Computational Methods for inferring Transcription Factor Binding Sites

Vyacheslav Morozov

Thesis submitted to the Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of Master of Science in Mathematics [1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

[1] The M.Sc. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Abstract

Position weight matrices (PWMs) have become a tool of choice for the identification of transcription factor binding sites in DNA sequences. PWMs are compiled from experimentally verified and aligned binding sequences. PWMs are then used to computationally discover novel putative binding sites for a given protein. DNA-binding proteins often show degeneracy in their binding requirement, the overall binding specificity of many proteins is unknown and remains an active area of research. Although PWMs are more reliable predictors than consensus string matching, they generally result in a high number of false positive hits. A previous study introduced a novel method to PWM training based on the known motifs to sample additional putative binding sites from a proximal promoter area. The core idea was further developed, implemented and tested in this thesis with a large scale application. Improved mono- and dinucleotide PWMs were computed for *Drosophila melanogaster*. The Matthews correlation coefficient was used as an optimization criterion in the PWM refinement algorithm. New PWMs keep an account of non-uniform background nucleotide distributions on the promoters and consider a larger number of new binding sites during the refinement steps. The optimization included the PWM motif length, the position on the promoter, the threshold value and the binding site location. The obtained predictions were compared for mono- and dinucleotide PWM versions with initial matrices and with conventional tools. The optimized PWMs predicted new binding sites with better accuracy than conventional PWMs.

Acknowledgements

I wish to thank my thesis advisors Prof. Ilya Ioshikhes and Prof. Stéphane Aris-Brosou for their guidance and constructive feedback during the duration of my project. I am particularly grateful to Prof. Stéphane Aris-Brosou also for the cost of his extra time for his thorough reading of my thesis, marking and comments which greatly enhanced the quality of the final document.

I am thankful to have been part of Prof. Ilya Ioshikhes's laboratory, and would like to acknowledge the lab members for their support during my studies.

The research described in this thesis was partly funded Natural Sciences and Engineering Research Council of Canada (NCERC, grant number RGPIN/ 372240-2009) and Canada Foundation for Innovation Leaders Opportunity Fund / Ontario Research Fund (grant number 22880).

Contents

Chapter 1

Introduction

1.1 Eukaryotic transcription factors in gene regulation

This thesis focuses on studying the biological phenomenon of transcriptional regulation using methods of computational biology. The overall objective is to improve upon the existing computational methods for identifying novel transcription factor binding sites. A key result is the demonstration that significant improvement of computational prediction is possible without resorting to additional cost and labor intensive experiments and thereby presenting an alternative to molecular modeling.

During transcription, a gene, (segment of DNA molecule that encodes a functional product) is transcribed into RNA which, in the case of a protein-coding gene, is then translated into a protein product that performs a specific function in the cell. Gene expression is a process by which information from a gene is used in the synthesis of a functional gene product.

Transcription is a directional process in which the DNA code is read from the 5' end to the 3' end of a single DNA strand. Eukaryotic organisms have two stranded DNA molecules which can organize loosely-packed protein-DNA formations named

chromatin. A tighter packing of chromatin in cell nucleus leads to the formation of individual chromosomes. Different species can have different haploid chromosome numbers. This means, for instance, that *Drosophila melanogaster* (further *Drosophila*) has five, while *Homo sapiens* has 23 distinct chromosomes.

Since transcription is a directional process, always occurring from 5' to 3', the term "upstream" refers to sequences in the 5' direction of a region of interest on a single DNA strand, while the term "downstream" refers to the 3' direction.

Numerous proteins have been shown to participate in the regulation of gene expression and these proteins are termed transcription factors (TFs). In eukaryotes, TFs recognize target nucleotide sequences on DNA and bind their segments typically upstream of target genes. When a TF binds to the DNA, it then recruits or blocks an RNA polymerase (specified in the next paragraph) to effect initiation or prevention of transcription of target genes, respectively. A defining feature of TFs as protein molecules is that they contain one or more DNA-binding domains, which attach to specific sequences of DNA. TFs are biologically specific for genes they regulate. This means that each of them binds certain short regions of DNA responsible for certain gene regulation that are termed transcription factor binding site (TFBS). Defined DNA-binding domains in TFs have, as measured experimentally, up to 10^6 fold higher affinity for their target sequences (TFBSs) than for the remainder of the DNA strand [86]. The next two paragraphs give a closer look onto structural component involved in DNA transcription.

The RNA polymerase is a huge factory with many moving parts. Although three types of RNA polymerase exist in eukaryotes, here we focus on TFs that act with polymerase II. An RNA polymerase is a protein complex (enzyme) which forms a machine that surrounds DNA strands, unwinds them, and builds an RNA strand based on the information held inside the DNA. Once transcription gets started, RNA polymerase marches from 5' to 3' of the DNA, copying RNA strands thousands of nucleotides long. In case of polymerase II, the product of the DNA transcription is

an mRNA (messenger RNA), which is a versatile one-stranded molecule. In its most familiar role, mRNA acts as an intermediary, carrying genetic information from the DNA to the machinery of protein synthesis.

Transcription factors are often classified based on the sequence similarity and hence the tertiary structure of their DNA-binding domains [71]. Each TF as a molecule has a characteristic three-dimensional protein structure, characterized by structural motifs. Each structural motif is a relatively small portion of the protein organized into a certain geometric form. Structural motifs are used to define specific classes of TFs. Among specific examples of TF proteins with defined motifs are (Figure 1.1): zinc finger proteins, steroid receptors, leucine zippers, lambda-repressor proteins, etc. [112]. TF-DNA binding can be seen as a stochastic process in which a TF can have multiple sequence specificities and bind a number of different recognition sequences with variable binding affinities. Intrinsic sequence-recognition property of any TF implies that a TF is able to recognize at least one smaller oligonucleotide sub-sequence within a corresponding gene regulatory region. The sub-sequence is not necessarily contiguous.

A TF can be involved in gene regulation in two distinct ways: either through cis-acting elements or through trans-acting elements. In the context of transcriptional regulation, a trans-acting element is usually a DNA sequence that contains a gene. This gene codes for a diffusible molecule such as a TF protein which is then used in the regulation of another target gene that can be located anywhere in the genome [117]. In contrast, cis-acting elements take part in gene control without encoding a diffusible molecule. Cis-acting elements are termed as such because they are located on the same DNA molecule with the target gene they regulate. Trans-acting elements can control genes on distant DNA molecules (for instance, on another chromosome) or on the same molecule always by means of a diffusible molecule such as TF that binds to cis-elements in the vicinity of target genes.

A cis-acting element can be located upstream of the 5' end of the coding sequence

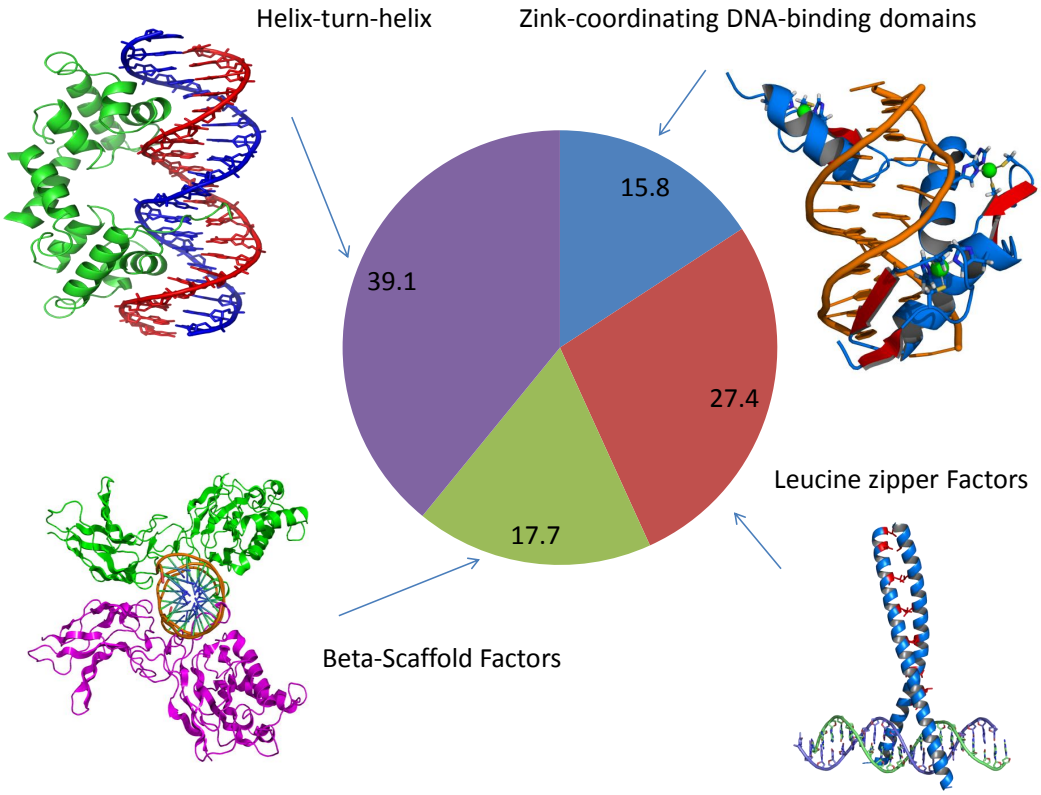

Figure 1.1: Four classes of transcription factors. Transcription factors are generally grouped into four distinct classes. Left-up, helix-turn-helix; right-up, zinc-coordinating DNA-binding domains; left-bottom, beta-scaffold factors; right-bottom, leucine zipper factors. The proportions were calculated based on TRANSFAC (v.7.0) (Adapted with changes from [122])

of a gene it controls (in a so called promoter region or further upstream of the promoter), within the gene, or downstream of the 3' end of the genes coding sequence. Cis-acting elements (such as enhancers) are often located many kilobases away from the basal transcription complex of the gene. Therefore their 3' or 5' orientation with respect to the distant gene is unimportant [25] or may vary for different enhancers (from private conversation with Dr. Ioshikhes). In such a case these elements are able to function because of the flexible geometry of DNA molecules which can bend onto the attached transcription initiation complex.

Cis-acting elements in higher eukaryotes are organized into modular units named cis-regulatory modules (CRMs) of a few hundred baise pairs [8]. A common feature of CRMs is the presence of multipple binding sites. In other words, TFBS in higher eukaryotes have a tendency to be clustered. The term "module" has a functional significance: cis-regulatory modules respond to given regulatory states by producing a unique regulatory output, the result of the individual interactions at their clustered internal target sites. In particular, the cis-modules use TFs as inputs to generate the command given to the transcription machinery, which in turn determines the rate of gene transcription or whether a gene is turned on (promoting transcription as an activator) or off (blocking transcription or acting as a transcriptional repressor) [25]. One cis-regulatory element can regulate several genes [105] and in turn, one gene can have several cis-regulatory modules [25].

Experimental [40] and computational studies [121] demonstrated that sequences of regulatory DNA that bind TFs can exhibit many different types of architecture. CRMs can be homotypic, containing multiple sites for one particular TF or heterotypic, containing one or more binding sites for multiple TFs [114]. The cooperative interactions among transcription factors are very frequent phenomena in eukaryotic transcriptional regulation. Many publications report experimentally verified co-occurrences of TFBS consistent with pairwise co-occurencces of TFs [60]. Bioinformatics studies indicate that antagonistic factors often bind to overlapping sites [65]

whereas synergetic factors are often positioned within a fixed distance [70], often close to the multiple of 10.2 base pairs (bp) of the DNA double-helix pitch value (pitch is the number of bp per complete turn of the helix) [65, 13].

TFs can regulate the expression of genes in several ways. General TFs are those TFs that are necessary for the transcription of all genes transcribed by RNA polymerase II [57]. These TFs are involved in the formation of preinitiation complexes. TFs may also regulate transcription more specifically than shown in Figure 1.2. For instance, they can be tissue-specific and expressed in a defined set of tissues or cell types (so called spatial regulation); TF may be expressed at specific times during development (temporal regulation); TFs may require modification (phosphorylation) prior to performing their function; TFs may be activated by binding with signal triggering molecules, which bind to a site on the target TF (ligand binding); TFs may be activated by cooperation with other factors, by participating in a protein complex.

TFs are thereby controlling the flow of genetic information from DNA to mRNA and perform their functions alone or with other factors (proteins or small molecules) in a complex. Moreover, TFs are likely only one of the means by which our cells express different combinations of genes. Working in this way, they allow for differentiation into various types of cells, tissues, and organs that make up our bodies. TFs not only control transcription to regulate the amount of gene products such as RNA and proteins, but they also regulate the production of themselves by a feedback loop (positive, negative or both, depending on the factor) that controls the transcription of other TFs.

Prediction of TFBS is therefore an important step toward understanding the process of gene regulation. As evidenced from published studies, with small variations in numbers, the lengths of binding sequence motifs ranged between 5 and 20 bp [19, 17, 53]. The small length of TFBS is a challenging factor confounding the discovery of new sites, because TFBS sequences may appear on DNA sequences just by chance. Wasserman showed that typical detection tools will detect a hit every 500 – 5000 bp

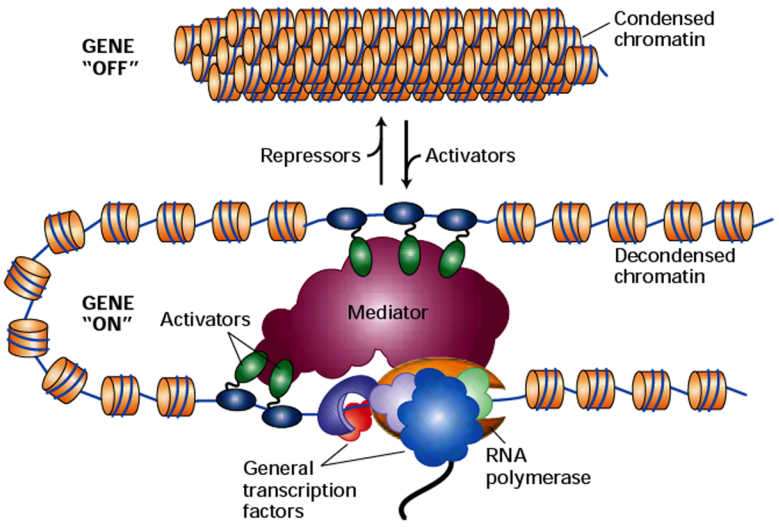

Figure 1.2: Illustration of the Polymerase complex with the transcription factors involved in transcription. The initiation complex, containing the general transcription factors and RNA polymerase II is bound to the promoter. Transcription is simulated by an activator, which binds to a distant enhancer sequence. The activator interacts with the initiation complex and stimulates the rate of transcription. The illustration shows that the activator(s) can interact with the initiation complex via a mediator (also called a co-activator). The activator is bound to an enhancer that is as far as 1000 bp upstream from the promoter. The distant activators are brought into close proximity of the initiation complex by the looping of the DNA. (Adapted from Lodish H. 2000, Molecular cell biology, Freeman, NY with permission).

depending on parameter settings [115]. This obstacle is not that critical for using these tools as predictor tools because most of functional TFBSs are concentrated in specific areas of promoters where the search can be localized.

A proximal promoter is the proximal sequence typically upstream of the gene that tends to contain primary regulatory elements. This area usually comprises from 100 bp downstream to 100 bp upstream and might be up to 250 bp upstream of the transcription start site (TSS). In contrast, a distal promoter is the distal sequence upstream of the gene that may contain additional regulatory elements, often with a weaker influence than the proximal promoter. A core promoter is the minimal portion of a promoter required to properly initiate transcription. The core promoters include the TSSs, where transcription of RNA begins for a particular gene, and other cis-elements upstream such as binding sites for general TFs.

In eukaryotes, promoters are extremely diverse in terms of their nucleotide composition and are therefore difficult to characterize. Typically they are located upstream of the genes they regulate and can have regulatory elements several kilo base pairs away from the TSS. Widely known examples of such cis-elements are enhancers, which can bind TFs and regulate transcription of target genes. In eukaryotes, transcriptional complexes can cause the DNA to bend which allows direct communication between enhancers and basic transcriptional complexes.

There are several types of promoter elements discovered in eukaryotic genomes. Between 10-20% of all genes contain a TATA box, which is typically located in the area between 35th and 25th positions upstream of TSS [35]. The TATA box binds TATA-binding protein (TBP) which is an element of TFIID (Transcription factor II D) - one of the general TFs involved in the formation of the RNA polymerase transcription complex [98]. Other known experimentally confirmed core promoter elements are Initiator (Inr), Downstream Promoter Element (DPE), TFIIB recognition element (BRE), which are known to act in several synergetic combinations [35, 56].

In higher eukaryotic organisms such as mammals, transcription is a more complex

process, so that the binding landscape becomes more diverse. In particular, regulation of transcription in eukaryotes is a result of the combined effects of structural properties such as how DNA is packaged and of TFs-TFs and TFs-DNA interactions. The most important structural difference between eukaryotic and prokaryotic DNA is the formation of chromatin in eukaryotes [83]. Nucleosomes are complexes comprising segments of DNA coiled around histone protein cores as showed in Figure 1.2. Nucleosome packaging plays a role in the regulation of transcription in eukaryotes.

The increasing volume of available genome sequence data gives rise to a mine of information about location of regulatory elements by comparing different genomes. Based on comparative approaches, it was found that functionally constrained sequences are often conserved in evolution, more so than nonfunctional sequences. The validity of this assumption is evident from the success of such methods as phylogenetic footprinting [11].

1.2 Binding sites in comparative genomics

Functional annotation of eukaryotic genomic sequences represents one of the greatest challenges in modern biology. A promising class of computational methods for gene and cis-regulatory element prediction (such as TFBS) is based on comparative sequence analysis. Numerous methods for motif-based TFBS and TF target gene prediction have been developed that utilize sets of orthologous sequences from multiple species.

In many studies, highly conserved noncoding sequences (CNS) in plants and animals have proved to be indicators of many regulatory elements [41, 62]. CRM are generally more conserved than their flanking intergenic regions [10]. In particular, working with *Drosophila* genome, Berman *et al.* showed that CNSs tend to be spatially clustered with conserved spacing between CNSs [8], and such clusters can be used to predict enhancer sequences without prior knowledge of TFBSs [20].

On the other hand, one promising approach takes advantage of the observation that enhancers often contain multiple TFBSs [37].

In fact, the area of comparative genomics is built entirely on methods of comparative sequence analysis that leverage conserved sequences to annotate genomes from different species and discover novel coding and regulatory elements in newly-sequenced genomes. Due to the importance of sequence conservation in genome annotation studies, its accurate quantification is essential for the discovery and characterization of novel functional elements.

The recent refinements of computational techniques for identifying binding sites have generated considerable interest from the field in the development of validation analysis. Evolutionary constrain is used not only to identify sites but also to distinguish real motifs from false positives [12] and to discern potentially functional sites from neutral DNA [94, 55].

In our study we also used comparative genomics of orthologous TFs (elements derived from a common ancestor through a speciation event) as a feasible method of validation of newly discovered putative TFBSs. Noncoding sequences conserved in the evolution of orthologous gene pairs in diverged species, which are also shared by groups of co-expressed genes, are likely to correspond to cis-regulatory modules [20].

1.3 Modeling the binding specificity of eukaryotic transcription factors

Although the structure of a large number of protein-DNA complexes is known, the mechanisms underlying their specific binding to DNA are poorly understood. Studies have shown that basically in terms of binding affinities there is no simple one-to-one correspondence between amino acid residues and nucleotide sequences in protein-DNA complexes [90]. Meanwhile, the majority of available data indicates that protein-DNA

binding should be considered at a single base pair level to be biologically relevant for gene regulatory process [16]. Accordingly, each base pair within the binding site contributes to the protein-DNA contact. Substitutions of a single base pair may dramatically affect the binding affinity between the TF and the DNA [90, 46, 106]. Several studies suggest a long range dependency of the motif pattern of TFs, in which additional protein factors that participate in the transcription complex may influence the TF-TFBS binding [39, 104].

The discovery of the biochemical mechanisms that underlie TF binding site occupancy remain a fundamental challenge of genomic analysis [53]. To identify transcription factor binding sites (TFBS), a variety of laboratory (*in vivo* and *in vitro*) experimental techniques has been developed. These techniques contribute to our knowledge of TF binding specificity [31]. Methods for the *in vitro* detection of DNA binding motifs include ElectorMobility Shift Assay (EMSA) [33], Systematic Evolution of Ligands by EXponential enrichment (SELEX) [111, 87] and protein-binding DNA microarrays [80]. A common genome wide technique for *in vivo* identification of TFBS for a TF of interest is chromatin immunoprecipitation of bound DNA followed by either hybridization (ChIP-chip) or sequencing (ChIP-seq) [82]. Purified or *in vitro* synthesized proteins can be used in Protein Binding Microarray (PBM) experiments that involve hybridizations of the protein to arrays of oligonucleotide probes [7, 80].

Laboratory *in vivo* binding experiments show that known eukaryotic TFs bind to thousands of DNA sites throughout the genome [53, 64, 78]. Many of them show strong binding within cis-regulatory modules and are evolutionarily conserved [64, 14, 53]. In addition, studies also indicate that other regions in the genome bind TFs at a lower affinity and do not appear to be functionally relevant. Moreover, TFs with unrelated biochemical properties can frequently bind to the same genomic regions [77, 53]. Many scientists argued that a great degree of discrepancy between binding patterns predicted by *in vivo* and *in vitro* experiments can be explained by several biochemical mechanisms, such as competitive inhibition of binding at regions with

overlapping binding sites on the DNA or by nucleosomes and chromatin-associated proteins, which modulate DNA accessibility [105, 53, 65, 112].

Stormo G.D. and Zhao Y. [103] reviewed the methods for determining the specificity of protein-DNA interactions and suggested that in a real cell environment, the binding affinity of a TF is not as crucial as its specificity because TF-DNA affinity depends on the TF concentration. Based on biophysical measurements, they suggested that the affinity in cell nucleus is always higher for binding sites that do not display a high affinity *in vitro*. The binding of TFs to specific regions on the DNA (specificity) requires that the TFs be able to distinguish these sites from the vast majority of non-regulatory DNA sequences. Characterizing the binding patterns of TFs requires some prior knowledge of a representative set of binding sites for a TF of interest that can be used to construct models that allow the complete specificity to be estimated.

TF binding sites occur with a high frequency across the genome and all of them may not be detectable as binding sequences *in vivo*. Due to non-specific binding that involves low-affinity interactions between TFs and non-specific regions on the DNA, *in vivo* binding assays (such as ChIP-seq) require the use of a signal threshold to identify true binding sites [21]. This kind of processing always results in a number of false positives. In contrast, methods for computational TFBS predictions that are based on known sequence-based binding models often rely on a collection of experimentally-determined binding sites and aim to generalize the available information in terms of a deducted regulatory code or pattern.

Limited and/or biased information about the specificity of TF-DNA binding results in an experimental bias, which is a serious bottleneck for computational modeling and identification of binding events. Researchers have developed several strategies to overcome biases in TFBS prediction. These include the parallel processing of gene expression and sequencing data [22] and correlations of cis-regulatory modules with nucleosome positioning [106, 118, 48, 72] and TF cooperation [36]. The

role of nucleosome positioning became evident in the determination of functional and non-functional sites. Approximately 75-90% of eukaryotic DNA is organized into nucleosomes [95].

Many mathematical descriptions of gene regulation are performed in the framework of one-dimensional DNA lattice models. These methods are based conceptually on the assumptions of Markovian properties in nucleotide DNA sequences [68, 32] or Biophysical models [49, 75].

An early study of binding specificity in terms of binding motifs variability and their statistical properties was by Berg and von Hippel in 1987. This study argued that statistical mechanics theory could be used to predict TFBS. Other investigations formalized the biophysical properties of TFs and to refine computational methods to discover new TFBSs [27].

Since the Berg and Hippel work, advanced sequencing techniques have emerged in genomics and many studies have focused on understanding the complexity of TF-TFBS binding. Specifically, identifying binding patterns for TFs of interest has become an area of active investigation in bioinformatics, especially given the high volume of next generation sequencing data.

The current information about the regulatory regions in the genome is collected in a variety of databases. Experimentally determined binding sites are compiled in databases such as TRANSFAC, a commercially supported database for gene regulation [71] and an open source JASPAR [89]. These resources are commonly used for training the algorithms for new TFBS prediction [34, 35, 45].

From the 1980s, PWMs have become widespread and useful tools for computational prediction of protein, DNA or RNA sequences. PWMs are a simple yet powerful model for TF-DNA binding specificity. Although researchers still argue that PWMs are a viable approximations of binding specificity of TFs, it is unknown to which extent this is true for each particular TF [103, 94, 27, 16, 6]. Incomplete information regarding the TF-DNA binding interaction remains a challenge from either the

DNA or the protein side. Even for some of the most extensively studied eukaryotic organisms (i.e., human, rat and mouse), only one-sixth of all proteins with annotated DNA-binding domains have been characterized experimentally [94, 2].

Another common name for PWMs is the Position Sensitive Scoring Matrices (PSSMs). The next chapters of this thesis describe statistical and computational aspects of inferring TFBSs using generic PWMs as well as matrix refinement techniques to improve the prediction of TF binding specificity. Then new PWMs are used for the prediction of new binding sites on proximal promoters of the *Drosophila melanogaster* genome.

1.4 Statistical models of binding specificity

Given a set of sequences, $S_1, S_2, ..., S_K$, such as those listed in Table 1.1, the two goals of this section are:

1. to identify the substrate binding segments, called "elements",

2. to estimate the parameters of the product multinomial model that describes the collection of the most similar or aligned elements. We call "motif" a model which generalizes in a certain way sequence information from known cis-acting elements. Such motifs appear to be useful as a descriptive model of binding specificity. A DNA region is a binding "site" if it is described by the motif in terms of the model and a "non-site" otherwise.

One of the first aggregative models for quantitative binding specificity was built 20 years ago on the basis of the consensus sequence, which provides the most common pattern based on statistically most overrepresented nucleotides in a motif [26].

PWM model was introduced by G. Stormo and used to describe and find translation initiation sites in the RNA. A consensus sequence is easily represented as a weight matrix by simply giving equal scores to each allowed nucleotide at each position, but a weight matrix cannot be represented as a consensus sequence without sacrificing

some information. The matrix approach is therefore more general and is recommended [102]. The main advantage of PWMs over the initial attempts based on consensus sequences is that PWMs allow different mismatches from the preferred sequence to have different effects on the binding site predictions. Later Stromo and collaborators extended the PWM approach to the modeling of TFBSs in DNA sequences [42]. In contrast to consensus sequences, PWMs are based on scores to measure the level of similarity between aligned elements and to provide a more sophisticated formalization of possible binding scenarios.

A very intuitive and basic definition of PWMs can be made using only the site independence assumption. If we know the probability of each positionally independent nucleotide s_i in a DNA sequence S, then the probability for a single sequence $S = s_1, ..., s_L$ is

$$P(S) = \prod_{j=1}^{L} P_j(s_j) = \prod_{j=1}^{L} w_j(s_j), \tag{1.4.1}$$

where the probability of each nucleotide base can depend on the position j. This reflects the i.i.d. (independent and identically distributed) assumption, which will be used throughout this thesis unless otherwise specified (for dinucleotide PWM).

Thus, having in possession possible weights $\mathbf{W} = (\boldsymbol{w}_j)_j$ as in (1.4.1) we can define PWMs and estimate the single sequence probability $P(S|\mathbf{W})$ of the sequence being a binding site. In addition, in more complex cases of a given set of sequences (we use the calligraphy font to denote sets) $\mathcal{S} = \{S_1, S_2, ..., S_K\}$ we can answer the question about $\mathbf{W}$ in the form of *a posteriori* probability $P(\mathbf{W}|\mathcal{S})$.

A descriptive model of binding specificity considered together with a PWM gives a tool for computational prediction of new TFBSs and their comprehensive characterization. Before we can define any formal model of binding specificity we need to give several definitions.

Let $\mathbf{u}$ and $\mathbf{v}$ be vectors with components either in $\mathbb{R}$ or in $\mathbb{N}$, such as $\mathbf{u} = (u_1, u_2, u_3, u_4)$ and $\mathbf{v} = (v_1, v_2, v_3, v_4)$, and Γ be a special gamma function. Using the

language of vectors helps us to avoid unnecessary indexes in many cases. Some of the operations are not very common and used exclusively in the thesis to omit indexes in statements where it does not contradict with the common definitions.

$$\begin{aligned} \mathbf{u}+\mathbf{v} &= (u_1+v_1, u_2+v_2, u_3+v_3, u_4+v_4), \\ |\mathbf{v}| &= |v_1|+|v_2|+|v_3|+|v_4|\,, \\ \mathbf{u}/\mathbf{v} &= (u_1/v_1, u_2/v_2, u_3/v_3, u_4/v_4), \\ \log(\mathbf{u}) &= (\log u_1, \log u_2, \log u_3, \log u_4), \\ \mathbf{uv} &= (u_1v_1, u_2v_2, u_3v_3, u_4v_4), \\ \mathbf{u}\cdot\mathbf{v} &= u_1v_1+u_2v_2+u_3v_3+u_4v_4, \end{aligned} \tag{1.4.2}$$

where we assume that u_i and v_i belong to domains of definition of corresponding scalar functions. We can notice from the last two rows here, that we use two kind of multiplications for vectors: by-coordinates and the dot-product.

In general, the sequence data from Table 1.1 can be represented as a list of strings $\mathcal{S}$ or a cell array of symbols, not necessary aligned by position.

$$\mathcal{S} = \{s_{kl}; k = 1, ...K; l = 1, ..., L_k\} \tag{1.4.3}$$

The symbols s_{kl} in this array have as a state space the alphabet of four letters A, T, G, C which represent the four nucleotides (also called nucleotide bases): Adenine (A), Thymine (T), Guanine (G), Cytosine (C). In this thesis we consider the four letter as an ordered set: A, T, G, C, so that the correspondence between nucleotide bases and numbers one to four can be established.

Although the exact location of TFBS(s) is usually unknown, sequences identified by laboratory experiments are assumed to have binding properties. Thus the location of TFBS needs to be determined *in silico*. However, in order to simplify the problem, the length of common binding sites (motif length) is usually assumed to be constant and no gaps are permitted. An example of these experimentally determined sequences

is presented in Table 1.1.

Since binding sites are short segments similar to known experimental sequences S_i (see Table 1.1), we consider the problem of the determination of TBFS location as one of searching for subsequences most conserved over the whole set $\mathcal{S}$. As this problem is similar to local sequence alignment, we can apply conventional tools developed for this latter purpose. The Smith-Waterman [99] algorithm is based on dynamic programming and is a common algorithm used for local sequence alignment. Locally aligned similar segments are shown in upper cases in Table 1.1 where the motif length is $L = 9$.

In order to go further, we simplify the notation used in (1.4.3). Depending on the problem at hand, we can refer to the elements of array $\mathcal{S}$ either as nucleotides or the corresponding indexes. For any subset of indexes $\mathcal{I} \subseteq \mathcal{S}$, we denote $\mathcal{S}_\mathcal{I} = \{s_{kl} : (k,l) \in \mathcal{I}\}$. In particular, the set of aligned subsequences is denoted with a bar sign: $\bar{\mathcal{S}}_\mathcal{I}$. Meanwhile, a complementary set is automatically identified as $\mathcal{I}^c = \mathcal{S} \setminus \mathcal{I}$ and the aligned subsequences $\bar{\mathcal{S}} \subseteq S$ are also defined as

$$\bar{\mathcal{S}} = \{\boldsymbol{S}_k\}_k = \{s_{kl} | l = 1, ..., L; (k,l) \in \bar{\mathcal{S}}_\mathcal{I}\}, k = 1, ...K \tag{1.4.4}$$

Indexes $\mathcal{I}$ followed by $\bar{\mathcal{S}}$ are able to capture more complex patterns that contain gaps (i.e. sequences containing insertions and deletions relative to each other) or in which different positions are correlated with each other [43]. However here, recalling the assumption from that fundamental work of Hertz and Stormo [43], we assume that the positions of an alignment function independently according to whatever biochemical criteria were used to select the underlying, functionally related sequences.

From here, we will refer to the coordinates of nucleotides of DNA sequences. The local coordinates of nucleotide positions in an aligned set $\bar{\mathcal{S}}_\mathcal{I}$ increase from left to right assuming that this corresponds to the downstream direction. All positions are centered at the TSS position point (coordinate of "zero").

Table 1.1: The table illustrates initial sequences of different length that are used to construct a statistical model of the most similar parts of the sequences. The data were taken from the JASPAR database for transcription factor OVO, which controls germline and epidermis differentiation in *Drosophila*. Upper cases indicate the most similar segments after alignment

Sequence S_1	gTGTAACTGTctttctaccgccaaggcctctg
Sequence S_2	ccagctggtaTGTAACCGTagtacacctgagc
Sequence S_3	tgaCGTAACAGTgagcataaggggggcaattgg
Sequence S_4	agtccaAGTAACTGCaacacctttagaccact
Sequence S_5	ttAGAAACAGTggtagtgttggggacctgaac
Sequence S_6	tcgtcccacTGAAACAGTaccggggctttgtg
Sequence S_7	gcgtccccctttaccaaaACTAACTGTcccca
Sequence S_8	ccgactcattAGTAACGGActcccatacgcgg
Sequence S_9	ctcgtgtAGCAACTGTgagaaccttaacagat
Sequence S_{10}	ggcataggaaattccTATAACTGTgcagatcc
Sequence S_{11}	ggttatggtttaggtcATTAACAGCgagcggt
Sequence S_{12}	cTGTTACTGTaatccgggccggactgga
Sequence S_{13}	tTCGAACAGTtgagatctcctaggggcccatg
Sequence S_{14}	ggtctgttcagtccTCTAACGGCcatttgtac
Sequence S_{15}	caggactgttcgctagctcttAGAAACCGAct
Sequence S_{16}	gtaggctactaggctactaggctcatAGAAACCGAcc
Sequence S_{17}	gccgtaatAGCAACCGAcacgagatctttcga
Sequence S_{18}	ctcgcatcccttatCGTAACCGGtccacag
Sequence S_{19}	cttcgACGAACAGCacacccccaatagccgca
Sequence S_{20}	gtcgtgtagcaactatgacaacCTTAACAGAt
Sequence S_{21}	caccagcacccctaatcagaCGTAACGGGcc

Thus, the aligned sequences $\bar{\mathcal{S}}$ split the set $\mathcal{S}$ into two parts - the binding site and the non-site positions. Another data structure, constituting the alignment, is a set of positions a_k for each k from 1 to K, for the starting positions of the elements within each sequence S_k.

Let each m_{kj} be the number of nucleotide k at position j in an aligned set of indexes $\bar{\mathcal{S}}$. Then we define an integer matrix $\mathbf{M}$

$$\mathbf{M} = (\mathbf{m}_1, ..., \mathbf{m}_L) = \begin{pmatrix} m_{11} & m_{12} & \dots & m_{1L} \\ m_{21} & m_{22} & \dots & m_{2L} \\ m_{31} & m_{32} & \dots & m_{2L} \\ m_{41} & m_{42} & \dots & m_{2L} \end{pmatrix} \tag{1.4.5}$$

The matrix $\mathbf{M}$ has the format of a PWM and is used in the TRANSFAC database to represent the binding specificities of TFs where they are called position frequency matrices (PFMs) [71]. $\mathbf{M}$ is also frequently called an alignment matrix [42].

Under the i.i.d. assumption defined above, the *a priori* probability of the sequence of L letters is the product of the prior probabilities of the individual letters. Each column is assumed to have its own prior. Our most immediate goal is to model the kinds of distributions that most likely generate the actual observed occurrence counts $\mathbf{m}_j$.

The probabilities of the individual letters in equation (1.4.1) can be estimated using the overall frequencies of the letters within all sequences of genome. A simple approach would be to estimate probabilities for each count column $\mathbf{m}_j$ using the frequencies of letters in the aligned sequences $\bar{\mathcal{S}}$. As shown later in (1.4.16), under the assumption of independence, each single vector $\mathbf{m}_j$ of nucleotide counts is interpreted as the vector-parameter of a multinomial distribution with unknown probabilities. Now we can estimate these probabilities as parameters from $\bar{\mathcal{S}}$ using the usual maximum likelihood estimator (MLE), i.e. by finding the $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_L$ as

$$\begin{gathered}\boldsymbol{\theta}_j = \arg\max_{\boldsymbol{\theta}_j}(Prob(\mathbf{m}_j \mid \boldsymbol{\theta}_j)), \\ \Sigma_i \theta_{ij} = 1; j = 1, ..., L\end{gathered} \tag{1.4.6}$$

This equation leads to the MLE $\boldsymbol{\theta}_j = \mathbf{m}_j / \sum_l m_{lj}$ [47], which is a normalized summary of aligned data (since in most cases $\sum_l m_{lj} = const$). MLE provides poor (due to a bias if only a small number of binding sites is available) although widely used estimates for the actual underlying probability distributions [15, 102, 120].

The MLE is sensitive to the correctness of underlying probability model specification and some authors suggest other estimators for multinomial distribution instead of (1.4.6). A particular reason for doing so is that distinct positions of multinomial space, representing different combinations of nucleotides, should have high prior probabilities. To correct the bias, some authors propose using a numeric value, called a pseudocount [79]. The pseudocount is usually allocated for each of L positions, and its fraction according to the background base composition is added to each $\boldsymbol{\theta}_j$. Another example to address this small sample size deficiency of probabilistic model is the work of Brown *et al.* where Dirichlet mixture priors were proposed [47]. Although the mixture priors are beyond the scope of the thesis we will discuss single Dirichlet priors a few paragraphs later because of its methodological importance.

Returning to equation (1.4.6), $4 \times L$ matrix of weights ordered as A, T, G, C (also called mononucleotides as opposed to dinucleotides defined and used in chapter 2 is one of the simplest PWMs

$$\mathbf{W} = \{w_{ij}, i = 1, .., 4; j = 1, ...L\} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_L) = \begin{pmatrix} \theta_{11} & \theta_{12} & \dots & \theta_{1L} \\ \theta_{21} & \theta_{22} & \dots & \theta_{2L} \\ \theta_{31} & \theta_{32} & \dots & \theta_{2L} \\ \theta_{41} & \theta_{42} & \dots & \theta_{2L} \end{pmatrix} \tag{1.4.7}$$

There are several interpretations and approaches that exist to compute weights will

consider in this thesis.

Such an example was introduced in 1984 by Staden. Working independently from Stormo, he proposed to use PWM with *log*-frequencies, calculated as $\boldsymbol{\theta}_j = -\log \frac{\mathbf{m}_j}{\sum_l m_{lj}}$ [100]. *Log*-frequencies were just a first step toward broader class of PWMs, where computation of individual weights were motivated by a variety of technical and statistical metaphors. One of such computational PWM types was mainly used in the thesis and therefore is discussed separately in section 1.5.

Unlike *log*-frequencies $\boldsymbol{\theta}_j$, vectors of nucleotide occurrences $\mathbf{m}_j$ keep the binding specificity information from known TFBSs unchanged and leave a question of how to compute $\boldsymbol{\theta}_j$ to the researcher's discretion. The intuitive definition of PWMs as in (1.4.1) can be extended from a sequence to any set of sequences. That is, the probability to obtain a set of K length-L sequences $\mathcal{S}'$ when sampling K sequences from $\mathbf{W}$ is given by the joint probability

$$P(\mathcal{S}'|\mathbf{W}) = \prod_{s\in\mathcal{S}'} P(s|\mathbf{W}) = \prod_{j=1}^{L}\prod_{l=1}^{4}(\theta_{lj})^{m_{lj}(\mathcal{S}')}, \tag{1.4.8}$$

where $m_{jl}(\mathcal{S}')$ is defined as the number of times the letter s occurs at position i in the aligned sequences $\mathcal{S}'$. Probability P is defined for the whole sample space of $4 \times L$ sequences. Thus, the probability of obtaining sequences $\mathcal{S}'$ when sampling from $\mathbf{W}$ depends only on the counts $m_{lj}(\mathcal{S}')$. Using Bayes' theorem, the posterior probability $P(\mathbf{W}|\mathcal{S}')$ for $\mathbf{W}$ given a set of sites $\mathcal{S}'$ is given by

$$P(\mathbf{W}|\mathcal{S}') = \frac{P(\mathcal{S}'|\mathbf{W})P(\mathbf{W})}{P(\mathcal{S}')} \tag{1.4.9}$$

In this equation, $P(\mathbf{W})$ is the prior probability of $\mathbf{W}$. The denominator is a normalizing constant. The prior $P(\mathbf{W})$ represents our prior information about $\mathbf{W}$ before looking at the data. As will become clear later, the computations are analytically most easily tractable if we use the Dirichlet prior that has the following general form

$$P(\mathbf{W}) = \prod_{j=1}^{L} P(w_j) = \prod_{j=1}^{L} c_j \prod_{l=1}^{4} (\theta_{lj})^{\alpha_{lj}-1}, \tag{1.4.10}$$

where c_j is a normalizing constant for column j, described in (1.4.15), and the α_{jl} are constants that determine the prior. There are no special requirements for the selection of α_{jl}, except the trivial case when $\alpha_{jl} = 1$. In this last case, all θ_{lj} are the ignored source of information and all positions in the motif are uniformly distributed. However, Stormo *et al.* [103] argued that efforts made to estimate θ_{lj} using laboratory measurements showed that columns with equal probabilities $\theta_{lj} = 0.25$ are not very likely and thus the product-Dirichlet model (1.4.10) allows us to account for more scenarios with variables θ_{lj} and parameters α_{jl}. Moreover, α_{jl} is actually controlled by one free parameter $\alpha = \sum_{lj} \alpha_{jl}$ that follows from properties of Dirichlet distribution.

In this paragraph we give the exact formula to compute $P(\mathbf{W}|\mathcal{S}')$. To do this we put the likelihood (1.4.8) and our Dirichlet prior (1.4.10) into (1.4.9). Taking into account that the denominator in (1.4.9) does not depend on $\mathbf{W}$ and re-grouping the multipliers in the resulting formula gives the product Dirichlet distribution

$$P(\mathbf{W}|\mathcal{S}') \propto \prod_{j=1}^{L} \prod_{l=1}^{4} (\theta_{lj})^{m_{lj}(\mathcal{S}')+\alpha_{lj}-1} \tag{1.4.11}$$

Adding a multiplicative constant C to transform equation (1.4.11) into identity and applying integration over all possible $\mathbf{W}$ to left and right sides of resulting equation, we obtain

$$1 = \int_{\mathcal{D}} P(\mathbf{W}|\mathcal{S}')d\mathbf{W} = \int_{\Delta_1} \cdots \int_{\Delta_L} C \prod_{j=1}^{L} \prod_{l=1}^{4} (\theta_{lj})^{m_{lj}(\mathcal{S}')+\alpha_{lj}-1} d\theta_{l1} \ldots d\theta_{lL}, \tag{1.4.12}$$

where $\mathcal{D} = \{\mathbf{W} = (\boldsymbol{\theta}_j)_j | \sum_l \theta_{lj} = 1; \forall j = 1, \ldots, L\}$; the integration domain Δ_j for each j is an independent copy of simplex in $\mathbb{R}^4$, such as $\sum_l \theta_{lj} = 1$. The solution of the integral (1.4.12) is given as the product of simpler multiple integrals over the

same simplex domain, namely as shown in [73, 119], the exact solution is

$$\int_{\Delta} (x_1)^{y_1-1} \dots (x_n)^{y_n-1} dx_1 \dots dx_n = \frac{\prod_{j=1}^{4} \Gamma(y_j)}{\Gamma(\sum_j y_j)} \tag{1.4.13}$$

Finally, the complete formula to compute (1.4.11) is given by

$$P(\mathbf{W}|\mathcal{S}') = \left(\prod_{j=1}^{L} \frac{\Gamma(m_{\bullet j}(\mathcal{S}') + \alpha_{\bullet j})}{\prod_{l=1}^{4} \Gamma(m_{lj}(\mathcal{S}') + \alpha_{lj})} \right) \prod_{j=1}^{L} \prod_{l=1}^{4} (\theta_{lj})^{m_{lj}(\mathcal{S}')+\alpha_{lj}-1}, \tag{1.4.14}$$

where $m_{\bullet j}(\mathcal{S}') = \sum_{l=1}^{4} m_{lj}(\mathcal{S}')$ and $\alpha_{\bullet j} = \sum_{l=1}^{4} \alpha_{lj}$.

Normalizing constants c_i of the Dirichlet distribution (1.4.10) can be computed in a similar way as C. It gives

$$c_j = \frac{\Gamma(\sum_{l=1}^{4} \alpha_{lj})}{\prod_{l=1}^{4} \Gamma(\alpha_{lj})} \tag{1.4.15}$$

The joint probability of occurrences in the alignment matrix (1.4.5) with the aforementioned i.i.d. assumption over nucleotide positions follows the product-multinomial distribution [42]

$$P(|\boldsymbol{m}_1|, ..., |\boldsymbol{m}_L|; \mathbf{W}) = \prod_{j=1}^{L} fM(|\mathbf{m}_j|; K, \boldsymbol{\theta}_j) = \prod_{j=1}^{L} \frac{K!}{\prod_{k=1}^{4} m_{kj}!} \prod_{l=1}^{4} p_l^{m_{lj}}, \tag{1.4.16}$$

where $\mathbf{m}_j$ is the i.i.d. random variable following the multinomial distribution, fM, p_l is the probability of single l base, and m_{kj} are components of $\boldsymbol{m}_j$.

From equation (1.4.16) we can make a simple but meaningful conclusion for the alignment matrix $\mathbf{M}$. As a result of the i.i.d. assumption, despite the possible positional differences in the frequency distribution θ_{kj}, any permutation of columns in matrix $\mathbf{M}$ would not change the probability of composition (1.4.16). To implement the possibility of positional dependencies between neighboring nucleotides, we used

a dinucleotide PWM model that can be built preserving all PWM models built for mononucleotide PWM with two differences, (1) the state space is the 16-letter alphabet of dinucleotides and (2) the length L becomes $(L-1)$. In the next chapter we work with both mono and dinucleotide matrices.

In conclusion, we provided here a theoretical background which builds a basis for probabilistic modeling and parameter estimation of PWM, which is presented either as a product multinomial distribution of nucleotide occurrences or as a product Dirichlet distribution of base probabilities. Although each of these model types included probabilities of bases as an input, it is computationally unreasonable to use such models to scan large promoter areas due to non-homogeneous background nucleotide distribution discussed later.

1.5 Heuristic methods for computational prediction of binding sites

Before describing more specific PWM types, we will discuss the existing classification of PWMs in terms of biological data mining, a computational approach based on heuristic methods. Computational methods for the prediction of TFBSs fall into two broad classes: *de novo* methodologies in which upstream regions of genes are analyzed for overrepresented motifs [58, 67]; and training-based methodologies in which existing experimental data are used to identify instances of similar transcription factor binding sites. The main difference between the two is that the former methods emphasize an overrepresentation as a prime factor in modeling and the latter methods start from known binding sites and attempt to find similar TFBSs by using statistical information and similarity metrics to make predictions.

Although typical models discussed above belong to supervised leaning methods, a variety of existing ways to capture statistical information from training is too large

to be examined here in details. Practical applications may also include the variety of schemes used to train matrices, so that the number can be even larger. It is worth mentioning that the training-based methodology itself assumes a broad variation of interpretations. For example, many authors often use the training metaphor to address simple PWM computations from set of aligned sequences $\mathcal{S}'$. Unlike these authors, we preserved the term "training" in the context of more immersive machine learning framework, which involves several steps of biological signal detection on the promoter sequences and optimization steps with feedback.

The availability of whole genome sequences gave rise to the development of new computational methods for searching for TF binding sites using training data based on experimentally-verified TFBSs stored in databases. However, the number of experimentally-determined TFBSs in such databases is typically small. Thus, estimates (1.4.6) in many PWMs may be poor, because they may describe random error or noise instead of actual relationships. This is referred to as a PWM over-fitting problem in which generic PWMs are constructed from an insufficient amount of data. From a practical computational point of view, another limitation of using known TFBSs to identify new instances of these sites is that the probability distributions in PWM models can be hardly used for large scale predictions of new putative binding sites because of computational sampling complexity.

Another reason for the transition to other types of specificity models is that the probabilistic models described in the previous section are not always the best choice for discovering new TFBSs, as the trade-off between computational complexity, predictability and amount of binding sequences needs to be considered in PWM.

The development of methods that are reliable and have low false positive and false negative rates is of paramount importance. In contrast to models studied above, Stormo and Fields [101] used a weight matrix defined as the "information content" that defined w_{lj} with respect to the binding affinity. It is then interpreted as a measure of binding energy. A similar heuristic was studied in the work of Djordjevic et

al. [27]. In addition many other interpretations of PWMs also exist (some of them are following), for example log-odds scores (ratios), log-likelihood of the substring under a product multinomial distribution, log-likelihood ratio, binding energy, measure of information content, Kullback-Leibler divergence, relative entropy etc. In the thesis we follow the notorious interpretation of PWMs defined as the log-odds scores matrices as they tested suitable for training [15, 35].

A widespread graphical presentation of aligned sequences $\bar{\mathcal{S}}$ in equation (1.4.4) or PFM in equation (1.4.5) is a sequence Logo, which can fingerprint any PWM. Sequence Logo was constructed by Schneider in 1986 who also defined "information content" based on Shannon's entropy [93]. Equation (1.5.1) gives the original presentation of sequence Logo for gap-less sequence S of nucleotides.

$$\boldsymbol{h_j} = \boldsymbol{\theta}_j(2 - H_S(j) + e(n)); H_S(j) = -\boldsymbol{\theta}_j \cdot \log_2 \boldsymbol{\theta}_j, \tag{1.5.1}$$

where $H_S(j)$ is the Shannon's entropy (uncertainty); $2 - H_S(j) + e(n)$ is the information content, and $\boldsymbol{h_j}$ is the high of chart at position j; $e(n)$ is pseudo-weight correction for small samples. The unit of $H_S(j)$ is called "bit". Thus, $\boldsymbol{h_j}$ measures bits of information and varies between 0 and 2.

Each Logo (in version for DNA sequences) consists of stacks of nucleotide symbols, one stack for each position in the sequence. The overall height of the stack is proportional to the information content at that position, while the height of symbols within the stack indicates the relative frequency of each nucleotide at that position. In general, a sequence Logo provides a richer and more precise description of, for example, a binding site, than would a consensus sequence. We use Logo as a visualization tool to represent a whatever collection of TFBSs do we have, such as the initial set of TFBSs or the new set of predicted TFBSs. We also call "Logo" a different metric than information where y-axis means not entropy but nucleotide proportions at each particular position. Examples of Logos of both types can be found in Figure 3.1 and

in Figures 3.4 and for the remaining TFs in Figures 5.2-5.6.

With regards to statistical description, the characteristics of the motif nucleotide distribution are fully defined by the nucleotide frequencies $(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_L)$ as used in (1.4.7). So far, we have assumed that background nucleotide frequencies are equal. However, a complication arises in eukaryotic genomes where regulatory elements usually have non-homogeneous nucleotide content. For instance, some areas in the genome can be GC-rich. As a result a PWM built on frequencies trained on one particular nucleotide content may not be useful in areas with a different background nucleotide distribution. As a workaround, Stormo proposed to compare the base frequencies with a vector-parameter of background frequencies $\boldsymbol{\theta}_0 = (\theta_{1,0}, \theta_{2,0}, \theta_{3,0}, \theta_{4,0})^T$ (in future we omit the coma between indexes) established from the whole genome or at least from sequences significantly longer than the length of a motif [101]. Background frequencies $\boldsymbol{\theta}_0$ can be estimated in a straightforward way when they are compiled from a large area which includes target TFBSs .

In equation (1.5.1), "2" corresponds to complete uncertainty where any of four nucleotides can be at a certain position. For skewed genome sequences the information content should be corrected to $\boldsymbol{\theta}_j \cdot log_2 \frac{\boldsymbol{\theta}_j}{\boldsymbol{\theta}_0} + e(n)$. This dot-product is used in the literature as a component of certain kinds of PWMs. For instance, Stormo used this normalized log-likelihood ratio as a component of his PWM $\boldsymbol{w_j} = \boldsymbol{\theta}_j log_2 \frac{\boldsymbol{\theta}_j}{\boldsymbol{\theta}_0}$ which is also known as Kullback-Leibler information of relative entropy [42]. Both of these are used as PWM metaphors.

Notably, for the sake of simplicity, the abbreviation "PWM" as will be used in the next chapter refers to any type of matrices regardless of the type of coefficients. In general terms all PWMs can be regarded as matrices, where components are computed using a variety of schemes. In this thesis we avoid multiple terms for different matrix types. Thus, the abbreviation "PWM" is preserved for any matrix type representing a model of binding specificity based on a collection of known binding sites, despite the difference in domains they may have. Now, the word "weight" in the abbreviation does

not necessarily refer to the components and the weights are interpreted as proportions and thus bounded by zero and one. Particularly, in chapter 2 we use PWMs with integer components (PFM). In cases where the type of PWM may contradict the readers' understanding we will specify the type of numeric domain where they are defined, as discussed in the next section.

In the next two sections we discuss the main advantages and common drawbacks of PWMs and describe some possible ways to improve their predictive performance.

1.6 On the use and misuse of PWM to predict TFBSs

Arguably, PWMs are the most commonly used statistical model to depict the binding specificity of a given TF, however, we here describe the benefits and limitations of PWMs as a family of methods.

Although PWMs were shown by many authors to be simple and successful computational tools for motif specificity prediction [16, 101, 103], they are based on many biological and computational assumptions that often remain to be evaluated experimentally. Some of these commonly accepted assumptions include:

1. A PWM assumes that the recognition sequence is of a fixed length, a hypothesis that has been seriously questioned for a number of eukaryotic TFs [15].

2. Individual bases of the recognition sequence contribute independently of each other to the total binding energy of the DNA-protein complex. For sake of simplicity in a PWM, each position of a binding site is modeled as making an independent contribution to the overall binding affinity of the site. Experimental evidence suggests that this assumption of independence is not always valid, but researchers argue that in most cases it provides a good approximation of the

true nature of the specific protein-DNA interaction [103, 36, 35, 27, 6]. Whereas many authors study dependencies between nucleotides [4], some suggest that the i.i.d. assumption still gives a good approximation to describe TF-DNA binding [6, 103].

3. It is also questionable whether the matching score of PWMs relates to the actual affinity of target DNA sequences for a TF [103, 88]. These scores are presumed to reflect the affinity of a TF for the bound sequence, but in almost all applications, a cutoff score is chosen to distinguish between functional and non-functional binding sites. Given the variety of TFs, it is unlikely that the use of a common threshold for all TFs is appropriate [63]. For this reason many authors incorporate additional heuristic information into PWMs to strengthen their biological relevance. The heuristics include training on affinity-threshold models or the use of evolutionary patterns or models of nucleotide distributions on promoters.

A PWM-based computational method to predict new TFBSs usually follows the standard strategy: given a DNA sequence and a PWM denoted by $\mathbf{W}$, the method uses a score that measures the quality of the match between the sequence and the matrix. The promoter (or another target sequence) is scanned and the binding motif of maximal score is found. Denote (temporarily) the value of this maximal score as c. We then need to decide whether c is high enough to call the corresponding sequence a "hit", i.e. predict that the TF binds at the corresponding position. This decision is taken on the basis of the comparison of c with a threshold $c_0 = c(\mathbf{W})$: if $c > c_0$, there is a match (further also called "hit") and a putative binding site of the TF is identified. All methods that use PWMs have to deal with the question of determining an optimal threshold for each TF.

PWM models built from available in database TFBSs (usually a small number) may be ineffective in distinguishing a true motif from a random segment. A less

stringent c_0 results in a larger number of false positives whereas a more stringent c_0 eliminates true positives and generates false negatives. This problem makes the genome-wide motif identification using PWMs less useful. In the next section of the introduction we sketch out a PWM refinement approach which opens up a way for a large-scale PWM applications that have better combination of true and false positives.

1.7 PWM refinement to improve prediction

Consider two possible scenarios. In the first, we only have 10 sequences from which to estimate the parameters of the model. In the alignment of these 10 sequences we have a column containing only As, and no other nucleotides. With such a small sample, we cannot rule out the possibility that similar binding sites may have different nucleotides at this position. In particular, we presume that we know that A is commonly substituted by G in some promoters. Thus, our estimate of expected probability distribution at this position would include these nucleotides, and perhaps the other nucleotides (based on our biological intuition), albeit with smaller probabilities.

In the second scenario, we have an alignment of 100 divergent sequences and again the second column containing only A, and no other nucleotides. In this case, we have much more evidence that As and no other nucleotides are possible at this position in the motif. Indeed we have much more evidence that A is functionally conserved at this position. Therefore, generalizing the distribution of nucleotides at this position to include "similar" nucleotides is likely not biologically relevant. In this situation, it makes more sense to give less attention to prior beliefs about similarities among nucleotides (biological intuition), and more attention to the actual counts observed.

As mentioned in section 1.3, many transcription factors have low specificity for their binding motifs and this degeneracy makes it hard to estimate how many di-

vergent sequences are necessary to assure that certain positions are conserved (as described in the second scenario). This observation points to machine learning methods and other integrative approaches capable of incorporating more specifc biological information to improve PWMs as statistical models of TF-DNA binding specificity [59].

The high false-positive rates in TFBSs prediction using a variety of generic PWMs led to various attempts to draw in extra information to improve predictive performance of PWM [47, 107, 59, 61]. The topic of this thesis remains in focus of such earlier attempts which are furthered in next chapters [15, 35, 36].

We mention a few computational strategies used to improve the performance of PWMs as predictive tools. The simplest way to improve the performance of conventional PWMs is to take into account the dependencies among the adjacent positions [4, 96]. This model (also termed the weight array model) is represented by a dinucleotide PWMs [96, 36]. Besides the obvious advantage of involving higher-order statistics, this model may capture longer lengths of motifs. In this thesis I used dinucleotide matrices in addition to mononucleotide PWM and compated the performace of the two approaches.

Examples of successful approaches to improve performance of PWM include: genetic algorithm that maximizes the area under ROC curve [61], or the detection of locally positioned dinucleotides, identified from known sites using a genetic algorithm with discriminant analysis [59].

A radically different approach to PWM refinement was introduced by Gershenzon [35], an approach that was implemented in my thesis project for mononucleotide and dinucleotide matrices. Several steps are involved to detect potential putative binding sites on promoter sequences. While searching for a new auxiliary TFBSs, this approach also implements optimization steps to estimate the motif length more precisely, optimize matching cut-off and the location on the promoters. Figure 1.3 shows a sketch of the algorithm's steps, which are examined in the next chapter, where

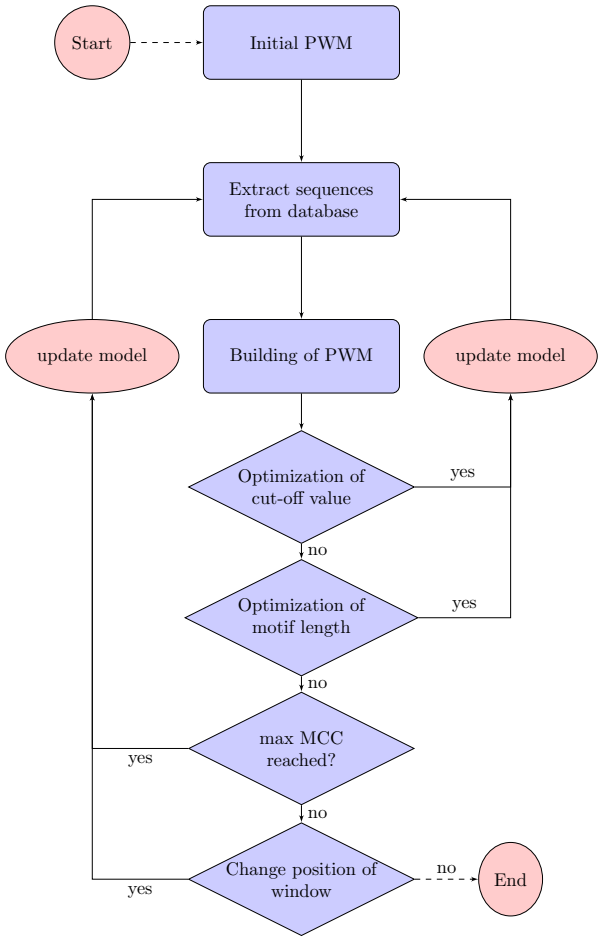

Figure 1.3: The flowchart for PWM optimization process. *MCC* means Mattews correlation coefficient (Adapted with modifications from [35]).

we start introducing its theoretical background. This algorithm is referred to as the Bucher's algorithm for weight matrix refinement where the local over-representation of binding motifs is used as optimization criteria [15]. Unlike Bucher, Gershenzon used Matthews correlation coefficient which is used throughout all the algorithm's steps as the optimization criteria [35].

1.8 Motivation and scope of the thesis

In this work we were motivated to continue systematic development of the approach for PWM refinement introduced by Gershenson *et al.* We aimed to extend this approach to a large number of TFs from *Drosophila*. The main goal of our study was the construction of refined PWMs for as large as possible numbers of available PWMs from TRANSFAC. The complementary goals were to characterize the predictability and analyze the performance of the refined PWMs using direct tests, comparisons with existing methods, and literature evidence. In order to achieve our goals, we resolved a range of technical problems, such as the selection of methods and tools to estimate statistical significance of TFBS, development of tests and criteria for motif over-representation. We created a variety of custom software and scripts that are available to other academic researchers faced with similar challenges.

Previous work by Bucher and Gerchenzon *et al.* implemented a PWM refinement approach for a few single promoter elements. Although these two studies were based on the same idea of using promoter sequences, they used two different algorithms for PWM refinement. In our study we use Gershenzons *et al.* algorithm for PWM refinement with additional modifications and corrections to allow for batch processing.

Based on the work quoted above, from our preliminary analysis of PWM the matching score distribution for TATA-box, Inr and Sp1, we found that PWM for these elements returned strong matching scores, which would not be expected to occur by chance in a large scale application involving many TFs.

These conditions from the last two paragraphs challenged us to finding a trade-off between best features of already published technique and additional improvements arisen from a large scale application, especially if PWM matching signals are weak or noisy. To reach this goal we had to revise all steps of the core idea of the PWM refinement. Based on Gershenzon *et al.* we continued using both mono and dinucleotide PWM versions. There were three reasons for using both mono and dinucleotide PWMs, (1) it was unclear which PWM type would perform better for a particular TF and (2), methodologically it was expected to be helpful to use both types as an additional cross-validation instrument during the software development stage and (3) using both types of PWMs may bring additional biological insight into TF-DNA binding.

To represent the iterative PWM refinement technique of Gershenzon *et al.* we used the supervised machine learning metaphor [74]. The method which was implemented in the thesis interpolates smoothly between the reliance on the prior information concerning likely nucleotide distributions, in the absence of large amounts of accurate data, and the confidence in the nucleotides frequencies observed at each position. In such an integrative approach, a supervised PWM refinement technique can be compared with a strategy of searching for "similar" motifs from promoter sequences. As we show in this thesis, the refined PWMs of both types mono and dinucleotide have better performance, compared to conventional methods, biologically relevant and can be used as computational tools for prediction of new TFBSs.

Chapter 2

Computational models of PWM scores

2.1 Model of PWM with promoter background compensation

As described in section 1.4 of the introduction, the probabilistic models of binding specificity tend to be more suitable to study their formal properties than to contribute to the computational prediction of putative TFBSs on empirical data. In addition to what we already said on page 25, an important drawback of the probabilistic models is that they are simplified. In particular, the models do not take into account the heterogeneous nucleotide distributions within cis-regulatory elements. In section 1.2 we mentioned the evolutionary and functional conservation of regulatory sequences that result in non-uniform distribution of nucleotides. Computational identification of functional TFBS in regulatory modules becomes a challenging problem. A heatmap of nucleotide distribution for *Drosophila* promoters taken from EPD database (`http://epd.vital-it.ch/`) is present in Figure 2.1. This figure indicates the zero position of TSSs. The annotation of specific TFBSs is complicated by the fact that these

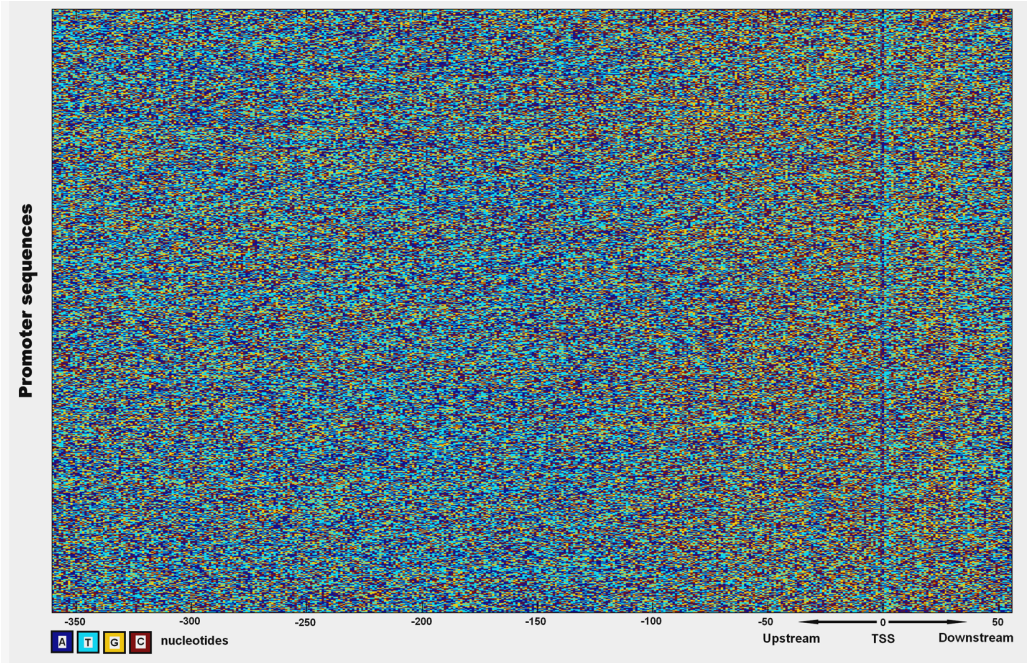

Figure 2.1: Distribution of nucleotides on an aligned set of 1919 *Drosophila* proximal promoter sequences shown as a heatmap. Nucleotides A, T, G, C are color-coded as shown in the legend. The distribution of nucleotides has non-random pattern, especially around TSS. An example of dinucleotide distribution for GG, GC, CG, CC shown in Figure 5.1 in appendix

short, degenerate sequences may frequently be conserved by chance rather than by functional constraint.

We exploit the idea that motifs, which are necessary for transcriptional regulation, should be overrepresented in a particular area of the genome. We call a binding site overrepresented if it occurs more frequently than expected under the conditions of a certain statistical model. We refer to the nucleotide distribution over a much bigger area than the motif length as background nucleotide distribution (or background nucleotide frequencies mentioned on page 27). In this regard, we can say that the "expected" distribution is a nonuniform background nucleotide distribution (when no TFBS is present), which can be positionally site-dependent. While it is generally accepted that an overrepresentation of particular oligonucleotides in the DNA sequence

is biologically relevant, ways of quantifying this overrepresentation are quite variable and range from exhaustive direct searches to numerical filtering.

Unlike exhaustive enumeration of sequences, many authors use a variety of scores and statistics to measure the abundance of nucleotide letters ("words") [58, 18, 91, 67]. Another straightforward approach to detecting overepresentations of DNA words is based on TF binding specificity model, such as a matrix or a PWM model, also illustrated in Figure 2.2. The PWM based approach to detect overepresentations can be applied with different types of PWM, based on how the matching score and cut-off values are computed. We used slightly corrected log-odds scores PWM coefficients compiled from nucleotide frequencies in this thesis.

The standard strategy to filter out TFBSs using PWM (see page 29) suffers from a large number of false positives. It is *a priori* impossible to completely overcome this problem due to the short length of TFBSs. Even so, some authors have implemented additional indicators such as z-scores or other statistical characteristics of nucleotide distributions [44, 35, 15] to narrow down search area(s) with potential target sequences. We also follow a similar strategy defined later in section 3.2.2.

Although any PWM is constructed from known TFBSs, a PWM properly adapted to promoter areas such as shown in Figure 2.1 has to take into account the background nucleotide distribution around these areas. More specific aspects arising for the score-based type of PWM are considered next.

Score-based PWMs for biological sequence analysis were first introduced by Stormo and colleagues in 1982 [102] and Staden in 1984 [100]. These authors defined the PWM as a two-dimensional array of components, which represents the scores for finding each nucleotide at each position. Bucher in 1990 [15] noticed that there is some controversy about how weights in PWM should relate to frequencies $\boldsymbol{\theta}_j$. He proposed the PWM format, which is very close to that defined in equation (2.2.2) and accounts for the background nucleotide distribution by considering $\boldsymbol{\theta}_0$.

After first works of Stormo and Bucher the ratio of nucleotide frequencies became

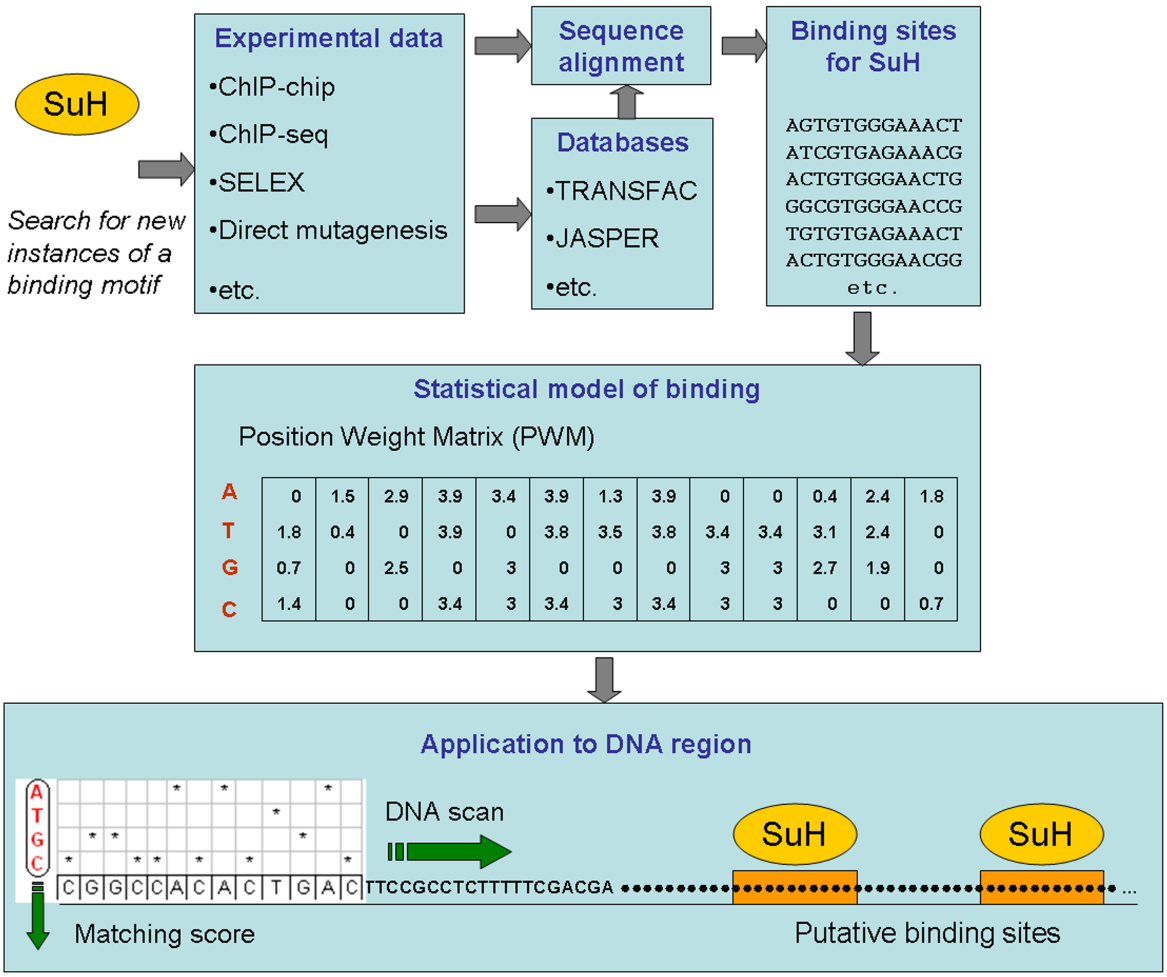

Figure 2.2: Illustration of the process of searching for new putative TFBS for *Drosophila* SuH TF using a mononucleotide PWM

a notorious part of many score-based PWMs. In section 2.2 of this chapter we provide accustomed to our needs specifications of logg-ods scores that used in our software, but here we need to address the important common issues with ratios.

Let parameters $\mathbf{m}_j$, $\boldsymbol{\theta}_j$,$\boldsymbol{\theta}_0$ be those valid for a given aligned set of sequences $\mathcal{S}'$. Then we define an integer matrix of vector-scores: $\mathbf{W} = (\boldsymbol{w}_1, ..., \boldsymbol{w}_L)$ such as

$$\boldsymbol{w}_j = \log \frac{\boldsymbol{\theta}_j}{\boldsymbol{\theta}_{0,j}} \tag{2.1.1}$$

In this formula $\boldsymbol{\theta}_j$ is a vector where each element represents the number of times with which a specific nucleotide base occurs at the j-th position in the motif of length L, and $\boldsymbol{\theta}_{0,j}$ is the background base frequency. We preserved index j in the denominator because the expected frequency is a piecewise constant vector-function of index j. The numerator and denominator in equation (2.1.1) can change values non-synchronously depending on the background model used. For instance, Stormo and Fields [101] used estimated probability of base occurrence in the genome, so, that $\boldsymbol{\theta}_{0,j} = \boldsymbol{\theta}_0 = const$.

The idea to model the background nucleotide distribution is not novel. For instance, several researchers used Markov models (usually of order zero to three) to describe the background nucleotide distributions $\boldsymbol{\theta}_{0,j}$. More specifically, Tompa *et al.* used a Markov chain of order three [109]. New-generation sequencing technologies provided an opportunity to use large number of DNA sequences available for training of such models. A drawback of this approach is that the "correct" stochastic process that nature uses is unknown, and so we would be introducing biases while constructing PWMs based on preconceived background models. Nevertheless, Tompa *et al.* mentioned that accuracy tests on computational TFBSs prediction depends on the background model. For reasons mentioned above we are not going to set up Markovian properties directly but rather to construct a machine learning technique which is able to capture existing nucleotide dependencies.

Having been asked by someone "What kind of PWM model is better of describing

TF-DNA binding specificity?" it would be fair to answer the following. Whereas PFM is the most common format of collecting an aligned TFBSs, the integer-valued PFM *a priori* is a poor instrument for TFBS predictions (at least because PFM model does not account for the background nucleotide irregularity). In turn, real-valued score based PWMs (such as Stormo's log-likelihood ratios or log-odds scores, see page 1.5), constructed to be computational predictors of new TFBSs, are not suitable for representing all known TFBSs since this information is hidden in weights. Finally, any particular PWM type that has its own advantage suffers from common weaknesses such as listed in the introduction. Therefore, the choice of PWM depends on the purpose of PWM selection.

These considerations provide extra motivation for my research project. When one aims to compile a PWM which is expected to be a proper hands-on predictor of new TFBSs, the optimization metaphor is definitely a good choice. The basic methods do not show which cut-off should be taken and how to pick the representative set of sites with which to construct PWMs. Answering both questions is important since it is not uncommon to have false positives among "known" TFBSs [103]. Optimized PWMs would be helpful in finding the answers but the optimization problem was always considered as one of the most challenging computational problem, especially if the optimization model is based on heuristics and the optimization domain is mathematically poorly formalized.

Before deliberating the systematic implementation of matrix refinement techniques, one remark about the implementation of score-based PWMs should be made. Assuming that each column of PWM equally contributes to the total matching score $\tilde{P}$ of sequence $S = s_1, ..., s_L$, the total matching score is defined as the sum of individual scores at each position in the sequence

$$\tilde{P}(S) = \sum_{j=1}^{L} w_{s_j j} \tag{2.1.2}$$

Here w_{s_jj} is a component of the j-column of matrix $\mathbf{W}$ that indicates the score of letter s_j. The denotation $\tilde{P}$ with a tilde for the score is reminiscent of its probabilistic analog (1.4.1), as they both play a similar role expressing the likelihood (or log-likelihood) of a putative TFBS. This analogy becomes more apparent when extra $\log w_{s_jj}$ is applied either in (2.1.2) or directly to the components of matrix $\mathbf{W}$ as we do in the remaining part of this thesis. Similarly, equation (2.1.2) can be extended for a dinucleotide PWM, but the summation would be only for $L-1$ dinucleotide positions in sequence S.

$$\tilde{P}^d(S) = \sum_{j=1}^{L-1} w^d_{s_jj}, \tag{2.1.3}$$

where index j now refers to the position of j-dinucleotide in the sequence S.

Thus, the additional natural attribute for predictive tools either in (2.1.2) or in (2.1.3) versions would be the matching score threshold parameter, which we must tune up. To complete an analogy in designing mono and dinucleotide PWMs we must make an additional i.i.d. assumption for dinucleotides and their positions. Figure 2.2 shows the workflow of detecting putative binding site candidates using the matching score (2.1.2) or (2.1.3). The general schema of PWM application does not depend on the type of PWM used.

2.2 Bucher's type PWM model with modifications

Pursuing one of our main goals to compute refined PWMs in this section we have to specify models and formulas used in our PWM refinement algorithm. Recall log-odds score matrices, firstly introduced by Bucher [15] Gershenzon *et al.* proposed a modified version which they used to compute mono and dinucleotide matrices, both designed from log-odds scores essentially defined in equation (2.1.1).

Following in general the approach of Gershenzon *et al.* in this thesis, we used the mean as an estimator of the nucleotide frequencies within the promoter area of

length 600nt including the TSS. Bucher [15] used the geometrical mean normalization for $\boldsymbol{\theta}_{0,j}$ within the motif length before he used them in the formula for weights $\boldsymbol{w}_j$, although the benefit of doing so has never been shown.

Since the core method [35] uses the log-ratio of occurrences, one additional PWM column-specific positive constant under the logarithm is ultimate: $\boldsymbol{\lambda}_j$ - a position-dependent fudge factor to avoid the logarithm to reach infinity. Although the necessity of such a constant is trivial computationally, the choice of particular $\boldsymbol{\lambda}_j$ is important, since when $\boldsymbol{\theta}_j$ equals zero, the value of $\boldsymbol{\lambda}_j$ defines the relative importance of different nucleotides. Since $\boldsymbol{\lambda}_j$ additively appears under the logarithm its role might relate to pseudo-counts. Different approaches were proposed to define the $\boldsymbol{\lambda}_j$ value [24].

The second constant (the translation constant) was chosen to render the column maximum as zero. After such a translation was done, an overrepresented nucleotide at a given position should have the score zero, similarly to the log-probability of 100% of a certain nucleotide at that position in the probabilistic model (equation (1.4.1)). Each column of such a hypothetical PWM has in theory the same co-domain for its components as the logarithm of base probabilities, but distribution of individual base scores might be different. So, as a result such a translation defines a common scale for score comparison.

2.2.1 The mononucleotide version of PWM model

The mononucleotide version of a PWM model with positionally dependent background compensation is a matrix of size $4 \times L$ with columns $\boldsymbol{w}_j$:

$$\begin{gathered}\boldsymbol{w}_j = ln\Big(\frac{\boldsymbol{\theta}_j}{\boldsymbol{\theta}_{0,j}} + \boldsymbol{\lambda}_j\Big) + \gamma_j \boldsymbol{I}; j \in [1, L], \\ \gamma_j = -max\Big\{ln\Big(\frac{\boldsymbol{\theta}_j}{\boldsymbol{\theta}_{0,j}} + \boldsymbol{\lambda}_j\Big)\Big\}, \\ \boldsymbol{\lambda}_j = \begin{cases} 0, & \text{if } \frac{\boldsymbol{\theta}_j}{\boldsymbol{\theta}_{0,j}} > 0.01 \cdot \bar{n}_j \boldsymbol{I} \\ 0.01 \cdot \bar{n}_j \boldsymbol{I} / \boldsymbol{\theta}_{0,j}, & \text{otherwise} \end{cases}\end{gathered} \tag{2.2.1}$$

where L is the length of motif, b is the index of a nucleotide from ordered set A, T, G, C, $\bar{n}_j = \sum_{b=1}^{4} n_{b,j}/4$ - the expected fraction of bases b, $\boldsymbol{I}$ is the identity vector.

Positionally-dependent constants $\boldsymbol{\lambda}_j$ and γ_j are included, as before, to redefine the logarithm and ensure that in each j-th column the maximum value is zero. Values $\boldsymbol{\lambda}_j$ are responsible for the position weights of very rare mononucleotides.

2.2.2 The dinucleotide version of PWM model

This type of matrix builds on the basis of 16 combinations of letters (AA, AT, AG, AC, TA, TT, TG, TC, GA, GT, GC, GG, CA, CT, CG, CC), unlike the mononucleotide matrix that uses single four letters: A, T, G, C. As in (2.2.1), the dinucleotide version of a PWM model with positionally dependent background compensation can

be presented as a matrix of size $16 \times (L-1)$ with columns $\boldsymbol{w}_j^d$

$$\begin{aligned} \boldsymbol{w}_j^d &= ln\Big(\frac{\boldsymbol{\theta}_j^d}{\boldsymbol{\theta}_{0,j}^d} + \boldsymbol{\lambda}_i^d\Big) + \gamma_j^d \boldsymbol{I}; j \in [1, L-1], \\ \gamma_j^d &= -max\Big\{ln\Big(\frac{\boldsymbol{\theta}_j^d}{\boldsymbol{\theta}_{0,j}^d} + \boldsymbol{\lambda}_j^d\Big)\Big\}, \\ \boldsymbol{\lambda}_j^d &= \begin{cases} 0, & \text{if } \frac{\boldsymbol{\theta}_j^d}{\boldsymbol{\theta}_{0,j}^d} > 0.01 \cdot \bar{n}_j^d \boldsymbol{I} \\ 0.01 \cdot \bar{n}_j^d \boldsymbol{I} / \boldsymbol{\theta}_{0,j}^d, & \text{otherwise} , \end{cases} \end{aligned} \tag{2.2.2}$$

where vectors have length 16 and $\bar{n}_j^d = \sum_{d=1}^{16} n_{d,j}/16$ - the expected fraction of dinucleotide d at dinucleotide position j. The frequencies $\boldsymbol{\theta}_j^d$ and $\boldsymbol{\theta}_{0,j}^d$ play he same role as the frequencies of mononucleotide version.

We chose in the software to count a local coordinate of dinucleotide in a sequence as the coordinate of its upstream (the first) nucleotide. The same rule used to address position of any other subsequence of length $k, k < L$ (named "k"-mer).

Positionally-dependent positive constants $\boldsymbol{\lambda}_j^d$ and γ_j^d are included as before to support the correct area of logarithm and to ensure that in each j-th column the maximum value is zero. Values $\boldsymbol{\lambda}_j^d$ are responsible for the position weights of very rare dinucleotides.

Dinucleotide matrices provide more reliable models considering the presence of pairwise dependencies at some positions in motif [16], which cannot be explained using only mononucleotide approximation of binding specificity[35]. However, while providing a better binding model, they (similarly to mononucleotide matrices) do not address the binding motif specificity for a variety of TFs. One could expect that a trinucleotide matrix would be even better than the dinucleotide matrix, but for such models we do not have sufficient amount of accurate TFBS data to make confident estimation of all 64 triplet frequencies. Moreover, some researchers suggest

that dinucleotide matrix is a better choice than a trinucleotide matrix because it reflects the geometry of DNA grooves (personal conversation with Dr. I. Ioshikhes).

2.3 Statistical evaluation of binding motifs

In this section we consider PWM **W** as a probabilistic model of binding motifs that encapsulates possible patterns of binding sequences for a certain TF. Searching for motif occurrences in a DNA sequence (such as illustrated in Figure 2.2, disregarding the mono- or dinicleotide PWM type) presumes the assignment of a PWM matching score cut-off parameter.

An inappropriately designated cut-off value will proceed either without matches or with an overwhelmingly large number of random hits. In this section, we evaluate the cut-off value t of weight matrix **W** linking it with a probability of random motif selection assuming that the background nucleotide distribution is known. The approach provided in this section is to construct a P-value function, which measures a statistical significance of a putative TFBS according to its score t.

Given t, the matrix **W** indicates the motif "occurrence" in the sequence S at the certain position on the promoter sequence, the matching score for this sequence $\tilde{P}(S) \geq t$. Therefore, in accordance with our PWM model of TFBSs prediction, we need to define the probability $P(t, \mathbf{W})$ for which the background model can randomly achieve the same or a better score than the cut-off value t. This will represent the P-value as the proportion of strings whose matching scores (compiled using matrix **W**) that is greater than a designated cut-off t.

The connection between the score cut-off and its significance can be conceived as two single problems. One is called the P-value problem: given t_0 we look for the corresponding P-value $P(t_0)$. The opposite problem (called the threshold problem: given a P-value $P \in [0, 1]$, find the cut-off) can be also of interest, but is beyond our current goal.

We assume that the background nucleotides can be described in terms of the probabilistic model which defines the probability of each nucleotide s_i within the search area of potential TFBS locations. Then, theoretically, for any given sequence S of length L with independent nucleotides s_j at any position j, the probability of observing a sequence S with score $\tilde{P}(S, \mathbf{W}) = t$ is calculated as

$$P(\tilde{P}(S, \mathbf{W}) = t|\mathbf{W}) = \sum_{\tilde{P}(S,\mathbf{W})=t} \prod_{i=1}^{L} P_i(s_i); S = s_i, ..., s_{i+L-1}, \tag{2.3.1}$$

where the summation is done all across such sequences S, for which the matching score $\tilde{P}(S, \mathbf{W}) = t$. Note that unlike equation (0.4.1) where the probabilities $P_j(s_j)$ come from PWM weights, here the probabilities $P_i(s_i)$ come from the non-uniform distribution of the background nucleotides.

The previous equation indicates that not all values t are accessible as the matching scores, assuming a certain $\mathbf{W}$. The simple reason of the inaccessibility is because t could be out of the matching score range for $\mathbf{W}$. Another reason is because any possible matching score counts as a sum of certain L components of matrix $\mathbf{W}$.

Thus, we consider that any t should be

$$t \in [\tilde{P}_{min}(\mathbf{W}), \tilde{P}_{max}(\mathbf{W})], \tag{2.3.2}$$

where

$$\tilde{P}_{min}(\mathbf{W}) \stackrel{\text{def}}{=} \sum_{i=1}^{L} min(w_{b_i i}); \tilde{P}_{max}(\mathbf{W}) \stackrel{\text{def}}{=} \sum_{i=1}^{L} max(w_{b_i i})$$

The problem of finding accessible cut-offs t will be resolved in the next section by dealing with only accessible scores. The number of such scores is a critical parameter of computational complexity of an algorithm. The proposed algorithm assumes that $\mathbf{W}$ has all integer components and follows the generic TRANSFAC PFM format (equation (0.4.5)).

To solve the P-value problem for matching score t_0, we need to find a probability of a set of sites, whose scores are above or equal to t_0 for any t_0 within the boundary defined in equation (1.3.2)

$$P(t_0) \stackrel{\text{def}}{=} P(S|\tilde{P}(S,\mathbf{W}) \geq t_0) \tag{2.3.3}$$

The P-value problem (2.3.3) is known to be an $\mathcal{NP}$-hard computational problem [1, 110], and so, a polynomial algorithm for its exact solution does not exist. Since the problem is very laborious, some approximation or heuristic algorithms must be considered.

2.4 Solution to the P-value problem using dynamic programming

A binding motif of length L, which is considered as being a very short biological sequence in the genome, is one of all 4^L possible sequences of equal length. Fortunately, it is up to 4^{20} sequences for longest motif length and thus a dynamic programming approach is a feasible solution if we can find a way to break the $\mathcal{NP}$-hard P-value problem down into simpler ones.

The following heuristic method takes inspiration from [5, 66, 110]. To implement this idea we assume that $\mathbf{W}$ has integer components such as the matrix PFM defined in (1.4.5). For matrices such as real-valued scoring PWM defined in equations (2.1.1) and (2.2.1) or (2.2.2), a rescaling to integer components is required. Averaging to nearest integers is not the best choice because the larger the scale, the more precise P-value estimates.

Before we can write an iterative process, let us define a sub-matrix $\mathbf{W}(b,k)$, $k \leq L$ that coincides with the first k columns of $\mathbf{W}$. Then we define a series of functions $Q_k(t)$ for all $k \in [0, L-1]$. These functions map the nucleotide scores in

W to the probabilities in a top-to-bottom way. The last $(L-1)$ probability function uses the whole set of **W** columns as

$$Q_{L-1}(t) \stackrel{\text{def}}{=} P(\tilde{P}(S, \mathbf{W}) = t|\mathbf{W}), \tag{2.4.1}$$

where the rigth-hand side of the equation resembles the left-hand side of (2.3.1) with highlighted dependence from the whole matrix **W**.

The remaining $Q_k(t)$ is defined as

$$\begin{gathered} Q_k(t) \stackrel{\text{def}}{=} \sum_b Q_{k-1}(t - \mathbf{W}(b,k))P(b), \\ Q_{-1}(t) \stackrel{\text{def}}{=} \begin{cases} 1, & \text{if } t = 0 \\ 0, & \text{otherwise} \end{cases} \end{gathered} \tag{2.4.2}$$

Values t for each k are selected within the interval

$$\tilde{P}_{min}(\mathbf{W}(k)) \leq t \leq \tilde{P}_{max}(\mathbf{W}(k)) \tag{2.4.3}$$

In this two-sided inequality, $\tilde{P}_{min}$ and $\tilde{P}_{max}$ values define minimal and maximal matching scores of sub-matrix $\mathbf{W}(b,k)$, similarly to those shown in equation (2.3.2).

Given k, function $Q_k(t)$ equals zero almost everywhere except the series of integers w_{ij} taken from $\mathbf{W}(b, k-1)$. At those values, function $Q_k(t)$ is equal to frequencies of corresponding bases.

Given the matching score t_0 and probability P defined in equation (2.3.3), the probability-value is obtained from the relation

$$P(t_0) = \sum_{t \geq t_0} Q_{L-1}(t) \tag{2.4.4}$$

Thus pursuing the goal to estimate P-value (2.4.4), we only need to compute

equations (2.4.1) and (2.4.2).

We implemented this approach as an *ad hoc* R program for two reasons. First, from our research we did not find any existing function from open source file depositories (such as CRAN, CPAN, MathWorks file exchange) to solve the PWM P-value problem that would run under the R or Matlab/Octave or Perl runtime environment. Second, we were looking for a tool to analyze the statistical significance of putative TFBS in this thesis project. The R-scripts written on the basis of dynamic programming are available on the attached CD. An example of the distribution of the matching score under the null hypothesis is shown in Figure 2.3; such a distribution is naturally used to compute the relevant P-value. Since the speed of a P-value algorithm does not matter due to a small problem size (short length of PWM) we do not consider any existing tricks or heuristics methods that could provide a better performance.

In support of matrix refinement, Figure 2.3B illustrates the intuitive expectation that closer neighboring matching scores of the refined PFM should have closer likelihoods. In opposite, Figure 2.3A shows the probability for non-refined PFM which is likely multi-modal and the relationships between scores and their probabilities are more ambiguous. The cumulative P-value distribution is usually used instead of exact P-value distribution. The former shows smoother performance and reliably corresponds to P-values with scores [66].

2.5 Matthews correlation coefficient as an optimization criterion

Unlike many existing implementations where this coefficient was used to estimate the quality of binary classifications, we implemented this coefficient following Gershenzon *et al.* as a part of the machine learning algorithm where it plays a role of an optimizing

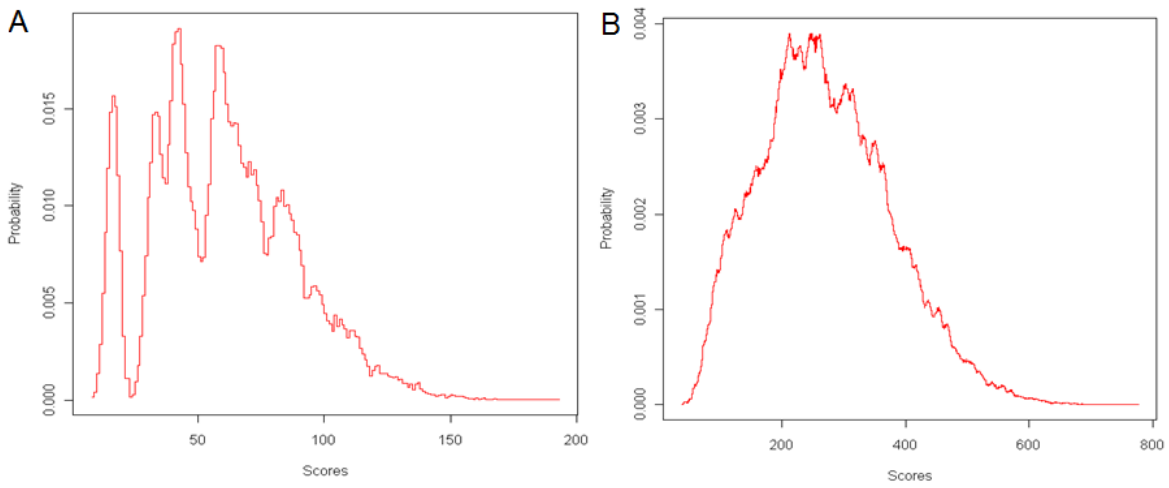

Figure 2.3: Exact P-value for matching score distribution for PFM matrices for TF Kr. A - the initial PFM (from TRANSFAC), B - the refined PFM mononucleotide version

criterion at each step of PWM refinement (see the flowchart in Figure 1.3) [35, 3].

Matthews correlation coefficient is generally regarded as being one of the best measures to represent true and false positives and negatives, describing a so called 2×2 confusion matrix using a single number. Similarly, Matthews coefficient is a convenient form to integrate characteristics of sensitivity and specificity. Sensitivity and specificity are defined by

$$Sens = \frac{TP}{TP + FN}; Spec = \frac{TN}{TN + FP}; \tag{2.5.1}$$

where TP, FP, TN and FN are the numbers of true positives, false positives, true negatives and false negatives respectively.

As shown by Baldi et al., Matthews correlation coefficient in essence is the Pearson's correlation coefficient between the observed and predicted binary classifications [3].

$$Cor = \frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TP + FN) \cdot (TN + FP)(TP + FP)(TN + FN)}}, \tag{2.5.2}$$

where Cor returns a value between -1 and $+1$ and the bigger the value the better the prediction.

Here is another interpretation of correlation coefficient. Cor, although non-linearly, is proportional to the area under the Receiver Operating Characteristic (ROC) curve, or AUC. It can be shown that

$$\begin{aligned} Cor &= \pm\sqrt{IM} = \pm\sqrt{(2AUC-1)M}, \\ M &\stackrel{\text{def}}{=} \tfrac{TP}{TP+FP} + \tfrac{TN}{TN+FN} - 1, \\ 2AUC - 1 &= I = \tfrac{TP}{TP+FN} + \tfrac{TN}{TN+FP} - 1 \end{aligned} \tag{2.5.3}$$

In the algorithm for finding similar motifs described in the section 3.2.3 Cor was computed with TP, FP, TN and FN which were defined at each step comparing hits produced by two matrices: those resulting from the previous and current iteration. The matrix preserved from the previous step is dubbed "old" and is considered as a "truth" matrix for the current iteration. (Note: We used the term "old" because we start from the initial, old matrix). The matrix employed at the current step plays the role of "new". So, TP is the number of sites composing the "old matrix" positively identified by the "new" matrix in the given functional window of length L; FP is the difference between the number of sites positively identified by the new matrix in the functional window at all considered promoter sequences and TP. FN is the difference between the number of sites positively identified by the old matrix in the given functional window and TP. TN is the complement of sum TP, FP and FN to the total expected number of all sites given interval of length L for the whole set of aligned promoter sequences.

Now that we have all the necessary tools in place, the remaining parts of the thesis apply the tools described above to the analysis of actual data. We will in particular concentrate on the performance of our new PWMs based on a heuristic refinement algorithm.

Chapter 3

A refinement technique to optimize PWM performance

3.1 Summary of achievements and of the road ahead

Bucher [15] introduced the idea of optimizing PWMs in two steps: (1) optimize the length of the binding motif in a pre-selected area and (2) optimize the PWM cut-off value; these two steps are reiterated until no further improvement is obtained. Bucher used this algorithm to predict four major eukaryotic promoter elements ("TATA"-box, cap-signal, "CCAAT"- and GC-box) for 502 vertebrate and non-vertebrate promoters from the EPD database. These four elements exhibit clear positional preferences for specific promoter regions and this fact facilitates the problem of PWM training.

Gershenzon *et al.* aimed to extend Bucher's idea so that the algorithm can detect weaker matching scores (signals) and elucidate more degenerated motif sequences. Two major changes they proposed include the concept of a functional window and the introduction of Matthews correlation coefficient as an optimization criterion instead of the overrepresentation parameter used by Bucher. The result of the new implemen-

tation was successfully tested on the GC-box and applied for the Sp1 TF. Another novelty of Gershenzon *et al.* was an application of the idea of PWM refinement to dinucleotide PWMs. Both algorithms were published as flowcharts (the Gershenzon algorithm is found in Figure 1.3), but are limited to the analysis of TFBSs that have very strong signals, with well-defined patterns such as the GC-box.

In this thesis we were motivated to revise the Gershenzon *et al.* version of PWM refinement to make it applicable to a larger number of PWMs regardless of the exact TFBS location. We also allowed variable motif length and variable quality of the binding signal.

3.2 Heuristic method of PWM refinement

An assessment of the abundance or rarity of TFBS S comprising L nucleotides can be estimated by comparing its observed frequency to that expected in a certain area on promoters. Unlike exhaustive sequence-based enumeration [58], the PWM approach to counting frequencies involves calculated z-scores for a number of sequence occurrences (hits) in that promoter area by applying the initial PWM $\mathbf{W}$ with certain matching cut-off c, together summarized as a suite model $< \mathbf{W}, c >$. This search involves several biological contents including the matrix specificity model and localization on promoters.

The implementation of the models for a large-scale application consists of the following three stages which might be iteratively repeated until a final solution is found.

1. Selection of binding sites and preprocessing: inspection of a representative list of TFBSs avoiding gaps and sequence replicates. If necessary, the preprocessing includes sequence alignment and truncation to a fixed length. The final stage of preprocessing is compiling the initial PWM which accounts for the background

nucleotide distribution.

2. Biological signal detection: preliminary scanning of promoter sequences to find potential areas of new TFBSs. This stage estimates the initial cut-off, cut-off boundary values and the functional window (see Figure 3.1).

3. Finding similar motifs in promoter sequences. This is an iterative processing for similar motifs in the functional window, using the range of PWM cut-offs; compiling the refined PWMs from discovered auxiliary new sites; optimizing the functional window, motif length, and cut-off. Matthews correlation coefficient *Cor* (equation (2.5.2)) evaluates monotonically until it reaches the maximum or when increase terminates (whichever comes first). The process is illustrated in Figure 3.1.

The first and second stages are the initiating stage followed by the core, third stage of the iterative matrix refinement process. The automated second and third stages are implemented in the original core MATLAB/Octave code not invoking any specialized toolboxes. A Perl script computes the PWM matrix using BioPerl package for sequence handling. All code is available in the attached CD.

3.2.1 Stage 1: Selection of binding sites and preprocessing

Since our software is designed to evaluate equally sized binding sites without gaps, preprocessing consists of mostly manual selection of binding sequences from the database. We preserved existing TFBSs alignment if they were published in the database. However, in cases of E74A, Ubx, Dl (Dorsal), MtTFA, DREF, the TFBSs were listed with different length and we performed our own alignments. For some TFs such as Ubx, available in TRANSFAC with 88 sites, we computed matrices from distinct common aligned parts trying to preserve most of the known binding specificity, but the remaining TFs were preprocessed as they were. The preprocessing stage ends with initial

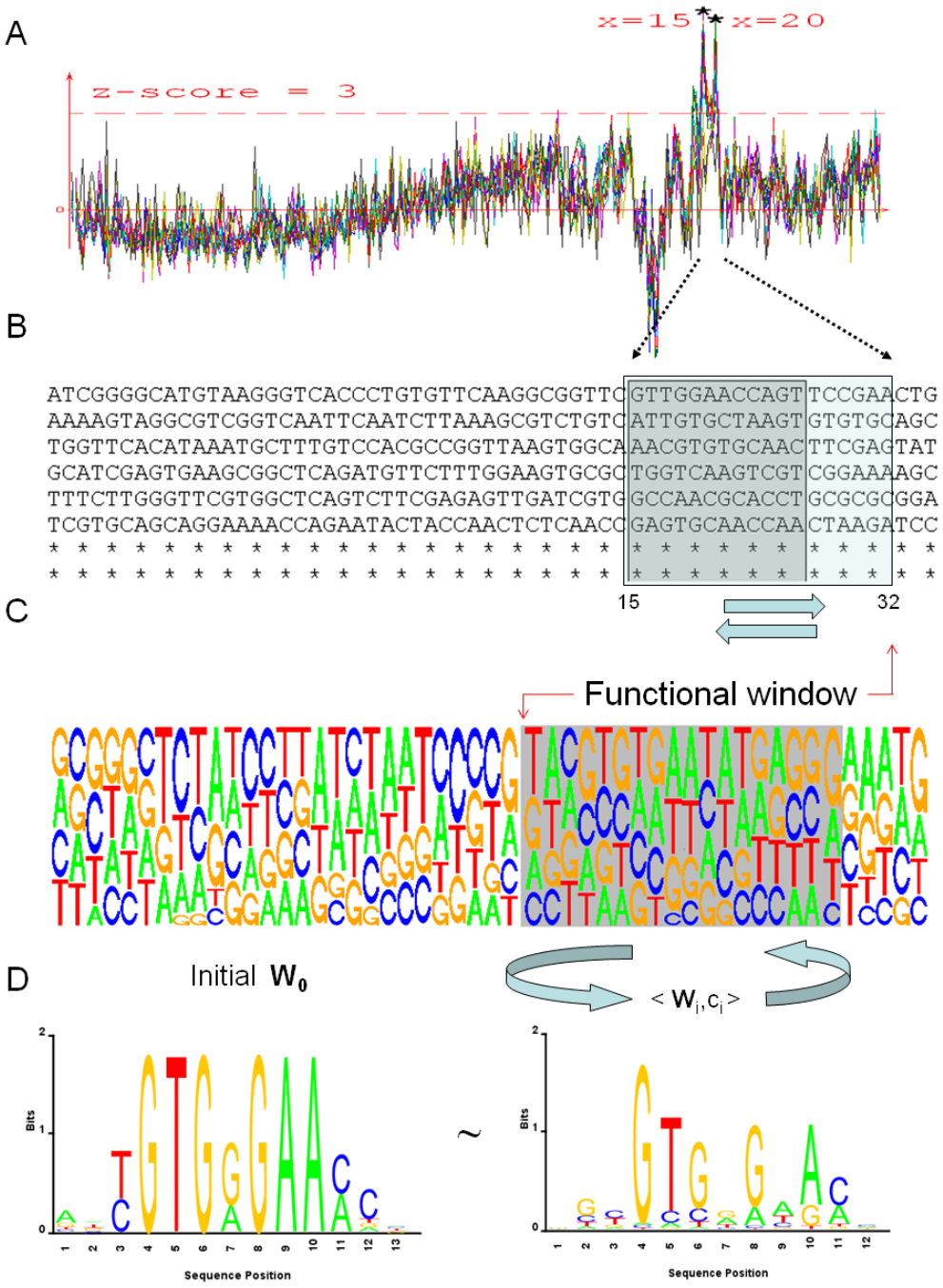

Figure 3.1: Illustration of the PWM refinement process for *Drosophila* SuH TF using a mononucleotide PWM. A: Biological signal detection: Scanning of promoters with ranged cut-offs; localization of functional window; B: A fragment of proximal area on promoters with irregular distribution of nucleotides; C: A fragment of promoters shown in terms of nucleotide frequencies; D: Searching for similar motifs in functional window for the optimized cut-off, the functional window and the motif length

mononucleotide and dinucleotide matrices that are subjected to further training for optimized performance.

3.2.2 Stage 2: Biological signal detection using z-scores

The PWM matching score, which is a standalone measure of overepresentation of potential TFBSs, results in a large number of random matches (see Introduction). To limit the number of false positive TFBSs, we used an additional criterion, the z-score, in order to (i) locate a functional window and (ii) estimate an initial PWM detection cut-off. We calculated z-scores (3.2.1) from number of hits detected above the cut-off based on equations (2.1.2) or (2.1.3), for the mono and dinucleotide case, respectively. The z-score was used as a measure of the overrepresentation of hits in a window of length L gradually sliding downstream the aligned promoter sequences.

$$z(S) = \frac{O_S - E_S}{\sigma(O_S)}, \tag{3.2.1}$$

where σ and E are symbols of standard deviation and expectation of the positional hit count, respectively. Z-score is a measure of standardized difference: observed O_S minus expected E_S over the standard deviation $\sigma(O_S)$. The expected frequency is estimated as the mean of all hits over all positions in the promoter set divided by the numbers of these positions and promoters.

An area of motif overrepresentation with $z \geq 3$ (a threshold established empirically in reminiscence of asymptomatically normal z-score distribution) on TSS-centered promoters with qualified outliers has been estimated as an initial functional window. The length of the initial functional window is L, which is subject to further optimization. Since the z-score is sensitive to cut-off value, which is unknown *a priori*, we performed simple analysis of outlier distribution around the expected functional window with a range of consecutive cut-offs to elucidate the initial matrix cut-off and the most comprehensible functional window (Figure 3.2).

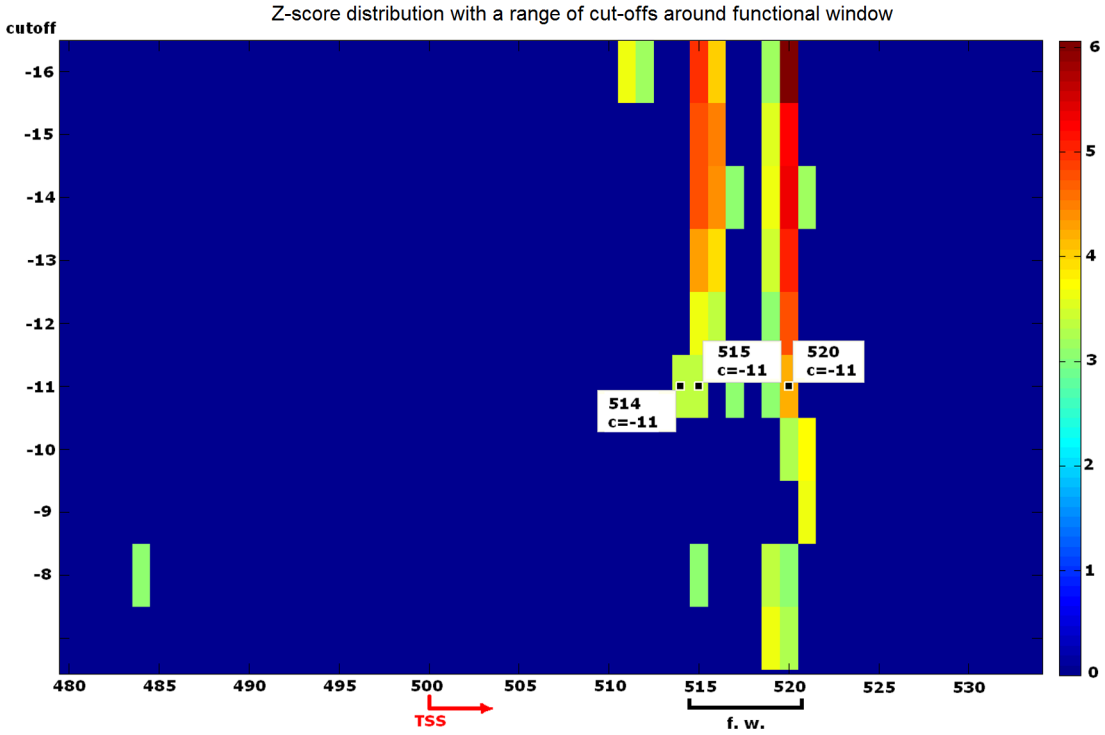

Figure 3.2: Demonstration of how the initial PWM cut-off c has been estimated. Z-score distribution on proximal promoter context of 1919 promoters produced by initial mononucleotide PWM for Su(H) TF. The colors show z-scores ≥ 3 around the TSS position of sliding windows. "Cutoff" shows cut-off values ranging from less (top) to most stringent (bottom). Only area proximal to TSS is shown. "f.w." indicates the functional window.

There is no one-to-one dependency between matrix cut-off values and true binding sites. While we expect that an accurate TFBS has a high z-score and is obtained under a stringent cut-off, in reality, a more degenerate site can be identified with a less stringent cut-off [103]. A small number of accurate sites used to compute an initial PWM might also cause a statistical bias in score estimates and the resulting cut-off value. Another reason why it is impossible to establish one-to-one dependency between matrix cut-off and matching score is because PWM matching score is an additive composition of constant weights of individual nucleotides (equation (2.1.2)). As a result it is impossible to reach any given cut-of value and there is no one-to-one dependency between matrix cut-off and matching score.

In such circumstances, applying a spectrum of cut-off values to filter corresponding z-scores helps us to find the location of most sustainable signals. As seen from the example in Figure 3.2, the z-score at nucleotide positions 515-516 and 519-520 has the magnitude ≥ 3 (selected threshold), which is sustained until a cut-off of -11. Because the distance between these two signals is smaller than the motif length (which is 13 nt), stage 2 fuses these positions into a single widow.

Some authors have suggested the z-score as a measure of statistical significance to be used as an instrument to search for putative binding motifs [97, 58, 85]. Unlike z-scores used to detect motifs our approach simply uses z-scores to detect overrepresentation (outliers on a promoter area with highest z-score). For this purpose we do not need to compare the z-scores from different areas. However, we could be able to associate with a z-score a P-value but that would be beyond our approach. For descriptive purpose we performed non-parametric Lilliefors (Kolmogorov-Smirnov) test of normality for z-scores of arbitrary TFs. The tests showed that the right tail of z-score distribution contributes mainly to deviation from normality as shown in Figure 3.3.

An accurate location of functional window is a big challenge due to two obstacles: A small number of TFBSs is available in the database and the short length of binding

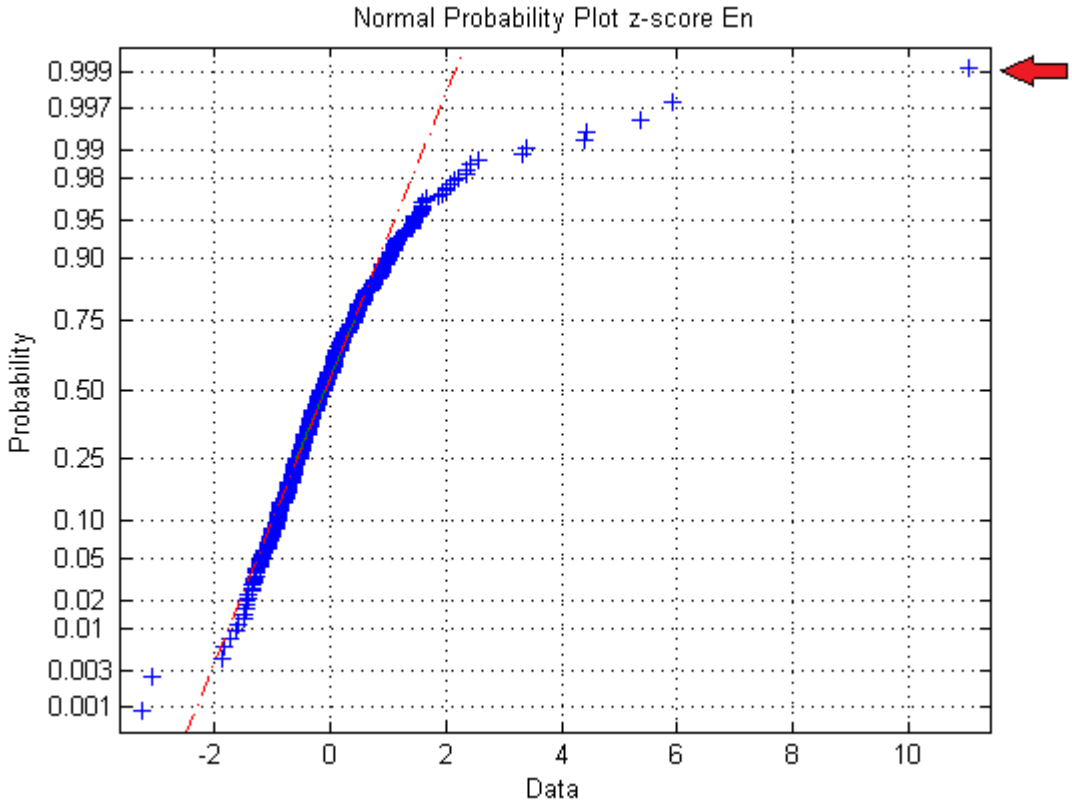

Figure 3.3: Q-Q plot for z-score values (showed as Data) distributed along the promoters and computed for TF En with the initial mononucleotide matrix. The maximum z-score (≥ 10) was selected as a signal indicator (see arrow).

motifs may result in a high number of false positive hits. In a case when we detected more than one functional window we selected the window which gives the maximum correlation Cor with respect to hits detected by the original PWM in this window and PWM after the first iteration.

We followed a similar heuristic to optimize the initial matrix cut-off, so that the selected initial cut-off generates high z-scores and does not predict too big a number of new putative binding sites (more than a few *per* promoter). An estimated functional window was used in our study to process both types of PWM matrices and to define optimized cut-offs, although initial cut-offs were estimated prior to the optimization process.

To be more specific, we found from our preliminary computational experiments, the estimation of functional window and setting the initial cut-off is crucial at the first step of iterative process (stage 3). The number of "similar" (measured by value of Cor) sites should be moderated. A very small number of mined sites leads to a trivial ($Cor = 1$) non-improved PWM. Empirically for all initial PWMs in this thesis (at stage 2) all the initial cut-offs, size and position of functional windows was tuned to follow up with a Cor value ranging between 0.66-0.85 before we transferred to stage 3. Following this rule we were able to further improve PWM through the steps of monotonically increasing Cor values until a maximum was reached for all factors from Table 3.2.

When optimization occurs in a global manner, less similar sequences often need to be considered in order to identify new putative binding sites in a larger sampling area. This process is fairly formalized because during training the number of discovered auxiliary sites depends on the variability of the sites found in previous optimization stages. The introduction of these heuristic rules made it possible to keep under control the number of additional sites found during the refinement process. These rules are one of the most important parts of our implementation of the heuristic algorithm, especially in the beginning with the initial set of accurate TFBSs.

3.2.3 Stage 3: Finding similar motifs in promoter sequences

The search for motifs is implemented in a completely automated routine. The method deals equally with mononucleotide and dinucleotide matrices and uses initial PWM within pre-estimated functional window and matrix cut-off. The goal of motif similarity search is to find new putative binding sites with maximum value of Cor. It starts from initial matrix/cut-off suite: $< \mathbf{W}_0, c_0 >$ and passes through the following steps:

1. Apply $< \mathbf{W}_0, c_0 >$ to search for new sites in a functional window.

2. Extend the initial set of sites with the new sites from step 1 and compute a new PWM matrix $\mathbf{W}_i$.

3. Apply matrix $\mathbf{W}_i$ with cut-offs values c on the grid points ranging from c_1 to c_2 and find $< \mathbf{W}_i, c_i >$ with maximum Cor for initial combination $< \mathbf{W}_0, c_0 >$.

4. Use the new $< \mathbf{W}_i, c_i >$ to optimize motif length L. For this purpose, apply $< \mathbf{W}_i, c_i >$ and find a functional window shorter and longer motifs applying +/- one nucleotide variation to original motif length. This step consists of four substeps used sequentially, all within the same iteration step:

   - Use $< \mathbf{W}_i, c_i >$ to find hits and take corresponding sites truncated from left (abbreviated as "cl") by one nucleotide. Compute $\mathbf{W}_{i_{cl}}$
   - Use $< \mathbf{W}_i, c_i >$ to find hits and take corresponding sites truncated from right (abbreviated as "cr") by one nucleotide. Compute $\mathbf{W}_{i_{cr}}$
   - Use $< \mathbf{W}_i, c_i >$ to find hits and take corresponding sites extended to left ("el") for one nucleotide. Compute $\mathbf{W}_{i_{el}}$
   - Use $< \mathbf{W}_i, c_i >$ to find hits and take corresponding sites extended to right ("er") for one nucleotide. Compute $\mathbf{W}_{i_{er}}$

5. The resulting matrix (with the shorter or longer length) should be applied to a functional window with variable cut-offs valued from c_1 to c_2 and find the one c which yields maximum Cor with the previous pair $< \mathbf{W}_i, c_i >$. Take this best as an initial matrix with cut-off value for the next optimization cycle, which resemble all the aforementioned steps beginning from step 1.

6. Reassign $< \mathbf{W}_i, c_i >$ as the new $< \mathbf{W}_0, c_0 >$, increment over i and repeat all aforementioned steps 1–5 until value Cor reaches the maximum or when increase terminates (whichever comes first).

The final result of this procedure is a PWM with optimized length l, and cut-off value c. The aforementioned refinement process is continued either until $Cor = 1$ first time in the cycle or it reaches its maximum (usually 3 to 8 cycles). Each cycle brings a portion of new putative TFBSs overrepresented in this particular window and excludes some non typical TFBS. Each cycle consequently increases the influence of "similar" sites from the functional window. This influence is strongly supported by the requirement of keeping the magnitude of correlation coefficient Cor at a higher level. For functional windows with equally big signals, all aforementioned steps are tested for each window and the window with the highest optimized Cor is selected (an example is shown in Figure 3.2).

In cases where the selection of a functional window is ambiguous, based on the observation of z-score spatial distribution on the promoters, we performed additional computations with variable window sizes around the area of potential signal and selected the window with maximum value of Cor for the initial $< \mathbf{W}_0, c_0 >$ to prevent a high number of noisy sites at the initial step. This sort of heuristic was used to estimate the initial cut-off value as well. Changing the length of the motif is a typical part of the optimization procedure and its processing was performed independently from those of the functional window.

3.3 Databases of gene regulation

The quality of the computational predictions of new TFBSs builds upon the quality and quantity of available binding specificity information. Selection of database resources was made in a very conservative manner based on previous work [35]. We had two reasons for making such a selection: (1) results should be comparable with previous implementations of the algorithm and (2) the most comprehensive and stable source of sequence specific information is required [35, 36]. Based on the described algorithm to construct PWMs we needed two different kinds of information - promoter sequences and TFBSs collections for as large as possible number of TFs.

To compute a collection of refined PWMs for *Drosophila*, two databases were used: TRANSFAC (`http://www.gene-regulation.com`, release 2009.4) and the Eukaryotic Promoter Database (EPD) (`http://epd.vital-it.ch`, release 105) [92, 71]. To be able to compute both mono and dinucleotide types of PWMs we utilized 33 TRANSFAC entries for *Drosophila* where both sequences and PFMs were available.

In addition to these two main resources, we also used other databases for the testing of our matrices and, occasionally, other resources for validation of discovered observations. In this section we discuss mainly the databases included in our regular framework for PWM training and PWM testing.

Following the terminology provided in machine learning we refer to the initial TFBS set used for PWM refinement as a training set and the other distinct TFBS set, where we tested the predictions, as a testing set [74].

JASPAR, used for PWM testing, is a database `http://jaspar.cgb.ki.se/` which contains a curated, non-redundant set of profiles, derived from published collections of experimentally defined TFBSs for eukaryotes [89]. The JASPAR collection provided us with an extra challenge when used for PWM testing of predictability because most of the sequences are not present in TRANSFAC and have longer binding motifs.

Due to differences between sequence contents of TRANSFAC and JASPAR, the former was mapped against JASPAR's core collection. For each of the 33 TF entries from TRANSFAC we tried to find the corresponding TF from JASPAR core collection (using matrix accession code). Rare cases of ambiguity in names were resolved using NCBI official gene names and their aliases. We found 14 common TFs and used them to construct our testing sets, as described in subsection 3.3.1.

Testing the predictive power of our new matrices presumes a generic computational tool to compare results. We used Match^{TM} which is provided with TRANSFAC for this purpose [54]. An additional benefit of using Match^{TM} was a comprehensive access to whole capacity library of mononucleotide matrices provided with the commercial version of TRANSFAC we used, TRANSFAC professional.

Finally, our PWM training included the Eukaryotic Promoter Database (EPD) of experimentally characterized eukaryotic polymerase II promoters, a high quality source of biological content with experimentally determined TSSs. The EPD database was employed as a source of biologically confirmed information to search for new statistically significant binding motif sequences. The 1919 EPD promoter sequences for *Drosophila*, each 600bp long (-499 to +100 bp aligned and centered on the position of a TSS), were used in two ways: (1) as a target set to mine new putative TFBSs (these sequences were scanned during PWM refinement), (2) EPD sequences were reshuffled to simulate randomized background sequences used in PWM definition [35]. Promoters containing unknown nucleotides (denoted as "n" or "N") were excluded from these computational experiments and were not included in those 1919. The length of proximal to TSSs areas was selected to include potential location of all target TFs selected for the study.

3.3.1 A structure of synthetic tests from JASPAR data

A synthetic test sequence was constructed for each TF tested (14 in total) and provided as a testing set. Each JASPAR sequence candidate for this testing was assessed for its non-redundancy against the corresponding TRANSFAC sequences. We used non-trimmed JASPAR original sequences as carrying relevant biological content, regardless of the fact that the length could be variable within a TF entry. The original JASPAR alignment of these sequences (shown in red in Table 3.1) was preserved. Then, to embed the aligned and non-trimmed sequences into a constant-length promoters (which was 600nt for all tests, not shown in Table 3.1) we extended the sequences with random flanks (shown as "n" in Table 3.1).

This design automatically identified the position of the first upstream nucleotide in an aligned part that facilitated matching hit counts (finding true positives). An area identified by the longest left and right flanks (marked green in Table 3.1) is the area where we could expect an irregular background nucleotide distribution due to the irregularity of nucleotide content on original flanks. To capture this background information, we used an option to reshuffle the nucleotide content at each position in these areas (up and downstream from the aligned red parts).

The most important benefit of such a synthetic promoter data set is an opportunity to construct more realistic promoter models that preserve the background nucleotide distribution.

3.3.2 A review of input information

In this section we highlight the main differences between sequence information that was retrieved from TRANSFAC and an input information of PWM refinement algorithm.

Table 3.2 reviews the collection of TFs from *Drosophila* obtained from the TRANSFAC (release 2009.4) and the corresponding PWMs used in the thesis. The last three

Table 3.1: A fragment of the synthetic data test example constructed from JASPAR sequences distinct from those used in the TRANSFAC matrix. Alignment of JASPAR sites is highlighted in red, and start of the alignment is centered at a fixed known position throughout the current promoter set at position 27 in this example. Equally distributed random nucleotides "n" extend each promoter sequence to achieve equal length.

nnnnnnnnnnnnnnnnnnnnnnnnngTGTAACTGTctttctaccgccaaggcctctg
nnnnnnnnnnnnnnnnccagctggtaTGTAACCGTagtacacctgagcnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnntgaCGTAACAGTgagcataagggggcaattggn
nnnnnnnnnnnnnnnnnnnnagtccaAGTAACTGCaacacctttagaccactnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnttAGAAACAGTggtagtgttggggacctgaacn
nnnnnnnnnnnnnnnnntcgtcccacTGAAACAGTaccggggctttgtgnnnnnnnn
nnnnnnnngcgtccccctttaccaaaACTAACTGTccccannnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnccgactcattAGTAACGGActcccatacgcggnnnnnnnnn
nnnnnnnnnnnnnnnnnnnctcgtgtAGCAACTCGgagaaccttaacagatnnnnnn
nnnnnnnnnnnggcataggaaattccTATAACTGTgcagatccnnnnnnnnnnnnnn
nnnnnnnnnnggttatggtttaggtcATTAACAGCgagcggtnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnnnnnncTGTTACTGTaatccgggccggactggannnn
nnnnnnnnnnnnnnnnnnnnnnnnntTCGAACAGTtgagatctcctaggggcccatg
nnnnnnnnnnnnggtctgttcagtccTCTAACGGCcatttgtacnnnnnnnnnnnnn
nnnnncaggactgttcgctagctcttAGAAACCGActnnnnnnnnnnnnnnnnnnnn
gtaggctactaggctactaggctcatAGAAACCGAccnnnnnnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnngccgtaatAGCAACCGAcacgagatctttcgannnnnnn
nnnnnnnnnnnnctcgcatcccttatCGTAACCGGtccacagnnnnnnnnnnnnnnn
nnnnnnnnnnnnnnnnnnnnncttcgACGAACAGCacacccccaatagccgcannnn
nnnngtcgtgtagcaactatgacaacCTTAACAGAtnnnnnnnnnnnnnnnnnnnnn
nnnnncaccagcaccccctaatcagaCGTAACGGGccnnnnnnnnnnnnnnnnnnnn

columns of Table 3.2 contain numbers separated by colons. Numbers before and after colons show numbers of sites and their length, respectively. The third column presents this information for TFBSs which were selected for our analysis from all available TFBSs (third column). The names of TFs are provided in the second column.

As evident from Table 3.2, TFBSs for the corresponding TFs have the motif length between 7 and 23 nucleotides ($mean = 11.3$, $st.dev = 4$). Several entries in the table have the same names which come with different TRANSFAC accession codes (first column). That means that those TFs are represented in the database by more than one collection of TFBSs. In the case of TF Sn, we could align TFBSs in two different ways and process with both versions. Ambiguous entries of Table 3.2 have not been considered in the performance analysis.

At the time of our study, the quantitative information about confirmed TF-DNA-binding specificities was limited. As an example, Table 3.2 shows that most of the *Drosophila* TFs are represented (in TRANSFAC) by a small number of available binding sites ($mean = 18.1$; $st.dev = 15.78$ for column (3) and $mean = 15.7$; $st.dev = 10.4$ for aligned and used, column (4)). Many TF entries in TRANSFAC have binding information represented exclusively as PFM and excluded from our analysis and not shown in Table 3.2.

Some TFs (such as Ovo, Kr, Sn) originally had two matrices and we optimized them both. For Dorsal (Dl) TF, we excluded the second entry M00043 from the analysis because the alignment of 22 sites appears to involve the sixth unidentified position and this ambiguity may be functional, because as suggested in [69] where binding sequences for Dl conform to two degenerate sequences: $GGG(W)_4CCM$ and $GGGWDWWWCCM$ (IUPAC nucleotide code for "W" is A or T; "D" is any but "C"; "M" is A or C). As it is the case in most typical PWM applications, we could not address gaps and variable lengths in sequences in the current version of the algorithm.

Table 3.2: Summary of TFBSs from TRANSFAC with the addition of putative sites, predicted using PWMs refined with our approach. Accession numbers correspond to TRANSFAC; "Avail." and "Used" provide the number of sites published and used for matrix training; the numbers in front and after colons correspond to the numbers of predicted sites and their length L, respectively, for mononucleotide and dinucleotide PWMs refined using our method with an optimized cut-off.

Matrix (1)	TF name (2)	Avail. (3)	Used:L (4)	Mono found:L (5)	Di found:L (6)
M01083	Abd A	40	40:10	209:8	801:8
M01094	Abd B	7	7:7	39:7	155:6
M00171	Adf-1	7	7:21	15:21	69:20
M01095	AP	14	14:8	768:7	232:8
M01096	Brk	10	10:7	57:8	51:8
M01097	CAD	13	13:10	408:9	41:9
M01087	CEBP	12	12:23	27:22	26:22
M00120	dl	13	13:11	3:10	28:12
M00488	DREF	10	10:13	60:12	13:12
M00110	Elf1	5	5:16	166:13	8:16
M00696	En	11	11:7	94:10	56:9
M00396	En-1	10	10:7	84:7	142:8
M00020	Ftz	9	9:12	26:14	95:11
M00022	Hb	16	16:10	1082:9	341:11
M00021	Kr	6	6:10	427:9	21:9
M01089	Kr	30	30:12	34:10	29:10
M01090	Mad	9	9:8	83:7	152:8
M00487	mtTFA	11	9:10	262:9	25:10
M00461	Ovo	21	21:15	17:12	28:13
M01101	Ovo	9	9:8	791:7	36:8
M01091	PRD	9	9:7	115:7	92:8
M01102	SD	19	19:7	518:7	277:6
M01098	CF1A	11	11:16	135:15	94:11
M00112	CF1	38	38:9	26:9	78:10
M00044	Sn	9	9:14	28:14	42:11
M00060	Sn	40	12:13	253:10	18:12
M00060	Sn	40	22:10	36:11	58:9
M00666	Sry beta	5	5:9	44:11	57:11
M00234	Su(H)	10	10:13	68:12	118:12
M01092	TCF	25	25:16	89:13	16:15
M00679	Tll	10	10:8	44:7	355:6
M01103	TWI	11	11:14	24:12	8:14
M00018	Ubx	88	48:10	1476:5	535:7
M00283	Z	24	24:11	92:9	31:11
M01099	KNI	32	32:18	91:17	35:15
M00016	E74A	14	14:12	120:13	109:14
M01088	DEAF1	22	22:7	284:7	130:7

88 TFBSs for Ubx have a very short common length and we were not able to use this set as a whole. To preserve as many nucleotides as possible from sequences published for Ubx and MtTFA transcription factors, only the longest available binding sites were aligned and considered. For Ubx, it was argued that $TAAT$ bases (or$ATTA$ from reverse complementary strand) play a primary role in determining the affinity of binding [30]. Therefore, we tried to select sites with $TAAT/ATTA$ patterns with flanks, which were confirmed in the later publication [29].

The transcription factor Sn (Snail, TRANSFAC entry M00060 and M00044) was present in two data sets with a slightly different consensus and a different motif length. In this case we could not make an ultimate decision when choosing the training set and decided to follow up with version M00060 with more sites, although M00060 is presented in two versions: first, with 22 sites of length 10, and second, with 12 sites of length 13. As an alternative to our solution, we could join M00060 and M00044 into one set, but decided to preserve reference to the TRANSFAC accession code.

Although PWMs may bias the comprehensiveness of respective experimental data, PWM refinement technique designed to deliver optimized predictive performance lessen this dependence. Our approach is independent from the experimental technology used to obtain the binding specificity data and from the validation method. Alternative TFBSs collections not published in TRANSFAC (such as JASPAR, ChIP-Seq-derived TFBSs or other relevant data sets) can be also used to compare results.

3.3.3 Advantages of TRANSFAC data over ChIP experiments

At the time of this study we asked about an alternative source of binding information other than TRANSFAC to provide us with an access to a richer number of available TFBSs. The advent of the high throughput sequencing data was the reason of such an inquiry. In addition to what we have already said at the beginning of this section,

we list below are some important remarks in favor of TRANSFAC.

We chose to use the TRANSFAC database because it assured an expert-curated experimentally confirmed set of TFs with corresponding TFBS entries. In addition, TRANSFAC is well-established in the field as a reference set for *in silico* TFBS determination and other genomic analyses.

The ChIP-seq technology is currently seen primarily as a higher resolution alternative to ChIP-chip, which requires microarrays for hybridization and therefore is restricted to a fixed number of probes [52]. Moreover, the ChIP technology itself is highly dependent on the quality of the antibody and does not work equally well for all TFs. In addition, although ChIP coupled with sequencing (ChIP-seq) is thought to exhibit less bias (such as regional biases along the genome) than ChIP-chip, the potential bias imposed by different sequencing platforms is not completely understood.

Computational methods that detect weak protein-binding signals in ChIP experiments while maintaining appropriately high specificity remain challenging [21]. As a result, an alternate confirmation of functional relevancy is a necessary step in any ChIP experiment [51]. *In silico* approaches to determining binding specificity are far less expensive, although they are limited to identifying new instances of TFBSs for TFs with experimentally validated DNA binding data. Since *in situ* DNA-binding data guarantee neither absence of noise nor binding completeness for certain TF, for training set, we sought data from a high quality source, which would combine different validation strategies and literature evidence.

3.4 An outline of results for refined PWMs

We optimized 33 TF entries (TRANSFAC accession codes shown in Table 3.2) for their PWMs in both mononucleotide and dinucleotide versions. Refinement of each new matrix in terms of Matthews correlation coefficient Cor was done for the parameters

such as motif length, location on the promoters and PWM matching cut-off.

We show in this section that improved matrices generated by the proposed refinement technique perform better than conventional TRANSFAC matrices (called initial PWMs). For this analysis, we used a summary of predictions on synthetic tests (Table 3.3) and a measure of information content presented as sequence Logos for discovered putative TFBSs (Figure 3.4). A more comprehensive analysis, including comparison with Match^{TM} tool and validation of a biological content of some of our findings, is performed in the following chapter.

The length of the initial matrices is equal to the common length of aligned TFBSs. TRANSFAC sites in most cases have already been aligned in the database. However, the resulting mononucleotide PWMs did not necessarily preserve the initial lengths. This happens because the matrix refinement procedure offers the motif length to find matches with largest Cor. For this reason, we can observe shifts in positioning toward the most conserved nucleotides within the initial sequence, which accounts for the irregularity of background nucleotide distribution on promoter sequences used for training. In particular, for Brk, Ftz, Dl, En, Hb, one or both optimized matrices have longer lengths than the initial sequences.

As seen from Table 3.2, for each particular TF, the number of discovered sites was not larger than the total number of promoter sequences used for PWM training. Employing an arbitrary lower cut-off value with the original PWM would apparently produce more matches, most of which would be false positives. Low specificity is also a known issue, that hampers the application of conventional PWMs for TFBS prediction [109]. Remarkably, using optimized PWM with optimized cut-offs a number of matches automatically (without additional cut-off control or filtering) gives an average number of matches ≤ 1 *per* promoter area used in the study.

Taking into account our limited knowledge, some optimized matrices can be expected with better or worse (when compared to the original PWMs) specificity or sensitivity in a single analysis. Someone who is searching for a tool involving many

PWMs can benefit from better averaged total performance of optimized matrices. In particular, result of the comparison presented in Table 3.3 (page 73) provides extra insight into the differences between optimized mono- and dinucleotide matrices when they are applied to a completely new promoter content, that was not used in matrix training. In the next paragraph we will give more comprehensive quantitative analysis of PWM tests, but here we provide a big picture of the predictive performance of our method.

The following selected result comes from further analysis of synthetic tests constructed from biological binding sequences taken from JASPAR (Table 4.3). Since TFs have different numbers of TP and FP matches (the columns "TP" and "FP" in Table 3.3 reproduce the rows "TP" and "FP" in Table 4.3) we characterized them as proportions of all predicted TP (or FP). All true and false positive matches discovered in JASPAR synthetic tests in Table 3.3 were summarized in each of three target groups by the type of matrix: In total, the initial matrices give 77 true positives against 127 false positives; optimized mono matrices give 123 true positives against 73 false positives; finally, optimized dinucleotide matrices give 106 true positives against 76 false positives. Because both optimized mono- and dinucleotide matrices were tested on the same testing sets, we can easily calculate the percentage of improvement when using optimized matrices against the initial matrices calculated from TRANSFAC.

The total number of TPs improved (increased) by 59.7% (out of total 77) for the mononucleotide matrices and by 37.7% for the dinucleotide matrices compared to the result from initial matrices used in TRANSFAC. Similarly, for the number of FPs a global picture comes as follows: for the mononucleotide PWMs the total number of FPs decreased by 42.5% (out of total 127) and for the dinucleotide matrices it decreased by 40.2%. The comparison was done with a fixed cutoff of the corresponding optimized matrix. More comparisons of the results from Table 4.3 using statistical measures are provided in the following subsection (a summary is shown in Table 4.3 and in Figure 4.3).

Table 3.3: Summary from Table 4.3 with additions about performance of initial PWMs on synthetic tests constructed from JASPAR sites.

TF names	JASPAR test files	Initial PWM		Opt mono PWM		Opt di PWM		Test size
		TP	FP	TP	FP	TP	FP	
Abd-A	MA0206.1.sites	3	1	4	3	8	20	23
AP Q6	MA0209.1.sites	1	3	0	3	0	0	20
DEAF1	MA0185.1.sites	3	11	1	3	2	7	10
E74A	MA0026.1.sites	11	0	11	0	0	0	17
En	MA0220.1.sites	0	20	4	7	8	0	23
Hb	MA0049.1.sites	15	11	16	31	11	17	16
KNI	MA0451.1.sites	11	0	2	0	0	0	26
KR Q6	MA0452.1.sites	19	7	15	0	11	0	31
PRD	MA0239.1.sites	0	7	2	7	1	5	37
Sn	MA0086.1.sites	0	0	2	0	0	0	40
SuH	MA0085.1.sites	10	0	10	0	10	0	10
Tll	MA0459.1.sites	2	33	29	1	31	14	34
Ubx	MA0094.2.sites	2	18	14	17	14	13	20
Z	MA0255.1.sites	0	16	13	1	10	0	41
TOTAL:		77	127	123	73	106	76	348

Sequence Logos in Figure 3.4 show how well nucleotide content is conserved in the putative binding sequences found here. The sequence Logo, constructed from equation (1.5.1), shows not only the information content in bits but also determines the consensus sequence and relative frequency of bases at every position in a binding site[93]. For reference, the Logo of the initial TRANSFAC sequences was included in the first column. Putative TFBSs newly discovered by mono- and dinucleotide optimized matrices with optimal cut-off are placed in the second and third columns respectively. PWM refinement method is not an exhaustive method for identifying the overrepresented motifs in promoter sequences. Figure 3.4 emphasizes similarity of consensus patterns visible at top nucleotide, which is an indicator of overrepresented patterns. Sequence Logos for the remaining TFs are placed in the appendix (Figures 5.2-5.6).

The prevalence of the most conserved nucleotides at the top of the Logos (Figure

3.4) for some TFs does not coincide with the nucleotide order presented in the initial consensus. The first two or rarely three top nucleotides sometimes are permuted or changed. For instance, that happens at positions 5 and 11 in Z, positions 1, 4, 7 in En, and position 1 in Abd-B. (positions numbered in left-to-right order as shown in Figure 3.4 with reference to position numbers in column A). This variability might indicate those nucleotide positions with significant degeneracy. Other examples can be found in sequence Logos from the appendix (Figures 5.2-5.6), which also includes Logos of discovered sites for other TFs. From such examples we can see that often the initial consensus pattern was reproduced at the top nucleotides although with fewer entropy. The positions where this happened should be considered more stable than positions where swapping has occurred. More details on this matter are provided in the discussion section.

In addition to that, our results (Figure 3.4) shows that some nucleotide positions in putative sites appeared to be more conserved by the information content [93] than expected from initial sequences, for instance: Dl at positions 4, 5, 6; en at positions 1, 2, 7; Brk at position 2, and Abd-B at positions 2, 4, 6 (all positions related to the initial sequence consensus starting from left to right).

Inspection of the sequence Logos from predicted sequences on the aligned promoters shows that sequences, discovered on the promoters where dinucleotide matrices were applied, are in general more similar to those used in training and thus they are likely biased by the content of initial sequences. This can be seen when comparing top nucleotides from most respective Logos in columns C to A vs. B to A in Figure 3.4 with the exception of Dl, which might demand more special investigation (as mentioned for entry M00043). We suggest (without estimating the quantitative effect, which is beyond the current study) that this visible property might result from a stronger selective pressure when similar sequences were searched in the training process. In particular, the matching score takes into account the co-occurrence of neighboring nucleotides during dinucleotide matrix optimization.

Engrailed (En) is a homeobox TF that plays an important role in *Drosophila* segmentation. Using a mononucleotide optimized matrix, we detected En putative binding sites upstream of the TSS in several promoters. Computational discovery of these sites using PWM is a challenging task because at least three nucleotides in the consensus En binding motif coincide exactly with those in the following overlapping sequence $TCAGT$, also known to be one of the overrepresented sequences in *Drosophila* genome, mostly at Initiator (Inr) sites [23]. Both optimized matrices recognized noisy matching signal from En TFBS upstream of the Inr site, in agreement with prior work [81]. Comparing dinucleotide with mononucleotide PWM versions, one may see that "G" is absent in putative binding sites obtained by the dinucleotide matrix. Figure 3.4 shows the routine dependency analysis between mono and dinucleotide occurrence frequencies for En TF. This figure shows that adjacent nucleotide positions are mostly dependent and hence the dinucleotide matrix predicts them better.

From these results we see that the refined PWMs have a better predictive power of new TFBSs than the initial matrices, despite of the differences in prediction of TFBSs for individual PWMs. This analysis addresses the hypothetical question: "What kind of general improvements can we expect just from optimizing PWMs on available sequence content?" In this "proof of concept analysis" we used the same definitions of all PWM types, particularly, the background nucleotide distribution was estimated on the same promoter sequences. Additional insight on the refined PWMs will be given using Match$^{\text{TM}}$, which uses different types of PWMs.

Besides the better predictability on synthetic tests, the refined PWMs with optimized motif length and cut-off show biological relevancy on real promoter sequences.

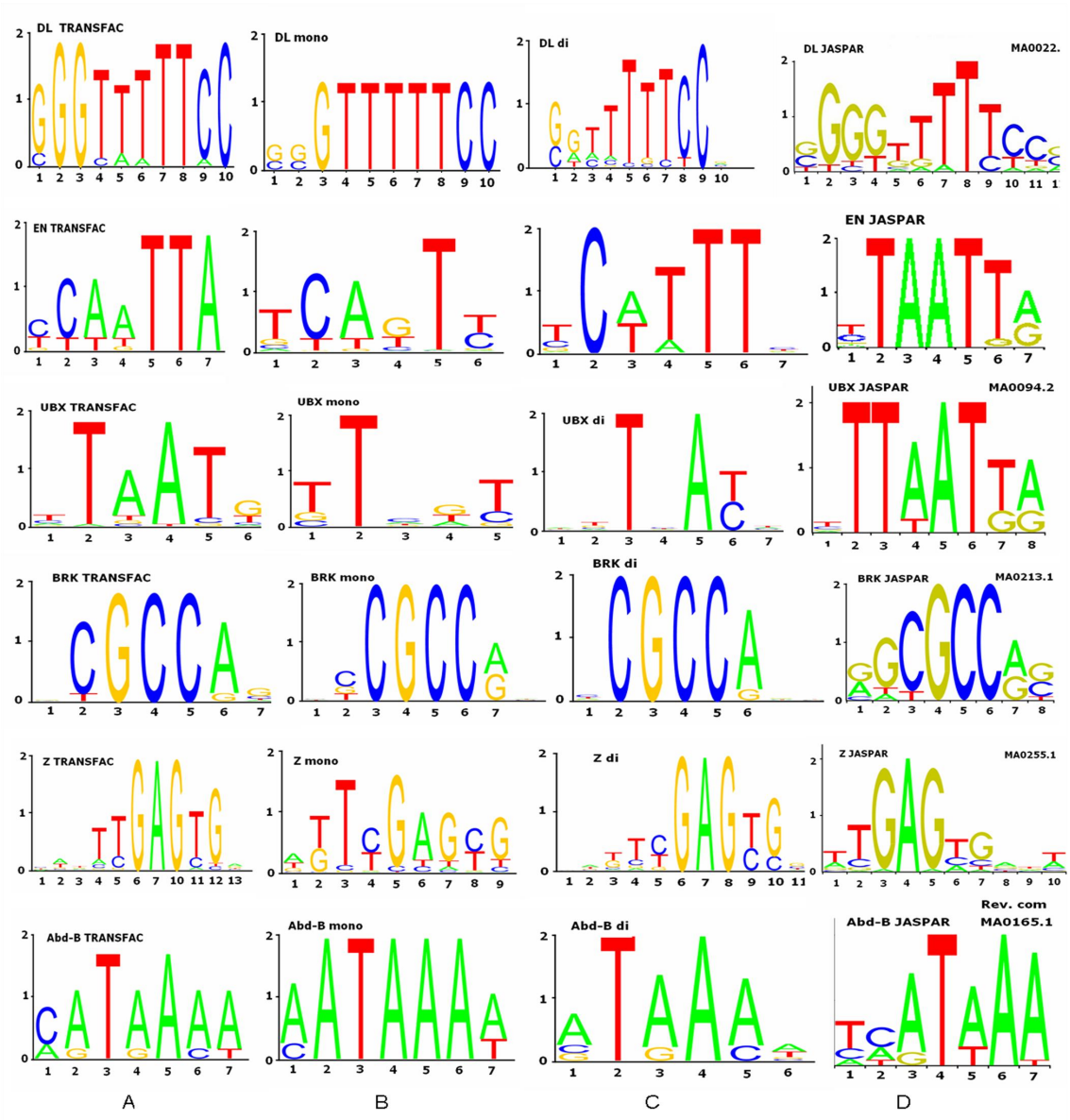

Figure 3.4: Sequence Logos for six TFs discussed in text. Highs of letters distributed according to the measure of information content of new sites found by the optimized matrices compared to TRANSFAC and JASPAR sites. (A) Initial TRANSFAC sites. (B) From new discovered sites (mononucleotide matrix). (C) From new discovered sites (dinucleotide matrix). (D) From JASPAR CORE collection. Remaining sequence Logos are showed in appendix in Figures 5.2-5.6.

Chapter 4

Refined PWMs outperform Match^{TM} on synthetic and PBM data

4.1 Refined PWMs are sensitive search tools, supported by the literature

In silico identification of TFBSs is more cost-effective than laboratory methods. However, it requires independent confirmations of the results by alternative approaches. To investigate whether motifs found by our analysis were bilogically relevant, we used the UniPROBE database, which is generated by protein binding microarrays (PBM) for a range of proteins from different organisms [80, 7]. Consequently the UniPROBE database is relatively accurate collection of data on TF-DNA binding.

Each entry for a protein in UniPROBE provides quantitative preferences for all possible nucleotide sequence variants ("words") of length k (called "k-mers"). The modest number of proteins collected in UniPROBE (currently 406) prevents us from performing a comprehensive annotation of all putative TFBSs identified using

refined PWMs. Thus, we used UniPROBE for illustrative purposes to estimate the capabilities of our optimized PWMs to predict new biologically meaningful TFBSs. As such we were able to map some of our putative binding sites to known orthologous TFs that had similar DNA binding preferences based on the data in UniPROBE.

For this analysis we selected non-redundant putative TFBSs of three TFs: Ubx, En and TCF which cleared from redundant sites (further we refer to selected sites as PWM-sites for brevity). Selection of these three TFs was established arbitrarily. Ubx and En are homeodomain proteins and homeodomains have conserved amino acid sequences across most eukaryotic organisms. This high degree of conservation makes them an ideal system for studies attempting to elucidate specific protein-DNA interactions. The third TF was TCF (Pangolin) with High Mobility Group box (HMG-box) sequence specific binding domain. Given the evolutionary conservation of both binding sites and protein sequences, we reasoned that these three sets of PWM-sites would be similar to known motifs for other binding proteins with similar DNA binding domain structures, and that finding such a similarity would provide additional support that matched new PWM-sites are likely to represent genuine TFBSs.

At the time of the study, UniPROBE contained non-redundant DNA binding proteins from a diverse collection of organisms, but excluding *Drosophila*. For this reason we were not able to look for exact matches with *Drosophila* TFs.

We used the UniPROBE search tool available at `http://the_brain.bwh.harvard.edu/uniprobe/` with default settings and queried putative PWM-sites for Ubx, En and TCF on the promoter data set to find TFs that bind similar DNA (oligonucleotide) motifs in other organisms. The upper boundary for E-values was set to 0.001. This upper boundary was selected based on a range of E-values obtained from queried TFBSs used to construct the original PWMs for all three TFs. The E-value is utilized not as a criterion for statistical inference but rather as a ranking tool to identify protein with a known binding site most similar to the submitted TFBS. In total, we queried UniPROBE with 296, 39 and 16 putative non-redundant TFBSs

Table 4.1: Three examples of UniPROBE queries with putative PWM-sites, which were discovered using optimized PWMs. Only the best matches shown for three tested TF Ubx, En and TCF.
Abbreviations: "Ce" means *C. elegans*; "Sc" means *Saccharomyces cerevisiae*; "Mm" means *Mus musculus*; "Hs" means *Homo sapiens*.

Ubx			En			TCF Q6		
Similar motif	organism	best E-val	Similar motif	organism	best E-val	Similar motif	organism	best E-val
Hoxa6	Mm	0.003571	CEH-22	Ce	0.000305	Sox-4	Hs	0.008115
Tea1	Sc		Tea1	Sc				

for Ubx, En and TCF, respectively. The numbers correspond to the total number of detected non-redundant PWM-sites for these TFs. The complete list of sites can be found in "Optimized TRANSFAC" folder on the attached CD.

In the remaining part of this section we investigate the similarity between all top UniPROBE matches to the three TFs of interest. To reach this goal, we use a conventional sequence alignment tool with graphical representation of aligned amino acids sequences to indicate most conserved areas, and the location of protein binding domains. We also use evidence from publications and databases regarding the binding property of these proteins and their closest homologues.

Querying UniPROBE with PWM-sites for Ubx yielded 24 proteins, the PWM-sites for En yielded two, and the PWM-sites for TCF yielded only one. Table 4.1 lists the top matches from this analysis according to their E-values. Notably, each of the matched TFs CEH-22 and Sox-4 appeared in UniPROBE query report several times, all below the E-value threshold. From this result we can assume that En and TCF have closer binding specificities with CEH-22 and Sox-4 TFs than with others in UniPROBE.

We hypothesized that proteins identified by the UniPROBE analysis were related in terms of the amino acid sequences of their DNA-binding domains. To test this possibility we aligned the amino acid sequences of retrieved proteins using ClustalW2 available at `http://www.ebi.ac.uk/Tools/msa/clustalw2` [38]. The aligned amino

acid sequences were retrieved from the UniProt database (although the name is similar, this is not the UniPROBE database) `www.uniprot.org` [108]. When UniProt protein had several isoforms, the protein sequence labeled as "canonical" was retrieved. The results were visually assessed with JalView available online with ClustalW2 [116]. A standalone Jalview application is available at `http://www.jalview.org/`.

The three panels in Figure 4.1 present the three alignments among the amino acid sequences of the best matches and target proteins from Table 4.1. For comparison, we added the known homologs Ceh-16 and Hoxa-7 also found in UniProt. The bars at the bottom of the aligned sequences are placed according to the amino acid coordinates of the binding domains published in UniProt. The numbers at the top of each panel show the amino acid coordinates of the first protein from the proteins submitted to ClustalW2. Three conservation histograms (feature of JalView) in Figure 4.1 show a high similarity among the aligned protein sequences including the targets En, TCF and Ubx. In particular, from the conservation histogram in panel A, more than 50% of positions showed at least 70% sequence identity for En, Ceh-16 and Ceh-22; similar results are shown in panels B and C. These alignments suggest that Hoxa-6, CEH-22 and Sox-4 have similar binding domains and therefore are likely to have similar DNA-binding preferences, providing support for the UniPROBE analysis.

The smaller length (seven) and the bigger number of queried PWM-sites for Ubx (than for En and TCF) are consistent with the bigger number of UniPROBE matches. Although we did not quantify dependencies between the number of UniPROBE matches and their quality, we visualized the result of the alignment as a rooted UPGMA (Unweighted Pair Group Method with Arithmetic Mean) tree constructed based on the percentage of identity, as computed by JalView (Figure 4.2). For this analysis, we used proteins of similar length as Ubx and excluded proteins marked "non-characterized" or "predicted"; Ubx sequence was included as a reference. The tree confirmed the E-value result for Ubx from the UniPROBE report in Table 4.1, in that Hoxa-6 is closely related to Ubx, based on the amino acid sequence alignment.

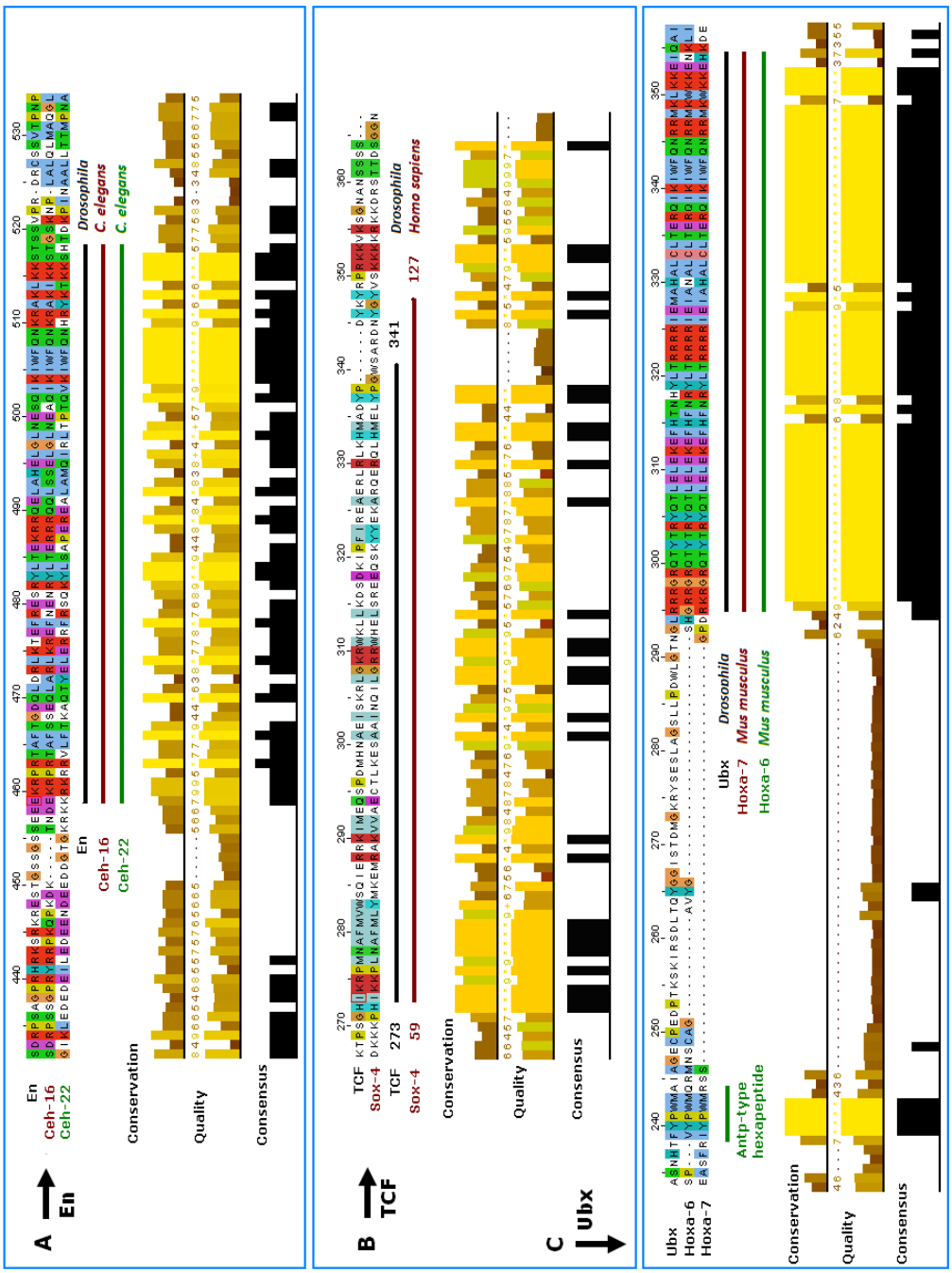

Figure 4.1: Results of the UniPROBE search querying putative TFBSs for three selected TFs. The search allows checking homology to known binding sites from different species. The results of sequence alignment are shown for three TFs: En (A), TCF (B) and Ubx (C) with most similar binding proteins reported from UniPROBE. Independently from UniPROBE analysis we found that Ubx has the reported homolog Hoxa-7, which we included in (C) for reference. Equally, the Ceh-16 orthолog was added to (A) for En. Three bars under the alignment show that all three proteins have a common protein binding domain.

The analysis above indicates that TFs identified by UniPROBE as having similar DNA binding preferences to TFs of interest also had similar amino acid sequences. Based on this similarity, we reasoned that these TFs may also share similar biological functions. To investigate the biological roles of TFs identified by UniPROBE, the TFs used in sequence alignment and their homologs, we conducted literature reviews.

Beyond the sequence alignment which indicated closely overlapping protein binding domains in Figure 4.2, we could not find experimentally verified homologs between TCF and Sox-4, although work by Van de Wetering *et al.* mentioned that TCF-1 and Sox-4 are highly homologous factors and members of the same protein family [113].

Another TF, Ceh-16 was reported to be an ortholog of Ceh-22 at `http://www.genecards.org/index.php?path=/Search/keyword/`; Hoxa-7 (Mm) was reported at `http://www.genecards.org/cgi-bin/carddisp.pl?gene=HOXA7` as homolog of Ubx. Three proteins, Ceh-16, Ceh-22 and En, have homeobox DNA binding domains as annotated in UniProt. Hoxa-6 (from UniPROBE) and Hoxa-7 (was not in UniPROBE) a member of the Antp homeobox family and also has a co-localized DNA-binding domain, which was confirmed from our analysis as shown in Figure 4.2.

During this project we asked if Match™ [54] (see page 64), would be able to identify the same TFs UniPROBE reported on our queries for PWM-sites. To address this question, the same kind of analysis was conducted for Ubx, En and TCF using putative binding sites predicted by Match™. We again queried UniPROBE with Match™ predicted sites using the UniPROBE search tool to find corresponding TFs. Although in both cases close homologs were found (Figure 4.1), all PWM-sites for optimized PWM resulted in a much bigger number of hits and smaller E-values to similar motifs (from UniPROBE collection of TFs) than sequences discovered by Match™. E-values of Match™ as indicated in the summary Table 4.2 actually are very poor.

We then further assessed the biological relevance of the new predicted TFBSs. To do so, we performed a mutual comparative analysis of DNA motifs and amino

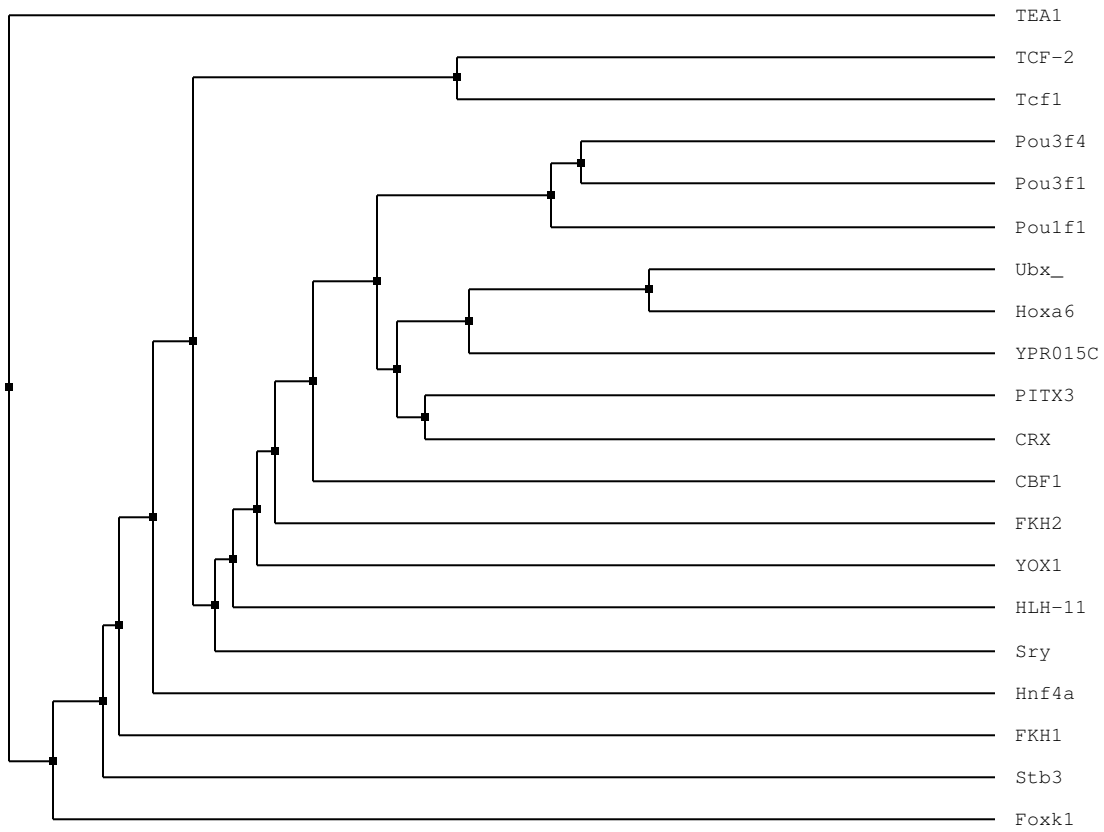

Figure 4.2: The tree shows alignment distances between TFs found in UniPROBE for Ubx PWM-sites. Numbers show the branch lengths. The entry for Ubx was not present in UniPROBE at the time of analysis. However, proteins Hoxa-6 and YPR015c mapped to queried putative TFBSs for Ubx with best E-values and showed closest sequence identity to Ubx. Moreover, we found Hoxa-7 (*Mus musculus*) was reported as Ubx homolog (see links in text).

Table 4.2: Summary of analysis Match™ vs. optimized PWM vs. UniPROBE search for similar motifs.

sequences found	Match™		Opt PWM		Matched
for TF	num. of hits	best E-val	num. of hits	best E-val	with
En	1	0.0429540	12	0.000305	CEH-22
Ubx	1	0.361303	7	0.003571	Hoxa6
TCF Q6	2	0.107297	5	0.008115	Sox4

acid sequences using a collection of DNA binding motifs from protein binding microarrays, which have a high resolution. This kind of analysis is limited (therefore potentially biased) to available information found either in databases or in publications. Nevertheless, for three TFs we were able to find evidence from different sources of information that at least some of the new putative PWM-sites carry binding properties discovered in homologous or even orthologous TFs. These findings were obtained from querying the UniPROBE database by new putative TFBSs and confirmed by aforementioned publications. For this reason we can consider that the refined PWMs are a reasonable as tool for computational prediction of new TFBSs. In remaining section, we assess the quality of refined PWMs as predictors using mostly quantitative comparative analysis with Match™.

4.2 Optimized PWMs show an increased accuracy in simulations

The sequence Logos that we obtained in chapter 3 (Figure 3.4), revealed that we could find new putative PWM-sites that might be relevant to TF-DNA binding. However, this analysis did not provide us with a comprehensive examination of the predictive performance of the new refined matrices. Therefore, here we describe an accuracy analysis that evaluates the predictive performance of refined PWMs. For this purpose we constructed tests on synthetic sequences, a popular *de facto* instrument for testing of new computational methods [109].

In the introduction we discussed the Match™ software, integrated with TRANSFAC database and used with original PWMs. Here we used Match™ with the commercial version of TRANSFAC that contains a larger number of matrices. Predictions from optimized matrices (further in text and table indicated with the prefix "OPT") were compared separately for mono- and dinucleotide versions with generic mononu-

cleotide matrices used in Match™. The goal of the testing was to compare the number of sites predicted by both methods at expected locations. A summary of all test results is shown in Table 4.3.

As was mentioned in sections 3.1, all synthetic test data were constructed from the JASPAR collection of sequences. The JASPAR collection of TFs is smaller to that of TRANSFAC. Match™ and OPT PWMs were applied in parallel for each of these 14 synthetic samples and the results are summarized in Table 4.3.

The impact of the sequence background was reduced by taking ten replicates, each of which obtained by the random shuffling of the nucleotide background within the inserted flanks. Unaligned parts were preserved in replicates, although we were able to permute them at each position ("green" nucleotides in Table 4.3). To include possible shifted matches, we enlarged the search area in favor of Match™ during hit counts. This means that if Match™ detected a hit within a half of the motif length +/-1nt from closest motif edge we considered this hit as correct.

Since Match™ recommended three distinct types of cut-off values for different tasks we performed tests with all of them. The rational for the three cut-off values was three-fold: 1) to minimize the number of biologically relevant binding sites which are missed by Match™ (false negative or FN), 2) to minimize the number of random matches found (named false positive or FP), and 3) to minimize these combined error rates (FN+TP).

We compared the means of total TP (in odd rows) from columns "both" in Table 4.3 for (1) Match™ vs. optimized mononucleotide matrix ("MATCH" block vs. "OPT matr m") and (2) Match™ vs. optimized dinucleotide matrix ("MATCH" vs. "OPT matr d"). Since each of these pairs is comparable across rows (because the number of FP is fixed) we performed paired two-tailed t-tests on the means. In both cases we found significant differences (at 5% level) showing that refined PWMs have better performances. Namely, for (1) the P-value is 0.0077 and for (2) the t-test yields the P-value of 0.0176.

Table 4.3: A summary of experiments with synthetic data tests. The results for Match™ ("MATCH") and optimized matrices (OPT) are organized in blocks. "MATCH" consists of three tested types of settings for the Match™: "min FP" - minimum number of false positives; "min FN" - minimum number of false negatives; "both" - min of sum of both criteria; "m" and "di" - refer to the number of hits after the application of mono- and dinucleotide matrices, respectively. TP (FP=const) block shows the number of TP hits assuming that both methods (Match™ and optimized) have the same stabilized FP value (presented in the first "MATCH" block for each transcription factor in "FP" row). Cut-off values for optimized matrices were selected as they appeared after optimization, except blocks with "TP (FP=const)" labels where cutoffs were adjusted to the corresponding FP rate. Two last blocks show numbers of hits found on the randomized synthetic test content using Match™ and optimized matrices; cut-off values were the same as in the non-randomized test set.

		MATCH			OPT matr		OPT matr m TP(FP=const)			OPT matr d TP(FP=const)			MATCH rand			OPT rand	
		min FP	min FN	both	m	di	min FP	min FN	both	min FP	min FN	both	min FP	min FN	both	m	di
Kr Q6	TP	0	19	13	15	11	15	31	23	22	31	23					
	FP	0	28	4	0	15							0	0	0	0	0
SuH	TP	1	9	3	10	10	10	10	10	10	10	10					
	FP	1	37	8	0	0							0	18	0	0	0
Abd-A	TP	0	0	0	4	8	0	9	5	0	8	1					
	FP	0	18	4	3	20							0	12	2	5	2
Hb	TP	0	0	0	16	11	0	16	11	0	14	3					
	FP	0	25	4	31	17							0	17	4	15	2
En	TP	0	0	0	4	8	0	8	2	0	22	9					
	FP	0	16	4	7	0							0	13	0	0	0
PRD	TP	0	0	0	2	0	0	9	3	1	7	2					
	FP	2	141	31	7	5							0	116	20	10	0
Z	TP	0	28	20	13	10	13	22	18	13	22	18					
	FP	0	168	30	1	0							0	110	10	7	0
Ubx	TP	0	0	0	14	15	0	13	0	0	15	2					
	FP	0	15	1	17	18							0	20	1	30	1
Tll	TP	0	0	0	29	31	29	34	31	31	31	31					
	FP	1	147	30	1	14							0	110	19	3	1
Sna	TP	0	0	0	2	0	8	32	22	5	37	22					
	FP	9	106	44	0	0							0	49	7	0	0
KNI	TP	0	0	0	2	0	4	21	11	4	14	9					
	FP	0	32	5	0	0							0	32	5	0	0
E74A	TP	3	16	9	11	0	11	17	16	4	16	9					
	FP	0	38	6	0	0							0	25	2	0	0
DEAF1	TP	0	0	0	1	2	0	3	1	0	6	1					
	FP	0	61	3	3	7							0	25	2	0	0
Ap Q6	TP	0	0	0	0	0	0	0	0	0	3	0					
	FP	0	13	3	3	0							0	14	3	0	0

Another analysis was then performed to compare TP discovery rates for Match™ and OPT methods. For this purpose we took the proportion of correctly identified hits over the total number of sequences in synthetic data tests. To make two types of results comparable, we used OPT matrices with such cut-off values both methods achieved the same number of FPs (shown in even rows of Table 4.3) and counted the corresponding numbers of TPs for the OPT method. In Table 4.3 a summary of these results is placed in columns indicated as "OPT TP (FP=const)" assuming constant false positive rate, which may vary for different settings of Match™.

Figure 4.3 shows that optimized matrices perform better than Match™ in terms of the number of TPs regardless of the Match™ settings used, with the one exception of Z (Zeste) TF. Zeste precessed with "min FN" settings in Match™ gives a 10% better result than OPT.

To understand why Match™ prediction looks better on that particular TF, we investigated the properties of Zeste motif that might be favorable to the Match™ method. This investigation also explained the benefit of using a combination of both mono and dinucleotide matrices.

For this task we used integer-valued matrices supplemented with refined log-odds matrices. From components of the mononucleotide occurrence matrix we compiled the products of single nucleotide probabilities for each possible combination of dinucleotides. From the dinucleotide occurrence matrix we derived *a priori* dinucleotide counts. Thus, the prior probabilities of dinucleotides can be assessed directly. When the adjacent nucleotides in a dinucleotide are independent then the ratio of these two probabilities should be close to one. Numerator and denominator of this ratio are visualized for each pair as adjacent cells in the heatmap presented in Figure 4.4.

We note that a similarity pattern between odd and even rows (carrying the same dinucleotide) in the heatmap (Figure 4.4) is apparently variable at each position within a binding motif. For example, most positions in the En motif (Figure 4.4, left panel) are more likely to be different (symbol "D", dependent) than similar (symbol

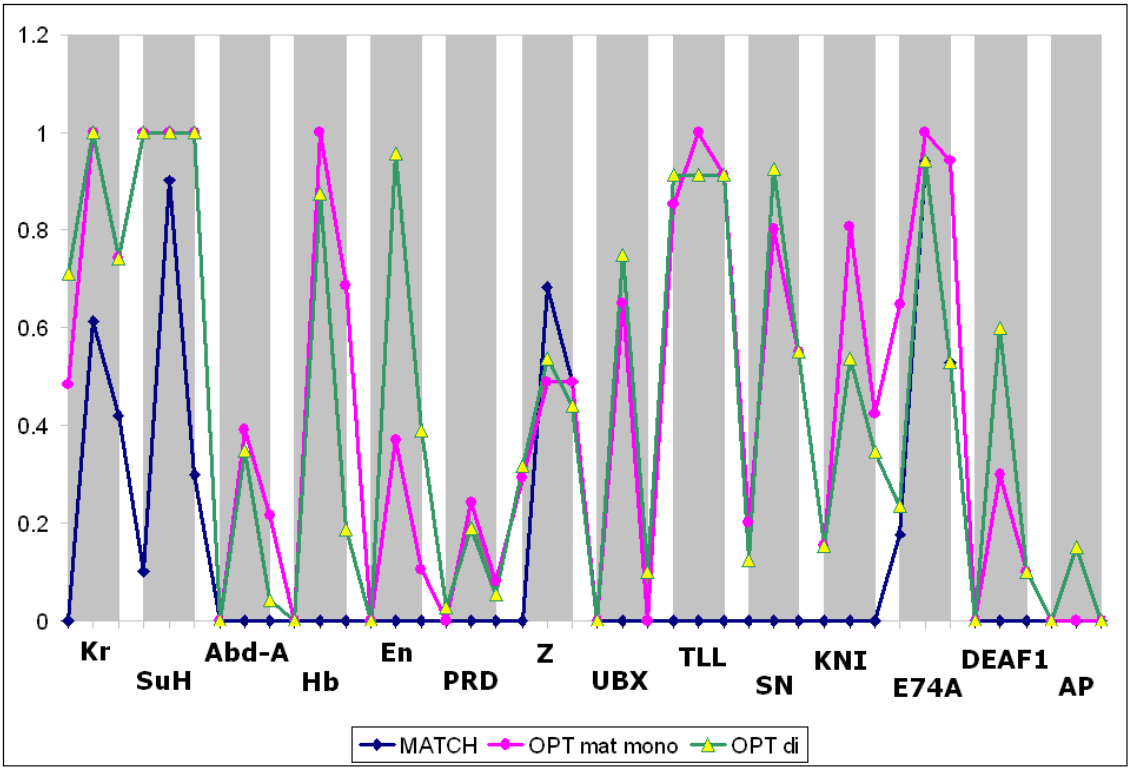

Figure 4.3: Illustration of the superiority of the predictive ability of the optimized matrices compared to Match^{TM} tool. The data are taken from rows marked as "TP" in Table4.3. Y – axis shows the proportion of found (TPs) to the number of all sequences in the synthetic test promoter data set (constant in each row, not shown, see supplementary Table 1) calculated for each of the 14 TFs from Table 4.3. Three groups: "MATCH", "OPT matr. mono" and "OPT matr. di" are compared by this ratio assuming MATCH FP=const for each of the three settings for a given TF. The order of settings "min FP", "min FN", "both crit" is left-center-right in each painted bar.

"S", independent). In other words, the mononucleotide matrix for En does not provide enough information to reproduce sequence content yielded by the dinucleotide matrix.

As seen from Figure 4.4 a high degree of similarity between some rows in Zeste panel indicates that the corresponding nucleotides are likely independent (such as for TT at 2^{nd} and 3^{rd} and CG at 4^{th} positions). It would be questionable to visually evaluate the positions 4-7 where more accurate tests are required. At the same time the left panel shows that all the adjacent nucleotide positions in consensus of the Engrailed TF are mutually dependent and dinucleotide matrix brings additional information that cannot be explained using mononucleotide counts.

To evaluate the heat maps quantitatively we quantified differences between the even and odd rows using two-sample z-test to test differences between proportions. We assume the null hypothesis $\{H_0 : P_1 = P_2\}$, where probabilities P_1 and P_2 correspond to the numerator and denominator used in the ratio. We computed the z-score as

$$z = \frac{P_1 - P_2}{\sqrt{P(1-P)(1/n_1 + 1/n_2)}}; P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}, \tag{4.2.1}$$

where P is the pooled sample proportion, n_1 is the size of the first sample (in odd rows), and n_2 is the size of the second sample (in even rows).

Before we assess TF Zeste, which is the main focus of this test, we estimated two apparently closest dinucleotides in a heat map of En TF. This example could serve as a benchmark for other cells in this heat map. We tested AT and CA at positions 3 and 2 respectively. They yielded two-tailed z-test P-values less than 10^{-4} which provides evidence of the significant difference between two proportions (with $n_1 = 2864; n_2 = 311$).

For TF Zeste the picture is more variable than for En. From equation (4.2.1) we obtained P-values less than 10^{-4} in two-tailed tests for z-score (with $n_1 = 411; n_2 = 50$) when it was computed for GA, AG, GT dinucleotides at their corresponding positions #5, #6, #7, respectively. We can conclude that we should reject the null

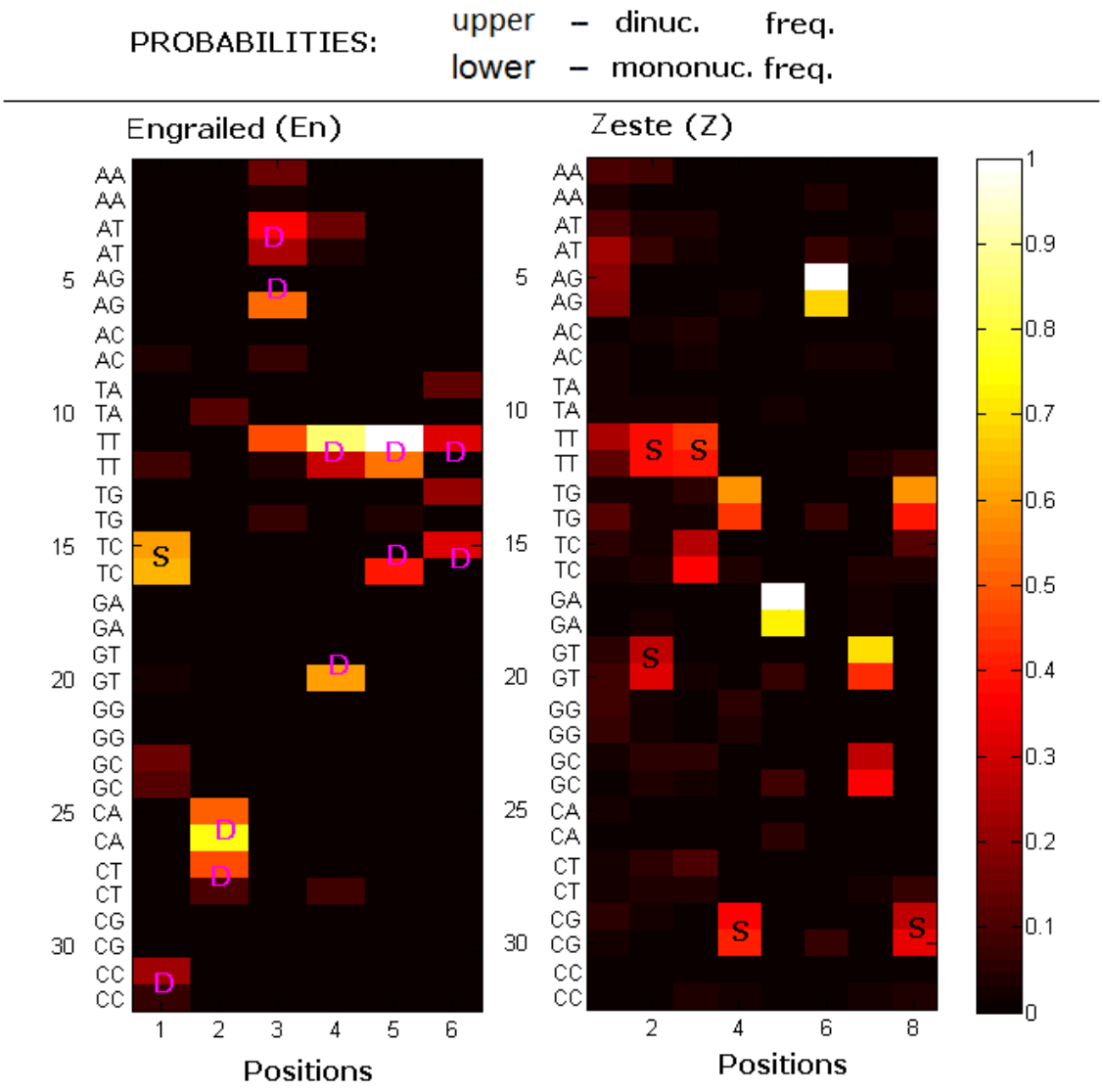

Figure 4.4: Heatmap shows probabilities of each dinucleotide at each position in the consensus for En (Engrailed, left panel) and Z (Zeste, right panel) TFs. Both panels show two probabilities of replicated dinucleotides: estimated from counting dinucleotide frequencies (upper) and from mononucleotide probabilities (lower) assuming independency of the adjacent positions. Letter "S" indicates that the two estimations are more likely similar than different whereas "D" signalizes significant difference between these values. Differences in expressions of upper and lower dinucleotides at positions 4-8 for Zeste are explained in text.

hypothesis for those three dinucleotides which are likely composed of dependent nucleotides at these positions. Thus, as GAG sequence is very conserved in Zeste motif and the positions 4-7 (and possibly 8, because TG's P-value is only 0.0.0143) form a set of dependent nucleotides. In contrast with these dinucleotides which carry dependencies, the dinucleotide TG at position 4^{th} does not show any significant difference in tested proportions and thus, it can be composed from mononucleotide predictions.

We suggest that Match™ method (looking for five most conserved nucleotides in a motif) might be sensitive enough for such a condition, but this superiority of Match™ was only observed in one out of the fourteen cases considered here, and might therefore be rare.

4.3 Discussion

Results of our computational experiments show that optimized matrices can successfully detect binding sites on a data set constructed independently of the training data sets. This demonstrates that the machine learning approach implemented to get additional sites for PWM refinement resulted in PWMs with better predictive performance than generic TRANSFAC matrices.

Another insight from our results is the demonstration that our adjusted heuristic framework makes the optimization approach implementable in a large scale for TRANSFAC TFBSs collection for *Drosophila*, and results in a better global performance compared to the conventional approach [84].

The further development of the refinement technique was a logical continuation of the previous work of Gershenzon *et al.* where they implemented the core idea to the PWM of the GC-box binding site for Sp1 TF. Its application to a larger scale was, however, limited by the manual inspection of promoter area for the presence of signal and demanded converged optimization process. As a result, this heuristic procedure was applicable only for a limited number of successful PWMs. In our

improved algorithm, which included the biological signal detection part and controlled steps over all optimization process, we provided a set of heuristic rules combined into one stable and recyclable model. Optimized matrices were used to predict up to a thousand new putative binding site candidates in the SIB-EPD set of promoter DNA sequences for *Drosophila* (as shown in Table 3.2, columns 5 and 6).

The practical method presented in the thesis involves the training of PWM on biological sequences that are extended by statistical similarity. Unlike in sequence alignment, which is central to comparative studies, the similarity we are exploiting in the algorithm comes from the possible statistical variability of positionally dependent nucleotides in proximal areas of promoters (in detected functional window). We predicted such areas using the z-score values. The location of the most sustainable signals was obtained by a closer inspection of z-score performance. The applied range of sequential cut-offs gave a stable and biologically relevant interpretation on the presence of similar motifs with closest matching scores. Biological relevancy was validated with a common standard assumption in binding site analysis that over-represented nucleotide patterns are considered as an evidence of cis-elements.

It is worth mentioning that having within 1-2nt variation of motif length, the number of sequences predicted by optimized dinucleotide matrices is roughly half the number of sequences predicted by optimized mononucleotide matrices (two last columns in Table 3.2). In terms of nucleotide counts, it means that the optimized mononucleotide matrices gain as many as twice more nucleotides per TF compared to nucleotides in sequences discovered by the dinucleotide matrices. In contrast, the count of nucleotides in the initial training sequences ("S.Used:L" column in Table 3.2) was ten times less than that after the application of optimized mononucleotide matrices with an optimal cut-off. Since we used the same functional windows in both matrices, this phenomenon illustrates that mono- and dinucleotide matrices complement each other in a way that the mononucleotide matrix aims to extend the diversity of nucleotide content, whereas the dinucleotide matrix works to preserve

biological relationship between neighboring positions and thus both goals are achieved by using both matrices. As a result, dinucleotide matrices predict sequences that look more similar to the original ones (compare Logo's pictures (B to A) vs. (C to A) in Figure 3.4). The entropy measure expressed in sequence Logos supports the interpretation of that similarity.

For promoters synthesized from the JASPAR data, we can see how quantitative and qualitative aspects of optimized mono- and dinucleotide matrices complement each other. As discussed in chapter 3, we found that OPT dinucleotide matrices give only 37.7% of improvement vs. 59.7% improvement for the OPT mononucleotide matrices with respect to the number of true positives (TPs). At the same time, the performance in terms of specificity or the number of false positives (FPs) shows a 40.2% improvement for OPT dinucleotide vs. 42.5% improvement for OPT mononucleotide matrices (Table 3.3).

Therefore, our method of constructing optimized matrices performed with higher sensitivity and specificity compared to the conventional approaches. Since the same data sets were used for each TF, we could easily evaluate specificities by comparing the number of false positives in each even row in the first two blocks "MATCH" and "OPT" from Table 4.3. For five TFs (SuH, Tll, Sna, KNI, E74A), both the mononucleotide and dinucleotide versions performed with fewer number of FPs. For other six TFs (En, Kr, PRD, Z, DEAF1 and AP) we obtained a smaller number of FPs for both mono- and dinucleotide versions. Better specificity compared to conventional approaches was reported as a feature of optimized matrices in earlier publications [35, 36] for single TFs. Here we extend this observation to batch processing of TFs. Low specificity is a common limitation of conventional TFBS prediction algorithms, and therefore the improvement in the specificity (while maintaining a descent sensitivity) of our method is a significant accomplishment.

Our study confirmed that finding similar biologically relevant binding sites is an achievable goal. We used computational methods to show that the PWMs pre-

dicted by our approach are likely real binding sites and warrant further experimental investigation.

From the inspection of discovered sequences, we note that putative binding sites contain variable nucleotide contents at certain positions as a result of more or less significant degeneracy in binding motifs at that position, which can bring additional insight into the binding affinity of certain TFs. Potentially, one can expect that the degeneracy of known PWMs to be either confirmed or not by additional sequence data because it can be an artifact of insufficient information in the initially available sequences. For transcription factor Z (Zeste), for example, nucleotides next to and immediately after the GAG pattern are interchangeable (T or C), although the occurrence of pyrimidines at those positions was confirmed [76]. Figure 3.4 shows for Ubx that the position after the first T in $TAAT$ is more degenerate than the remaining nucleotides.

Optimized matrices are able to deliver comprehensive sequence content consistent with what is biologically proven for binding affinity. To confirm this observation we considered transcription factor Zeste (Z). Mutagenesis studies determined that $(T/C/g)GAGTG(A/G/c)$ is the consensus Zeste recognition sequence [76]. Similarly, our results were able partly to confirm this consensus based on both types of matrices, as we found the following nine sequences (out of 92 using optimized mononucleotide matrix) on the set of proximal promoters. These motifs differ only by one nucleotide C immediately after GAG:

AGTTGAGCG
ATTCGAGCG
CGTCGAGCG
CTCCGAGCG
GCTTGAGCG
GGTCGAGCG
GTTTGAGCG
TGCCGAGCG
TTTGGAGCG

Consistently with what was found by mononucleotide PWM, the optimal dinucleotide matrix brings seven additional sites (out of 31 discovered):

AGTTTGAGCGC
CAGTTGAGCGC
CGGTCGAGCGG
CTTTGGAGCGA
TAGCTGAGCGG
TTTGCGAGCGG
TCGTCGAGCGG

In agreement with our finding, two publications [9, 78] also mentioned that the $TTGAGCG$ site we also found was among the experimentally characterized binding sites. Thus, we can suggest that optimized PWMs are able to bring new information and new insight on binding specificity of TFs that can be consistent with biological possibilities examined in literature. The putative binding sites we found complement to four TRANSFAC sequences (among 24 initially available) that we used for training:

TCACTGAGCGA
TTATTGAGCGG
TTTTTGAGCGC
GTTTTGAGCGT

For the Ubx TFs, we predicted sequences with $ATTA$ and $TAAT$ motifs which were reported as a preferential combination for Ubx homeodomain protein [30, 29, 28, 50] when it binds *in vitro* in a sequence- specific way. This type of motifs is observed in

both dinucleotide and mononucletide (although shortened) sets of PWM-sites, which advocates in favor of their biological relevancy. In this example we capture $TAAT$ and $ATTA$ patterns in the discovered sites. However, for some reasons they were missed in the TRANSFAC although we detected such sequences with refined mononucleotide PWM.

Intuitively, one can expect that a hypothetical method capable of finding more similarities to the initial binding sequences would perform more favorably compared to a method that does not find such similarities. This statement, for example, might be affected by stochastic irregularity of promoter content, by completeness of TFBSs collection used for training and their biological degeneracy (positional variability), and thus, in fact, the declaration is not true in all cases.

Z-score is used as a tool to detect a promoter area with significant peak matching scores. Some promoters are enriched with potential binding sites predicted by initial matrices. In other words, an area where the difference between the observed and expected hit-count frequencies gets three folds over the standard deviation presumably contains the majority of functional binding sites for a given transcription factor and hence is dubbed by us a "functional window". This scenario is a valuable part of the heuristic as we tested it on Drosophila proximal promoter data sets. The z-score can be approximated by the normal distribution on a chosen promoter length of 600nt. In addition, as shown in [115], due to the short and degenerate nature of TFBS, a typical PWM will detect a hit every 500-5000nt depending on the parameter settings.

Conceptually, the z-score used for hit counts is similar to the averaged positional distribution of elements along the aligned promoter sequences published [35], where the expected occurrence frequency is taken for a randomized (shuffled) promoter content. Our experiments with z-scores pointed to the same promoter areas, though indicating smoother distribution with fewer and more expressed peaks that facilitates large scale implementation. The number of available TFs would have increased to 64 TRANSFAC entries at the time of the study if we also had included the data where

only (mononucleotide PFM) frequency tables were present. Thus, In our study the presence of TFBSs in the database was a limiting factor in selection of TFs.

We complemented commonly used mononucleotide matrices with dinucleotide versions of PWMs for the following two reasons:

(1) a dinucleotide PWM creates additional source of information indicating dependencies between adjacent nucleotides; dinucleotide binding model infers binding scores from commonly used thermodynamic models that minimize binding energy of protein DNA-binding motif;

(2) dinucleotide PWMs can be calculated using the same computational algorithm as mononucleotide PWM. In addition, used in pairs both matrices allow us to investigate the capability of the computational schema to generate similar putative sites based on the two mutually complementary concepts: assuming or discarding the Markovian property between neighboring nucleotides. An answer was not obvious from the onset of our analysis; in particular, those similar solutions could be found for all matrices of interest using both concepts especially when considering irregular nucleotide background distributions near TSS. We found more or less similar patterns for all 32 TFs considered where biologically verified sequences were available.

Binding sequences used for training are limited in quantity (Table 3.2, column 3) for most TFs used. Therefore, we did not expect good sampling while performing matrix training because good sampling presumes confident knowledge of the whole range of binding specificities. Surprisingly enough, in all cases we were able to reasonably extend the number of sites which show matches with alternative biological source. As an example, for dinucleotide matrices, we found new putative sites that are in good agreement with experimentally confirmed patterns. This is illustrated by the observation of two most abundant nucleotides at each position. Moreover, the dinucleotide patterns are more often similar to those derived from the initial sequences. This provides evidence that additional information brought from dinucleotide pairs is able to compensate for the scarce information if only a few binding entries are present

in the data. Consistent with this is the observation that the top of the Logo pattern of the dinucleotide version is more conserved compared to the initial Logo pattern. On the other hand, mononucleotide versions often show more variability at certain positions. Examples of less variability: for TF Abd-B strong nucleotide T at position 3; for TF Ubx – G at position 4; for TF KRQ6 - C at position 5 (Appendix Figures 5.2-5.6). (The numbers correspond to the nucleotide order in the initial TRANSFAC sequences).

We used data from TRANSFAC commercial database for training purpose and JASPAR binding sequence entries for testing. Although the JASPAR public data collection is not as large as the commercial TRANSFAC collection, the former database delivers smaller numbers of manually curated sites suitable for testing. In many cases (for example En, Ubx, BRK) we found discrepancies between the JASPAR and the TRANSFAC sets, as seen in the sequence Logos, which was the source of an extra challenge for optimized matrices to predict sequences listed in JASPAR in such cases, as our optimized matrices are based on the TRANSFAC training data.

Tompa *et al.* argued that type of promoter content greatly impact TFBS predictions [109]. Using their classification of promoter models as "real", "generic" and "Markov" we can present our result as the follow. Optimized matrices, trained on one set of binding sequences with a "real" promoter content, were able to identify biologically relevant sequences on an alternative "generic" promoter content. When composition of fair and effective tests is an extra challenge for computational biologists, the capability of optimized matrices was significant for all 14 refined matrices from TRANFAC where we were able to perform independent testing. In particular, as shown on Figure 4.3, optimized matrices computed tests better than the conventional Match™ tool.

4.4 Conclusion and future directions

A detailed understanding of mechanisms of gene regulation may be achievable in the post-genomic era with the availability of genome sequences from various species. Technological advances have contributed to the generation of increasing volumes of sequence data that need to be processed computationally in order to annotate functionally important regions. For instance, the knowledge of TFBSs can be used to build a model of TF-DNA binding specificity that can help reveal mechanisms of gene regulation, including the coordinated regulation by multiple transcription factors acting together. In addition, the knowledge of TFBSs provides an effective way to annotate mutations that may disrupt regulatory mechanisms. Therefore, the ability to predict undiscovered novel TFBSs is crucial for understanding normal and diseased processes in a living cell.

Current computational algorithms for identifying TFBSs suffer from high rates of false positive predictions. The identification of regulatory signals such as TFBSs in the DNA depends on the nature and quality of the patterns of representative sequences that are constructed from training sets of sequences by means of probabilistic models. These models either assume independence between positions or suffer from considerable computational complexity. Therefore, new models are needed for more accurate prediction of novel regulatory sequences.

We re-implemented and modified the approach of Gershenzon *et al.* for PWM refinement. This approach is based on the supervised machine learning paradigm to search for similar putative TFBSs in defined areas on promoters. These TFBSs are then used to train PWMs resulting in the PWM refinement. We showed that our method can substantially improve the performance of all PWMs regardless of the initial quality published in TRANSFAC even with small number of TFBSs. Moreover, unlike the previous work by Gershenzon *et al.*, the updated version of the algorithm performs well on all TFs. We empirically determined the best combination

of heuristics used in the algorithm enabling consecutive robust improvement of the optimization criteria without drifting away from the initial TFBS pattern.

The refined PWMs in mono and dinucleotide PWM versions for *Drosophila* were compared with TRANSFAC-integrated conventional tool Match™ for computational prediction. This comparison showed that the new refined PWMs with optimized cut-off, motif length and functional window performed favorably with respect to Match™. This better performance included better sensitivity and specificity on synthetic promoter data sets, constructed from real binding sequences, different from those used in training. We arbitrarily selected several newly predicted PWM-sites and further investigated them for biological relevance using high quality databases such as UniProt and UniPROBE. In particular we found that some of the newly predicted PWM-sites mapped to homologous TF from other eukaryotic species. This analysis suggested that putative binding site predictions should be followed up by future experimental studies, and that the computational predictions delivered new reliable putative binding site candidates thus economizing time and resources.

With increasing availability of accurate data, our software should prove useful for improving PWMs for genome-wide identification of TFBSs. The improved identification can be used in an in-depth bioinformatics study of cooperative synergistic action of promoter chromatin architecture with other elements of transcription regulation machinery (TFs and their respective binding sites). As follows from recent discoveries in the area, the knowledge of nucleosome positioning around promoters is crucial for understanding of mechanisms for chromatin remodeling and gene expression in general. A first step towards this will be the study of cooperation between nucleosomes and other promoter elements based on their mutual positioning in the promoters.

An original toolbox to perform computation of refined PWM is available on attached CD.

Chapter 5

Appendix

5.1 Additional results and charts

Figures 5.2-5.6 reproduced sequence Logos for sequences mined from the set of promoter sequences described in text using refined PWMs of both types. These Figures are logical continuation of Figure 3.4.

Refined PWMs can be found in "Optimized TRANSFAC" folder on attached CD and organized by TF names. All refined PWMs also supplied with electronic versions of sequence Logos (reproducing the Figures 5.2-5.6).

5.2 Software

The software is divided into three sections and is included into "Software" folder on the attached CD.

1. "Folder 1" contains the R scripts to compute P-value for PFM,
2. "Folder 2" contains the Matlab/Octave codes to compute refined PWMs and the set of functions to perform all necessary operations with strings,
3. "Folder 3" includes the shell scripts to run the tests.

Each folder contains examples. All software is provided "as is".

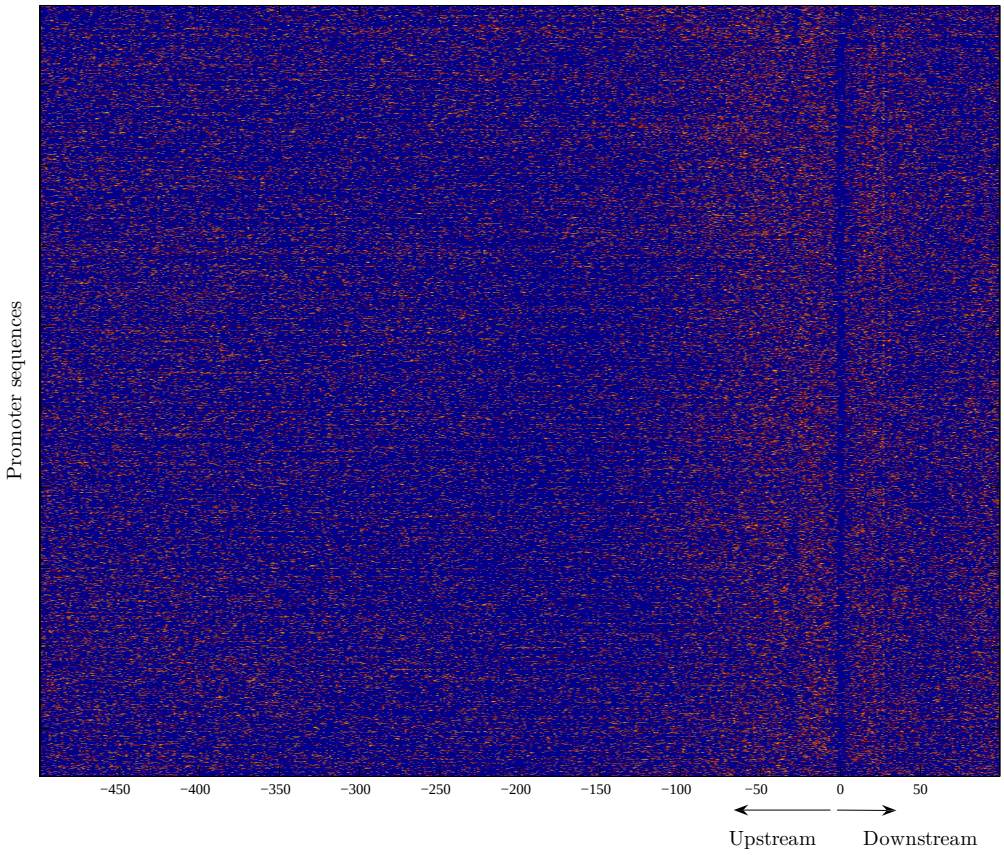

Figure 5.1: CC, GC, CG, GG dinucleotide distribution filtered from 1919 promoter sequences. The remaining dinucleotide content painted "blue". This Figure complements Figure 2.1.

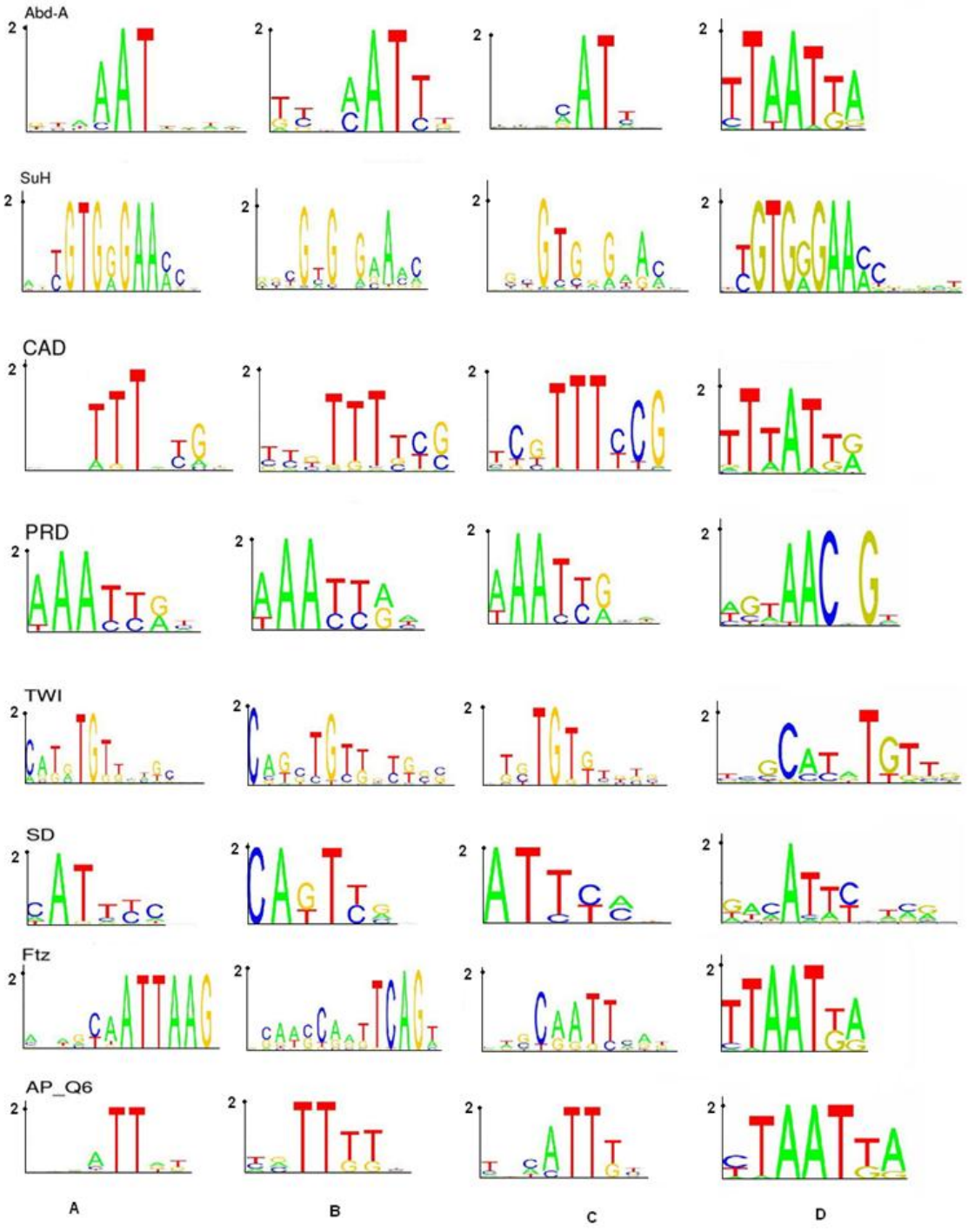

Figure 5.2: Logos of predicted sequences using optimized matrices compared to TRANSFAC and JASPAR sites. (A) Initial TRANSFAC sites. (B) From new discovered sites (mononucleotide matrix). (C) From new discovered sites (dinucleotide matrix). (D) From JASPAR CORE collection.

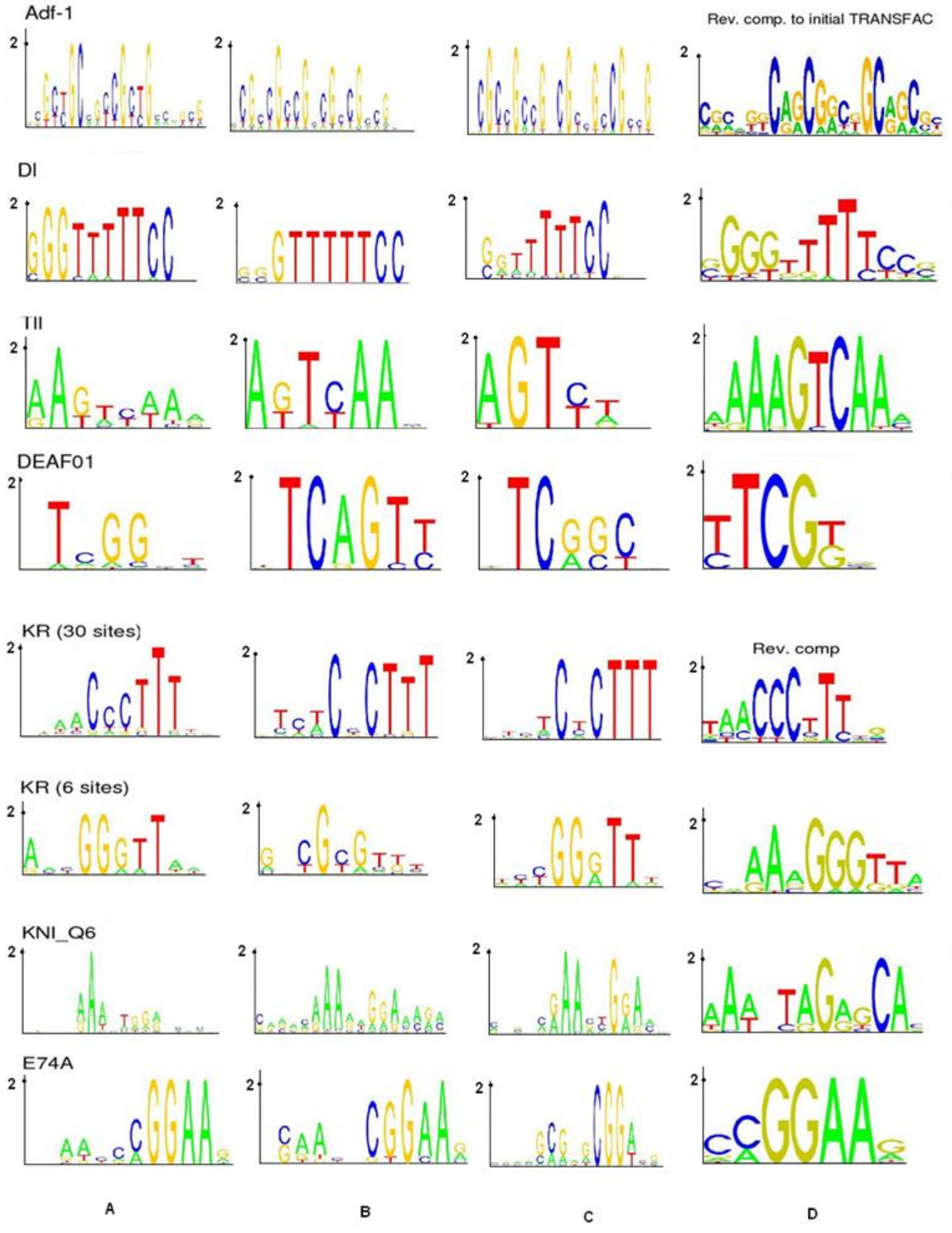

Figure 5.3: Logos of predicted sequences using optimized matrices compared to TRANSFAC and JASPAR sites. (A) Initial TRANSFAC sites. (B) From new discovered sites (mononucleotide matrix). (C) From new discovered sites (dinucleotide matrix). (D) From JASPAR CORE collection.

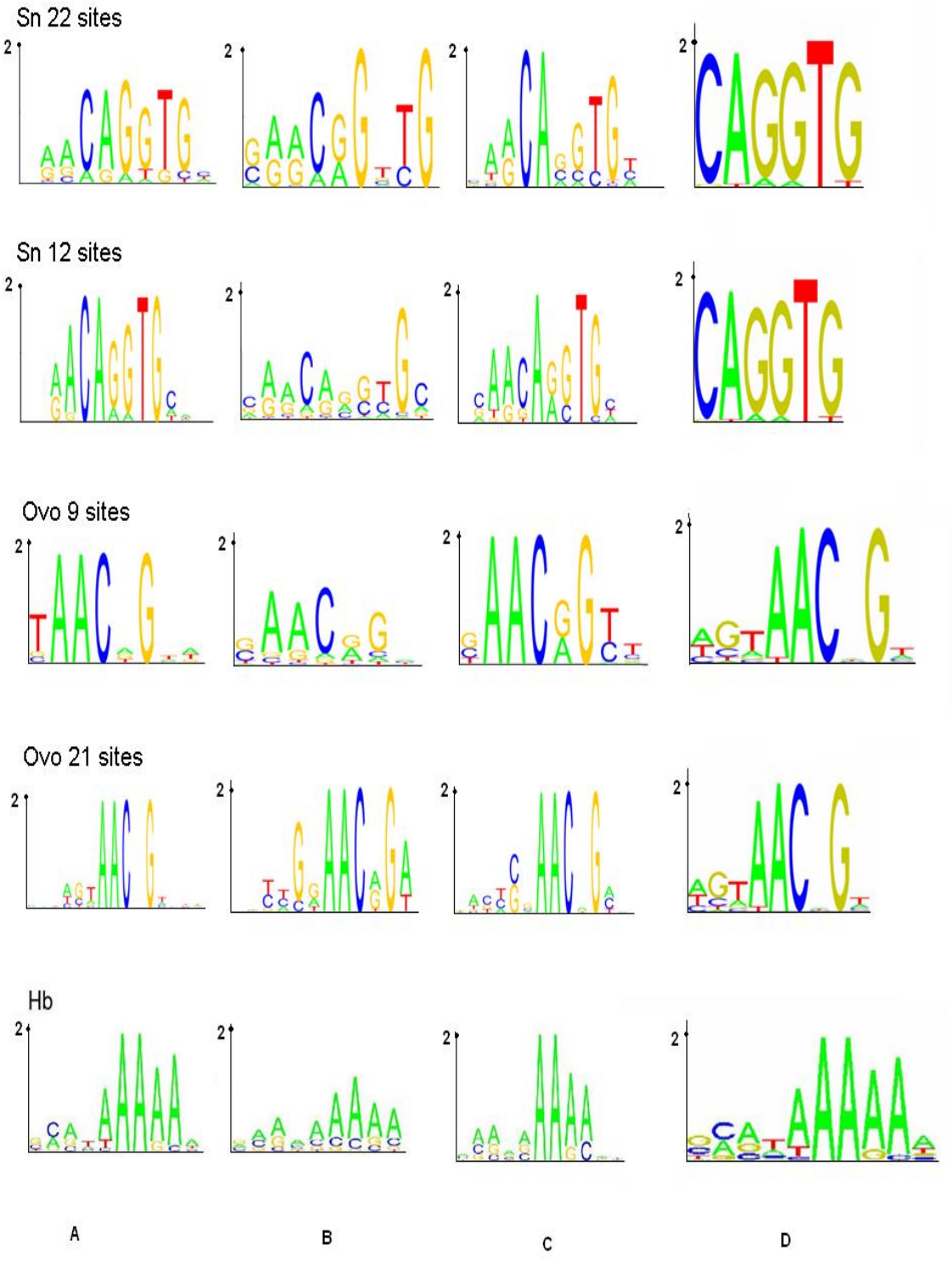

Figure 5.4: Logos of predicted sequences using optimized matrices compared to TRANSFAC and JASPAR sites. (A) Initial TRANSFAC sites. (B) From new discovered sites (mononucleotide matrix). (C) From new discovered sites (dinucleotide matrix). (D) From JASPAR CORE collection.

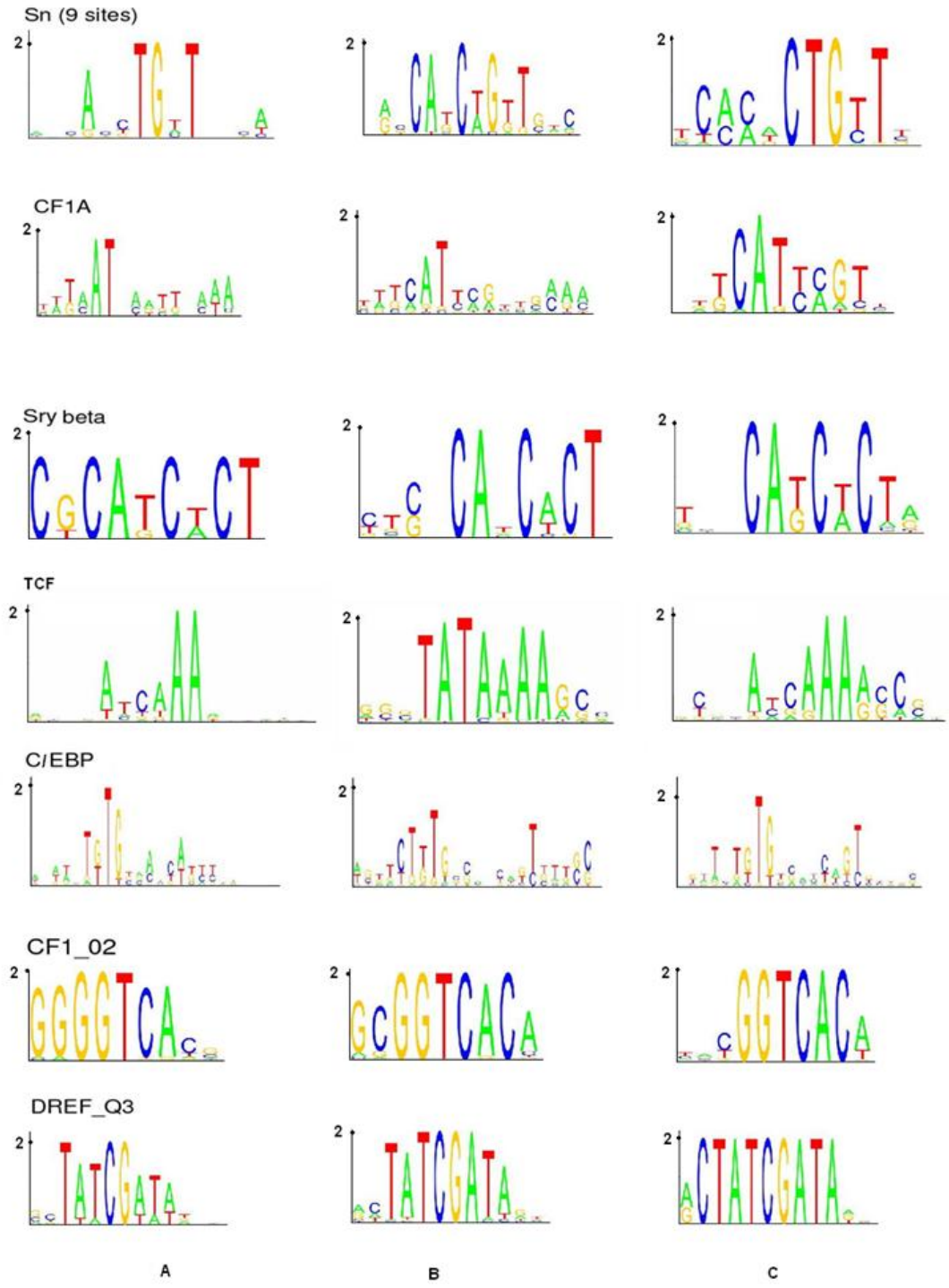

Figure 5.5: Logos of predicted sequences using optimized matrices compared to TRANSFAC and JASPAR sites. (A) Initial TRANSFAC sites. (B) From new discovered sites (mononucleotide matrix). (C) From new discovered sites (dinucleotide matrix).

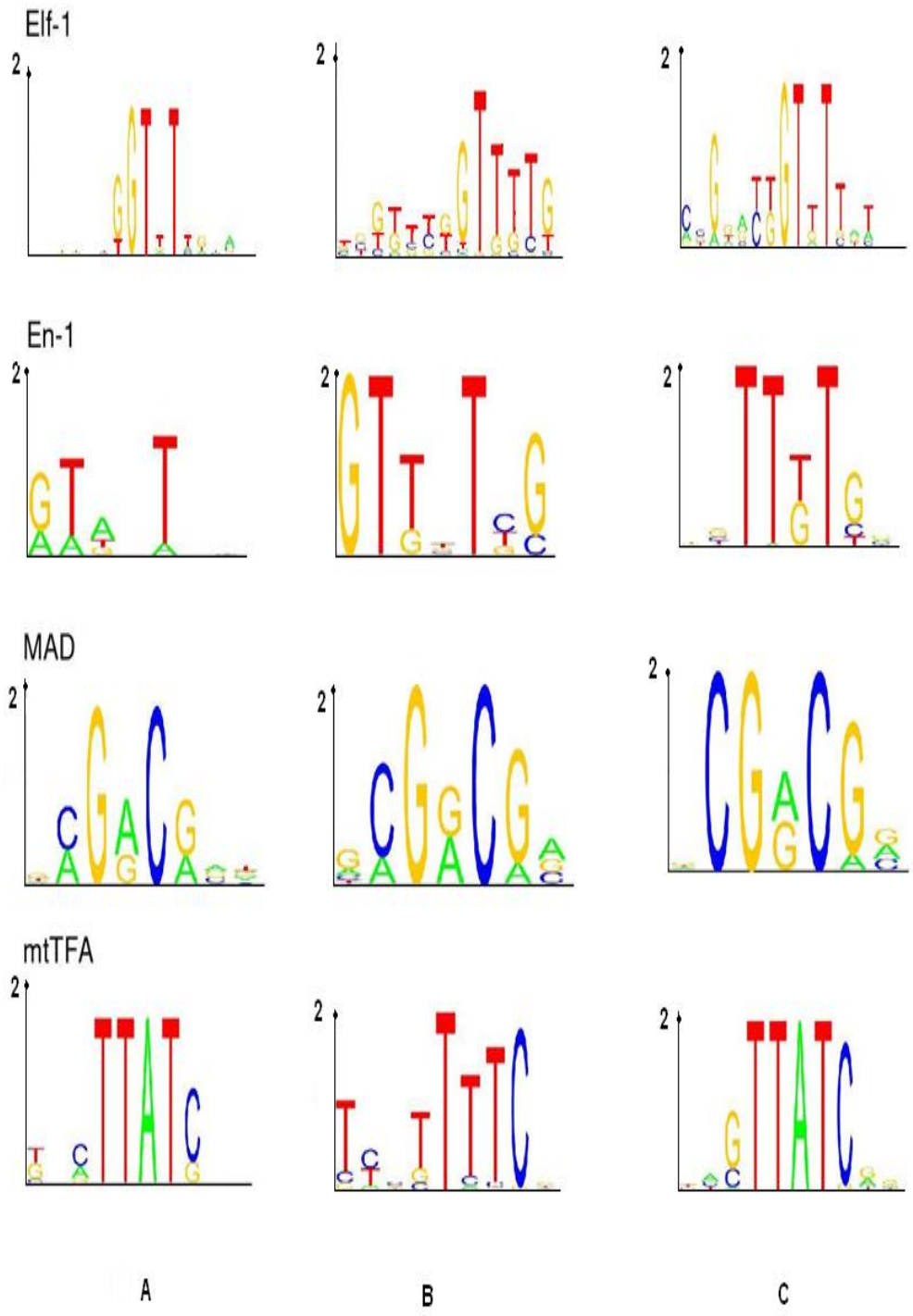

Figure 5.6: Logos of predicted sequences using optimized matrices compared to TRANSFAC and JASPAR sites. (A) Initial TRANSFAC sites. (B) From new discovered sites (mononucleotide matrix). (C) From new discovered sites (dinucleotide matrix).

Bibliography

[1] T. Akutsu, H. Bannai, S. Miyano, and S. Ott. On the complexity of deriving position specific score matrices from positive and negative sequences. *Discrete Applied Mathematics*, 155(6-7):676–685, Apr. 2007.

[2] D. Alamanova, P. Stegmaier, and A. Kel. Creating PWMs of transcription factors using 3D structure-based computation of protein-DNA free binding energies. *BMC bioinformatics*, 11:225, Jan. 2010.

[3] P. Baldi, S. Brunak, Y. Chauvin, C. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics (Oxford, England)*, 16(5):412–24, May 2000.

[4] Y. Barash, G. Elidan, N. Friedman, and T. Kaplan. Modeling dependencies in protein-DNA binding sites. In *Proceedings of the seventh annual international conference on Computational molecular biology - RECOMB '03*, pages 28–37, New York, New York, USA, Apr. 2003. ACM Press.

[5] M. Beckstette, R. Homann, R. Giegerich, and S. Kurtz. Fast index based algorithms and software for matching position specific scoring matrices. *BMC bioinformatics*, 7(1):389, Jan. 2006.

[6] P. Benos, M. Bulyk, and G. Stormo. Additivity in protein-DNA interactions:

how good an approximation is it? *Nucleic acids research*, 30(20):4442–51, Oct. 2002.

[7] M. Berger and M. Bulyk. Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. *Methods in molecular biology (Clifton, N.J.)*, 338:245–60, Jan. 2006.

[8] B. Berman, Y. Nibu, B. Pfeiffer, P. Tomancak, S. Celniker, M. Levine, G. Rubin, and M. Eisen. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2):757–62, Jan. 2002.

[9] M. Biggin, S. Bickel, M. Benson, V. Pirrotta, and R. Tjian. Zeste encodes a sequence-specific transcription factor that activates the Ultrabithorax promoter in vitro. *Cell*, 53(5):713–22, June 1988.

[10] M. Blanchette, A. 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 research*, 16(5):656–68, May 2006.

[11] M. Blanchette, B. Schwikowski, and M. Tompa. Algorithms for phylogenetic footprinting. *Journal of computational biology : a journal of computational molecular cell biology*, 9(2):211–23, Jan. 2002.

[12] M. Blanchette and S. Sinha. Separating real motifs from their artifacts. *Bioinformatics (Oxford, England)*, 17 Suppl 1:S30–8, Jan. 2001.

[13] V. Boeva, J. Clément, M. Régnier, M. Roytberg, and V. Makeev. Exact p-value calculation for heterotypic clusters of regulatory motifs and its application in

computational annotation of cis-regulatory modules. *Algorithms for molecular biology : AMB*, 2(1):13, Jan. 2007.

[14] R. Bradley, X.-Y. Li, C. Trapnell, S. Davidson, L. Pachter, H. C. Chu, L. Tonkin, M. Biggin, and M. Eisen. Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. *PLoS biology*, 8(3):e1000343, Mar. 2010.

[15] P. Bucher. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *Journal of molecular biology*, 212(4):563–78, Apr. 1990.

[16] M. Bulyk, P. Johnson, and G. Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic acids research*, 30(5):1255–61, Mar. 2002.

[17] H. Burden and Z. Weng. Identification of conserved structural features at sequentially degenerate locations in transcription factor binding sites. *Genome informatics. International Conference on Genome Informatics*, 16(1):49–58, Jan. 2005.

[18] C. Burge, A. Campbell, and S. Karlin. Over- and under-representation of short oligonucleotides in DNA sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 89(4):1358–62, Feb. 1992.

[19] M. Carey, C. Peterson, and S. Smale. *Transcriptional Regulation in Eukaryotes: Concepts, Strategies, and Techniqes*. Cold Spring Harbor Laboratory Press; 2 edition, 2008.

[20] S. Celniker and G. Rubin. The Drosophila melanogaster genome. *Annual review of genomics and human genetics*, 4:89–117, Jan. 2003.

[21] K.-B. Chen and Y. Zhang. A varying threshold method for ChIP peak-calling using multiple sources of information. *Bioinformatics (Oxford, England)*, 26(18):i504–10, Sept. 2010.

[22] X. Chen, L. Guo, and Z. Fan. Learning position weight matrices from sequence and expression data. *Science And Technology*, pages 1–12, 2007.

[23] L. Cherbas and P. Cherbas. The arthropod initiator: the capsite consensus plays an important role in transcription. *Insect biochemistry and molecular biology*, 23(1):81–90, Jan. 1993.

[24] J.-M. Claverie and S. Audic. The statistical significance of nucleotide position-weight matrix matches. *Computer applications in the biosciences : CABIOS*, 12(5):431–9, Oct. 1996.

[25] E. Davidson. *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*. Academic Press, 1 edition edition, 2006.

[26] W. Day and F. McMorris. Critical comparison of consensus methods for molecular sequences. *Nucleic acids research*, 20(5):1093–9, Mar. 1992.

[27] M. Djordjevic, A. Sengupta, and B. Shraiman. A biophysical approach to transcription factor binding site discovery. *Genome research*, 13(11):2381–90, Nov. 2003.

[28] A. Dorn, M. Affolter, M. Müller, W. Gehring, and W. Leupin. Distamycin-induced inhibition of homeodomain-DNA complexes. *The EMBO journal*, 11(1):279–86, Jan. 1992.

[29] S. Ekker, D. von Kessler, and P. Beachy. Differential DNA sequence recognition is a determinant of specificity in homeotic gene action. *The EMBO journal*, 11(11):4059–72, Nov. 1992.

[30] S. Ekker, K. Young, D. von Kessler, and P. Beachy. Optimal DNA sequence recognition by the Ultrabithorax homeodomain of Drosophila. *The EMBO journal*, 10(5):1179–86, May 1991.

[31] L. Elnitski, V. Jin, P. Farnham, and S. Jones. Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome research*, 16(12):1455–64, Dec. 2006.

[32] W. Feller. *An introduction to Probability Theory and Its applications*. John Wiley, New York, 3 edition, 1968.

[33] M. Fried and D. Crothers. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic acids research*, 9(23):6505–25, Dec. 1981.

[34] M. Frith, U. Hansen, and Z. Weng. Detection of cis-element clusters in higher eukaryotic DNA. *Bioinformatics (Oxford, England)*, 17(10):878–89, Oct. 2001.

[35] N. Gershenzon, G. Stormo, and I. Ioshikhes. Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. *Nucleic acids research*, 33(7):2290–301, Jan. 2005.

[36] N. Gershenzon, E. Trifonov, and I. Ioshikhes. The features of Drosophila core promoters revealed by statistical analysis. *BMC genomics*, 7:161, Jan. 2006.

[37] V. Gotea, A. Visel, J. Westlund, M. Nobrega, L. Pennacchio, and I. Ovcharenko. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome research*, 20(5):565–77, May 2010.

[38] M. Goujon, H. McWilliam, W. Li, F. Valentin, S. Squizzato, J. Paern, and R. Lopez. A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic acids research*, 38(Web Server issue):W695–9, July 2010.

[39] S. Hannenhalli. Eukaryotic transcription factor binding sites–modeling and integrative search methods. *Bioinformatics (Oxford, England)*, 24(11):1325–31, June 2008.

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

[41] R. Hardison. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends in genetics : TIG*, 16(9):369–72, Sept. 2000.

[42] G. Hertz and G. Stormo. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics (Oxford, England)*, 15(7-8):563–77, 1999.

[43] G. Hertz and G. Stromo. Identi cation of Consensus Patterns in Unaligned DNA and Protein Sequences: a Large-Deviation Statistical Basis for Penalizing Gaps. In L. H. C. and Cantor, editors, *Proceedings of the 3rd International Conference on Bioinformatics and Genome Research*, pages 201–216, Singapore, 1995. World Scienti c Publishing Co., Ltd.

[44] L. Hertzberg, S. Izraeli, and E. Domany. STOP: searching for transcription factor motifs using gene expression. *Bioinformatics*, 23(14):1737–1743, 2007.

[45] D. Holloway, M. Kon, and C. DeLisi. Integrating Genomic Data to Predict Transcription Factor Binding. *Genome Informatics*, 16(1):83–94, 2005.

[46] B. Hoopes, J. LeBlanc, and D. Hawley. Contributions of the TATA box sequence to rate-limiting steps in transcription initiation by RNA polymerase II. *Journal of molecular biology*, 277(5):1015–31, Apr. 1998.

[47] R. Hughey, M. Brown, A. Krogh, D. Haussler, and S. Mian. Using Dirichlet mixture priors to derive hidden Markov models for protein families. pages 47–55, 1993.

[48] I. Ioshikhes, I. Albert, S. Zanton, and B. Pugh. Nucleosome positions predicted through comparative genomics. *Nature genetics*, 38(10):1210–5, Oct. 2006.

[49] E. Ising. Beitrag sur Theorie des Ferromagnetismus. *Zeit. fur Physik*, 31:253–258, 1925.

[50] F. Johnson and M. Krasnow. Stimulation of transcription by an Ultrabithorax protein in vitro. *Genes & development*, 4(6):1044–52, June 1990.

[51] W. Johnson, W. Li, C. Meyer, R. Gottardo, J. Carroll, M. Brown, and X. Liu. Model-based analysis of tiling-arrays for ChIP-chip. *Proceedings of the National Academy of Sciences of the United States of America*, 103(33):12457–62, Aug. 2006.

[52] R. Jothi, S. Cuddapah, A. Barski, K. Cui, and K. Zhao. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic acids research*, 36(16):5221–31, Sept. 2008.

[53] T. Kaplan, X.-Y. Li, P. Sabo, S. Thomas, J. Stamatoyannopoulos, M. Biggin, and M. Eisen. Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. *PLoS genetics*, 7(2):e1001290, Jan. 2011.

[54] A. Kel, E. Gössling, I. Reuter, E. Cheremushkin, O. Kel-Margoulis, and E. Wingender. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic acids research*, 31(13):3576–9, July 2003.

[55] D. King, J. Taylor, L. Elnitski, F. Chiaromonte, W. Miller, and R. Hardison. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome research*, 15(8):1051–60, Aug. 2005.

[56] D.-H. Lee, N. Gershenzon, M. Gupta, I. Ioshikhes, D. Reinberg, and B. Lewis. Functional characterization of core promoter elements: the downstream core element is recognized by TAF1. *Molecular and cellular biology*, 25(21):9674–86, Nov. 2005.

[57] T. Lee and R. Young. Transcription of eukaryotic protein-coding genes. *Annual review of genetics*, 34:77–137, Jan. 2000.

[58] M. Leung, G. Marsh, and T. Speed. Over- and underrepresentation of short DNA words in herpesvirus genomes. *Journal of computational biology : a journal of computational molecular cell biology*, 3(3):345–60, Jan. 1996.

[59] V. Levitsky, E. Ignatieva, E. Ananko, I. Turnaev, T. Merkulova, N. Kolchanov, and T. Hodgman. Effective transcription factor binding site prediction using a combination of optimization, a genetic algorithm and discriminant analysis to capture distant interactions. *BMC bioinformatics*, 8:481, Jan. 2007.

[60] S. Levy, S. Hannenhalli, and C. Workman. Enrichment of regulatory signals in conserved non-coding genomic sequence. *Bioinformatics (Oxford, England)*, 17(10):871–7, Oct. 2001.

[61] L. Li, Y. Liang, and R. Bass. GAPWM: a genetic algorithm method for optimizing a position weight matrix. *Bioinformatics (Oxford, England)*, 23(10):1188–94, May 2007.

[62] G. Loots and I. Ovcharenko. rVISTA 2.0: evolutionary analysis of transcription

factor binding sites. *Nucleic acids research*, 32(Web Server issue):W217–21, July 2004.

[63] R. Lusk and M. Eisen. Use of an evolutionary model to provide evidence for a wide heterogeneity of required affinities between transcription factors and their binding sites in yeast. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 489–500, Jan. 2008.

[64] S. MacArthur, X.-Y. Li, J. Li, J. Brown, H. C. Chu, L. Zeng, B. Grondona, A. Hechmer, L. Simirenko, S. Keränen, D. Knowles, M. Stapleton, P. Bickel, M. Biggin, and M. Eisen. Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome biology*, 10(7):R80, Jan. 2009.

[65] V. Makeev, A. Lifanov, A. Nazina, and D. Papatsenko. Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. *Nucleic acids research*, 31(20):6016–26, Oct. 2003.

[66] K. Malde and R. Giergerich. Calculating PSSM probabilities with lazy dynamic programming. *Journal of Functional Programming*, 16(01):75–81, Jan. 2006.

[67] L. Mariño Ramírez, J. Spouge, G. Kanga, and D. Landsman. Statistical analysis of over-represented words in human promoter sequences. *Nucleic acids research*, 32(3):949–58, Jan. 2004.

[68] A. Markov. Investigation of a specific case of dependent observations. *Izv. Imper. Akad. Nauk. St.-Petersburg*, 3:61–80, 1908.

[69] M. Markstein, P. Markstein, V. Markstein, and M. Levine. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the

Drosophila embryo. *Proceedings of the National Academy of Sciences of the United States of America*, 99(2):763–8, Jan. 2002.

[70] M. Markstein, R. Zinzen, P. Markstein, K.-P. Yee, A. Erives, A. Stathopoulos, and M. Levine. A regulatory code for neurogenic gene expression in the Drosophila embryo. *Development (Cambridge, England)*, 131(10):2387–94, May 2004.

[71] 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 research*, 34(Database issue):D108–10, Jan. 2006.

[72] T. Mavrich, I. Ioshikhes, B. Venters, C. Jiang, L. Tomsho, J. Qi, S. Schuster, I. Albert, and B. Pugh. A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome research*, 18(7):1073–83, July 2008.

[73] T. Minka. Bayesian inference, entropy, and the multinomial distribution, 2003.

[74] T. Mitchell. *Machine Learning [Paperback]*. McGraw-Hill; New edition edition, 1997.

[75] Q. Mo. A fully Bayesian hidden Ising model for ChIP-seq data analysis. *Biostatistics (Oxford, England)*, pages kxr029–, Sept. 2011.

[76] L. Mohrmann, A. Kal, and C. Verrijzer. Characterization of the extended Myb-like DNA-binding domain of trithorax group protein Zeste. *The Journal of biological chemistry*, 277(49):47385–92, Dec. 2002.

[77] C. Moorman, L. Sun, J. Wang, E. de Wit, W. Talhout, L. Ward, F. Greil, X.-J. Lu, K. White, H. Bussemaker, and B. van Steensel. Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America*, 103(32):12027–32, Aug. 2006.

[78] A. Moses, D. Pollard, D. Nix, V. Iyer, X.-Y. Li, M. Biggin, and M. Eisen. Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS computational biology*, 2(10):e130, Oct. 2006.

[79] D. Mount. *Bioinformatics: Sequence and Genome Analysis, Second Edition*. Cold Spring Harbor Laboratory Press.

[80] S. Mukherjee, M. Berger, G. Jona, X. Wang, D. Muzzey, M. Snyder, R. Young, and M. Bulyk. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nature genetics*, 36(12):1331–9, Dec. 2004.

[81] M. Orihara, C. Hosono, T. Kojima, and K. Saigo. Identification of engrailed promoter elements essential for interactions with a stripe enhancer in Drosophila embryos. *Genes to cells : devoted to molecular & cellular mechanisms*, 4(4):205–18, Apr. 1999.

[82] P. Park. ChIP-seq: advantages and challenges of a maturing technology. *Nature reviews. Genetics*, 10(10):669–80, Oct. 2009.

[83] T. Phillips. Regulation of Transcription and Gene Expression in Eukaryotes. *Nature Education*, 1(1):2008–2010, 2008.

[84] L. Pickert, I. Reuter, F. Klawonn, and E. Wingender. Transcription regulatory region analysis using signal detection and fuzzy clustering. *Bioinformatics*, 14(3):244–251, Apr. 1998.

[85] B. Prum, F. Rodolphe, and E. de Turckheim. Finding Words with Unexpected Frequencies in Deoxyribonucleic Acid Sequences. Oct. 1995.

[86] A. Reményi, H. Schöler, and M. Wilmanns. Combinatorial control of gene expression. *Nature structural & molecular biology*, 11(9):812–5, Sept. 2004.

[87] E. Roulet, S. Busso, A. Camargo, A. Simpson, N. Mermod, and P. Bucher. High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nature biotechnology*, 20(8):831–5, Aug. 2002.

[88] E. Roulet, I. Fish, P. Bucher, and N. Mermod. Evaluation of computer tools for the prediction of transcription factor binding sites on genomic DNA. *In Silico Biology*, 1:21–28, 1998.

[89] A. Sandelin, W. Alkema, P. Engström, W. Wasserman, and B. Lenhard. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, 32(Database issue):D91–4, Jan. 2004.

[90] A. Sarai and H. Kono. Protein-DNA recognition patterns and predictions. *Annual review of biophysics and biomolecular structure*, 34:379–98, Jan. 2005.

[91] S. Schbath. An efficient statistic to detect over- and under-represented words in DNA sequences. *Journal of computational biology : a journal of computational molecular cell biology*, 4(2):189–92, Jan. 1997.

[92] 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 research*, 34(Database issue):D82–5, Jan. 2006.

[93] T. Schneider and R. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–6100, Oct. 1990.

[94] A. Schröder, J. Eichner, J. Supper, J. Eichner, D. Wanke, C. Henneges, and A. Zell. Predicting DNA-binding specificities of eukaryotic transcription factors. *PloS one*, 5(11):e13876, Jan. 2010.

[95] E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thå ström, Y. Field, I. Moore, J.-P. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–8, Aug. 2006.

[96] R. Siddharthan. Dinucleotide weight matrices for predicting transcription factor binding sites: generalizing the position weight matrix. *PloS one*, 5(3):e9722, Jan. 2010.

[97] S. Sinha and M. Tompa. Discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic acids research*, 30(24):5549–60, Dec. 2002.

[98] S. Smale and J. Kadonaga. The RNA polymerase II core promoter. *Annual review of biochemistry*, 72:449–79, Jan. 2003.

[99] T. Smith and M. Waterman. Identification of Common Molecular Subsequences Identification of Common Molecular Subsequences. *J . M o l . Biol.*, pages 195–197, 1981.

[100] R. Staden. Computer methods to locate signals in nucleic acid sequences. *Nucleic acids research*, 12(1 Pt 2):505–19, Jan. 1984.

[101] G. Stormo and D. Fields. Specificity, free energy and information content in protein-DNA interactions. *Trends in biochemical sciences*, 23(3):109–13, Mar. 1998.

[102] G. Stormo, T. Schneider, L. Gold, and A. Ehrenfeucht. Use of the 'Perceptron' algorithm to distinguish transitional initiation sites in E. coli. *Nucleic Acids Research*, 10:2997–3012, 1982.

[103] G. Stormo and Y. Zhao. Determining the specificity of protein-DNA interactions. *Nature reviews. Genetics*, 11(11):751–60, Nov. 2010.

[104] A. Tanay, I. Gat-Viks, and R. Shamir. A global view of the selection forces in the evolution of yeast cis-regulation. *Genome research*, 14(5):829–34, May 2004.

[105] V. Teif. Predicting gene-regulation functions: lessons from temperate bacteriophages. *Biophysical journal*, 98(7):1247–56, Apr. 2010.

[106] V. Teif and K. Rippe. Calculating transcription factor binding maps for chromatin. *Briefings in bioinformatics*, July 2011.

[107] D. Thakurta. Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic acids research*, 34(12):3585–98, Jan. 2006.

[108] The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic acids research*, 40(Database issue):D71–5, Jan. 2012.

[109] M. Tompa, N. Li, T. Bailey, G. Church, B. De Moor, E. Eskin, A. Favorov, M. Frith, Y. Fu, W. Kent, V. Makeev, A. Mironov, W. 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. *Nature biotechnology*, 23(1):137–44, Jan. 2005.

[110] H. Touzet and J.-S. Varré. Efficient and accurate P-value computation for Position Weight Matrices. *Algorithms for molecular biology : AMB*, 2(1):15, Jan. 2007.

[111] C. Tuerk and L. Gold. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science (New York, N.Y.)*, 249(4968):505–10, Aug. 1990.

[112] B. Turner. *Chromatin and Gene Regulation: Molecular Mechanisms in Epigenetics*. Wiley-Blackwell, 2002.

[113] M. van de Wetering, M. Oosterwegel, K. van Norren, and H. Clevers. Sox-4, an Sry-like HMG box protein, is a transcriptional activator in lymphocytes. *The EMBO journal*, 12(10):3847–54, Oct. 1993.

[114] A. Wagner. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics (Oxford, England)*, 15(10):776–84, Oct. 1999.

[115] W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nature reviews. Genetics*, 5(4):276–87, Apr. 2004.

[116] A. M. Waterhouse, J. B. Procter, D. M. A. Martin, M. Clamp, and G. J. Barton. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics (Oxford, England)*, 25(9):1189–91, May 2009.

[117] J. Watson, R. Myers, A. Caudy, and J. Witkowski. *Recombinant DNA: Genes and Genomes - A Short Course, Third Edition (Watson, Recombinant DNA)*. W. H. Freeman, 2006.

[118] X. Yi, Y.-D. Cai, Z. He, W. Cui, and X. Kong. Prediction of nucleosome positioning based on transcription factor binding sites. *PloS one*, 5(9):7, Jan. 2010.

[119] Y.-K. Yu and S. Altschul. The complexity of the dirichlet model for multiple alignment data. *Journal of computational biology : a journal of computational molecular cell biology*, 18(8):925–39, Aug. 2011.

[120] Y. Zhao, D. Granas, and G. Stormo. Inferring binding energies from selected binding sites. *PLoS computational biology*, 5(12):e1000590, Dec. 2009.

[121] Z. Zhu, J. Shendure, and G. Church. Discovering functional transcription-factor combinations in the human cell cycle. *Genome research*, 15(6):848–55, June 2005.

[122] Q. Ziliang, H. Zhisong, and C. Yudong. Applying Machine Learning Strategy in Transcription Factor DNA Binding Site Prediction. In G. P. Fung, editor, *A Practical Guide to Bioinformatics Analysis*, chapter 13, pages 184–194. iConcept Press, 2010.