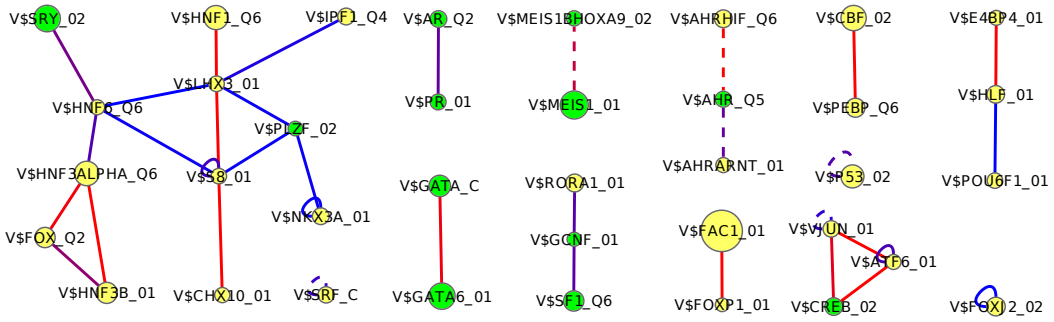


Figure 6.29: Predicted regulatory network in mouse testis expressed genes, CpG depleted upstreams. TFBS co-occurrences are overrepresented for a window size of 100bp at a scanning threshold for a balanced threshold. Edge color represents PWM similarity as a gradient from blue (dissimilar) to red (similar).

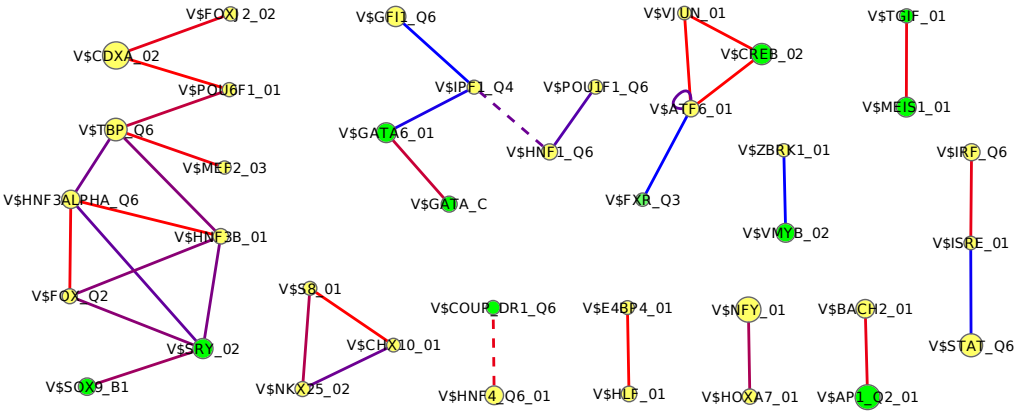


Figure 6.30: Predicted regulatory network in mouse testis expressed genes, CpG enriched upstreams. TFBS co-occurrences are overrepresented for a window size of 100bp at a scanning threshold for a balanced threshold. Edge color represents PWM similarity as a gradient from blue (dissimilar) to red (similar).

Roider et al. [280] find that the enrichment analysis for individual TFs in the CpG enriched sets does not lead to the identification of known tissue-specific TFs and explain this with the localization of tissue-specific TFBSs further away from the TSS in the case of CpG enriched upstream sequences. Our results suggest tissue-specificity of rather TF combinations than individual binding sites also for the CpG enriched regions. This will need further proof.

6.5 Comparing the Co-occurrence Score with a Theoretical Measure

6.5.1 Synopsis

The co-occurrence score described in Chapter 4 relies on expected pair counts calculated using a permutation procedure for the TFBSs. Here we compare the performance of this empirical background model with a theoretical model for expected co-occurrence events by Pape et al. [253] called *costat*, which is based on a symmetric i.i.d. background sequence model and described in Section 3.2.2.

6.5.2 Dataset and Application of costat

We create a sequence set analoguous to Section 6.2.2: We extract putative regulatory regions from yeast ranging from -150 to +50 bp relative to the annotated TSS and subsequently mask repeats. Afterwards we annotate TFBSs using the MacIsaac PWM set at a fixed false-positive rate of 0.05.

The *costat* method provides the probability of a co-occurrence event of binding sites belonging to two different PWMs. A co-occurrence event is a sequence window with at least one hit for each of the two involved PWMs. In the simple case *costat* calculates the probabilities for non-overlapping windows. Overlapping windows lead to approximation errors for which Pape et al. [253] present a quantification based on the Chen-Stein method. Since the resulting errors are high, we calculate values for non-overlapping windows and adjust the parameters for the calculation of the co-occurrence score in such a way that the results are comparable. We assess the results using heterotypic direct interactions as a positive set for a ROC curve as described before.

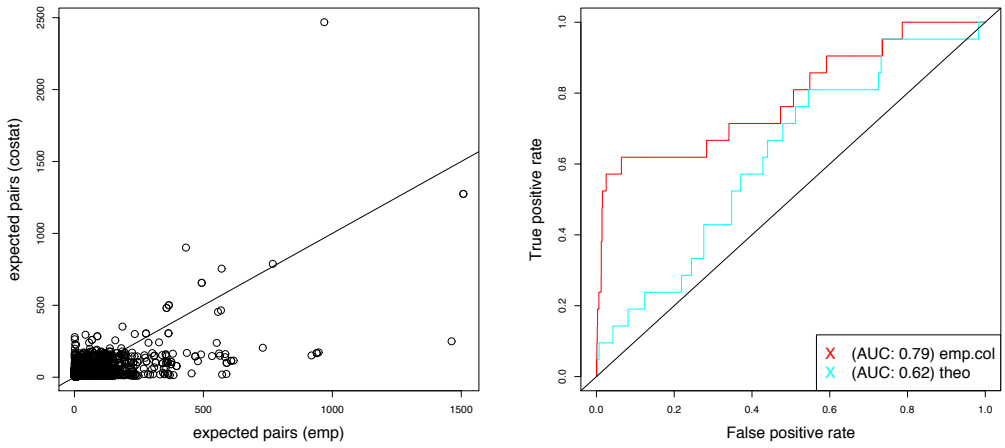
We calculate the *costat* co-occurrence probabilities for all PWM pairs for a fixed false positive cutoff of 0.05 and a window size of 200bps. We set the GC content for the calculation to the empirical GC content of 37% of our sequence set. We multiply the co-occurrence probability with the number of total non-overlapping 200bp windows in our sequence set to obtain the expected number of co-occurrence events. Analogous to our co-occurrence score we use a log-odds score to identify overrepresented TFBS combinations in the dataset:

$$S_{costat} = \log \frac{N_{pairs}}{p_{costat} \times N_{windows}} \tag{6.1}$$

We calculate our co-occurrence score for a window size of 200bps, so that the counting windows effectively do not overlap. Moreover we count overlapping TFBSs to retrieve comparable results for both methods.

6.5.3 Comparison of Performance

First we compare the expected number of TF pairs calculated using the permutation procedure described in Section 3.2.2 with the expected costat paircount. The two values show only moderate correlation (Pearson correlation coefficient of 0.61, Figure 6.31 a).



a) *b)*

Figure 6.31: Comparison of performance with COSTAT overrepresentation

The AUC for the ROC curves is 0.79 in the case of the empirical co-occurrence score, while it is 0.62 for the *costat* co-occurrence score. The empirically derived expected counts perform better in the detection of known interactions. The limitation of *costat* is probably related to the hypothesis of a symmetric i.i.d. background. This background distribution prevents the consideration of higher-order sequence features like CpG islands or other sequence properties specific for regulatory regions. Since these features are not ignored by the permutation model, it retrieves better results. However, Pape et al. [253] proposed an extension of the *costat* method to higher-order Markov models, which would apparently improve the results, but is not straight-forward.

Chapter 7

Predicting Regulatory Regions in Human

In Chapter 5 we presented a new method for the prediction of regulatory regions. It uses binding site graphs to represent combinations of transcription factors and their possible interactions, and weights the combinations based on their co-occurrence score in a set of training sequences. In Section 7.1 we apply the various regulatory potential scores to well characterized regulatory regions. We assess the discrimination performance of the regulatory potential scores on various data sets in Section 7.2. The tests include the comparison of regulatory sequences against a shuffled annotation set and against non-regulatory regions.

7.1 Calculation of Regulatory Potential for Known Regulatory Regions

7.1.1 Synopsis

Here we show the application of various regulatory potentials on known regulatory regions. We start with the regulatory regions of the gene *Pax6* and continue with the characterization of experimentally verified enhancers from the *VISTA* dataset.

7.1.2 Murine Pax 6

The transcription factor *Paired box gene 6* or short *Pax6* is highly conserved among vertebrates and also other bilaterian species. It plays an important role in development of a number of organs, among them brain, eye, nervous system [247, 144]. Therefore the transcriptional regulation of *Pax6* itself has been a topic of intense research. The factor exists in several splice forms and has three known promoters, and at least six experimentally confirmed enhancers with different specificity. Morgan [228] reviews the regulatory regions of *Pax6* in mouse.

In Figure 7.1 we show the murine *Pax6* locus and the transcripts from EnsEMBL 46. The regions marked in yellow correspond to the regulatory regions from the review of Morgan [228]. `P0`, `P1`, and $\texttt{P}_\alpha$ are the known alternative promoters. Note that EnsEMBL does not have annotation for transcripts starting at $\texttt{P}_\alpha$. The enhancers `E2` to `E5` correspond to the enhancers in Morgan [228], Figure 2. Enhancer `E2` influences transcription in lens, cornea, lacrimal gland, and conjunctiva. Enhancer `E3` regulates transcription in dorsal telencephalon, hindbrain, and spinal cord. `E4` activates transcription in non-terminally differentiated neurons, and `E5` in amacrine cells, iris, ciliary body, neural retina, and the

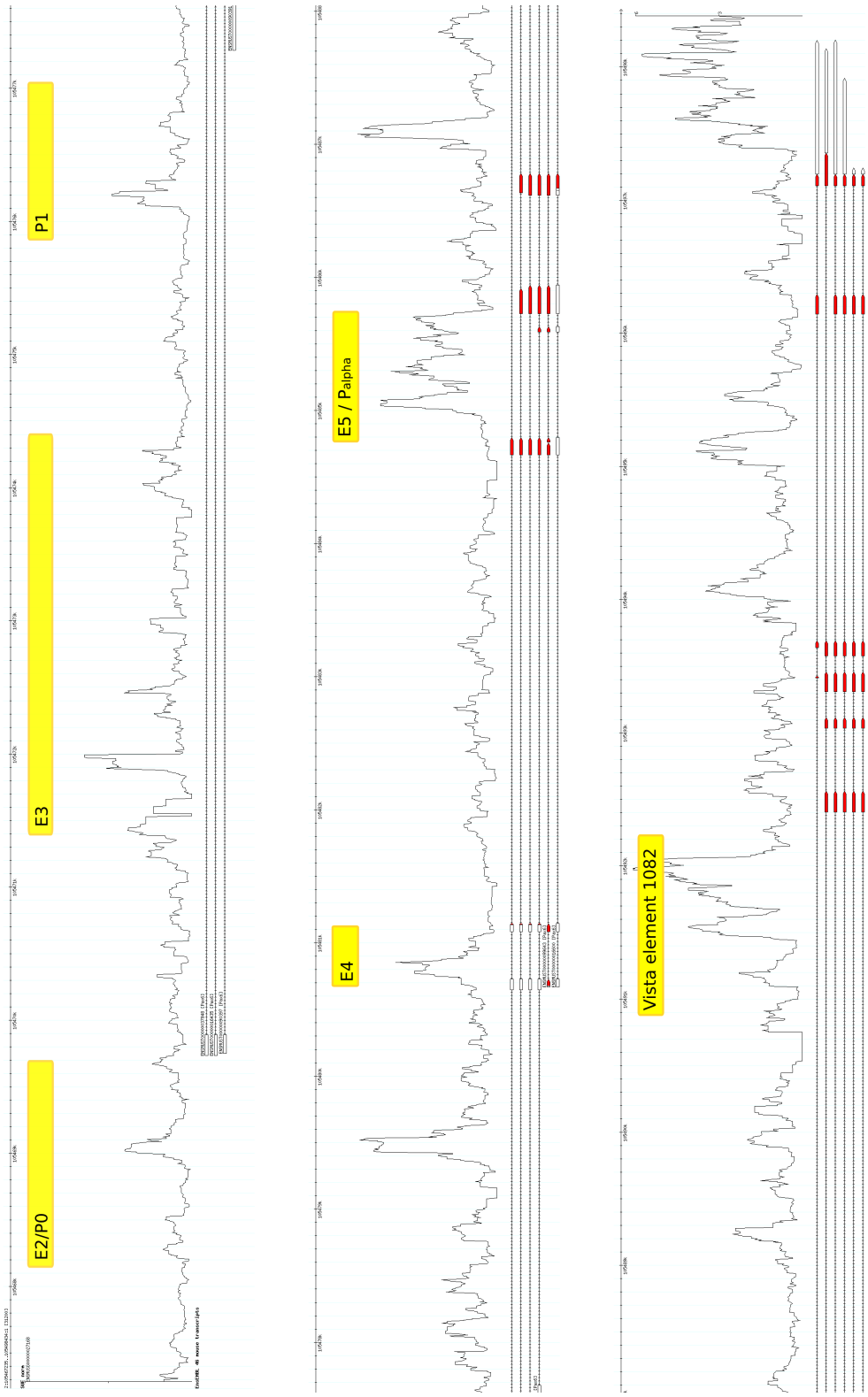


Figure 7.1: Genomic region of the murine *Pax6* gene on chromosome 2 with annotation of regulatory potential $\mathcal{R}_{\mathrm{SUE}}^{\mathrm{norm.e}}$. Regulatory regions in yellow boxes correspond to Morgan [228] and were curated to the updated structure of the locus for EnsEMBL 46.

pigmented layer of retina. The peak upstream of exon 7 marked with "VISTA element 1082" corresponds to an enhancer, shown to have regulatory activity in hindbrain and neural tube. We discuss the homologous human region in the next section.
We annotate the region with the $\mathcal{R}_{\mathrm{SUE}}^{\mathrm{norm.e}}$ regulatory potential. The score peaks in the known regulatory regions. Another peak is located upstream of exon 2 at the start of an EnsEMBL transcript not annotated in Morgan [228]. We suggest that this peak is related to the promoter region of the transcript starting 5' of known region `E4`. Apart from that, the 3' region of the locus obtains a high regulatory potential, which agrees with the findings of Cawley et al. [55], that the 3' regions of genes contain a high density of real TFBSs. Another striking peak downstream of exon 6 is visible, for which a regulatory function is not known yet. The exons of *Pax6* coincide with regions of low regulatory potential.
For *Pax6* the top regions of interest as predicted with the $\mathcal{R}_{\mathrm{SUE}}^{\mathrm{norm.e}}$ mostly coincide with known regulatory elements.

7.1.3 Human Enhancers

In this section we apply the calculation of the regulatory potential scores on examples taken from the VISTA Enhancer Database [345]. Visel et al. [345] detected candidate enhancers based on comparative genomics by conservation from human to chicken, frog, puffer fish, or zebra fish. Subsequently they verify *in vivo* activity by reporter gene expression. We restrict ourselves to the analysis of candidate enhancers for which Visel et al. [345] confirmed regulatory activity.

We show the regulatory potentials $\mathcal{R}_{\mathrm{MWM}}$, based on maximum weighted matching in the binding site graph, and the normalized $\mathcal{R}_{\mathrm{SAE}}^{\mathrm{norm.e}}$, based on the summation of all edges, both calculated based on an annotation with a balanced TFBS scanning cutoff, and a respective co-occurrence matrix for the non-redundant set of vertebrate PWMs described in Section 6.3.2. We calculate the score using a sliding window with a step size of 20bp. We extend the enhancer regions by 2,000bps up- and downstream, and additionally display transcript information from EnsEMBL v. 46 [153]. Additionally we include a track showing CAGE tags from the FANTOM3 project [54], providing evidence for transcriptional activity.

VISTA element 1082 in Figure 7.2 is located in the intragenic region of the transcription factor *PAX6*, 5' of an exon on chromosome 11 (31,773,028-31,774,997). Its size is roughly 2,000bps. The element influences expression in hindbrain and neural tube.
The enhancer region shows various peaks in the $\mathcal{R}_{\mathrm{MWM}}$ regulatory potential. The surrounding region contains peaks in the score as well, but none are higher than the values achieved inside the enhancer region. The peaks 3' (left) of the enhancer region coincide with introns of the *PAX6* gene, while the score dips in exonic regions as expected. The normalized $\mathcal{R}_{\mathrm{SAE}}^{\mathrm{norm.e}}$ potential achieves high scores in the same positions as the first potential, although here the distinction between enhancer region and surrounding region is not as strong as in the first case. Due to the summation of all edge weights in the binding site graph, the $\mathcal{R}_{\mathrm{SAE}}^{\mathrm{norm.e}}$ potential also penalizes' TFBS combinations, which are uncommon in the training set for the co-occurrence scores, later used as edge weights. This happens for example in the area marked in blue.

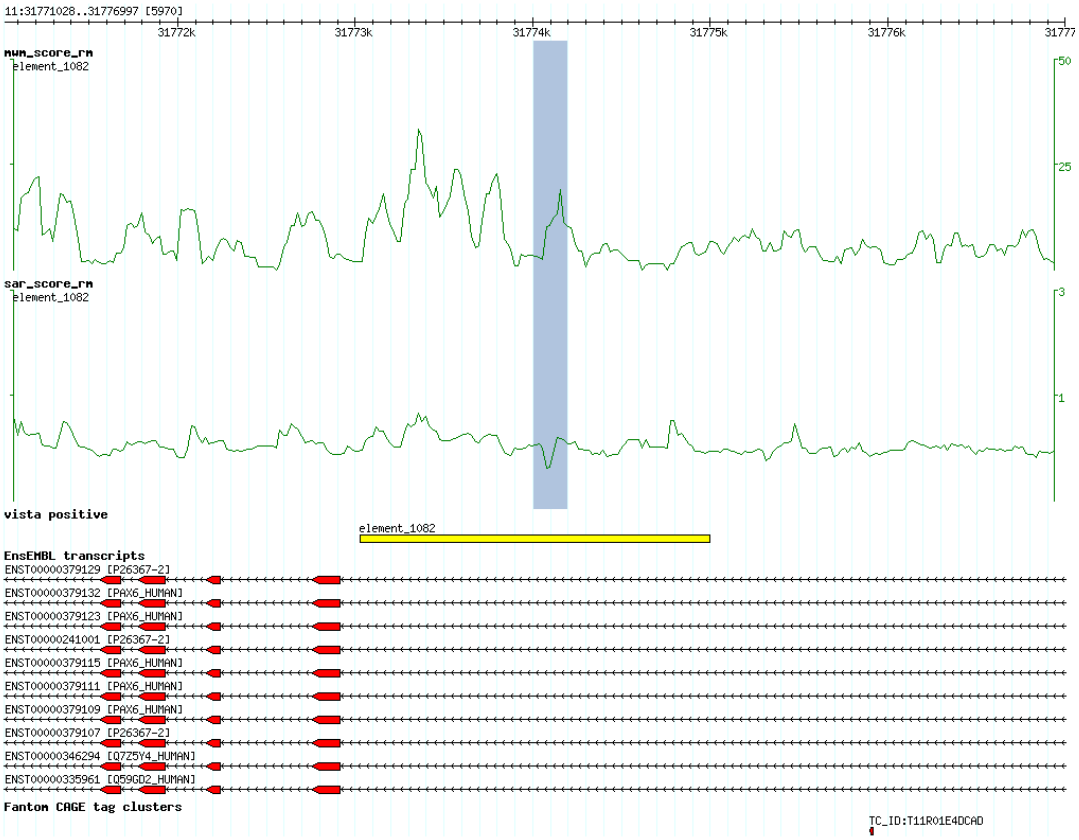


Figure 7.2: VISTA element 1082 is a roughly 2,000bp long region, located upstream of a *PAX6* exon. The $\mathcal{R}_{\mathrm{MWM}}$ regulatory potential (upper panel) shows peaks with the highest value in the complete region inside the enhancer element. The $\mathcal{R}_{\mathrm{SAE}}^{\mathrm{norm.e}}$ regulatory potential (lower panel) mostly shows peaks at the same positions as the $\mathcal{R}_{\mathrm{MWM}}$ potential. The distinction between enhancer and non-enhancer region is not as strong as for the $\mathcal{R}_{\mathrm{MWM}}$ potential. In some cases peaks in the upper potential correspond to negative dips in the lower one. This cooresponds to a region with a high number of TFBS combinations, which are underrepresented in the training set for the co-occurrence score matrices.

VISTA element 627 in Figure 7.3 is located in a roughly 20,000bp long intergenic region between the genes *NEUROD2* and *PPP1R1B* on chromosome 17 (35,028,011-35,028,514) and has a size of around 500bp. The element influences gene expression in midbrain. Here we see a single high peak in the $\mathcal{R}_{\text{MWM}}$ regulatory potential as well as the $\mathcal{R}_{\text{SEA}}^{\text{norm.e}}$ potential, which surmounts all values in the surrounding regions.

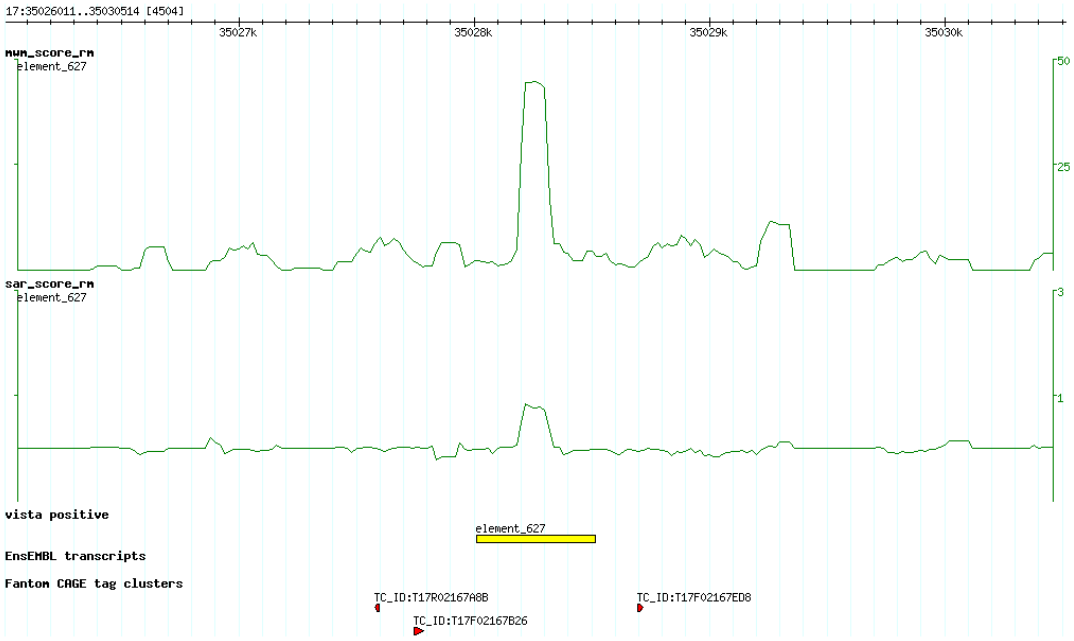
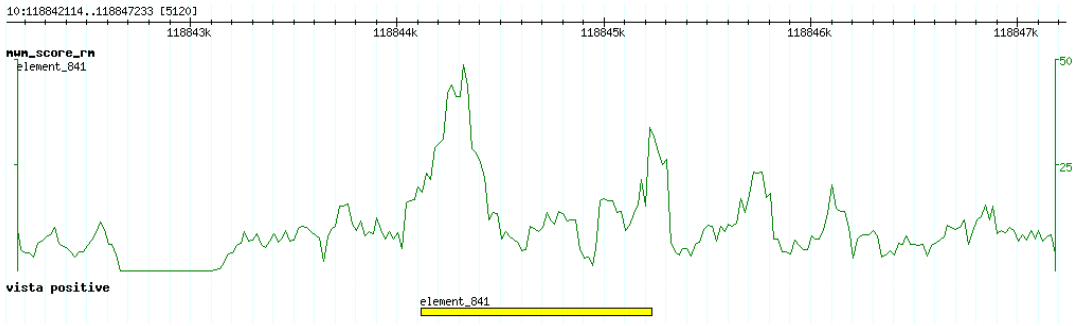


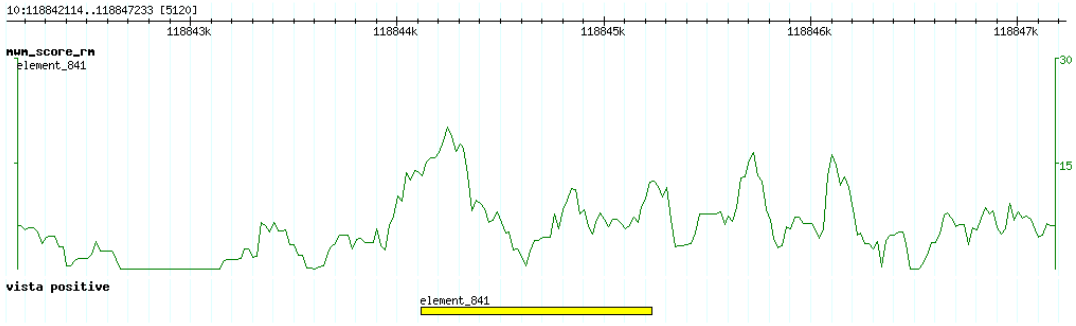
Figure 7.3: VISTA element 627 has a size of approximately 500bp and is located in the intergenic region between the two genes *NEUROD2* and *PPP1R1B*.

Influence of PWM Clustering Recall the two sets of PWMs from Section 6.3.2, one of which consisting of clustered PWMs. With respect to known interactions, the clustered PWM set performed slightly worse than the non-redundant, but unclustered vertebrate PWM set. The clustered PWM set has the disadvantage, that overrepresented motif combinations are harder to assign to a single TF than in case of the unclustered variant, making a biological interpretation more difficult. When we use the co-occurrence scores in the context of binding site graphs, this does not matter though, since we do not interpret the individual combination of binding sites, but combinations of TFBSs, which are typical for regulatory regions. The motivation to use the clustered PWM set for the binding site graphs lies in anticipated problems due to redundancy in the original PWM set.

Figure 7.4 shows the vista enhancer element 841 with regulatory activity in midbrain and forebrain, annotated with the $\mathcal{R}_{\text{MWM}}$ regulatory potential. Figure 7.4 a) shows the values calculated for the non-redundant vertebrate PWM set, and Figure 7.4 b) the values for the clustered PWM set. The plot shows a common case in the comparison of values for the different PWM sets: the peaks are in the same position, but peak heights differ relative to each other. The absolute scores differ from each other as well, since the non-redundant set contains 142 PWMs while the clustered set only contains 45 PWMs. This leads to less



a) non-redundant vertebrate PWMs



b) clustered vertebrate PWMs

Figure 7.4: VISTA element 841 annotated with $\mathcal{R}_{\mathrm{MWM}}$ with two different PWM sets and the respective co-occurrence scores. a) annotation using the non-redundant vertebrate PWM set. b) annotation using the clustered vertebrate PWM set

predicted TFBSs per window which influences the $\mathcal{R}_{\mathrm{MWM}}$ value as well as the different co-occurrence scores calculated for the two sets.

Complete VISTA Positive Set The evaluation of the complete set of 495 VISTA enhancers with regulatory function results in in roughly 200 regions for which the highest peak of the $\mathcal{R}_{\mathrm{MWM}}$ regulatory potential is inside the known enhancer region. For another 70 regions we find a peak inside the enhancer and a second distinct peak somewhere in the surrounding region. About 180 have the highest peak clearly outside of the known enhancer. Roughly 40 of the regions have a low regulatory potential in general with no striking peaks.
Among the regions with the peak outside the known enhancer are cases, which have CAGE clusters annotated at the higher peaks outside the enhancer, suggesting that the high $\mathcal{R}_{\mathrm{MWM}}$ values are not necessarily false positives. Peaks with annotation neither in the VISTA set nor in the FANTOM data might correspond to other regulatory regions. In a few cases they correspond to core promoter regions of annotated EnsEMBL genes.
The peaks we observe are often high only relative to the surrounding region. The height of the highest peak in one example region might correspond only medium height peaks in other regions. Often peaks in the positive set span only a fraction of the annotated enhancer, implying that transcription factor binding inside the enhancers does not stretch over the complete region. This is not surprising, since almost 70% of the sequences in the positive set are longer than 1,000bps.

7.2 Large Scale Assessment of Regulatory Potentials

7.2.1 Synopsis

In this section we assess the performance of the regulatory potentials on larger sequence sets assembled from the human genome. We calculate regulatory potentials on two promoter and one enhancer set as positive sets, and a random intergenic, as well as a shuffled TFBS set as negative sets. Subsequently we perform ROC analyses as described in Section 3.6.1, and use the AUC as a measure for the quality of separation between positive and negative sets.

7.2.2 Sequence Sets

Positive Sets with Regulatory Function

Promoter Sets We define two different positive sets consisting of putative promoters. The FANTOM3 project [54] applied CAGE technology for the detection of transcripts in human and in mouse. The project resulted in millions of genome-wide CAGE tags, plenty of which occur in four different cluster classes, which Carninci et al. [54] related to different types of TSSs. The *single peak* (SP) class shows a sharp peak of CAGE tags, hinting towards a well-defined TSS. The *broad* shape (BR) class hints to a region with multiple weak TSSs. Apart from that, the *bimodal/multimodal* (MU) shape class implies several well defined TSS within a short distance, and the *broad with dominant shape* (PB) class combines a sharp

TSS with several weakly defined.

We take the SP as well as the BR class for the human data set as input. The SP and BR classes contains 1,549 and 1,513 sequences, respectively. We extract the region between -350 and +50 relative to the 5' end of each TSS region, assuming that we capture the promoter region belonging to the TSS. We sample 5,000 regions of size 200bps from each set. Furthermore we mask repeats in both sets.

Enhancer Set The enhancer set is based on the VISTA enhancer set [345] already used in Section 7.1.3. From this sequence set, we sample 5,000 regions of size 200bps. We mask repeats on the enhancer set.

Negative Sets without Regulatory Function

Artificial Negative Set We create an artificial negative set to test the influence of TFBS combinations alone on the performance of the different regulatory potentials. To that aim, we take the SP promoter set described above and shuffle the TFBS labels. As a consequence, we retain the total and individual TFBS count, the binding site positions, and density is the same as in the original annotation, but the information in the combination of TFBSs is lost.

Intergenic Set As a real negative set we aim to select non-regulatory sequences. Nevertheless, it is hard to create a true negative set with regard to regulatory function. We can not rule out, that some randomly selected piece of sequence does harbor regulatory elements. Our approach to compile a negative set starts with the removal of all known genes annotated in EnsEMBL v. 46 the 5,000bp up- and 5,000bp downstream of each gene. Furthermore we remove the surrounding 1,000bps up- and downstream of annotated CAGE tags from the FANTOM3 project from the set. Although we can not guarantee that the remaining intergenic sequence does not contain enhancer regions, we have a set with relatively low probability of regulatory activity.
We sample 5,000 regions of size 200bps and use it as a negative set after masking repeats.

7.2.3 Performance Assessment

Promoters vs. Artificial Negative Set In Figure 7.5 we show the ROC curve calculated for the regulatory potentials and the *CAGE SP* promoter set and the shuffled negative set.

The regulatory potentials which separate the two data sets best are $\mathcal{R}_{\mathrm{SAE}}$ and $\mathcal{R}_{\mathrm{SAE}}^{\mathrm{norm.e}}$. Both regulatory potentials penalize combinations of TFBSs which are uncommon in the training set, because they take into account negative edge weights. Other measures of the regulatory potential, which only take into account edges with positive weight, do not perform particularly well. The shuffling procedure leads to a high number of TFBS combinations, which are untypical for regulatory regions. These combinations obtain negative co-occurrence scores, which in turn lead to negative values of the $\mathcal{R}_{\mathrm{SAE}}$ and its normalized version $\mathcal{R}_{\mathrm{SAE}}^{\mathrm{norm.e}}$.

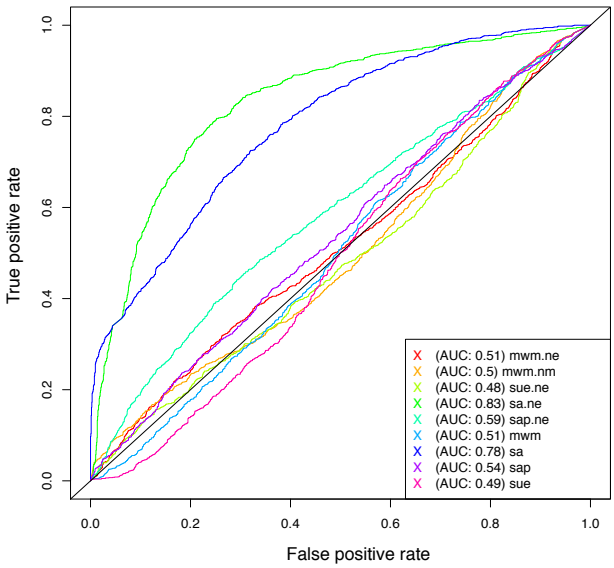


Figure 7.5: ROC curve for comparison of *SP* promoter set vs. shuffled promoter set.

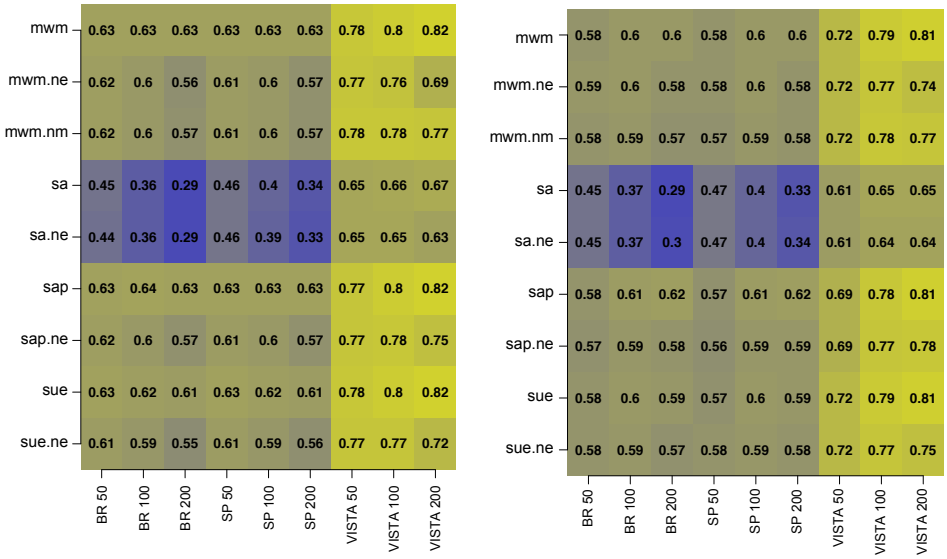
Regulatory Regions vs. Random Intergenic To survey the performance of a regulatory potential, we take each of the positive sequence sets described above together with the random intergenic negative set and calculate all the regulatory potentials for each sequence. Subsequently we generate ROC curves as described in Section 3.6.1. A high AUC corresponds to pronounced ability of the regulatory potential to separate the positive from the negative set. We perform the calculations using two PWM sets described in Section 6.3.2, one non-redundant vertebrate set and a clustered version.

The matrix in Figure 7.6 a) shows the AUCs achieved with regulatory potentials calculated with the non-redundant PWM set on the complete positive and negative sets. The row annotation of the plot correspond to the different regulatory potentials, and the column annotation to the various comparisons of positive versus a negative set. The comparisons differ in the positive sets and the window size used. The three comparisons for window sizes 50, 100, and 200bps for each sequence set are located next to each other. The first block of three columns belongs to the *CAGE BR* set, the second block to the *CAGE SP* set and the third block to the *VISTA* set.
The results on the two promoter sets largely behave similar. An increase of the window size leads to a worse performance in the case of the promoter sets. Except for the scores $\mathcal{R}_{\mathrm{SAE}}$ and $\mathcal{R}_{\mathrm{SAE}}^{\mathrm{norm.e}}$, based on the summation of all edges in the binding site graph, the different regulatory potentials perform comparable on the different sequence sets.

Figure 7.6 b) contains the AUCs calculated using the clustered PWM set. Row and column order is identical to the left part of the figure. The performance for the regulatory potentials is very similar to the results of the non redundant PWM set. Differences in scores are

smaller 0.03 in almost all cases.

The separation of the promoter sets from the intergenic sets performs moderately at AUCs around 0.6. For the separation of enhancer regions and the negative set the regulatory potentials perform better at AUCs between 0.72 and 0.82 for all potentials except $\mathcal{R}_{\mathrm{SAE}}$ and $\mathcal{R}_{\mathrm{SAE}}^{\mathrm{norm.e}}$.



a) non − redundant vertebrate PWMs b) clustered PWMs

Figure 7.6: AUC from ROC curves for all regulatory potentials. Positive and negative sets contain sequences from the *complete range of GC content*. The left image contains AUCs for annotation and co-occurrence scores for the *non-redundant vertebrate PWM set*. The right image contains AUCs for annotation and co-occurrence scores for the *clustered vertebrate PWM set*. Positive sets are derived from *CAGE SP* and *CAGE BR* promoter sets and *VISTA* enhancer set. Negative set: random intergenic. Calculation of regulatory potential is carried out for window sizes of 50, 100, and 200bps for repeat masked sequence sets. Suffixes of regulatory potential names correspond to the respective normalization: ne represents normalization with number of edges in the graph, nm normalization with maximum possible number of edges in a matching.

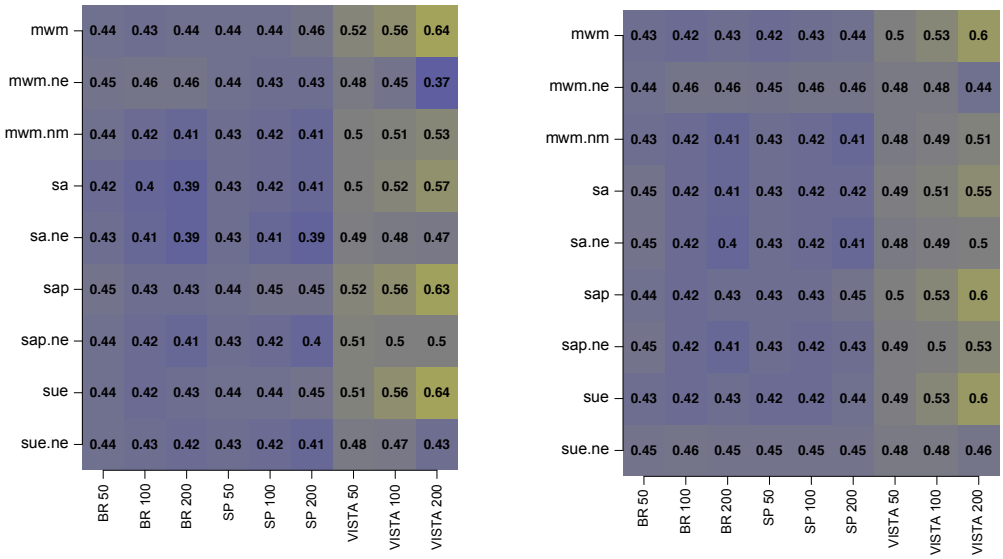
Division into High and Low GC Content Sequence Sets Regulatory regions often have a high GC content (see also Section 3.3.1). This fact is often used to discern regulatory from non-regulatory sequences. In our test set the GC content difference between the promoter sets and the intergenic sequence set is large enough to discriminate between positive and negative set at AUCs between 0.8 and 0.85. The GC content difference between the enhancer set and the intergenic set is not as large, but the AUC is still roughly 0.65. To test whether our regulatory potentials are able to separate sequences from the positive set from sequences from the negative set even without the strong influence of the GC content, we divide the sequence sets into two classes, one with sequences with a GC content ≤ 0.5 (*low GC*), the

other with sequences with a GC content > 0.5 (*high GC*). We repeat the calculation of the AUCs for all of the comparisons regarding the sequence classes. The top row of Figure 7.7 contains the AUC matrices for the low GC set, again with the non-redundant vertebrate PWM set in the left matrix, and the clustered PWM set in the right matrix, the bottom row contains the respective AUC matrices for the high GC set.
On the low GC set, the promoter sequences are not distinguishable from the intergenic sequences, irrespective of the regulatory potential, the window size, or used PWM set. For the enhancer set, separation is possible, albeit with a number of wrong predictions at AUCs between 0.57 and 0.64 for the window size of 200bps. The performance is better for the non-normalized potentials, while the normalized potentials obtain AUCs around 0.5.
On the high GC set on the other hand the regulatory potentials can separate positive and negative sets. The best AUCs obtained are around 0.7 for the discrimination between promoters and intergenic sequences and 0.73 for the discrimination between enhancers and intergenic sequences. Again the normalized scores perform worse than the non-normalized ones. For the promoters, the AUCs obtained for high GC sequences are bigger than in the comparison of the complete sequence set. In case of the enhancers, the obtained AUCs are slightly smaller for the high GC sequence set than for the complete sequence set.

7.3 Discussion

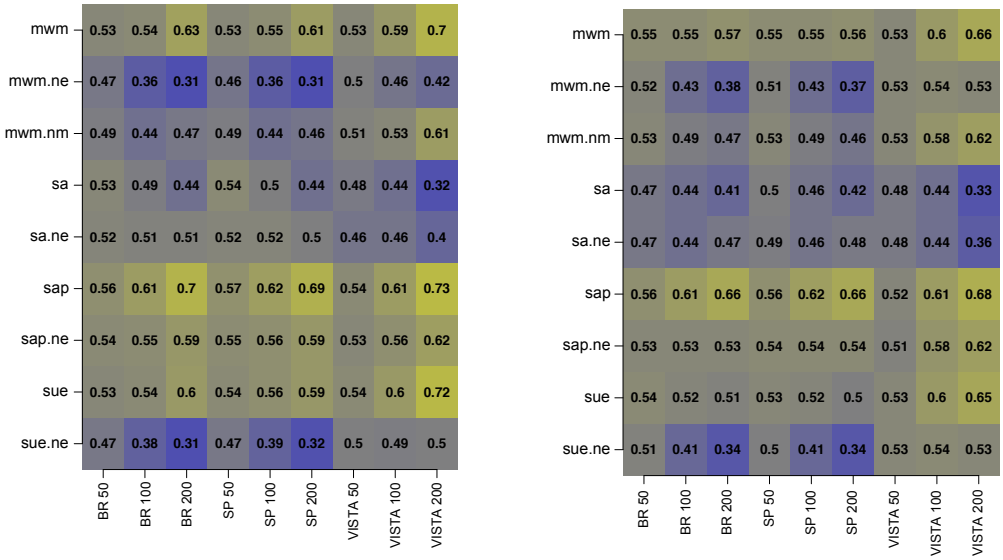
Known Regulatory Regions In Section 7.1 we showed examples sequences containing known regulatory regions and annotated them with various regulatory potentials. We could show that in known enhancers and promoter regions the regulatory potentials yield high values. On the VISTA enhancer set we identify the highest peak of the enhancer region plus surrounding areas inside the known enhancer for about 41% of the cases, and another 14% with a distinct peak inside the known region inside and a second distinct peak outside. The remaining enhancer sequences either contain higher peaks outside the known enhancer or show an unclear pattern. High peaks outside of the experimentally verified regions do not necessarily imply false predictions though, since other regions than the annotated once might exert yet uncharacterized regulatory function.
An important observation is that usually there are distinct peaks of the regulatory potential in regulatory regions, which do not cover the whole known regulatory region. This does *not* imply that only parts of the regulatory elements are needed for the respective regulatory function though. Some parts of the experimentally verified regions harbors a higher density of real TFBSs, while other parts might be important because of distance, nucleosome, or histone related circumstances.

Large Scale Assessment In Section 7.2 we assessed the regulatory potentials on different positive and negative sequence sets. First we test the performance on a promoter set and a shuffled TFBS set, so that the TFBS combinations of the regulatory region are destroyed. We find that those regulatory potentials perform best, which incorporate also negative edges from the binding site graphs in the resulting score. In the later comparisons of regulatory regions against intergenic regions exactly potentials incorporating negative edge weights perform worst. We conclude that the TFBS combinations responsible for the low scores in the artificial set are also not that common in *real* non-regulatory sequences. For example, we do not expect combinations of TFBSs with a high difference in the GC content by



a) non − redundant vertebrate PWMs, low GC

b) clustered PWMs, low GC



c) non − redundant vertebrate PWMs, high GC

d) clustered PWMs, high GC

Figure 7.7: AUC from ROC curves for all regulatory potentials. Positive and negative sets contain sequences of with a *GC content* < 0.5 in the top row, sets with a *GC content* ≥ 0.5 in the bottom row. The left column image contains AUCs for annotation and co-occurrence scores for the *non-redundant vertebrate PWM set*, and the right column image contains AUCs for annotation and co-occurrence scores for the *clustered vertebrate PWM set*. Positive sets are derived from *CAGE SP* and *CAGE BR* promoter sets and *VISTA* enhancer set. Negative set: random intergenic. Calculation of regulatory potential is carried out for window sizes of 50, 100, and 200bps on repeat masked sequence sets. Suffices of regulatory potential names correspond to the respective normalization: ne represents normalization with number of edges in the graph, nm normalization with maximum possible number of edges in a matching.

chance. These combinations can result from shuffling, but are unlikely to show up without a functional reason in the *intergenic* negative set.

For the large scale analyses we used a non-redundant vertebrate PWM set as well as a clustered PWM set. We obtain AUCs in comparable ranges in all of the cases, sometimes with slightly lower values for the clustered set. This implies, that the impact of problems due to redundancy of the PWM set is small.

Normalization of the regulatory potentials turns out to be unsuccessful. First indications for that show up already in the individual enhancer examples, and this impression is reinforced in the large scale analyses. The normalization lessens the influence of the raw number of TFBSs in a piece of sequence, which as such contributes to a strong signal for regulatory regions.

The small peak width compared to the total size of regulatory regions seen in individual examples in Section 7.1 suggests, that due to the random selection of a sequence subset, our positive sets in the large scale analysis might not contain the part of the sequence with the highest peaks of the respective regions. Overall this would lead to a lower AUC then theoretically could be achieved. Again, we can also not rule out that the negative set contains real regulatory elements, which would dilute the results.

The regulatory potentials are able to discriminate regulatory regions from intergenic regions, although false predictions exist.
We calculated the regulatory potentials using co-occurrence scores determined on putative *promoter* regions as described in Section 6.3. In this context it is surprising, that the separation performance on promoters and intergenic regions is worse than on enhancers and intergenic regions. A possible reason might be a cleaner data set in case of the *VISTA* enhancers due to experimental verification, whereas the promoters regions were defined relative to TSSs, which might contain some false positive sequences.

While in the complete enhancer set the regulatory potentials achieve slightly better AUCs then the GC content as a classifier, on promoter sets, the regulatory potentials perform worse than using GC. This is in general agreement with the performance tests of promoter prediction tools by Abeel et al. [2], although we can not compare the results directly. Here, the majority of tools achieve considerably worse results than using the GC content alone, while only a few lie level or outperform the GC content. This indicates that the problem of prediction of regulatory regions stays complicated.

To test how strong the GC content influences our regulatory potentials, we split the sequence sets and test, if the regulatory potentials are able to separate positive and negative sets *within* low GC and high GC content sequences. Separation is not possible for promoters versus the intergenic sequences. The separation of enhancers from intergenic regions obtains an AUC of 0.64 with the non-normalized scores at larger window sizes. In case of the high GC sequences, discrimination between positive and negative sets obtains better results, for both the promoters and the enhancers against the intergenic set. When using larger window sizes, the AUCs reaches 0.73.

Since non-regulatory regions with a high GC content are relatively rare, the higher performance in the high GC set does not immediately suggest a general applicability of the regulatory potentials. Nevertheless, in case of sequence regions which might be viewed as candidates for regulatory regions because of directly using the GC content or some other measure dependent on it, our regulatory potentials can help to lower the number of false-positive predictions.

Some of the regulatory potential achieve very low AUCs $<< 0.5$ in some of the comparisons. Especially $\mathcal{R}_{\mathrm{SAE}}$ and $\mathcal{R}_{\mathrm{SAE}}^{\mathrm{norm.e}}$ show this behaviour at the window size of 200bps. This unexpected result implies the existence of a signal, which makes distinction between the two sequence sets possible. However, this signal is based on the counter-intuitive finding, that the negative set sequences contain binding site combinations, that better resemble the combinations described by the co-occurrence scores, than the positive set. This affects especially the potentials which penalize uncommon TFBS combinations. We suspect that at window sizes in the matching procedure that are larger than the window size for the calculation of the co-occurrence score, the number of uncommon TFBS combinations increases disproportionately.

Summary In brief, the usage of one of the non-normalized regulatory potentials $\mathcal{R}_{\mathrm{MWM}}$, $\mathcal{R}_{\mathrm{SUE}}$, and $\mathcal{R}_{\mathrm{SAP}}$ helps in the detailed investigation of individual genomic regions. Apart from the identification of putative promoter or enhancer regions, the regulatory potentials also point out parts of the sequence with high numbers of putative binding sites.
The discrimination performance of the regulatory potantials lags behind the GC content as a classifier on general data sets. Within high GC sequence sets, the regulatory potentials perform reasonably well. Hence we suggest the application of the regulatory potentials as a means to lower the rate of false positive predictions obtained by other methods, which are rely mostly on the GC content.

Part IV

Summary and Conclusions

Chapter 8

Summary and Conclusions

Transcription factor interactions are crucial for understanding the mode of action of individual factors as well as the regulatory network in general. The presence, absence, or exchange of interaction partners can lead to a change of a factor's role between activation, repression, specificity, modulation of strength, or even non-functionality. Although large scale experimental data are available, they are far from covering the complete interaction space, leading to the need for computational methods for the prediction of TF interactions. Methods for the prediction of protein-protein interactions in general perform only moderately for TFs, due to various TF specific properties like low expression values, or the evolutionary origin of TFs from a small number of factor classes. Furthermore, apart from direct physical interactions, functional interactions are of a specific interest for TFs. Apart from methods for the prediction of TF interactions that require experimental data, a group of approaches exist that rely on combinations of motifs in regulatory regions.
Another complicated problem in the field of gene regulation is the identification of regulatory regions in the genome. While candidate regions for putative promoter sequences are readily available upstream of transcriptional start sites, these regions neither define promoters exhaustively nor do they contain regulatory regions only. Changing the focus to enhancer regions, the problem gets even more complicated. For both types of regions several approaches exist. The majority implicitly or explicitly exploit low-level sequence properties like GC content or CpG rate for the prediction of regulatory regions, leading to great difficulties in the identification of low GC / low CpG regulatory regions.

Method Summary In this thesis we addressed the defined problems in two related approaches. Both approaches are based on the assumption that in regulatory regions, pairs of binding sites of functionally interacting TFs are overrepresented with regard to their expected numbers.
In the first method we adopt this hypothesis to design a method for the calculation of a *co-occurrence score* to predict interacting TFs. Initially, we conceived a counting method for TFBS co-occurrences, that uses a sliding window and takes into account TFBS specific properties, such as binding site clustering. Then we devised a log-odds score of observed pair counts over expected pair counts, with the expected pair counts estimated from a permutation procedure. The co-occurrence score enabled us to identify TFBS pairs overrepresented in regulatory regions.
The second approach exploits the over-representation of TFBS combinations to build binding site graphs of sequences of unknown function. The binding site graphs contain the predicted TFBSs of a sequence as vertices. All vertices are connected to all others with

edges, to which we assign the respective co-occurrence score as weight. Based on these binding site graphs, we propose several different *regulatory potentials*, which measure how well the present TFBS combinations resemble the situation in known regulatory regions.

Prediction of Transcription Factor Interactions In Section 6.1 we evaluated our co-occurrence score on artificial data and showed that it detects enriched TFBS combinations even at very low frequencies. While the sensitivity is lower for TFBS combinations of highly abundant TFBSs, it performs well for specific TFBS combinations.
Subsequently, we apply our method to putative regulatory regions in yeast in Section 6.2. First we ascertained the best combinations of parameters of the counting procedure for a set of known TF interactions, leading to an AUC of 0.8 for the best parameter combination. Based on the co-occurrence scores calculated at the optimal parameters, we identified putative interacting TFs. For roughly half of the predicted interactions we find experimental evidence in the literature. There are PWM pairs similar to each other among the binding motifs of putative TF interactions, indicating competition for binding sites or overlapping binding sites for TFs in a complex. Moreover we find some known interactions, whose binding site combinations are underrepresented in the yeast genome. One possible explanation for this observation is a high specificity and rare usage of the respective factor combination.
In Section 6.3 we employed the co-occurrence score method on putative human promoters. Here we included upstream size and sequence conservation into the parameter optimization. While a variation of upstream size and conservation does not result in big differences in the AUC, we can identify window sizes around 100bps and the relatively permissive balanced scanning cutoff as the best parameter combination, which leads to an AUC of 0.76. The survey of the combinations with the highest co-occurrence scores yielded slightly more than a third of the predicted interactions known from the literature. Again, we detected pairs of PWMs which are similar as well as dissimilar to each other.
We also separately predicted interacting TFs for genes expressed in human embryonic kidney cells, the upstream regions of which have a low and a high normalized CpG content. Again we found many known interactions in the top interactions, some of which have key roles in kidney or embryonic development. Additionally we performed a similar analysis in tissue-specific sequences in mouse, leading to known interactions among the top pairs. We found factors with known roles in the respective tissues in the CpG depleted sets as well as in the CpG enriched sets, suggesting tissue-specificity of TF combinations instead of individual TFs.
We discovered a high number of combinations of homeodomain TFs in the genome-wide data and in the specific gene sets, which was initially surprising. Ruling out possible artefacts due to PWM similarity within the group of homeodomain TFs, we concluded that the finding agrees with the overall importance of homeodomain TFs in development and adulthood.
In all settings, taking into account overlapping TFBSs lead to better results. Due to a number of known examples of overlapping binding sites on the one hand and TF competition on the other hand, this is not surprising.
Finally, in Section 6.5 we compared the performance of our co-occurrence score with the *costat* method and showed, that our permutation-based background model is more suitable than the theoretical one.

Prediction of Regulatory Regions We applied the regulatory potentials in Chapter 7. In Section 7.1 we first showed individual examples of regulatory regions in human and mouse, in which peaks in the regulatory potentials coincide with known promoters and enhancers. Usually the peak width of the regulatory potential is only a fraction of the size of the known regulatory region, indicating that parts of the regulatory regions contain a high density of regulatory elements.
In Section 7.2 we created a set of positive and negative test sets for the performance assessment of the regulatory potentials in the separation of sequence sets. As positive sets, we defined two promoter sets and an enhancer set, and as negative sets an artificial one and an intergenic set. We calculated the regulatory potentials for the sequences in all of the sets. The subsequent analysis with ROC curves resulted only in average separation performance for promoters vs. the intergenic set and with good performance for the enhancers. In the case of the promoters, the GC content performs better than the regulatory potentials, while for the enhancers the regulatory potentials is slightly better than GC. This is in agreement with the results of genome-wide promoter prediction with other tools.
To assess whether the regulatory potentials are able to discriminate between positive and negative sets without the large influence of the GC content, we divide the sequence sets into high and low GC sets and repeat the analysis. The regulatory potentials are not able to separate positive and negative sets in the low GC case, while they are able to discriminate the sets in the high GC sequences.
The regulatory potentials are suitable to identify regions of interest in the in-depth analysis of potential regulatory regions. Besides this, although the problem of prediction of regulatory regions remains complicated, our regulatory potentials can be used as an additional filter to remove false positive predictions from other methods that depend more strongly on the GC content.

Outlook We showed the prediction of TF interactions based on the co-occurrence score. While we found many known interactions among the highest scoring TF pairs in yeast and in vertebrates, a number of TF pairs achieved high co-occurrence scores, which are not known to interact. For these, experimental verification would be needed.
The prediction of tissue-specific interactions suffered from the calculation of expected pair counts on the same sequence set, potentially leading to not detecting interactions of TFs, which are abundant in the tissue in question. One might circumvent this problem by calculating a background based on genome-wide occurrences of TFBSs, which in turn requires proper normalization.
The resulting networks of high scoring interactions often contain more than one predicted interaction for a TF, which does *not* imply that the respective interactions are based on co-occurrences in the same genes. To deduce information about TF interactions involving multiple factors at once, one could extend the co-occurrence score to frequent item sets instead of TF pairs, including suitable background models.

Overall, the regulatory potentials show a performance in the separation between regulatory and other regions comparable to the GC content. In sequences with a high GC content they are able to filter out some fraction of false positives. Hence, a combination of a regulatory potential with GC content or other similar measures, like the normalized CpG

frequency, appears like a natural extension. Furthermore, using co-occurrence scores for known regulatory regions with a low GC content might improve the performance of the regulatory potential in the case of sequences with a low GC content.
Another application of the regulatory potentials worth to check is the detection of tissue-specific regulatory regions. As a first step, co-occurrence scores derived from known tissue-specific regulatory regions could be used for the construction of the binding site graphs. Moreover, for well known tissues, it could be favourable to restrict the TFs for which one calculates co-occurrence scores and which one uses for the binding site graphs to the ones, which play a role in the respective tissue.

Appendix

Appendix A

German Summary

Transkriptionelle Regulation und Transkriptionsfaktor-Interaktionen Transkriptionsfaktoren, welche die Rate der Transkription von Genen durch Bindung an spezifische Motive auf der DNA und durch Interaktion mit der Transkriptionsmaschinerie regulieren, erfüllen ihren Aufgaben im Zellkern in Kombination mit anderen Transkriptionsfaktoren. Diese Interaktionen können auf der einen Seite sehr spezifisch sein, so daß ein bestimmter Faktor einen bestimmten anderen Faktor benötigt, um seine Funktion auszuführen. Auf der anderen Seite kann die Bandbreite an Interaktionspartnern eines Faktors sehr groß sein, wobei die Funktion verschiedener Komplexe redundant sein, oder sich mit wechselnden Partnern verändern kann.

Für viele Transkriptionsfaktoren sind Bindemotive bekannt, so daß eine bioinformatische Vorhersage von potentiellen Bindungsstellen möglich ist. Wegen verschiedener anderer Ebenen der Regulation ist diese ist im Normalfall jedoch fehlerbehaftet und führt zu einer großen Menge falsch-positiver Vorhersagen.

In der vorliegenden Arbeit nutzen wir vorhergesagte Transkriptionsfaktor-Bindungsstellen (TFBS), um zunächst anhand von häufig beobachteten TFBS-Kombinationen in bekannten regulatorischen Regionen mögliche Interaktionspartner von Transkriptionsfaktoren zu identifizieren. In einem weiteren Schritt nutzen wir das gewonnene Wissen über gehäuftes Auftreten der TFBS-Kombinationen, um mit Hilfe von Bindungsstellen-Graphen DNA-Sequenzen bezüglich ihres regulatorischen Potentials zu charakterisieren.

Arbeitshypothese Die unserer Arbeit zugrunde liegende Annahme ist, daß die vorhergesagten Bindungsstellen interagierender Transkriptions-Faktoren häufiger in Nähe zueinander auftreten, als durch Zufall zu erwarten wäre. Wir rechnen mit einer erschwerten Detektion interagierender Faktoren durch die hohe Falsch-Positiven-Rate bei der Vorhersage individueller Bindungsstellen, gehen jedoch davon aus, daß die vorhandene Information auf der DNA groß genug ist.

Grundlegende Methoden Wir behandeln die biologischen Grundlagen der transkriptionellen Regulation, und experimentelle Methoden zur Ermittlung von Transkriptionsfaktor-Bindungsstellen, Transkriptionsfaktor-Interaktionen und regulatorischen Regionen in Kapitel 2.

In Abschnitt 3.1 erläutern wir den bioinformatischen Weg von der Beschreibung einer experimentell ermittelten Transkriptionsfaktor-Bindungsstelle zu der Suche nach weiteren potentiellen Bindungsstellen eines Faktors. Abschnitt 3.2.2 enthält einen Überblick über verschiedene Ansätze zur Vorhersage von Transkriptionsfaktor-Interaktionen. In Abschnitt 3.3 stellen wir gängige Methoden zur Detektion von regulatorischen Regionen vor.

Vorhersage von Transkriptionsfaktor-Interaktionen Wir beschreiben unsere Methode zur Vorhersage von Transkriptionsfaktor-Interaktionen in Kapitel 4. Sie besteht im einzelnen aus einer Zählmethode, welche auf einem sich über die Sequenz bewegenden Fenster, und speziellen Behandlungsweisen für homotypische Bindungsstellen-Häufungen und überlappende Bindungsstellen basiert. Zur Identifikation von überrepräsentierten Bindungsstellen-Kombinationen berechnen wir die *co-occurrence score* als *log-odds score* der beobachteten und durch Zufall erwarteten Anzahl an Bindungsstellen-Paaren.

In Kapitel 6 testen wir unsere Vorhersagemethode zunächst auf künstlich generierten Datensätzen (Abschnitt 6.1) und können zeigen, daß die Methode auch TFBS-Paare identifizieren kann, die nur schwach überrepräsentiert sind. Die Anwendung der Methode in Abschnitt 6.2 auf Promotor-Sequenzen aus der Bäckerhefe ergibt unter den am stärksten überrepräsentierten Kombinationen einen sehr großen Anteil bereits in der Literatur beschriebener Interaktionen. Darüberhinaus untersuchen wir die Ähnlichkeit der Bindungsstellen interagierender Transkriptionsfaktoren. In Abschnitt 6.5 vergleichen wir die *co-occurrence score* mit der *costat*-Methode, die in Pape et al. [253] vorgestellt wurde. Anschließend untersuchen wir potentielle Transkriptionsfaktor-Interaktionen in Vertebraten. Trotz der weitaus größeren Komplexität des regulatorischen Netzwerkes in Vertebraten finden wir unter den Transkriptionsfaktor-Paaren mit den höchsten *co-occurrence scores* eine große Zahl bereits bekannter Interaktionen – dies sowohl in einer genomweiten Untersuchung auf humanen Promotor-Sequenzen (Abschnitt 6.3), als auch in den Promotoren von HEK-exprimierten Genen (Abschnitt 6.4.2) und gewebspezifisch exprimierten Genen in Maus (Abschnitt 6.4.3).

Vorhersage von Regulatorischen Regionen Die meisten Methoden zur Vorhersage regulatorischer Regionen nutzen explizit oder implizit Sequenzeigenschaften wie den GC-Gehalt oder CpG-Inseln. Unser Ziel ist es, eine Methode zu entwickeln, die weniger von Merkmalen auf niedriger Ebene abhängt, und nutzen aus diesem Grund Informationen über Über- und Unterrepräsentation von Bindungsstellen-Paaren in bekannten regulatorischen Regionen, um das regulatorische Potential einer DNA-Sequenz zu beschreiben. Wir stellen vorhergesagte Bindungsstellen in einem Stück Sequenz durch Knoten in einem Bindungsstellen-Graphen dar. Zunächst werden alle Knoten mit allen anderen durch Kanten verbunden, die als Gewicht die den Endpunkten entsprechende *co-occurrence score* zugewiesen bekommen. Die *co-occurrence scores* stammen aus der vorher in Abschnitt 4 beschriebenen Methode und wurden auf bekannten regulatorischen Regionen des entsprechenden Organismus berechnet. Basierend auf diesem Bindungsstellen-Graphen berechnen wir verschiedene Kantengewicht-basierte regulatorische Potentiale, die die Häufigkeit des Auftretens Promotor-typischer Bindungsstellen-Kombinationen beschreiben. Wir beschreiben diesen Ansatz ausführlich in Kapitel 5.
In Kapitel 7 wenden wir die Methoden auf bekannten regulatorischen Regionen an. Wir berechnen regulatorische Potentiale für die gut untersuchten regulatorischen Regionen des *Pax6* Gens in Maus und für Enhancer-Regionen aus dem VISTA-Datensatz. In Abschnitt

7.2 bewerten wir die Zuverlässigkeit unserer Methode für genomweite Vorhersagen regulatorischer Regionen basierend auf verschiedenen Test-Datensätzen. Diese enthalten echte Promoter- und Enhancer-Sequenzen in verschiedenen Positiv-Sets, und künstliche und intergenische Regionen als Negativ-Set. Unsere Ergebnisse zeigen, dass die unterschiedlichen Scores in der Lage sind, nicht-regulatorische von regulatorischen Sequenzen zu unterscheiden. Obwohl der Faktor mit dem größten Einfluß auf die Vorhersage regulatorischer Funktion nach wie vor der GC-Gehalt ist, ermöglichen es die regulatorischen Potentiale, wegen hohem GC-Gehalts falsch-positive Vorhersagen einer Sequenz herauszufiltern.

In Kapitel 8 fassen wir die Arbeit zusammen und diskutieren die Ergebnisse im Überblick.

Appendix B

Short Curriculum Vitae

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten.

Appendix C

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Berlin, Dezember 2009

Bibliography

[1] I. Abaza and F. Gebauer. Functional domains of drosophila unr in translational control. *RNA*, 14(3):482–490, Mar 2008. doi: /rna.802908. URL http://dx.doi.org//rna.802908.

[2] T. Abeel, Y. Saeys, E. Bonnet, P. Rouzé, and Y. V. de Peer. Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Res*, 18:310–23, Dec 2008. doi: 10.1101/gr.6991408. URL http://dx.doi.org/10.1101/gr.6991408.

[3] I. Abnizova and W. R. Gilks. Studying statistical properties of regulatory DNA sequences, and their use in predicting regulatory regions in the eukaryotic genomes. *Brief Bioinform*, 7(1):48–54, Mar 2006.

[4] A. Acharya, V. Rishi, J. Moll, and C. Vinson. Experimental identification of homodimerizing b-zip families in homo sapiens. *J Struct Biol*, 155(2):130–139, Aug 2006. doi: 10.1016/j.jsb.2006.02.018. URL http://dx.doi.org/10.1016/j.jsb.2006.02.018.

[5] S. Aerts, P. V. Loo, G. Thijs, Y. Moreau, and B. D. Moor. Computational detection of cis -regulatory modules. *Bioinformatics*, 19 Suppl 2:ii5–i14, Oct 2003. URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/19/suppl_2/ii5.

[6] S. Aerts, G. Thijs, M. Dabrowski, Y. Moreau, and B. D. Moor. Comprehensive analysis of the base composition around the transcription start site in metazoa. *BMC Genomics*, 5(1):34, Jun 2004. doi: 10.1186/1471-2164-5-34. URL http://dx.doi.org/10.1186/1471-2164-5-34.

[7] W. C. Aird, J. D. Parvin, P. A. Sharp, and R. D. Rosenberg. The interaction of gata-binding proteins and basal transcription factors with gata box-containing core promoters. a model of tissue-specific gene expression. *J Biol Chem*, 269(2):883–889, Jan 1994.

[8] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell. 4th ed.* Garland Publishing, 2002. URL http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowTOC&rid=mboc4.TOC&depth=2.

[9] W. B. L. Alkema, O. Johansson, J. Lagergren, and W. W. Wasserman. MSCAN: identification of functional clusters of transcription factor binding sites. *Nucleic Acids Res*, 32(Web Server issue):W195–8, Jul 2004. doi: 10.1093/nar/gkh387. URL http://dx.doi.org/10.1093/nar/gkh387.

[10] S. Aparicio, A. Morrison, A. Gould, J. Gilthorpe, C. Chaudhuri, P. Rigby, R. Krumlauf, and S. Brenner. Detecting conserved regulatory elements with the model genome of the japanese puffer fish, fugu rubripes. *Proc Natl Acad Sci U S A*, 92(5):1684–1688, Feb 1995.

[11] M. N. Arbeitman, E. E. M. Furlong, F. Imam, E. Johnson, B. H. Null, B. S. Baker, M. A. Krasnow, M. P. Scott, R. W. Davis, and K. P. White. Gene expression during the life cycle of drosophila melanogaster. *Science*, 297(5590):2270–2275, Sep 2002. doi: 10.1126/science.1072152. URL `http://dx.doi.org/10.1126/science.1072152`.

[12] M. Arnone and E. Davidson. The hardwiring of development: organization and function of genomic regulatory systems. *Development*, 124(10):1851–64, May 1997.

[13] M. M. Babu, N. M. Luscombe, L. Aravind, M. Gerstein, and S. A. Teichmann. Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol*, 14 (3):283–291, Jun 2004. doi: 10.1016/j.sbi.2004.05.004. URL `http://dx.doi.org/10.1016/j.sbi.2004.05.004`.

[14] G. D. Bader and C. W. V. Hogue. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotechnol*, 20(10):991–997, Oct 2002. doi: 10.1038/nbt1002-991. URL `http://dx.doi.org/10.1038/nbt1002-991`.

[15] M. Baes, H. Castelein, L. Desmet, and P. E. Declercq. Antagonism of coup-tf and ppar alpha/rxr alpha on the activation of the malic enzyme gene promoter: modulation by 9-cis ra. *Biochem Biophys Res Commun*, 215(1):338–345, Oct 1995. doi: 10.1006/bbrc.1995.2471. URL `http://dx.doi.org/10.1006/bbrc.1995.2471`.

[16] Y. Bai, Q. Ge, Q. Liu, T. Li, J. Wang, and Z. Lu. A free-labeled method for dna-binding protein detection using a double-stranded dna microarray. *J Nanosci Nanotechnol*, 5(8):1216–1219, Aug 2005.

[17] T. L. Bailey and W. S. Noble. Searching for statistically significant regulatory modules. *Bioinformatics*, 19 Suppl 2:ii16–ii25, Oct 2003. URL `http://bioinformatics.oxfordjournals.org/cgi/content/abstract/19/suppl_2/ii16?ck=nck`.

[18] A. S. Bais, S. Grossmann, and M. Vingron. Incorporating evolution of transcription factor binding sites into annotated alignments. *J Biosci*, 32(5):841–850, Aug 2007.

[19] A. S. Bais, S. Grossmann, and M. Vingron. Simultaneous alignment and annotation of cis-regulatory regions. *Bioinformatics*, 23(2):e44–e49, Jan 2007. doi: 10.1093/bioinformatics/btl305. URL `http://dx.doi.org/10.1093/bioinformatics/btl305`.

[20] V. B. Bajic and S. H. Seah. Dragon gene start finder: an advanced system for finding approximate locations of the start of gene transcriptional units. *Genome Res*, 13(8): 1923–1929, Aug 2003. doi: 10.1101/gr.869803. URL `http://dx.doi.org/10.1101/gr.869803`.

[21] V. B. Bajic and S. H. Seah. Dragon gene start finder identifies approximate locations of the 5' ends of genes. *Nucleic Acids Res*, 31(13):3560–3563, Jul 2003.

[22] V. B. Bajic, S. L. Tan, Y. Suzuki, and S. Sugano. Promoter prediction analysis on the whole human genome. *Nat Biotechnol*, 22(11):1467–73, Nov 2004. doi: 10.1038/nbt1032. URL `http://dx.doi.org/10.1038/nbt1032`.

[23] M. E. Bakkoury, E. Dubois, and F. Messenguy. Recruitment of the yeast mads-box proteins, argri and mcm1 by the pleiotropic factor argriii is required for their stability. *Mol Microbiol*, 35(1):15–31, Jan 2000.

[24] J. E. Balmer and R. Blomhoff. Anecdotes, data and regulatory modules. *Biol Lett*, 2(3):431–434, Sep 2006. doi: 10.1098/rsbl.2006.0484. URL `http://dx.doi.org/10.1098/rsbl.2006.0484`.

[25] N. Banerjee and M. Q. Zhang. Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res*, 31(23):7024–7031, Dec 2003. URL `http://nar.oxfordjournals.org/cgi/content/abstract/31/23/7024`.

[26] Z. Bar-Joseph, G. K. Gerber, T. I. Lee, N. J. Rinaldi, J. Y. Yoo, F. Robert, D. B. Gordon, E. Fraenkel, T. S. Jaakkola, R. A. Young, and D. K. Gifford. Computational discovery of gene modules and regulatory networks. *Nat Biotechnol*, 21(11):1337–1342, Nov 2003. doi: 10.1038/nbt890. URL `http://dx.doi.org/10.1038/nbt890`.

[27] M. Beckstette, R. Homann, R. Giegerich, and S. Kurtz. Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics*, 7:389, 2006. doi: 10.1186/1471-2105-7-389. URL `http://dx.doi.org/10.1186/1471-2105-7-389`.

[28] M. A. Beer and S. Tavazoie. Predicting gene expression from sequence. *Cell*, 117(2):185–198, Apr 2004. URL `http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WSN-4C56FTV-7&_coverDate=04%2F16%2F2004&_alid=321907170&_rdoc=1&_fmt=&_orig=search&_qd=1&_cdi=7051&_sort=d&view=c&_acct=C000056216&_version=1&_urlVersion=0&_userid=2122024&md5=412dfb58951690fea77fb98d0995c5fe`.

[29] D. D. Belsham and P. L. Mellon. Transcription factors oct-1 and c/ebpbeta (ccaat/enhancer-binding protein-beta) are involved in the glutamate/nitric oxide/cyclic-guanosine 5'-monophosphate-mediated repression of mediated repression of gonadotropin-releasing hormone gene expression. *Mol Endocrinol*, 14(2):212–228, Feb 2000.

[30] D. R. Bentley. Whole-genome re-sequencing. *Curr Opin Genet Dev*, 16(6):545–552, Dec 2006. doi: 10.1016/j.gde.2006.10.009. URL `http://dx.doi.org/10.1016/j.gde.2006.10.009`.

[31] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. K. Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. J. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. D. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. C. E. Catenazzi, S. Chang, R. N. Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. F. Fajardo, W. S. Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish,

C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. H. Jones, G.-D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. L. Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. C. Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. C. Rodriguez, P. M. Roe, J. Rogers, M. C. R. Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. E. S. Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, Nov 2008. doi: 10.1038/nature07517. URL `http://dx.doi.org/10.1038/nature07517`.

[32] O. G. Berg and P. H. von Hippel. Selection of dna binding sites by regulatory proteins. statistical-mechanical theory and application to operators and promoters. *J Mol Biol*, 193(4):723–750, Feb 1987.

[33] M. F. Berger, G. Badis, A. R. Gehrke, S. Talukder, A. A. Philippakis, L. Peña-Castillo, T. M. Alleyne, S. Mnaimneh, O. B. Botvinnik, E. T. Chan, F. Khalid, W. Zhang, D. Newburger, S. A. Jaeger, Q. D. Morris, M. L. Bulyk, and T. R. Hughes. Variation in homeodomain dna binding revealed by high-resolution analysis of sequence preferences. *Cell*, 133(7):1266–1276, Jun 2008. doi: 10.1016/j.cell.2008.05.024. URL `http://dx.doi.org/10.1016/j.cell.2008.05.024`.

[34] B. P. Berman, Y. Nibu, B. D. Pfeiffer, P. Tomancak, S. E. Celniker, M. Levine, G. M. Rubin, and M. B. Eisen. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proc Natl Acad Sci U S A*, 99(2):757–62, Jan 2002. doi: 10.1073/pnas.231608898. URL `http://dx.doi.org/10.1073/pnas.231608898`.

[35] B. E. Bernstein, C. L. Liu, E. L. Humphrey, E. O. Perlstein, and S. L. Schreiber. Global nucleosome occupancy in yeast. *Genome Biol*, 5(9):R62, 2004. doi: 10.1186/gb-2004-5-9-r62. URL `http://dx.doi.org/10.1186/gb-2004-5-9-r62`.

[36] Y. Bilu and N. Barkai. The design of transcription-factor binding sites is affected by combinatorial regulation. *Genome Biology*, 6(R103), 2005.

[37] B. L. Black, K. L. Ligon, Y. Zhang, and E. N. Olson. Cooperative transcriptional activation by the neurogenic basic helix-loop-helix protein mash1 and members of the myocyte enhancer factor-2 (mef2) family. *J Biol Chem*, 271(43):26659–26663, Oct 1996.

[38] M. Blanchette, B. Schwikowski, and M. Tompa. Algorithms for phylogenetic footprinting. *J Comput Biol*, 9(2):211–223, 2002. doi: 10.1089/10665270252935421. URL `http://dx.doi.org/10.1089/10665270252935421`.

[39] M. Blanchette, A. R. Bataille, X. Chen, C. Poitras, J. Laganière, C. Lefèbvre, G. Deblois, V. Giguère, V. Ferretti, D. Bergeron, B. Coulombe, and F. Robert. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res*, Apr 2006. doi: 10.1101/gr.4866006. URL `http://dx.doi.org/10.1101/gr.4866006`.

[40] C. Blaschke, M. A. Andrade, C. Ouzounis, and A. Valencia. Automatic extraction of biological information from scientific text: protein-protein interactions. *Proc Int Conf Intell Syst Mol Biol*, pages 60–67, 1999.

[41] D. Boffelli, J. McAuliffe, D. Ovcharenko, K. D. Lewis, I. Ovcharenko, L. Pachter, and E. M. Rubin. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299(5611):1391–1394, Feb 2003. doi: 10.1126/science.1081331. URL `http://dx.doi.org/10.1126/science.1081331`.

[42] D. Bomgardner, B. T. Hinton, and T. T. Turner. 5' hox genes and meis 1, a hox-dna binding cofactor, are expressed in the adult mouse epididymis. *Biol Reprod*, 68(2):644–650, Feb 2003.

[43] P. M. Bowers, S. J. Cokus, D. Eisenberg, and T. O. Yeates. Use of logic relationships to decipher protein network organization. *Science*, 306(5705):2246–2249, Dec 2004. doi: 10.1126/science.1103330. URL `http://dx.doi.org/10.1126/science.1103330`.

[44] P. M. Bowers, B. D. O'Connor, S. J. Cokus, E. Sprinzak, T. O. Yeates, and D. Eisenberg. Utilizing logical relationships in genomic data to decipher cellular processes. *FEBS J*, 272(20):5110–5118, Oct 2005. doi: 10.1111/j.1742-4658.2005.04946.x. URL `http://dx.doi.org/10.1111/j.1742-4658.2005.04946.x`.

[45] K. E. Boyd, J. Wells, J. Gutman, S. M. Bartley, and P. J. Farnham. c-myc target gene specificity is determined by a post-dnabinding mechanism. *Proc Natl Acad Sci U S A*, 95(23):13887–13892, Nov 1998.

[46] T. C. Brelje, M. W. Wessendorf, and R. L. Sorenson. Multicolor laser scanning confocal immunofluorescence microscopy: practical application and limitations. *Methods Cell Biol*, 38:97–181, 1993.

[47] P. Bucher. Weight matrix descriptions of four eukaryotic rna polymerase ii promoter elements derived from 502 unrelated promoter sequences. *J Mol Biol*, 212(4):563–578, Apr 1990.

[48] P. Bucher and E. N. Trifonov. Compilation and analysis of eukaryotic pol ii promoter sequences. *Nucleic Acids Res*, 14(24):10009–10026, Dec 1986.

[49] M. J. Buck and J. D. Lieb. Chip-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83(3):349–360, Mar 2004.

[50] M. J. Buck and J. D. Lieb. A chromatin-mediated mechanism for specification of conditional transcription factor targets. *Nat Genet*, 38(12):1446–1451, Dec 2006. doi: 10.1038/ng1917. URL `http://dx.doi.org/10.1038/ng1917`.

[51] M. Bulyk, E. Gentalen, D. Lockhart, and G. Church. Quantifying DNA-protein interactions by double-stranded DNA arrays. *Nat Biotechnol*, 17(6):573–7, Jun 1999. doi: 10.1038/9878. URL `http://dx.doi.org/10.1038/9878`.

[52] M. L. Bulyk. Computational prediction of transcription-factor binding site locations. *Genome Biol*, 5(1):201, 2003. doi: 10.1186/gb-2003-5-1-201. URL `http://genomebiology.com/2003/5/1/201`.

[53] T. W. Burke and J. T. Kadonaga. Drosophila tfiid binds to a conserved downstream basal promoter element that is present in many tata-box-deficient promoters. *Genes Dev*, 10(6):711–724, Mar 1996.

[54] P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. M. Semple, M. S. Taylor, P. G. Engström, M. C. Frith, A. R. R. Forrest, W. B. Alkema, S. L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawaji, C. Kai, M. Nakamura, H. Konno, K. Nakano, S. Mottagui-Tabar, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S. M. Grimmond, C. A. Wells, V. Orlando, C. Wahlestedt, E. T. Liu, M. Harbers, J. Kawai, V. B. Bajic, D. A. Hume, and Y. Hayashizaki. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*, 38(6):626–635, Jun 2006. doi: 10.1038/ng1789. URL `http://dx.doi.org/10.1038/ng1789`.

[55] S. Cawley, S. Bekiranov, H. H. Ng, P. Kapranov, E. A. Sekinger, D. Kampa, A. Piccolboni, V. Sementchenko, J. Cheng, A. J. Williams, R. Wheeler, B. Wong, J. Drenkow, M. Yamanaka, S. Patel, S. Brubaker, H. Tammana, G. Helt, K. Struhl, and T. R. Gingeras. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding rnas. *Cell*, 116(4):499–509, Feb 2004.

[56] S. Cereghini, S. Saragosti, M. Yaniv, and D. H. Hamer. Sv40-alpha-globulin hybrid minichromosomes. differences in dnase i hypersensitivity of promoter and enhancer sequences. *Eur J Biochem*, 144(3):545–553, Nov 1984.

[57] E. Chargaff. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia*, 6(6):201–209, Jun 1950.

[58] G. Chaurasia, S. Malhotra, J. Russ, S. Schnoegl, C. Hänig, E. E. Wanker, and M. E. Futschik. Unihi 4: new tools for query, analysis and visualization of the human protein-protein interactome. *Nucleic Acids Res*, 37(Database issue):D657–D660, Jan 2009. doi: 10.1093/nar/gkn841. URL `http://dx.doi.org/10.1093/nar/gkn841`.

[59] L. Chen, J. N. Glover, P. G. Hogan, A. Rao, and S. C. Harrison. Structure of the DNA-binding domains from NFAT, Fos and Jun bound specifically to DNA. *Nature*, 392(6671):42–8, Mar 1998. doi: 10.1038/32100. URL `http://dx.doi.org/10.1038/32100`.

[60] X.-W. Chen and M. Liu. Prediction of protein-protein interactions using random decision forest framework. *Bioinformatics*, 21(24):4394–4400, Dec 2005. doi: 10.1093/bioinformatics/bti721. URL `http://dx.doi.org/10.1093/bioinformatics/bti721`.

[61] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell*, 2(1):65–73, Jul 1998.

[62] G. Cochrane, R. Akhtar, J. Bonfield, L. Bower, F. Demiralp, N. Faruque, R. Gibson, G. Hoad, T. Hubbard, C. Hunter, M. Jang, S. Juhos, R. Leinonen, S. Leonard, Q. Lin, R. Lopez, D. Lorenc, H. McWilliam, G. Mukherjee, S. Plaister, R. Radhakrishnan, S. Robinson, S. Sobhany, P. T. Hoopen, R. Vaughan, V. Zalunin, and E. Birney. Petabyte-scale innovations at the european nucleotide archive. *Nucleic Acids Res*, 37 (Database issue):D19–D25, Jan 2009. doi: 10.1093/nar/gkn765. URL `http://dx.doi.org/10.1093/nar/gkn765`.

[63] E. N. C. O. D. E. P. Consortium. Identification and analysis of functional elements in 1genome by the encode pilot project. *Nature*, 447(7146):799–816, Jun 2007. URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=17571346`.

[64] F. A. N. T. O. M. Consortium, H. Suzuki, A. R. R. Forrest, E. van Nimwegen, C. O. Daub, P. J. Balwierz, K. M. Irvine, T. Lassmann, T. Ravasi, Y. Hasegawa, M. J. L. de Hoon, S. Katayama, K. Schroder, P. Carninci, Y. Tomaru, M. Kanamori-Katayama, A. Kubosaki, A. Akalin, Y. Ando, E. Arner, M. Asada, H. Asahara, T. Bailey, V. B. Bajic, D. Bauer, A. G. Beckhouse, N. Bertin, J. Björkegren, F. Brombacher, E. Bulger, A. M. Chalk, J. Chiba, N. Cloonan, A. Dawe, J. Dostie, P. G. Engström, M. Essack, G. J. Faulkner, J. L. Fink, D. Fredman, K. Fujimori, M. Furuno, T. Gojobori, J. Gough, S. M. Grimmond, M. Gustafsson, M. Hashimoto, T. Hashimoto, M. Hatakeyama, S. Heinzel, W. Hide, O. Hofmann, M. Hörnquist, L. Huminiecki, K. Ikeo, N. Imamoto, S. Inoue, Y. Inoue, R. Ishihara, T. Iwayanagi, A. Jacobsen, M. Kaur, H. Kawaji, M. C. Kerr, R. Kimura, S. Kimura, Y. Kimura, H. Kitano, H. Koga, T. Kojima, S. Kondo, T. Konno, A. Krogh, A. Kruger, A. Kumar, B. Lenhard, A. Lennartsson, M. Lindow, M. Lizio, C. Macpherson, N. Maeda, C. A. Maher, M. Maqungo, J. Mar, N. A. Matigian, H. Matsuda, J. S. Mattick, S. Meier, S. Miyamoto, E. Miyamoto-Sato, K. Nakabayashi, Y. Nakachi, M. Nakano, S. Nygaard, T. Okayama, Y. Okazaki, H. Okuda-Yabukami, V. Orlando, J. Otomo, M. Pachkov, N. Petrovsky, C. Plessy, J. Quackenbush, A. Radovanovic, M. Rehli, R. Saito, A. Sandelin, S. Schmeier, C. Schönbach, A. S. Schwartz, C. A. Semple, M. Sera, J. Severin, K. Shirahige, C. Simons, G. S. Laurent, M. Suzuki, T. Suzuki, M. J. Sweet, R. J. Taft, S. Takeda, Y. Takenaka, K. Tan, M. S. Taylor, R. D. Teasdale, J. Tegnér, S. Teichmann, E. Valen, C. Wahlestedt, K. Waki, A. Waterhouse, C. A. Wells, O. Winther, L. Wu, K. Yamaguchi, H. Yanagawa, J. Yasuda, M. Zavolan, D. A. Hume, R. O. S. Center, T. Arakawa, S. Fukuda, K. Imamura, C. Kai, A. Kaiho, T. Kawashima, C. Kawazu, Y. Kitazume, M. Kojima, H. Miura, K. Murakami, M. Murata, N. Ninomiya, H. Nishiyori, S. Noma, C. Ogawa, T. Sano, C. Simon, M. Tagami, Y. Takahashi, J. Kawai, and Y. Hayashizaki. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat Genet*, 41(5):553–562, May 2009. doi: 10.1038/ng.375. URL `http://dx.doi.org/10.1038/ng.375`.

[65] I. H. G. S. Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, Oct 2004.

[66] U. Consortium. The universal protein resource (uniprot) 2009. *Nucleic Acids Res*, 37(Database issue):D169–D174, Jan 2009. doi: 10.1093/nar/gkn664. URL `http://dx.doi.org/10.1093/nar/gkn664`.

[67] A. Cornish-Bowden. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res*, 13(9):3021–3030, May 1985.

[68] R. H. Costa, V. V. Kalinichenko, and L. Lim. Transcription factors in mouse lung development and function. *Am J Physiol Lung Cell Mol Physiol*, 280(5):L823–L838, May 2001.

[69] G. E. Crawford, I. E. Holt, J. Whittle, B. D. Webb, D. Tai, S. Davis, E. H. Margulies, Y. Chen, J. A. Bernat, D. Ginsburg, D. Zhou, S. Luo, T. J. Vasicek, M. J. Daly, T. G. Wolfsberg, and F. S. Collins. Genome-wide mapping of dnase hypersensitive sites using massively parallel signature sequencing (mpss). *Genome Res*, 16(1):123–131, Jan 2006. doi: 10.1101/gr.4074106. URL `http://dx.doi.org/10.1101/gr.4074106`.

[70] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, Aug 1970.

[71] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner. Weblogo: a sequence logo generator. *Genome Res*, 14(6):1188–1190, Jun 2004. doi: 10.1101/gr.849004. URL `http://dx.doi.org/10.1101/gr.849004`.

[72] M. Crossley, M. Merika, and S. Orkin. Self-association of the erythroid transcription factor GATA-1 mediated by its zinc finger domains. *Mol Cell Biol*, 15(5):2448–56, May 1995.

[73] N. Daraselia, A. Yuryev, S. Egorov, S. Novichkova, A. Nikitin, and I. Mazo. Extracting human protein interactions from medline using a full-sentence parser. *Bioinformatics*, 20(5):604–611, Mar 2004. doi: 10.1093/bioinformatics/btg452. URL `http://dx.doi.org/10.1093/bioinformatics/btg452`.

[74] D. Das, N. Banerjee, and M. Q. Zhang. Interacting models of cooperative gene regulation. *Proc Natl Acad Sci U S A*, 101(46):16234–16239, Nov 2004. doi: 10.1073/pnas.0407365101. URL `http://dx.doi.org/10.1073/pnas.0407365101`.

[75] E. H. Davidson. *Genomic regulatory systems: development and evolution.* Academic Press, 2001.

[76] E. H. Davidson. *The Regulatory Genome. Gene Regulatory Networks in Development and Evolution*. Academic Press, San Diego, CA, 2006.

[77] R. Davuluri, I. Grosse, and M. Zhang. Computational identification of promoters and first exons in the human genome. *Nat Genet*, 29(4):412–7, Dec 2001. doi: 10.1038/ng780. URL `http://dx.doi.org/10.1038/ng780`.

[78] W. H. Day and F. R. McMorris. Threshold consensus methods for molecular sequences. *J Theor Biol*, 159(4):481–489, Dec 1992.

[79] J. R. de Wet, K. V. Wood, M. DeLuca, D. R. Helinski, and S. Subramani. Firefly luciferase gene: structure and expression in mammalian cells. *Mol Cell Biol*, 7(2): 725–737, Feb 1987.

[80] E. L. Decker, N. Nehmann, E. Kampen, H. Eibel, P. F. Zipfel, and C. Skerka. Early growth response proteins (egr) and nuclear factors of activated t cells (nfat) form heterodimers and regulate proinflammatory cytokine gene expression. *Nucleic Acids Res*, 31(3):911–921, Feb 2003.

[81] S. Delgado, M. Gómez, A. Bird, and F. Antequera. Initiation of dna replication at cpg islands in mammalian chromosomes. *EMBO J*, 17(8):2426–2435, Apr 1998. doi: 10.1093/emboj/17.8.2426. URL `http://dx.doi.org/10.1093/emboj/17.8.2426`.

[82] C. Dieterich, H. Wang, K. Rateitschak, H. Luz, and M. Vingron. CORG: a database for COmparative Regulatory Genomics. *Nucleic Acids Res*, 31(1):55–7, Jan 2003.

[83] C. Dieterich, S. Rahmann, and M. Vingron. Functional inference from non-random distributions of conserved predicted transcription factor binding sites. *Bioinformatics*, 20 Suppl 1:I109–I115, Aug 2004. doi: 10.1093/bioinformatics/bth908. URL `http://dx.doi.org/10.1093/bioinformatics/bth908a`.

[84] C. Dieterich, S. Grossmann, A. Tanzer, S. Ropcke, P. Arndt, P. Stadler, and M. Vingron. Comparative promoter region analysis powered by CORG. *BMC Genomics*, 6 (1):24, Feb 2005. doi: 10.1186/1471-2164-6-24. URL `http://dx.doi.org/10.1186/1471-2164-6-24`.

[85] B. Dongol, Y. Shah, I. Kim, F. J. Gonzalez, and M. C. Hunt. The acyl-coa thioesterase i is regulated by pparalpha and hnf4alpha via a distal response element in the promoter. *J Lipid Res*, 48(8):1781–1791, Aug 2007. doi: 10.1194/jlr.M700119-JLR200. URL `http://dx.doi.org/10.1194/jlr.M700119-JLR200`.

[86] M. O. Dorschner, M. Hawrylycz, R. Humbert, J. C. Wallace, A. Shafer, J. Kawamoto, J. Mack, R. Hall, J. Goldy, P. J. Sabo, A. Kohli, Q. Li, M. McArthur, and J. A. Stamatoyannopoulos. High-throughput localization of functional elements by quantitative chromatin profiling. *Nat Methods*, 1(3):219–225, Dec 2004. doi: 10.1038/nmeth721. URL `http://dx.doi.org/10.1038/nmeth721`.

[87] T. A. Down and T. J. P. Hubbard. Computational detection and location of transcription start sites in mammalian genomic dna. *Genome Res*, 12(3):458–461, Mar 2002. doi: 10.1101/gr.216102. URL `http://dx.doi.org/10.1101/gr.216102`.

[88] T. Drewes, L. Klein-Hitpass, and G. U. Ryffel. Liver specific transcription factors of the hnf3-, c/ebp- and lfb1-families interact with the a-activator binding site. *Nucleic Acids Res*, 19(23):6383–6389, Dec 1991.

[89] R. Drouin, J. P. Therrien, M. Angers, and S. Ouellet. In vivo dna analysis. *Methods Mol Biol*, 148:175–219, 2001. doi: 10.1385/1-59259-208-2:175. URL `http://dx.doi.org/10.1385/1-59259-208-2:175`.

[90] K. N. Duncliffe, A. G. Bert, M. A. Vadas, and P. N. Cockerill. A t cell-specific enhancer in the interleukin-3 locus is activated cooperatively by oct and nfat elements within a dnase i-hypersensitive site. *Immunity*, 6(2):175–185, Feb 1997.

[91] L. Duret and P. Bucher. Searching for regulatory elements in human noncoding sequences. *Curr Opin Struct Biol*, 7(3):399–406, Jun 1997.

[92] J. Edmonds. Paths, trees, and flowers. *Canad. J. Math.*, 17:449–467, 1965.

[93] D. Eisenberg, E. M. Marcotte, I. Xenarios, and T. O. Yeates. Protein function in the post-genomic era. *Nature*, 405(6788):823–826, Jun 2000. doi: 10.1038/35015694. URL http://dx.doi.org/10.1038/35015694.

[94] R. Elkon, C. Linhart, R. Sharan, R. Shamir, and Y. Shiloh. Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res*, 13(5):773–780, May 2003. doi: 10.1101/gr.947203. URL http://dx.doi.org/10.1101/gr.947203.

[95] L. Elnitski, R. C. Hardison, J. Li, S. Yang, D. Kolbe, P. Eswara, M. J. O'Connor, S. Schwartz, W. Miller, and F. Chiaromonte. Distinguishing regulatory DNA from neutral sites. *Genome Res*, 13(1):64–72, Jan 2003. doi: 10.1101/gr.817703. URL http://dx.doi.org/10.1101/gr.817703.

[96] L. Elnitski, V. X. Jin, P. J. Farnham, and S. J. M. Jones. Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res*, 16(12):1455–1464, Dec 2006. doi: 10.1101/gr.4140006. URL http://dx.doi.org/10.1101/gr.4140006.

[97] A. J. Enright, I. Iliopoulos, N. C. Kyrpides, and C. A. Ouzounis. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757):86–90, Nov 1999. doi: 10.1038/47056. URL http://dx.doi.org/10.1038/47056.

[98] S. Ercan and R. T. Simpson. Global chromatin structure of 45,000 base pairs of chromosome iii in a- and alpha-cell yeast and during mating-type switching. *Mol Cell Biol*, 24(22):10026–10035, Nov 2004. doi: 10.1128/MCB.24.22.10026-10035.2004. URL http://dx.doi.org/10.1128/MCB.24.22.10026-10035.2004.

[99] M. D. Ermolaeva, O. White, and S. L. Salzberg. Prediction of operons in microbial genomes. *Nucleic Acids Res*, 29(5):1216–1221, Mar 2001.

[100] J. Fang, R. J. Haasl, Y. Dong, and G. H. Lushington. Discover protein sequence signatures from protein-protein interaction data. *BMC Bioinformatics*, 6:277, 2005. doi: 10.1186/1471-2105-6-277. URL http://dx.doi.org/10.1186/1471-2105-6-277.

[101] M. Fatemi, M. M. Pao, S. Jeong, E. N. Gal-Yam, G. Egger, D. J. Weisenberger, and P. A. Jones. Footprinting of mammalian promoters: use of a cpg dna methyltransferase revealing nucleosome positions at a single molecule level. *Nucleic Acids Res*, 33 (20):e176, 2005. doi: 10.1093/nar/gni180. URL http://dx.doi.org/10.1093/nar/gni180.

[102] T. Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27:861–74, 2006.

[103] J. Fickett and A. Hatzigeorgiou. Eukaryotic promoter recognition. *Genome Res*, 7(9): 861–78, Sep 1997. URL http://www.genome.org/cgi/content/full/7/9/861.

[104] J. W. Fickett. Coordinate positioning of mef2 and myogenin binding sites. *Gene*, 172 (1):GC19–GC32, Jun 1996.

[105] D. S. Fields, Y. He, A. Y. Al-Uzri, and G. D. Stormo. Quantitative specificity of the mnt repressor. *J Mol Biol*, 271(2):178–194, Aug 1997.

[106] S. Fields and O. Song. A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230):245–246, Jul 1989. doi: 10.1038/340245a0. URL `http://dx.doi.org/10.1038/340245a0`.

[107] M. X. Fitzgerald, J. R. Rojas, J. M. Kim, G. B. Kohlhaw, and R. Marmorstein. Structure of a leu3-dna complex: recognition of everted cgg half-sites by a zn2cys6 binuclear cluster protein. *Structure*, 14(4):725–735, Apr 2006. doi: 10.1016/j.str.2005.11.025. URL `http://dx.doi.org/10.1016/j.str.2005.11.025`.

[108] P. C. FitzGerald, A. Shlyakhtenko, A. A. Mir, and C. Vinson. Clustering of DNA sequences in human promoters. *Genome Res*, 14(8):1562–74, Aug 2004. doi: 10.1101/gr.1953904. URL `http://dx.doi.org/10.1101/gr.1953904`.

[109] T. Fitzwater and B. Polisky. A selex primer. *Methods Enzymol*, 267:275–301, 1996.

[110] M. Fried and D. M. Crothers. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Res*, 9(23):6505–6525, Dec 1981.

[111] M. C. Frith, J. L. Spouge, U. Hansen, and Z. Weng. Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res*, 30(14):3214–24, Jul 2002.

[112] M. C. Frith, M. C. Li, and Z. Weng. Cluster-Buster: Finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res*, 31(13):3666–8, Jul 2003.

[113] H. Fukuda, N. Sano, S. Muto, and M. Horikoshi. Simple histone acetylation plays a complex role in the regulation of gene expression. *Brief Funct Genomic Proteomic*, 5 (3):190–208, Sep 2006. doi: 10.1093/bfgp/ell032. URL `http://dx.doi.org/10.1093/bfgp/ell032`.

[114] M. E. Futschik, G. Chaurasia, and H. Herzel. Comparison of human protein-protein interaction maps. *Bioinformatics*, 23(5):605–611, Mar 2007. doi: 10.1093/bioinformatics/btl683. URL `http://dx.doi.org/10.1093/bioinformatics/btl683`.

[115] H. N. Gabow. *Implementation of algorithms for maximum matching on nonbipartite graphs*. PhD thesis, Stanford University Stanford, CA, USA, 1974.

[116] D. J. Galas and A. Schmitz. Dnase footprinting: a simple method for the detection of protein-dna binding specificity. *Nucleic Acids Res*, 5(9):3157–3170, Sep 1978.

[117] Z. Galil. Efficient algorithms for finding maximum matching in graphs. *ACM Comput. Surv.*, 18(1):23–38, 1986. ISSN 0360-0300. doi: http://doi.acm.org/10.1145/6462.6502.

[118] T. K. B. Gandhi, J. Zhong, S. Mathivanan, L. Karthick, K. N. Chandrika, S. S. Mohan, S. Sharma, S. Pinkert, S. Nagaraju, B. Periaswamy, G. Mishra, K. Nandakumar, B. Shen, N. Deshpande, R. Nayak, M. Sarker, J. D. Boeke, G. Parmigiani, J. Schultz, J. S. Bader, and A. Pandey. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet*, 38(3):285–293, Mar 2006. doi: 10.1038/ng1747. URL `http://dx.doi.org/10.1038/ng1747`.

[119] R. Gangal and P. Sharma. Human pol ii promoter prediction: time series descriptors and machine learning. *Nucleic Acids Res*, 33(4):1332–1336, 2005. doi: 10.1093/nar/gki271. URL `http://dx.doi.org/10.1093/nar/gki271`.

[120] M. Gardiner-Garden and M. Frommer. Cpg islands in vertebrate genomes. *J Mol Biol*, 196(2):261–282, Jul 1987.

[121] M. Garner and A. Revzin. 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 Res*, 9(13):3047–60, Jul 1981.

[122] A. Gautier-Stein, C. Domon-Dell, A. Calon, I. Bady, J.-N. Freund, G. Mithieux, and F. Rajas. Differential regulation of the glucose-6-phosphatase tata box by intestine-specific homeodomain proteins cdx1 and cdx2. *Nucleic Acids Res*, 31(18):5238–5246, Sep 2003.

[123] B. Gazit and H. Cedar. Nuclease sensitivity of active chromatin. *Nucleic Acids Res*, 8(22):5143–5155, Nov 1980.

[124] M. Gilchrist, V. Thorsson, B. Li, A. G. Rust, M. Korb, K. Kennedy, T. Hai, H. Bolouri, and A. Aderem. Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4. *Nature*, 441(7090):173–8, May 2006. doi: 10.1038/nature04768. URL `http://dx.doi.org/10.1038/nature04768`.

[125] C. S. Goh, A. A. Bogan, M. Joachimiak, D. Walther, and F. E. Cohen. Co-evolution of proteins with their interaction partners. *J Mol Biol*, 299(2):283–293, Jun 2000. doi: 10.1006/jmbi.2000.3732. URL `http://dx.doi.org/10.1006/jmbi.2000.3732`.

[126] L. Gordon, A. Y. Chervonenkis, A. J. Gammerman, I. A. Shahmuradov, and V. V. Solovyev. Sequence alignment kernel for recognition of promoter regions. *Bioinformatics*, 19(15):1964–1971, Oct 2003.

[127] Y. H. Grad, F. P. Roth, M. S. Halfon, and G. M. Church. Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in drosophila melanogaster and d.pseudoobscura. *Bioinformatics*, 20(16):2738–2750, Nov 2004. doi: 10.1093/bioinformatics/bth320. URL `http://dx.doi.org/10.1093/bioinformatics/bth320`.

[128] C. Gregori, A. Kahn, and A. L. Pichard. Competition between transcription factors hnf1 and hnf3, and alternative cell-specific activation by dbp and c/ebp contribute to the regulation of the liver-specific aldolase b promoter. *Nucleic Acids Res*, 21(4): 897–903, Feb 1993.

[129] A. Grigoriev. On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Res*, 31(14):4157–4161, Jul 2003.

[130] D. S. Gross and W. T. Garrard. Nuclease hypersensitive sites in chromatin. *Annu Rev Biochem*, 57:159–197, 1988. doi: 10.1146/annurev.bi.57.070188.001111. URL `http://dx.doi.org/10.1146/annurev.bi.57.070188.001111`.

[131] D. GuhaThakurta and G. Stormo. Identifying target sites for cooperatively binding factors. *Bioinformatics*, 17(7):608–21, Jul 2001.

[132] M. Gupta and J. S. Liu. De novo cis-regulatory module elicitation for eukaryotic genomes. *Proc Natl Acad Sci U S A*, 102(20):7079–84, May 2005. doi: 10.1073/pnas.0408743102. URL http://dx.doi.org/10.1073/pnas.0408743102.

[133] J. B. Gurdon and P. Y. Bourillot. Morphogen gradient interpretation. *Nature*, 413 (6858):797–803, Oct 2001. doi: 10.1038/35101500. URL http://dx.doi.org/10.1038/35101500.

[134] U. Güldener, M. Münsterkötter, M. Oesterheld, P. Pagel, A. Ruepp, H.-W. Mewes, and V. Stümpflen. Mpact: the mips protein interaction resource on yeast. *Nucleic Acids Res*, 34(Database issue):D436–D441, Jan 2006. doi: 10.1093/nar/gkj003. URL http://dx.doi.org/10.1093/nar/gkj003.

[135] N. Haiminen, H. Mannila, and E. Terzi. Determining significance of pairwise co-occurrences of events in bursty sequences. *BMC Bioinformatics*, Preprint, 2008.

[136] M. S. Halfon, Y. Grad, G. M. Church, and A. M. Michelson. Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res*, 12(7):1019–28, Jul 2002. doi: 10.1101/gr.228902. URL http://dx.doi.org/10.1101/gr.228902.

[137] O. Hallikas, K. Palin, N. Sinjushina, R. Rautiainen, J. Partanen, E. Ukkonen, and J. Taipale. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell*, 124(1):47–59, Jan 2006. doi: 10.1016/j.cell.2005.10.042. URL http://dx.doi.org/10.1016/j.cell.2005.10.042.

[138] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, Apr 1982.

[139] S. E. Hanlon and J. D. Lieb. Progress and challenges in profiling the dynamics of chromatin and transcription factor binding with dna microarrays. *Curr Opin Genet Dev*, 14(6):697–705, Dec 2004. doi: 10.1016/j.gde.2004.09.008. URL http://dx.doi.org/10.1016/j.gde.2004.09.008.

[140] S. Hannenhalli and S. Levy. Promoter prediction in the human genome. *Bioinformatics*, 17 Suppl 1:S90–S96, 2001.

[141] S. Hannenhalli and S. Levy. Predicting transcription factor synergism. *Nucleic Acids Res*, 30(19):4278–84, Oct 2002. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=12364607.

[142] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J.-B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, Sep 2004. doi: 10.1038/nature02800. URL http://dx.doi.org/10.1038/nature02800.

[143] S. C. Harrison. A structural taxonomy of dna-binding domains. *Nature*, 353(6346): 715–719, Oct 1991. doi: 10.1038/353715a0. URL http://dx.doi.org/10.1038/353715a0.

[144] N. Haubst, J. Berger, V. Radjendirane, J. Graw, J. Favor, G. F. Saunders, A. Stoykova, and M. Götz. Molecular dissection of pax6 function: the specific roles of the paired domain and homeodomain in brain development. *Development*, 131(24):6131–6140, Dec 2004. doi: 10.1242/dev.01524. URL `http://dx.doi.org/10.1242/dev.01524`.

[145] L. M. Hellman and M. G. Fried. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat Protoc*, 2(8):1849–61, 2007. doi: 10.1038/nprot.2007.249. URL `http://dx.doi.org/10.1038/nprot.2007.249`.

[146] G. Z. Hertz and G. D. Stormo. Identifying dna and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7-8):563–577, 1999.

[147] A. Hewetson, E. C. Hendrix, M. Mansharamani, V. H. Lee, and B. S. Chilton. Identification of the rush consensus-binding site by cyclic amplification and selection of targets: demonstration that rush mediates the ability of prolactin to augment progesterone-dependent gene expression. *Mol Endocrinol*, 16(9):2101–2112, Sep 2002.

[148] T. Hiesberger, X. Shao, E. Gourley, A. Reimann, M. Pontoglio, and P. Igarashi. Role of the hepatocyte nuclear factor-1beta (hnf-1beta) c-terminal domain in pkhd1 (arpkd) gene transcription and renal cystogenesis. *J Biol Chem*, 280(11):10578–10586, Mar 2005. doi: 10.1074/jbc.M414121200. URL `http://dx.doi.org/10.1074/jbc.M414121200`.

[149] Y. Ho, A. Gruhler, A. Heilbut, G. D. Bader, L. Moore, S.-L. Adams, A. Millar, P. Taylor, K. Bennett, K. Boutilier, L. Yang, C. Wolting, I. Donaldson, S. Schandorff, J. Shewnarane, M. Vo, J. Taggart, M. Goudreault, B. Muskat, C. Alfarano, D. Dewar, Z. Lin, K. Michalickova, A. R. Willems, H. Sassi, P. A. Nielsen, K. J. Rasmussen, J. R. Andersen, L. E. Johansen, L. H. Hansen, H. Jespersen, A. Podtelejnikov, E. Nielsen, J. Crawford, V. Poulsen, B. D. Sørensen, J. Matthiesen, R. C. Hendrickson, F. Gleeson, T. Pawson, M. F. Moran, D. Durocher, M. Mann, C. W. V. Hogue, D. Figeys, and M. Tyers. Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868):180–183, Jan 2002. doi: 10.1038/415180a. URL `http://dx.doi.org/10.1038/415180a`.

[150] M. L. Howard and E. H. Davidson. cis-regulatory control circuits in development. *Dev Biol*, 271(1):109–118, Jul 2004. doi: 10.1016/j.ydbio.2004.03.031. URL `http://dx.doi.org/10.1016/j.ydbio.2004.03.031`.

[151] J. Hu, H. Hu, and X. Li. Mopat: a graph-based method to predict recurrent cis-regulatory modules from known motifs. *Nucleic Acids Res*, 36(13):4488–4497, Aug 2008. doi: 10.1093/nar/gkn407. URL `http://dx.doi.org/10.1093/nar/gkn407`.

[152] Z. Hu, B. Hu, and J. F. Collins. Prediction of synergistic transcription factors by function conservation. *Genome Biol*, 8(12):R257, 2007. doi: 10.1186/gb-2007-8-12-r257. URL `http://dx.doi.org/10.1186/gb-2007-8-12-r257`.

[153] T. J. P. Hubbard, B. L. Aken, S. Ayling, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, L. Clarke, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Graf, S. Haider, M. Hammond, R. Holland, K. Howe, A. Jenkinson, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker,

B. Pritchard, D. Rios, M. Schuster, G. Slater, D. Smedley, W. Spooner, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, S. Wilder, A. Zadissa, E. Birney, F. Cunningham, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, A. Kasprzyk, G. Proctor, J. Smith, S. Searle, and P. Flicek. Ensembl 2009. *Nucleic Acids Res*, 37(Database issue):D690–D697, Jan 2009. doi: 10.1093/nar/gkn828. URL `http://dx.doi.org/10.1093/nar/gkn828`.

[154] J. Hughes, P. Estep, S. Tavazoie, and G. Church. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol*, 296(5):1205–14, Mar 2000. doi: 10.1006/jmbi.2000.3519. URL `http://dx.doi.org/10.1006/jmbi.2000.3519`.

[155] W. Humphrey, A. Dalke, and K. Schulten. Vmd: visual molecular dynamics. *J Mol Graph*, 14(1):33–8, 27–8, Feb 1996.

[156] M. Huynen, B. Snel, W. Lathe, and P. Bork. Exploitation of gene context. *Curr Opin Struct Biol*, 10(3):366–370, Jun 2000.

[157] I. P. Ioshikhes and M. Q. Zhang. Large-scale human promoter mapping using cpg islands. *Nat Genet*, 26(1):61–63, Sep 2000. doi: 10.1038/79189. URL `http://dx.doi.org/10.1038/79189`.

[158] H. Ishida, K. Ueda, K. Ohkawa, Y. Kanazawa, A. Hosui, F. Nakanishi, E. Mita, A. Kasahara, Y. Sasaki, M. Hori, and N. Hayashi. Identification of multiple transcription factors, hlf, ftf, and e4bp4, controlling hepatitis b virus enhancer ii. *J Virol*, 74 (3):1241–1251, Feb 2000.

[159] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98(8):4569–4574, Apr 2001. doi: 10.1073/pnas.061034498. URL `http://dx.doi.org/10.1073/pnas.061034498`.

[160] S. Itzkovitz, T. Tlusty, and U. Alon. Coding limits on the number of transcription factors. *BMC Genomics*, 7:239, 2006. doi: 10.1186/1471-2164-7-239. URL `http://dx.doi.org/10.1186/1471-2164-7-239`.

[161] V. R. Iyer, C. E. Horak, C. S. Scafe, D. Botstein, M. Snyder, and P. O. Brown. Genomic binding sites of the yeast cell-cycle transcription factors sbf and mbf. *Nature*, 409(6819):533–538, Jan 2001. doi: 10.1038/35054095. URL `http://dx.doi.org/10.1038/35054095`.

[162] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644):449–453, Oct 2003. doi: 10.1126/science.1087361. URL `http://dx.doi.org/10.1126/science.1087361`.

[163] A. G. Jegga, S. P. Sherwood, J. W. Carman, A. T. Pinski, J. L. Phillips, J. P. Pestian, and B. J. Aronow. Detection and visualization of compositionally similar cis-regulatory element clusters in orthologous and coordinately controlled genes. *Genome Res*, 12 (9):1408–17, Sep 2002. doi: 10.1101/gr.255002. URL `http://dx.doi.org/10.1101/gr.255002`.

[164] F. Jenkins, P. N. Cockerill, D. Bohmann, and M. F. Shannon. Multiple signals are required for function of the human granulocyte-macrophage colony-stimulating factor gene promoter in t cells. *J Immunol*, 155(3):1240–1251, Aug 1995.

[165] O. Johansson, W. Alkema, W. Wasserman, and J. Lagergren. Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm. *Bioinformatics*, 19 Suppl 1:i169–76, 2003.

[166] R. Jothi, P. F. Cherukuri, A. Tasneem, and T. M. Przytycka. Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. *J Mol Biol*, 362(4):861–875, Sep 2006. doi: 10.1016/j.jmb.2006.07.072. URL `http://dx.doi.org/10.1016/j.jmb.2006.07.072`.

[167] J. T. Kadonaga and R. Tjian. Affinity purification of sequence-specific dna binding proteins. *Proc Natl Acad Sci U S A*, 83(16):5889–5893, Aug 1986.

[168] S.-H. L. Kang, K. Vieira, and J. Bungert. Combining chromatin immunoprecipitation and dna footprinting: a novel method to analyze protein-dna interactions in vivo. *Nucleic Acids Res*, 30(10):e44, May 2002.

[169] M. Kato, N. Hata, N. Banerjee, B. Futcher, and M. Q. Zhang. Identifying combinatorial regulation of transcription factors and binding motifs. *Genome Biol*, 5 (8):R56, 2004. doi: 10.1186/gb-2004-5-8-r56. URL `http://dx.doi.org/10.1186/gb-2004-5-8-r56`.

[170] J. Kawai, A. Shinagawa, K. Shibata, M. Yoshino, M. Itoh, Y. Ishii, T. Arakawa, A. Hara, Y. Fukunishi, H. Konno, J. Adachi, S. Fukuda, K. Aizawa, M. Izawa, K. Nishi, H. Kiyosawa, S. Kondo, I. Yamanaka, T. Saito, Y. Okazaki, T. Gojobori, H. Bono, T. Kasukawa, R. Saito, K. Kadota, H. Matsuda, M. Ashburner, S. Batalov, T. Casavant, W. Fleischmann, T. Gaasterland, C. Gissi, B. King, H. Kochiwa, P. Kuehl, S. Lewis, Y. Matsuo, I. Nikaido, G. Pesole, J. Quackenbush, L. M. Schriml, F. Staubli, R. Suzuki, M. Tomita, L. Wagner, T. Washio, K. Sakai, T. Okido, M. Furuno, H. Aono, R. Baldarelli, G. Barsh, J. Blake, D. Boffelli, N. Bojunga, P. Carninci, M. F. de Bonaldo, M. J. Brownstein, C. Bult, C. Fletcher, M. Fujita, M. Gariboldi, S. Gustincich, D. Hill, M. Hofmann, D. A. Hume, M. Kamiya, N. H. Lee, P. Lyons, L. Marchionni, J. Mashima, J. Mazzarelli, P. Mombaerts, P. Nordone, B. Ring, M. Ringwald, I. Rodriguez, N. Sakamoto, H. Sasaki, K. Sato, C. Schönbach, T. Seya, Y. Shibata, K. F. Storch, H. Suzuki, K. Toyo-oka, K. H. Wang, C. Weitz, C. Whittaker, L. Wilming, A. Wynshaw-Boris, K. Yoshida, Y. Hasegawa, H. Kawaji, S. Kohtsuki, Y. Hayashizaki, R. I. K. E. N. G. E. R. G. P. I. Team, and the FANTOM Consortium. Functional annotation of a full-length mouse cdna collection. *Nature*, 409(6821): 685–690, Feb 2001.

[171] A. Kel, O. Kel-Margoulis, V. Babenko, and E. Wingender. Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells. *J Mol Biol*, 288(3):353–76, May 1999. doi: 10.1006/jmbi.1999.2684. URL `http://dx.doi.org/10.1006/jmbi.1999.2684`.

[172] A. E. Kel, E. Gössling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, and E. Wingender. Match: A tool for searching transcription factor binding sites in dna sequences. *Nucleic Acids Res*, 31(13):3576–3579, Jul 2003.

[173] O. V. Kel-Margoulis, A. E. Kel, I. Reuter, I. V. Deineko, and E. Wingender. TRANSCompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res*, 30(1):332–4, Jan 2002.

[174] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423 (6937):241–254, May 2003. doi: 10.1038/nature01644. URL `http://dx.doi.org/10.1038/nature01644`.

[175] P. Kheradpour, A. Stark, S. Roy, and M. Kellis. Reliable prediction of regulator targets using 12 drosophila genomes. *Genome Res*, 17(12):1919–1931, Dec 2007. doi: 10.1101/gr.7090407. URL `http://dx.doi.org/10.1101/gr.7090407`.

[176] S. Khorasanizadeh and F. Rastinejad. Transcription factors: the right combination for the dna lock. *Curr Biol*, 9(12):R456–R458, Jun 1999.

[177] G. Khoury and P. Gruss. Enhancer elements. *Cell*, 33(2):313–314, Jun 1983.

[178] J.-Y. Kim, S.-M. Moon, H.-J. Ryu, J.-J. Kim, H.-T. Kim, C. Park, K. Kimm, B. Oh, and J.-K. Lee. Identification of regulatory polymorphisms in the TNF-TNF receptor superfamily. *Immunogenetics*, 57(5):297–303, Jun 2005. doi: 10.1007/s00251-005-0800-8. URL `http://dx.doi.org/10.1007/s00251-005-0800-8`.

[179] C. Kimura-Yoshida, K. Kitajima, I. Oda-Ishii, E. Tian, M. Suzuki, M. Yamamoto, T. Suzuki, M. Kobayashi, S. Aizawa, and I. Matsuo. Characterization of the pufferfish otx2 cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. *Development*, 131(1):57–71, Jan 2004. doi: 10.1242/dev.00877. URL `http://dx.doi.org/10.1242/dev.00877`.

[180] C. V. Kirchhamer and E. H. Davidson. Spatial and temporal information processing in the sea urchin embryo: modular and intramodular organization of the cyiiia gene cis-regulatory system. *Development*, 122(1):333–348, Jan 1996.

[181] S. Klein-Hessling, G. Schneider, A. Heinfling, S. Chuvpilo, and E. Serfling. Hmg i(y) interferes with the dna binding of nf-at factors and the induction of the interleukin 4 promoter in t cells. *Proc Natl Acad Sci U S A*, 93(26):15311–15316, Dec 1996.

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

[183] J. Klug. Ku autoantigen is a potential major cause of nonspecific bands in electrophoretic mobility shift assays. *Biotechniques*, 22(2):212–4, 216, Feb 1997.

[184] P. S. Knoepfler, K. R. Calvo, H. Chen, S. E. Antonarakis, and M. P. Kamps. Meis1 and pknox1 bind dna cooperatively with pbx1 utilizing an interaction surface disrupted in oncoprotein e2a-pbx1. *Proc Natl Acad Sci U S A*, 94(26):14553–14558, Dec 1997.

[185] K. C. Knower, S. Kelly, and V. R. Harley. Turning on the male–sry, sox9 and sex determination in mammals. *Cytogenet Genome Res*, 101(3-4):185–198, 2003. doi: 10.1159/000074336. URL `http://dx.doi.org/10.1159/000074336`.

[186] S. Knudsen. Promoter2.0: for the recognition of polii promoter sequences. *Bioinformatics*, 15(5):356–361, May 1999.

[187] A. Koike and T. Takagi. Prediction of protein-protein interaction sites using support vector machines. *Protein Eng Des Sel*, 17(2):165–173, Feb 2004. doi: 10.1093/protein/gzh020. URL `http://dx.doi.org/10.1093/protein/gzh020`.

[188] D. Kolbe, J. Taylor, L. Elnitski, P. Eswara, J. Li, W. Miller, R. Hardison, and F. Chiaromonte. Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome Res*, 14(4):700–707, Apr 2004. doi: 10.1101/gr.1976004. URL `http://dx.doi.org/10.1101/gr.1976004`.

[189] G. Kreiman. Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes. *Nucleic Acids Res*, 32(9):2889–900, 2004. doi: 10.1093/nar/gkh614. URL `http://dx.doi.org/10.1093/nar/gkh614`.

[190] W. Krivan and W. Wasserman. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res*, 11(9):1559–66, Sep 2001. doi: 10.1101/gr.180601. URL `http://dx.doi.org/10.1101/gr.180601`.

[191] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Res Logist. Quart*, 2:253–258, 1955.

[192] L. Laricchia-Robbio, R. Fazzina, D. Li, C. R. Rinaldi, K. K. Sinha, S. Chakraborty, and G. Nucifora. Point mutations in two evi1 zn fingers abolish evi1-gata1 interaction and allow erythroid differentiation of murine bone marrow cells. *Mol Cell Biol*, 26 (20):7658–7666, Oct 2006. doi: 10.1128/MCB.00363-06. URL `http://dx.doi.org/10.1128/MCB.00363-06`.

[193] G. M. Lawson, B. J. Knoll, C. J. March, S. L. Woo, M. J. Tsai, and B. W. O'Malley. Definition of 5' and 3' structural boundaries of the chromatin domain containing the ovalbumin multigene family. *J Biol Chem*, 257(3):1501–1507, Feb 1982.

[194] T. I. Lee and R. A. Young. Transcription of eukaryotic protein-coding genes. *Annu Rev Genet*, 34:77–137, 2000. doi: 10.1146/annurev.genet.34.1.77. URL `http://dx.doi.org/10.1146/annurev.genet.34.1.77`.

[195] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J.-B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298(5594):799–804, Oct 2002. doi: 10.1126/science.1075090. URL `http://dx.doi.org/10.1126/science.1075090`.

[196] B. Lemon and R. Tjian. Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev*, 14(20):2551–2569, Oct 2000.

[197] M. Levine and R. Tjian. Transcription regulation and animal diversity. *Nature*, 424 (6945):147–51, Jul 2003. doi: 10.1038/nature01763. URL `http://dx.doi.org/10.1038/nature01763`.

[198] S. Levy, S. Hannenhalli, and C. Workman. Enrichment of regulatory signals in conserved non-coding genomic sequence. *Bioinformatics*, 17(10):871–7, Oct 2001. URL `http://bioinformatics.oupjournals.org/cgi/reprint/17/10/871`.

[199] T. Li, Y. Jin, A. K. Vershon, and C. Wolberger. Crystal structure of the mata1/matalpha2 homeodomain heterodimer in complex with dna containing an a-tract. *Nucleic Acids Res*, 26(24):5707–5718, Dec 1998.

[200] Z. Li, S. V. Calcar, C. Qu, W. K. Cavenee, M. Q. Zhang, and B. Ren. A global transcriptional regulatory role for c-myc in burkitt's lymphoma cells. *Proc Natl Acad Sci U S A*, 100(14):8164–8169, Jul 2003. doi: 10.1073/pnas.1332764100. URL `http://dx.doi.org/10.1073/pnas.1332764100`.

[201] A. P. Lifanov, V. J. Makeev, A. G. Nazina, and D. A. Papatsenko. Homotypic regulatory clusters in Drosophila. *Genome Res*, 13(4):579–588, Apr 2003. doi: 10.1101/gr.668403. URL `http://dx.doi.org/10.1101/gr.668403`.

[202] R. P. Lifton, M. L. Goldberg, R. W. Karp, and D. S. Hogness. The organization of the histone genes in drosophila melanogaster: functional and evolutionary implications. *Cold Spring Harb Symp Quant Biol*, 42 Pt 2:1047–1051, 1978.

[203] R. Liu and D. J. States. Consensus promoter identification in the human genome utilizing expressed gene markers and gene modeling. *Genome Res*, 12(3):462–469, Mar 2002. doi: 10.1101/gr.198002. URL `http://dx.doi.org/10.1101/gr.198002`.

[204] X. Liu, D. M. Noll, J. D. Lieb, and N. D. Clarke. Dip-chip: rapid and accurate determination of dna-binding specificity. *Genome Res*, 15(3):421–427, Mar 2005. doi: 10.1101/gr.3256505. URL `http://dx.doi.org/10.1101/gr.3256505`.

[205] P. V. Loo and P. Marynen. Computational methods for the detection of cis-regulatory modules. *Brief Bioinform*, Jun 2009. doi: 10.1093/bib/bbp025. URL `http://dx.doi.org/10.1093/bib/bbp025`.

[206] P. V. Loo, S. Aerts, B. Thienpont, B. D. Moor, Y. Moreau, and P. Marynen. Moduleminer - improved computational detection of cis-regulatory modules: are there different modes of gene regulation in embryonic development and adult tissues? *Genome Biol*, 9(4):R66, Apr 2008. doi: 10.1186/gb-2008-9-4-r66. URL `http://dx.doi.org/10.1186/gb-2008-9-4-r66`.

[207] N. Luscombe, S. Austin, H. Berman, and J. Thornton. An overview of the structures of protein-DNA complexes. *Genome Biol*, 1(1):REVIEWS001, 2000. URL `http://genomebiology.com/2000/1/1/REVIEWS/001`.

[208] K. D. MacIsaac, T. Wang, D. B. Gordon, D. K. Gifford, G. D. Stormo, and E. Fraenkel. An improved map of conserved regulatory sites for Saccharomyces cerevisiae. *BMC Bioinformatics*, 7:113, 2006. doi: 10.1186/1471-2105-7-113. URL `http://dx.doi.org/10.1186/1471-2105-7-113`.

[209] S. Mahony, P. E. Auron, and P. V. Benos. Dna familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS Comput Biol*, 3(3):e61, Mar 2007. doi: 10.1371/journal.pcbi.0030061. URL `http://dx.doi.org/10.1371/journal.pcbi.0030061`.

[210] Y. Makita, M. Nakao, N. Ogasawara, and K. Nakai. Dbtbs: database of transcriptional regulation in bacillus subtilis and its contribution to comparative genomics. *Nucleic Acids Res*, 32(Database issue):D75–D77, Jan 2004. doi: 10.1093/nar/gkh074. URL `http://dx.doi.org/10.1093/nar/gkh074`.

[211] M. Mannervik. Target genes of homeodomain proteins. *Bioessays*, 21(4):267–270, Apr 1999. doi: 3.0.CO;2-C. URL http://www3.interscience.wiley.com/journal/61001929/abstract.

[212] C. J. V. Marcotte and E. M. Marcotte. Predicting functional linkages from gene fusions with confidence. *Appl Bioinformatics*, 1(2):93–100, 2002.

[213] E. M. Marcotte, M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751–753, Jul 1999. URL http://www.sciencemag.org/cgi/content/full/285/5428/751.

[214] E. R. Mardis. Chip-seq: welcome to the new frontier. *Nat Methods*, 4(8):613–614, Aug 2007. doi: 10.1038/nmeth0807-613. URL http://dx.doi.org/10.1038/nmeth0807-613.

[215] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437 (7057):376–380, Sep 2005. doi: 10.1038/nature03959. URL http://dx.doi.org/10.1038/nature03959.

[216] S. Martin, D. Roe, and J.-L. Faulon. Predicting protein-protein interactions using signature products. *Bioinformatics*, 21(2):218–226, Jan 2005. doi: 10.1093/bioinformatics/bth483. URL http://dx.doi.org/10.1093/bioinformatics/bth483.

[217] G. A. Maston, S. K. Evans, and M. R. Green. Transcriptional Regulatory Elements in the Human Genome. *Annu Rev Genomics Hum Genet*, 7:29–59, Sep 2006. doi: om.7.080505.115623. URL http://dx.doi.org/om.7.080505.115623.

[218] S. Mathivanan, B. Periaswamy, T. K. B. Gandhi, K. Kandasamy, S. Suresh, R. Mohmood, Y. L. Ramachandra, and A. Pandey. An evaluation of human protein-protein interaction data in the public domain.pon. *BMC Bioinformatics*, 7 Suppl 5: S19, 2006. doi: 10.1186/1471-2105-7-S5-S19. URL http://dx.doi.org/10.1186/1471-2105-7-S5-S19.

[219] V. Matys, O. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A. Kel, and E. Wingender. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34(Database issue): D108–10, Jan 2006. doi: 10.1093/nar/gkj143. URL http://dx.doi.org/10.1093/nar/gkj143.

[220] M. McArthur, S. Gerum, and G. Stamatoyannopoulos. Quantification of dnasei-sensitivity by real-time pcr: quantitative analysis of dnasei-hypersensitivity of the mouse beta-globin lcr. *J Mol Biol*, 313(1):27–34, Oct 2001. doi: 10.1006/jmbi.2001.4969. URL `http://dx.doi.org/10.1006/jmbi.2001.4969`.

[221] A. D. McLachlan. Analysis of gene duplication repeats in the myosin rod. *J Mol Biol*, 169(1):15–30, Sep 1983.

[222] K. Mehlhorn and S. Näher. *LEDA: a platform for combinatorial and geometric computing*. Cambridge University Press, November 1999.

[223] K. Mehlhorn and G. Schaefer. Implementation of o(nm logn)o(nmlogn) weighted matchings in general graphs. the power of data structures. *Lecture Notes in Computer Science*, 1982:23, 2001.

[224] H. Mewes, D. Frishman, C. Gruber, B. Geier, D. Haase, A. Kaps, K. Lemcke, G. Mannhaupt, F. Pfeiffer, C. Schüller, S. Stocker, and B. Weil. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*, 28(1):37–40, Jan 2000. URL `http://nar.oxfordjournals.org/cgi/content/abstract/28/1/37`.

[225] H. Mewes, D. Frishman, U. Güldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Münsterkötter, S. Rudd, and B. Weil. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*, 30(1):31–4, Jan 2002. URL `http://nar.oxfordjournals.org/cgi/content/abstract/30/1/31`.

[226] H. W. Mewes, D. Frishman, K. F. X. Mayer, M. Münsterkötter, O. Noubibou, P. Pagel, T. Rattei, M. Oesterheld, A. Ruepp, and V. Stümpflen. Mips: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res*, 34(Database issue):D169–D172, Jan 2006. doi: 10.1093/nar/gkj148. URL `http://dx.doi.org/10.1093/nar/gkj148`.

[227] G. Moreno-Hagelsieb and J. Collado-Vides. A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics*, 18 Suppl 1:S329–S336, 2002.

[228] R. Morgan. Conservation of sequence and function in the Pax6 regulatory elements. *Trends Genet*, 20(7):283–7, Jul 2004. doi: 10.1016/j.tig.2004.04.009. URL `http://dx.doi.org/10.1016/j.tig.2004.04.009`.

[229] R. Morgan. Hox genes: a continuation of embryonic patterning? *Trends Genet*, 22 (2):67–69, Feb 2006. doi: 10.1016/j.tig.2005.11.004. URL `http://dx.doi.org/10.1016/j.tig.2005.11.004`.

[230] S. Mukherjee, M. F. Berger, G. Jona, X. S. Wang, D. Muzzey, M. Snyder, R. A. Young, and M. L. Bulyk. Rapid analysis of the dna-binding specificities of transcription factors with dna microarrays. *Nat Genet*, 36(12):1331–1339, Dec 2004. doi: 10.1038/ng1473. URL `http://dx.doi.org/10.1038/ng1473`.

[231] C. W. Müller and B. G. Herrmann. Crystallographic structure of the t domain-dna complex of the brachyury transcription factor. *Nature*, 389(6653):884–888, Oct 1997. doi: 10.1038/39929. URL `http://dx.doi.org/10.1038/39929`.

[232] F. Müller, D. W. Williams, J. Kobolák, L. Gauvry, G. Goldspink, L. Orbán, and N. Maclean. Activator effect of coinjected enhancers on the muscle-specific expression of promoters in zebrafish embryos. *Mol Reprod Dev*, 47(4):404–412, Aug 1997. doi: 3.0.CO;2-O. URL `http://dx.doi.org/3.0.CO;2-O`.

[233] N. Nagamine, Y. Kawada, and Y. Sakakibara. Identifying cooperative transcriptional regulations using protein-protein interactions. *Nucleic Acids Res*, 33(15):4828–4837, 2005. doi: 10.1093/nar/gki793. URL `http://dx.doi.org/10.1093/nar/gki793`.

[234] L. Narlikar and I. Ovcharenko. Identifying regulatory elements in eukaryotic genomes. *Brief Funct Genomic Proteomic*, Jun 2009. doi: 10.1093/bfgp/elp014. URL `http://dx.doi.org/10.1093/bfgp/elp014`.

[235] V. Neduva, R. Linding, I. Su-Angrand, A. Stark, F. de Masi, T. J. Gibson, J. Lewis, L. Serrano, and R. B. Russell. Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol*, 3(12):e405, Dec 2005. doi: 10.1371/journal.pbio.0030405. URL `http://dx.doi.org/10.1371/journal.pbio.0030405`.

[236] E. V. Nimwegen. Scaling laws in the functional content of genomes. *Trends Genet*, 19 (9):479–484, Sep 2003.

[237] C. Notredame. Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics*, 3(1):131–144, Jan 2002. doi: 10.1517/14622416.3.1.131. URL `http://dx.doi.org/10.1517/14622416.3.1.131`.

[238] A. Nourani, M. Wesolowski-Louvel, T. Delaveau, C. Jacq, and A. Delahodde. Multiple-drug-resistance phenomenon in the yeast saccharomyces cerevisiae: involvement of two hexose transporters. *Mol Cell Biol*, 17(9):5453–5460, Sep 1997.

[239] D. T. Odom, N. Zizlsperger, D. B. Gordon, G. W. Bell, N. J. Rinaldi, H. L. Murray, T. L. Volkert, J. Schreiber, P. A. Rolfe, D. K. Gifford, E. Fraenkel, G. I. Bell, and R. A. Young. Control of pancreas and liver gene expression by HNF transcription factors. *Science*, 303(5662):1378–81, Feb 2004. doi: 10.1126/science.1089769. URL `http://dx.doi.org/10.1126/science.1089769`.

[240] U. Ohler, S. Harbeck, H. Niemann, E. Nöth, and M. Reese. Interpolated markov chains for eukaryotic promoter recognition. *Bioinformatics*, 15(5):362–9, May 1999.

[241] U. Ohler, G. chun Liao, H. Niemann, and G. M. Rubin. Computational analysis of core promoters in the drosophila genome. *Genome Biol*, 3(12):RESEARCH0087, 2002.

[242] A. R. Oliphant, C. J. Brandl, and K. Struhl. Defining the sequence specificity of dna-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast gcn4 protein. *Mol Cell Biol*, 9(7):2944–2949, Jul 1989.

[243] T. Onizuka, S. Endo, M. Hirano, S. Kanai, and H. Akiyama. Design of a fluorescent electrophoretic mobility shift assay improved for the quantitative and multiple analysis of protein-dna complexes. *Biosci Biotechnol Biochem*, 66(12):2732–2734, Dec 2002.

[244] T. Ono, H. Hishigaki, A. Tanigami, and T. Takagi. Automated extraction of information on protein-protein interactions from the biological literature. *Bioinformatics*, 17 (2):155–161, Feb 2001.

[245] V. Orlando. Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem Sci*, 25(3):99–104, Mar 2000.

[246] G. Orphanides, T. Lagrange, and D. Reinberg. The general transcription factors of rna polymerase ii. *Genes Dev*, 10(21):2657–2683, Nov 1996.

[247] N. Osumi, H. Shinohara, K. Numayama-Tsuruta, and M. Maekawa. Concise review: Pax6 transcription factor contributes to both embryonic and adult neurogenesis as a multifunctional regulator. *Stem Cells*, 26(7):1663–1672, Jul 2008. doi: 10.1634/stemcells.2007-0884. URL `http://dx.doi.org/10.1634/stemcells.2007-0884`.

[248] P. Pagel, P. Wong, and D. Frishman. A domain interaction map based on phylogenetic profiling. *J Mol Biol*, 344(5):1331–1346, Dec 2004. doi: 10.1016/j.jmb.2004.10.019. URL `http://dx.doi.org/10.1016/j.jmb.2004.10.019`.

[249] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stümpflen, H.-W. Mewes, A. Ruepp, and D. Frishman. The mips mammalian protein-protein interaction database. *Bioinformatics*, 21(6):832–834, Mar 2005. doi: 10.1093/bioinformatics/bti115. URL `http://dx.doi.org/10.1093/bioinformatics/bti115`.

[250] X. Pan and J. Heitman. Protein kinase a operates a molecular switch that governs yeast pseudohyphal differentiation. *Mol Cell Biol*, 22(12):3981–3993, Jun 2002.

[251] U. J. Pape and M. Vingron. Statistics for co-occurrence of dna motifs. In *International Workshop on Applied Probability*, 2008.

[252] U. J. Pape, S. Rahmann, and M. Vingron. Natural Similarity Measures between Position Frequency Matrices with an Application to Clustering. *Bioinformatics*, 24 (3):350–357, Jan 2008. doi: 10.1093/bioinformatics/btm610. URL `http://dx.doi.org/10.1093/bioinformatics/btm610`.

[253] U. J. Pape, H. Klein, and M. Vingron. Statistical detection of co-operative transcription factors with similarity adjustment. *Bioinformatics*, Mar 2009. doi: 10.1093/bioinformatics/btp143. URL `http://dx.doi.org/10.1093/bioinformatics/btp143`.

[254] F. Pazos and A. Valencia. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng*, 14(9):609–614, Sep 2001.

[255] A. G. Pedersen, P. Baldi, Y. Chauvin, and S. Brunak. Dna structure in human rna polymerase ii promoters. *J Mol Biol*, 281(4):663–673, Aug 1998. doi: 10.1006/jmbi.1998.1972. URL `http://dx.doi.org/10.1006/jmbi.1998.1972`.

[256] A. G. Pedersen, P. Baldi, Y. Chauvin, and S. Brunak. The biology of eukaryotic promoter prediction–a review. *Comput Chem*, 23(3-4):191–207, Jun 1999.

[257] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, 96(8):4285–4288, Apr 1999.

[258] H. Peretz and D. Elson. Synthesis of a cleavable protein-crosslinking reagent for the investigation of ribosome structure. *Eur J Biochem*, 63(1):77–82, Mar 1976.

[259] D. Pfeifer, R. Kist, K. Dewar, K. Devon, E. S. Lander, B. Birren, L. Korniszewski, E. Back, and G. Scherer. Campomelic dysplasia translocation breakpoints are scattered over 1 mb proximal to sox9: evidence for an extended control region. *Am J Hum Genet*, 65(1):111–124, Jul 1999. doi: 10.1086/302455. URL `http://dx.doi.org/10.1086/302455`.

[260] A. A. Philippakis, F. S. He, and M. L. Bulyk. Modulefinder: a tool for computational discovery of cis regulatory modules. *Pac Symp Biocomput*, pages 519–30, 2005.

[261] E. M. Phizicky and S. Fields. Protein-protein interactions: methods for detection and analysis. *Microbiol Rev*, 59(1):94–123, Mar 1995.

[262] Y. Pilpel, P. Sudarsanam, and G. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat Genet*, 29(2):153–9, Oct 2001. doi: 10.1038/ng724. URL `http://dx.doi.org/10.1038/ng724`.

[263] L. Ponger and D. Mouchiroud. Cpgprod: identifying cpg islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics*, 18 (4):631–633, Apr 2002.

[264] E. Portales-Casamar, S. Kirov, J. Lim, S. Lithwick, M. I. Swanson, A. Ticoll, J. Snoddy, and W. W. Wasserman. Pazar: a framework for collection and dissemination of cis-regulatory sequence annotation. *Genome Biol*, 8(10):R207, 2007. doi: 10.1186/gb-2007-8-10-r207. URL `http://dx.doi.org/10.1186/gb-2007-8-10-r207`.

[265] T. Pramila, S. Miles, D. GuhaThakurta, D. Jemiolo, and L. L. Breeden. Conserved homeodomain proteins interact with mads box protein mcm1 to restrict ecb-dependent transcription to the m/g1 phase of the cell cycle. *Genes Dev*, 16(23):3034–3045, Dec 2002. doi: 10.1101/gad.1034302. URL `http://dx.doi.org/10.1101/gad.1034302`.

[266] D. S. Prestridge. Computer software for eukaryotic promoter analysis. *Methods Mol Biol*, 130:265–295, 2000.

[267] Y. Qi, Z. Bar-Joseph, and J. Klein-Seetharaman. Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, 63(3):490–500, May 2006. doi: 10.1002/prot.20865. URL `http://dx.doi.org/10.1002/prot.20865`.

[268] J. Qian, J. Lin, N. M. Luscombe, H. Yu, and M. Gerstein. Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics*, 19(15):1917–26, Oct 2003.

[269] C. Queen, M. N. Wegman, and L. J. Korn. Improvements to a program for dna analysis: a procedure to find homologies among many sequences. *Nucleic Acids Res*, 10(1):449–456, Jan 1982.

[270] S. Rahmann, T. Mueller, and M. Vingron. On the power of profiles for transcription factor binding site detection. *Statistical Applications in Genetics and Molecular Biology*, 2(1):Art. 7, 2003. URL `http://www.bepress.com/sagmb/vol2/iss1/art7/`.

[271] N. Rajewsky, M. Vergassola, U. Gaul, and E. D. Siggia. Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. *BMC Bioinformatics*, 3(1):30, Oct 2002. URL `http://www.biomedcentral.com/1471-2105/3/30`.

[272] K. Rateitschak, T. Mueller, and M. Vingron. Annotating significant pairs of transcription factor binding sites in regulatory DNA. *In Silico Biol*, 4(3):0040, 8 2004. URL `http://www.bioinfo.de/isb/2004/04/0040/`.

[273] M. G. Reese. Application of a time-delay neural network to promoter annotation in the drosophila melanogaster genome. *Comput Chem*, 26(1):51–56, Dec 2001.

[274] W. Reith, C. Ucla, E. Barras, A. Gaud, B. Durand, C. Herrero-Sanchez, M. Kobr, and B. Mach. Rfx1, a transactivator of hepatitis b virus enhancer i, belongs to a novel family of homodimeric and heterodimeric dna-binding proteins. *Mol Cell Biol*, 14(2): 1230–1244, Feb 1994.

[275] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of dna binding proteins. *Science*, 290(5500):2306–2309, Dec 2000. doi: 10.1126/science.290.5500.2306. URL `http://dx.doi.org/10.1126/science.290.5500.2306`.

[276] G. Rigaut, A. Shevchenko, B. Rutz, M. Wilm, M. Mann, and B. Séraphin. A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*, 17(10):1030–1032, Oct 1999. doi: 10.1038/13732. URL `http://dx.doi.org/10.1038/13732`.

[277] G. Robertson, M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones. Genome-wide profiles of stat1 dna association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*, 4(8): 651–657, Aug 2007. doi: 10.1038/nmeth1068. URL `http://dx.doi.org/10.1038/nmeth1068`.

[278] K. A. Robinson, J. I. Koepke, M. Kharodawala, and J. M. Lopes. A network of yeast basic helix-loop-helix interactions. *Nucleic Acids Res*, 28(22):4460–4466, Nov 2000.

[279] J. C. Rodríguez, J. A. Ortiz, F. G. Hegardt, and D. Haro. The hepatocyte nuclear factor 4 (hnf-4) represses the mitochondrial hmg-coa synthase gene. *Biochem Biophys Res Commun*, 242(3):692–696, Jan 1998.

[280] H. G. Roider, B. Lenhard, A. Kanhere, S. A. Haas, and M. Vingron. Cpg-depleted promoters harbor tissue-specific transcription factor binding signals–implications for motif overrepresentation analyses. *Nucleic Acids Res*, Sep 2009. doi: 10.1093/nar/gkp682. URL `http://dx.doi.org/10.1093/nar/gkp682`.

[281] J. W. Rooney, Y. L. Sun, L. H. Glimcher, and T. Hoey. Novel nfat sites that mediate activation of the interleukin-2 promoter in response to t-cell receptor stimulation. *Mol Cell Biol*, 15(11):6299–6310, Nov 1995.

[282] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak,

R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, Oct 2005. doi: 10.1038/nature04209. URL `http://dx.doi.org/10.1038/nature04209`.

[283] R. P. Ryseck and R. Bravo. c-jun, jun b, and jun d differ in their binding affinities to ap-1 and cre consensus sequences: effect of fos proteins. *Oncogene*, 6(4):533–542, Apr 1991.

[284] H. Salgado, S. Gama-Castro, M. Peralta-Gil, E. Díaz-Peredo, F. Sánchez-Solano, A. Santos-Zavaleta, I. Martínez-Flores, V. Jiménez-Jacinto, C. Bonavides-Martínez, J. Segura-Salazar, A. Martínez-Antonio, and J. Collado-Vides. Regulondb (version 5.0): Escherichia coli k-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res*, 34(Database issue):D394–D397, Jan 2006. doi: 10.1093/nar/gkj156. URL `http://dx.doi.org/10.1093/nar/gkj156`.

[285] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Res*, 32(Database issue):D449–D451, Jan 2004. doi: 10.1093/nar/gkh086. URL `http://dx.doi.org/10.1093/nar/gkh086`.

[286] A. Sandelin and W. W. Wasserman. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J Mol Biol*, 338(2): 207–215, Apr 2004. doi: 10.1016/j.jmb.2004.02.048. URL `http://dx.doi.org/10.1016/j.jmb.2004.02.048`.

[287] S. Saxonov, P. Berg, and D. L. Brutlag. A genome-wide analysis of cpg dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A*, 103(5):1412–1417, Jan 2006. doi: 10.1073/pnas.0510310103. URL `http://dx.doi.org/10.1073/pnas.0510310103`.

[288] E. Scarano, M. Iaccarino, P. Grippo, and E. Parisi. The heterogeneity of thymine methyl group origin in dna pyrimidine isostichs of developing sea urchin embryos. *Proc Natl Acad Sci U S A*, 57(5):1394–1400, May 1967.

[289] G. Schaefer. Weighted matchings in general graphs. Master's thesis, Universität des Saarlandes, 2000.

[290] M. Scherf, A. Klingenhoff, and T. Werner. Highly specific localization of promoter regions in large genomic sequences by promoterinspector: a novel context analysis approach. *J Mol Biol*, 297(3):599–606, Mar 2000. doi: 10.1006/jmbi.2000.3589. URL `http://dx.doi.org/10.1006/jmbi.2000.3589`.

[291] C. D. Schmid and P. Bucher. Chip-seq data reveal nucleosome architecture of human promoters. *Cell*, 131(5):831–2; author reply 832–3, Nov 2007. doi: 10.1016/j.cell.2007.11.017. URL `http://dx.doi.org/10.1016/j.cell.2007.11.017`.

[292] C. D. Schmid, R. Perier, V. Praz, and P. Bucher. Epd in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Res*, 34(Database issue):D82–D85, Jan 2006. doi: 10.1093/nar/gkj146. URL `http://dx.doi.org/10.1093/nar/gkj146`.

[293] T. Schneider and R. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18(20):6097–100, Oct 1990.

[294] T. D. Schneider. Information content of individual genetic sequences. *J Theor Biol*, 189(4):427–441, Dec 1997. doi: 10.1006/jtbi.1997.0540. URL `http://dx.doi.org/10.1006/jtbi.1997.0540`.

[295] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J.-P. Z. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 442 (7104):772–778, Aug 2006. doi: 10.1038/nature04979. URL `http://dx.doi.org/10.1038/nature04979`.

[296] R. Sekido, I. Bar, V. Narváez, G. Penny, and R. Lovell-Badge. Sox9 is up-regulated by the transient expression of sry specifically in sertoli cell precursors. *Dev Biol*, 274 (2):271–279, Oct 2004. doi: 10.1016/j.ydbio.2004.07.011. URL `http://dx.doi.org/10.1016/j.ydbio.2004.07.011`.

[297] R. Sharan, A. Ben-Hur, G. G. Loots, and I. Ovcharenko. CREME: Cis-Regulatory Module Explorer for the human genome. *Nucleic Acids Res*, 32(Web Server issue): W253–6, Jul 2004. doi: 10.1093/nar/gkh385. URL `http://dx.doi.org/10.1093/nar/gkh385`.

[298] J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang. Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci U S A*, 104(11):4337–4341, Mar 2007. doi: 10.1073/pnas.0607879104. URL `http://dx.doi.org/10.1073/pnas.0607879104`.

[299] J. Shendure and H. Ji. Next-generation dna sequencing. *Nat Biotechnol*, 26(10):1135–1145, Oct 2008. doi: 10.1038/nbt1486. URL `http://dx.doi.org/10.1038/nbt1486`.

[300] J. Shendure, G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra, and G. M. Church. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309(5741):1728–1732, Sep 2005. doi: 10.1126/science.1117389. URL `http://dx.doi.org/10.1126/science.1117389`.

[301] T. Shiraki, S. Kondo, S. Katayama, K. Waki, T. Kasukawa, H. Kawaji, R. Kodzius, A. Watahiki, M. Nakamura, T. Arakawa, S. Fukuda, D. Sasaki, A. Podhajska, M. Harbers, J. Kawai, P. Carninci, and Y. Hayashizaki. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A*, 100(26):15776–15781, Dec 2003. doi: 10.1073/pnas.2136655100. URL `http://dx.doi.org/10.1073/pnas.2136655100`.

[302] B. A. Shoemaker and A. R. Panchenko. Deciphering protein-protein interactions. part ii. computational methods to predict protein and domain interaction partners. *PLoS Comput Biol*, 3(4):e43, Apr 2007. doi: 10.1371/journal.pcbi.0030043. URL `http://dx.doi.org/10.1371/journal.pcbi.0030043`.

[303] H. Siemen, M. Nix, E. Endl, P. Koch, J. Itskovitz-Eldor, and O. Brüstle. Nucleofection of human embryonic stem cells. *Stem Cells Dev*, 14(4):378–383, Aug 2005. doi: 10.1089/scd.2005.14.378. URL `http://dx.doi.org/10.1089/scd.2005.14.378`.

[304] P. Simpson. Evolution of development in closely related species of flies and worms. *Nat Rev Genet*, 3(12):907–917, Dec 2002. doi: 10.1038/nrg947. URL `http://dx.doi.org/10.1038/nrg947`.

[305] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):3940–3941, Oct 2005. doi: 10.1093/bioinformatics/bti623. URL `http://dx.doi.org/10.1093/bioinformatics/bti623`.

[306] S. Sinha, E. van Nimwegen, and E. D. Siggia. A probabilistic method to detect regulatory modules. *Bioinformatics*, 19 Suppl 1:i292–301, 2003.

[307] S. Sinha, M. D. Schroeder, U. Unnerstall, U. Gaul, and E. D. Siggia. Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in drosophila. *BMC Bioinformatics*, 5:129, Sep 2004. doi: 10.1186/1471-2105-5-129. URL `http://dx.doi.org/10.1186/1471-2105-5-129`.

[308] L. Skrabanek, H. K. Saini, G. D. Bader, and A. J. Enright. Computational prediction of protein-protein interactions. *Mol Biotechnol*, 38(1):1–17, Jan 2008. doi: 10.1007/s12033-007-0069-2. URL `http://dx.doi.org/10.1007/s12033-007-0069-2`.

[309] S. T. Smale and J. T. Kadonaga. The rna polymerase ii core promoter. *Annu Rev Biochem*, 72:449–479, 2003. doi: 10.1146/annurev.biochem.72.121801.161520. URL `http://dx.doi.org/10.1146/annurev.biochem.72.121801.161520`.

[310] A. Smit, R. Hubley, and P. Green. Repeatmasker open-3.0. http://www.repeatmasker.org, 1996-2004.

[311] G. P. Smith. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*, 228(4705):1315–1317, Jun 1985.

[312] E. S. Snitkin, A. M. Gustafson, J. Mellor, J. Wu, and C. DeLisi. Comparative assessment of performance and genome dependence among phylogenetic profiling methods. *BMC Bioinformatics*, 7:420, 2006. doi: 471-2105-7-420. URL `http://dx.doi.org/471-2105-7-420`.

[313] M. J. Solomon, P. L. Larsen, and A. Varshavsky. Mapping protein-dna interactions in vivo with formaldehyde: evidence that histone h4 is retained on a highly transcribed gene. *Cell*, 53(6):937–947, Jun 1988.

[314] V. V. Solovyev and I. A. Shahmuradov. Promh: Promoters identification using orthologous genomic sequences. *Nucleic Acids Res*, 31(13):3540–3545, Jul 2003.

[315] J.-Y. Springael and M. J. Penninckx. Nitrogen-source regulation of yeast gamma-glutamyl transpeptidase synthesis involves the regulatory network including the gata zinc-finger factors gln3, nil1/gat1 and gzf3. *Biochem J*, 371(Pt 2):589–595, Apr 2003. doi: 10.1042/BJ20021893. URL `http://dx.doi.org/10.1042/BJ20021893`.

[316] E. Sprinzak and H. Margalit. Correlated sequence-signatures as markers of protein-protein interaction. *J Mol Biol*, 311(4):681–692, Aug 2001. doi: 10.1006/jmbi.2001.4920. URL `http://dx.doi.org/10.1006/jmbi.2001.4920`.

[317] E. Sprinzak, Y. Altuvia, and H. Margalit. Characterization and prediction of protein-protein interactions within and between complexes. *Proc Natl Acad Sci U S A*, 103 (40):14718–14723, Oct 2006. doi: 10.1073/pnas.0603352103. URL `http://dx.doi.org/10.1073/pnas.0603352103`.

[318] M. Srivastava, Y. Torosyan, M. Raffeld, O. Eidelman, H. B. Pollard, and L. Bubendorf. Anxa7 expression represents hormone-relevant tumor suppression in different cancers. *Int J Cancer*, 121(12):2628–2636, Dec 2007. doi: 10.1002/ijc.23008. URL `http://dx.doi.org/10.1002/ijc.23008`.

[319] R. Staden. Methods for calculating the probabilities of finding patterns in sequences. *Comput Appl Biosci*, 5(2):89–96, Apr 1989.

[320] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksöz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. E. Wanker. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–968, Sep 2005. doi: 10.1016/j.cell.2005.08.029. URL `http://dx.doi.org/10.1016/j.cell.2005.08.029`.

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

[322] G. D. Stormo and D. S. Fields. Specificity, free energy and information content in protein-dna interactions. *Trends Biochem Sci*, 23(3):109–113, Mar 1998.

[323] W. M. Strauss. Transfection of mammalian cells via lipofection. *Methods Mol Biol*, 54:307–327, 1996.

[324] K. Struhl. Molecular mechanisms of transcriptional regulation in yeast. *Annu Rev Biochem*, 58:1051–1077, 1989. doi: 10.1146/annurev.bi.58.070189.005155. URL `http://dx.doi.org/10.1146/annurev.bi.58.070189.005155`.

[325] P. Sudarsanam, Y. Pilpel, and G. M. Church. Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in Saccharomyces cerevisiae. *Genome Res*, 12(11):1723–31, Nov 2002. doi: 10.1101/gr.301202. URL `http://dx.doi.org/10.1101/gr.301202`.

[326] M. Sultan, M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S. O'Keeffe, S. Haas, M. Vingron, H. Lehrach, and M.-L. Yaspo. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321(5891):956–960, Aug 2008. doi: 10.1126/science.1160342. URL `http://dx.doi.org/10.1126/science.1160342`.

[327] Y. Suzuki, T. Tsunoda, J. Sese, H. Taira, J. Mizushima-Sugano, H. Hata, T. Ota, T. Isogai, T. Tanaka, Y. Nakamura, A. Suyama, Y. Sakaki, S. Morishita, K. Okubo, and S. Sugano. Identification and characterization of the potential promoter regions of 1031 kinds of human genes. *Genome Res*, 11(5):677–684, May 2001. doi: 10.1101/gr.164001. URL `http://dx.doi.org/10.1101/gr.164001`.

[328] Y. Tabach, R. Brosh, Y. Buganim, A. Reiner, O. Zuk, A. Yitzhaky, M. Koudritsky, V. Rotter, and E. Domany. Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site. *PLoS One*, 2(8):e807, 2007. doi: 10.1371/journal.pone.0000807. URL `http://dx.doi.org/10.1371/journal.pone.0000807`.

[329] R. L. Tatusov, S. F. Altschul, and E. V. Koonin. Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc Natl Acad Sci U S A*, 91(25):12091–12095, Dec 1994.

[330] S. A. Teichmann and M. M. Babu. Conservation of gene co-regulation in prokaryotes and eukaryotes. *Trends Biotechnol*, 20(10):407–10; discussion 410, Oct 2002.

[331] S. A. Teichmann and M. M. Babu. Gene regulatory network growth by duplication. *Nat Genet*, 36(5):492–496, May 2004. doi: 10.1038/ng1340. URL `http://dx.doi.org/10.1038/ng1340`.

[332] J. D. Thompson, D. G. Higgins, and T. J. Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–4680, Nov 1994.

[333] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. D. Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Régnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23(1):137–144, Jan 2005. doi: 10.1038/nbt1053. URL `http://dx.doi.org/10.1038/nbt1053`.

[334] F. Tronche, F. Ringeisen, M. Blumenfeld, M. Yaniv, and M. Pontoglio. Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. *J Mol Biol*, 266(2):231–245, Feb 1997. doi: 10.1006/jmbi.1996.0760. URL `http://dx.doi.org/10.1006/jmbi.1996.0760`.

[335] A. M. Tsankov, C. R. Brown, M. C. Yu, M. Z. Win, P. A. Silver, and J. M. Casolari. Communication between levels of transcriptional control improves robustness and adaptivity. *Mol Syst Biol*, 2:65, 2006. doi: 10.1038/msb4100106. URL `http://dx.doi.org/10.1038/msb4100106`.

[336] R. Y. Tsien. The green fluorescent protein. *Annu Rev Biochem*, 67:509–544, 1998. doi: 10.1146/annurev.biochem.67.1.509. URL `http://dx.doi.org/10.1146/annurev.biochem.67.1.509`.

[337] P. A. Tsonis. *Anatomy Of Gene Regulation: A Three-dimensional Structural Analysis*. Cambridge University Press, 2003.

[338] S. C. Tucker and R. Wisdom. Site-specific heterodimerization by paired class homeodomain proteins mediates selective transcriptional responses. *J Biol Chem*, 274(45):32325–32332, Nov 1999.

[339] C. Tuerk and L. Gold. Systematic evolution of ligands by exponential enrichment: Rna ligands to bacteriophage t4 dna polymerase. *Science*, 249(4968):505–510, Aug 1990.

[340] J.-V. Turatsinze, M. Thomas-Chollier, M. Defrance, and J. van Helden. Using rsat to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat Protoc*, 3(10):1578–1588, 2008. doi: 10.1038/nprot.2008.97. URL `http://dx.doi.org/10.1038/nprot.2008.97`.

[341] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770):623–627, Feb 2000. doi: 10.1038/35001009. URL `http://dx.doi.org/10.1038/35001009`.

[342] J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet*, 10(4):252–263, Apr 2009. doi: 10.1038/nrg2538. URL `http://dx.doi.org/10.1038/nrg2538`.

[343] S. Vardhanabhuti, J. Wang, and S. Hannenhalli. Position and distance specificity are important determinants of cis-regulatory motifs in addition to evolutionary conservation. *Nucleic Acids Res*, 35(10):3203–3213, 2007. doi: 10.1093/nar/gkm201. URL `http://dx.doi.org/10.1093/nar/gkm201`.

[344] A. K. Vershon. Protein interactions of homeodomain proteins. *Curr Opin Biotechnol*, 7(4):392–396, Aug 1996.

[345] A. Visel, S. Minovitsky, I. Dubchak, and L. A. Pennacchio. VISTA Enhancer Browser–a database of tissue-specific human enhancers. *Nucleic Acids Res*, 35(Database issue): D88–92, Jan 2007. doi: 10.1093/nar/gkl822. URL `http://dx.doi.org/10.1093/nar/gkl822`.

[346] D. Vlieghe, A. Sandelin, P. J. D. Bleser, K. Vleminckx, W. W. Wasserman, F. van Roy, and B. Lenhard. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles. *Nucleic Acids Res*, 34(Database issue):D95–7, Jan 2006. doi: 10.1093/nar/gkj115. URL `http://dx.doi.org/10.1093/nar/gkj115`.

[347] C. von Mering, L. J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Krüger, B. Snel, and P. Bork. String 7–recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res*, 35(Database issue):D358–D362, Jan 2007. doi: 10.1093/nar/gkl825. URL `http://dx.doi.org/10.1093/nar/gkl825`.

[348] A. Wagner. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, 15(10):776–84, Oct 1999. - statistical technique to detect homotypic/heterotypic cooperativity - shown for two TF in yeast - goal: detection of "best candidate genes" for regulation by one or more TFs -statistical model: null hypothesis: distribution of TFBS(i) Poisson distributed some problems: * short sites, nucleotide ambiguities, repetitive structure will not follow Poisson distribution even in random DNA - test if TFBS distribution is consistent with Poisson distribution on pseudorandom DNA * assumption of a constant factor lambda for Poission over whole genome unrealistic solution here: incorporation of global and local DNA composition into statistical analysis [plot of probability of observing a TFBS based on dinucleotide composition within 5kb windows].

[349] H. Wakaguri, R. Yamashita, Y. Suzuki, S. Sugano, and K. Nakai. DBTSS: database of transcription start sites, progress report 2008. *Nucleic Acids Res*, 36(Database issue): D97–101, Jan 2008. doi: 10.1093/nar/gkm901. URL `http://dx.doi.org/10.1093/nar/gkm901`.

[350] W. Wasserman and J. Fickett. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol*, 278(1):167–81, Apr 1998.

[351] W. Wasserman, M. Palumbo, W. Thompson, J. Fickett, and C. Lawrence. Human-mouse genome comparisons to locate regulatory sites. *Nat Genet*, 26(2):225–8, Oct 2000. doi: 10.1038/79965. URL `http://dx.doi.org/10.1038/79965`.

[352] W. W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet*, 5(4):276–87, Apr 2004. doi: 10.1038/nrg1315. URL `http://dx.doi.org/10.1038/nrg1315`. futility theorem tp:fp - 1:1000.

[353] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, Apr 1953.

[354] A. S. Weinmann and P. J. Farnham. Identification of unknown target genes of human transcription factors using chromatin immunoprecipitation. *Methods*, 26(1):37–47, Jan 2002. doi: 10.1016/S1046-2023(02)00006-3. URL `http://dx.doi.org/10.1016/S1046-2023(02)00006-3`.

[355] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, J. Ostell, K. D. Pruitt, G. D. Schuler, M. Shumway, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 36(Database issue):D13–D21, Jan 2008. doi: 10.1093/nar/gkm1000. URL `http://dx.doi.org/10.1093/nar/gkm1000`.

[356] K. P. White, S. A. Rifkin, P. Hurban, and D. S. Hogness. Microarray analysis of drosophila development during metamorphosis. *Science*, 286(5447):2179–2184, Dec 1999.

[357] R. M. Williams, M. Primig, B. K. Washburn, E. A. Winzeler, M. Bellis, C. S. de Menthiere, R. W. Davis, and R. E. Esposito. The ume6 regulon coordinates metabolic and meiotic gene expression in yeast. *Proc Natl Acad Sci U S A*, 99(21):13431–13436, Oct 2002. doi: 10.1073/pnas.202495299. URL `http://dx.doi.org/10.1073/pnas.202495299`.

[358] C. Wolberger. Homeodomain interactions. *Curr Opin Struct Biol*, 6(1):62–68, Feb 1996.

[359] W. S. W. Wong and R. Nielsen. Finding cis-regulatory modules in drosophila using phylogenetic hidden markov models. *Bioinformatics*, 23(16):2031–2037, Aug 2007. doi: 10.1093/bioinformatics/btm299. URL `http://dx.doi.org/10.1093/bioinformatics/btm299`.

[360] C. Woodbury and P. von Hippel. On the determination of deoxyribonucleic acid-protein interaction parameters using the nitrocellulose filter-binding assay. *Biochemistry*, 22(20):4730–7, Sep 1983.

[361] G. A. Wray, M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman, and L. A. Romano. The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol*, 20(9):1377–419, Sep 2003. doi: 10.1093/molbev/msg140. URL `http://dx.doi.org/10.1093/molbev/msg140`.

[362] W. E. Wright, M. Binder, and W. Funk. Cyclic amplification and selection of targets (casting) for the myogenin consensus binding site. *Mol Cell Biol*, 11(8):4104–4110, Aug 1991.

[363] C. Wu. The 5' ends of drosophila heat shock genes in chromatin are hypersensitive to dnase i. *Nature*, 286(5776):854–860, Aug 1980.

[364] T. D. Wu, C. G. Nevill-Manning, and D. L. Brutlag. Fast probabilistic analysis of sequence function using scoring matrices. *Bioinformatics*, 16(3):233–244, Mar 2000.

[365] X. Xie, J. Lu, E. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, and M. Kellis. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, Feb 2005. doi: 10.1038/nature03441. URL `http://dx.doi.org/10.1038/nature03441`.

[366] J. Yoo, L. E. Herman, C. Li, S. B. Krantz, and D. Tuan. Dynamic changes in the locus control region of erythroid progenitor cells demonstrated by polymerase chain reaction. *Blood*, 87(6):2558–2567, Mar 1996.

[367] H. Yu, X. Zhu, D. Greenbaum, J. Karro, and M. Gerstein. Topnet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. *Nucleic Acids Res*, 32(1):328–337, 2004. doi: 10.1093/nar/gkh164. URL `http://dx.doi.org/10.1093/nar/gkh164`.

[368] X. Yu, J. Lin, T. Masuda, N. Esumi, D. J. Zack, and J. Qian. Genome-wide prediction and characterization of interactions between transcription factors in Saccharomyces cerevisiae. *Nucleic Acids Res*, 34(3):917–27, 2006. doi: 10.1093/nar/gkj487. URL `http://dx.doi.org/10.1093/nar/gkj487`.

[369] X. Yu, J. Lin, D. J. Zack, and J. Qian. Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic Acids Res*, 34(17):4925–36, 2006. doi: 10.1093/nar/gkl595. URL `http://dx.doi.org/10.1093/nar/gkl595`.

[370] X. Yu, J. Lin, D. J. Zack, and J. Qian. Identification of tissue-specific cis-regulatory modules based on interactions between transcription factors. *BMC Bioinformatics*, 8:437, 2007. doi: 10.1186/1471-2105-8-437. URL `http://dx.doi.org/10.1186/1471-2105-8-437`.

[371] G.-C. Yuan, Y.-J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler, and O. J. Rando. Genome-scale identification of nucleosome positions in s. cerevisiae. *Science*, 309(5734):626–630, Jul 2005. doi: 10.1126/science.1112178. URL `http://dx.doi.org/10.1126/science.1112178`.

[372] C. H. Yuh, H. Bolouri, and E. H. Davidson. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science*, 279(5358):1896–1902, Mar 1998.

[373] Q. Zhou and W. H. Wong. CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc Natl Acad Sci U S A*, 101(33):12114–9, Aug 2004. doi: 10.1073/pnas.0402858101. URL `http://dx.doi.org/10.1073/pnas.0402858101`.

[374] J. Zhu and M. Q. Zhang. Scpd: a promoter database of the yeast saccharomyces cerevisiae. *Bioinformatics*, 15(7-8):607–611, 1999.

[375] Z. Zhu, J. Shendure, and G. M. Church. Discovering functional transcription-factor combinations in the human cell cycle. *Genome Res*, 15(6):848–55, Jun 2005. doi: 10.1101/gr.3394405. URL `http://dx.doi.org/10.1101/gr.3394405`.