Statistical methods for motif hit enrichment in DNA sequences

Dissertation

eingereicht beim Fachbereich Mathematik und Informatik
der Freien Universität Berlin

von

Wolfgang Kopp

Berlin, 2016

Erstgutachter: Prof. Dr. Martin Vingron
Zweitgutachter: Prof. Dr. Sven Rahmann

Tag der Disputation: 25.04.2017

Selbstständigkeitserklärung

Hiermit erläre ich, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsquellen und Quellen verwendet habe. Ich erkläre weiterhin, dass ich die vorliegende Arbeit oder deren Inhalt nicht in einem früheren Promotionsverfahren eingereicht habe.

Wolfgang Kopp, Berlin, den 29. September 2016

Acknowledgments

I sincerely thank Prof. Dr. Martin Vingron for giving me the opportunity to obtain my PhD in his department. He gave me room for creativity and supported me with valuable and fruitful feedback. I want to thank Dr. Peter Arndt for proofreading parts of the thesis. I also want to thank all former and current colleagues in the Vingron department and in the IMPRS-CBCS for creating such a nice research and social environment and, in particular, Matt Huska for being a great office mate and for many valuable discussions. I want to thank the IMPRS-CBCS coordinators Kirsten Kelleher and Fabian Feutlinske for their great help with organizational issues. Finally, I want to thank my wife Hyun Jung and my parents for all their love and support.

Wolfgang Kopp

Contents

Chapter 1

Introduction

More than 150 years ago, the scientific contributions of Mendel and Darwin helped to shape the understanding of biology fundamentally, much of which remains still valid until today. Around 1856, Gregor Mendel conducted the famous cross-breeding experiments in peas which made apparent that biological traits are inherited from one generation to the next. Around that time Charles Darwin also proposed a theory of biological evolution and natural selection which attempted to explain how species evolve by acquiring ever so slight random variations that are inherited to the next generation and might be advantageous or disadvantageous [14].

Based on these findings, the notion of a gene was established as a piece of information that is inherited to its offspring and which eventually defines biological traits. However, genes were merely an abstract concept until Watson and Crick discovered the structure of the DNA in 1953 [57] - the single most important molecule for life (see Figure 1.1). The DNA is the molecule that contains all heritable information and a copy of which is present in (almost) every living cell. One key aspects of the DNA is that it is composed

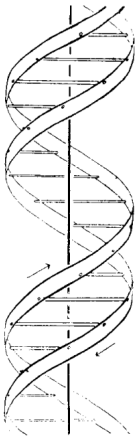


Figure 1.1: DNA double helix from [57]

of four nucleotides (Adenine, Cytosine, Guanine and Thymine) which are arranged to form a linear biopolymer sequence. Moreover, in each cell the DNA is always present in terms of two sequences that are complementary to one another so that they form a special hybrid structure which has become known as the DNA double helix (see Figure 1.1).

Among other functional parts, genes have been identified as distinct regions within the DNA. They do not perform molecular functions themselves (e.g. to catalyze biochemical reactions), but rather do they provide a description of molecular tools that can be build if necessary to catalyze biochemical reactions, unfold into structural components (e.g. cytoskeleton), or regulate developmental processes, to mention just a few. Those molecular tools, which are also referred to as gene products, are in most cases proteins or RNA molecules. They are generated in a process called gene expressions.

An important subgroup of genes are the protein-coding genes (that is, genes whose products are proteins). Those are expressed by an universal two staged molecular mechanism. In the first stage, the *transcription*, the DNA sequence of a gene serves as input

to an enzyme (the RNA polymerase) that produces an RNA molecule (the messenger RNA) as output. RNA molecules are molecularly similar to the DNA, however, they are much shorter and essentially serve as a short term copy of gene sequences that can be transported to other parts of the cell for further processing steps. RNA molecules are frequently referred to as transcripts and the collection of all transcripts within a cell is termed the transcriptome. In the second stage, the *translation*, transcripts are used as input by the ribosome which produces amino acid sequences, which eventually make up proteins [1].

Both, transcription and translation are highly conserved across all living species, from bacteria to mammals and are critical for maintaining life along with a third important process, the DNA replication. DNA replication generates an identical copy of the DNA in order to inherit one copy to its daughter cell upon cell division. Due to their outstanding importance for the existence of life, transcription, translation and DNA replication have become known as the ”central dogma of molecular biology” [12] (see Figure 1.2).

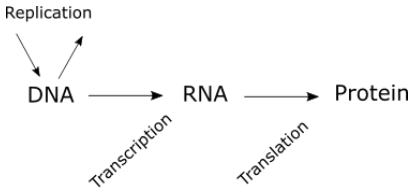


Figure 1.2: Central dogma of molecular biology (Adapted from [12]).

To this day we have acquired a good understanding of the anatomy of genes and the genome size (where genome refers to the ensemble of genes in a given species) for numerous species through DNA sequencing endeavors, including the Bacterophage MS2 - the first sequenced virus in 1976 [18], the first Escherichia coli sequence in 1997 [8],

the first eukaryote sequenced - Saccharomyces cerevisiae in 1996 [20] and the human genome project in 2001 [29, 11, 55], to mention just a few. In recent years, spawned by the advent of next generation sequencing technologies, a large number of fully sequenced organisms has been newly acquired, including by the Genome 10K project [24].

Among the most surprising findings after the completion of the sequencing of eukaryotic organisms were, first, that only a relatively small proportion of the DNA encodes for protein-coding genes. For instance in human, protein-coding genes account only for as little as $\sim$ 3% of the entire DNA sequence which had raised the question as to whether the remaining fraction would be biologically relevant [11, 10]. Second, prior to the publication of the human genome a much larger number of protein-coding genes had been anticipated in complex multicellular organisms (e.g. humans) compared to simpler ones (e.g. yeast). However, work by Ewing *et al.* [17] estimated only about 35000 human genes, while more recent estimates indicate an even lower number of protein-coding genes (see Table 1.1). To put this into perspective, the human genome is only about four times larger the yeast genome, yet humans seem to be much more complex organisms compared to yeast [17, 11, 10]. This observation raises the question, how the large differences in complexity between organisms can be explained, if not by the number of genes.

Table 1.1: Numbers of genes in humans (GENCODE Version 23 - March 2015 freeze, GRCh38 [23]).

Gene types	Numbers of genes
Protein-coding	19797
Long non-coding RNAs	15931
Small non-coding RNAs	9882
Pseudogenes	14477

One explanation that has been proposed is based on alternative splicing. Alternative splicing is a mechanism that selectively alters RNA molecules depending on the cellular context. Thereby parts of the transcript are cleaved out or maintained with respect to the gene sequence, which, as a consequence, may alter the protein function. Alternative splicing may potentially amplify the protein diversity significantly and thus, offers an explanation for the additional complexity [34, 4].

Another explanation rests on a more elaborate gene regulatory network in higher organisms compared to simpler ones (e.g human compared to yeast gene regulation) [1]. While both, alternative splicing and gene regulation, are currently under intensive investigation in the biology and bioinformatics community, we shall focus only on gene regulatory aspects for the remainder of this thesis.

In general, gene regulation refers to the process that establishes and maintains selective expression of genes. Gene regulation can be grouped in gene activation and gene repression. Gene activation is employed if the product of a certain gene is demanded for the proper function of the cell in a given context. As a consequence, the gene becomes expressed. On the other hand, gene repression is employed when it is undesired to express a certain gene in a given cellular state, in which case the gene product becomes absent. Gene regulation may be enacted at various steps throughout the production, maintenance and degradation of transcripts and proteins. In this thesis, we concentrate on the initial step of gene regulation, which is also known as transcriptional regulation. More precisely, we focus only on the *transcription factor*-mediated regulation, while, *epigenetic modifications*, such as DNA methylation and histone modifications [1], are beyond the scope in this thesis.

We continue by briefly reviewing the key aspects of transcription and transcriptional

regulation.

1.1 Key aspects of transcriptional regulation

In a very simplistic view, transcription is regulated by the interplay between two molecular layers: First, by numerous proteins (such as RNA polymerase, various transcription factors, etc.) and second, by regulatory regions within the DNA that contain information about where a gene starts and when its expression is required (see Figure 1.3 and 1.4).

Regarding the proteins that are involved in transcription, RNA polymerase is responsible for generating RNA molecules based on the gene sequences. To this end, initially, RNA polymerase is recruited to the *promoter* of a gene, which refers to the DNA segment that is flanking the start of a gene. Subsequently, upon an activation stimulus from other proteins, RNA polymerase moves along the gene body and produces a transcript of the gene.

Both the recruitment of RNA polymerase to a promoter region and the subsequent transcriptional initiation signal are tightly regulated by a class of proteins that are referred to as *transcription factors* (TFs) which can be roughly grouped into general TFs and context-specific (or cell-type-specific) TFs. General TFs, which include *TFIIA*-*TFIIH*, are proteins that assemble along with RNA polymerase at all promoters (for genes that are expressed). The main purpose of the general TFs is to help RNA polymerase to position correctly and initially pull the DNA strands apart from one another [1]. While, the complex of RNA polymerase and general TFs constitutes a minimal setup that is necessary for transcription, it is often not sufficient to initiate transcription. In most cases, context-specific TFs are additionally required to ultimately trigger transcription.

Context-specific TFs function by binding to regulatory regions (to *promoters* or *enhancers*; see Figure 1.3-1.4) that are in the vicinity of their target genes and further act on the RNA polymerase/general TFs complex (see Figure 1.5). Context-specific TFs are grouped into activators or suppressors and may therefore trigger or suppress gene expression, respectively. Consequently, context-specific TFs play a key role in establishing and maintaining selectivity in the gene expression program, including cell-type specific gene expression, or regulation of developmental processes.

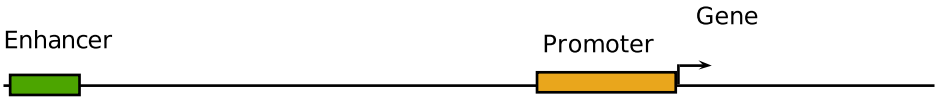


Figure 1.3: Regulatory regions in the DNA can be grouped into promoters and enhancers. Whereas, promoters immediately flank the start of genes, enhancers are usually located some 10kb-100kb up- or downstream of their target genes.

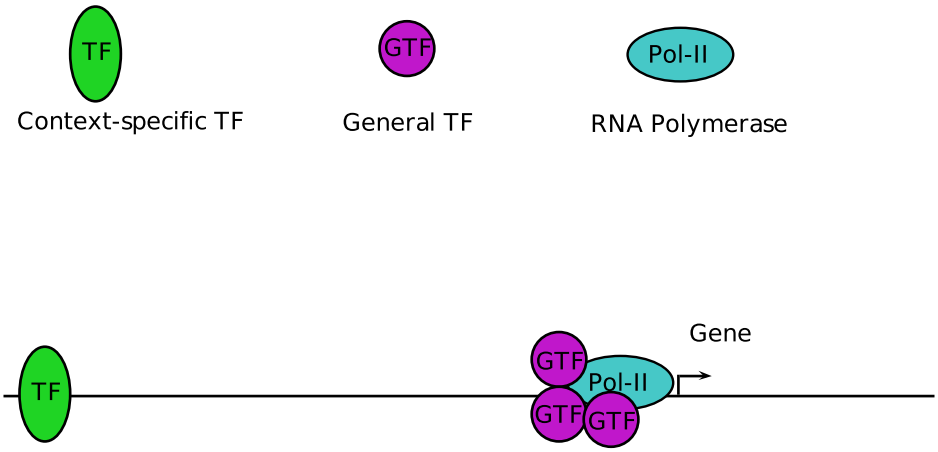


Figure 1.4: Regulatory regions are subjected to binding of several proteins in order to regulate transcription. The most important of these can be summarized as RNA Polymerase II, general transcription factors and context-specific transcription factors.

1.1.1 Transcription factors

In this Section, we briefly discuss a class of proteins that are of particular importance when it comes to gene regulation - the transcription factors. As we have discussed in the previous Section, TFs can be grouped in general and context-specific TFs. For the remainder of this treatment, however, we shall solely focus on context-specific transcription factors, since they determine the specificity of gene regulation. For this reason, we shall use the acronym TFs to mean context-specific TFs unless explicitly mentioned otherwise.

TFs act by migrating into the nucleus and binding to specific sequences in the DNA, so-called transcription factor binding sites (TFBSs), that are contained in the regulatory regions of the DNA. When binding to the DNA, they also form complexes with other proteins, including other transcription factors, RNA polymerase and general TFs and thereby trigger or prevent gene expression.

While, general TFs are always positioned at the promoter regions, context-specific TFs not only can be located at or close to the promoters, but also rather distantly at so-called enhancers, which are found some 10 kb - 100 kb up- or downstream of the respective target gene promoters. Even though enhancers may be far away from a gene start site with respect to the linear DNA sequence, DNA looping is hypothesized to bring enhancer-binding TFs together with the proteins that reside at promoters in the 3D space (see Figure 1.5).

A TF finds its binding site in the genome by its inherent DNA-binding specificity, which is brought about by a variety of chemical interaction mechanisms [35], the most prominent of which is the contact between the TF and the major groove of the DNA double

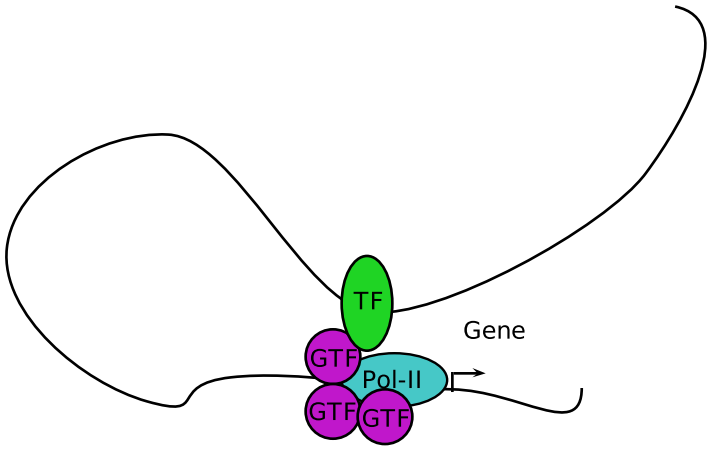


Figure 1.5: DNA looping is facilitated in order to bring context-specific TFs in close proximity to their target promoter regions where they for a complex with the general TFs and RNA Polymerase.

helix. Thereby hydrogen bonds between its residues and the nucleotides in the bound region are formed (see Figure 1.6). More recently, TFs have also been shown to interact with the minor groove by reading the electrostatic profile [42]. Depending on their physical interaction mechanism a TF binds to more or less specific short DNA sequences which are usually between 5bp and 20bp in length. For instance, the human transcription factor *E12* was found to bind the sequences summarized in an alignment in Table 1.2. From the alignment, one can observe that *E12* has a strong preference to bind to the DNA sequence word GCAGGTG at positions 4-10, while it apparently tolerates more variability at the remaining positions.

Even though the DNA binding affinity has been identified already for a large number of TFs [58], predicting which genomic regions are subject to TF binding remains a difficult task. There are several reasons for that: First, most TFs bind not only to one specific sequence, but are rather tolerant to sequence variations at bound sites (compare Table 1.2). Second, not all potential TFBS instances throughout the genome are bound

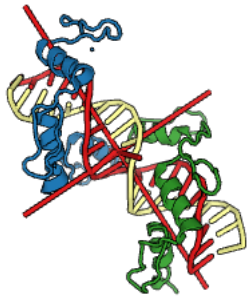


Figure 1.6: Structure of DNA-binding of tandem Zif268 [37]. The DNA double helix is shown in yellow and red, while Zif268 is depicted in green and blue. Structure analysis of the DNA-protein complex reveals that the protein contacts the major groove and thereby senses the DNA sequence.

Table 1.2: Alignment of sequences which are bound by *E12* [48]: Each row corresponds to a TFBS of *E12* and each column corresponds to position within the sequences for that alignment.

G	C	T	G	C	A	G	G	T	G	T	T	C	T	-
A	A	G	G	C	A	G	G	T	G	G	C	C	C	A
G	C	A	G	C	A	G	G	T	G	T	T	C	C	C
A	G	G	G	C	A	G	G	T	G	T	C	C	G	G
G	G	G	G	C	A	G	G	T	G	G	T	G	G	T
C	G	T	A	C	A	G	A	T	G	T	G	G	A	T
A	C	C	G	C	A	G	C	T	G	G	C	C	C	T
C	A	A	G	C	A	G	G	T	G	T	A	T	C	C
C	C	C	A	C	A	G	C	T	G	G	G	A	T	C
C	C	A	G	C	A	G	G	T	G	T	G	T	G	C
A	G	G	G	C	A	G	G	T	G	T	C	C	G	G

by the TF, but only a faction of them. For instance, only a small fraction of matches with the *E12*-bound sequences contained in Alignment 1.2 are bound throughout the genome. Third, not all regions that are bound by a TF actually exhibit TFBSs for that TF. Sometimes, TFs indirectly bind to a genomic locus by tethering to other proteins, in which case, a TFBS is not necessary [46]. Forth, many TFs are part of TF families, including *Fox*-family TFs or *Hox*-family TFs, where TFs in a family share a remarkably similar DNA-binding affinity. Nevertheless, they may still regulate different aspects and therefore bind to distinct regulatory regions. For instance, Crocker *et al.* [13] found that while *Hox*-family TFs share highly similar binding profiles *in vitro*, they mostly occupy non-overlapping regulatory regions *in vivo*.

The discrepancy of predicting TF binding purely from motif matches in the DNA sequence is likely caused by the fact that TFs not only bind to the DNA, but also cooperate with other proteins (e.g. other TFs) at their binding sites. Therefore, the binding strength of a single TF only partially contributes to the binding affinity of an entire complex consisting perhaps of multiple TFs [46]. This in turn means that predicting truly bound sites across the genome based on TFBSs occurrence of one TF alone performs only poorly. However, when the entire sequence context is considered to predict truly bound sites, accurate predictions are possible [2].

1.1.2 Regulatory regions

Regulatory regions refer to the non-coding proportion of DNA that dictates context-specific gene expression. They contain specific DNA features (e.g. TFBSs) that allow transcription factors to recognize and bind them. In general, regulatory sequences are

grouped into promoters and enhancers (see Figure 1.3). Promoter refers to the region that flanks the start of a gene and while enhancer refers to distantly situated regions (e.g. some 10 kb - 100 kb in distance from the gene start site). Promoters can further be categorized into high CpG promoters (that is a promoter that overlaps with a CpG island) and low CpG promoters according to the abundance of CG dinucleotides in the region [44]. Generally, genes attached to high CpG promoters have been found to be expressed ubiquitously across different cell-types and are associated with housekeeping functions [44, 1]. On the other hand, low CpG promoters tend to be attached to cell-type specific genes and often contain a DNA sequence feature that is referred to as the TATA-box, which is recognized by general TFs [1]. Besides sequence features such as the CpG content or the TATA-box, promoters may also contain TFBSs for cell-type specific TFs [54, 46], which ultimately define context-specific gene expression. However, although promoters contain cell-type specific TFBSs to some extend, such DNA features appear to be more frequently present in enhancer regions. Therefore, enhancers have recently been suggested to play a predominant role in the cell-type specific gene regulatory program, whereas promoters predominantly contain general features [16].

In order to understand context-specific gene regulation, it is important to understand 1) which sequences bear the potential of becoming promoters or enhancers and 2) when the regulatory regions become activated (that is, bound by TFs).

With the advent of high throughput sequencing technologies, it has become possible to interrogate the activity of regulatory sequences depending on their context or cell type. For instance, ChIP-seq offers a means to identify subsets of regulatory regions that are bound by a specific protein of interest in a given condition or cell-type. Moreover, since certain histone modifications are known to coincide with active regulatory regions in

general [25], ChIP-seq has also been used to reveal genomic loci that appear to act as enhancers or promoters [10, 7]. Another strategy for identifying putative regulatory regions is based on the observation that most of the DNA is present in a highly compacted form, the heterochromatin, which is inaccessible to e.g. RNA polymerase or TFs. Only a few specific regions, on the other hand, are maintained to be accessible (e.g. due to the binding of TFs) in order to perform regulatory functions. Both, DNase-seq and ATAC-seq, provide a means to determine chromatin accessibility throughout the genome in a given condition and therefore, may be used to identify regulatory regions [45, 10, 7]. In case, chromatin accessibility was interrogated by DNase-seq, accessible regions are referred to as DNase-1-hypersensitive sites (DHSs).

While, computationally predicting putative TFBSs in the entire genome is still very difficult (as we have discussed in the last Section), predicting TFBSs in putative regulatory regions (e.g. obtained by ChIP-seq or DNase-seq) is more promising, because taken together, these regions are comparably short relative to the entire genome size. As a consequence, the signal-to-noise ratio for identifying true TFBSs is substantially increased by discarding the majority of the genomic sequence.

Common bioinformatics strategies, after having identified putative enhancers and promoters, therefore are to identify TFBSs, assess the abundance of TFBSs across a set of regulatory regions, or, studying the co-occurrences between TFBSs for two or more TFs.

We shall dedicate the remainder of this thesis to discussing statistical models for analyzing the presence of putative TFBSs and drawing conclusions about their abundance in a set of sequences.

1.2 Statistical models for sequence analysis

We start the statistical discussion about sequential data (e.g. DNA sequences) by formally introducing a set of letters (or the alphabet) denoted by $\mathcal{A}$. For nucleotides, the alphabet is given by $\mathcal{A} = \{A, C, G, T\}$. A sequence of length N on this alphabet is denoted by $\mathbf{w} = w_1 w_2 \cdots w_N$ where $\forall i : w_i \in \mathcal{A}$. Throughout the remainder of the thesis, we shall discuss the specific case of DNA sequences, although adaptation to e.g. peptide sequences is straight forward. Moreover, since we primarily study TFBSs occurrence, we denote the length of a generic TFBS by M.

In the following, we review the statistical foundations for this thesis, which includes representations of the DNA sequences (the TF motif and the background model), the definition of a motif hits, a random process for generating motif hits, periodicity of DNA words and the statistics related to the number of motif hits.

1.2.1 Representations of transcription factor motifs

Transcription factors prefer to bind to short and more or less specific DNA sequences. In order to reveal the binding preference of a TF, we can build an alignment across a collection of bound sequences similar to the one illustrated in Table 1.2 for the human transcription factor *E12*. A simplifying representation of such an alignment is given by a count matrix which is an $|\mathcal{A}| \times M$ matrix. The entries of the count matrix represent the number of times a certain nucleotide was found at a certain position within the

alignment. For example,

$$\begin{array}{lcccccccccccccc} A| & 4 & 2 & 3 & 2 & 0 & 11 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ C| & 4 & 5 & 2 & 0 & 11 & 0 & 0 & 2 & 0 & 0 & 0 & 4 & 6 & 4 & 4 \\ G| & 3 & 4 & 4 & 9 & 0 & 0 & 11 & 8 & 0 & 11 & 4 & 3 & 2 & 4 & 2 \\ T| & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 11 & 0 & 7 & 3 & 2 & 2 & 3 \end{array} \tag{1.1}$$

represents the corresponding count matrix for the alignment that is illustrated in Table 1.2. From the count matrix, a position frequency matrix (PFM) can be determined by normalizing each column such that it sums up to one. This also renders the PFM a probabilistic description of the TF binding affinity, and thus, may serve as a generative process. A general PFM is defined by

$$\begin{bmatrix} m_{A1} & m_{A2} & \cdots & m_{AM} \\ m_{C1} & m_{C2} & \cdots & m_{CM} \\ m_{G1} & m_{G2} & \cdots & m_{GM} \\ m_{T1} & m_{T2} & \cdots & m_{TM} \end{bmatrix} \tag{1.2}$$

subject to the constraints

$$\forall k : \sum_{w \in \mathcal{A}} m_{wk} = 1 \tag{1.3}$$

$$\forall k \forall w : m_{wk} \geq 0 \tag{1.4}$$

which we shall refer to as a *TF motif* throughout the remainder of this thesis. Note that in the latter representation, each position is governed by an independent multinomial probability over the letters in $\mathcal{A}$.

An example for the TF motif for *E12* is given as follows

$$\begin{array}{lllllllllllllll} A| & 0.36 & 0.18 & 0.28 & 0.18 & 0 & 1 & 0 & 0.1 & 0 & 0 & 0 & 0.1 & 0.1 & 0.1 & 0.1 \\ C| & 0.36 & 0.46 & 0.18 & 0 & 1 & 0 & 0 & 0.18 & 0 & 0 & 0 & 0.36 & 0.54 & 0.36 & 0.4 \\ G| & 0.28 & 0.36 & 0.36 & 0.82 & 0 & 0 & 1 & 0.72 & 0 & 1 & 0.36 & 0.28 & 0.18 & 0.36 & 0.2 \\ T| & 0 & 0 & 0.18 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0.64 & 0.28 & 0.18 & 0.18 & 0.3. \end{array} \tag{1.5}$$

A TF motif can additionally be characterized by its per-position information content which is defined by

$$IC_i = \log_2 |\mathcal{A}| + \sum_{a \in \mathcal{A}} m_{a,i} \log_2(m_{a,i}) \tag{1.6}$$

where $\sum_{a \in \mathcal{A}} m_{a,i} \log_2(m_{a,i})$ denotes the negative entropy of the model at position i [47]. The information content is measured in bits and ranges from zero to two for DNA sequences, with IC_i being close to zero in case the TF tolerates large sequence variation at position i and IC_i being close to $\log_2(|\mathcal{A}|)$ if it is highly specific to one nucleotide. The total information content across the motif is given by

$$IC = \sum_{i=1}^{M} IC_i.$$

The information content is further leveraged in a visual representation of the TF motif that is called the *sequence logo*, where the nucleotide letters with high information content are depicted larger than nucleotides with low information content throughout the TF motif. Accordingly, the height of each nucleotide letter is scaled by $m_{a,i} \times IC_i$ for $a \in \mathcal{A}$, where Definition (1.2) and (1.6) are used [47]. An example of a sequence logo

for *E12* is depicted in Figure 1.7.

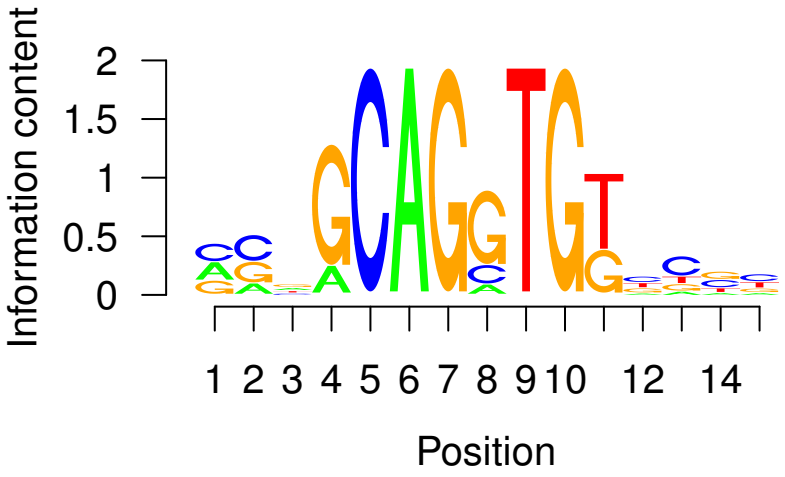


Figure 1.7: Sequence logo of E12

Finally, we state the likelihood of a sequence $\mathbf{a} \in \mathcal{A}^M$ under the TF motif by

$$P_M(\mathbf{a}) = \prod_{i=1}^{M} m_{a_i i}. \tag{1.7}$$

where the subscript M indicates the likelihood with respect to the motif. The likelihood serves as a measure to determine how probable a sequence $\mathbf{a}$ is under the model. Furthermore, a *consensus sequence* refers to a sequence $\mathbf{a}^*$ for which the likelihood function is maximized, such that

$$\mathbf{a}^* = \arg\max_{\mathbf{a} \in \mathcal{A}^M} P_M(\mathbf{a}). \tag{1.8}$$

Note that the *consensus sequence* does not necessarily need to be unique.

1.2.2 Modeling the background sequences

Having introduced the statistical foundation for modeling TF motifs in the previous Section, we shall turn to formalizing what is known as the background model in this Section. In contrast to the TF motifs, which models rather short DNA sequence stretches in length (e.g. 5-20bps), the background model is tailored towards modeling relative long stretches of DNA (e.g. the entire regulatory region which might be some 100 bps in length) by recapitulating some basic DNA features, including nucleotide or dinucleotide frequencies.

The background model is of fundamental importance when it comes to identifying TFBSs and counting the number of TFBSs in the sequence, because it serves as a null model for generating sequences. Accordingly, we study how many TFBSs would to be found just by chance based on the background model, which gives rise to a statistical test as we shall see later.

We shall start our formal discussion by defining the background model as an ergodic and homogeneous order-d Markov model that is parametrized by the transition probabilities

$$\pi(w_{-d} \cdots w_{-1}, w_0) = P(w_0|w_{-1}, ..., w_{-d}) \quad \forall w_{-d} \cdots w_0 \in \mathcal{A}^{d+1}.$$

According to the Markov assumption, drawing a random letter w_0 only depends on the d previous letters, while it is independent of any letters that were observed prior to position $-d$. Furthermore, note that due to the ergodicity assumption, the Markov chain is guaranteed to give rise to a unique stationary distribution denoted by $\mu(w_1 \cdots w_d)$ [26].

The likelihood of a sequence $\mathbf{w} \in \mathcal{A}^N$ under the background model is given by

$$P_B(\mathbf{w}) = \mu(w_1 \cdots w_d) \prod_{i=d+1}^{N} \pi(w_{i-d} \cdots w_{i-1}, w_i) \tag{1.9}$$

where the subscript B on the left hand side indicates the probability with respect to the background model.

The parameters of the background model can be obtained by the maximum likelihood procedure on a set of training sequences (e.g. promoter or enhancer sequences) according to

$$\hat{\pi}(w_{-d} \cdots w_{-1}, w_0) = \frac{f(w_{-d} \cdots w_{-1}, w_0)}{f(w_{-d} \cdots w_{-1})} \tag{1.10}$$

where $f(\mathbf{w})$ denotes the frequency with which a certain word $\mathbf{w}$ is observed in the training set [41].

While at this point the background model is already applicable for modeling single-stranded DNA sequences, when it comes to studying double-stranded sequences, we have yet to impose additional constraints on the model. For the sake of the following discussion, let the forward strand sequence be denoted by $\mathbf{w} = w_1 \cdots w_N$ and its reverse complement by $\mathbf{w}' = w'_1 \cdots w'_N$ as shown in the following example

$$\begin{array}{lccccccc} 3' \leftarrow & w'_1 & w'_2 & \cdots & w'_{N-1} & w'_N & \leftarrow 5' \\ 5' \rightarrow & w_1 & w_2 & \cdots & w_{N-1} & w_N & \rightarrow 3'. \end{array}$$

In many cases (e.g. for regions acquired by ChIP-seq) it is not possible to tell whether the directionality of the region should be read with respect to the forward strand or the reverse strand. Therefore, when searching for instance for TFBSs in the region, we are

indifferent about searching the forward strand or the reverse strand. This issue is solved by simply scanning both strands for TFBSs. Without any constraints on the background model, however, the same sequence is generally weighted by different probabilities on the forward and the reverse strand, respectively. This is equivalent to the observation that a sequence and its reverse complement may be associated with different probabilities (e.g. $P_B(\mathbf{w} = AA) \neq P_B(\mathbf{w} = TT)$). For this reason, it is necessary to enforce that the probability of any word is the same as the probability of its reverse complement, which is given by

$$P_B(\mathbf{w}) = P_B(\mathbf{w}'). \tag{1.11}$$

In general, Equation (1.11) is satisfied if the following condition holds true

$$\mu(w_{-d} \cdots w_{-1})\pi(w_{-d} \cdots w_{-1}, w_0) = \mu(w'_0 \cdots w'_{-d+1})\pi(w'_0 \cdots w'_{-d+1}, w'_{-d}) \tag{1.12}$$

which is reminiscent of the detailed balance condition - a concept that is widely used in statistical physics [53].

Another detail about treating double-stranded sequences is that when $\mathbf{w}$ is generated from 5' to 3', $\mathbf{w}'$ is generated from 3' to 5' at the same time, so, from the end to the beginning. While generating the sequence in the forward direction is simply done by employing $\pi(w_{-d} \cdots w_{-1}, w_0)$, generating the sequence in the reverse direction, on the other hand, requires a different set of transition probabilities $\pi'(w_0 \cdots w_{-d+1}, w_{-d})$ which are determined according to the Bayes theorem

$$\pi'(w_0 \cdots w_{-d+1}, w_{-d}) = \frac{\mu(w_{-d} \cdots w_{-1})\pi(w_{-d} \cdots w_{-1}, w_0)}{\sum_{w_{-d}} \mu(w_{-d} \cdots w_{-1})\pi(w_{-d} \cdots w_{-1}, w_0)} \tag{1.13}$$

Importantly, in general $\pi'(w_0 \cdots w_{-d+1}, w_{-d}) \neq \pi(w_0 \cdots w_{-d+1}, w_{-d})$. However, in order to simplify the further algorithmic treatment of the model, we impose an additional constraint on the model

$$\mu(w_{-d} \cdots w_{-1})\pi(w_{-d} \cdots w_{-1}, w_0) = \mu(w_0 \cdots w_{-d+1})\pi(w_0 \cdots w_{-d+1}, w_{-d}) \tag{1.14}$$

which enforces $\pi(w_0 \cdots w_{-d+1}, w_{-d}) \equiv \pi'(w_0 \cdots w_{-d+1}, w_{-d})$. Comparisons between the background models with and without Constraint 1.14 have shown that the results are very similar (data not shown).

In order to adopt the constraints (1.12) and (1.14) into the maximum likelihood estimation scheme given by Equation (1.10), we constrain the respective k-mer counts according to

$$f(w_0 \cdots w_d) = f(w_d \cdots w_0) = f(w'_0 \cdots w'_d) = f(w'_d \cdots w'_0). \tag{1.15}$$

1.2.3 Log-likelihood ratio and motif hit identification

In this Section, we shall employ the TF model and the order-d background model that we have introduced in the previous sections in order to decide how well each model represents any given sequence $\mathbf{a} \in \mathcal{A}^M$. Throughout the remainder of this thesis, we shall assume that the Markov order $d < M$. This principle is used to identify potential TFBSs in a given DNA sequence which we refer to as *motif hit identification* problem.

We introduce the log-likelihood ratio (also referred to as *score* or *motif score*) as

$$s(\mathbf{a}) := \log\Big(\frac{P_M(\mathbf{a})}{P_B(\mathbf{a})}\Big) \tag{1.16}$$

$$= \sum_{i=1}^{M} \log(m_{a_i i}) - \log(\mu(a_1 \cdots a_d)) - \sum_{i=d+1}^{M} \log(\pi(a_{i-d} \cdots a_{i-1}, a_i)) \tag{1.17}$$

where we make use of Equation (1.7) and (1.9) and the fact that $\log(1/a) = -\log(a)$ and $\log(ab) = \log(a) + \log(b)$ [47]. We assume that both $P_M(\mathbf{a}) > 0$ and $P_B(\mathbf{a}) > 0$ for all $\mathbf{a} \in \mathcal{A}^M$ so as to ensure that the score is always finite, which is achieved by adding pseudo-counts when estimating the TF motif as well as the background model. The *score* represents a measure that intuitively results in positive values if the TF motif provides a better explanation for $\mathbf{a}$ compared to the background and negative values otherwise. Based on the *score*, we can state the corresponding statistical hypothesis test, which we refer to as to the *motif score test*

$$H_0: \quad \text{"}\mathbf{a} \text{ was generated from the background"}$$
$$H_1: \quad \text{"}\mathbf{a} \text{ was generated from the TF motif".}$$

Using a predefined *score threshold* t_s, we reject H_0 if $s(\mathbf{a}) \geq t_s$. We refer to the set of words for which H_0 is rejected as the set of *compatible words* [59], which is defined by

$$C(t_s) = \{\mathbf{a} \in \mathcal{A}^M : \quad s(\mathbf{a}) \geq t_s\}. \tag{1.18}$$

Furthermore, the probability of drawing a false positive TFBS is given by

$$P(s(\mathbf{a}) \geq t_s | H_0) = \alpha. \tag{1.19}$$

In order to determine t_s for an associated significance level α, it is necessary to estimate $P(S|H_0)$ - the *score distribution* under H_0, which depends on both, the TF motif and

the background. We shall discuss the evaluation of the score distribution under the null hypothesis in Chapter 2.

1.2.4 A random process for generating motif hits

In this Section, we consider evaluating the scores at each position along a DNA sequence that is generated according to the background. Theses scores are then tested for significance by utilizing the score threshold t_s as described in the previous Section. Due to the uncertainty about DNA sequence, the scores and the outcomes of significance tests become random variables as well. Importantly, the resulting distribution over the outcomes of the significance test along the sequence is of fundamental importance throughout the remainder of this thesis, since it allows us to establish a null model over the number of motif hits. Therefore, in this Section, we shall formally introduce the random process for producing motif hits.

To start our discussion, let $\mathbf{a} \in \mathcal{A}^M$ be a sequence that was generated by the background model, we introduce the indicator random variable that reflects the outcome of the *motif score test* as follows

$$Y := \mathbf{1}[s(\mathbf{a}) \geq t_s] \tag{1.20}$$

where we made use of Definition (1.17) and $\mathbf{1}[.]$, which denotes the indicator function that evaluates to one only if its argument is true and otherwise to zero [56]. We shall refer to the outcomes of $Y = 1$ and $Y = 0$ as a *motif hit* and a *non-motif hit*, respectively. Note that, while a motif hit represents an instance of a false positive test outcome, a non-motif hit represents a true negative test outcome. According to the previous Section, a

motif hit occurs with probability $P(Y = 1) = \alpha$, while a non-motif hit occurs with probability $P(Y = 0) = 1 - \alpha$.

Next, suppose that $\mathbf{w} = w_1 \cdots w_N$, such that $N > M$, was generated from the background model. Scanning the sequence for occurrences of motif hits (e.g. TFBSs) yields a score $s_i = s(w_i \cdots w_{i+M-1})$, for each position $1 \leq i \leq N - M + 1$ throughout the sequence. For simplicity, we employ Definition 1.17 regardless of the position and the order of the background model. Note, however, that the computation of s_i at a position $i > d$ with an order-d background model ideally requires to replace $\mu(a_1 \cdots a_d)$ by $\prod_{0<i\leq d} \pi(a_{-d+1+i} \cdots a_{i-1}, a_i)$ in Definition (1.17) in order to correctly account for the sequence prefix.

Applying the score threshold on top of s_i yields a Bernoulli process $Y_i = \mathbf{1}[s_i \geq t_s]$ for all i, which shall compactly be denote as

$$\mathbf{Y}_{[1:N-M+1]} := Y_1 \cdots Y_{N-M+1}. \tag{1.21}$$

While above, we have discussed motif hits as a result of scanning a single strand of the DNA sequence, in many cases it is necessary to scan both strands of a given double-stranded DNA sequence, since it is initially not known on which strand a motif hit might reside. To this end, we scan $\mathbf{w}$ twice, once with the original TF motif and once with the reverse complemented TF motif. The result of the latter operation are the additional score values s'_i and the test outcomes Y'_i, where the prime indicates that they were evaluated with respect to the reverse strand. We denote the random process for

generating motif hits due to scanning both strands as

$$\mathbf{Y}_{[1:N-M+1]} := Y_1 \cdots Y_{N-M+1} Y'_1 \cdots Y'_{N-M+1}. \tag{1.22}$$

Probabilistic relationship between the DNA sequence and motif hits

To express the deterministic relationship between the underlying random DNA sequence $\mathbf{w}$ and the outcomes of the significance tests $\mathbf{Y}_{[1:N-M+1]}$, we introduce the conditional probability

$$P(Y_i = 1 | w_i \cdots w_{i+M-1}) := \begin{cases} 1 & \text{if } s_i \geq t_s \\ 0 & ow. \end{cases} \tag{1.23}$$

Due to the uncertainty about the DNA sequence $\mathbf{w}$, however, the events $\mathbf{Y}_{[1:N-M+1]}$ become random variables as well. Therefore, the joint probability over both the DNA sequence and the motif hit indicators is given by

$$P(\mathbf{Y}_{[1:N-M+1]}, w_1 \cdots w_N) = P(\mathbf{Y}_{[1:N-M+1]} | w_1 \cdots w_N) P_B(w_1 \cdots w_N) \tag{1.24}$$

with

$$P(\mathbf{Y}_{[1:N-M+1]} | w_1 \cdots w_N) = \prod_{i=1}^{N-M+1} P(Y_i | w_1 \cdots w_N). \tag{1.25}$$

By averaging over the latent DNA sequence, we obtain

$$P(\mathbf{Y}_{[1:N-M+1]}) = \sum_{w_1 \cdots w_N} P(\mathbf{Y}_{[1:N-M+1]}, w_1 \cdots w_N) \tag{1.26}$$

which denotes the random process for generating motif hits. The probability given by Equation (1.26) is the single most important quantity throughout of the remainder of this thesis. If we could compute this quantity analytically, we could draw exact statistical conclusions about e.g. the number of motif hits, as we shall see later. Unfortunately, however, the explicit averaging over $\mathbf{w}$ is intractable since not only does Equation (1.26) require to enumerate over exponentially many sequences $\mathbf{w}$ with respect to the sequence length, but also does the number of possible assignments of $\mathbf{Y}_{[1:N-M+1]}$ grow exponentially with increasing N.

In order to bypass the need for performing the explicit summation over $\mathbf{w}$ and storing the complete joint probability $P(\mathbf{Y}_{[1:N-M+1]})$, we seek to approximate $P(\mathbf{Y}_{[1:N-M+1]})$. In doing so, we treat $\mathbf{Y}_{[1:N-M+1]}$ as the observed random variable, while the underlying DNA sequence $\mathbf{w}$ represents a latent random variable which induces complicated statistical dependences between the events in $\mathbf{Y}_{[1:N-M+1]}$. Our goal is to 1) identify (approximate) independence relationships about the events Y_i and Y_j in $\mathbf{Y}_{[1:N-M+1]}$ which allow us to factorize $P(\mathbf{Y}_{[1:N-M+1]})$ into simpler constituent parts and 2) to estimate the statistical association between potentially highly correlated events Y_i and Y_j, both of which shall be discussed in Chapter 3.

Bayesian Network representation

Bayesian networks are an excellent tool for illustrating the relationship about random variables, in particular, when it comes to highlighting statistical independence. In Bayesian networks, random variables are depicted as nodes and arrows represent a direct influence of a random variable on another one. Additionally, nodes might be depicted transparent or shaded in gray which corresponds to latent and observed random vari-

ables. For a comprehensive treatment on this subject the reader is referred to Koller *et al.* [28].

We shall use Bayesian networks throughout the thesis in order to depict the relationship between the underlying DNA sequence $\mathbf{w} \in \mathcal{A}^N$, their induced scores or the outcomes of the *motif score tests* $\mathbf{Y}_{[1:N-M+1]}$ to highlight statistical independence properties, because such properties can be leveraged to derive efficient dynamic programming-based algorithms in turn.

The probabilistic relationship between the DNA sequence $\mathbf{w}$ and the outcomes of the significance tests $\mathbf{Y}_{[1:N-M+1]}$ that we have discussed in the previous section are illustrated in terms of a Bayesian network in Figure 1.8. Figures 1.8a and 1.8c depict the induction of uncertainty for the test outcomes $\mathbf{Y}_{[1:N-M+1]}$ when a single strand and both strands are scanned, respectively. Moreover, Figure 1.8b illustrates the fact that conditioning on the underlying DNA sequence not only would determine the sequence of outcomes deterministically, but also the become independent of one another as defined by Equation (1.25).

Clump start and overlapping hits

When a DNA sequence of length N is scanned for motif hits with a motif of length M, in general, motif hits occur in terms of *clumps*, where a clump refers to one or more overlapping motif hit instances [41]. Accordingly, in a clump, there are two types motif hits that might arise: 1) A motif hit that starts the clump and which does not overlap with any motif hits to its left and 2) overlapping motif hits which are preceded by the clump start hit or an overlapping hit. That is, for a clump start hit at position i we have $M - 1$

preceding non-hits $Y_{i-1} = 0, \cdots Y_{i-M+1} = 0$. When both strands of a DNA sequence are scanned the clump start hit requires non-hits on both strands $Y_{i-1} = 0, \cdots Y_{i-M+1} = 0$ and $Y'_{i-1} = 0, \cdots Y'_{i-M+1} = 0$. By contrast, for an overlapping hit to occur at position i we have observed that at least one hit among $Y_{i-1}, \cdots, Y_{i-M+1}$. Likewise, if both DNA strands are scanned the reverse strand outcomes $Y'_{i-1}, \cdots, Y'_{i-M+1}$ need to be taken into consideration as well

1.2.5 Periodicity of word patterns

The phenomenon of self-overlapping words has been described by the concept of *periodicity* in the word pattern community, including in Reinert *et al.* [41], which we shall briefly review in this Section.

A word $\mathbf{a} = a_1 \cdots a_M$ may overlap with itself if and only if $\mathbf{a}$ is periodic. Periodicity holds if there exists an integer $1 \leq p < M$ such that $\forall i \in \{1, \cdots M - p\} : a_i = a_{i+p}$. Moreover, the set of all periods that is associated with $\mathbf{a}$ is given by

$$\mathcal{P}(\mathbf{a}) = \{p \in \{1, \cdots M - 1\} : a_i = a_{i+p}, \forall i \in \{1, \cdots M - p\}\}. \tag{1.27}$$

If the set $\mathcal{P}(\mathbf{a})$ is empty, the word is not self-overlapping [41].

As an example of a self-overlapping word, consider 'AAA' which is associated with the periods $\mathcal{P}('AAA') = \{1, 2\}$.

In addition to periods, the notion of *principal periods* describe all periods $p \in \mathcal{P}(\mathbf{a})$ that cannot be formed as integer multiples of other periods [41]. In other words, principal periods relate to periods by being their root cause. Thus, while the periods may contain

redundant information about self-overlapping words, principal periods describe self-overlapping words non-redundantly. The set of principal periods associated with a word $\mathbf{a}$ is denoted by $\mathcal{P}'(\mathbf{a})$, where $\mathcal{P}'(\mathbf{a}) \subseteq \mathcal{P}(\mathbf{a})$.

For the case of the word 'AAA', we have $\mathcal{P}'('AAA') = \{1\}$, because the second period ($p = 2$) is brought about by concatenating two instances of the principal period $p = 1$.

The discussion on periods so far can be extended to overlapping instances of multiple distinct words $\mathbf{a}^1, \cdots, \mathbf{a}^m$ [41]. In this case, overlap refers to both overlap of any $\mathbf{a}^l$ with itself as well as overlap between $\mathbf{a}^l$ and $\mathbf{a}^k$, with $l \neq k$. Accordingly, the set of periods is given by

$$\mathcal{P}(\mathbf{a}^l, \mathbf{a}^k) = \{p \in \{1, \cdots M-1\} : a_i^l = a_{i+p}^k, \forall i \in \{1, \cdots M-p\}\} \tag{1.28}$$

Analogously, we denote the set of principal periods as $\mathcal{P}'(\mathbf{a}^l, \mathbf{a}^k) \subseteq \mathcal{P}(\mathbf{a}^l, \mathbf{a}^k)$. As before, the principal periods form the root cause for all periods across the set of words $\mathbf{a}^1, \cdots, \mathbf{a}^m$.

The set of principal periods, in both the case of single word patterns and multiple word patterns, have been important to model the tendency of a word to form *clumps*, which was leveraged for instance in the compound Poisson model for describing the distribution over number of exact word matches in a random DNA sequence [41]. Theoretically, the periodicity concept can be applied to studying motif hit occurrence as well, where the periodicity across all compatible words $C(t_s)$ needs to be established. However, as enumerating all compatible words in general is computationally intractable [59], the *peridicity*-concept cannot be adapted exactly to motif hits that are obtained by the *motif score test* defined by (1.17). However, we shall extend this concept approximately to

motif hits in Chapter 3.

1.2.6 Counting motif hits and motif hit enrichment

Based on the fact that a given TF preferentially binds towards a more or less specific sequence (e.g. TFBSs), a set of regulatory regions that is bound by the protein of interest is often enriched for specific TFBSs. In order to measure the level of abundance of TFBSs, we count the number of motif hits which gives rise to the *motif hits count*. We introduce the *motif hits count* random variable that is based on the random process $P(\mathbf{Y}_{[1:N-M+1]})$ as

$$X = \sum_{i=1}^{N+M-1} Y_i. \tag{1.29}$$

Analogously, when scanning both DNA strands for the presence of motif hits we have

$$X = \sum_{i=1}^{N+M-1} Y_i + Y_i'. \tag{1.30}$$

We employ the following statistical hypothesis test in order to quantify the level of abundance of TFBSs in the DNA segments

H_0 : "X hits were produced by the background model"

H_1 : "X hits were not produced by the background model".

Accordingly, we judge whether the number of hits X is well explained by H_0, and reject H_0 if it appears to be unlikely that it has produced X hits. While, this statistical test is

an instance of a two-sided test, in which case H_1 represents both, observing too many or too few motif hits, we are often specifically interested in the one-sided test where we ask whether there are too many motif hits emitted. This latter statistical test shall be referred to as the *motif hit enrichment* test in which we reject H_0 if the number of motif hits exceeds a threshold count x. The false positive probability for the *motif hit enrichment* test is given by $P(X \geq x|H_0)$.

Unfortunately, $P(X|H_0)$ cannot be computed analytically in an efficient way, since it rests on the computation of $P(\mathbf{Y}_{[1:N-M+1]})$. Nevertheless, a number of approximative distributions have been proposed, which achieve more or less accurate results in practice. In the following Sections, we shall briefly introduced approximations of $P(X|H_0)$.

Simulation of the motif hits count distribution

The perhaps most straight forward approximation to the motif hits count distribution can be obtained by sampling many DNA sequences according to the background model and count how many motif hits arise. The resulting distribution corresponds to the *empirical motif hits count distribution*.

With an increasing number of drawn DNA sequences, the empirical distribution eventually converges to the true distribution $P(X|H_0)$. However, the drawback of this approach is that it may require significant computational resources for obtaining accurate estimates of the $P(X|H_0)$. This is in particular true for stringent choices of α, because in this case, motif hits occur only rarely which require a substantial number of samples to accurately estimate the empirical distribution.

This drawback also motivated the development of sophisticated analytic models of the

motif hits count distribution, that at the same time yield accurate results and require significantly less computational resources.

Nevertheless, we shall use the empirical motif hits count distribution in order to validate the accuracy of several analytical models that we discuss throughout Chapter 4 and 6.

Binomial distribution

The simplest analytic approximation for $P(X|H_0)$ is given by the binomial distribution

$$P_{Binom}(x) = Binom(x, N - M + 1, p) \tag{1.31}$$

where x denotes the number of motif hits, N denotes the length of the DNA, M the length of the TF motif and p denotes the success rate for obtaining a motif hit [56, 41]. Depending on whether we scan a single-strand only or both DNA strands, we have $p = \alpha$ or $p = 2\alpha$, respectively. The binomial distribution, rests on the assumption that the outcome of Y_i is independent of the outcome Y_j if $i \neq j$, in the process $\mathbf{Y}_{[1:N-M+1]}$. While, the violation of the independence assumption does not pose an big problem for non-self-overlapping motifs and sufficiently stringent significance levels (with $pN \leq O(1)$ [40]), for self-overlapping motifs (such as palindromes or repetitive motifs), for which observing $Y_i = 1$ likely induces another (overlapping) hit $Y_j = 1$, the assumption incurs significant biases, so that $P_{Binom}(x)$ may not represent the true distribution $P(X|H_0)$ accurately anymore. In turn, the *motif hit enrichment* tests that are drawn on $P_{Binom}(x)$ may yield misleading results [36].

Compound Poisson model

A more sophisticated approximative model for $P(X|H_0)$ is given by the compound Poisson model [56, 41]. The compound Poisson model incorporates the tendency of a TF motif to yield mutually self-overlapping motif hits, which makes the model applicable to a broader class of TF motifs (e.g. self-overlapping motifs and non-self-overlapping motifs).

To treat self-overlapping hits in the compound Poisson model, we make use of the definition of a *clump* as one or more motif hits that overlap one another [41]. While motif hits within clumps are self-overlapping, clumps cannot overlap one another, since in that case, they would simply merge to a clump that that combines all hits. Furthermore, a *k-clump* is defined as a clump that contains exactly k motif hits.

As an example for the concept of clumps, consider a motif hit identification setup in which there exists only one compatible word, namely $C(t_s) = \{'AA'\}$. The fictive sequence "ACGAATCAAAA" then contains 4 motif hits that constitute a 1-*clump* starting at position 4 and a 3-*clump* starting at position 8, respectively.

The compound Poisson model attempts to capture both the rate of *clump* formation as well as the distribution over the clump sizes, by two independent random variables, in order to reconstruct the motif hits count distribution. To this end, assuming that motif hits occur only rarely, the number of *k-clump* occurrences in the DNA sequence is approximately distributed according to

$$C_k \sim \text{Poisson}(\lambda_k) \tag{1.32}$$

where λ_k denotes the occurrence rate of obtaining *k-clumps* and clumps with different k occur independently of one another [41]. Subsequently, the motif hits count can be expressed as

$$X = \sum_{k>0} kC_k. \tag{1.33}$$

The aim of the compound Poisson framework is to obtain accurate estimates of λ_k, which is the subject of Chapter 4, where we discuss a compound Poisson approximation that is based on Pape *et al.* [36].

1.3 Positioning within pattern matching problems

In this Section, we summarize bioinformatics resources and tools that have been established in order to study the sequence content of regulatory regions of with respect to TFBSs. The bioinformatics approaches for studying TFBSs in regulatory regions can be broadly grouped into *motif discovery*, *motif hit identification* and *motif hit enrichment*, where the first one is just briefly mentioned for completeness.

Motif discovery refers to the task of identifying one or more unknown TF motifs from a set of DNA sequences. The result of this task is an alignment, similar to the one shown in Table 1.2 without prior knowledge about the TF affinity. Therefore, motif discovery can be used to reveal novel TF motifs and their respective TFBSs in the DNA sequence. There exist a plethora of tools that implement motif discovery [15], including MEME [3], GibbsSampler [30] and RSAT [49]. Alternative to the TF motif given by a PFM, a motif discovery tools also exist for motifs represented as generalized strings

(e.g. IUPAC strings) [32].

These days, numerous TFs have already been associated with their DNA binding preferences. Representations for these TF motifs can be found in databases, including Transfac [58] and Jaspar [43] across many TFs and species. Such TF motifs provide a valuable resource for analyzing regulatory regions with respect to the presence and abundance of TFBSs, since that may indicate the involvement of a regulatory region in a biological function.

Given a known TF motif, one might be interested in identifying putative TFBSs in a given DNA sequence, which we referred to as the *motif hit identification* problem. Identifying motif hits has been done based on several approaches, including by searching for occurrences of consensus sequences of a TF motif [52], by searching for generalized strings (e.g. IUPAC strings of DNA) [38], based on the information content [9] and based on the log-likelihood ratio [21, 49]. While the first three approaches only employ information contained in the motif itself in order to identify putative TFBSs (e.g. by employing string matching algorithms), the log-likelihood ratio additionally requires the specification of a background model (see Section 1.2.3) [47]. While, the TF motif defines the binding preference, a background model accounts for typical sequence features in the regulatory sequence, for instance, mono- or di-nucleotide composition. The role of the background model is to ensure that commonly occurring DNA sequences are less likely to give rise to putative TFBSs.

Background models have often been represented as order-0 or order-d (sometimes referred to as higher-order) Markov models in order to represent common DNA sequence features. Whereas, an order-0 background model solely accounts for nucleotide frequency in the DNA (e.g. as used in FIMO [21]), order-d Markov models additionally

account for d + 1-mer frequencies (e.g. as used in RSAT [49]). Higher-order Markov models are more adequate for representing DNA segments that include common higher-order features compared to order-0 models. Ignoring such higher-order DNA features may affect *motif hit identification* and *motif hit enrichment* and yield an excess of false positive predictions (see Chapter 4). An example of regions that include higher-order features are CpG islands, which are characterized by an unusually high abundance of CG dinucleotides and which often overlap with regulatory regions (e.g. promoters [44]).

In order to decide whether a DNA region gives rise to a putative TFBS, a score threshold t_s needs to be specified in advance. Each score threshold t_s is associated with a significance level such that $P(S \geq t_s) \leq \alpha$. That is, a threshold t_s gives rise to a false positive prediction with probability α. Therefore, it is desired to compute the score distribution $P(S \geq t_s)$ based on which the score threshold can be chosen. Based on the work of Rahmann *et al.* [40] there exist efficient and exact algorithms for determining the score distribution, $P(S|H_0)$, and subsequently a desired score threshold t_s. We discuss algorithms for computing the score distribution based on both order-0 and higher-order Markov models in Chapter 2.

The third class of methods concerns *motif hit enrichment* in a DNA sequence, which may answer whether TFBSs arise more frequently in a given set of DNA sequences than a given background model is able to explain. Originally, this problem was phrased as finding the distribution of the number of exact word matches in a random sequence, under the restriction that only a relatively small number of words were subject to investigation. A number of approximative distributions concerning the number word occurrences have been proposed which are summarized in an excellent review by Reinert *et al.* [41], including the compound Poisson approximation, the Poisson approximation

or the normal approximation. More recently, additionally, various dynamic programming algorithms have been proposed which achieve exact distributions of the number of word matches in random sequences, including Zhang *et al.* [59] which assumes an order-1 Markov model as background and Marschall [32] which addresses enrichment of generalized strings (e.g. IUPAC strings of DNA).

The above stated approaches on the word-count statistics yield accurate (or even exact) results if one is interested in the enrichment of single words or relatively few words. Unfortunately, many TFs not only bind to one or a few DNA sequences, but rather to a potentially large set of distinct sequence with moderate affinity. This is especially the case, for long TF recognition sites (e.g. $>$ 20bps). As the runtimes of the above stated algorithms depend on the enumeration of all potential binding sites instances, computing the distributions might become intractable if the set of words is too big. For instance, Zhang *et al.* [59] has shown, that enumerating the set of compatible words is in general *NP-hard* for fixed score thresholds. In another publication, Marschall [32] came to a similar conclusion for motifs represented as deterministic finite automatons whos construction requires exponential runtime with respect to the motif length.

To bypass the need for enumerating all compatible words, Pape *et al.* [36] instead proposed to exploit the score distribution in order to efficiently derive a set of overlapping hit probabilities, using an adaptation of the dynamic programming approach stated in Rahmann *et al.* [40]. In their work, they further showed that the overlapping hit probabilities can be used to approximate the clump size distribution, and subsequently, the compound Poisson distribution of the number of motif hits. Therefore, *motif hit enrichment* can be determined efficiently and accurately, even for long TF motifs and with fairly large numbers of compatible words for which the word-based approaches would

be too time-consuming.

1.4 Outline

Throughout the thesis, our foremost interests rest on 1) the identification of TFBSs in regulatory regions (the *motif hit identification* problem) and 2) studying of the abundance of TFBSs in regulatory regions (the *motif hit enrichment* problem).

In Chapter 2, we discuss the score distribution and review methods for computing it efficiently based on dynamic programming using general order-d background models [40, 49]. The score distribution is leveraged to find a score threshold t_s for a desired significance level α, which is subsequently used to identify putative TFBSs in the DNA.

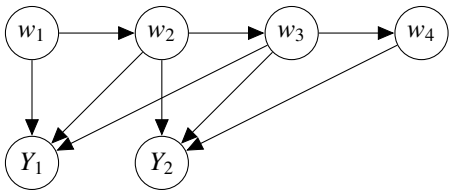
Chapter 3 discusses statistical dependence and independence properties between the outcomes in the process $P(\mathbf{Y}_{[1:N-M+1]})$. We shall discuss why observing $Y_i = 1$ approximately yields independence between Y_j and Y_k with $j < i < k$, which shall be exploited later in the thesis. Moreover, we study correlations between non-overlapping and overlapping events Y_i and Y_j. While, we argue that associations between non-overlapping events are often negligible, this is not the case for mutually overlapping events. Therefore, we discuss algorithms to quantify associations between overlapping outcomes Y_i and Y_j.

In Chapter 4, we follow up on the compound Poisson approximation framework that was put forward by Pape *et al.* [36] and discuss improvements upon the previously proposed approximation. Most importantly, we incorporate general order-d background models as well as refined version for determining overlapping hit probabilities into the

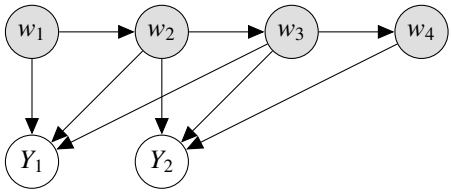
compound Poisson framework. Thorough comparisons between several analytic models show that the novel compound Poisson model generally yields advantageous result across a range of parameters and motifs.

In Chapter 5, we propose a novel Markov model for analyzing the Bernoulli process $\mathbf{Y}_{[1:N-M+1]}$. The Markov model uses a correspondence between its state space and patterns of outcomes in $\mathbf{Y}_{[1:N-M+1]}$. Accordingly, one can analyze the Markov model in order to answer questions about the underlying Bernoulli process $\mathbf{Y}_{[1:N-M+1]}$. We shall use the Markov model in particular to identify the *clump start probability*, which, as opposed to *overlapping hit probabilities*, are associated with hits that are not overlapped by a previously occurring hit to the left. The resulting *clump start probability* is of critical importance for the *combinatorial model*, which is introduced in the final Chapter.

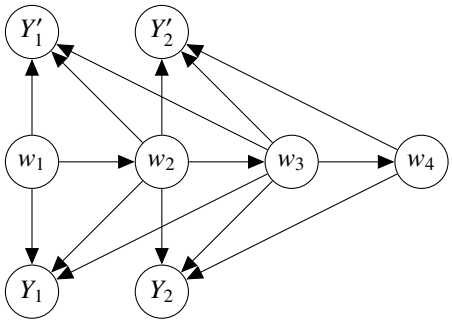
In Chapter 6, we propose a novel statistical model for the motif hits count distribution, which we term the *combinatorial model*. In contrast to the compound Poisson model, which is an asymptotic distribution, the *combinatorial model* efficiently sums over all possible ways of obtaining x hits in a given process $\mathbf{Y}_{[1:N-M+1]}$ by means of dynamic programming, and thereby, explicitly models finite-length sequences. In addition, it does not rely on the validity of the Poisson assumption, which demands motif hits to only occur rarely, and which is required for the compound Poisson model. Through systematic comparisons, we show that the *combinatorial model* generally yields similar results to the compound Poisson model. However, the *combinatorial model* is particularly more accurate in the regime where the violated Poisson assumption affects the accuracy of the compound Poisson model.



(a) A generative process for motif hits



(b) Conditioned on the underlying sequence, the events $\mathbf{Y}_{1:N-M+1}$ become independent of one another.



(c) A generative process for motif hits in double stranded sequences

Figure 1.8: Bayesian network representation of a random nucleotide sequence w_i that is generated by an order-1 background model which induces random outcomes of the significance tests, indicated by $\mathbf{Y}_{[1:N-M+1]}$. In the example, a generic TF motif of length $M = 3$ is shown. Further details on Bayesian networks can be found in Koller *et al.* [28].

Chapter 2

The score distribution

In this Chapter, we discuss algorithms for computing the score distribution $P(S|H_0)$. We shall use the outcome of this computation for identifying a score threshold t_s that is associated with a desired significance level α, which is in turn used to identify motif hits in the DNA.

An algorithm to determine the score distribution based on an order-0 background model was given by Rahmann *et al.* [40]. Multiple sources have also pointed out the possibility to extend this algorithm for the use of general order-d background models, including Beckstette *et al.* [6] and Touzet *et al.* [51]. An implementation of the algorithm for computing the score distribution based on an order-d background model was reported by the authors of RSAT [49]. Whereas, the algorithm was briefly outlined in the manual of RSAT, to our knowledge, a detailed description of the algorithm has never been published.

Because of the central role of the score distribution in my thesis we shall discuss the

approach of Rahmann *et al.* [40] and the extension to higher-order background models in detail in this Chapter.

2.1 Discretization of the score values

Our goal in this Chapter is to analytically determine the score distribution that is produced by generating words $\mathbf{a} \in \mathcal{A}^M$ from the background model. Since the score $s(\mathbf{a})$ for each word $\mathbf{a}$ is real-valued, there are potentially exponentially many distinct score values across all words $\mathbf{a}$ of length M. Therefore, in order to represent and determine the exact score distribution, both exponentially increasing memory and runtime usage would be required with increasing word length M, which renders the exact computation intractable.

However, since the scores are real-valued, we can approximate the score distribution by discretizing the score range into a fixed number of bins. In this way, we can represent the score distribution even for rather long motifs, e.g. $M > 20$, which provides room for the development of efficient algorithms that run in polynomial time, as we shall see in the remainder of this Chapter.

We discretized the range of scores into

$$G = \frac{\max_{\mathbf{a}\in\mathcal{A}^M}(s(\mathbf{a})) - \min_{\mathbf{a}\in\mathcal{A}^M}(s(\mathbf{a}))}{\Delta s} \tag{2.1}$$

non-overlapping bins, which comes about by dividing the total score range by a predefined score granularity Δs. Note that the maximum and minimum score in Equation (2.1) can obtained efficiently in $O(M|\mathcal{A}|^{d+1})$.

The entire scores range is subsequently represented by G discrete bins. Accordingly, the g^{th} bin subject to $0 \leq g < G$, represents any score within the interval

$$(\min_{\mathbf{a}\in\mathcal{A}^M}(s(\mathbf{a})) + g\Delta s, \min_{\mathbf{a}\in\mathcal{A}^M}(s(\mathbf{a})) + (g+1)\Delta s]. \tag{2.2}$$

Even though the discretization can theoretically be chosen such that Equation (2.5) is exactly equivalent [40], for most practical cases, the score range will be discretized such that the distribution of the discretized score values amounts an approximation of the distribution of continuous scores. The discretization error increases for increasing score granularity Δs. A rigorous analysis of the discretization error that results from this approach is given elsewhere [51].

2.2 Computation of the score distribution

Naively, the construction of the score distribution can be achieved by enumerating all sequences $\mathbf{a} \in \mathcal{A}^M$, computing their associated scores $s(\mathbf{a})$ and likelihood values $P_B(\mathbf{a})$ and adding their result to the final score distribution. However, the brute force approach is in general intractable, since the number of sequence to be enumerated grows exponentially with increasing M. Fortunately, due to the statistical independence assumptions on the TF motif and the background model, it is possible to employ dynamic programming in order to construct $P(S|H_0)$.

2.2.1 Score distribution based on an order-0 background model

In this Section, we review the approach by Rahmann *et al.* [40] for computing the score distribution based on an order-0 background model. The probability of producing any sequence $\mathbf{a} \in \mathcal{A}^M$ according to the order-0 background is given by

$$P_B(\mathbf{a}) = \prod_{m=1}^{M} \mu(a_m) \tag{2.3}$$

where we made use of Definition (1.9) in which $\mu(a)$ denotes the probability of observing nucleotide $a \in \mathcal{A}$ irrespective of its position. Furthermore, the log-likelihood ratio is given by

$$s(\mathbf{a}) = \log\left(\frac{P_M(\mathbf{a})}{P_B(\mathbf{a})}\right) \approx \sum_{m=1}^{M} l_m(a_m) \tag{2.4}$$

where the summands on the right hand side of Equation (2.4) are given by

$$l_m(a) := \lfloor \log(m_{am}) - \log(\mu(a)) \rfloor. \tag{2.5}$$

We refer to Definition (2.5) as to the *local score contributions* which express the per position log-likelihood ratios. The floor operator in Definition (2.5) explicitly indicate the score discretization, which for most practical cases is chosen such that score distribution is an approximation rather than exact.

As we have described in Section 1.2.4, even though the local score contribution l_i is deterministically given for a specific nucleotide a_i for $1 \leq i \leq M$, the uncertainty about a_i induces stochasticity into the scores as well. Therefore, we introduce the stochastic local score contribution L_i to highlight the stochastic relationship between the DNA

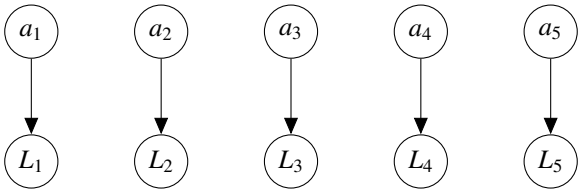


Figure 2.1: Bayesian network representation of the nucleotide sequence $a_1 \cdots a_5$ drawn from an order-0 background model and the respective local score contributions $L_1 \cdots L_5$ for a TF motif of length $M = 5$. The graph highlights that the uncertainty about a_i induces an uncertainty about L_i. Moreover, it shows that neighboring positions are independent of one another.

sequence $\mathbf{a} \in \mathcal{A}^M$ and $L_1, \cdots, L_M$. The relationship between a nucleotide a and L_m is defined by the following conditional local score contribution probability

$$P(L_m = l|a) := \begin{cases} 1 & \text{if } l = l_m(a) \\ 0 & ow. \end{cases} \tag{2.6}$$

which subsequently yields the joint distribution between a and L_m

$$P(L_m, a) = P(L_m|a)\mu(a)$$

that is depicted in terms of a Bayesian network to illustrate the independence structure and causality (see Figure 2.1).

Our goal is to determine the distribution of $S = L_1 + \cdots + L_M$, which requires to average over the random variables $a_1 \cdots a_M$. Since, the nucleotides at each position are independent from one another, the summations can be performed individually for each position, which results in

$$P(L_m) = \sum_{a \in \mathcal{A}} P(L_m|a)\mu(a) \tag{2.7}$$

Figure 2.2: Illustration of Bayesian network from Figure 2.1 after individually summing over the nucleotides $a_1 \cdots a_5$. Note that the local score contributions $L_1 \cdots L_5$ defined by Equation (2.7) remain independent from one another.

and which is further illustrated in Figure 2.2. Averaging over the underlying nucleotides still renders L_i independent of L_j for any $i \neq j$, which allows to compute the distribution of the sum $L_i + L_j$ by employing the convolution between the individual distributions $P(L_i)$ and $P(L_j)$ [22].

In order to determine the score distribution, the following sequence of convolution operations needs to be determined

$$P(S\,|H_0) = P(L_1) * P(L_2) \cdots * P(L_M) \tag{2.8}$$

where $*$ denotes the convolution operation. Equation (2.8) can further be reduced to a recursive algorithm, since the convolution operation is associative. To this end, to initialize the recursive algorithm, let $Q_1(S) = P(L_1)$. Subsequently, we evaluate the following recursion

$$Q_{i+1}(S) = Q_i(S) * P(L_{i+1}) \tag{2.9}$$

which establishes the desired score distribution

$$P(S\,|H_0) = Q_M(S). \tag{2.10}$$

The algorithm represents an instance of dynamic programming and requires a runtime of

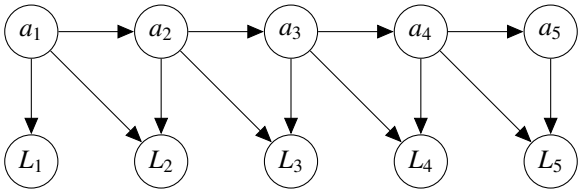


Figure 2.3: Bayesian network representation of the nucleotide sequence $a_1 \cdots a_5$ that is generated from an order-1 Markov model and their respective local score contributions $L_1 \cdots L_5$, for a TF motif of length $M = 5$ [28]. Due to the dependence between nucleotides, the local score contributions are no longer independent of one another.

$O(|\mathcal{A}|MG)$ where $|\mathcal{A}|$ denotes the size of the alphabet, M denotes the length of the motif and G denotes the number of bins for discetizing the scores [40]. Accordingly, whereas, reducing Δs for discretizing the score range yields increasingly accurate results, it also increases the number of bins G and therefore, the runtime requirement.

2.2.2 Score distribution based on an order-d background model

In this Section, we discuss the computation of the score distribution using an order-d Markov model as background. With an order-0 background model local score contributions L_i and L_j for $i \neq j$ were statistical independent of one another which allowed the formulation of an efficient dynamic programming algorithm. However, statistical independence between L_i and L_j no longer holds for order-d Markov models. Fortunately, however, local score contributions are still conditionally independent given the sequence context of the d previous nucleotides according to the background model.

According to an order-d Markov model, the probability $P_B(\mathbf{a})$ of a word $\mathbf{a} \in \mathcal{A}^M$ is given by Definition (1.9) and its score according to Definition (1.17) by

$$s(\mathbf{a}) := \log\left(\frac{P_M(\mathbf{a})}{P_B(\mathbf{a})}\right) \approx l_d(a_1 \cdots a_d) + \sum_{m=d+1}^{M} l_m(a_{m-d} \cdots a_m) \tag{2.11}$$

where

$$l_j(a_{j-d}\cdots a_j) := \begin{cases} 0 & \text{if } j < d \\ \lfloor \log\left(\frac{\prod_{i=1}^{d} m_{a_i i}}{\mu(a_1\cdots a_d)}\right)\rfloor & \text{if } j = d \\ \lfloor \log\left(\frac{m_{a_j j}}{\pi(a_{j-d}\cdots a_{j-1}, a_j)}\right)\rfloor & \text{if } d < j \leq M \\ 0 & \text{if } j > M \end{cases} \tag{2.12}$$

denotes the *local score contribution*. As in the previous section, the floor operator represents the score discretization, which for most practical cases is chosen such that score distribution is an approximation rather than exact. For the same reason as in the previous Section, we express the deterministic dependence of the local score contribution on a nucleotide context of $d + 1$ nucleotides as follows

$$P(L_i|a_{i-d}\cdots a_i) := \begin{cases} 1 & \text{if } L_i = l_i(a_{i-d}\cdots a_i) \\ 0 & ow. \end{cases} \quad \text{for } d \leq i \leq M. \tag{2.13}$$

Therefore, although L_i is deterministically given by the nucleotide context, the uncertainty about the sequence $a_1 \cdots a_M$ induces a score distribution in turn (see Figure 2.3).

In contrast to the order-0 background, where we could immediately sum over the nucleotides at each position individually, this is no longer possible in the general order-d case. Instead, however, notice that L_i and L_{i+k} with $k > 0$ become conditionally independent given the context word $w_{i-d+1} \cdots w_i$ (see Figure 2.4). The conditional independence, in turn, allows to exploit convolution once again to determine the distribution of the sum $L_i + L_{i+k}$ with the only difference that one needs to condition on a prefix word of length d.

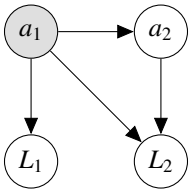


Figure 2.4: Bayesian network illustration of the conditional independence between local score contributions L_1 and L_2 for an order-1 Markov model and $M = 2$ [28]. The gray shaded node a_1 represents that it is observed, which prevents information from passing between L_1 and L_2. That is, L_1 and L_2 are conditionally independent given a_1. Consequently, convolution can be employed to compute the distribution of the sum $L_1 + L_2$ if we condition on a_1.

We proceed by stating an efficient recursive algorithm for determining the score distribution with an order-d background model. In the initialization step of the algorithm, let

$$Q_d(S, a_1 \cdots a_d) = P(L_d|a_1 \cdots a_d)\mu(a_1 \cdots a_d) \quad \forall a_1 \cdots a_d \in \mathcal{A}^d. \tag{2.14}$$

Subsequently, we proceed by recursively evaluating for $a_1 a_2 \cdots a_{d+1} \in \mathcal{A}^{d+1}$ and x ranging over the discretized scores

$$Q_{i+1}(S, a_2 \cdots a_{d+1}) := \sum_{a_1 \in \mathcal{A}} \sum_{x} Q_i(S - x, a_1 \cdots a_d) \times P(L_{i+1} = x|a_1 \cdots a_{d+1}) \times \pi(a_1 \cdots a_d, a_{d+1}). \tag{2.15}$$

Note that Equation (2.15) performs two operations at once: First, it employs convolution (through the summation over l) while conditioning on a given sequence context $a_1 \cdots a_d$. Second, the nucleotide a_1 is summed out, since it only directly influences nucleotide $a_2 \cdots a_{d+1}$. Whereas, all nucleotides beyond position $d + 1$ are not directly due to the Markov assumption of the background mode..

Ultimately, the final recursion step yields $Q_M(S, a_1 \cdots a_d)$ and subsequently, upon summing over all words $a_1 \cdots a_d$, the score distribution $P(S|H_0)$ for an order-d background model is established.

2.2.3 Runtime

The runtime for the algorithm to determine the score distribution using an order-d Markov model is given by $O(G|\mathcal{A}^{d+1}|M)$ [6, 51], which is dominated by the choice of Δs and d. While, decreasing Δs and increasing d increases the accuracy of the algorithm, at the same time, they increase the runtime. We found that usually choosing $\Delta s = 0.1$ and Markov model orders $d \in \{0, 1, 2\}$ in most cases yield accurate results at reasonable computational cost.

2.3 Determining the score threshold

Prior to motif hit identification with a given TF motif in a given DNA sequence, one needs to decide on a score threshold t_s, which is used for the *motif score test* to predict TFBSs (see Section 1.2.3). The ability to determine the score distribution allows us to choose a score threshold based on a significance level α. Therefore, given a desired significance level, e.g. $\alpha = 0.01$, one simply needs to identify its associated quantile t_s such that $P_B(S \geq t_s) = \alpha$. Due to the discrete nature of the scores, this relation almost never holds true exactly. Therefore, we determine t_s according to

$$\arg\min_{t_s} \left\{ \sum_{s=t_s}^{\infty} P_B(S = s) \leq \alpha \right\}. \tag{2.16}$$

A score threshold that is chosen based on Equation 2.16, consequently, guarantees that the false positive probability of obtaining a motif hit from the background model is at most α.

2.4 Examples of the score distributions

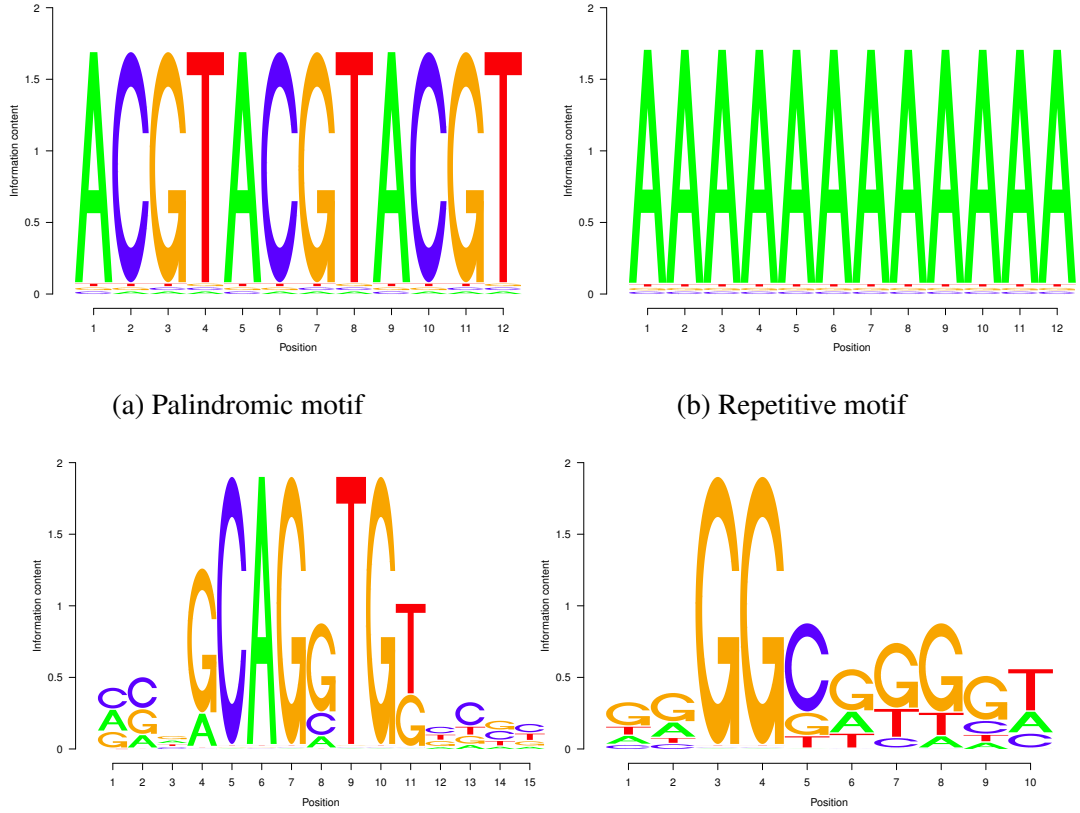
In this Section, we illustrate the score distribution associated with several example motifs (see the motifs in Figure 2.5 and their associated score distributions in Figure 2.6). In each case, an order-1 background model was estimated on DNase-I hypersensitive sites from ENCODE [50], that is on hypothetical regulatory regions.

The score distributions were computed with the procedure from the last Section. We used a score granularity of $\Delta s = 0.1$ and a significance level of $\alpha = 0.05$.

As is apparent from inspecting the score distributions, the shape of the score distribution as well as the score threshold for a fixed significance level depend heavily on the parametrization of both the TF motif and the background model. Concerning the shape of the score distribution, one may appreciate that for TF motifs with high overall information content, the score distribution appears to be highly multimodal, while for low information content motifs the score distribution approaches a Gaussian distribution. Each mode can be approximately attributed to the number of matches between nucleotides in a given DNA sequence and the consensus sequence.

With increasing motif length, the score distribution eventually approaches a Gaussian distribution, due to the central limit theorem. That is even the case of high information content TF motifs where the modes eventually merge to an overall Gaussian envelop

distribution.



(a) Palindromic motif

(b) Repetitive motif

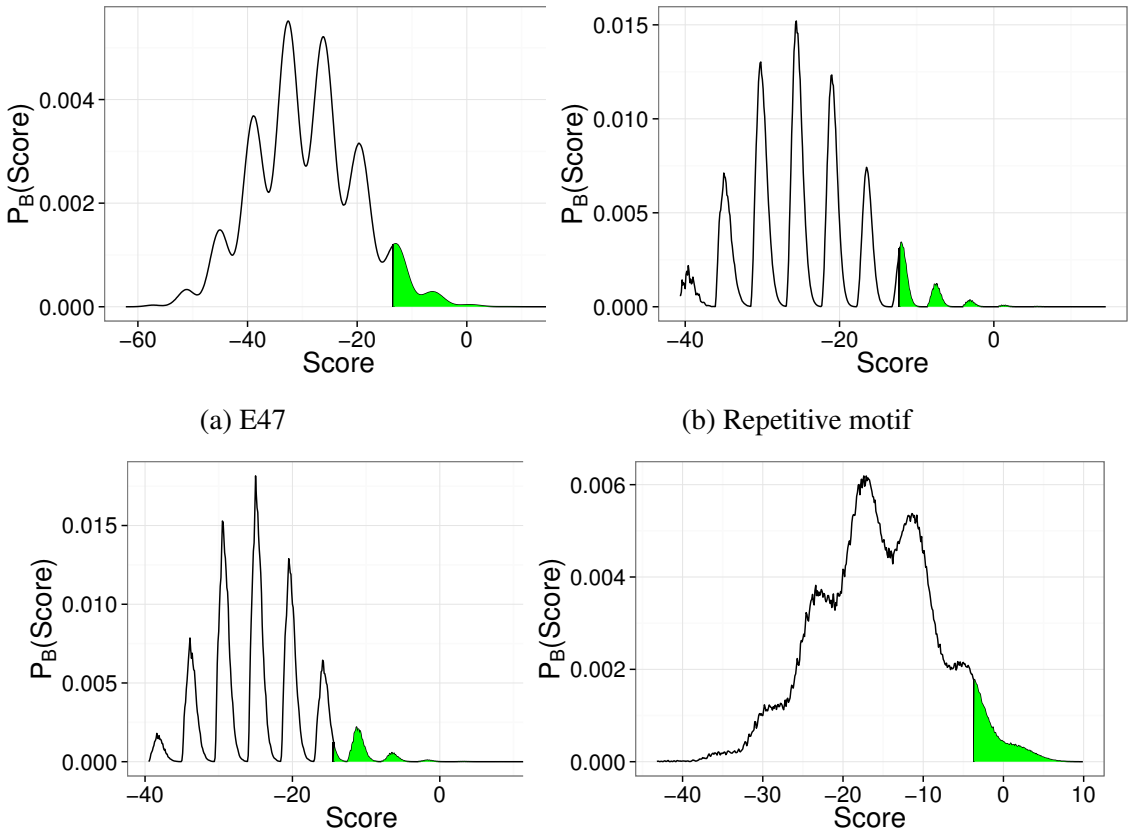
(c) E47 motif from Transfac

(d) SP1SP3 motif from Transfac

Figure 2.5: DNA motifs

2.5 Discussion

Previously, the score has been suggested to call putative motif hits (e.g. TFBSs) in the DNA based on a predefined score threshold t_s [47]. In order to be able to evaluate how many false positive predictions α are brought about by choosing a certain t_s, the score distribution needs to be determined. In this Chapter, we have reviewed that by



(a) E47

(b) Repetitive motif

(c) Palindrome

(d) SP1SP3

Figure 2.6: Score distributions for several DNA motif structures with a score granularity of $\Delta s = 0.1$ (see Figure 2.5). The greenly shaded area indicates to the 5% significance level for calling a motif hit.

discretizing the score range into a fixed number of bins, the score distribution can be determined efficiently using dynamic programming based on order-d Markov models regardless of the length of the TF motif M. In other words, the discretized score range allows us to efficiently deal with the exponentially increasing set of compatible words $C(t_s)$ with increasing M [59].

Similar as in the order-0 case, where the convolution operation is recursively employed to derive the score distribution [40], in the order-d background model case we take

advantage of the conditional independence property to employ convolution recursively. While the RSAT suite [49] previously reported to determine the score distribution based on order-d background models, to our knowledge, the precise algorithmic details of the RSAT implementation were never published. Because of this and because of its central role for the remainder of this thesis, we provided a detailed description of the algorithm.

Finally, we exemplified the score distributions for several motifs in order to illustrate their dependence on the TF motif and the background model (see Figure 2.6). From those distributions, it is also apparent that for a fixed significance level α, t_s changes across the motifs.

Chapter 3

Statistical dependence between motif hits

In the previous Chapter we have discussed the score distribution and how it can be employed in order to derive a score threshold t_s for a desired significance level α. The score threshold is in turn used to identify putative TFBSs in a given DNA sequence in the process of *motif hit identification*. In this process, we are solely interested in localizing motif hits, irrespective of any potential relationship between two or more motif hits.

When it comes to *motif hit enrichment*, however, the statistical relationship between motif hits becomes an issue. The reason for that is that the distribution of the number of motif hits $P(X|H_0)$, not only is determined by the false positive rate of producing motif hits α, but also by the correlation between the events in the process $\mathbf{Y}_{[1:N-M+1]}$, Therefore, we discuss the relationship between motif hits that are produced by the process $\mathbf{Y}_{[1:N-M+1]}$ in this Chapter.

We shall start by exploring the information imbalance between observing a motif hit and a non-hit. As a result, we obtain an important statistical independence assumption that allows us to factorize $P(\mathbf{Y}_{[1:N-M+1]})$ for a give realization.

Moreover, we analyze the statistical dependence between motif hits, which can be grouped into *non-overlapping motif hits* and *overlapping motif hits*. While, non-overlapping motif hits may be linked through the background model, which may induce long-range interactions between distantly located hits, overlapping motif hits depend not only on the background model, but also on the structure of the motif. In the latter situation, motif hits occurring in $\mathbf{Y}_{[1:N-M+1]}$ might be strongly coupled even if the underlying DNA sequence was drawn from an order-0 background model (e.g. for repetitive motifs) [56].

We give reasons why the coupling between non-overlapping motif hits can be neglected in many practical situations, whereas overlapping motif hits need to be accounted for, which leads us to study overlapping motif hits in more detail.

We discuss the qualitative classes of overlapping motif hits that might occur when scanning a single-stranded sequences (e.g. RNA sequences) as well as when both strands of the DNA are scanned for motif hits. Moreover, we discuss overlapping motif hit probabilities and how one computes them. Therefore, we discuss two different types overlapping hit probabilities: On the one hand, we adopted the approach proposed by Pape *et al.* [36], in order to compute the overlapping hit probability via the score distribution, which give rise to the *marginal overlapping hit probabilities*. On the other hand, we adopt the concept of *periodicity* to overlapping motif hits which allows us to discuss motif hits at principal periods. We refer to the latter overlapping motif hits as *principal overlapping hit* for which we developed a novel algorithm to approximate the *principal overlapping hit probabilities*.

The resulting algorithms of this Chapter are critical for further developments that we discuss in Chapters 4, 5 and 6.

3.1 Information imbalance between hits and non-hits

As we have discussed in Section 1.2.4, the latent DNA sequence induces complicated statistical dependences between the events in $\mathbf{Y}_{[1:N-M+1]}$. Exact computational analysis of $P(\mathbf{Y}_{[1:N-M+1]})$ is unfortunately prohibitively expensive, however, in this section, we discuss an important (approximate) independence property that holds for the events in $\mathbf{Y}_{[1:N-M+1]}$, which is associated with observing motif hits.

Typically, when we identify motif hits, the significance level α is set to fairly stringent values so as to avoid false positive predictions of motif hits (e.g. $\alpha = 10^{-2} - 10^{-4}$). As a consequence, the set of compatible words $C(t_s)$ contains far less elements than its complementary set $\bar{C}(t_s) = \mathcal{A}^M \setminus C(t_s)$. In addition, not only is $|C(t_s)| \ll |\bar{C}(t_s)|$, but also do the compatible words share common features, rather than just being a random subset of $\mathcal{A}^M$. That is, they are all more or less well explained by the TF motif. By contrast, $\bar{C}(t_s)$ lacks any common features between its words.

Now, consider the task of inferring the underlying DNA sequence after observing the state of Y by using Definition (1.9) and (1.23) according to Bayes' theorem

$$P(w_1 \cdots w_M|Y) \propto P(Y|w_1 \cdots w_M)P_B(w_1 \cdots w_M).$$

In the case of observing $Y = 1$, due to the common structure of the compatible words and since $|C(t_s)| \ll |\mathcal{A}^M|$, the underlying DNA sequence can be predicted rather accurately

from $P(w_1 \cdots w_M|Y = 1)$, depending on the stringency level of α and the per-position information content IC_i of the motif.

On the other hand, the complementary set $\bar{C}(t_s)$ lacks a common structure between its words. Therefore, $P(w_1 \cdots w_M|Y = 0)$ will result in a rather flat posterior distribution of a large proportion of words of length M, which implies a rather high uncertainty about the latent sequence $w_1 \cdots w_M$.

Recall that if we would observe the underlying DNA sequence $\mathbf{w}$, the outcomes in $\mathbf{Y}_{[i+1:i+M-1]}$ would be conditionally independent of one another, which is illustrated in Figure 1.8b (see also Equation (1.25)). However, in our case, we observe $\mathbf{Y}_{[i+1:i+M-1]}$ and treat $\mathbf{w}$ as latent events. Therefore, the probabilities of of the outcomes $\mathbf{Y}_{[i+1:i+M-1]}$ do not factorize in general if $\mathbf{w}$ is not observed. Nevertheless, since observing a motif hit $Y = 1$ is rather informative about the underlying DNA sequence, we suggest that the following approximate statistical independence assumption still holds approximately

$$P(\mathbf{Y}_{[i+1:i+M-1]}|Y_i = 1, \mathbf{Y}_{[1:i-1]}) = P(\mathbf{Y}_{[i+1:i+M-1]}|Y_i = 1) \tag{3.1}$$

which says that once a motif hit is observed (at position i), the events following Y_i are independent of any events prior to Y_i. Or equivalently, a motif hit $Y_i = 1$ prevents information from passing between Y_j to Y_k with $j < i < k$.

Note that this assumption was implicitly used by Pape *et al.* [36] to derive the extension factors, however, its validity was not explained there.

In the extreme case, where the set of compatible words $C(t_s)$ contains only a single word and the background model order was $d \in \{0, 1\}$, the motif hit $Y = 1$ is deterministically associated with that $\mathbf{w} \in C(t_s)$. As a consequence, observing $Y = 1$ means that at the

same time $\mathbf{w}$ is observed in which case the independence assumption is satisfied exactly. However, for $|C(t_s)| > 1$ or $d > 1$, the Equation (3.1) represent an approximation.

We also want to stress that Assumption (3.1) is an instance of a context-specific independence [28], which refers to a form of independence which depends on the particular assignment of a set of random variables (e.g. $Y = 1$) rather than just observing the variable Y regardless of its assignment (which would be an instance of conditional independence). Adopting the notation from Koller *et al.* [28], context-specific independence is denoted by the $\perp\!\!\!\perp_c$ symbol. Accordingly, we can express Assumption (3.1) as

$$(Y_k \perp\!\!\!\perp_c Y_j | Y_i = 1) \quad \text{for } j < i < k. \tag{3.2}$$

For more details on context-specific independence see Koller *et al.* [28].

When a DNA sequence is scanned for motif hits on both strands, of a given DNA sequence, we additionally have the following context-specific independence assertions

$$(Y_k \perp\!\!\!\perp_c Y_j | Y_i = 1) \quad \text{for } j < i < k \tag{3.3}$$

$$(Y_k \perp\!\!\!\perp_c Y_j | Y'_i = 1) \quad \text{for } j \leq i < k \tag{3.4}$$

$$(Y_k \perp\!\!\!\perp_c Y'_j | Y_i = 1) \quad \text{for } j < i < k \tag{3.5}$$

$$(Y_k \perp\!\!\!\perp_c Y'_j | Y'_i = 1) \quad \text{for } j < i < k \tag{3.6}$$

$$(Y'_k \perp\!\!\!\perp_c Y_j | Y_i = 1) \quad \text{for } j < i \leq k \tag{3.7}$$

$$(Y'_k \perp\!\!\!\perp_c Y_j | Y'_i = 1) \quad \text{for } j \leq i < k \tag{3.8}$$

which essentially expresses that any motif hit, regardless of its strandedness, blocks

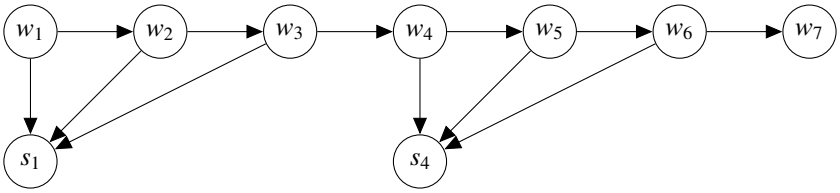
the information from passing between events before and after the hit. Independence assumptions (3.2)-(3.8) play an integral part throughout the rest of the thesis, because they will be exploited to develop efficient algorithms for otherwise computationally infeasible problems, including for the derivation of the probabilities of obtaining overlapping hits at principal periods (see Section 3.3.3) or in order to derive the extension factors for computing the clump size probabilities (see Chapter 4).

3.2 Statistical dependence between non-overlapping motif hits

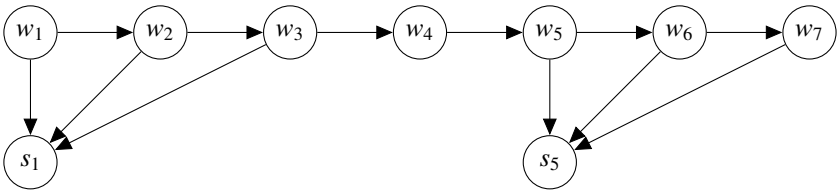
This Section illuminates the statistical dependence between non-overlapping motif hits, $Y_i = 1$ and $Y_{i+k} = 1$ such that $k > M - 1$, where M denotes the length of a TF motif.

First, lets consider drawing a sequence $w_1 \cdots w_N$ from an order-0 background model. Since each nucleotide w_i is drawn independently from any other one, any two non-overlapping words (e.g. $w_i \cdots w_{i+M-1}$ and $w_{i+k} \cdots w_{i+k+M-1}$ with $k > M - 1$) are by definition independent as well. Consequently, observing a motif hit $Y_i = 1$ conveys no information about the presence of an additional non-overlapping motif hit $Y_{i+k} = 1$.

By contrast, consider $w_1 \cdots w_N$ generated by an order-d Markov model with $d > 0$. In this case, although the subsequences $w_i \cdots w_{i+M-1}$ and $w_{i+k} \cdots w_{i+k+M-1}$ with $k > M - 1$ do not share a common subsegment, there is still information passed between w_{i+M-1} and w_{i+k}, either directly or transitively. Direct dependence would be given if the context of the transition probability $\pi(w_{-d} \cdots w_{-1}, w_0)$ spans two non-overlapping words, while transitive passing of information refers to correlations that are indirectly estab-



(a) Direct influence of s_1 on s_4



(b) Transient influence of s_1 on s_5

Figure 3.1: Bayesian network representation of scores s_i that are determined based on a DNA sequence $w_1 \cdots w_7$ which were generated by an order-1 Markov model. In the example, the motif length equals $M = 3$. Due to the latent DNA sequences, information about s_1 is passed on to s_4 and s_5, even though the underlying subsequences that are used to determine s_1 and s_4 (or s_5) do not overlap.

lished through a chain of intermediate random nucleotides (see Figure 3.1). Correlation between the nucleotides implies a correlation between the scores s_i and s_{i+k} and in turn an association between motif hits $Y_i = 1$ and $Y_{i+k} = 1$. Therefore, our next question addresses how strongly the letters w_i and w_{i+k} may depend on each other.

Both, direct and indirect (or transient) dependence between non-overlapping words are determined by the parametrization of the background model, that is $\pi(w_{-d} \cdots w_{-1}, w_0)$. Direct dependence affects only d positions after the current position and can be appreciated from $\pi(w_{-d} \cdots w_{-1}, w_0)$ by comparing the probability of obtaining a_0 across different contexts $w_{-d} \cdots w_{-1}$, while, transient dependence theoretically affects all subsequent positions. To analyze direct dependence, one possibility is to compare the transition probabilities $\pi(w_{-d} \cdots w_{-1}, w_0)$, whereby, weak dependence is given if the probabilities of observing a letter w_0 are similar across the d nucleotide prefixes. On the other hand, the dependence is pronounced if the probabilities are very different across the prefixes.

In the following, we shall concentrate on transient dependence between nucleotides drawn from an order-d Markov model.

Transient coupling between nucleotides can be captured by eigenvalue decomposition of the transition matrix of the Markov model, which can be leveraged to investigate how quickly the influence of a letter w_1 on another letter w_i with $i > 1$ decays as the stochastic process progresses [19].

To this end, let $\mathbf{T}$ be an $|\mathcal{A}^d| \times |\mathcal{A}^d|$ matrix, termed the transition matrix, contains the transition probabilities of the order-d Markov model such that having observed the word $\mathbf{w} \in \mathcal{A}^d$, the word $\mathbf{w}' \in \mathcal{A}^d$ is observed in the next step. Since the transition matrix contains only real elements and assuming that its eigenvectors can be chosen to be a basis for the whole space, we can subject the transition matrix to eigenvalue decomposition such that

$$\mathbf{T} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1} \tag{3.9}$$

where $\mathbf{Q}$ denotes a matrix consisting of the eigenvectors and $\mathbf{\Lambda}$ denotes a diagonal matrix consisting of the eigenvalues on the diagonal which need not be distinct [26]. For convenience the eigenvalues shall be ordered descendingly such that $\lambda_1 \geq \lambda_2 \geq \cdots$. Since $\mathbf{T}$ is a stochastic matrix, at least one eigenvalue is equal to one and all eigenvalues satisfy $|\lambda_i| \leq 1$ [26]. Furthermore, due to the ergodicity assumption for $\mathbf{T}$, it is guaranteed that the largest eigenvalue is unique. That is, all remaining eigenvalues are absolutely smaller $|\lambda_i| < \lambda_1 = 1$ for $i > 1$. Moreover, the associated eigenvector for the largest eigenvalue (λ_1) is equal to the stationary distribution of the Markov model [26].

Next, consider running the Markov chain according to the background model starting

from an arbitrary initial distribution $\mathbf{p}_1$ of all words of length d (e.g. $p_1(\text{"}AAA\text{"}) = 1$ where $d = 3$). As discussed in Karlin *et al.* [26] for general Markov chains, we can express the distribution of the words of length d after running the process for i steps according to

$$\mathbf{p}_i = \mathbf{T} \cdot \mathbf{p}_{i-1}. \tag{3.10}$$

By unrolling the recursion stated in Equation (3.10) and, subsequently, substituting Equation (3.9) we obtain

$$\mathbf{p}_i = \mathbf{T}^{i-1} \cdot \mathbf{p}_1 \tag{3.11}$$

$$= \left(\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}\right)^{i-1} \cdot \mathbf{q}_1. \tag{3.12}$$

Due to the fact that $\mathbf{Q}^{-1}\mathbf{Q} = \mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix, Equation (3.12) becomes

$$\mathbf{p}_i = \mathbf{Q}\mathbf{\Lambda}^{i-1}\mathbf{Q}^{-1} \cdot \mathbf{p}_1 \tag{3.13}$$

Since $1 = \lambda_1 > |\lambda_2| \geq \cdots$, Equation (3.13) highlights that the eigenvalues of all but the largest value decrease exponentially. Therefore, eventually, as $i \to \infty$, $\mathbf{p}_i$ approaches the stationary distribution denoted by $\boldsymbol{\mu}$.

Approaching the stationary distribution $\boldsymbol{\mu}$ has an important implication: The influence of w_1 on w_i decays exponentially with increasing i. Or alternatively, w_1 and w_i becoming more and more statistically independent from one another and, in the limit, w_1 does not convey any information about w_∞ at all.

However, $\mathbf{p}_i$ might still be considerably different from $\boldsymbol{\mu}$ for small i (e.g. $i = 5$), in which case the nucleotides might considerably depend one another. Therefore, we seek to quantify how quickly $\mathbf{p}_i$ approaches $\boldsymbol{\mu}$ as we run the Markov chain.

One way to estimate how quickly the knowledge about a state w_1 dissipates through running the random process is by exploring the magnitude of the second largest eigenvalue $|\lambda_2|$ [26]. The smaller $|\lambda_2|$, the faster the Markov chain will approach the stationary distribution. In contrast, if $|\lambda_2| \approx 1$, the chain would maintain long-range interactions between fairly distantly located nucleotides, which might cause distantly located motif hits to be correlated as well.

In the following Section, we provide a case-study of the association between distantly located nucleotides w_1 and w_i for increasing i using an order-1 background model trained on human accessible chromatin sites.

3.2.1 An order-1 background model on human accessible chromatin regions

In this Section, we investigate the transition matrix of an order-1 Markov model that was estimated on a subset of human DNase-I hypersensitive sites [50]. To this end, we determined the order-1 Markov model using the approach that was discussed in

Section 1.2.2. The resulting transition matrix is given by

$$\mathbf{T} = \begin{bmatrix} \pi(A,A) & \pi(C,A) & \pi(G,A) & \pi(T,A) \\ \pi(A,C) & \pi(C,C) & \pi(G,C) & \pi(T,C) \\ \pi(A,G) & \pi(C,G) & \pi(G,G) & \pi(T,G) \\ \pi(A,T) & \pi(C,T) & \pi(G,T) & \pi(T,T) \end{bmatrix} = \begin{bmatrix} 0.280 & 0.263 & 0.299 & 0.187 \\ 0.249 & 0.292 & 0.146 & 0.284 \\ 0.284 & 0.146 & 0.292 & 0.249 \\ 0.187 & 0.299 & 0.263 & 0.280 \end{bmatrix} \tag{3.14}$$

Comparing the entries of the transition matrix across each row of the matrix can be used to judge direct dependence. As the entries differ by less than 15%, the dependence is rather mild. In comparison, the most extreme scenarios would be either 1) statistical independence (that is all rows are identical) and/or deterministic dependence (e.g. the identity matrix).

Next, we address the question as to how strongly non-overlapping nucleotides are transiently associated with one another for the given example. To this end, we study how fast the influence of a nucleotide w_1 decays on another nucleotide w_i as i is increased.

We employ eigenvalue decomposition of $\mathbf{T}$ which yields

$$[\lambda_1, \lambda_2, \lambda_3, \lambda_4] = [1, 0.164, -0.095, 0.075].$$

According to the discussion in the last Section, $|\lambda_2|$ allows us to assess how quickly the information degrades as i increases, which equals 0.164^{i-1}. Therefore, after running the Markov chain for only 10 steps, the influence drops substantially ($0.164^{10} = 1.39e{-}8$), rendering nucleotides even in relative close vicinity for all practical purposes independent.

In summary, we argue that dependencies of non-overlapping words are usually negligi-

ble in most practical situations (e.g. for analyzing regulatory regions) when background orders are relatively small (e.g. $d < 4$).

3.3 Statistical dependence between overlapping motif hits

While, non-overlapping motif hits may be statistically coupled by the background model, overlapping motif hits are coupled through both, the TF motif and the background. For that reason, even if the nucleotide sequence $w_1 \cdots w_N$ was drawn from an order-0 model, observing a motif hits might be indicative of obtaining an additional (overlapping) motif hit depending on the TF motif [56].

In this Section, we shall start by discussing the qualitative aspects of overlapping motif hits, especially with respect to analyzing both the forward and the reverse strand of a DNA sequence. Subsequently, we turn to quantifying overlapping motif hits. We discuss the *marginal overlapping hit probabilities* as the probabilities of obtaining overlapping hits, which can be derived based on the score distribution, as proposed by Pape *et al.* [36]. While previously an order-0 background was assumed to derive the score distribution, we describe an improved algorithm that is applicable for general order-d background models.

Finally, we discuss the concept of *periodicity*, which was originally used in the word pattern community [41]. We propose a novel adaptation of the *periodicity*-concept to TF motif hits based on the log-likelihood ratio. In particular, we establish the novel *principal overlapping hit probabilities* by extending the notion of *principal periods* to TF motif based hits.

3.3.1 The notion of overlapping motif hits

Scanning for TF motif hits in a DNA sequence $w_1 \cdots w_N$ involves the computation of s_i at each position throughout the sequence and subsequently, deciding on motif hits based on the score threshold t_s (see Section 1.2.3). For a motif of length M, we evaluate the scores at the positions i and $i + k$. If $0 < k < M$, the score values s_i and s_{i+k} depend on $M - k$ shared nucleotides, which induces a statistical dependence between the scores (see Figure 3.2). Likewise, the outcomes of the *motif score* tests Y_i and Y_{i+k} depend on one another too. In the following we shall focus on a special kind of relationship, namely that obtaining a motif hit $Y_i = 1$ informs us about an additional overlapping hit $Y_{i+k} = 1$ (see Section 1.2.4). Such a scenario causes clumping of motif hits which affects *motif hit enrichment* analysis, as we shall see in later chapters.

In general, overlapping motif hits emerge in a sequence if $Y_i = 1$ and $Y_{i+k} = 1$ such that $0 < k < M$. When it comes to scanning double-stranded DNA sequences (that is scanning both strand for the occurrence of motif hits), however, overlapping motif hits may not only arise from overlapping hits on one strand of the DNA, but also due to base pair complementarity between hits on complementary strands. That is, a motif hit on the forward strand $Y_i = 1$ might overlap with a motif hit on the reverse strand $Y'_j = 1$ and vice versa.

Assuming that we encounter motif hits by successively traversing double-stranded DNA sequences from 5' to 3' with respect to the forward strand and that we encounter forward strand events prior to reverse strand events at the same position, there are four scenarios of obtaining overlapping hits: (1) $Y_i = 1$ and $Y_{i+k} = 1$, (2) $Y'_i = 1$ and $Y'_{i+k} = 1$, (3) $Y_i = 1$ and $Y'_{i+k} = 1$ and (4) $Y'_i = 1$ and $Y_{i+k} = 1$, such that $0 < k < M$. Note that the first

and second case are identical due to symmetry and can therefore be reduce to a single case, namely overlapping hits on the same strand. Moreover, regarding the overlap due to base pair complementarity, we refer to the cases (3) and (4) as to 3'-end and 5'-end overlapping hits, respectively, as they result in overlaps of the 3' and 5'-ends of the respective compatible words (see Figure 3.3b and 3.3c). 3' and 5'-end overlapping hits are in general not symmetric (see examples below). Therefore, in total there may be three types of overlapping hits (see Figure 3.3). In the remainder of this Section, we illustrate the three types of overlapping hits.

To illustrate an overlapping motif hit on the same strand, consider scanning for a motif whose set of compatible words is given by $C(t_s) = \{'AAAAA'\}$: Whenever a motif hit is encountered in the sequence, there is a high probability that another motif hit might arise at the next position, because 4 out of 5 'A's are already in common and only one extra 'A' is required. Therefore, a DNA sequence 'AAAAAA' would produce two mutually overlapping motif hits, which is also an instance of a 2-clump.

When it comes to overlapping hits due to base pair complementarity, 3' and 5'-end overlapping hits may be admissible exclusively or simultaneously. For example, for the compatible words $C(t_s) = \{'TCG'\}$ and $C(t_s) = \{'CGT'\}$, 3'-end and 5'-end complementarity hold exclusively, respectively. As a consequence, 3' and 5'-end overlapping hits are not symmetric. To see this, consider the DNA sequence 'TCGA' in which the word 'TCG' leads an 3'-end overlapping matches, as can be seen in the following example

$$\begin{array}{cccccc} & & G & C & T & \leftarrow \\ 5' & \rightarrow T & C & G & A & \rightarrow & 3' \end{array}$$

By contrast, a 5'-end overlapping match is not possible for 'TCG'. In comparison, the

sequence 'ACGT' would emit 5'-end overlapping hits of 'CGT', as can be seen in

$$\begin{array}{cccccccc} & \leftarrow & T & G & C & & \leftarrow & \\ 5' & \rightarrow & A & C & G & T & \rightarrow & 3'. \end{array}$$

Likewise, for 'CGT', a 3'-end overlapping hit is not possible. In both examples, base pair complementarity is restricted to only one end of the underlying words (e.g. 'TCG' and 'CGT'). Therefore, at most two overlapping motif hits are possible.

A compatible word can also be complementary to itself simultaneously at both ends. In this case, more than two overlapping motif hits are feasible. An example of such a case is given by the compatible word $C(t_s) = \{'CGC'\}$ which occurs in the DNA sequence 'CGCGCGC' at position 1, 3 and 5 on the forward strand well as at position 2 and 4 on the reverse strand

$$\begin{array}{cccccccc} & & C & G & C & G & C & \leftarrow \\ \rightarrow & C & G & C & G & C & G & \end{array}.$$

By concatenating arbitrarily many 'CG' dinucleotides to the sequence shown above, the sequence of overlapping motif hits can be extended infinitely many times.

Lastly, we discuss a special case for an overlapping hit, namely the palindromic motif hit. In this case, the motif hits $Y_i = 1$ and $Y_i' = 1$ occur simultaneously at the same position, but on opposite strands. In case of palindromic overlap, a compatible word not only partially, but fully complements itself. An example of a palindromic word is given

by

$$\begin{array}{cccccc} & T & G & C & A & \leftarrow \\ \rightarrow & A & C & G & T & \end{array} .$$

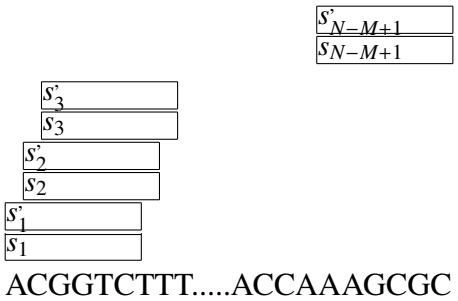


Figure 3.2: Evaluating the scores on either strand in a sliding windows fashion leads to overlapping windows and that induce a statistical coupling between neighboring scores. The coupling between overlapping scores might subsequently induced overlapping motif hits. The scores at each position are indicated by the position in the subscript and the strandedness, where the prime denotes scores on the reverse strand and the absence of the prime denotes scores obtained on the forward strand.

3.3.2 Marginal overlapping hit probabilities

In Section 3.3.1, we have identified three qualitative types of overlapping motif hits that might arise when both strands of the DNA are scanned for motif hits. We continue by quantitatively studying how probable each of the three overlapping hit scenarios might occur for a particular TF motif, a background model and a score threshold t_s.

Starting from the random process $\mathbf{Y}_{[1:N-M+1]}$ that was described in Section 1.2.4, we introduce the joint probabilities of all pairs of mutually overlapping events in $\mathbf{Y}_{[1:N-M-1]}$

as

$$P(Y_{i+k}, Y_i) = \sum_{\mathbf{y} \in \mathbf{Y}_{[1:N-M-1]} \setminus \{Y_{i+k}, Y_i\}} P(Y_{i+k}, Y_i, \mathbf{y}) \qquad 1 \leq k < M \tag{3.15}$$

$$P(Y'_{i+k}, Y_i) = \sum_{\mathbf{y} \in \mathbf{Y}_{[1:N-M-1]} \setminus \{Y'_{i+k}, Y_i\}} P(Y'_{i+k}, Y_i, \mathbf{y}) \qquad 0 \leq k < M \tag{3.16}$$

$$P(Y_{i+k}, Y'_i) = \sum_{\mathbf{y} \in \mathbf{Y}_{[1:N-M-1]} \setminus \{Y_{i+k}, Y'_i\}} P(Y_{i+k}, Y'_i, \mathbf{y}) \qquad 1 \leq k < M. \tag{3.17}$$

which are derived by summing out all random events in $\mathbf{Y}_{[1:N-M-1]}$ except for the two under consideration (see Figure 3.3). The shift between the events in Equation (3.15)-(3.17) is denoted by k. By conditioning on the preceding event and because the background model is homogeneous, we further yield the *marginal overlapping hit probabilities* which are defined by

$$\gamma_k := \frac{P(Y_k = 1, Y_0 = 1)}{P(Y_0 = 1)} = P(Y_k = 1 | Y_0 = 1) = P(Y'_k = 1 | Y'_0 = 1) \qquad 1 \leq k < M \tag{3.18}$$

$$\gamma_{3',k} := \frac{P(Y'_k = 1, Y_0 = 1)}{P(Y_0 = 1)} = P(Y'_k = 1 | Y_0 = 1) \qquad 0 \leq k < M \tag{3.19}$$

$$\gamma_{5',k} := \frac{P(Y_k = 1, Y'_0 = 1)}{P(Y'_0 = 1)} = P(Y_k = 1 | Y'_0 = 1) \qquad 1 \leq k < M \tag{3.20}$$

where γ_k denotes the overlapping hit probability on the same strand, $\gamma_{3',k}$ denotes a forward strand hit which is followed by a partially or fully overlapping reverse strand hit (the 3'-end overlapping hit) and $\gamma_{5',k}$ denotes a reverse strand hit which is followed by a partially overlapping forward strand hit (the 5'-end overlapping hit). In each case, the

overlapping is shifted by k positions from a previous hit at position zero. The attribute "marginal" refers to the fact that the indicator random variables in between position 0 and k have been summed out in this representation. That is

$$\gamma_k = \sum_{\substack{y_1 \cdots y_{k-1} \in \{0,1\}^{k-1} \\ y'_0 \cdots y'_{k-1} \in \{0,1\}^k}} P(Y_k = 1, y_1 \cdots y_{k-1}, y'_0 \cdots y'_{k-1} | Y_0 = 1) \qquad 1 \leq k < M$$

$$\gamma_{3',k} = \sum_{\substack{y_1 \cdots y_k \in \{0,1\}^k \\ y'_0 \cdots y'_{k-1} \in \{0,1\}^k}} P(Y_k = 1, y_1 \cdots y_k, y'_0 \cdots y'_{k-1} | Y_0 = 1) \qquad 0 \leq k < M$$

$$\gamma_{5',k} = \sum_{\substack{y_1 \cdots y_k \in \{0,1\}^k \\ y'_0 \cdots y'_{k-1} \in \{0,1\}^k}} P(Y_k = 1, y_1 \cdots y_k, y'_0 \cdots y'_{k-1} | Y_0 = 1) \qquad 1 \leq k < M$$

By definition, the *marginal overlapping hit probabilities* quantify how probable a motif hit is overlapped by another hit without considering intermediate events. In Chapter 4, 5 and 6, they shall be the basis for deriving the necessary statistics for testing *motif hit enrichment* and are therefore fundamental for this thesis. In the following, we discuss how the *marginal overlapping hit probabilities* can be derived.

The two-dimensional score distribution for an order-d background model

Pape *et al.* [36] have proposed previously that the overlapping hit probability can be determined based on a two-dimensional score distribution $P(S, S'|H_0)$, where S and S' are associated with the scores at mutually overlapping positions. To this end, the distribution of S and S' was obtained simultaneously using an extension of the algorithm stated by Rahmann *et al.* [40]. In their work, the algorithm computes the two-dimensional score distribution based on a GC-background model (an order-0 Markov model with

matched AT and CG frequencies) by leveraging a two-dimensional convolution operation.

As part of this thesis, we sought to improve upon the previous algorithm by computing $P(S, S'|H_0)$ based on a general order-d background model. This algorithm may also be viewed as an extension of the score distribution algorithm for order-d background models that we have discussed earlier (see Chapter 2). As a result, the algorithm allows us to determine the following two-dimensional score distributions

$$P(S_k, S_0|H_0) \quad 1 \leq k < M \tag{3.21}$$

$$P(S'_k, S_0|H_0) \quad 0 \leq k < M \tag{3.22}$$

$$P(S_k, S'_0|H_0) \quad 1 \leq k < M \tag{3.23}$$

from which, after applying the score threshold t_s, we directly obtain the *marginal overlapping hit probabilities*.

In the following discussion, the two score dimensions spanned by S and S' depend on which distribution is being processed ((3.21), (3.22) or (3.23)). Therefore, generic variables S and S' are interpreted as forward strand or reverse strand scores at respective shifts depending on the concrete overlap configuration.

For the same reason as discussed in Section 2.1, we start the discussion of the algorithm by the discretization of the score plane spanned by S and S'. Accordingly, the scores are split into

$$G := \frac{\max_{\mathbf{a}\in\mathcal{A}^M}(s(\mathbf{a})) - \min_{\mathbf{a}\in\mathcal{A}^M}(s(\mathbf{a}))}{\Delta s} \tag{3.24}$$

$$G' := \frac{\max_{\mathbf{a}\in\mathcal{A}^M}(s'(\mathbf{a})) - \min_{\mathbf{a}\in\mathcal{A}^M}(s'(\mathbf{a}))}{\Delta s} \tag{3.25}$$

bins, respectively, where we use the same score granularity Δs for both dimensions. Both, the maximal and minimal score can be obtained in $O(|\mathcal{A}|^{d+1}M)$, respectively. Accordingly, a pair of integers (g, g') such that $0 \leq g < G, 0 \leq g' < G'$ represents the real-valued score interval

$$(s, s') \in \mathcal{R}^2 : (g\Delta s \leq s < (g+1)\Delta s) \text{ and } (g'\Delta s \leq s' < (g'+1)\Delta s) \tag{3.26}$$

assuming that we have subtracted $(\min_{\mathbf{a}\in\mathcal{A}^M}(s(\mathbf{a})), \min_{\mathbf{a}\in\mathcal{A}^M}(s'(\mathbf{a})))$ from (s, s'). As mentioned in the previous chapter, theoretically, the score granularity Δs can be chosen such that the score distribution can be determined exactly. However, choosing Δs too small leads to substantial runtime and memory requirements. As a tradeoff, we shall thus use a course-grained score granularity (e.g. $\Delta s = 0.1$) such that an accurate approximate score distribution can be computed quickly.

Next, consider a particular sequence $a_1 \cdots a_{M+k}$ on which the pair of scores s and s' is determined with a shift of k nucleotides. The *motif scores* are given by

$$s(a_1 \cdots, a_{M+k}) = \log\left(\frac{P_M(a_1 \cdots a_{M+k})}{P_B(a_1 \cdots a_{M+k})}\right) \approx l_d(a_1 \cdots a_d) + \sum_{i=d+1}^{M} l_i(a_{i-d} \cdots a_i) \tag{3.27}$$

$$s'(a_1 \cdots, a_{M+k}) = \log\left(\frac{P_M(a_{1+k} \cdots a_{M+k})}{P_B(a_{1+k} \cdots a_{M+k})}\right) \approx l'_d(a_{1+k} \cdots a_{d+k}) + \sum_{i=k+d+1}^{M} l'_i(a_{i-d} \cdots a_i) \tag{3.28}$$

which make use of the respective *local score contributions*

$$l_j(w_{j-d}\cdots w_j) := \begin{cases} 0 & \text{if } j < d \\ \lfloor \log\left(\frac{\prod_{i=1}^{d} m_{w_i i}}{\mu(w_1\cdots w_d)}\right)\rfloor & \text{if } j = d \\ \lfloor \log\left(\frac{m_{w_j j}}{\pi(w_{j-d}\cdots w_{j-1}, w_j)}\right)\rfloor & \text{if } d < j \leq M \\ 0 & \text{if } j > M \end{cases} \tag{3.29}$$

$$l'_j(w_{j-d}\cdots w_j) := \begin{cases} 0 & \text{if } j < d \\ \lfloor \log\left(\frac{\prod_{i=1}^{d} m'_{w_i i}}{\mu(w_1\cdots w_d)}\right)\rfloor & \text{if } j = d \\ \lfloor \log\left(\frac{m'_{w_j j}}{\pi(w_{j-d}\cdots w_{j-1}, w_j)}\right)\rfloor & \text{if } d < j \leq M \\ 0 & \text{if } j > M \end{cases} \tag{3.30}$$

where the floor operator denotes the discretization of the score values. As described in Section 2.1, the discretization will induce an error in most practical situations. Even though it is theoretically possible to choose the score granularity Δs such that the score distribution is exact, we shall use a course-grain score granularity Δs (e.g. $\Delta s = 0.1$). Since, s and s' depend on the strandedness, need to employ m_{wi} and m'_{wi} according to the concrete strandedness in Definition (3.29) and (3.30). While, for the forward strand score, the original TF motif is used, for the reverse strand score, the reverse complemented TF motif must be used.

Next, we take advantage of a probabilistic representation of the local score contributions

according to

$$P(L_i = l|w_{i-d}\cdots w_i) := \begin{cases} 1 & \text{if } l = l_i(w_{i-d}\cdots w_i) \\ 0 & ow \end{cases} \quad \text{for } d \leq i \leq M \tag{3.31}$$

$$P(L'_i = l|w_{i-d}\cdots w_i) := \begin{cases} 1 & \text{if } l = l'_i(w_{i-d}\cdots w_i) \\ 0 & ow \end{cases} \quad \text{for } d \leq i \leq M. \tag{3.32}$$

to highlight the local score contribution depends on the underlying DNA sequence which is subsequently induced by the background model (see Figure 3.4).

Along a similar argument as presented in Section 2, the Bayesian network representation highlights the conditional independence relationship between L_i and L_{i+k} as well as L'_i and L'_{i+k} with $k > 0$ given the underlying DNA sequence. Consequently, conditional independence can be exploited in order to compute the distribution of the sum of random variables by convolving the distributions of the respective summands.

Computation of the two-dimensional score distribution for an order-d background model

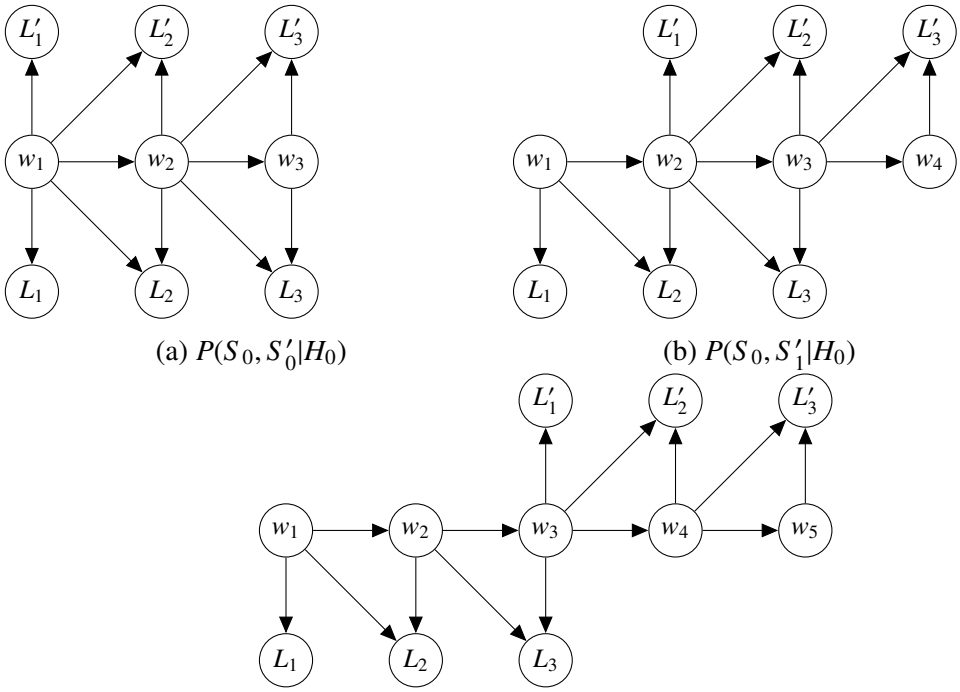
We shall now state the recursive algorithm for determining the two-dimensional score distribution. Note that in case an order-0 Markov model is assumed, the following algorithm corresponds exactly to the algorithm that was previously proposed by Pape *et al.* [36]. However, in the variant that we present in this thesis, the algorithm was generalized to an order-d Markov model as background. The algorithm that we discuss in this section can also be viewed as a generalization of the score distribution algorithm presented in Chapter 2, where instead of only considering a one-dimensional score dis-

(a) Overlapping hits on the same strands

(b) 3' overlapping hits on opposite strands

(c) 5' overlapping hits on opposite strands

Figure 3.3: Types of overlapping hit with a shift of k in double stranded DNA. The arrows pointing to the right and left represent forward (5'→3') and reverse (3'←5') strand hits, respectively. Forward strand hits are called on the first position while reverse strand hits are called on the last position.



(a) $P(S_0, S'_0|H_0)$

(b) $P(S_0, S'_1|H_0)$

(c) $P(S_0, S'_2|H_0)$

Figure 3.4: Bayesian network representation of a nucleotide sequence $w_1 \cdots w_5$ that is generated from an order-1 Markov model and which induces the local score contributions $L_1 \cdots L_3$ and $L'_1 \cdots L'_3$ for a motif of length $M = 3$. Panel a, b and c represent the shifts between the motif starts of $k = 0, 1$ and 2, respectively.

tribution, we next study the distribution of scores at two positions and/or complementary strands simultanously.

Using Definitions (3.31)-(3.32) as well as the transition probability and the stationary distribution from the background model (Equation (1.9)), the algorithm is initialized such that the two-dimensional score distribution Q_d at position d where $d \leq M$ is equal to

$$Q_d(S, S', \mathbf{a}) = P(L_d|\mathbf{a})P(L'_{d-k}|\mathbf{a})\mu(\mathbf{a}) \quad \forall \mathbf{a} \in \mathcal{A}^d. \tag{3.33}$$

That is, for each word **a**, only a single entry of the score plane is non-zero, while the rest of the plane contains zeros.

We continue by iteratively computing the distribution Q_i starting from $i = d + 1$ until $i = M + k$ from the previously determined distribution Q_{i-1}. To this end, we first apply the convolution operation over the discretized score plane s and s' between the previous score distribution Q_{i-1} and the next local score contribution $P(L_i|a_1 \cdots a_{d+1}) \times P(L_i|a_1 \cdots a_{d+1}) \times \pi(a_1 \cdots a_d, a_{d+1})$ for each word $a_1 a_2 \cdots a_{d+1} \in \mathcal{A}^{d+1}$. This operation is justified, because Q_{i-1} is conditionally independent from $P(L_i|a_1 \cdots a_{d+1}) \times P(L_i|a_1 \cdots a_{d+1}) \times \pi(a_1 \cdots a_d, a_{d+1})$ for a given word $a_1 a_2 \cdots a_{d+1}$ (see Figure 3.4). Then, we sum over all nucleotides a_0 to establish the score distribution Q_i, because a_0 does not directly interact with a_i such that $i > d$, due to the Markov independence assumption. This leads us to the following recursive equation

$$\begin{aligned} Q_i(s, s', a_2 \cdots a_{d+1}) = \sum_{a_1 \in \mathcal{A}} \sum_{x} \sum_{x'} & Q_{i-1}(s - x, s' - x', a_1 \cdots a_d) \times \\ & P(L_i = x|a_1 \cdots a_{d+1}) \times P(L'_{i-k} = x'|a_1 \cdots a_{d+1}) \times \end{aligned}$$

$$\pi(a_1 \cdots a_d, a_{d+1}) \tag{3.34}$$

The final score distribution is obtained by summing over the sequence context $\mathbf{a} \in \mathcal{A}^d$ for Q_{M+k}

$$P(S, S'|H_0) = \sum_{\mathbf{a}\in\mathcal{A}^d} Q_{M+k}(S, S', \mathbf{a}). \tag{3.35}$$

From the two-dimensional score distribution we proceed by determining the *marginal overlapping hit probabilities* defined by (3.18)-(3.20) according to

$$P(S' \geq t_s|S \geq t_s, H_0) = \frac{P(S \geq t_s, S' \geq t_s|H_0)}{P_B(S \geq t_s|H_0)} \tag{3.36}$$

where we employ the predefined score threshold t_s.

Runtime

The asymptotic runtime of the algorithm for computing the two-dimensional score distribution is given by $O(GG'|\mathcal{A}|^{d+1}(M+k))$ where G and G' denote the numbers of discrete score bins, $|\mathcal{A}|$ denotes the size of the alphabet, d denotes the order of the background model, M the length of the motif and k the shift between the motif hits.

Although this algorithm runs in polynomial time, it might still use considerable computational resources. The major determinants for the runtime and memory usage are the choices of d and Δs. Increasing d and decreasing Δs yield increasingly accurate results, however, they imply also an increase in the runtime.

Apart from the asymptotic runtime of the algorithm, there are several ways by which

the runtime can be significantly reduced in practice. The most important optimization strategy, which was proposed by Pape *et al.* [36], concerns performing the convolution only on the part the score-plane which is informative about obtaining overlapping motif hits. The rationale behind this is that the intermediate distribution $Q_i(s, s', a_1 \cdots a_d)$ after step $i < M+k$ might assign positive probabilities to scores which cannot establish $s \geq t_s$ at the end of the motif (e.g. $i = M+k$), even in the case when the best-scoring remaining subsequence is extended afterwards. Intermediate scores for which this is the case, can be dropped for the convolution, which reduces the effective size of the score plane. On the other hand, there might also be achievable scores according to $Q_i(s, s', a_1 \cdots a_d)$ for which it is guaranteed that they exceed the score threshold at the end of the motif regardless of which sequence appears afterwards. In this case, such sub-scores can be immediately aggregated, which additionally reduces the area of the plane over which the two-dimensional convolution needs to processed.

Another way of optimizing the runtime can be achieved by avoiding redundant computation during the initialization phase of the algorithm, because, for computing e.g. $P(S_0, S_1|H_0)$, $P(S_0, S_2|H_0)$, ..., $P(S_0, S_{M-1}|H_0)$, the initial one-dimensional convolution operations, until the start position of the second overlapping motif, are redundant. Therefore, instead of initializing the algorithm with Equation (3.33) each time from scratch, it is possible to start from the intermediate distribution $Q_i(s, a_1 \cdots a_d)$ which was obtained from one run of the algorithm for the one-dimensional score distribution. A similar optimization, was also suggested by Pape *et al.* [36], which however, addressed the redundant computational effort at the end of the motif.

A third, and to our knowledge novel way of optimizing the runtime of the algorithm can be exploited by using a simple enumeration algorithm (instead of the dynamic pro-

gramming algorithm) for initializing the score distribution $Q_i(S, S', \mathbf{a})$. To this end, for a some small number i, all possible sequences $w_1 \cdots w_i \in \mathcal{A}^i$ are enumerated in order to determine the their subscore up to position i and their likelihood which are used to initialize Q_i. Although it might seem counter-intuitive to use the enumeration method instead of the dynamic programming variant, because of its exponential asymptotic runtime with respect to the sequence length, for short motifs or sub-motifs, the runtime of the enumeration method is usually less than the runtime of its dynamic programming counter-part. The reason for this is that for very short motifs or sub-motifs of length x, the number of words $|\mathcal{A}|^x$ is still rather small, which means that at most $|\mathcal{A}|^x$ distinct scores can be potentially achieved. However, the score plane of size $G \times G'$ at the recursive step x might already be bigger than $|\mathcal{A}|^x$, in which case it would only be sparsely populated with non-zero entries. Consequently, performing a convolution in the early recursive steps amounts to performing many multiplications with zero, which is a waste of computational resources.

As an example, suppose $x = 4$ and $G = G' = 100$ for $Q_x(s, s', \mathbf{a})$. In this case the convolution requires 10000 multiplication operations, while the brute force algorithm can initialize the score-plane in only 256 steps.

Importantly, it is possible to evaluate the sub-motif length x for which the brute force method can be done faster than the dynamic programming variant. Therefore, we can use the brute force initialization up to position x in the motif and use the dynamic programming variant for all positions greater than x.

3.3.3 Periodicity of motif hits

The concept of *periodicity* was established in the word pattern community and was used to describe the possibility of obtaining overlapping word matches in sequential data, which we have reviewed in Section 1.2.5 [56, 41]. In this Section, we shall extend the concept of *periodicity* to motif hits that were identified using the log-likelihood ratio.

In the case of TF motif, a motif hit is observed upon obtaining any word $\mathbf{a} \in C(t_s)$ in a given DNA sequence, where each word might be periodic or aperiodic according to the definition stated in Section 1.2.5. The precise periodicity relationship across the entire set $C(t_s)$ can be studied by enumerating all words in $C(t_s)$ and by determining their individual periodicity properties. However, as the set of compatible words $C(t_s)$ with fixed t_s grows exponentially with increasing motif length M, enumerating all words is too time consuming in general.

In this Section, we concern with the probabilities of obtaining motif hits at their respective periods. To this end, we study the approximate periodicity relationship as the ensemble average probability of the words in $C(t_s)$ to obtain a motif hit at a given period.

We proceed by establishing a link between the *marginal overlapping hit probabilities* (which were discussed in the previous Section) and the probabilities of obtaining overlapping hits at *periods*. Finally, we propose a novel algorithm for establishing the probability of obtaining an overlapping motif hit at so-called *principal periods* for TF motifs.

Relation between periods and marginal overlapping hit probabilities

In this Section, we illustrate that the *marginal overlapping hit probabilities* bear inherently redundant information about overlapping hits and therefore reflect the probabilities of obtaining motif hits at *periods* (rather than *principal periods*).

As a motivating example, consider the set of compatible words to be $C(t_s) = \{'AAA'\}$ and an order-d background model with $d \in \{0, 1\}$. For the given example, the *marginal overlapping hit probabilities* yield $\gamma_1 > 0$ and $\gamma_2 > 0$ (both being strictly positive). However, γ_1 and γ_2 are inherently related as shown below

$$\begin{aligned}
\gamma_2 &= P(Y_2 = 1 | Y_0 = 1) && (3.37)\\
&= \sum_{y \in \{0,1\}} P(Y_2 = 1, Y_1 = y | Y_0 = 1) && (3.38)\\
&= \underbrace{P(Y_2 = 1, Y_1 = 0 | Y_0 = 1)}_{=0} + \underbrace{P(Y_2 = 1, Y_1 = 1 | Y_0 = 1)}_{P(Y_1=1|Y_0=1)^2} && (3.39)\\
&= \gamma_1^2. && (3.40)
\end{aligned}$$

Since, it is not possible to obtain a hit at period 2 without an intermediate hit at period 1, the first term in Equation (3.39) equates to zero. On the other hand, the second term corresponds to obtaining two successive hits with period 1, respectively (see Section 3.1). Therefore, the hit at period 2 occurs with probability $\gamma_2 = \gamma_1^2$. In this example, since, $C(t_s)$ contains exactly one element, Equation (3.40) is exactly satisfied for the background model orders $d \in \{0, 1\}$.

In most situations, however, the cardinality of $|C(t_s)|$ is greater than one, in which case Equation (3.40) does not hold exactly anymore. Nevertheless, if the score threshold was

chosen stringently enough, the independence assertion (3.2) still holds approximately, so that we can use $P(Y_2 = 1|Y_1 = 1, Y_0 = 1) \approx P(Y_2 = 1|Y_1 = 1)$ to yield $\gamma_2 \approx \gamma_1^2$.

Moreover, for background model orders $d > 1$, the relationship between γ_2 and γ_1 stated above does not hold in general, because the required context for $\pi(w_{-d} \cdots w_{-1}, w_0)$ depends on nucleotides upstream of the start of γ_1. For simplicity, we shall nevertheless assume that $P(Y_2 = 1|Y_1 = 1, Y_0 = 1) \approx P(Y_2 = 1|Y_1 = 1)$ even if $d > 1$.

In conclusion, we argue that the *marginal overlapping hit probabilities* can be interpreted as the probabilities of obtaining an overlapping motif hit at periods, but not principal periods, due to the fact that we have implicitly averaged over all combination of obtaining intermediate hits. Those intermediate hits, however, might represent the causes (or principal periods) of the marginal overlapping hits, so that the marginal overlapping hits convey redundant information about overlapping hits.

In the next Section, we attempt to remove the redundancy of overlapping hits by establishing the overlapping hit probabilities at principal periods.

Principal periodicity of motif hits

While, the *marginal overlapping hit probabilities* reflect the properties of obtaining overlapping hits at periods, the concept of principal periods has been lacking in TF motif hit-based literature. In this Section, we introduce the *principal periods* for TF motif hits which explain all overlapping hits in a non-redundant fashion. We shall refer to the probability of obtaining an overlapping hit at principal periods as to the *principal overlapping motif hit probabilities*.

We start by defining the *principal overlapping motif hit probability* when scanning only

a single-strand of a DNA sequence as

$$\beta_k := P(Y_k = 1, \mathbf{Y}_{[1:k-1]} = \mathbf{0} | Y_0 = 1) \qquad 1 \leq k < M. \tag{3.41}$$

where $\mathbf{Y}_{[1:k-1]} = \mathbf{0}$ indicates that all outcomes between 0 and k evaluate to zero (non-hits). An analogous expression for exact word matches was also used in Marschall, 2011 [32]. Note that Definition 3.41, only accounts for overlapping hits due to principal periods. To see this, recall that by definition, all periods are made up by a combination of principal periods or they are themselves principal periods. Since Definition 3.41 excludes the possibility of obtaining intermediate hits (which may explain the overlapping hit) it cannot be a combination of other principal periods. Therefore, it solely expresses the probabilities of obtaining hits at principal periods.

To consider overlapping hits that are obtained by scanning both strands of a double-stranded DNA sequence (see Section 3.3.1), we extend the notion of *principal overlapping motif hit probabilities* as follows

$$\beta_k := P(Y_k = 1, \mathbf{Y}_{[1:k-1]} = \mathbf{0}, Y'_0 = 0 | Y_0 = 1) \qquad 1 \leq k < M \tag{3.42}$$

$$\beta_{3',0} := P(Y'_0 = 1 | Y_0 = 1) \tag{3.43}$$

$$\beta_{3',k} := P(Y'_k = 1, \mathbf{Y}_{[1:k-1]} = \mathbf{0}, Y'_0 = 0, Y_k = 0 | Y_0 = 1) \qquad 1 \leq k < M \tag{3.44}$$

$$\beta_{5',k} := P(Y_k = 1, \mathbf{Y}_{[1:k-1]} = \mathbf{0} | Y'_0 = 1) \qquad 1 \leq k < M. \tag{3.45}$$

The fact that β_k, $\beta_{3',k}$ and $\beta_{5',k}$ quantify overlapping hits non-redundantly, can also be seen from the fact that each term (across different k) represents an event that is mutually exclusive with respect to all the other terms. For example, while $\beta_{3',0}$ includes the

palindromic hit, β_k, $\beta_{3',k}$ and $\beta_{5',k}$ for $1 \leq k < M$ explicitly exclude the palindromic hit. Likewise, this is the case for all other overlapping positions.

Since, the *principal overlapping hit probabilities* are associated with mutually exclusive events, we obtain the probability of an overlapping hit regardless of its position for scanning a single stranded sequence according to

$$\beta = \sum_{i=1}^{M-1} \beta_i \tag{3.46}$$

and, likewise, for scanning both DNA strands as

$$\beta = \sum_{i=1}^{M-1} \beta_i \tag{3.47}$$

$$\beta_{3'} = \sum_{i=0}^{M-1} \beta_{3',i} \tag{3.48}$$

$$\beta_{5'} = \sum_{i=1}^{M-1} \beta_{5',i} \tag{3.49}$$

for another hit on the same strand, another reverse strand hit that follows a forward strand hit and another forward strand hit that follows a reverse strand hit, respectively.

The probability of a stretch of non-hits following a hit

In the previous Section, we have introduced the *principal overlapping hit probabilities*, which inherently define the clump size distribution. In this section, in particular, we describe the probability that a clump is finished due to the absence of any further overlapping hits.

First, we define the probability that until position j after the last motif hit in a clump, no

further overlapping hits occur. Consequently, when a single DNA strand is scanned, the absence of overlapping hits is defined by

$$\delta_j := P(\mathbf{Y}_{[1:j]} = \mathbf{0} | Y_0 = 1) \tag{3.50}$$

and when both DNA strands are scanned, we define

$$\delta_j := P(\mathbf{Y}_{[1:j]} = \mathbf{0}, Y'_0 = 0 | Y_0 = 1) \tag{3.51}$$

$$\delta'_j := P(\mathbf{Y}_{[1:j]} = \mathbf{0} | Y'_0 = 1). \tag{3.52}$$

In the remainder of the thesis, it will be clear from the context whether δ_j refers to the single stranded or double stranded case.

Next, we turn to deriving the probabilities for Definition (3.50)-(3.52) based on the previously introduced *principal overlapping hit probabilities*. As mentioned above, the *principal overlapping hit probabilities* are associated with mutually exclusive events. If a single DNA strand is scanned for motif hits, we can compute the probability that no overlapping hit occurs until position j by subtracting the *principal overlapping hit probabilities* from one as

$$\delta_j = 1 - \sum_{k=1}^{j} \beta_k \tag{3.53}$$

Likewise, if both DNA strands are scanned for motif hits, we determine the probabilities that no further overlapping hits occur until position j as

$$\delta_j = 1 - \sum_{k=1}^{j} \beta_k - \sum_{k=0}^{j} \beta_{3',k} \tag{3.54}$$

$$\delta'_j = 1 - \sum_{k=1}^{j} \beta_k - \sum_{k=1}^{j} \beta_{5',k}. \tag{3.55}$$

Symmetry relationship for scanning both DNA strands

In this Section, we derive an important relationship between overlapping motif hits that occur on the same strand. As a premise recall that since we assume that the underlying DNA sequence is generated from an order-d Markov chain that starts in its stationary distribution, the probability to obtain a motif hit is identical regardless of the position within the DNA sequence and the strandedness

$$P(Y_i = 1) = P(Y'_i = 1) = \alpha \qquad \text{for } 1 \leq i \leq N - M + 1.$$

Morover, overlapping motif hits on the same strand and palindromic motif hits are symmetrical to one another such that the following cases are equivalent

$$\begin{aligned} P(Y_k = 1, \mathbf{Y}_{[1:k-1]} = \mathbf{0}, Y_0 = 1, Y'_0 = 0) &= P(Y'_k = 1, \mathbf{Y}_{[1:k-1]} = \mathbf{0}, Y'_0 = 1, Y_0 = 0) \\ &= P(Y_k = 1, \mathbf{Y}_{[1:k-1]} = \mathbf{0}, Y_0 = 1, Y'_k = 0) \\ &= P(Y'_k = 1, \mathbf{Y}_{[1:k-1]} = \mathbf{0}, Y'_0 = 1, Y_k = 0). \end{aligned} \tag{3.56}$$

An illustration of these four cases is depicted in Figure 3.5.

Therefore, we can derive the equivalence of between β_k and slightly different conditional probabilities as shown in the following

$$\beta_k = P(Y_k = 1, \mathbf{Y}_{[1:k-1]} = \mathbf{0}, Y'_0 = 0 | Y_0 = 1)$$

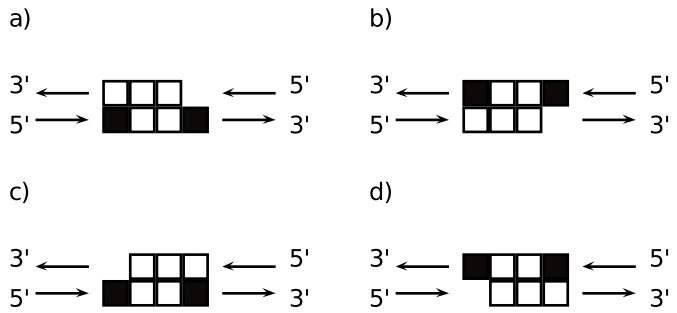


Figure 3.5: Illustration of neighboring motif hits on the same strand (black boxes), with a palindromic non-hit (white box). By symmetry, all of these cases in Panel a)-d) occur with the same probability.

$$
\begin{aligned}
&= \frac{P(Y_k = 1, \mathbf{Y}_{[1:k-1]} = \mathbf{0}, Y'_0 = 0 | Y_0 = 1) P(Y_0 = 1)}{P(Y_0 = 1)} \\
&= \frac{P(Y_k = 1, \mathbf{Y}_{[1:k-1]} = \mathbf{0}, Y'_0 = 0, Y_0 = 1)}{P(Y_0 = 1)} \\
&= \frac{P(Y'_k = 1, \mathbf{Y}_{[1:k-1]} = \mathbf{0}, Y_k = 0, Y'_0 = 1)}{P(Y_0 = 1)} \\
&= \frac{P(Y'_k = 1, \mathbf{Y}_{[1:k-1]} = \mathbf{0}, Y_k = 0 | Y'_0 = 1) P(Y'_0 = 1)}{P(Y_0 = 1)} \\
&= \frac{P(Y'_k = 1, \mathbf{Y}_{[1:k-1]} = \mathbf{0}, Y_k = 0 | Y'_0 = 1) \alpha}{\alpha} \\
&= P(Y'_k = 1, \mathbf{Y}_{[1:k-1]} = \mathbf{0}, Y_k = 0 | Y'_0 = 1). \qquad (3.57)
\end{aligned}
$$

We shall use this result in the next section, in order to approximately determine the *principal overlapping hit probabilities*.

Computation of the principal overlapping hit probabilities

Our aim in this Section is to determine the *principal overlapping hit probabilities*, defined by Equation (3.41) - (3.45). While, the *marginal overlapping hit probabilities* were analytically determined based on the score distribution, this approach is not feasi-

ble for computing the *principal overlapping hit probabilities* since this would require to evaluate a multi-dimensional score distribution

$$\beta_k \propto P(Y_k = 1, \mathbf{Y}_{[1:k-1]} = \mathbf{0}, Y'_0 = 0, Y_0 = 1) \tag{3.58}$$

$$= \sum_{s_k \geq t_s} \sum_{s_{k-1} < t_s} \cdots \sum_{s_0 \geq t_s} P(s_k, s_{k-1}, \cdots s_0 | H_0) \tag{3.59}$$

which is even for small k too time and memory consuming.

Since, an efficient and exact way of determining *principal overlapping hit probabilities* is, to our knowledge, not known, we suggest to approximate the desired quantities. Accordingly, we proceed by proposing an efficient, recursive approach for approximately determining the *principal overlapping hit probabilities* based on the *marginal overlapping hit probabilities*, by successively removing redundancy.

We start by discussing the *principal overlapping hit probabilities*, defined by Equation (3.41) where a single DNA strand is scanned for motif hits. First, since there is no intermediate event for γ_1 between position 0 and 1, we trivially obtain the base case

$$\beta_1 := \gamma_1. \tag{3.60}$$

We proceed by stating the following theorem that is used to derive the *principal overlapping hit probabilities* β_k for $k > 0$.

Theorem 1. *Assuming that the marginal overlapping hit probabilities* γ_i *for* $1 \leq i < M$ *are based on an order-d background model that starts in the stationary distribution (as described in the Introduction) and assuming Assertion* (3.2)*. The principal overlapping*

hit probabilities for $k > 1$ is obtained by

$$\beta_k = \gamma_k - \sum_{i=1}^{k-1} \gamma_{k-i}\beta_i. \tag{3.61}$$

Proof. We show Equation (3.61) by induction. The base case β_1 is immediately obtained by Definition (3.60). Next, we assume that we have already derived $\beta_1, \cdots, \beta_{k-1}$ and show that we can express γ_k in terms of

$$\gamma_k = \beta_k + \sum_{i=1}^{k-1} \gamma_{k-i}\beta_i$$

from which, by rearrangement, we establish β_k.

$$\begin{aligned}
\gamma_k =& P(Y_k = 1|Y_0 = 1) && \text{By Definition (3.18)} \\
=& \sum_{y_1\cdots y_{k-1}\in\{0,1\}^{k-1}} P(Y_k = 1, y_1 \cdots y_{k-1}|Y_0 = 1) && \text{Since } Y_1 \cdots Y_{k-1} \text{ are latent random variables} \\
=& \sum_{y_2\cdots y_{k-1}\in\{0,1\}^{k-2}} P(Y_k = 1, y_2 \cdots y_{k-1}, Y_1 = 0|Y_0 = 1) && \text{Split up the sum into } Y_1 = 1 \text{ and } Y_1 = 0 \\
&+ \sum_{y_2\cdots y_{k-1}\in\{0,1\}^{k-2}} P(Y_k = 1, y_2 \cdots y_{k-1}, Y_1 = 1|Y_0 = 1) \\
=& \sum_{y_2\cdots y_{k-1}\in\{0,1\}^{k-2}} P(Y_k = 1, y_2 \cdots y_{k-1}, Y_1 = 0|Y_0 = 1) \\
&+ \sum_{y_2\cdots y_{k-1}\in\{0,1\}^{k-2}} P(Y_k = 1, y_2 \cdots y_{k-1}|Y_1 = 1)P(Y_1 = 1|Y_0 = 1) && \text{By Assertion (3.2)} \\
=& \sum_{y_2\cdots y_{k-1}\in\{0,1\}^{k-2}} P(Y_k = 1, y_2 \cdots y_{k-1}, Y_1 = 0|Y_0 = 1) \\
&+ \underbrace{P(Y_{k-1} = 1|Y_0 = 1)P(y_1 = 1|Y_0 = 1)}_{=:\gamma_{k-1}\beta_1}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{y_3 \cdots y_{k-1} \in \{0,1\}^{k-3}} P(Y_k = 1, y_3 \cdots y_{k-1}, Y_2 = 0, Y_1 = 0 | Y_0 = 1) && \text{Next, split } Y_2 = 1 \text{ and } Y_2 = 0 \\
&\quad + \sum_{y_3 \cdots y_{k-1} \in \{0,1\}^{k-3}} P(Y_k = 1, y_2 \cdots y_{k-1}, Y_2 = 1, Y_1 = 0 | Y_0 = 1) \\
&\quad + \gamma_{k-1}\beta_1 \\
&= \sum_{y_3 \cdots y_{k-1} \in \{0,1\}^{k-3}} P(Y_k = 1, y_3 \cdots y_{k-1}, Y_2 = 0, Y_1 = 0 | Y_0 = 1) \\
&\quad + \underbrace{P(Y_{k-2} = 1 | Y_0 = 1) P(Y_2 = 1, Y_1 = 0 | Y_0 = 1)}_{=:\gamma_{k-2}\beta_2} && \text{By Assertion (3.2)} \\
&\quad + \gamma_{k-1}\beta_1 \\
& && \text{Repeated until } Y_{k-1} \\
&= \sum_{y_{k-1} \in \{0,1\}} P(Y_k = 1, y_{k-1}, Y_{k-2} = 0 \cdots Y_1 = 0 | Y_0 = 1) \\
&\quad + \sum_{i=1}^{k-2} \gamma_{k-i}\beta_i \\
&= \underbrace{P(Y_k = 1, Y_{k-1} = 0, Y_{k-2} = 0 \cdots Y_1 = 0 | Y_0 = 1)}_{=:\beta_k} && \text{Split } Y_{k-1} = 0 \text{ and } Y_{k-1} = 1 \\
&\quad + \underbrace{P(Y_1 = 1 | Y_0 = 1) P(Y_{k-1} = 1, Y_{k-2} = 0 \cdots Y_1 = 0 | Y_0 = 1)}_{=:\gamma_1\beta_{k-1}} && \text{By Assertion (3.2)} \\
&\quad + \sum_{i=1}^{k-2} \gamma_{k-i}\beta_i \\
\gamma_k &= \beta_k + \sum_{i=1}^{k-1} \gamma_{k-i}\beta_i.
\end{aligned}
$$

Solving for β_k establishes Equation (3.41) and finishes the proof. □

Next, we derive the *principal overlapping hit probabilities* defined by Equation (3.42)-(3.45) for the case when both DNA strands are scanned for motif hits. Analogously to

the single stranded case, we define the bases cases as

$$\beta_{3',0} := \gamma_{3',0} \tag{3.62}$$

$$\beta_{5',1} := \gamma_{5',1} \tag{3.63}$$

which represent the overlapping hit probabilities such that there exists no intermediate events Y.

We next state the theorem for deriving the *principal overlapping hit probabilities* β_k, $\beta_{3',k}$ and $\beta_{5',k}$ for $k > 0$.

Theorem 2. *Assuming that the marginal overlapping hit probabilities* γ_i, $\gamma_{3',i}$ *and* $\gamma_{5',i}$ *for* $1 \leq i < M$ *and* $\gamma_{3',0}$ *are based on an order-d background model that starts in the stationary distribution (as described in the Introduction) and assuming Assertions* (3.3)-(3.8)*. The principal overlapping hit probabilities and for scanning both DNA strands are obtained by*

$$\beta_k = \gamma_k - \sum_{i=1}^{k-1} \gamma_{k-i}\beta_i - \sum_{i=0}^{k-1} \gamma_{5',k-i}\beta_{3',i} \qquad \text{for } 1 \leq k < M \tag{3.64}$$

$$\beta_{3',k} = \gamma_{3',k} - \sum_{i=1}^{k} \gamma_{3',k-i}\beta_i - \sum_{i=0}^{k-1} \gamma_{k-i}\beta_{3',i} \qquad \text{for } 1 \leq k < M \tag{3.65}$$

$$\beta_{5',k} = \gamma_{5',k} - \sum_{i=1}^{k-1} \gamma_{5',k-i}\beta_i - \sum_{i=1}^{k-1} \gamma_{k-i}\beta_{5',i} \qquad \text{for } 1 < k < M. \tag{3.66}$$

Proof. The induction proof for Equation (3.64)-(3.66) is similar to the proof for Theorem 1, where we started from the fact that the intermediate events $Y_1 \cdots Y_{k-1}$ are averaged out for each γ_k and $k > 1$ and showed that the explicit summation over the intermediate events can be recursively reformulated in terms of the γ's and β's. The main difference for the double stranded situation is that we now reduce the summation over

the intermediate positions in the following order $Y_i Y'_i Y_{i+1} Y'_{i+1} \cdots Y_{i+k} Y'_{i+k}$ where $k > 0$.

The base cases are immediately given by Definition (3.62) and (3.63). Assuming that we have already derived $\beta_1, \cdots, \beta_{k-1}$ and $\beta_{3',0} \cdots \beta_{3',k-1}$ we start by showing that γ_k can be expressed in terms of

$$\gamma_k = \beta_k + \sum_{i=1}^{k-1} \gamma_{k-i}\beta_i + \sum_{i=0}^{k-1} \gamma_{5',k-i}\beta_{3',i}$$

which by rearrangement establishes Equation (3.64).

$$\begin{aligned}
\gamma_k =& P(Y_k = 1 | Y_0 = 1) && \text{By Definition (3.18)} \\
=& \sum_{\substack{y_1 \cdots y_{k-1} \in \{0,1\}^{k-1} \\ y'_0 \cdots y'_{k-1} \in \{0,1\}^k}} P(Y_k = 1, y_1 \cdots y_{k-1} y'_0 \cdots y'_{k-1} | Y_0 = 1) && \text{Since } Y_1 \cdots Y_{k-1} Y'_0 \cdots Y'_{k-1} \text{ are latent RVs.} \\
=& \sum_{\substack{y_1 \cdots y_{k-1} \in \{0,1\}^{k-1} \\ y'_1 \cdots y'_{k-1} \in \{0,1\}^{k-1}}} P(Y_k = 1, y_1 \cdots y_{k-1} y'_1 \cdots y'_{k-1}, Y'_0 = 0 | Y_0 = 1) && \text{Split up } Y'_0 = 1 \text{ and } Y'_0 = 0 \\
&+ \underbrace{\sum_{\substack{y_1 \cdots y_{k-1} \in \{0,1\}^{k-1} \\ y'_1 \cdots y'_{k-1} \in \{0,1\}^{k-1}}} P(Y_k = 1, y_1 \cdots y_{k-1} y'_1 \cdots y'_{k-1} | Y'_0 = 1) P(Y'_0 = 1 | Y_0 = 1)}_{= P(Y_k = 1 | Y'_0 = 1) P(Y'_0 = 1 | Y_0 = 1) =: \gamma_{5',k}\beta_{3',0}} && \text{and use Assertion (3.4)} \\
=& \sum_{\substack{y_2 \cdots y_{k-1} \in \{0,1\}^{k-2} \\ y'_1 \cdots y'_{k-1} \in \{0,1\}^{k-1}}} P(Y_k = 1, y_2 \cdots y_{k-1} y'_1 \cdots y'_{k-1}, Y_1 = 0, Y'_0 = 0 | Y_0 = 1) && \text{Next, split } Y_1 = 1 \text{ and } Y_1 = 0 \\
&+ \underbrace{\sum_{\substack{y_2 \cdots y_{k-1} \in \{0,1\}^{k-2} \\ y'_1 \cdots y'_{k-1} \in \{0,1\}^{k-1}}} P(Y_k = 1, y_2 \cdots y_{k-1} y'_1 \cdots y'_{k-1}, Y_1 = 1, Y'_0 = 0 | Y_0 = 1)}_{= P(Y_{k-1} = 1 | Y_0 = 1) P(Y_1 = 1, Y'_0 = 0 | Y_0 = 1) =: \gamma_k \beta_1} && \text{and use Assertion (3.3)} \\
&+ \gamma_{5',k}\beta_{3',0}
\end{aligned}$$

Repeat for $Y'_1 Y_2 Y'_2 Y_3 \cdots Y_{k-1}$

$$= \sum_{y'_{k-1} \in \{0,1\}} P(Y_k = 1, y'_{k-1}, Y_{k-1} = 0, \cdots Y'_1 = 0, Y_1 = 0, Y'_0 = 0 | Y_0 = 1)$$
$$+ \sum_{i=0}^{k-2} \gamma_{5',k-i} \beta_{3',i} + \sum_{i=i}^{k-1} \gamma_{k-i} \beta_i$$

Split $Y'_{k-1} = 1$ and $Y'_{k-1} = 0$

$$= \underbrace{P(Y_k = 1, Y'_{k-1} = 0, Y_{k-1} = 0, \cdots Y'_1 = 0, Y_1 = 0, Y'_0 = 0 | Y_0 = 1)}_{=:\beta_k}$$
$$+ \underbrace{P(Y_k = 1, Y'_{k-1} = 1, Y_{k-1} = 0, \cdots Y'_1 = 0, Y_1 = 0, Y'_0 = 0 | Y_0 = 1)}_{=:\gamma_{5',1}\beta_{3',k-1}}$$
$$+ \sum_{i=0}^{k-2} \gamma_{5',k-i} \beta_{3',i} + \sum_{i=i}^{k-1} \gamma_{k-i} \beta_i$$
$$\gamma_k = \beta_k + \sum_{i=0}^{k-1} \gamma_{5',k-i} \beta_{3',i} + \sum_{i=i}^{k-1} \gamma_{k-i} \beta_i$$

Solving the equation for β_k establishes Equation (3.64).

Next, assuming that we have already derived $\beta_1, \cdots, \beta_k$ and $\beta_{3',0} \cdots \beta_{3',k-1}$, $\gamma_{3',k}$ can be expressed in terms of

$$\gamma_{3',k} = \beta_{3',k} + \sum_{i=1}^{k} \gamma_{3',k-i} \beta_i + \sum_{i=0}^{k-1} \gamma_{k-i} \beta_{3',i}$$

which by rearrangement establishes Equation (3.65).

$$\gamma_{3',k} = P(Y'_k = 1 | Y_0 = 1)$$

By Definition (3.19)

$$= \sum_{\substack{y_1 \cdots y_k \in \{0,1\}^k \\ y'_0 \cdots y'_{k-1} \in \{0,1\}^k}} P(Y'_k = 1, y_1 \cdots y_k y'_0 \cdots y'_{k-1} | Y_0 = 1)$$

Since $Y_1 \cdots Y_k Y'_0 \cdots Y'_{k-1}$ are latent RVs.

$$= \sum_{\substack{y_1 \cdots y_k \in \{0,1\}^k \\ y'_1 \cdots y'_{k-1} \in \{0,1\}^{k-1}}} P(Y'_k = 1, y_1 \cdots y_k y'_1 \cdots y'_{k-1}, Y'_0 = 0 | Y_0 = 1)$$

Split up $Y'_0 = 1$ and $Y'_0 = 0$

$$+ \sum_{\substack{y_1 \cdots y_k \in \{0,1\}^k \\ y'_1 \cdots y_{k-1} \in \{0,1\}^{k-1}}} P(Y'_k = 1, y_1 \cdots y_k y'_1 \cdots y'_{k-1} | Y'_0 = 1) P(Y'_0 = 1 | Y_0 = 1)$$

and use Assertion (3.8)

$$= \sum_{\substack{y_1 \cdots y_k \in \{0,1\}^k \\ y'_1 \cdots y'_{k-1} \in \{0,1\}^{k-1}}} P(Y'_k = 1, y_1 \cdots y_k y'_1 \cdots y'_{k-1}, Y'_0 = 0 | Y_0 = 1)$$

$$+ \underbrace{P(Y'_k = 1 | Y'_0 = 1) P(Y'_0 = 1 | Y_0 = 1)}_{= P(Y_k = 1 | Y_0 = 1) P(Y'_0 = 1 | Y_0 = 1) =: \gamma_k \beta_{3',0}}$$

$$= \sum_{\substack{y_2 \cdots y_k \in \{0,1\}^{k-1} \\ y'_1 \cdots y'_{k-1} \in \{0,1\}^{k-1}}} P(Y'_k = 1, y_2 \cdots y_{k-1} y'_1 \cdots y'_{k-1}, Y_1 = 0, Y'_0 = 0 | Y_0 = 1)$$

Next, split $Y_1 = 1$ and $Y_1 = 0$

$$+ \underbrace{\sum_{\substack{y_2 \cdots y_k \in \{0,1\}^{k-1} \\ y'_1 \cdots y'_{k-1} \in \{0,1\}^{k-1}}} P(Y'_k = 1, y_2 \cdots y_{k-1} y'_1 \cdots y'_{k-1}, Y_1 = 1, Y'_0 = 0 | Y_0 = 1)}_{= P(Y'_{k-1} = 1 | Y_0 = 1) P(Y_1 = 1, Y'_0 = 0 | Y_0 = 1) =: \gamma_{3',k} \beta_1}$$

and use Assertion (3.7)

$$+ \gamma_k \beta_{3',0}$$

Repeat the same way for $Y'_1 Y_2 Y'_2 Y_3 \cdots Y'_{k-1}$

$$= \sum_{y_k \in \{0,1\}} P(Y'_k = 1, y_k, Y_{k-1} = 0, \cdots Y'_1 = 0, Y_1 = 0, Y'_0 = 0 | Y_0 = 1)$$

$$+ \sum_{i=0}^{k-1} \gamma_{k-i} \beta_{3',i} + \sum_{i=i}^{k-1} \gamma_{3',k-i} \beta_i$$

$$= \underbrace{P(Y'_k = 1, Y_k = 0, Y'_{k-1} = 0, Y_{k-1} = 0, \cdots Y'_1 = 0, Y_1 = 0, Y'_0 = 0 | Y_0 = 1)}_{=: \beta_{3',k}}$$

Split $Y_k = 1$ and $Y_k = 0$

$$+ \underbrace{P(Y'_k = 1, Y_k = 1, Y'_{k-1} = 0, Y_{k-1} = 0, \cdots Y'_1 = 0, Y_1 = 0, Y'_0 = 0 | Y_0 = 1)}_{=: \gamma_{3',0} \beta_k}$$

$$+ \sum_{i=0}^{k-1} \gamma_{k-i} \beta_{3',i} + \sum_{i=i}^{k-1} \gamma_{3',k-i} \beta_i$$

$$\gamma_{3',k} = \beta_{3',k} + \sum_{i=0}^{k-1} \gamma_{k-i}\beta_{3',i} + \sum_{i=i}^{k} \gamma_{3',k-i}\beta_i.$$

Solving the equation for $\beta_{3',k}$ establishes Equation (3.65).

Finally, assuming that we have already derived $\beta_1, \cdots, \beta_{k-1}$ and $\beta_{5',1} \cdots \beta_{5',k-1}$, we show that $\gamma_{5',k}$ can be expressed in terms of

$$\gamma_{5',k} = \beta_{5',k} + \sum_{i=1}^{k-1} \gamma_{5',k-i}\beta_i + \sum_{i=1}^{k-1} \gamma_{k-i}\beta_{5',i}$$

which by rearrangement establishes Equation (3.66).

$$\begin{aligned}
\gamma_{5',k} &= P(Y_k = 1 | Y'_0 = 1) && \text{By Definition (3.20)} \\
&= \sum_{\substack{y_1 \cdots y_{k-1} \in \{0,1\}^{k-1} \\ y'_1 \cdots y'_{k-1} \in \{0,1\}^{k-1}}} P(Y_k = 1, y_1 \cdots y_{k-1} y'_0 \cdots y'_{k-1} | Y'_0 = 1) && \text{Since } Y_1 \cdots Y_{k-1} Y'_1 \cdots Y'_{k-1} \text{ are latent RVs.} \\
&= \sum_{\substack{y_2 \cdots y_{k-1} \in \{0,1\}^{k-2} \\ y'_1 \cdots y'_{k-1} \in \{0,1\}^{k-1}}} P(Y_k = 1, y_2 \cdots y_{k-1} y'_1 \cdots y'_{k-1}, Y_1 = 0 | Y'_0 = 1) && \text{Split up } Y_1 = 1 \text{ and } Y_1 = 0 \\
&\quad + \sum_{\substack{y_2 \cdots y_{k-1} \in \{0,1\}^{k-2} \\ y'_1 \cdots y_{k-1} \in \{0,1\}^{k-1}}} P(Y_k = 1, y_1 \cdots y_{k-1} y'_1 \cdots y'_{k-1} | Y_1 = 1) P(Y_1 = 1 | Y'_0 = 1) && \text{and use Assertion (3.5)} \\
&= \sum_{\substack{y_2 \cdots y_{k-1} \in \{0,1\}^{k-2} \\ y'_1 \cdots y'_{k-1} \in \{0,1\}^{k-1}}} P(Y_k = 1, y_1 \cdots y_k y'_1 \cdots y'_{k-1}, Y_1 = 0 | Y'_0 = 1) \\
&\quad + \underbrace{P(Y_{k-1} = 1 | Y_0 = 1) P(Y_1 = 1 | Y'_0 = 1)}_{=:\gamma_{k-1}\beta_{5',1}} \\
&= \sum_{\substack{y_2 \cdots y_{k-1} \in \{0,1\}^{k-2} \\ y'_2 \cdots y'_{k-1} \in \{0,1\}^{k-2}}} P(Y_k = 1, y_2 \cdots y_{k-1} y'_1 \cdots y'_{k-1}, Y'_1 = 0, Y_1 = 0 | Y'_0 = 1) && \text{Next, split } Y'_1 = 1 \text{ and } Y'_1 = 0
\end{aligned}$$

$$+ \underbrace{\sum_{\substack{y_2\cdots y_{k-1}\in\{0,1\}^{k-2}\\ y'_2\cdots y'_{k-1}\in\{0,1\}^{k-2}}} P(Y_k = 1, y_2\cdots y_{k-1}y'_2\cdots y'_{k-1}, Y'_1 = 1, Y_1 = 0|Y'_0 = 1)}_{=:P(Y_{k-1}=1|Y'_0=1)P(Y'_1=1,Y_1=0|Y'_0=1)}$$

and use Assertion (3.7)

$$+ \gamma_{k-1}\beta_{5',1}$$

$$= \sum_{\substack{y_2\cdots y_{k-1}\in\{0,1\}^{k-2}\\ y'_2\cdots y'_{k-1}\in\{0,1\}^{k-2}}} P(Y_k = 1, y_2\cdots y_{k-1}y'_1\cdots y'_{k-1}, Y'_1 = 0, Y_1 = 0|Y'_0 = 1)$$

$$+ \underbrace{P(Y_{k-1} = 1|Y'_0 = 1)\underbrace{P(Y'_1 = 1, Y_1 = 0|Y'_0 = 1)}_{=:\beta_1}}_{\gamma_{5',k-1}\beta_1}$$

Using Equation (3.57)

$$+ \gamma_{k-1}\beta_{5',1}$$

Repeat the same way for $Y_2Y'_2Y_3\cdots Y_{k-1}$

$$= \sum_{y'_{k-1}\in\{0,1\}} P(Y_k = 1, y'_{k-1}, Y_{k-1} = 0, \cdots Y'_1 = 0, Y_1 = 0|Y'_0 = 1)$$

$$+ \sum_{i=1}^{k-1} \gamma_{k-i}\beta_{5',i} + \sum_{i=i}^{k-2} \gamma_{5',k-i}\beta_i$$

$$= \underbrace{P(Y_k = 1, Y'_{k-1} = 0, Y_{k-1} = 0, \cdots Y'_1 = 0, Y_1 = 0|Y'_0 = 1)}_{=:\beta_{5',k}}$$

Finally, split $Y'_{k-1} = 1$ and $Y'_{k-1} = 0$

$$+ \underbrace{P(Y_k = 1, Y'_{k-1} = 1, Y_{k-1} = 0, \cdots Y'_1 = 0, Y_1 = 0|Y'_0 = 1)}_{=:\gamma_{5',k}\beta_k}$$

$$+ \sum_{i=0}^{k-1} \gamma_{k-i}\beta_{5',i} + \sum_{i=i}^{k-2} \gamma_{5',k-i}\beta_i$$

$$\gamma_{5',k} = \beta_{5',k} + \sum_{i=0}^{k-1} \gamma_{k-i}\beta_{5',i} + \sum_{i=i}^{k-1} \gamma_{5',k-i}\beta_i.$$

Solving the equation for $\beta_{5',k}$ establishes Equation (3.66) and completes the proof. □

3.4 Discussion

In this Chapter, we have dealt with statistical dependencies between motif hits that are emitted by the random process $\mathbf{Y}_{[1:N-M+1]}$.

We have elaborated on the similarity of the compatible words $C(t_s)$ and its complementary set $\bar{C}(t_s)$. Due to the fact that the compatible words are relatively similar to one another, obtaining a motif hits $Y = 1$ allows us to predict the underlying DNA sequence more accurately than if we had observed $Y = 0$ instead. Since, $Y = 1$ is informative of the underlying DNA sequence, we argue that the statistical independence assertions (3.2)-(3.8) hold approximately, based on the fact that if the underlying DNA sequence could be observed instead, these independence assertions would hold exactly. Subsequently, Assertions (3.2)-(3.8) formed the basis for approximating the *principal overlapping hit probabilities* and were also implicitly used in Pape *et al.* [36] for extending the clumps recursively by the extension factors (the ξ's), while their validity was not elaborated there.

Next, we focused on statistical dependences that may occur between non-overlapping motif hits as well as overlapping motif hits.

Non-overlapping motif hits are caused by the dependence between nucleotides in order-d background model with $d > 0$. We discussed the application of eigenvalue decomposition in order to quantify long-range statistical dependence between two distantly located nucleotides w_i and w_j. We illustrated that long-range interactions decay very quickly in a case study of human DNase-I-hypersensitive sites using an order-1 background model [50], suggesting that for studying regulatory regions, associations between non-overlapping hits can be ignored. However, when studying short repetitive

regions in the DNA with higher-order Markov models, the assumption might be violated, in which case non-overlapping motif hits may be considerably associated with one another. Studying repetitive regions is, however, beyond the scope of this thesis.

We have addressed overlapping motif hits from a qualitative and quantitative perspective. While, in general, motif hits may overlap with other motif hits due to the use of a sliding window approach, scanning both strands of a double-stranded DNA sequence also give rise to a special form of overlapping motif hits, namely overlapping motif hits due to base pair complementarity, which need to be accounted for as well. While, previously, two types of overlapping hits were advocated, 1) overlapping hits on the same strand and 2) 3'-end overlapping hits [36], we argue for the use of three canonical types of overlapping hits when both DNA strands are scanned: 1) Overlapping hits on the same strand, 2) 3'-end overlapping hits and 3) 5'-end overlapping hits.

We defined associated probabilities for the three canonical ways of forming overlapping motif hits, termed *marginal overlapping motif hits*, which were computed based on an algorithm for determining the two-dimensional score distributions, which was first suggested by Pape *et al.* [36] for an order-0 background model. As a part of this thesis, we extended the algorithm to general order-d background models.

Finally, we discussed the concept of *periodicity*, which was originally exploited in the word pattern statistics community to analyze the self-overlapping word matches [41]. and self-overlap of generalized strings [32]. As part of the thesis, we adopted the *periodicity*-concept for TF motif hits that are acquired based on the log-likelihood ratio (see Section 1.2.3). We argue that *marginal overlapping motif hit probabilities* correspond to the probabilities of obtaining motif hits at periods. However, along the line of Reinert *et al.* [41], the periods may contain redundant information about overlapping

hits. Therefore, it is desired to reduce the set of periods to a non-redundant set of so-called *principal periods*, which are the root cause for all periods. We established the probabilities of obtaining overlapping motif hits at principal periods, termed *principal overlapping motif hits*, by means of a novel approximative procedure (see Theorem 1 and 2). Moreover, we used the *principal overlapping motif hits* derive the probability that a clump ends.

The results of this Chapter provide the basis for the Chapters 4 and 6 where we turn to the distribution of the number of motif hits in DNA sequences.

Chapter 4

Compound Poisson distribution

In the previous Chapter we have established the notion of motif hits and overlapping hits. We described algorithms for quantifying the occurrence rate of individual motif hits as well as for quantifying the probability of obtaining overlapping motif hits. In this chapter, we turn to the question of how many motif hits are likely to be present in a DNA sequence of given length which shall be the basis for the *motif hit enrichment test*.

A model that has proved to be particularly well suited for modeling the motif hit counts distribution is the compound Poisson model which was initially advocated by the word count statistics community [41] and later adopted by Pape *et al.* [36], to describe the number of motif hits based on the log-likelihood ratio (see Section 1.2.3).

We shall start this Chapter by reviewing the compound Poisson model proposed by Pape *et al.* [36] as it provides the common basis between the previously proposed model the variants that we developed in the course of this thesis. Along the line of Pape *et al.*, we therefore consider scanning both strands of a DNA sequence for motif hits.

Subsequently, we discuss two variants for the computation of the *clump size distribution*. The first one attempts to improve upon the previously proposed algorithm [36], by utilizing three types of overlapping hits (on the same strand, by 3'-end complementarity and by 5'-end complementarity) that were discussed in Chapter 3, instead of only two types. Furthermore, overlapping hit probabilities shall be determined based on an order-d Markov model. Apart from that, the core of the methodology shall remain unchanged. The main incentive for the first model is to allow for fair comparisons with the second (further advanced) variant for computing the clump size distribution. For the second variant, we derive the clump size distribution based on the *principal overlapping hit probabilities* rather than directly based on the *marginal overlapping hit probabilities* (see Chapter 3). We refer to the compound Poisson model, that evaluates the clump size distribution based on the second variant, as the *novel* compound Poisson model. We notice that the latter approach is related to the approach proposed by Marschall [32], who developed the expected clump size based on a similar idea. However, Marschall [32] concerned with the enrichment of exact matches of (generalized) strings, whereas in this thesis motif hit enrichment is studied for motif hits that are called based on the log-likelihood ratio with a fixed score threshold t_s.

We systematically compare the compound Poisson variants with the binomial model to describe the distribution of the number of motif hits, which reveals that the novel compound Poisson approximation achieves accurate results across a larger range of parameters (e.g. α) and across different motif structures (e.g. self-overlapping and non-self-overlapping motifs).

4.1 Compound Poisson model

In this Section, we review the general framework of the compound Poisson model for modeling the number of motif hits in DNA sequences as proposed by Pape *et al.* [36]. The setup described in this section remains the same between the previous model, an improved model and a novel compound Poisson model that we discuss in this chapter. The differences between the models shall be discussed in Section 4.2.

As an alternative to Definition (1.33), where the number of motif hits X was expressed by the Poisson distributed numbers of *k-clump* occurrences C_k, we can equivalently express X by a single Poisson distributed total number of clump occurrences C (regardless of the clump size) and an independent random variable describing the size of the clump K

$$X = \sum_{k>0} kC_k = \sum_{c=1}^{C} K_c \tag{4.1}$$

where we have used the set of random variables $\{K_1, K_2, \cdots, K_C\}$ which represent the sizes of the first clump, second clump, etc. Accordingly, the indexed clump sizes K_i denote independently and identically distributed instances of the generic clump size variable random variable K [5].

Accordingly, the total number of clump occurrences $C = \sum_{k>0} C_k$ is parametrized by

$$C \sim \text{Poisson}(\lambda)$$

with

$$\lambda = \sum_{k>0} \lambda_k \tag{4.2}$$

where λ_k denotes the occurrence rate of a *k-clump*.

In principle, the probability of observing a clump size k is defined by the fraction of the number of *k-clumps* relative to the total number of clumps (regardless of their size) that are obtained from scanning an infinitely long sequence

$$\theta_k := P(K = k) = \lim_{C \to \infty} \frac{C_k}{C}. \tag{4.3}$$

We defer the computation of the clump size probability θ_k to the subsequent sections of this Chapter.

As elaborated in Pape *et al.* [36], because 1) $K_1 \cdots K_C$ are i.i.d. according to $P(K = k)$, 2) the random variables C and K are statistically independent, and 3) by linearity of the expectation operator we can express the expected number of motif hits as

$$\begin{aligned} \mathbf{E}[X] &= \mathbf{E}_C[\mathbf{E}_k[\sum_{c=1}^{C} K_c]]] \\ &= \mathbf{E}_C[C\mathbf{E}_K[K]] \\ &= \mathbf{E}_C[C]\mathbf{E}_K[K] \\ &= \mathbf{E}_C[C] \sum_{k>0} k\theta_k \end{aligned} \tag{4.4}$$

where, in the last line, we have used the definition of the expected clump size.

Alternatively, the expected number of motif hits can also be expressed as

$$\mathbf{E}[X] = 2 \times \alpha \times (N - M + 1) \tag{4.5}$$

where we make use of the false positive hit probability α from the *motif score test*, the length of the DNA sequence N and the length of the motif M. We multiply by two because both strands of the DNA are scanned for motif hit occurrences. By substituting the right hand side of Equation (4.5) into Equation (4.4) and solving for $\mathbf{E}_C[C]$ we obtain the rate of *clump* occurrence regardless of the *clump*-size

$$\mathbf{E}_C[C] = \lambda = \frac{2\alpha(N - M + 1)}{\sum_{k>0} k\theta_k}. \tag{4.6}$$

Finally, once we have determined the clump size probabilities θ_k (see next sections) and the rate of clump formation λ defined by (4.6), the final compound Poisson distribution can be determined according to Kemp, 1967 [27]

$$P(X = 0) = e^{-\lambda} \tag{4.7}$$

$$P(X = x + 1) = \frac{\lambda}{x + 1} \sum_{x'=1}^{x} (x - x' + 1)\theta_{x-x'+1} P(X = x'). \tag{4.8}$$

which was also advocated by Pape *et al.* [36].

4.2 Clump size probability

In the previous Section, we discussed the computation of the compound Poisson distribution which depends on the clump occurrence rate λ and the clump size probabilities θ_k

for $k > 0$. However, we have yet to discuss how θ_k is obtained. In this Section, we turn to computing the clump size probabilities θ_k, which is critical for obtaining an accurate compound Poisson model.

As suggested earlier, the clump size probability θ_k can be determined iteratively from the probability θ_{k-1} by successively extending additional hits at the end of the clump [36].

Therefore, we dedicate the next two subsections to discussing the computation of the clump size probability θ_k. First, we shall we discuss the approach that was proposed by Pape *et al.* [36] for computing the clump size distribution as well as the expected clump size for which we present an *improved* version upon the original version. Second, we propose a *novel* approach of determining the clump size probability which exploits the *principal overlapping hit probabilities* instead of the *marginal overlapping hit probabilities*.

4.2.1 Clump size probability - Improvement upon Pape *et al.*

The computation of the clump size probability θ_k for score-based motif hits was first proposed by Pape *et al* [36]. In this Section, we reiterate and improve upon the main idea of the original model for computing the probability of obtaining a *k-clump* θ_k. The improvements concern with 1) the use a general order-d background model for deriving the score-based statistics and 2) the use of three types of overlapping hits.

Probability of a 1-*clump*

A 1-*clump* is defined by a single motif hit which is not overlapped by any additional hits up- and downstream, neither on the forward strand nor one the reverse strand. When scanning both DNA strands, there are two types of 1-*clumps*: 1) A clump that contains a forward strand hit and 2) a clump that contains a reverse strand hit (see Figure 4.1). The probability of a 1-*clump* containing a forward strand and a reverse strand hit are

(a) 1-*clump* with forward strand hit (b) 1-*clump* with reverse strand hit

Figure 4.1: Examples of a 1-*clump* for a motif of length $M = 4$ with respect to scanning the forward and reverse strand of a DNA sequence. The black square denotes the position within the clump at which the hit occurs, whereas, the white boxes denote overlapping positions at which no hits occur.

respectively defined by

$$\theta_1^f = P(\mathbf{Y}_{[1:M-1]} = \mathbf{0}, Y'_0 = 0 | Y_0 = 1, \mathbf{Y}_{[-1:-(M-1)]} = \mathbf{0}) \tag{4.9}$$

$$\theta_1^r = P(\mathbf{Y}_{[1:M-1]} = \mathbf{0}, Y_0 = 0 | Y'_0 = 1, \mathbf{Y}_{[-1:-(M-1)]} = \mathbf{0}) \tag{4.10}$$

where $\mathbf{Y}_{[i:j]} = \mathbf{0}$ denotes a stretch of non-hits from position i to (including) j.

Since θ_1^f and θ_1^r is too difficult to determined directly, Pape *et al.* [36] proposed the simplification of assuming $Y_1 \cdots Y_{M-1} Y'_0 \cdots Y_{M-1}$ to be independent given the hits $Y_0 = 1$ and $Y'_0 = 1$. This facilitates the use of the *marginal overapping hit probabilities* defined by (3.18)-(3.20) in order to approximate θ_1^f and θ_1^r as

$$\theta_1^f \approx (1 - \gamma_{3',0}) \prod_{i=1}^{M-1} (1 - \gamma_i)(1 - \gamma_{3',i}) \tag{4.11}$$

$$\theta_1^r \approx (1-\gamma_{3',0}) \prod_{i=1}^{M-1} (1-\gamma_i)(1-\gamma_{5',i}). \tag{4.12}$$

Computing the probability of a *k-clump* for $k > 1$

As proposed earlier [36], the probability of a k-clump with $k > 1$ can be evaluated iteratively from the probability of a $k-1$-clump by appending another overlapping hit at the end of the clump. Therefore, we proceed by deriving expressions (so-called extension factors) that multiply in an additional hit at the end of the clump for θ_{k-1}^f and θ_{k-1}^r such that we subsequently obtain θ_k^f and θ_k^r.

First, consider a $k-1$-*clump* that ends in a forward strand hit and whose probability is given by θ_{k-1}^f. Every clump is finished off by a stretch of non-hits, which excludes further overlapping hits after the last hit in the clump (see Figure 4.2a). To append another forward strand hit i positions after the last hit, 1) the non-hits at and after position i occurring with probability $\prod_{j=i}^{M-1}(1-\gamma_j)(1-\gamma_{3',j})$ are divided out (see Figure 4.2b) and 2) a forward strand hit is multiplied in with γ_i along with a stretch of non-hits with probability $\prod_{j=1}^{M-1}(1-\gamma_j)\prod_{j=0}^{M-1}(1-\gamma_{3',j})$ so that the newly formed clump again ends with $M-1$ non-hits (see Figure 4.2c). This leads to the following extension factor (for the $f \to f$ scenario, which stands for forward strand hit followed by a forward strand hit) for position i

$$\xi_{f\to f}^i := \frac{\overbrace{\gamma_i}^{\text{Additional hit}} \overbrace{\prod_{j=1}^{M-1}(1-\gamma_j)\prod_{j=0}^{M-1}(1-\gamma_{3',j})}^{\text{New end of the clump}}}{\underbrace{\prod_{j=i}^{M-1}(1-\gamma_j)(1-\gamma_{3',j})}_{\text{Truncated previous end of the clump}}} \qquad \text{for } 1 \le i < M. \tag{4.13}$$

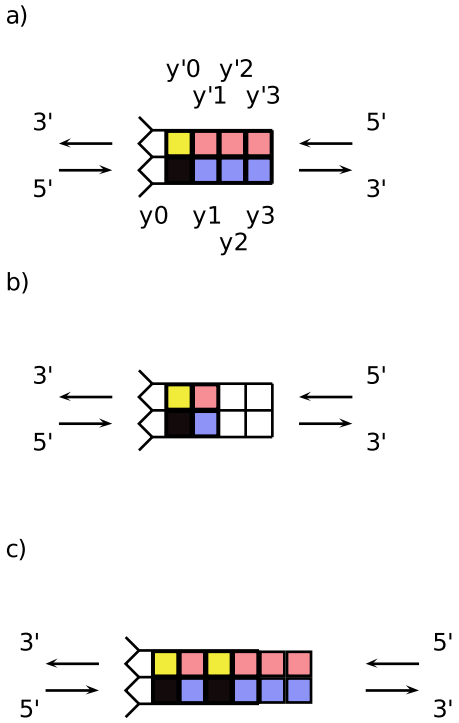


Figure 4.2: Extension of a $k-1$-*clump* for a motif of length $M = 4$ by another overlapping hit. a) A clump that ended in a forward strand hit (black box) can potentially be extended by another forward strand hit (blue boxes), a palindromic hit (yellow box) or a (non-palindromic) reverse strand hit (red boxes). b) In order to append another overlapping hit at position $y2$, the non-hit probabilities at $y2$, $y3$, $y'2$ and $y'3$ are divided out. c) Then, the probability of obtaining another overlapping hit at $y2$ is multiplied in along with the non-hit probabilities to finalize the new clump.

According to a completely analogous argument, the corresponding extension factors with respect to the remaining strandedness combinations for the last hit in the $k-1$-*clump* and the last hit in the *k-clump* are given by

$$\xi^i_{f\to r} := \frac{\gamma_{3',i}\prod_{j=1}^{M-1}(1-\gamma_j)\prod_{j=1}^{M-1}(1-\gamma_{5',j})}{\prod_{j=i}^{M-1}(1-\gamma_j)(1-\gamma_{3',j})} \qquad \text{for } 0 \le i < M \qquad (4.14)$$

$$\xi^i_{r\to f} := \frac{\gamma_{5',i}\prod_{j=1}^{M-1}(1-\gamma_j)\prod_{j=0}^{M-1}(1-\gamma_{3',j})}{\prod_{j=i}^{M-1}(1-\gamma_j)(1-\gamma_{5',j})} \qquad \text{for } 1 \le i < M \qquad (4.15)$$

$$\xi^i_{r\to r} := \xi^i_{f\to f} \qquad \text{for } 1 \le i < M. \qquad (4.16)$$

Note that in the case when a forward strand is followed by a reverse strand hit and vice versa, the denominator and numerator employ $(1-\gamma_{3',j})$ and $(1-\gamma_{5',j})$ such that the final hits in the $k-1$-*clump* and in the *k-clump* are correctly conditioned on.

While, the extension factors defined by (4.13)-(4.16), are used to extend a hit to the clump size probability θ_{k-1} for a specific position i, aggregating the $\xi^i_{f\to f}, \xi^i_{f\to r}, \xi^i_{r\to f}$ and $\xi^i_{r\to r}$ across all positions i yields the extension factors such that another hit occurs at any possible position

$$\xi_{f\to f} := \sum_{i=1}^{M-1} \xi^i_{f\to f} \tag{4.17}$$

$$\xi_{f\to r} := \sum_{i=0}^{M-1} \xi^i_{f\to r} \tag{4.18}$$

$$\xi_{r\to f} := \sum_{i=1}^{M-1} \xi^i_{r\to f} \tag{4.19}$$

$$\xi_{r\to r} := \xi_{f\to f}. \tag{4.20}$$

The extension factors defined by (4.17)-(4.20) can now be used to extend a $k-1$-*clump* by an additional final overlapping hit, regardless of the exact overlapping position and considering the previous and new strands of the final hits. This results in the following iterative approach that was introduced by Pape *et al.* [36] to obtain the probability of a *k-clump* according to

$$\begin{pmatrix} \theta^f \\ \theta^r \end{pmatrix}_k = \begin{bmatrix} \xi_{f\to f} & \xi_{r\to f} \\ \xi_{f\to r} & \xi_{r\to r} \end{bmatrix} \cdot \begin{pmatrix} \theta^f \\ \theta^r \end{pmatrix}_{k-1}. \tag{4.21}$$

Finally, due to the fact that θ^f_k and θ^r_k are associated with mutually exclusive events, the

total probability of observing a *k-clump* is simply obtained summing up the contributions w.r.t. the strandedness

$$\theta_k = \theta_k^f + \theta_k^r \tag{4.22}$$

and re-normalizing θ_k such that

$$\sum_{k>0} \theta_k = 1.$$

Notice that the update rule stated by Equation (4.21) only works if extension factors $\xi_{f\to f}$, $\xi_{f\to r}$, $\xi_{r\to f}$ and $\xi_{r\to r}$ are strictly positive and finite. However, for a perfect palindrome, it is the case that $\gamma_{3',0} = 1$, which means that $\xi^0_{f\to r}$ is undefined, because the denominator contains the factor $(1 - \gamma_{3',0})$. Consequently, also Equation (4.18) is undefined. This issue is caused by the fact that palindromic motifs only allow for even numbers of motif hits, while odd numbers occur with probability zero. We solve this issue by adding a small quantity ϵ (e.g. $\epsilon = 10^{-8}$) to $\gamma_{3',0}$ and $(1-\gamma_{3',0})$ and re-normalize those probabilites to enforce that the extention factors are strictly positive and finite.

Even though, conceptually this procedure that we discussed in this section is along the line of Pape *et al.* [36], we have improved the procedure in several ways. For a detailed comparison between the original version and the version present in this thesis the reader is referred to the discussion of this chapter.

4.2.2 Novel clump size probabilities

In the previous Section, we have discussed an iterative approach for determining the clump size probabilities according to the work of Pape *et al.* [36] with some slight im-

provements. In this section, we shall discuss a similar iterative strategy which, however, takes advantage of the newly derived *principal overlapping hit probabilities* instead of the *marginal overlapping hit probabilities* (see Section 3.3.3). The motivation for this is that the *marginal overlapping hit probabilities* are associated with the probabilities of obtaining hits at periods, which describe overlapping hits in a redundant fashion [41]. As a consequence, using the *marginal overlapping hit probabilities* directly for deriving the clump size probabilities induces biases to the model which might result in overly conservative clump size distributions. By contrast, the *principal overlapping hit probabilities* approximate the probabilities of obtaining overlapping hits at *principal periods*, which are defined to be non-redundant. That is, the principal periods form the root cause of all periods [41]. Thus, determining the clump size distribution based on the *principal overlapping hit probabilities* will in general yield more accurate results compared to when *marginal overlapping hit probabilities* are used. While, it has been advocated to use the principal periods directly for studying self-overlapping exact word matches [41, 32], to our knowledge, this concept was never used with TF motif hits that are derived using the log-likelihood ratio. We also notice that the following discussion is related to Marschall, 2011 [32], who used a similar approach to derive the expected clump size in the context of enrichment testing for generalized strings as motifs.

As in the previous section, we start our discussion with the 1-*clump* probability, which is defined by (4.9) and (4.10). We can derive the probability that a single forward strand hit and a single reverse strand hit with no further overlapping hits by employing the Definition of δ_{M-1}, δ'_{M-1} and $\beta_{3',0}$, defined by (3.51), (3.52) and (3.43), respectively, as

follows

$$\begin{aligned}\begin{pmatrix}\theta^f\\ \theta^r\end{pmatrix}_1 &= \begin{pmatrix}P(\mathbf{Y}_{[1:j]} = \mathbf{0}, Y'_0 = 0 | Y_0 = 1)\\ P(\mathbf{Y}_{[1:j]} = \mathbf{0}, Y_0 = 0 | Y'_0 = 1)\end{pmatrix}\\ &= \begin{pmatrix}P(\mathbf{Y}_{[1:j]} = \mathbf{0}, Y'_0 = 0 | Y_0 = 1)\\ P(\mathbf{Y}_{[1:j]} = \mathbf{0} | Y'_0 = 1) P(Y_0 = 0 | Y'_0 = 1)\end{pmatrix}\\ &= \begin{pmatrix}\delta_{M-1}\\ (1 - \beta_{3',0})\delta'_{M-1}\end{pmatrix}. \end{aligned} \tag{4.23}$$

where θ^f_1 is equivalent to δ_{M-1} by definition. Moreover, to express θ^r_1, we factorize the expression in the second line by applying Assertion (3.4). Since, palindromic hits occur symmetrically, we have $1 - \beta_{3',0} = P(Y'_0 = 0 | Y_0 = 1) = P(Y_0 = 0 | Y'_0 = 1)$ and recognize the remaining factor to be δ'_{M-1}.

Next, we derive the necessary expressions for extending a given $k-1$-*clump* by an additional hit i positions after the last hit, which establishes the k-*clump* for a specific i such that $0 \leq i < M$. The idea for extending motif hits follows a similar principle as discussed in the previous sections: First, the probability of the final non-hits with respect to the $k-1$-*clump* are divided out, second, an additional overlapping hit is multiplied in with probability, and third, the final non-hits are multiplied in with respect to the new hit with probability. This leads to the following extension factor for a specific overlapping hit position i and the case that the final forward strand hit is followed by another forward strand hit

$$\xi^{i*}_{f\to f} = \frac{\overbrace{\beta_i}^{\text{Additional hit}}}{\underbrace{\delta_{M-1}}_{\text{Previous end of clump}}} \times \underbrace{\delta_{M-1}}_{\text{New end of clump}} = \beta_i \qquad \text{For } 1 \leq i < M. \tag{4.24}$$

where the $*$ symbol indicates that the *principal overlapping hit probabilities* were used, as opposed to the *marginal overlapping hit probabilities*. Similarly, in the case where a reverse strand hit is followed by a reverse strand hit we have

$$\begin{aligned}
\xi^{i*}_{r\to r} &= \frac{\overbrace{P(Y'_i = 1, \mathbf{Y}_{[1:i-1]} = \mathbf{0}, Y_i = 0 | Y'_0 = 1)}^{\text{Additional hit}}}{\underbrace{\delta'_{M-1}}_{\text{Previous end of clump}}} \times \underbrace{\delta'_{M-1}}_{\text{New end of clump}} \\
&= \frac{\overbrace{P(Y_i = 1, \mathbf{Y}_{[1:i-1]} = \mathbf{0}, Y'_0 = 0 | Y_0 = 1)}^{=:\beta_i}}{\delta'_{M-1}} \times \delta'_{M-1} \\
&= \beta_i \qquad \text{For } 1 \le i < M. \qquad (4.25)
\end{aligned}$$

where the second line follows from Equation (3.57), as a consequence of symmetry. Therefore, $\xi^{i*}_{f\to f} = \xi^{i*}_{r\to r}$. Finally, we obtain the extension expressions for hits that are overlapped by hits on the complementary strand as

$$\xi^{i*}_{f\to r} = \frac{\beta_{3',i}}{\delta_{M-1}} \delta'_{M-1} \qquad \text{For } 0 \le i < M \qquad (4.26)$$

$$\xi^{i*}_{r\to f} = \frac{\beta_{5',i}}{\delta'_{M-1}} \delta_{M-1} \qquad \text{For } 1 \le i < M. \qquad (4.27)$$

We aggregate the extention factors across the specific overlapping hit position i which yields

$$\xi^{*}_{f\to f} = \sum_{i=1}^{M-1} \beta_i \qquad (4.28)$$

$$\xi^{*}_{f\to r} = \frac{\delta'_{M-1}}{\delta_{M-1}} \sum_{i=0}^{M-1} \beta_{3',i} \qquad (4.29)$$

$$\xi^{*}_{r\to f} = \frac{\delta_{M-1}}{\delta'_{M-1}} \sum_{i=1}^{M-1} \beta_{5',i} \qquad (4.30)$$

$$\xi^*_{r\to r} = \xi^*_{f\to f} \tag{4.31}$$

As discussed in the previous section, we determine the clump size probabilities θ^f_k and θ^r_k iteratively according to the following update rule

$$\begin{pmatrix} \theta^f \\ \theta^r \end{pmatrix}_k = \begin{bmatrix} \xi^*_{f\to f} & \xi^*_{r\to f} \\ \xi^*_{f\to r} & \xi^*_{r\to r} \end{bmatrix} \cdot \begin{pmatrix} \theta^f \\ \theta^r \end{pmatrix}_{k-1} . \tag{4.32}$$

The final clump size probabilities θ_k regardless of the strandedness are given by

$$\theta_k = \theta^f_k + \theta^r_k \tag{4.33}$$

which are renormalized such that $\sum_{k>0} \theta_k = 1$.

For the same reason as stated in the previous section, we also need to enforce that the extension factors $\xi^*_{f\to f}$, $\xi^*_{f\to r}$, $\xi^*_{r\to f}$ and $\xi^*_{r\to r}$ are always defined. As mentioned before, this is achieved by adding a small quantity ϵ (e.g. $\epsilon = 10^{-8}$) to $\beta_{3',0}$ and $(1-\beta_{3',0})$, which prevents $(1-\beta_{3',0})$ from being equal to zero. We re-normalize the obtained quantities to ensures that they are valid probabilities again.

We want to emphasize that the matrix that contains the extension factors in Equation (4.32) is conceptually analogous to the overlap matrix that was used in Marschall, 2011 [32], which was employed to determine the expected clump size.

4.3 Results

In this Section, we systematically compare the accuracy of the two compound Poisson model variants, which differ in the method for computing the clump size distribution. We refer to the variant that uses the *marginal overlapping hit probabilities* directly to approximate the clump size distribution as the *improved* model which is denoted by $P^{I}_{CP}(X)$, and we refer to the variant that is based on the *principal overlapping hit probabilities* as the *novel* model which is denoted by $P^{N}_{CP}(X)$. In addition, we compare the performance of the compound Poisson models with the simple binomial model, denoted by $P_{Bin}(X)$. As a reference for the comparisons, we determine the empirical distribution of the number of motif hits, denoted by $P_{E}(X)$, via a sampling strategy. In the following subsections, we shall first compared the accuracy of each model using different significance levels, motif structures and background model orders. Then we compared the performance between the models on a large collection of known motifs from Transfac in order to assess the practical relevance of the improvements. Finally, we studied how different background model orders for describing DNA sequences may affect the distribution of the number of motif hits.

4.3.1 Comparison between models for the motif hit counts

We systematically compared the accuracies of $P^{N}_{CP}(X)$, $P^{I}_{CP}(X)$ and $P_{Bin}(X)$ relative to $P_{E}(X)$ for various parameter settings, including various background model orders $d \in \{0, 1, 2\}$, various significance levels $\alpha \in \{10^{-2}, 10^{-3}\}$ (with respective DNA sequence lengths 1kb and 10kb) and three types of motif structures (a palindrome, a repetitive motif and a non-repetitive motif; see Figure 2.5a, 2.5b and 2.5c, respectively). We

estimated the background models on a subset of DNase-I-hypersensitive sites from ENCODE [50]. We used the improved compound Poisson model $P^I_{CP}(X)$ instead of the original compound Poisson model to ensure a fair comparison with the novel version.

The empirical distribution, which serves as a reference, is generated by sampling 100 batches of 1000 sequences according to the background model, with each sequence being 1 kb and 10 kb in length for $\alpha = 0.01$ and $\alpha = 0.001$, respectively. While the final empirical distribution is determined using the 100×1000 random samples as

$$P_E(X = x) = \frac{\text{\# of sequences that contain } x \text{ hits}}{100 \times 1000} \tag{4.34}$$

we additionally determine the 25 and 75 percentile of $P_E(X = x)$ per x by quantifying the variability that is caused by the sampling noise across the 100 batches (see error bars in Figure 4.3-4.8).

In order to measure the discrepancy between the analytically derived distributions and the empirical distribution, we employ the total variation distance measure

$$d(P_E, Q) := \sum_{x=0}^{X_{max}} |P_E(x) - Q(x)| \tag{4.35}$$

where we substitute $P^N_{CP}(X)$, $P^I_{CP}(X)$ or $P_{Bin}(X)$ for $Q(X)$, depending on the context. While, Equation (4.35) measures the similarity of the distribution on the entire range of motif hit counts x, we are particularly interested in how well the approximative distributions $Q(X)$ recapitulate the right tail of $P_E(X)$, since that regime is most critical for determining motif hit enrichment. Accordingly, we employ another variant of the total

variation distance

$$d_{5\%}(P_E, Q) := \sum_{x=x_{95\%}}^{X_{max}} |P_E(x) - Q(x)| \tag{4.36}$$

in order to quantify the accuracy of $Q(X)$ in the 5% significance region with respect to $P_E(X)$.

As expected, the motif hit counts distributions for the three different motif structures are markedly different with respect to the shapes of their distributions. While for non-self-overlapping motifs, the distribution is reminiscent of a normal distribution (see Figure 4.7), self-overlapping motifs generally exhibit heavier tails (see Figures 4.3 and 4.5). That is, their variances are increased compared to non-self-overlapping motifs. In addition to the increased variance, for palindromic motifs it becomes apparent that odd motif hit counts occur much less frequently than even motif hit counts, because forward and reverse strand hits are tightly coupled. In the extreme case, when a forward strand hit is *always* matched with a reverse strand hit (e.g. if $\gamma_{3',0} = 1$), odd numbers of motif hits are in fact impossible (see Figure 4.3).

As expected, the binomial distribution $P_{Bin}(X)$ only achieves accurate results for the non-self-overlapping motif, while it fails to capture the increased variance for the palindromic and the repetitive motif (see Figures 4.3, 4.5 and 4.7). The reason for that is that self-overlapping motif hits are ignored in the binomial setting. As a consequence, the use of $P_{Bin}(X)$ to evaluate *motif hit enrichment* causes a substantial excess of false positives for self-overlapping motifs.

By contrast, generally the compound Poisson models are well suited for non-self-overlapping as well as self-overlapping motif. Our comparison shows that while, $P^I_{CP}(X)$ achieves

accurate results for the palindromic motif and the non-self-overlapping motif (see Figures 4.3 and 4.7), it incurs biases for the repetitive motif (see Figure 4.5). These biases are caused by using the *marginal overlapping hit probabilities* instead of the *principal overlapping hit probabilities* for estimating the clump size distribution, which leads to a systematic overestimation of the large clump sizes and, consequently, an increase of the variance of $P^I_{CP}(X)$ compared to $P_E(X)$. On the other hand, our novel compound Poisson model $P^N_{CP}(X)$ represents the shape of the motif hit counts distribution accurately across all motif structures provided that the significance level is chosen stringently enough (see bold entries in Tables 4.1-4.3).

As the the compound Poisson models assume motif hits to occur only rarely (Poisson assumption), choosing a relaxed significance level α (e.g. $\alpha \geq 0.01$) leads to a biased approximation (Tables 4.1-4.3 for $\alpha = 0.01$). The violation of the Poisson assumption is evident for $\alpha = 0.01$ and the non-self-overlapping motif where the binomial model achieves more accurate results compared to the compound Poisson models. Even though, the binomial model also assumes motif hits to occur only rarely, it reacts less sensitively to violations of this assumptions (see Table 4.3 for $\alpha = 0.01$).

Next, we observe that different choices of background model orders d yield similarly accurate results (see Tables 4.1, 4.2, 4.3) This suggests that the simplifying assumptions that were made for dealing with order-d background models induced negligible biases. Those simplifying assumptions include neglecting the preceding context of length d (for higher-order background models) for computing the score (see Section 1.2.4) as well as assuming that non-overlapping outcomes Y_i and Y_{i+k} are independent with $k \geq M$ (see Section 3.2).

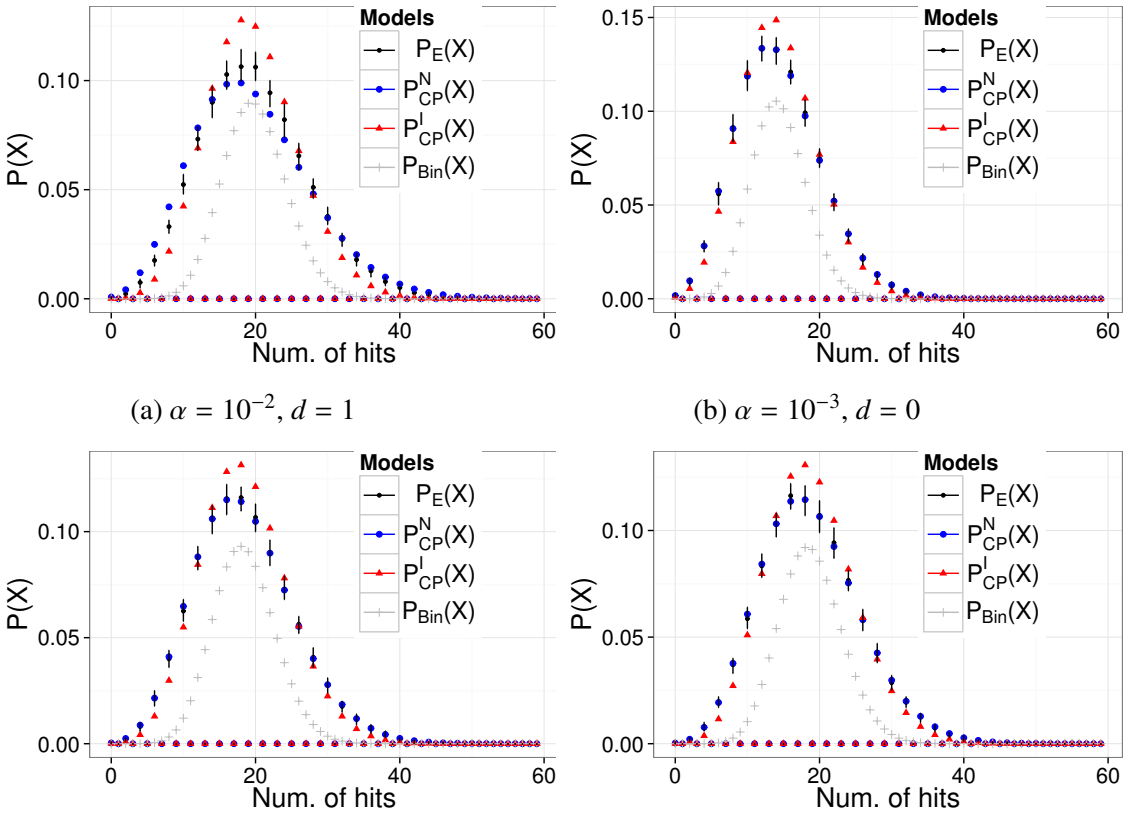
Then we asked whether the right tail of the motif hit counts distribution is represented

Table 4.1: Accuracy of the analytic models for the palindromic motif (Figure 2.5a) measured by Equation (4.35) - (4.36). The values marked in bold underline the most accurate analytic model (P^N_{CP}, P^I_{CP} or P_{Bin}) compared to $P_E(X)$.

d	α	$d(P_E, P^N_{CP})$	$d(P_E, P^I_{CP})$	$d(P_E, P_{Bin})$	$d_{5\%}(P_E, P^N_{CP})$	$d_{5\%}(P_E, P^I_{CP})$	$d_{5\%}(P_E, P_{Bin})$
1	10^{-2}	**0.101**	0.177	1	**0.0145**	0.0356	0.0748
0	10^{-3}	**0.0116**	0.103	1	**0.00254**	0.0198	0.0789
1	10^{-3}	**0.0135**	0.129	1	**0.00234**	0.0239	0.0708
2	10^{-3}	**0.0146**	0.122	1	**0.00349**	0.0228	0.0754

accurately, because this regime is important for the *motif hit enrichment test*. To this end, we employ Equation (4.36) to measure the distance between the analytically derived models and $P_E(x)$. We observe that the discrepancy measures defined by Equation (4.35) and (4.36) between the analytically derived models and the empirical distribution lead to highly agreeing results (compare columns 3-5 with columns 6-8 in Tables 4.3, 4.2 and 4.1; compare Figures 4.3, 4.5, 4.7 with Figures 4.4, 4.6, 4.8). This suggests that if the approximative distributions (the compound Poisson models or the binomial model) represent the empirical distribution accurately for the entire range of motif hit counts, it also represents the 5% significance level accurately which justifies *motif hit enrichment* testing.

In summary, our results suggests that the novel compound Poisson model $P^N_{CP}(X)$ accurately captures the motif hit counts distribution for a broad range of different motif structures and parameters as long as the significance level α for the *motif score test* is chosen stringently enough.



(a) $\alpha = 10^{-2}$, $d = 1$

(b) $\alpha = 10^{-3}$, $d = 0$

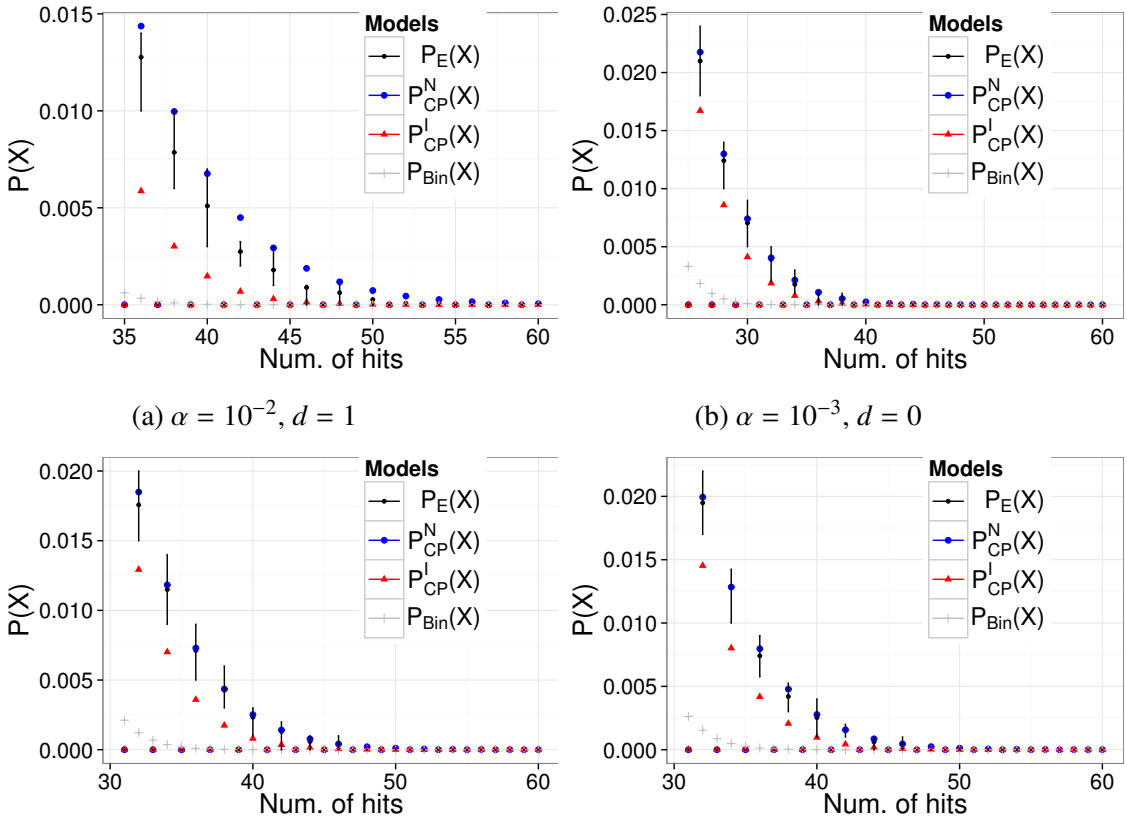
(c) $\alpha = 10^{-3}$, $d = 1$

(d) $\alpha = 10^{-3}$, $d = 2$

Figure 4.3: Motif hits count distribution for the palindromic motif (Figure 2.5a): (Black) Empirical distribution with error bars determined over 100 sample batches, (blue) novel compound Poisson distribution, (red) improved compound Poisson distribution, and (gray) binomial distribution.

4.3.2 Performance comparison on all Transfac motifs

In this Section, we elucidate the accuracy of $P^N_{CP}(X)$, $P^I_{CP}(X)$ and $P_{Bin}(X)$ on all Transfac motifs [58] that are at least 6 bps in length (of which there are in total 1015 motifs). The goal of this analysis is to estimate the practical relevance of the gain in accuracy of $P^N_{CP}(X)$ over $P^I_{CP}(X)$ and $P_{Bin}(X)$. We used a common order-1 background model that we estimated from ENCODE Dnase-I-hypersensitive sites [50] and determined the



(a) $\alpha = 10^{-2}$, $d = 1$

(b) $\alpha = 10^{-3}$, $d = 0$

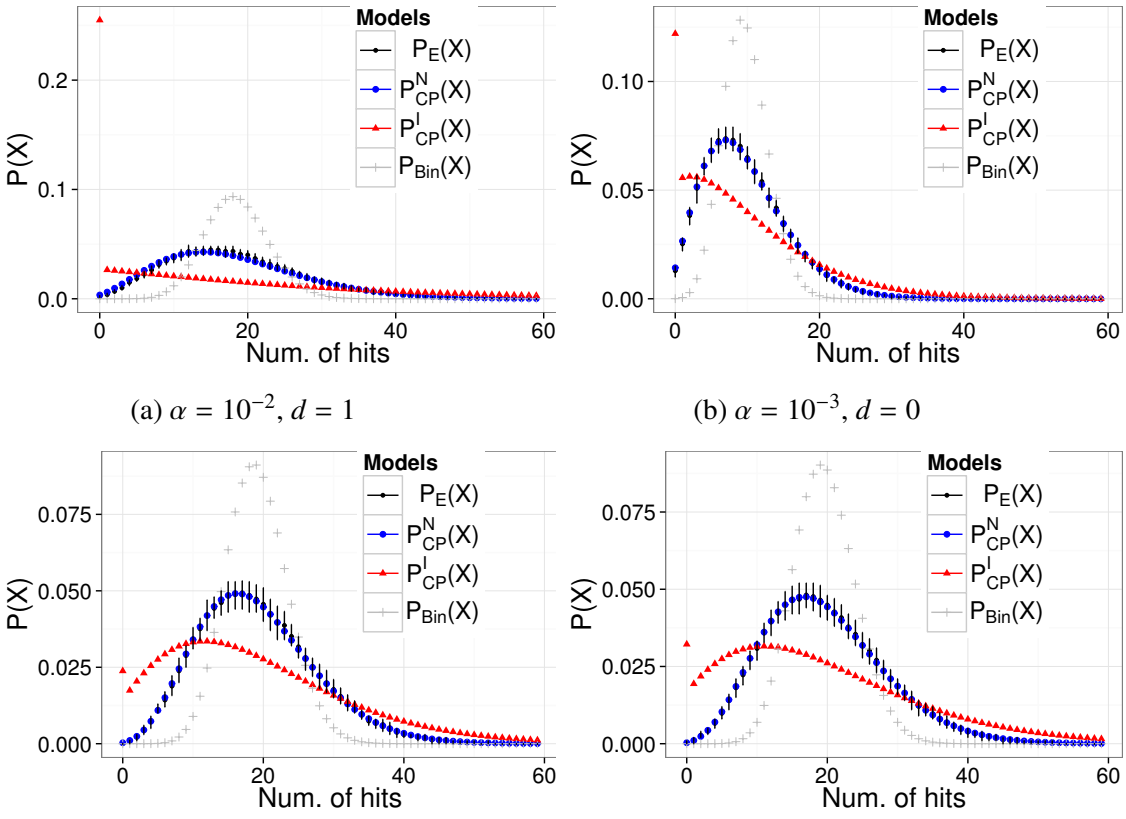
(c) $\alpha = 10^{-3}$, $d = 1$

(d) $\alpha = 10^{-3}$, $d = 2$

Figure 4.4: Motif hits count distribution for the palindromic motif (Figure 2.5a) shown for the 5% significance level: (Black) Empirical distribution with error bars determined over 100 sample batches, (blue) novel compound Poisson distribution, (red) improved compound Poisson distribution, and (gray) binomial distribution

motif hits count distribution for sequences of length 10 kb and $\alpha = 10^{-3}$, for all motifs. The empirical distribution $P_E(X)$ was generated by sampling 500 random sequences per motif and counting the number of motif hit occurrences. We tested the accuracy for all motifs using the total variation distance, defined by (4.35).

Across all motifs, both compound Poisson approximations $P^I_{CP}(X)$ and $P^N_{CP}(X)$ led to similar performances compared with $P_E(X)$ (see Figure 4.9a and 4.9b). By contrast, as expected, for the binomial model, there appeared to be numerous motifs for which the



(a) $\alpha = 10^{-2}$, $d = 1$

(b) $\alpha = 10^{-3}$, $d = 0$

(c) $\alpha = 10^{-3}$, $d = 1$

(d) $\alpha = 10^{-3}$, $d = 2$

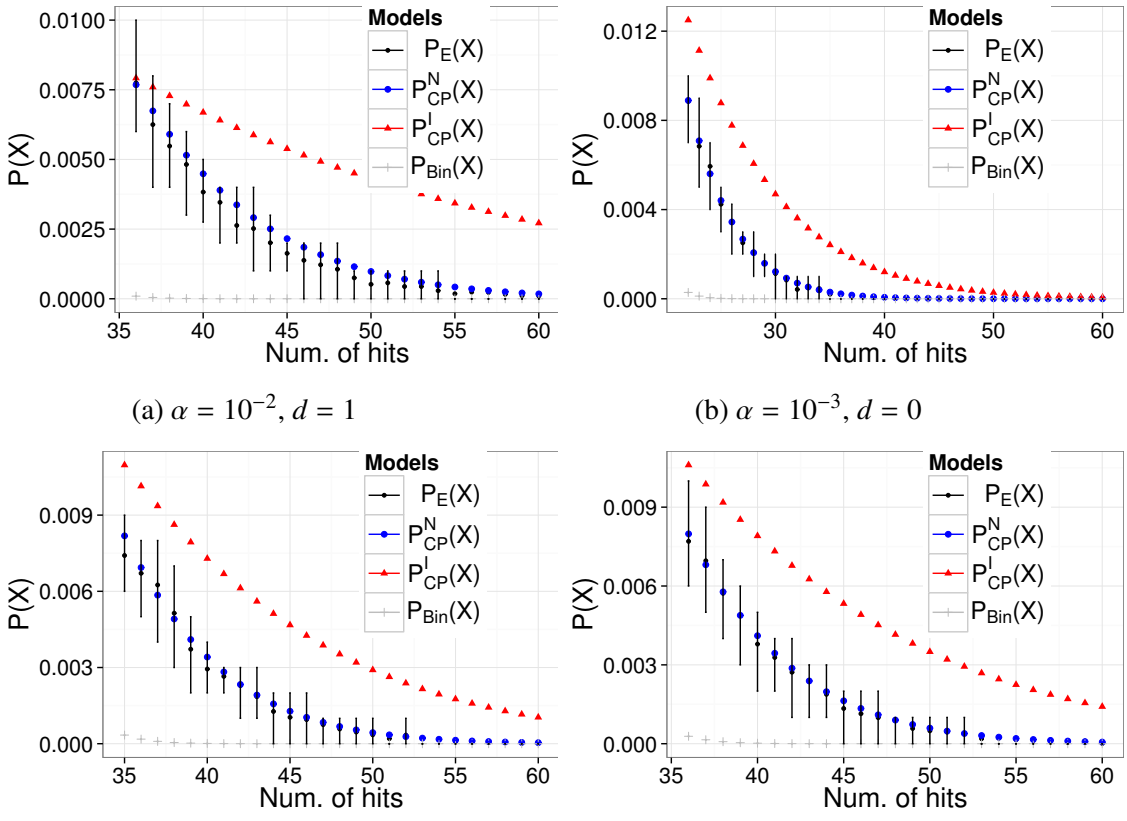
Figure 4.5: Motif hits count distribution for the repetitive motif (Figure 2.5b): (Black) Empirical distribution with error bars determined over 100 sample batches, (blue) novel compound Poisson distribution, (red) improved compound Poisson distribution, and (gray) binomial distribution.

distance to the empirical distribution was rather large. This can be appreciated by the heavier tail of the histogram shown in Figure 4.9c.

To further test how well the accuracy of $P^N_{CP}(X)$ compares to the accuracy of $P^I_{CP}(X)$ and $P_{Bin}(X)$, we contrasted the total variation distance according to

$$d(P^N_{CP}, P_E) - d(P^I_{CP}, P_E) \tag{4.37}$$

$$d(P^N_{CP}, P_E) - d(P_{Bin}, P_E). \tag{4.38}$$



(a) $\alpha = 10^{-2}$, $d = 1$

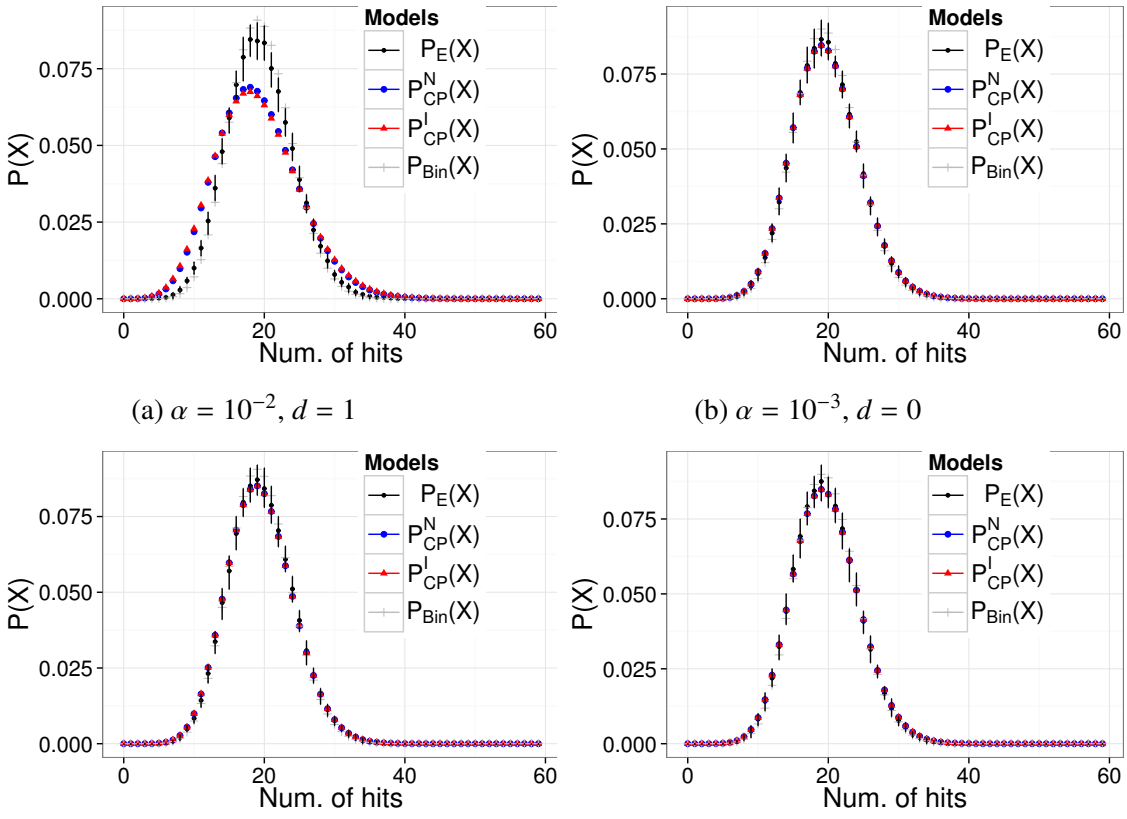
(b) $\alpha = 10^{-3}$, $d = 0$

(c) $\alpha = 10^{-3}$, $d = 1$

(d) $\alpha = 10^{-3}$, $d = 2$

Figure 4.6: Motif hits count distribution for the repetitive motif (Figure 2.5b) shown for the 5% significance level: (Black) Empirical distribution with error bars determined over 100 sample batches, (blue) novel compound Poisson distribution, (red) improved compound Poisson distribution, and (gray) binomial distribution.

Thus, negative values of (4.37) and (4.38) indicate that $P^N_{CP}(X)$ achieves more accurate results compared to $P^I_{CP}(X)$ and $P_{Bin}(X)$, respectively, while positive values indicate the opposite. Even though, according to the histograms both compound Poisson models achieve comparable results (see Figure 4.9a and 4.9b), using the measure defined by (4.37), we found that $P^N_{CP}(X)$ yields significantly more accurate results compared to $P^I_{CP}(X)$ across all Transfac motifs (Wilcoxon rank sum test: P-value=1.414×10^{-06}). In line with our observation from Figure 4.9a and 4.9c, we found that $P^N_{CP}(X)$ achieves



(a) $\alpha = 10^{-2}$, $d = 1$

(b) $\alpha = 10^{-3}$, $d = 0$

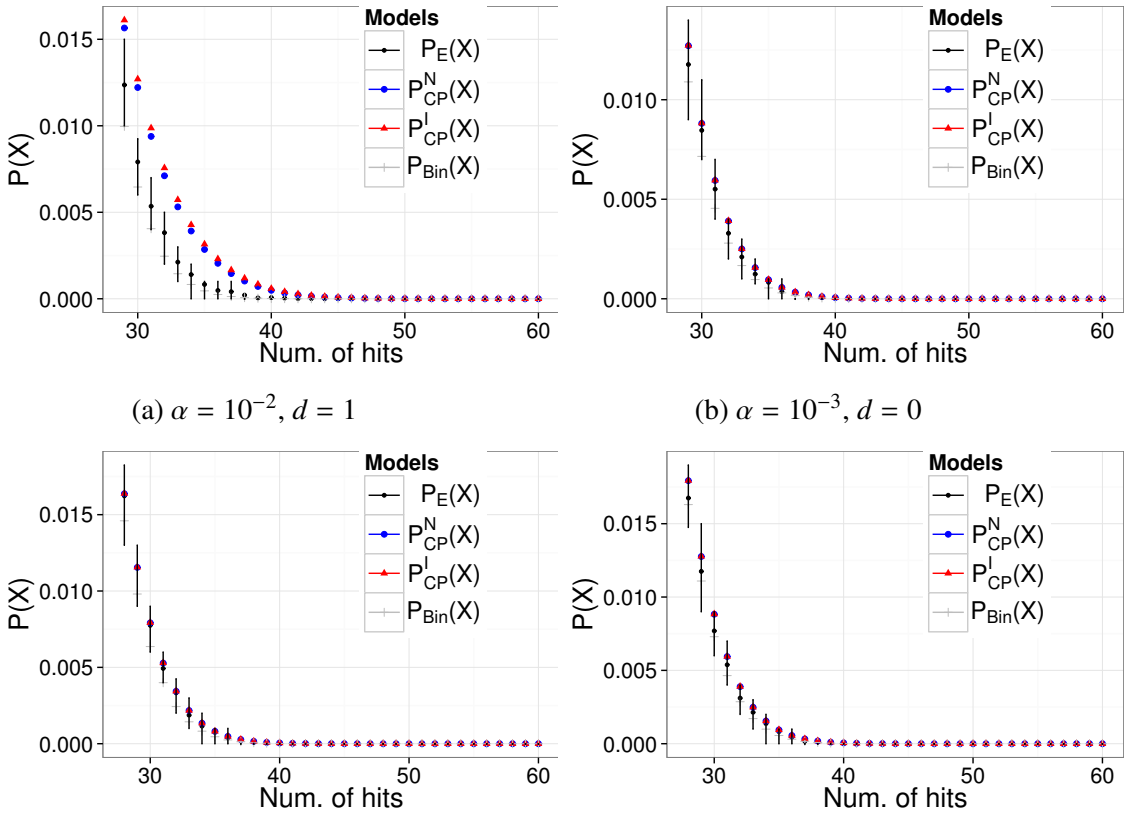
(c) $\alpha = 10^{-3}$, $d = 1$

(d) $\alpha = 10^{-3}$, $d = 2$

Figure 4.7: Motif hits count distribution for *E47* (Figure 2.5c): (Black) Empirical distribution with error bars determined over 100 sample batches, (blue) novel compound Poisson distribution, (red) improved compound Poisson distribution, and (gray) binomial distribution.

significantly more accurate results compared to $P_{Bin}(X)$ (Wilcoxon rank sum test: P-value$< 2.2 \times 10^{-16}$). This shows that in general $P^N_{CP}(X)$ yields more accurate results compared to $P^I_{CP}(X)$ and $P_{Bin}(X)$ across a large compendium of known transcription factor motifs.

Examples of Transfac motifs for which the discrepancy according to (4.37) and (4.38) are particularly high are depicted in Figures 4.10 and 4.11. Note that all of them represent instances of self-overlapping TF motifs.



(a) $\alpha = 10^{-2}$, $d = 1$

(b) $\alpha = 10^{-3}$, $d = 0$

(c) $\alpha = 10^{-3}$, $d = 1$

(d) $\alpha = 10^{-3}$, $d = 2$

Figure 4.8: Motif hits count distribution for *E47* (Figure 2.5c) shown for the 5% significance level: (Black) Empirical distribution with error bars determined over 100 sample batches, (blue) novel compound Poisson distribution, (red) improved compound Poisson distribution, and (gray) binomial distribution.

4.3.3 Influence of different choices of background models

In this Section, we emulate the effect of assuming a background model of inappropriate order to draw statistical conclusions for calling motif hits and about motif hit enrichment in a case study for human CpG islands and the SP1SP3 motif from Transfac (see Figure 2.5d). In particular, we shall assess how the distribution of the number of motif hits as well as the score distribution are affected by assuming an order-0 background

Table 4.2: Accuracy of the analytic models for the repetitive motif (Figure 2.5b) measured by Equation (4.35) - (4.36). The values marked in bold underline the most accurate analytic model (P^N_{CP}, P^I_{CP} or P_{Bin}) compared to $P_E(X)$.

d	α	$d(P_E, P^N_{CP})$	$d(P_E, P^I_{CP})$	$d(P_E, P_{Bin})$	$d_{5\%}(P_E, P^N_{CP})$	$d_{5\%}(P_E, P^I_{CP})$	$d_{5\%}(P_E, P_{Bin})$
1	10^{-2}	**0.0768**	0.846	0.731	**0.00817**	0.0729	0.0562
0	10^{-3}	**0.0186**	0.469	0.597	**0.00241**	0.0681	0.0492
1	10^{-3}	**0.0209**	0.46	0.616	**0.00439**	0.0736	0.0536
2	10^{-3}	**0.0184**	0.498	0.629	**0.00291**	0.0742	0.0552

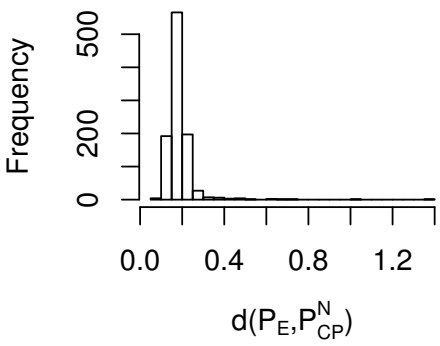
Table 4.3: Accuracy of the analytic models for the E47 motif (Figure 2.5c) measured by Equation (4.35) - (4.36). The values marked in bold underline the most accurate analytic model (P^N_{CP}, P^I_{CP} or P_{Bin}) compared to $P_E(X)$.

d	α	$d(P_E, P^N_{CP})$	$d(P_E, P^I_{CP})$	$d(P_E, P_{Bin})$	$d_{5\%}(P_E, P^N_{CP})$	$d_{5\%}(P_E, P^I_{CP})$	$d_{5\%}(P_E, P_{Bin})$
1	10^{-2}	0.228	0.249	**0.0787**	0.0307	0.0351	**0.0112**
0	10^{-3}	**0.026**	0.0265	0.0422	**0.00384**	0.00394	0.0065
1	10^{-3}	**0.0331**	0.0336	0.044	**0.00178**	0.00187	0.00938
2	10^{-3}	**0.0244**	0.0248	0.0458	0.00596	0.00605	**0.00495**

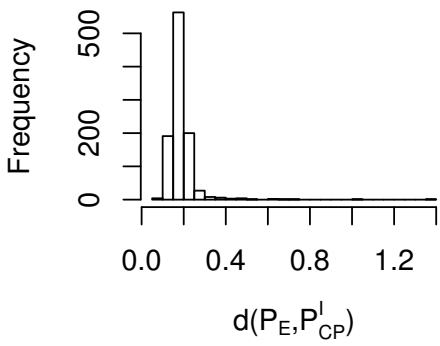
model, even though, the true underlying generative process for the DNA might be more complex.

We first investigate the effect of the background model choice on the score distribution. We model the DNA sequences obtained from human CpG islands using background models of various orders $d \in \{0, 1, 2\}$ and determine the score distribution in conjunction with the SP1SP3 motif from Transfac (see Figure 2.5d) according to Chapter 2. We denote the score distribution by $P_d(S)$ (to indicate the background model order).

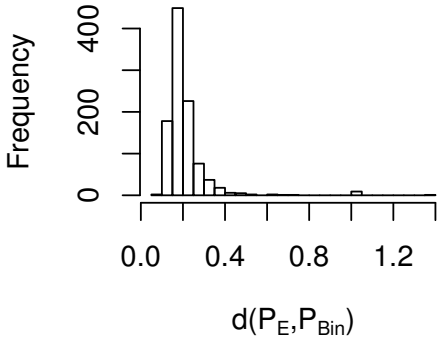
For the comparison, we determine the empirical score distribution by generating DNA sequences of length 10Mb from an order-d background model with $d \in \{0, 1, 2\}$, but where we nevertheless assume an order-0 background model to evaluate the log-likelihood ratio in all cases. We denote the empirical score distribution by $P^*_{E,d}(S)$. We shall measure the discrepancy between $P_0(S)$ and $P^*_{E,d}(S)$ for different d. Note that as the log-



(a)



(b)

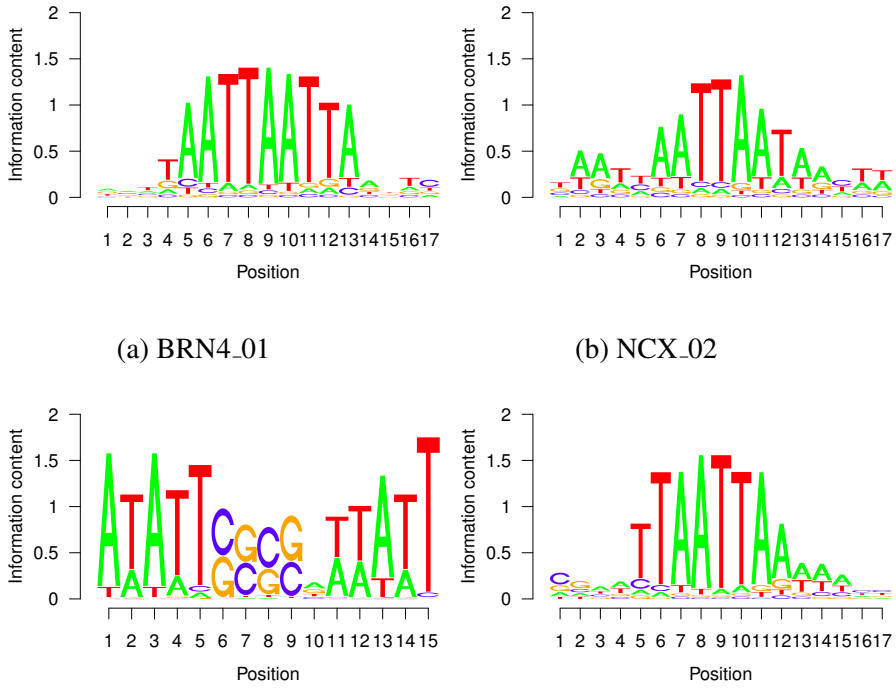


(c)

Figure 4.9: Distribution of distances according to Definition (4.35) between the analytic models and the empirical model.

likelihood ratio was determined in both cases assuming an order-0 model, the discrepany between $P_0(S)$ and $P^*_{E,d}(S)$ is directly attributable to the differences in the underlying generative processes for the DNA.

As expected, $P_0(S)$ and $P^*_{E,0}(S)$ are highly concordant, because the underlying gener-



(a) BRN4_01

(b) NCX_02

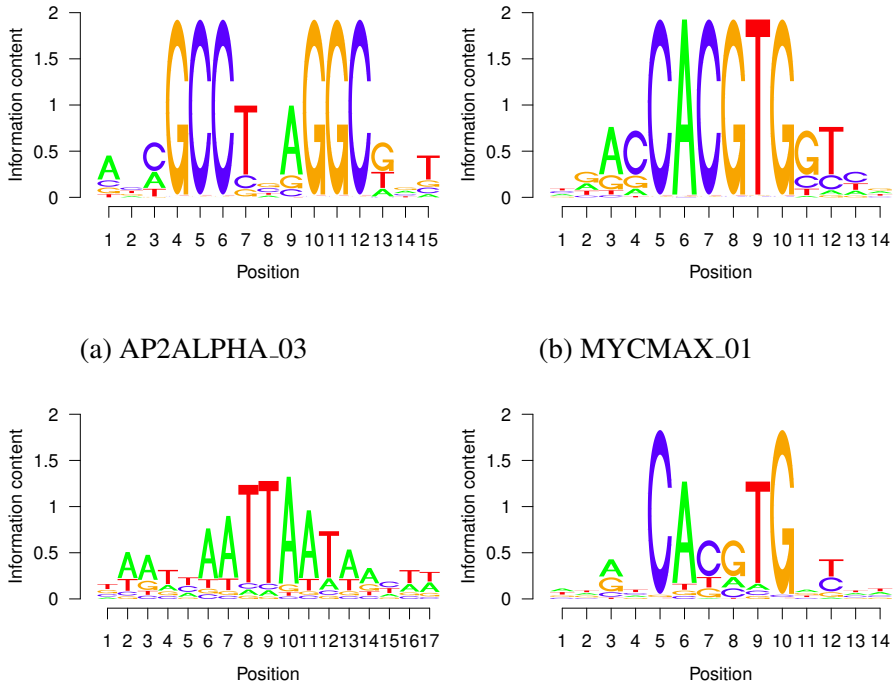
(c) HMGA2_01

(d) LMX1_01

Figure 4.10: Transfac motifs for which P^N_{CP} and P^I_{CP} disagree.

ative process is based on the same background model order (see Figure 4.12a). By contrast, if different background model orders are contrasted $P_0(S)$ and $P^*_{E,d}(S)$ such that $d > 0$, the score distributions become increasingly different (see Figures 4.12b - 4.12c). While, higher-order sequence features influence $P^*_{E,d}(S)$ through the generative process, such higher-order sequence features are ignored for $P_0(S)$.

To demonstrate that higher-order background models can take the sequence features of CpG islands better into account we next determined an empirical score distribution by generating DNA sequence of length 10Mb for which the log-likelihood ratio and the generative process are determined by the same order-d background model. This variant of the empirical distribution is denoted by $P_{E,d}(S)$ for $d \in \{1, 2\}$, In this case, the differ-



(a) AP2ALPHA_03

(b) MYCMAX_01

(c) NCX_02

(d) USF_02

Figure 4.11: Transfac motifs for which P^N_{CP} and P_{Bin} disagree.

ences between $P_d(S)$ and $P_{E,d}(S)$ for $d \in \{1, 2\}$ vanishes, because in both distributions take higher-order sequence features properly into account (see Figure 4.13).

To quantify the influence of assuming an inappropriate background model order on the effective false positive rate for calling motif hits, we measure the relative increase in the false positive rate (IFPR) between $P_0(S)$ and $P^*_{E,d}(S)$ for $d \in \{0, 1, 2\}$ according to

$$IFPR(t_s) = \frac{\sum_{s \geq t_s} P^*_{E,d}(s)}{\sum_{s \geq t_s} P_0(s)}. \tag{4.39}$$

For a desired significance level $\alpha = 0.01$, we found that ignoring the dinucleotide frequency in the background model increases the chance of obtaining a false positive motif

Table 4.4: Relative increase in false positive motif hits at $\alpha = 0.01$ (see Definition (4.39)) due to the use of inadequate background models.

d	$IFPR(0.01)$
0	0.99
1	1.26
2	1.297

hit effectively by 26%, whereas, ignoring trinucleotides increases the chance to obtain a false positive hit only slightly further to 29.7% (see Table 4.4). On the other hand, the slight decrease by 1% for the comparison of $P_{E,0}(S)$ with $P_0(S)$ can be attributed to the sampling noise.

Next, we study the influence of the background model choice on the distribution of the number of motif hits. To this end, we determine the compound Poisson distribution as described in Chapter 4 using different background model orders $d \in \{0, 1, 2\}$ along with a sequence of length $N = 10000$ kb and a significance level for calling motif hits of $\alpha = 0.01$, which we denote by $P^N_{CP,d}(X)$. For the comparison, we estimate the empirical distribution of the number of motif hits by sampling 10000 sequence from an order-d background model with $d \in \{0, 1, 2\}$ where we nevertheless assume an order-0 background model to compute log-likelihood ratio. The empirical distribution is denoted by $P^*_{E,d}(X)$. As expected, comparing $P^N_{CP,0}(X)$ and $P^*_{E,0}(X)$ shows that the two distributions are highly similar to one another, because, both distributions were generated by an order-0 background (see Figure 4.14a). However, the comparison of $P^N_{CP,0}(X)$ and $P^*_{E,d}(X)$ with $d > 0$ reveals that $P^N_{CP,0}(X)$ is shifted to the left with respect to $P^*_{E,d}(X)$, which signifies an excess of the effective false positive rate (see Figure 4.14b and 4.14c). This excess of false positives is caused by the systematic underestimation of word frequencies that give rise to motif hits due to assuming an order-0 background model instead of higher-order

models.

On the other hand, if the background model properly accounts for higher-order sequence features, those biases are expected to vanish. To show this, we estimated the empirical background model by sampling 10000 DNA sequences of length 10 kb each, denoted by $P_{E,d}(X)$ for $d > 0$, where for the generative model as well as for the log-likelihood ratio an order-d background model was assumed. Comparing $P^{N}_{CP,d}(X)$ and $P_{E,d}(X)$ for $d > 0$ shows that both distributions are highly similar (see Figure 4.15).

To summarize, this section demonstrates that the use of higher-order background models better accounts for the sequence complexity of natural sequences (e.g. CpG islands). This might reduce an excess of the false positive rate beyond the desired false positive rate α and, subsequently, more accurate *motif hit enrichment* statistics.

4.4 Discussion

In this Chapter, we have discussed the compound Poisson model for modeling the motif hits count distribution by accounting for motif hits on both strands of the DNA. The compound Poisson model was originally adopted to model the word count distribution of single words or sets of words [56, 41]. It was first adopted to approximating the distribution of the number of motif hits on both strands of the DNA in the context of motif hits that are called using the log-likelihood ratio and with a fixed score threshold by Pape *et al.* [36]. As part of this Chapter, we have refined the compound Poisson model that was suggested earlier. As a result, we presented two variants for approximating the clump size distribution: The first variant can be viewed as an *improvement* upon Pape *et al.* [36]. This variant directly employs the *marginal overlapping hit prob-*

abilities to approximate clump size distribution based on the assumption that the events within a clump are mutually independent of one another as proposed earlier [36] (see Equation (4.11) and (4.12)). We refined the originally proposed procedure to approximate the clump size distribution in two ways: 1) We utilized *marginal overlapping hit probabilities* that were derived based on a general order-d background model, instead of the previously proposed overlapping hit probabilities that are based on an order-0 background with matched AT and CG frequencies (also referred to as GC-background model). 2) The original compound Poisson model utilizes only two types of overlapping hit probabilities, namely overlapping hits on the same strand and 3'-strand overlapping hits on complementary strands. Those were originally denoted by γ_k and γ'_k and correspond to the *marginal overlapping hit probabilities* γ_k and $\gamma_{3',k}$ in this thesis, respectively (see Chapter 3). However, we advocate the use of three types of overlapping hit probabilities (γ_k, $\gamma_{3',k}$ and $\gamma_{5',k}$), because $\gamma_{3',k}$ and $\gamma_{5',k}$ are not symmetric in general. Apart from that, the derivation of the clump size probabilities is similar as described in Pape *et al.* [36]. The main purpose of this variant shall be to allow for a fair comparison with a further advanced variant of the compound Poisson approximation.

The second variant is facilitated by the newly derived *principal overlapping hit probabilities* instead of the *marginal overlapping hit probabilities*. We refer to this variant as the *novel* variant, because, to our knowledge, the probabilities associated with overlapping hits at *principal periods* was not discussed previously for motif hits that are called based on the log-likelihood ratio in conjunction with a fixed score threshold t_s. As advocated in Reinert *et al* [41], principal periods ought to be used over mere periods to study the clumping statistics, because periods might contain redundant information about overlapping hits. Therefore, using the overlapping hit probabilities at periods di-

rectly (which correspond to the *marginal overlapping hit probabilities*) which result in a biased clump size distribution. This is especially the case for repetitive motif structures.

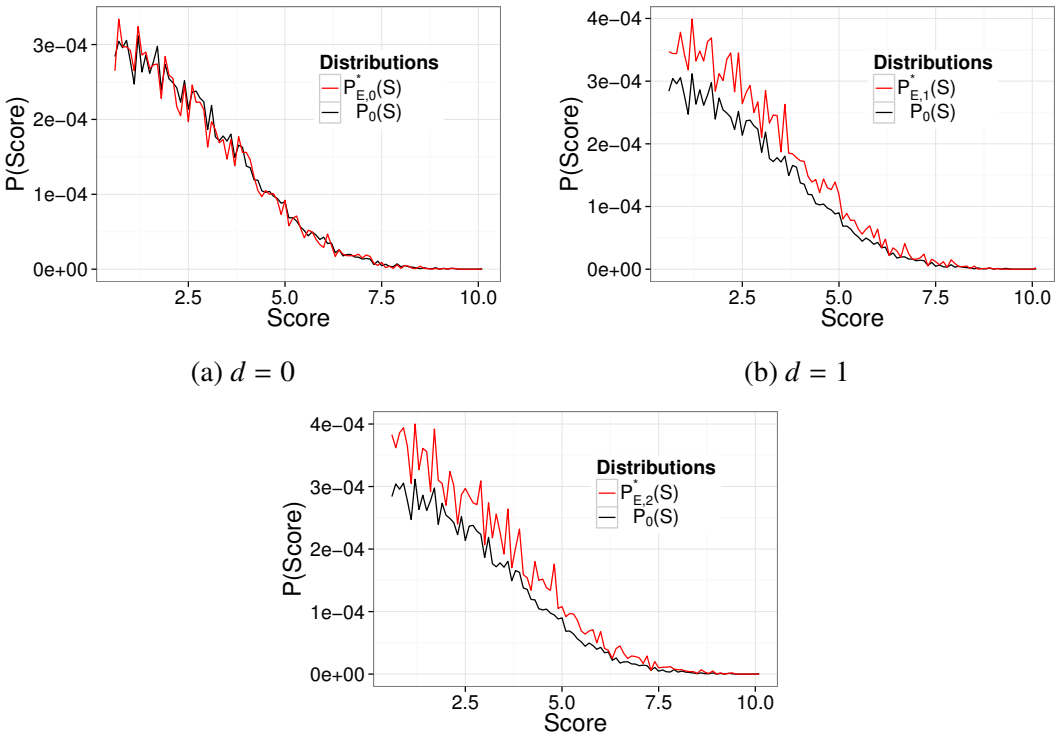
We notice that Marschall, 2011 [32] proposed a similar approach for deriving the expected clump size for the context of modeling the number of exact string matches for generalized strings. However, the approximation of the clump size distribution using the *principal overlapping hit probabilities* in the context of motif hits that are identified by the log-likelihood ratio with a given score threshold is, to our knowledge, novel and offers the advantage of being applicable even if the set of compatible words is very large, because the algorithms discussed in this thesis are independent of size of the compatible set $C(t_s)$.

We have systematically compared the two variants of the compound Poisson model and the binomial model for describing the distribution of number of motif hits. We found that the novel compound Poisson model performed advantageous compared to the other models for a wide range of parameters and across different motif structures. Even though, the improved and the novel compound Poisson variant yield similar results on non-self-overlapping and palindromic motifs, for repetitive motifs, the previous compound Poisson variant is less accurate compared to the novel one due to use of the *marginal overlapping hit probabilities* instead of the *principal overlapping hit probabilities*.

Furthermore, to assess the practical relevance and the accuracy of the models, we compared the compound Poisson model variants and the binomial model for all Transfac motifs by using a common background model and a common significance level. This analysis revealed that even though both compound Poisson models yield similar results across known motifs, the novel compound Poisson models achieves significantly more

accurate results compared to the binomial model and the improved compound Poisson model. The differences between the novel compound Poisson model and the other models are particularly pronounced for self-overlapping motif structures.

Finally, we investigated the effect of using an order-0 background model to identify motif hits and for studying the distribution of the number of motif hits when the actual driving generative model for the DNA is given by a general order-d Markov model with $d \in \{0, 1, 2\}$. This setup emulates the effect of ignoring higher-order sequence features, which however, are present in naturally occurring DNA sequences, for identifying motif hits and the motif hit counts. To that end, we studied human CpG islands that are scanned for the SP1SP3 motif [58]. We found that assuming an order-0 background model leads to substantially higher false positive rates of motif hits and in turn an excess of false positives for the motif hit enrichment test if the true generative process is driven by a higher-order background model. Accounting for the sequence complexity by utilizing an order-d background model offers the potential to reduce an increase in false positives and a reduced risk of statistical misinterpretations.

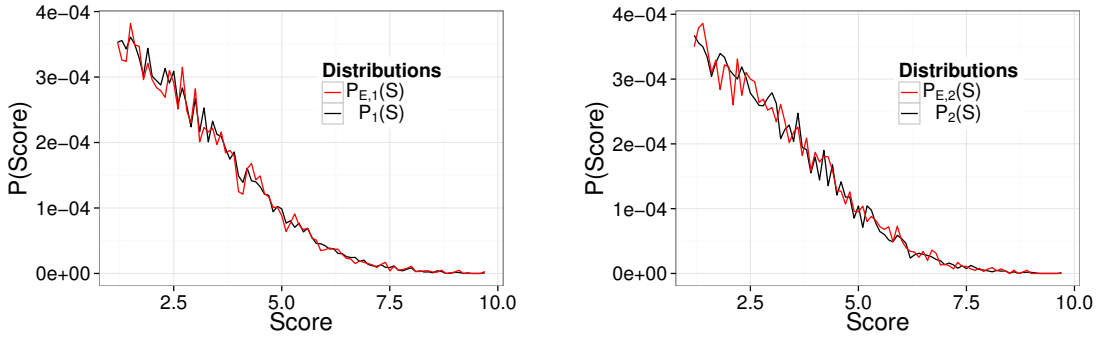


(a) $d = 0$

(b) $d = 1$

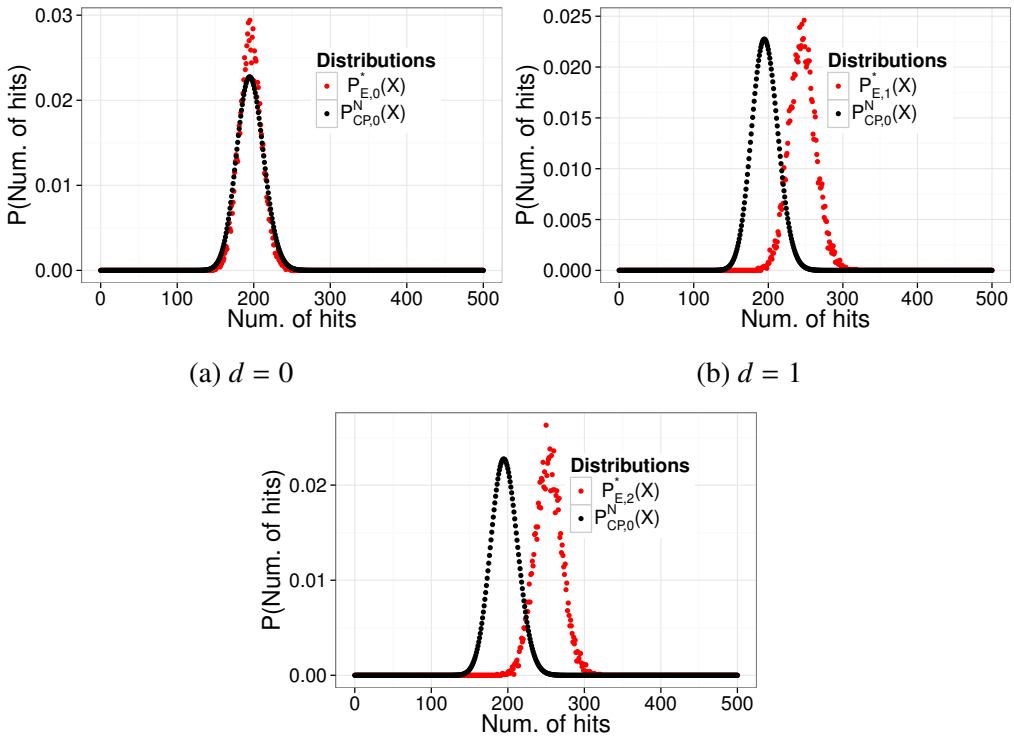
(c) $d = 2$

Figure 4.12: Score distribution for SP1SP3 in human CpG islands: We determined $P_0(S)$ (black) according to an order-0 background and the empirical distribution $P^*_{E,d}(S)$ (red) with $d \in \{0, 1, 2\}$ (shown in Panel a, b and c). As explained in the text, the log-likelihood ratio value for computing $P^*_{E,d}(S)$ was based on an order-0 background to facilitate comparability. With increasing discrepancy between the background model orders for generating the underlying DNA sequence, the score distributions become increasingly discordant.



(a) $d = 1$ (b) $d = 2$

Figure 4.13: Score distribution for SP1SP3 in human CpG islands: We determined $P_d(S)$ (black) according to an order-d background and the empirical distribution $P_{E,d}(S)$ (red) for $d \in \{1, 2\}$ (shown in Panel a and b). As the models are based on the same higher-order background models, the discrepancy between the distributions vanishes.

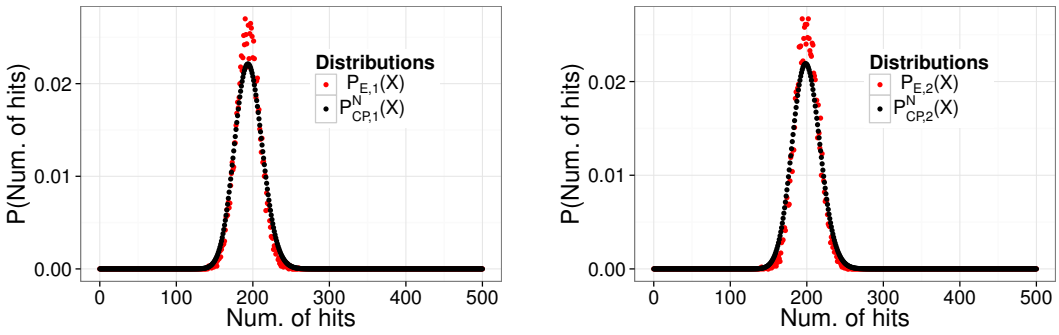


(a) $d = 0$

(b) $d = 1$

(c) $d = 2$

Figure 4.14: Motif hits count distribution for SP1SP3 in human CpG islands for hypothetical sequences of length 10kb and $\alpha = 0.01$. We determined the compound Poisson distribution $P^N_{CP,0}(X)$ (black) according to an order-0 background and the empirical motif hit counts distribution $P^*_{E,d}(X)$ (red) with $d \in \{0, 1, 2\}$ (shown in Panel a, b and c). As explained in the text, the log-likelihood ratios for computing $P^*_{E,d}(X)$ are based on an order-0 background to facilitate comparability (see text). With increasing discrepancy between the background model orders for generating the underlying DNA sequence, the motif hit counts distributions become increasingly discordant. As a consequence, compound Poisson distribution systematically underestimates the number of motif hits for $d > 0$.



(a) $d = 1$ (b) $d = 2$

Figure 4.15: Motif counts distribution for SP1SP3 in human CpG islands for hypothetical sequences of length 10kb and $\alpha = 0.01$. Each panel shows the empirical distribution $P_{E,d}(X)$ (red) and the compound Poisson distribution $P^N_{CP}(X)$ (black) with matched background model orders $d \in \{1, 2\}$. Using the matched background model order for the analytically as well as for the empirically derived distribution reduces the discrepancy between the distributions.

Chapter 5

Declumping - a Markov model approach

While in Chapter 3, we have dealt with the notion of overlapping motif hits, in this chapter, we study motif hits that are not preceded by an overlapping hit. Such a type of hit is associated with the start of a clump and, thus, we refer to its respective probability as *clump start probability* which is defined by

$$P(Y_0 = 1 | \mathbf{Y}_{-1:-M+1} = \mathbf{0}) \tag{5.1}$$

where $\mathbf{Y}_{-1:-M+1} = \mathbf{0}$ indicates that at position $-M+1, \cdots, -2, -1$ are all *zeros* . The derivation of *clump start probability* is motivated by the combinatorial model that we shall discuss in Chapter 6. Therefore, the main result of this chapter, a Markov model that is employed to determine the *clump start probability*, can be viewed as an auxiliary model which facilitates the combinatorial model.

We motivate the following discussion by reiterating the declumping approach given in Reinert, 2000 [41]. Suppose we scan a single DNA strand for motif hits with a motif of length M. Recall from the Section 1.2.2 that the probability of obtaining a motif hit equals $P(Y = 1) = \alpha$ because the underlying DNA sequence was generated from a homogeneous order-d Markov model which starts in the stationary distribution. The probability $P(Y = 1) = \alpha$ can be viewed as a mixture probability of observing 1) clump start hits and 2) overlapping hits. Accordingly, we have

$$\underbrace{P(Y_0 = 1)}_{\text{Probability of a hit} = \alpha} = \sum_{y_{-1}\cdots y_{-M+1} \in \{0,1\}^{M-1}} P(Y_0 = 1, y_{-1}, \cdots y_{-M+1}) \tag{5.2}$$

$$= P(Y_0 = 1, 0, \cdots, 0) + \sum_{\substack{y_{-1}\cdots y_{-M+1} \in \{0,1\}^{M-1}: \\ \sum_{k=1}^{M-1} y_{-k} > 0}} P(Y_0 = 1, y_{-1}, \cdots y_{-M+1}) \tag{5.3}$$

$$= \underbrace{P(Y_0 = 1|0, \cdots, 0)}_{\text{Clump start probability}} \quad \underbrace{P(y_{-1} = 0, \cdots y_{-M+1} = 0)}_{\text{Probability of observing } M-1 \text{ consecutive non-hits}} \quad +$$

$$+ \underbrace{\sum_{\substack{y_{-1}\cdots y_{-M+1} \in \{0,1\}^{M-1}: \\ \sum_{k=1}^{M-1} y_{-k} > 0}} P(Y_0 = 1, y_{-1}, \cdots y_{-M+1})}_{Y_0 = 1 \text{ overlaps with at least one hit to the left}} \tag{5.4}$$

Reinert *et al.* [41] discussed a declumping strategy by means of which $P(Y_0 = 1, Y_{-1} = 0 \cdots Y_{-M+1} = 0)$ can be determined for exact word matches. A similar strategy can also be adopted to motif hits that were called based on the *motif score test* with a fixed score threshold (see Section 1.2.3). As we are, however, primarily interested in the *clump start probability* $P(Y_0 = 1|0, \cdots, 0)$, rather than the joint probability, we propose an alternative declumping approach that is based on studying a Markov model that utilizes a correspondence relationship between patterns of outcomes of the Bernoulli process $\mathbf{Y}_{[1:N-M+1]}$ and the state space of the Markov model. Analyzing the stationary distribu-

tion of the Markov model allows us in turn to identify the *clump start probability* using the constraint that the total probability of obtaining a motif hit has to obey $P(Y = 1) = \alpha$.

A similar idea of translating sequence patterns into states was also pursued by Marschall *et al.* [33] in *probabilistic arithmetic automata*, with the aim of studying the distribution of the number of occurrences of generalized strings in sequential data. However, the application of a Markov model to declump motif hits that are called by using the log-likelihood ratio measure and by incorporating the chosen significance level α is new, to the best of our knowledge.

In the following we shall first introduce the Markov model for describing to the Bernoulli process $\mathbf{Y}_{[1:N-M+1]}$ in the case of scanning a single DNA strand for motif hits. We describe how the *clump start probability* can be obtained from studying the stationary distribution of the model. Subsequently, we extend the Markov model to the case of scanning both strands of the DNA and again utilize it to extract the *clump start probability*.

5.1 A Markov model for generating $\mathbf{Y}_{[1:N-M+1]}$ by scanning a single DNA strand

In this Section, we introduce a Markov model for studying the Bernoulli process $\mathbf{Y}_{[1:N-M+1]}$ that results in scanning a single DNA strand. We shall first discuss the state space of the Markov model by introducing a mapping between patterns in $\mathbf{Y}_{[1:N-M+1]}$ and states of the Markov model. Subsequently, each realization of the Bernoulli process $\mathbf{Y}_{[1:N-M+1]}$ can be uniquely mapped to a corresponding state sequence. Afterwards, we discuss the as-

sociated state and transition probabilities. Finally, we analyze the stationary distribution of the model in order to identify the previously unknown *clump start probability*.

5.1.1 The semantics of the states

We define the states Z_i that can be emitted at position i by the Markov model via their correspondence relationship with patterns in the underlying Bernoulli process as

$$\begin{array}{rcccccccc}
h & \widehat{=} & (Y_{i-M+2} = \bullet, & Y_{i-M+3} = \bullet, & \cdots, & Y_{i-2} = \bullet, & Y_{i-1} = \bullet, & Y_i = 1) \\
n_1 & \widehat{=} & (Y_{i-M+2} = \bullet, & Y_{i-M+3} = \bullet, & \cdots, & Y_{i-2} = \bullet, & Y_{i-1} = 1, & Y_i = 0) \\
n_2 & \widehat{=} & (Y_{i-M+2} = \bullet, & Y_{i-M+3} = \bullet, & \cdots, & Y_{i-2} = 1, & Y_{i-1} = 0, & Y_i = 0) \\
\vdots & & \vdots & & & & & & \\
n_{M-3} & \widehat{=} & (Y_{i-M+2} = \bullet, & Y_{i-M+3} = 1, & \cdots, & Y_{i-2} = 0, & Y_{i-1} = 0, & Y_i = 0) \\
n_{M-2} & \widehat{=} & (Y_{i-M+2} = 1, & Y_{i-M+3} = 0, & \cdots, & Y_{i-2} = 0, & Y_{i-1} = 0, & Y_i = 0) \\
n & \widehat{=} & (Y_{i-M+2} = 0, & Y_{i-M+3} = 0, & \cdots, & Y_{i-2} = 0, & Y_{i-1} = 0, & Y_i = 0).
\end{array} \tag{5.5}$$

As usual, hits and non-hits are represented by the outcomes 1 and 0 of the Bernoulli process, while, the '$\bullet$' symbol represents that *any* outcome is possible (hit or non-hit). For example, we observe the 'hit' state $Z_i = h$ at position i if $Y_i = 1$, regardless of the outcomes $Y_{i-M+2} \cdots Y_{i-1}$ and we observe the state $Z_i = n$ if a stretch of $M-1$ consecutive zeros end at position i in the Bernoulli process.

The states map to patterns of outcomes in $\mathbf{Y}_{[1:N-M+1]}$ such that any realizations of $\mathbf{Y}_{[1:N-M+1]}$ can be uniquely expressed by a sequence of state $Z_1 Z_2 \cdots Z_{N-M+1}$.

To further illustrate the association between the states and the outcomes of the Bernoulli process, consider a motif of length $M = 5$ for which the following state events may

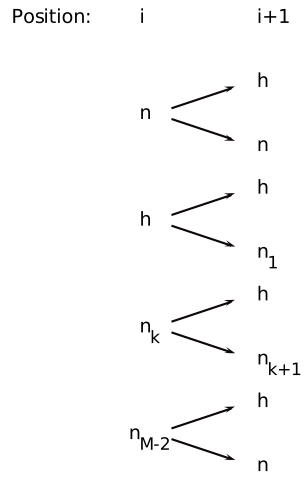


Figure 5.1: For any i, the possible state transitions from Z_i to Z_{i+1} are depicted for the case of scanning a single DNA strand for motif hits. For the third case from above, k ranges from $1 \leq k \leq M-3$.

occur: h, n_1, n_2, n_3 and n. Assuming that prior to the start of the Bernoulli process $Y_1 Y_2 \cdots$ only *zeros* were observed, the following realization of the Bernoulli process maps to the state sequence shown underneath

$$\begin{array}{lcccccccccccccc} Y_1Y_2Y_3\cdots = & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ Z_1Z_2Z_3\cdots = & n & n & n & h & h & n_1 & n_2 & h & n_1 & h & n_1 & n_2 & n_3 & n \end{array}$$

Finally, we shall stress that since the Bernoulli sequence is traversed with a step size of one, only a subset of states are reachable in one step from any given state. Figure 5.1 illustrates the feasible state transitions.

5.1.2 State and transition probabilities

Having defined the semantics of the states, we next introduce the associate state probabilities according to

$$P(Z_i = h) := P(Y_i = 1) \tag{5.6}$$

$$P(Z_i = n_k) := P(Y_i = 0, \cdots, Y_{i-k+1} = 0, Y_{i-k} = 1) \quad \text{for } 1 \leq k \leq M-2 \tag{5.7}$$

$$P(Z_i = n) := P(Y_i = 0, \cdots, Y_{i-M+2} = 0). \tag{5.8}$$

Since, the underlying DNA sequence was generated by a homogeneous order-d background model which starts in the stationary distribution, the probability to obtain a certain state is equivalent across all positions in the sequence. That is, $P(Z_0) = P(Z_i)$ for all i. Note that by definition, the probability of emitting the 'hit' state is given by $P(Z_0 = h) = P(Y = 1) = \alpha$. By contrast, the probabilities of observing $n, n_1 \cdots n_{M-2}$ are initially unknown.

According to the viable state transitions depicted in Figure 5.1, we next turn to defining the associated state transition probabilities $P(Z_{-1} \to Z_0)$ for observing a state Z_0 given that the previous state was Z_{-1} as follows

$$P(n \to h) := P(Y_0 = 1 | Y_{-1} = 0, \cdots Y_{-M+1} = 0) \tag{5.9}$$

$$P(n \to n) := 1 - P(n \to h) \tag{5.10}$$

$$P(h \to h) := P(Y_0 = 1 | Y_{-1} = 1) \tag{5.11}$$

$$P(h \to n_1) := 1 - P(h \to h) \tag{5.12}$$

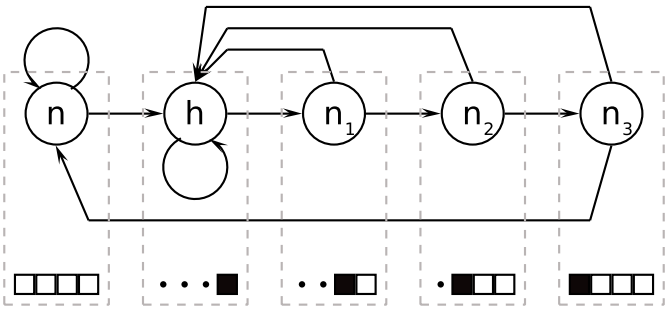


Figure 5.2: Illustration of the Markov model. The nodes denote the states of the model using a TF motif of length $M = 5$. Arrows indicate viable state transitions which may (but not necessarily) be associated with positive transition probabilities and assuming that the Bernoulli process is traversed in steps of size one (see transitions in Figure 5.1). Underneath each node, the associated pattern of the Bernoulli process is depicted where black and white boxes indicate ones and zeros in the process and the bullet denotes any outcome (one or zero) (see Correspondences (5.5)).

$$P(n_k \to h) := P(Y_0 = 1 | Y_{-1} = 0, \cdots, Y_{-k+1} = 0, Y_{-k} = 1) \qquad \text{for } 1 \leq k \leq M-2 \tag{5.13}$$

$$P(n_k \to n_{k+1}) := 1 - P(n_k \to h) \qquad \text{for } 1 \leq k \leq M-3 \tag{5.14}$$

$$P(n_{M-2} \to n) := 1 - P(n_{M-2} \to h) \tag{5.15}$$

An instance of the resulting Markov model in terms of a graphical representation is shown in Figure 5.2 for a TF motif of length $M = 5$ which consist of the states h, n, n_1, n_2 and n_3.

We proceed by discussing the individual transition probabilities, defined by (5.9), (5.11) and (5.13). The remaining transition probabilities are a consequence of those quantities. First, note that the transition of $Z_{-1} = n$ to $Z_0 = h$, defined by (5.9), corresponds to the definition of the *clump start probability*. This quantity is initially unknown. As we shall see below, the Markov model can be facilitated to identify the *clump start probability*.

Next, the probability that $Z_{-1} = h$ leads to $Z_0 = h$, defined by (5.11), is equivalent to

$$P(h \to h) = \beta_1 \tag{5.16}$$

where we made use of Definition (3.41). Finally, we derive the probability of $Z_0 = h$ given $Z_{-1} = n_k$, for $1 \leq k \leq M - 2$, defined by (5.13). Accordingly, we obtain

$$\begin{aligned} P(n_k \to h) &:= P(Y_0 = 1 | Y_{-1} = 0, \cdots, Y_{-k+1} = 0, Y_{-k} = 1) && (5.17) \\ &= P(Y_k = 1 | Y_{k-1} = 0, \cdots, Y_1 = 0, Y_0 = 1) && (5.18) \\ &= \frac{P(Y_k = 1, Y_{k-1} = 0, \cdots, Y_1 = 0 | Y_0 = 1)}{P(Y_{k-1} = 0, \cdots, Y_1 = 0 | Y_0 = 1)} && (5.19) \\ &= \frac{\beta_k}{\delta_{k-1}} && (5.20) \end{aligned}$$

where we made use of the *principal overlapping hit probabilities*, defined by (3.41), and the probability of a stretch of *zeros* that succeeds a hit, defined by (3.50). We shall assume that $\delta_{k-1} > 0$ always holds and therefore, the conditional probability is defined.

The transition probabilities can further be compactly represented in matrix form, which

gives rise to the transition matrix

$$M = \begin{bmatrix} P(h \to h) & P(n \to h) & P(n_1 \to h) & P(n_2 \to h) & \cdots & P(n_{M-2} \to h) \\ 0 & P(n \to n) & 0 & 0 & \cdots & P(n_{M-2} \to n) \\ P(h \to n_1) & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & P(n_1 \to n_2) & 0 & \cdots & 0 \\ 0 & 0 & 0 & P(n_2 \to n_3) & \cdots & 0 \\ \vdots & & & \ddots & \vdots & \\ 0 & 0 & 0 & 0 & \cdots & P(n_{M-3} \to n_{M-2}) \end{bmatrix} \tag{5.21}$$

We further assume $P(Z) > 0$ for all states which is established by choosing a significance level $0 < \alpha < 1$ such that $P(Z = n) > 0$ and $P(Z = h) > 0$. In that case, the Markov model consists of one recurrent class, which guarantees that there exists a unique stationary distribution over the state space, which is denoted by $\mu(Z)$ [26]. In general, one can then proceed by computing the the stationary distribution using eigenvalue decomposition of the transition matrix. The eigenvector associated with the largest eigenvalue (which is equal to one) represents the stationary distribution, for which we utilize the power method. However, since the unknown *clump start probability* $P(n \to h)$ parametrizes the transition matrix, we introduce the stationary distribution as a function of $P(n \to h)$. That is $\mu(Z, P(n \to h))$.

5.1.3 Identification of the clump start probability

In the following, we discuss the identification of the *clump start probability* $P(n \to h)$ which is based on two important insights: First, with the exception of $P(n \to h)$ and $P(n \to n)$, all other state transition probabilities are implicitly defined by the *principal overlapping hit probabilities* (see above). That is, the *clump start probability* is the only unknown parameter of the transition matrix. Second, we expect that in the stationary distribution, the 'hit' state $Z = h$ is observed with probability $\mu(h, P(n \to h)) = P(Y = 1) = \alpha$, which was prescribed as the significance level for calling motif hits. Therefore, our goal is to identify $P(n \to h)$ such that

$$\mu(h; P(n \to h)) = \alpha.$$

We proceed as follows: Starting from an initial guess about a *clump start probability* $P(n \to h)$ we measure the discrepancy between the desired probability of producing motif hits α and the model's current probability of producing motif hits $\mu(h; P(n \to h))$ by

$$(\alpha - \mu(h; P(n \to h)))^2$$

we refine the model such that the discrepancy measure is minimized, which leads us to the following optimization problem

$$\underset{P(n \to h)}{\arg\min} \quad (\alpha - \mu(h; P(n \to h)))^2.$$

To solve this problem, we employed a gradient descent approach where in each iteration we determine the numerical second-order derivative of the discrepancy measure

$$\frac{(\alpha - \mu(h; P(n \to h) + \epsilon))^2 - (\alpha - \mu(h; P(n \to h) - \epsilon))^2}{2\epsilon}$$

where ϵ is a small constant (e.g. $\epsilon = 10^{-8}$) and update $P(n \to h)$ such that the discrepancy is gradually decreased. The optimization routine is finished if the discrepancy measure falls below a certain tolerance threshold.

This procedure eventually yields the desired approximate *clump start probability* for the case of scanning a single DNA strand for motif hits.

5.2 A Markov model for generating $\mathbf{Y}_{[1:N-M+1]}$ by scanning both DNA strands

While, in the previous section, we have discussed the Markov model for identifying the *clump start probability* that is the result of scanning a single DNA strand for motif hits, in this section, we extend the Markov model to account for motif hits on both strands of a DNA sequence. As in the previous Section, we start by introducing a correspondence relationship between the states of the Markov model and patterns in the Bernoulli process $Y_1 Y'_1 Y_2 Y'_2 \cdots Y_{N-M+1} Y'_{N-M+1}$. The Markov model may then emit one of several states at each timepoint, denoted by Z, such that each realization of the Bernoulli process $Y_1 Y'_1 Y_2 Y'_2 \cdots$ uniquely maps to a state sequence $Z_1 Z_2 \cdots$. Subsequently, we discuss the association of the states with their respective probabilities as well as the transition probabilities of obtaining a state Z_0 given that the previous state was Z_{-1}.

Finally, we facilitate stationary distribution analysis of the Markov model in order to determine the *clump start probability*.

5.2.1 The semantics of the states

We introduce the correspondence relationship between the states of the Markov model which are denoted by Z and patterns that are emitted by the underlying Bernoulli process $Y_1 Y'_1 Y_2 Y'_2 Y_3 Y'_3 \cdots$ as

$$h_f \widehat{\equiv} \quad \begin{pmatrix} Y'_{-M+2} = \bullet, & Y'_{-M+3} = \bullet, & \cdots, & Y'_{-2} = \bullet, & Y'_{-1} = \bullet, & Y'_0 = \bullet \\ Y_{-M+2} = \bullet, & Y_{-M+3} = \bullet, & \cdots, & Y_{-2} = \bullet, & Y_{-1} = \bullet, & Y_0 = 1 \end{pmatrix} \tag{5.22}$$

$$h_r \widehat{\equiv} \quad \begin{pmatrix} Y'_{-M+2} = \bullet, & Y'_{-M+3} = \bullet, & \cdots, & Y'_{-2} = \bullet, & Y'_{-1} = \bullet, & Y'_0 = 1 \\ Y_{-M+2} = \bullet, & Y_{-M+3} = \bullet, & \cdots, & Y_{-2} = \bullet, & Y_{-1} = \bullet, & Y_0 = \bullet \end{pmatrix} \tag{5.23}$$

$$n_0 \widehat{\equiv} \quad \begin{pmatrix} Y'_{-M+2} = \bullet, & Y'_{-M+3} = \bullet, & \cdots, & Y'_{-2} = \bullet, & Y'_{-1} = \bullet, & Y'_0 = 0 \\ Y_{-M+2} = \bullet, & Y_{-M+3} = \bullet, & \cdots, & Y_{-2} = \bullet, & Y_{-1} = \bullet, & Y_0 = 1 \end{pmatrix} \tag{5.24}$$

$$n_1 \widehat{\equiv} \quad \begin{pmatrix} Y'_{-M+2} = \bullet, & Y'_{-M+3} = \bullet, & \cdots, & Y'_{-2} = \bullet, & Y'_{-1} = 0, & Y'_0 = 0 \\ Y_{-M+2} = \bullet, & Y_{-M+3} = \bullet, & \cdots, & Y_{-2} = \bullet, & Y_{-1} = 1, & Y_0 = 0 \end{pmatrix} \tag{5.25}$$

$$n_2 \widehat{\equiv} \quad \begin{pmatrix} Y'_{-M+2} = \bullet, & Y'_{-M+3} = \bullet, & \cdots, & Y'_{-2} = 0, & Y'_{-1} = 0, & Y'_0 = 0 \\ Y_{-M+2} = \bullet, & Y_{-M+3} = \bullet, & \cdots, & Y_{-2} = 1, & Y_{-1} = 0, & Y_0 = 0 \end{pmatrix} \tag{5.26}$$

$$\vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots$$

$$n_{M-3} \widehat{\equiv} \quad \begin{pmatrix} Y'_{-M+2} = \bullet, & Y'_{-M+3} = 0, & \cdots, & Y'_{-2} = 0, & Y'_{-1} = 0, & Y'_0 = 0 \\ Y_{-M+2} = \bullet, & Y_{-M+3} = 1, & \cdots, & Y_{-2} = 0, & Y_{-1} = 0, & Y_0 = 0 \end{pmatrix} \tag{5.27}$$

$$n_{M-2} \widehat{\equiv} \quad \begin{pmatrix} Y'_{-M+2} = 0, & Y'_{-M+3} = 0, & \cdots, & Y'_{-2} = 0, & Y'_{-1} = 0, & Y'_0 = 0 \\ Y_{-M+2} = 1, & Y_{-M+3} = 0, & \cdots, & Y_{-2} = 0, & Y_{-1} = 0, & Y_0 = 0 \end{pmatrix} \tag{5.28}$$

$$n'_1 \widehat{\equiv} \quad \begin{pmatrix} Y'_{-M+2} = \bullet, & Y'_{-M+3} = \bullet, & \cdots, & Y'_{-2} = \bullet, & Y'_{-1} = 1, & Y'_0 = 0 \\ Y_{-M+2} = \bullet, & Y_{-M+3} = \bullet, & \cdots, & Y_{-2} = \bullet, & Y_{-1} = \bullet, & Y_0 = 0 \end{pmatrix} \tag{5.29}$$

$$n'_2 \widehat{\equiv} \quad \begin{pmatrix} Y'_{-M+2} = \bullet, & Y'_{-M+3} = \bullet, & \cdots, & Y'_{-2} = 1, & Y'_{-1} = 0, & Y'_0 = 0 \\ Y_{-M+2} = \bullet, & Y_{-M+3} = \bullet, & \cdots, & Y_{-2} = \bullet, & Y_{-1} = 0, & Y_0 = 0 \end{pmatrix} \tag{5.30}$$

$$\vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots$$

$$n'_{M-3} \widehat{\equiv} \quad \begin{pmatrix} Y'_{-M+2} = \bullet, & Y'_{-M+3} = 1, & \cdots, & Y'_{-2} = 0, & Y'_{-1} = 0, & Y'_0 = 0 \\ Y_{-M+2} = \bullet, & Y_{-M+3} = \bullet, & \cdots, & Y_{-2} = 0, & Y_{-1} = 0, & Y_0 = 0 \end{pmatrix} \tag{5.31}$$

$$n'_{M-2} \widehat{\equiv} \quad \begin{pmatrix} Y'_{-M+2} = 1, & Y'_{-M+3} = 0, & \cdots, & Y'_{-2} = 0, & Y'_{-1} = 0, & Y'_0 = 0 \\ Y_{-M+2} = \bullet, & Y_{-M+3} = 0, & \cdots, & Y_{-2} = 0, & Y_{-1} = 0, & Y_0 = 0 \end{pmatrix} \tag{5.32}$$

$$n \widehat{\equiv} \quad \begin{pmatrix} Y'_{-M+2} = 0, & Y'_{-M+3} = 0, & \cdots, & Y'_{-2} = 0, & Y'_{-1} = 0, & Y'_0 = 0 \\ Y_{-M+2} = 0, & Y_{-M+3} = 0, & \cdots, & Y_{-2} = 0, & Y_{-1} = 0, & Y_0 = 0 \end{pmatrix} \tag{5.33}$$

Similar as before, hits and non-hits are represented by the outcomes 1 and 0 of the Bernoulli process, while, the '•' symbol represents *any* outcome (e.g. 0 or 1).

The main difference with respect to the previous section is that the pattern-to-state mapping now also respects the strandedness of hits. For the remainder of this discussion, we shall always traverse the realizations of the Bernoulli process in a specific order (namely from left to right and from forward to reverse strand) in order to generate a unique mapping between the realizations of $Y_1 Y'_1 Y_2 Y'_2 \cdots Y_{N-M+1} Y'_{N-M+1}$ and the state sequence $Z_1 Z_2 \cdots$.

Since, the Bernoulli process may exhibit two kinds of hits (forward and reverse) at each position of the underlying DNA sequence, each nucleotide position might be represented by one or two successive states in the state space. For example, observing a

palindromic hit is manifested by observing first the state $Z_{-1} = h_f$ and subsequently $Z_0 = h_r$. On the other hand, observing the state $Z_0 = n$ after $Z_{-1} = n$ corresponds to simultanously observing $Y = 0$ and $Y' = 0$. As a consequence, each realization of the Bernoulli process $Y_1 Y'_1 Y_2 Y'_2 \cdots Y_{N-M+1} Y'_{N-M+1}$ results in a state sequence $Z_1 Z_2 \cdots$ whose length depends on the particular realization of the Bernoulli process. That means that the state event Z_j with index j is not necessarily aligned with the outcome Y_j at position j of the Bernoulli process. Even though one could design the state correspondence relationship such that the indexes of the state sequence align with positions in the Bernoulli process, this is not relevant for our purpose, as we are only interested in the stationary distribution.

Moreover, the design of the Markov model was influenced by the design of the combinatorial model, since we seek to identify the *clump start probability* with the Markov model to facilitate the combinatorial model (see Chapter 6). In the combinatorial model, we shall traverse the realizations of the Bernoulli process in an analogous fashion.

To illustrate the association between the states $Z_1 Z_2 \cdots$ and the outcomes of the Bernoulli process, consider a motif of length $M = 5$ which consists of the following states: h_f, h_r, n_0, n_1, n_2, n_3, n'_1, n'_2, n'_3 and n. Assuming that prior to the start of the Bernoulli process $Y_1 Y'_1 Y_2 Y'_2 Y_3 Y'_3 \cdots$ only *zeros* were observed, the following realization of the Bernoulli process maps to the state sequence shown underneath

$$
\begin{aligned}
Y'_1 Y'_2 Y'_3 \cdots = &\quad 0 \;\; 0 \;\; 0 \;\; 1 \;\; 0 \;\; 0 \\
Y_1 Y_2 Y_3 \cdots = &\quad 0 \;\; 1 \;\; 0 \;\; 1 \;\; 0 \;\; 0 \\
Z_1 Z_2 Z_3 \cdots = &\quad n h_f n_0 n_1 h_f h_r n'_1 n'_2
\end{aligned}
$$

5.2.2 State probabilities and transition probabilities

Next, we introduce the associated state probabilities for the states defined by (5.22) as

$$P(Z = h_f) := P(Y_0 = 1) \tag{5.34}$$

$$P(Z = h_r) := P(Y'_0 = 1) \tag{5.35}$$

$$P(Z = n_k) := P\begin{pmatrix} Y'_{-k} = 0, & Y'_{-k+1} = 0, & \cdots, & Y'_{-1} = 0, & Y'_0 = 0 \\ Y_{-k} = 1, & Y_{-k+1} = 0, & \cdots, & Y_{-1} = 0, & Y_0 = 0 \end{pmatrix} \quad \text{for } 0 \le k \le M-2 \tag{5.36}$$

$$P(Z = n'_k) := P\begin{pmatrix} Y'_{-k} = 1, & Y'_{-k+1} = 0, & \cdots, & Y'_{-1} = 0, & Y'_0 = 0 \\ Y_{-k} = \bullet, & Y_{-k+1} = 0, & \cdots, & Y_{-1} = 0, & Y_0 = 0 \end{pmatrix} \quad \text{for } 1 \le k \le M-2 \tag{5.37}$$

$$P(Z = n) := P\begin{pmatrix} Y'_{-M+2} = 0, & Y'_{-M+3} = 0, & \cdots, & Y'_{-1} = 0, & Y'_0 = 0 \\ Y_{-M+2} = 0, & Y_{-M+3} = 0, & \cdots, & Y_{-1} = 0, & Y_0 = 0 \end{pmatrix}. \tag{5.38}$$

which are identical for each position throughout the underlying DNA sequence since the DNA sequence is generated by a homogeneous background model that starts in the stationary distribution. Recall from our previous discussion that motif hits occur with the same probability on both strands and at each position in the Bernoulli process (see Introduction). That is, the probability of obtaining a motif hit equals $P(h_f) = P(h_r) = \alpha$ regardless of the position in the DNA sequence. The remaining state probabilities are initially unknown.

Next, we introduce the transition probabilities that may arise in one step (e.g. from Z_{-1}

to Z_0) analogously to the previous section as

$$P(n \to h_f) := P\left(\begin{array}{l} Y'_0 = \bullet \\ Y_0 = 1 \end{array}\middle|\begin{array}{lll} Y'_{-1} = 0, & \cdots, & Y'_{-M+1} = 0 \\ Y_{-1} = 0, & \cdots, & Y_{-M+1} = 0 \end{array}\right) \tag{5.39}$$

$$P(n \to h_r) := P\left(\begin{array}{l} Y'_0 = 1 \\ Y_0 = 0 \end{array}\middle|\begin{array}{lll} Y'_{-1} = 0, & \cdots, & Y'_{-M+1} = 0 \\ Y_{-1} = 0, & \cdots, & Y_{-M+1} = 0 \end{array}\right) \tag{5.40}$$

$$P(n \to n) := 1 - P(n \to h_f) - P(n \to h_r) \tag{5.41}$$

$$P(h_f \to h_r) := P(Y'_0 = 1 | Y_0 = 1) = \beta_{3',0} \tag{5.42}$$

$$\begin{aligned} P(h_f \to n_0) &:= P(Y'_0 = 0 | Y_0 = 1) \\ &= 1 - P(h_f \to h_r) \end{aligned} \tag{5.43}$$

$$P(h_r \to h_f) := P(Y_0 = 1 | Y'_{-1} = 1) = \beta_{5',1} \tag{5.44}$$

$$P(h_r \to h_r) := P(Y'_0 = 1 | Y'_{-1} = 1) = \beta_1 \tag{5.45}$$

$$P(h_r \to n'_1) := P(Y'_0 = 0, Y_0 = 0 | Y'_{-1} = 1) \tag{5.46}$$

$$P(n_k \to h_f) := P\left(\begin{array}{l} Y'_0 = \bullet \\ Y_0 = 1 \end{array}\middle|\begin{array}{lll} Y'_{-1} = 0, & \cdots, & Y'_{-k-1} = 0 \\ Y_{-1} = 0, & \cdots, & Y_{-k-1} = 1 \end{array}\right) \quad \text{for } 0 \le k \le M-2 \tag{5.47}$$

$$P(n_k \to h_r) := P\left(\begin{array}{l} Y'_0 = 1 \\ Y_0 = 0 \end{array}\middle|\begin{array}{lll} Y'_{-1} = 0, & \cdots, & Y'_{-k-1} = 0 \\ Y_{-1} = 0, & \cdots, & Y_{-k-1} = 1 \end{array}\right) \quad \text{for } 0 \le k \le M-2 \tag{5.48}$$

$$P(n'_k \to h_f) := P\left(\begin{array}{l} Y'_0 = \bullet \\ Y_0 = 1 \end{array}\middle|\begin{array}{lll} Y'_{-1} = 0, & \cdots, & Y'_{-k-1} = 1 \\ Y_{-1} = 0, & \cdots, & Y_{-k-1} = \bullet \end{array}\right) \quad \text{for } 1 \le k \le M-2 \tag{5.49}$$

$$P(n'_k \to h_r) := P\left(\begin{array}{l} Y'_0 = 1 \\ Y_0 = 0 \end{array}\middle|\begin{array}{lll} Y'_{-1} = 0, & \cdots, & Y'_{-k-1} = 1 \\ Y_{-1} = 0, & \cdots, & Y_{-k-1} = \bullet \end{array}\right) \quad \text{for } 1 \le k \le M-2 \tag{5.50}$$

$$P(n_k \to n_{k+1}) := 1 - P(n_k \to h_f) - P(n_k \to h_r) \quad \text{for } 0 \le k \le M-3 \tag{5.51}$$

$$P(n'_k \to n'_{k+1}) := 1 - P(n'_k \to h_f) - P(n'_k \to h_r) \quad \text{for } 1 \le k \le M-3 \tag{5.52}$$

$$P(n_{M-2} \to n) := 1 - P(n_{M-2} \to h_f) - P(n_{M-2} \to h_r) \tag{5.53}$$

$$P(n'_{M-2} \to n) := 1 - P(n'_{M-2} \to h_f) - P(n'_{M-2} \to h_r) \tag{5.54}$$

Definition (5.41), (5.43), (5.51), (5.52), (5.53) and (5.54) are determined by the remaining transition probabilities, which we shall derive in the following. Definition (5.39) and (5.40) correspond to the probabilities of obtaining a clump start on the forward strand and the reverse strand, respectively. The clump start probabilities are initially unknown and we seek to identify them by using a similar optimization procedure as described in the previous Section. In order ensure that the optimization procedure is identifiable and gives rise to a plausible result, we approximate $P(n \to h_r)$ as a function of $P(n \to h_f)$ according to

$$\begin{aligned}
P(n \to h_r) &= P\left(\begin{matrix} Y'_0 = 1 \\ Y_0 = 0 \end{matrix}\middle|\, Y'_{-1} = 0, \quad \cdots, \quad Y'_{-M-1} = 0 \right) \\
&\approx P\left(\begin{matrix} Y'_0 = 0 \\ Y_0 = 1 \end{matrix}\middle|\, Y'_{-1} = 0, \quad \cdots, \quad Y'_{-M-1} = 0 \right) \\
&= P\left(\begin{matrix} Y'_0 = \bullet \\ Y_0 = 1 \end{matrix}\middle|\, Y'_{-1} = 0, \quad \cdots, \quad Y'_{-M-1} = 0 \right) \times (1 - P(Y'_0 = 1 | Y_0 = 1)) \\
&= P(n \to h_f) \times (1 - \beta_{3',0})
\end{aligned} \tag{5.55}$$

where we assume that palindromic events (conditioning on the upstream non-hits) are symmetric and use the independence Assertion (3.4). Note, however, that while for the unconditional case, $P(Y'_0 = 0, Y_0 = 1) = P(Y'_0 = 1, Y_0 = 0)$ indeed is symmetric, for the conditional case, the second line is merely an approximation.

We continue by deriving the quantities defined by (5.46)-(5.50) in terms of *priniple overlapping hit probabilities* (β_k, $\beta_{3',k}$, $\beta_{5',k}$) defined by (3.42)-(3.45) as well as the probability that no further overlapping hits occur (δ_k and δ'_k) defined by (3.51)-(3.52). Accordingly, we obtain

$$\begin{aligned} P(h_r \to n'_1) &:= 1 - P(Y_0 = 1 | Y'_{-1} = 1) - P(Y'_0 = 1, Y_0 = 0 | Y'_{-1} = 1) \\ &= \quad 1 - \beta_{5',1} - \beta_1 \end{aligned} \tag{5.56}$$

$$\begin{aligned} P(n_k \to h_f) &:= P\left(\begin{matrix} Y'_0 = \bullet \\ Y_0 = 1 \end{matrix} \middle| \begin{matrix} Y'_{-1} = 0, & \cdots, & Y'_{-k-1} = 0 \\ Y_{-1} = 0, & \cdots, & Y_{-k-1} = 1 \end{matrix} \right) \\ &= \frac{P(Y_0 = 1, Y_{-1} = 0 \cdots Y_{-k} = 0, Y'_{-1} = 0 \cdots Y'_{-k-1} = 0 | Y_{-k-1} = 1)}{P(Y_{-1} = 0 \cdots Y_{-k} = 0, Y'_{-1} = 0 \cdots Y'_{-k-1} = 0 | Y_{-k-1} = 1)} \\ &= \frac{P(Y_{k+1} = 1, Y_k = 0 \cdots Y_1 = 0, Y'_k = 0 \cdots Y'_0 = 0 | Y_0 = 1)}{P(Y_k = 0 \cdots Y_1 = 0, Y'_k = 0 \cdots Y'_1 = 0 | Y_0 = 1)} \\ &= \frac{\beta_{k+1}}{\delta_k} \qquad \text{for } 0 \le k \le M-2, \end{aligned} \tag{5.57}$$

$$\begin{aligned} P(n_k \to h_r) &:= P\left(\begin{matrix} Y'_0 = 1 \\ Y_0 = 0 \end{matrix} \middle| \begin{matrix} Y'_{-1} = 0, & \cdots, & Y'_{-k-1} = 0 \\ Y_{-1} = 0, & \cdots, & Y_{-k-1} = 1 \end{matrix} \right) \\ &= \frac{P(Y'_0 = 1, Y_0 = 0 \cdots Y_{-k} = 0, Y'_{-1} = 0 \cdots Y'_{-k-1} = 0 | Y_{-k-1} = 1)}{P(Y_{-1} = 0 \cdots Y_{-k} = 0, Y'_{-1} = 0 \cdots Y'_{-k-1} = 0 | Y_{-k-1} = 1)} \\ &= \frac{P(Y'_{k+1} = 1, Y_{k+1} = 0 \cdots Y_1 = 0, Y'_k = 0 \cdots Y'_0 = 0 | Y_0 = 1)}{P(Y_k = 0 \cdots Y_1 = 0, Y'_k = 0 \cdots Y'_0 = 0 | Y_0 = 1)} \\ &= \frac{\beta_{3',k+1}}{\delta_k} \qquad \text{for } 0 \le k \le M-2, \end{aligned} \tag{5.58}$$

$$
\begin{aligned}
P(n'_k \to h_f) &:= P\left(\begin{matrix} Y'_0 = \bullet \\ Y_0 = 1 \end{matrix}\middle|\begin{matrix} Y'_{-1} = 0, & \cdots, & Y'_{-k-1} = 1 \\ Y_{-1} = 0, & \cdots, & Y_{-k-1} = \bullet \end{matrix}\right) \\
&= \frac{P(Y_0 = 1, Y_{-1} = 0 \cdots Y_{-k} = 0, Y'_{-1} = 0 \cdots Y'_{-k} = 0 | Y'_{-k-1} = 1)}{P(Y_{-1} = 0 \cdots Y_{-k} = 0, Y'_{-1} = 0 \cdots Y'_{-k} = 0 | Y'_{-k-1} = 1)} \\
&= \frac{P(Y_{k+1} = 1, Y_k = 0 \cdots Y_1 = 0, Y'_k = 0 \cdots Y'_1 = 0 | Y'_0 = 1)}{P(Y_k = 0 \cdots Y_1 = 0, Y'_k = 0 \cdots Y'_1 = 0 | Y'_0 = 1)} \\
&= \frac{\beta_{5',k+1}}{\delta'_k} \qquad \text{for } 1 \le k \le M-2
\end{aligned} \tag{5.59}
$$

and

$$
\begin{aligned}
P(n'_k \to h_r) &:= P\left(\begin{matrix} Y'_0 = 1 \\ Y_0 = 0 \end{matrix}\middle|\begin{matrix} Y'_{-1} = 0, & \cdots, & Y'_{-k-1} = 1 \\ Y_{-1} = 0, & \cdots, & Y_{-k-1} = \bullet \end{matrix}\right) \\
&= \frac{P(Y'_0 = 1, Y_0 = 0 \cdots Y_{-k} = 0, Y'_{-1} = 0 \cdots Y'_{-k} = 0 | Y'_{-k-1} = 1)}{P(Y_{-1} = 0 \cdots Y_{-k} = 0, Y'_{-1} = 0 \cdots Y'_{-k} = 0 | Y'_{-k-1} = 1)} && (5.60) \\
&= \frac{P(Y'_{k+1} = 1, Y_{k+1} = 0 \cdots Y_1 = 0, Y'_k = 0 \cdots Y'_1 = 0 | Y'_0 = 1)}{P(Y_k = 0 \cdots Y_1 = 0, Y'_k = 0 \cdots Y'_1 = 0 | Y'_0 = 1)} && (5.61) \\
&= \frac{\beta_{k+1}}{\delta'_k} \qquad \text{for } 1 \le k \le M-2 && (5.62)
\end{aligned}
$$

where we made use of the symmetry relationship that we derived in Equation (3.57) to express $\beta_1 = P(Y'_0 = 1, Y_0 = 0 | Y'_{-1} = 1)$ in Equation (5.56) and $\beta_{k+1} = P(Y'_{k+1} = 1, Y_{k+1} = 0 \cdots Y_1 = 0, Y'_k = 0 \cdots Y'_1 = 0 | Y'_0 = 1)$ in Equation (5.61).

Note that in the case of a perfect palindrome, we have $\beta_{3',0} = 1$ and consequently, $\delta_0 = 0$, in which case $P(n_k \to h_f)$ as well as $P(n_k \to h_r)$ are undefined. However, since the states n_k for $0 \le k \le M-2$ cannot be reached in that case, we can ignore that.

The transition matrix

In the previous Section, we have defined the state probabilities as well as all transition probabilities for the Markov model that results in scanning both DNA strands for motif hits. In this Section, we shall construct a compact representation of the Markov model in matrix notation by introducing the transition matrix as follows

$$M = \begin{bmatrix} 0 & P(h_r \to h_f) & P(n \to h_f) & \mathbf{a}^\top & \mathbf{a'}^\top \\ P(h_f \to h_r) & P(h_r \to h_r) & P(n \to h_r) & \mathbf{b}^\top & \mathbf{b'}^\top \\ 0 & 0 & P(n \to n) & \mathbf{c}^\top & \mathbf{c}^\top \\ P(h_f \to n_0) & 0 & \cdots & & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{C} & \mathbf{0} \\ 0 & P(h_r \to n'_1) & 0 & \cdots & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{D} \end{bmatrix} \tag{5.63}$$

where the bold zeros denote sub-matrices containing only zeros and where we made use of the following constituent parts

$$\mathbf{a}^\top = \begin{bmatrix} P(n_0 \to h_f) & \cdots & P(n_{M-2} \to h_f) \end{bmatrix} \tag{5.64}$$

$$\mathbf{a'}^\top = \begin{bmatrix} P(n'_1 \to h_f) & \cdots & P(n'_{M-2} \to h_f) \end{bmatrix} \tag{5.65}$$

$$\mathbf{b}^\top = \begin{bmatrix} P(n_0 \to h_r) & \cdots & P(n_{M-2} \to h_r) \end{bmatrix} \tag{5.66}$$

$$\mathbf{b'}^\top = \begin{bmatrix} P(n'_1 \to h_r) & \cdots & P(n'_{M-2} \to h_r) \end{bmatrix} \tag{5.67}$$

$$\mathbf{c}^\top = \begin{bmatrix} 0 & \cdots & 0 & P(n_{M-2} \to n) \end{bmatrix} \tag{5.68}$$

$$\mathbf{c'}^\top = \begin{bmatrix} 0 & \cdots & 0 & P(n'_{M-2} \to n) \end{bmatrix} \tag{5.69}$$

$$\mathbf{C} = \begin{bmatrix} P(n_0 \to n_1) & 0 & & 0 \\ 0 & P(n_1 \to n_2) & & 0 \\ & & \ddots & \\ 0 & 0 & & P(n_{M-3} \to n_{M-2}) \end{bmatrix} \tag{5.70}$$

$$\mathbf{D} = \begin{bmatrix} P(n'_1 \to n'_2) & 0 & & 0 \\ 0 & P(n'_2 \to n'_3) & & 0 \\ & & \ddots & \\ 0 & 0 & & P(n'_{M-3} \to n'_{M-2}) \end{bmatrix}. \tag{5.71}$$

A graphical representation of the transition matrix stated by Equation (5.63) is depicted in Figure 5.3.

We shall assume that we have chosen the significance level α such that $0 < P(Z = n) < 1$, in which case, the Markov model consists of one recurrent class. In this case, it is guaranteed that there exists a unique stationary distribution for this model. We denote the stationary distribution by $\mu(Z)$. The stationary distribution can be computed by eigenvalue decomposition of the transition matrix for which we use the power method. The eigenvector that is associated with the largest eigenvalue (which is equal to one for stochastic matrices) equals the stationary distribution [26].

However, since the unknown *clump start probability* $P(n \to h_f)$ parametrizes the transition matrix (which is the only unknown parameter), we shall express the stationary distribution as a function of $P(n \to h_f)$. That is $\mu(Z, P(n \to h_f))$.

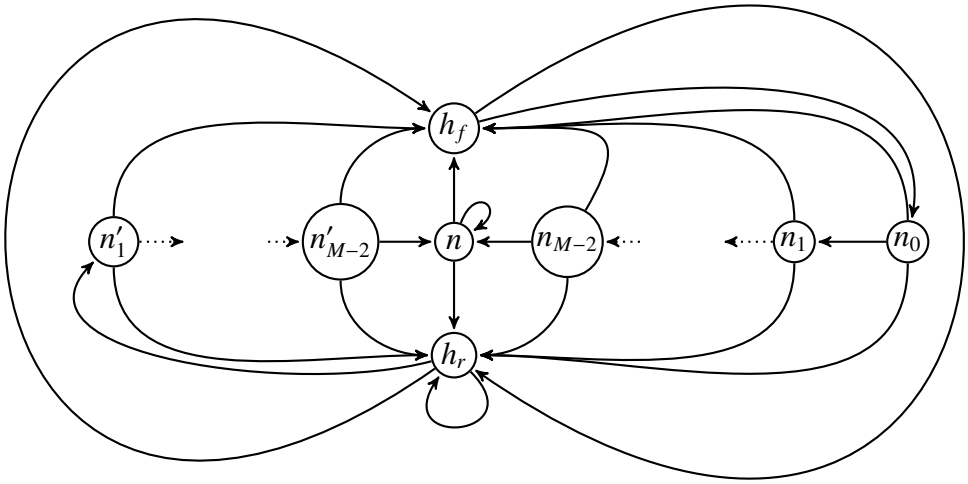


Figure 5.3: State transitions of the Markov model that can be achieved from Z_{-1} to Z_0. The correspondence relationship between patterns in the underlying Bernoulli process $\mathbf{Y}_{[1:N-M+1]}$ and the state space Z is defined by (5.33). The feasible state transitions are denoted by the arrows, which are associated with the transition probabilities defined by (5.39)-(5.54).

5.2.3 Identification of the clump start probability

We continue by explaining the identification of the *clump start probabilities* $P(n \to h_f)$ in this section, which is based by two insights: First, $P(n \to h_f)$ is the only unknown parameter of the transition matrix defined by (5.63). All remaining transition probabilities are determined by $P(n \to h_f)$ and/or the *principal overlapping hit probabilities* which are considered to be known. Second, due to the fact that the underlying DNA sequence was generated by a homogeneous order-d background model which starts in the stationary distribution, the probability of obtaining a motif hit is identical across all positions and both strands $P(Y_i = 1) = P(Y'_i = 1) = \alpha$. Consequently, the corresponding 'hit' states h_f and h_r are desired to be emitted with probability α in the stationary distribution, respectively, as well. Therefore, we seek to identify $P(n \to h_f)$ such that

$$\mu(h_f, P(n \to h_f)) + \mu(h_r, P(n \to h_f)) = P(Y = 1) + P(Y' = 1) = 2 \times \alpha.$$

To achieve our means, we define the discrepancy measure between the probability to obtain a motif hit 2α and the Markov model's current probability to producing hits $\mu(h_f; P(n \to h_f)) + \mu(h_r; P(n \to h_r))$

$$f(P(n \to h_f)) := \quad (2 \cdot \alpha - \mu(h_f; P(n \to h_f)) - \mu(h_r; P(n \to h_f)))^2.$$

We initialize $P(n \to h_f)$ (e.g. with $P(n \to h_f) = \alpha$) and iteratively refine the *clump start probability* using a gradient descent approach such that the discrepancy measure is gradually decreased. Therefore, in each iteration we determine the stationary distribution for the current $P(n \to h_f)$ using the power method and we determine the numerical second-order derivative of the discrepancy measure with respect to $P(n \to h_f)$ according

to

$$\frac{f(P(n \to h_f) + \epsilon) - f(P(n \to h_f) - \epsilon)}{2\epsilon}$$

where ϵ denotes a small constant (e.g. $\epsilon = 10^{-8}$).

We terminate the optimization procedure after a desired tolerance level has been reached, in order to estimate the *clump start probabilities* $P(n \to h_f)$ and $P(n \to h_r)$.

5.3 Discussion

In this Chapter, we have introduced Markov models whose states represent patterns in the Bernoulli process $\mathbf{Y}_{[1:N-M+1]}$ which resulted from scanning a random DNA sequence for motif hits using the log-likelihood ratio. Accordingly, a state sequence, denoted by $Z_1 Z_2 \cdots$ uniquely expresses a particular realization of $\mathbf{Y}_{[1:N-M+1]}$.

We first started by introducing the Markov model variant that represents scanning of a single DNA strand for motif hits. The main motivation for the Markov models is to approximately identifying the *clump start probability* (denoted by $P(n \to h)$) using a given TF motif, a background model and a significance level α. To this end, we seeks to identify the *clump start probability* such that the Markov model's stationary distribution $\mu(Z)$ emits the 'hit' states h with probability α, where we exploit the fact that the *clump start probability* is the only free parameter for defining the Markov model. All other transition probabilities are either zero or determined by the *principal overlapping hit probabilities*.

Subsequently, we extended the Markov model for the case when both DNA strands are scanned for motif hits. Again, the primary use of the Markov model is to evaluate

the *clump start probabilities*, denoted by $P(n \rightarrow h_f)$ and $P(n \rightarrow h_r)$ such that in the stationary distribution the 'hit' states $\mu(Z = h_f) + \mu(Z = h_r)$ are exhibited with probability 2α using an analogous optimization strategy. In order to render the optimization problem identifiable and enforce plausible results, we expressed $P(n \rightarrow h_r)$ in terms of $P(n \rightarrow h_f)$ such that the only free parameter becomes $P(n \rightarrow h_f)$.

The *clump start probability* becomes crucial for the combinatorial model that is introduce in Chapter 6. Therefore, the Markov model that was developed in this chapter can be viewed as an auxiliary model for the combinatorial model.

The Markov model may also be viewed as an alternative motif hit declumping approach. In contrast to the compound Poisson model, which rests on the Poisson assumption (or 'rare hit' assumption) [41], the Markov model does not rely on this assumption. The reason for this is that upon a clump starting hit, the memory states (e.g. n_k and n'_k) first must to be fully traversed such that n is again emitted before another clump start can possibly occur. This prevents clumps from mutually overlapping one another. On the other hand, the compound Poisson model lacks such a restriction, because each position in the DNA sequence is considered independent from one another, which might initiate a clump with a certain probability. That is, neighbouring positions might give raise to clumps simultanously, which eventually leads to biases if the stringency level α was chosen too relaxed.

While the Markov model does not rely on the Poisson assumption, it is instead based on the validity of the independence assumptions (3.2)-(3.8), which are violated for low stringency significance levels (e.g. $\alpha > 0.05$) and which would also incur biases. However, unless the stringency of the significance levels is extremely low (e.g. $\alpha > 0.05$), we found that the results are only mildly affected by a violation of this assumption, which

is studied in in Chapter 6.

As discussed above, the approximation of the *clump start probability* is based on two facts: First, the only parameter that is unknown is the *clump start probability*, while the remaining entries of the transition matrix are implicitly defined by the *principal overlapping hit probabilities* which were derived in Chapter 3 and are considered fixed. Second, we know that stationary probability of emitting a motif hit must equal α, due to the fixed significance level. However, since the *principal overlapping hit probabilities* are themselves approximated using the Assumptions (3.2)-(3.8), the identified *clump start probability* implicitly compensate for biases in the *principal overlapping hit probabilities* such that $\mu(h_f) + \mu(h_r)$ matches 2α. This features increases the accuracy of the combinatorial model which we introduce in Chapter 6.

Chapter 6

A combinatorial model for the number of motif hits

In this chapter, we shall introduce a novel statistical model for describing the number of motif hits in a DNA sequence, which we refer to as the *combinatorial model*. In the combinatorial model we attempt to approximate the distribution of the number of motif hits $P(X = x|H_0)$ by summing up the probabilities associated with all realizations of $\mathbf{Y}_{[1:N-M+1]}$ that contain exactly x motif hits, where N denotes the length of the DNA sequence and M denotes the TF motif length.

For didactical reasons we shall first address the combinatorial model for the case of scanning a single DNA strand. We show how, for each realization, $P(\mathbf{Y}_{[1:N-M+1]})$ can be approximately factorized into constituent parts that are defined by the *principal overlapping hit probabilities* defined by (3.41) and the *clump start probability* $P(Y_0 = 1|Y_{-1} = 0, \cdots, Y_{-M+1} = 0)$ which was identified by the Markov model in Chapter 5. Subsequently, we adopt the dynamic programming algorithm described by Liu and Lawrence

[31] to derive the distribution of the number of motif hits in a DNA sequence of finite length. Afterwards, we extend the *combinatorial model* to motif hits that were obtained by scanning both strands of the DNA.

We note that Zhang *et al.* [59] proposed a related dynamic programming approach for studying the enrichment of individual words and sets of words in DNA sequences. The previous approach yields the exact distribution of the number of words, which can only be applied for small sets of words, because its runtime depends on the number of words. By contrast, our dynamic programming algorithm does not depend on the number of words in the compatible set $C(t_s)$, because we use the log-likelihood ratio in conjunction with a score threshold as a proxy to avoid having to enumerate all compatible words individually.

We systematically compare the novel combinatorial model with the compound Poisson model and a binomial model that were described in the previous chapters. We find that while for high stringency score cut-offs, the combinatorial model and the compound Poisson model achieve similar results, for low stringency cut-offs, the combinatorial model generally yields more accurate results.

6.1 The combinatorial model for scanning a single DNA strand

In this section, we shall discuss the *combinatorial model* for scanning a single DNA strand which induces the Bernoulli process $Y_1 Y_2 \cdots Y_{N-m+1}$.

6.1.1 Factorization of $P(\mathbf{Y}_{[1:N-M+1]})$

Our main interest in this section rests on $P(\mathbf{Y}_{[1:N-M+1]})$ and its computation, because it forms the basis for the distribution of the number of motif hits $P(X|H_0)$.

Recall that the joint distribution $P(\mathbf{Y}_{[1:N-M+1]})$ is determined by summing over all possible underlying DNA sequences according to Equation (1.26), which induces complicated statistical correlations between the events in $\mathbf{Y}_{[1:N-M+1]}$. Therefore, in general, $P(\mathbf{Y}_{[1:N-M+1]})$ does not factorize exactly. Even though it would be desirable to compute $P(\mathbf{Y}_{[1:N-M+1]})$ directly according to Equation (1.26), this is in general prohibitively expensive since it requires to deal with exponentially many realizations of $\mathbf{Y}_{[1:N-M+1]}$ (see Section 1.2.4).

However, we can nevertheless approximately factorize $P(\mathbf{Y}_{[1:N-M+1]} = \mathbf{y})$ for a given realization $\mathbf{y}$ by utilizing 1) the independence assumption defined by (3.2), 2) the assumption that non-overlapping events Y_i and Y_j for $j \geq i + M$ are independent and 3) that the *clump start probability* $P(Y_0 = 1|Y_{-1} = 0, \cdots, Y_{-M+1} = 0)$ that was approximated using the Markov model (see Chapter 5), is constant throughout the sequence. That latter point assumes that prior to the start of the Bernoulli process $Y_1 Y_2 \cdots$, all events are zeros (e.g. $\mathbf{Y}_{[-1:-M+1]} = \mathbf{0}$).

These assumptions give rise to a set of factors that are multiplied together in order to express the joint probabability $P(\mathbf{Y}_{[1:N-M+1]} = \mathbf{y})$. For convenience, we shall abbreviate the *clump start probability* in the remainder of this chapter by

$$\alpha' := P(Y_0 = 1|Y_{-1} = 0, \cdots, Y_{-M+1} = 0). \tag{6.1}$$

Table 6.1: Conditional probabilities which are used to factorize $P(\mathbf{Y}_{[1:N-M+1]})$ for scanning a single DNA strand.

Probability	Abbreviation	Reference
$P(Y_0 = 1 \mid Y_{-1} = 0, \cdots Y_{-M+1} = 0)$	α'	see Definition (6.1)
$P(Y_0 = 0 \mid Y_{-1} = 0, \cdots Y_{-M+1} = 0)$	$1 - \alpha'$	see Definition (6.1)
$P(Y_k = 1, \mathbf{Y}_{[1:k-1]} = \mathbf{0} \mid Y_0 = 1)$	β_k	see Definition (3.41)
$P(\mathbf{Y}_{[1:k]} = \mathbf{0} \mid Y_0 = 1)$	δ_k	see Definition (3.50)

A summary of all quantities that are facilitated to factorize $P(\mathbf{Y}_{[1:N-M+1]})$ is given in Table 6.1.

Illustrations of the combination model

In this Section, we demonstrate the application of the combinatorial model for computing $P(X = x)$ by factorizing each realization of $P(\mathbf{Y}_{[1:N-M+1]})$ on a toy example. Therefore, consider a motif of length $M = 3$ that is used to scan a DNA sequence of length $N = 7$ for motif hits. We shall exemplify the computation of $P(X = 0)$, $P(X = 1)$ and $P(X = 2)$ in the following (see Figures 6.1 - 6.3). To this end, we sum over all combinations of placing 0, 1 and 2 *ones* in a sequence of binary events of length $N - M + 1 = 5$. Accordingly, this leads to the following computation

$$P(X = 0) = P(\mathbf{Y}_{[1:5]} = 00000) \tag{6.2}$$

$$\begin{aligned} P(X = 1) =& P(\mathbf{Y}_{[1:5]} = 10000) + P(\mathbf{Y}_{[1:5]} = 01000) + \\ & P(\mathbf{Y}_{[1:5]} = 00100) + P(\mathbf{Y}_{[1:5]} = 00010) + \\ & P(\mathbf{Y}_{[1:5]} = 00001) \end{aligned} \tag{6.3}$$

$$\begin{aligned} P(X = 2) =& P(\mathbf{Y}_{[1:5]} = 11000) + P(\mathbf{Y}_{[1:5]} = 10100) + \\ & P(\mathbf{Y}_{[1:5]} = 10010) + P(\mathbf{Y}_{[1:5]} = 10001) + \end{aligned}$$

$$P(\mathbf{Y}_{[1:5]} = 01100) + P(\mathbf{Y}_{[1:5]} = 01010) +$$
$$P(\mathbf{Y}_{[1:5]} = 01001) + P(\mathbf{Y}_{[1:5]} = 00110) +$$
$$P(\mathbf{Y}_{[1:5]} = 00101) + P(\mathbf{Y}_{[1:5]} = 00011). \tag{6.4}$$

We proceed by approximating each term on the right hand side of Equation (6.2)-(6.4) using the quantities summarized in Table 6.1.

First, consider $P(X = 0)$ from Equation (6.2), which represents the simplest case, because there is only one way to obtain only *zeros*. We approximate this case by

$$\begin{aligned}
P(X = 0) &= P(\mathbf{Y}_{[1:5]} = \mathbf{0}) \\
&\approx P(\mathbf{Y}_{[1:5]} = \mathbf{0} | Y_0 = 0, Y_{-1} = 0) \\
&= P(Y_1 = 0 | Y_0 = 0, Y_{-1} = 0) \times \\
&\quad \times P(Y_2 = 0 | Y_1 = 0, Y_0 = 0) \times \\
&\quad \times P(Y_3 = 0 | Y_2 = 0, Y_1 = 0) \times \\
&\quad \times P(Y_4 = 0 | Y_3 = 0, Y_2 = 0) \times \\
&\quad \times P(Y_5 = 0 | Y_4 = 0, Y_3 = 0) \times \\
&= P(Y_0 = 0 | Y_{-1} = 0, Y_{-2} = 0)^5 \\
&= (1 - \alpha')^5
\end{aligned}$$

where we made use of the assumption that prior to the start of the Bernoulli process, only *zeros* are emitted, in the second line. Furthermore, we used the complementary probability of the *clump start probability*, defined by (6.1), and the fact that the underlying background model is homogeneous and starts in the stationary distribution to

establish the approximation.

Second, for computing $P(X = 1)$, we need to evaluated each of the quantities on the right hand side of Equation (6.3) in turn and sum them up. We start by evaluating $P(\mathbf{Y}_{[1:5]} = 10000)$. Note that we not only need to account for the *clump start hit* at position one, but also that the events Y_2 and Y_3 are directly influenced by the occurrence of the hit, because of the overlap with those positions. We take this into account by δ_2 (see Table 6.1). Thus, we obtain the following approximation for this quantity

$$\begin{aligned} P(\mathbf{Y}_{[1:5]} = 10000) &\approx P(\mathbf{Y}_{[1:5]} = 10000 | Y_0 = 0, Y_{-1} = 0) \\ &= P(Y_1 = 1 | Y_0 = 0, Y_{-1} = 0) \times \\ &\quad \times P(Y_2 = 0, Y_3 = 0 | Y_1 = 1) \times \\ &\quad \times P(Y_4 = 0 | Y_3 = 0, Y_2 = 0) \times \\ &\quad \times P(Y_5 = 0 | Y_4 = 0, Y_3 = 0) \\ &= \alpha' \times \delta_2 \times (1 - \alpha')^2 \end{aligned} \tag{6.5}$$

Moreover, employing the assumption that $Y_0 = 0$ and $Y_{-1} = 0$, the following quantities are identically

$$\begin{aligned} P(\mathbf{Y}_{[1:5]} = 01000 | Y_0 = 0, Y_{-1} = 0) &= P(\mathbf{Y}_{[1:5]} = 00100 | Y_0 = 0, Y_{-1} = 0) \\ &= P(\mathbf{Y}_{[1:5]} = 10000 | Y_0 = 0, Y_{-1} = 0) \\ &= \alpha' \times \delta_2 \times (1 - \alpha')^2. \end{aligned}$$

Hence,

$$P(\mathbf{Y}_{[1:5]} = 01000) \approx P(\mathbf{Y}_{[1:5]} = 00100) \approx P(\mathbf{Y}_{[1:5]} = 10000) \approx \alpha' \times \delta_2 \times (1 - \alpha')^2.$$

We derive the remaining two realizations for Equation (6.3) in an analogous fashion as

$$\begin{aligned}
P(\mathbf{Y}_{[1:5]} = 00010) &\approx P(\mathbf{Y}_{[1:5]} = 00010 | Y_0 = 0, Y_{-1} = 0) \\
&= P(Y_1 = 0 | Y_0 = 0, Y_{-1} = 0) \times \\
&\quad \times P(Y_2 = 0 | Y_1 = 0, Y_0 = 0) \times \\
&\quad \times P(Y_3 = 0 | Y_2 = 0, Y_1 = 0) \times \\
&\quad \times P(Y_4 = 1 | Y_3 = 0, Y_2 = 0) \times \\
&\quad \times P(Y_5 = 0 | Y_4 = 1) \\
&= \alpha' \times \delta_1 \times (1 - \alpha')^3, \\
P(\mathbf{Y}_{[1:5]} = 00001) &\approx P(\mathbf{Y}_{[1:5]} = 00001 | Y_0 = 0, Y_{-1} = 0) \\
&= P(Y_1 = 0 | Y_0 = 0, Y_{-1} = 0) \times \\
&\quad \times P(Y_2 = 0 | Y_1 = 0, Y_0 = 0) \times \\
&\quad \times P(Y_3 = 0 | Y_2 = 0, Y_1 = 0) \times \\
&\quad \times P(Y_4 = 0 | Y_3 = 0, Y_2 = 0) \times \\
&\quad \times P(Y_5 = 1 | Y_4 = 0, Y_3 = 0) \\
&= \alpha' \times (1 - \alpha')^4.
\end{aligned}$$

Finally, the treatment of $P(X = 2)$ becomes slightly more complicated, since in addition to the hits that start a clump, we need to deal with overlapping hits as well. Thus,

the individual probabilities on the right hand side of Equation (6.4) are approximated according to

$$
\begin{aligned}
P(\mathbf{Y}_{[1:5]} = 11000) &\approx P(\mathbf{Y}_{[1:5]} = 11000 | Y_0 = 0, Y_{-1} = 0) \\
&= P(Y_1 = 1 | Y_0 = 0, Y_{-1} = 0) \times \\
&\quad \times P(Y_2 = 1 | Y_1 = 1) \times \\
&\quad \times P(Y_4 = 0, Y_3 = 0 | Y_2 = 1) \times \\
&\quad \times P(Y_5 = 0 | Y_4 = 0, Y_3 = 0) \\
&= \alpha' \times \beta_1 \times \delta_2 \times (1 - \alpha') \\
P(\mathbf{Y}_{[1:5]} = 01100) &\approx P(\mathbf{Y}_{[1:5]} = 01100 | Y_0 = 0, Y_{-1} = 0) \\
&= P(\mathbf{Y}_{[1:5]} = 11000 | Y_0 = 0, Y_{-1} = 0) \\
&= \alpha' \times \beta_1 \times \delta_2 \times (1 - \alpha') \\
P(\mathbf{Y}_{[1:5]} = 10100) &\approx P(\mathbf{Y}_{[1:5]} = 10100 | Y_0 = 0, Y_{-1} = 0) \\
&= P(Y_1 = 1 | Y_0 = 0, Y_{-1} = 0) \times \\
&\quad \times P(Y_3 = 1, Y_2 = 0 | Y_1 = 1) \times \\
&\quad \times P(Y_5 = 0, Y_4 = 0 | Y_3 = 1) \\
&= \alpha' \times \beta_2 \times \delta_2 \\
P(\mathbf{Y}_{[1:5]} = 10010) &\approx P(\mathbf{Y}_{[1:5]} = 10010 | Y_0 = 0, Y_{-1} = 0) \\
&= P(Y_1 = 1 | Y_0 = 0, Y_{-1} = 0) \times \\
&\quad \times P(Y_3 = 0, Y_2 = 0 | Y_1 = 1) \times \\
&\quad \times P(Y_4 = 1 | Y_3 = 0, Y_2 = 0) \times \\
&\quad \times P(Y_5 = 0 | Y_4 = 1)
\end{aligned}
$$

$$= \alpha'^2 \times \delta_2 \times \delta_1$$

$$P(\mathbf{Y}_{[1:5]} = 01010) \approx P(\mathbf{Y}_{[1:5]} = 01010 | Y_0 = 0, Y_{-1} = 0)$$

$$= P(Y_1 = 0 | Y_0 = 0, Y_{-1} = 0) \times$$

$$\times P(Y_2 = 1 | Y_1 = 0, Y_0 = 0) \times$$

$$\times P(Y_4 = 1, Y_3 = 0 | Y_2 = 1) \times$$

$$\times P(Y_5 = 0 | Y_4 = 1)$$

$$= \alpha' \times \beta_2 \times \delta_1 \times (1 - \alpha')$$

$$P(\mathbf{Y}_{[1:5]} = 00110) \approx P(\mathbf{Y}_{[1:5]} = 00110 | Y_0 = 0, Y_{-1} = 0)$$

$$= P(Y_1 = 0 | Y_0 = 0, Y_{-1} = 0) \times$$

$$\times P(Y_2 = 0 | Y_1 = 0, Y_0 = 0) \times$$

$$\times P(Y_3 = 1 | Y_2 = 0, Y_1 = 0) \times$$

$$\times P(Y_4 = 1 | Y_3 = 1) \times$$

$$\times P(Y_5 = 0 | Y_4 = 1)$$

$$= \alpha' \times \beta_1 \times \delta_1 \times (1 - \alpha')^2$$

$$P(\mathbf{Y}_{[1:5]} = 10001) \approx P(\mathbf{Y}_{[1:5]} = 10001 | Y_0 = 0, Y_{-1} = 0)$$

$$= P(Y_1 = 1 | Y_0 = 0, Y_{-1} = 0) \times$$

$$\times P(Y_3 = 0, Y_2 = 0 | Y_1 = 1) \times$$

$$\times P(Y_4 = 0 | Y_3 = 0, Y_2 = 0) \times$$

$$\times P(Y_5 = 1 | Y_4 = 0, Y_3 = 0) \times$$

$$= \alpha'^2 \times \delta_2 \times (1 - \alpha')$$

$$P(\mathbf{Y}_{[1:5]} = 01001) \approx P(\mathbf{Y}_{[1:5]} = 01001 | Y_0 = 0, Y_{-1} = 0)$$

$$= P(\mathbf{Y}_{[1:5]} = 10001 | Y_0 = 0, Y_{-1} = 0)$$

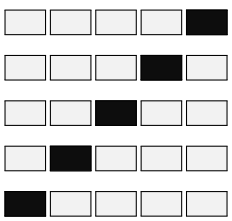
$$
\begin{aligned}
&= \alpha'^2 \times \delta_2 \times (1 - \alpha') \\
P(\mathbf{Y}_{[1:5]} = 00101) &\approx P(\mathbf{Y}_{[1:5]} = 00101 | Y_0, Y_{-1} = 0) \\
&= P(Y_1 = 0 | Y_0 = 0, Y_{-1} = 0) \times \\
&\quad \times P(Y_2 = 0 | Y_1 = 0, Y_0 = 0) \times \\
&\quad \times P(Y_3 = 1 | Y_2 = 0, Y_1 = 0) \times \\
&\quad \times P(Y_5 = 1, Y_4 = 0 | Y_3 = 1) \\
&= \alpha' \times \beta_2 \times \Omega^2 \\
P(\mathbf{Y}_{[1:5]} = 00011) &\approx P(\mathbf{Y}_{[1:5]} = 00011 | Y_0 = 0, Y_{-1} = 0) \\
&= P(Y_1 = 0 | Y_0 = 0, Y_{-1} = 0) \times \\
&\quad \times P(Y_2 = 0 | Y_1 = 0, Y_0 = 0) \times \\
&\quad \times P(Y_3 = 0 | Y_2 = 0, Y_1 = 0) \times \\
&\quad \times P(Y_4 = 1 | Y_3 = 0, Y_2 = 0) \times \\
&\quad \times P(Y_5 = 1 | Y_4 = 1) \\
&= \alpha' \times \beta_1 \times (1 - \alpha')^3.
\end{aligned}
$$

While, the main focus of this illustrative example relied on demonstrating how $P(\mathbf{Y}_{[1:N-M+1]})$ can be factorized for each realization. We employed an enumerative strategy to sum over all realizations of $P(\mathbf{Y}_{[1:N-M+1]})$ that contain x hits. However, the enumerative summation strategy is only possible for small $N - M + 1$ and small numbers of hits x, because the number of permutations for placing x hits in a sequence of length $N - M + 1$ is given by $\binom{N-M+1}{x}$. With increasing $N - M + 1$ and x, enumerating all possible combinations of placing x hits becomes a prohibitively expensive task.

In the next Section, we turn to describing an efficient dynamic programming approach for summing over all combinations of placing x hits in a binary sequence of length $n - M + 1$ in order to determine $P(X = x)$.

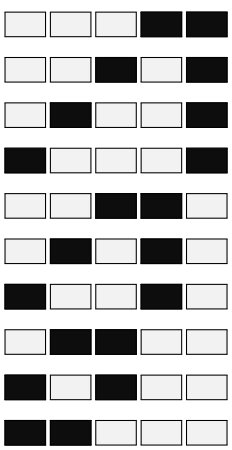
$= P(\mathbf{Y}_{[1:5]} = 00000)$

Figure 6.1: There is only one possibility how zero motif hits can be placed in binary sequence of length 5, where the white boxes represent non-hits.



$= P(\mathbf{Y}_{[1:5]} = 00001)$
$= P(\mathbf{Y}_{[1:5]} = 00010)$
$= P(\mathbf{Y}_{[1:5]} = 00100)$
$= P(\mathbf{Y}_{[1:5]} = 01000)$
$= P(\mathbf{Y}_{[1:5]} = 10000)$

Figure 6.2: Illustration of all combinations of placing one hit in a binary sequence of length 5, where white and black boxes denote non-hits and hits, respectively.



$= P(\mathbf{Y}_{[1:5]} = 00011)$
$= P(\mathbf{Y}_{[1:5]} = 00101)$
$= P(\mathbf{Y}_{[1:5]} = 01001)$
$= P(\mathbf{Y}_{[1:5]} = 10001)$
$= P(\mathbf{Y}_{[1:5]} = 00110)$
$= P(\mathbf{Y}_{[1:5]} = 01010)$
$= P(\mathbf{Y}_{[1:5]} = 10010)$
$= P(\mathbf{Y}_{[1:5]} = 01100)$
$= P(\mathbf{Y}_{[1:5]} = 10100)$
$= P(\mathbf{Y}_{[1:5]} = 11000)$

Figure 6.3: Illustration of all combinations of placing 2 hits in a sequence of length 5, where white and black boxes denote non-hits and hits, respectively.

6.1.2 Efficient summation over combinations of placing x motif hits

In this Section, we derive a dynamic programming algorithm that efficiently sums over the probabilities of all realizations $\mathbf{Y}_{[1:N-M+1]}$ that emit x motif hits. We shall start by first discussing a simplified scenario of the Bernoulli process, by assuming that all events Y_i in $\mathbf{Y}_{[1:N-M+1]}$ are independently and identically distributed. Subsequently, we extend the dynamic programming algorithm such that overlapping hit probabilities are adequatly accounted for.

Efficient summation assuming i.i.d events Y_i in $\mathbf{Y}_{[1:N-M+1]}$

We first derive the algorithm assuming that the events Y_i in the Bernoulli process $\mathbf{Y}_{[1:N-M+1]}$ are drawn independently and identically with $Y_i \sim P(Y)$ for all i where

$$P(Y_0 = 1) = P(Y_0 = 1 | Y_{-1} = 0, \cdots, Y_{-M+1} = 0) = \alpha'.$$

Note that in this situation, the overlapping hit probabilities which are summarized in Table 6.1 become obsolete. Our goal is to determine the probability $P(X = x)$ that x *ones* were emitted in $\mathbf{Y}_{[1:N-M+1]}$.

Due to the independence between all events Y_i in $\mathbf{Y}_{[1:N-M+1]}$, we could directly proceed by employing the binomial distribution

$$P(X = x) = \binom{N - M + 1}{x} \alpha'^{x} (1 - \alpha')^{N-M+1-x}$$

to establish the distribution of the number of successes (or *ones*) in the sequence.

Alternatively, however, we would like to stress a different strategy to achieve the same goal. The $\binom{N-M+1}{x}$ identical quantities ${\alpha'}^{x}(1-\alpha')^{N-M+1-x}$ can be aggregated by a variant of the dynamic programming strategy that was described by Liu and Lawrence [31] and which was originally employed to determine the distribution of the numbers of segment boundaries in the task of sequence segmentation. By slightly modifying the algorithm, we can address the task of computing the distribution of the number of *ones* in $\mathbf{Y}_{[1:N-M+1]}$.

To this end, we introduce the new random variable $X_{[1:j]}$ which represents the number of *ones* in the subsegment $\mathbf{Y}_{[1:j]}$ where $1 \leq j \leq N-M+1$. $X_{[1:j]}$ accounts for all combinations of placing the *ones* in $\mathbf{Y}_{[1:j]}$ (see Figure 6.1-6.3). Similarly as proposed by Liu and Lawrence [31], we determine the probability $P(X_{[1:j]} = x+1)$ by reusing the probabilities $P(X_{[1:i]} = x)$ for $i < j$ that were determined in the previous step and appending an additional hit. Accordingly, for each i, the segment ranging over $[1 : j]$ is split into two subsegments ranging over $[1 : i]$ and $[i+1 : j]$. For the first subsegment we assume that already x *ones* were observed, by utilizing $P(X_{[1:i]} = x)$. For the second segment, we assume that one more *one* is appended at the first position and the remainder of the segment is filled with *zeros* such that $P(\mathbf{Y}_{[i+1:j]} = 1, 0, \cdots, 0) = \alpha' \times (1-\alpha')^{j-i-1}$. By summing over all i the result is established according to

$$P(X_{[1:j]} = x+1) = \sum_{i=1}^{j-1} P(X_{[1:i]} = x) \times \alpha' \times (1-\alpha')^{j-i-1} \tag{6.6}$$

where we iterate from $x = 1$ to $x = x_{max}$ and from $j = 1$ to $j = N-M+1$. x_{max} denotes the predefined maximum number of hits. The procedure is initialized by computing

$$P(X_{[1:j]} = 0) = (1-\alpha')^{j} \tag{6.7}$$

for $1 \leq j \leq N - M + 1$.

Eventually, the dynamic programming approach establishes exactly the same result as the binomial distribution

$$P(X_{[1:N-M+1]} = x) = \binom{N - M + 1}{x} \alpha'^{x} (1 - \alpha')^{N-M+1-x}$$

in a computationally more demanding way. However, the benefit of the dynamic programming procedure is that it can be extended to incorporate correlation (e.g. by overlapping motif hits) between neighbouring events in the underlying Bernoulli processes $\mathbf{Y}_{[1:N-M+1]}$. An illustration of the dynamic programming procedure in Figure 6.4.

Efficient summation assuming dependent events Y_i in $\mathbf{Y}_{[1:N-M+1]}$ due to overlapping hits

In this Section, we shall extend the dynamic programming algorithm that was introduced in the previous section for the case that the events in $\mathbf{Y}_{[1:j]}$ are correlated due to sliding a motif of length M across the DNA sequence, which might cause overlapping motif hits. Therefore, we make full use of the quantities summarized in Table 6.1.

For the sake of this discussion, we introduce the random variable $X^a_{[1:j]}$ which, similar as in the previous section, denotes the number of *ones* that occurred in the Bernoulli process $\mathbf{Y}_{[1:j]}$, with an additional restriction for the position of the last *one* in the segment. Accordingly, if $1 \leq a < M$, the last hit was observed exactly at position $j - a + 1$ and if $a = M$, the last hit occurred at least $M - 1$ positions before position j. We shall use the indicator a in order to properly account for overlapping motif hits in $\mathbf{Y}_{[1:j]}$.

Figure 6.4: Illustration of the dynamic programming algorithm for $X_{[1:5]} = 2$. The algorithm makes use of the previously determined quantities $P(X_{[1:i]} = 1)$ for $i < 5$ (represented by the gray shaded realizations, that have already been aggregated in the previous step) and multiplies in an additional hit at position $i + 1$ as well as *zeros* for the remainder of the second segment. Therefore, $P(X_{[1:5]} = 2) = P(X_{[1:1]} = 1)P(1000) + P(X_{[1:2]} = 1)P(100) + P(X_{[1:3]} = 1)P(10) + P(X_{[1:4]} = 1)P(1)$ establishes the result. Note that reusing the quantities $P(X_{[1:i]})$, reduces the number of (redundant) operations substantially which results in an efficient algorithm.

To illustrate the use of the superscript a for a motif of length $M = 3$, consider the following associations: $\mathbf{Y}_{[1:5]} = 10000$, $\mathbf{Y}_{[1:5]} = 01000$, $\mathbf{Y}_{[1:5]} = 00100$, $\mathbf{Y}_{[1:5]} = 00010$ and $\mathbf{Y}_{[1:5]} = 00001$ correspond to $a = 3$, $a = 3$, $a = 3$, $a = 2$ and $a = 1$, respectively.

Note that the probabilities $P(X^a_{[1:j]})$ across different j and a are mutually disjoint events, because each realization of $\mathbf{Y}_{[1:j]}$ is attributable only to a single random event $X^a_{[1:j]}$. For example, the associated probabilities for $X^1_{[1:5]} = 2$, $X^2_{[1:5]} = 2$ and $X^3_{[1:5]} = 2$ with a motif of length $M = 3$ (see Figure 6.3) are given by

$$\begin{aligned} P(X^1_{[1:5]} = 2) =& P(\mathbf{Y}_{[1:5]} = 10001) \\ &+ P(\mathbf{Y}_{[1:5]} = 01001) \\ &+ P(\mathbf{Y}_{[1:5]} = 00101) \\ &+ P(\mathbf{Y}_{[1:5]} = 00011) \end{aligned} \tag{6.8}$$

$$\begin{aligned} P(X^2_{[1:5]} = 2) =& P(\mathbf{Y}_{[1:5]} = 10010) \\ &+ P(\mathbf{Y}_{[1:5]} = 01010) \\ &+ P(\mathbf{Y}_{[1:5]} = 00110) \end{aligned} \tag{6.9}$$

$$\begin{aligned} P(X^3_{[1:5]} = 2) =& P(\mathbf{Y}_{[1:5]} = 10100) \\ &+ P(\mathbf{Y}_{[1:5]} = 01100) \\ &+ P(\mathbf{Y}_{[1:5]} = 11000). \end{aligned} \tag{6.10}$$

As a consequence of the disjointness of $X^a_{[1:j]}$ for different j and a, we obtain the total probability of the number of *ones* in $\mathbf{Y}_{[1:j]}$ according to

$$P(X_{[1:j]}) = \sum_a P(X^a_{[1:j]}) \tag{6.11}$$

In the following, we discuss the modification of the dynamic programming algorithm defined by (6.6) such that overlapping motif hits are taken into account. We initialize the algorithm for the case of $x = 1$, for $1 \leq a \leq M$ and j such that $1 \leq j \leq N - M + 1$ according to

$$P(X^a_{[1:j]} = 1) = \begin{cases} (j - M + 1) \times (1 - \alpha')^{j-M} \alpha' \delta_{M-1} & \text{for } a = M \\ (1 - \alpha')^{j-a} \alpha' & \text{ow.} \end{cases} \tag{6.12}$$

Subsequently, the probability of obtaining $x+1$ *ones* in the segment $[1 : j]$ is established by splitting the segment into two parts: $[1 : i]$ and $[i + 1 : j]$ for all $i < j$, respectively. For the first subsegment, we assume that already x *ones* were observed and consequently reuse the probabilities $P(X^a_{[1:i]} = x)$ that were determined in the previous step. We add an additional *one* at the start of the second segment and fill the remaining positions with *zeros* according to $P(\mathbf{Y}_{[i+1:j]} = 1, 0, \cdots, 0)$. As the event $Y_{i+1} = 1$ might be influenced by overlapping hits, we use the indicator a to properly account for overlapping hits at different positions. This leads to the following dynamic programming formula

$$P(X^b_{[1:j]} = x + 1) = \sum_{i=1}^{j-1} \sum_{a=1}^{M} P(X^a_{[1:i]} = x) \cdot h(a) \cdot nh(j - i) \tag{6.13}$$

where

$$b := \begin{cases} j - i & \text{if } j - i < M \\ M & \text{o.w.} \end{cases} \tag{6.14}$$

$$h(a) := \begin{cases} \beta_a & \text{if } a < M \\ \alpha' & \text{o.w.} \end{cases} \tag{6.15}$$

$$nh(a) := \begin{cases} 1 & \text{if } a < M \\ \delta_{M-1} \cdot (1-\alpha')^{a-M} & \text{o.w.} \end{cases} \tag{6.16}$$

$h(a)$ serves for the purpose of multiplying in an additional hit at position $i+1$. Note that if $a < M$, we append an overlapping hit using the *principal overlapping hit probability* defined by (3.41), whereas, in case of $a = M$, a *clump start hit* defined by (6.1) is multiplied in. $nh(a)$ is used to fill the remaining $j - i - 1$ positions in the second subsegment with *zeros* if $a \geq M$. Otherwise, the remaining *zeros* are accounted for at the termination step.

Finally, the dynamic programming algorithm terminates by

$$P(X_{[1:N-M+1]} = x) = P(X^M_{[1:N-M+1]} = x) + \sum_{a=1}^{M-1} P(X^a_{[1:N-M+1]} = x)\delta_{a-1} \tag{6.17}$$

The probability for obtaining only *zeros*, given by $P(X_{[1:N-M+1]} = 0)$, is computed outside of the dynamic programming routine by $P(X_{[1:N-M+1]} = 0) = (1-\alpha')^{N-M+1}$, which concludes the derivation of the *combinatorial model* for the case of scanning a single DNA strand and establishes the desired distribution of the number of motif hits $P(X_{[1:N-M+1]})$.

6.2 The combinatorial model for scanning a both DNA strands

In this Section, we shall extend the combinatorial model for the purpose of modeling the number of motif hits that are obtained by scanning both DNA strands. Con-

sequently, the Bernoulli process $\mathbf{Y}_{[1:N-M+1]}$ which we study in this section becomes $Y_1 Y_2 \cdots Y_{N-M+1} Y'_1 Y'_2 \cdots Y'_{N-M+1}$.

6.2.1 Factorization of $P(\mathbf{Y}_{[1:N-M+1]})$

As we have described in Section 1.2.4, $P(\mathbf{Y}_{[1:N-M+1]})$ is obtained by averaging over all underlying DNA sequences. However, this is generally too difficult to compute. Even though, the events in $\mathbf{Y}_{[1:N-M+1]}$ may exhibit complicated statistical associations, we can still approximately factorize $P(\mathbf{Y}_{[1:N-M+1]})$ for a given realization based on a number of independence assumptions. Analogously to the previous section, we exploit 1) the independence assertions defined by (3.3)-(3.8), 2) the assumption that non-overlapping events $Y_i \perp\!\!\!\perp Y_j$, $Y'_i \perp\!\!\!\perp Y_j$ and $Y'_i \perp\!\!\!\perp Y'_j$ such that $j - M + 1 \leq i \leq j + M - 1$ are independent and 3) the assumption that prior to the start of the Bernoulli process only *zeros* were observed (e.g $\mathbf{Y}_{[0:-M+2]} = \mathbf{0}$). According to the last assumption, the *clump start probabilities* $P(n \to h_f)$ and $P(n \to h_r)$ that were derived using the Markov model in Chapter 5, remain constant throught the sequence. For convenience we shall denote the *clump start probability* with respect to the forward strand by

$$\begin{aligned} \alpha' &:= P(n \to h_f) \\ &= P(Y_0 = 1 | Y_{-1} = 0, \cdots, Y_{-M+1} = 0, Y'_{-1} = 0, \cdots, Y'_{-M+1} = 0). \end{aligned} \tag{6.18}$$

Recall from the previous chapter that the *clump start probability* with respect to the reverse strand is given by $P(n \to h_r) = \alpha'(1 - \beta_{3',0})$.

A summary of all quantities that are facilitated to factorize $P(\mathbf{Y}_{[1:N-M+1]})$ is given in Table 6.2.

Table 6.2: List of probabilities which are used to factorize $P(\mathbf{Y}_{[1:N-M+1]})$ when both DNA strands are scanned for motif hits.

Probability	Abbreviation	Evaluation
$P(Y_0 = 1 \mid \mathbf{Y}_{[-1:-M+1]} = \mathbf{0})$	α'	see Definition (6.18)
$P(Y'_0 = 1, Y_0 = 0 \mid \mathbf{Y}_{[-1:-M+1]} = \mathbf{0})$	$\alpha'(1-\beta_{3',0})$	see Definition (6.18) and (3.43)
$P(Y'_0 = 0, Y_0 = 0 \mid \mathbf{Y}_{[-1:-M+1]} = \mathbf{0})$	$1-\alpha'(2-\beta_{3',0})$	
$P(Y_k = 1, \mathbf{Y}_{[1:k-1]} = \mathbf{0}, Y'_0 = 0 \mid Y_0 = 1)$	β_k	see Definition (3.42)
$P(Y'_k = 1, \mathbf{Y}_{[1:k-1]} = \mathbf{0}, Y'_0 = 0 \mid Y_0 = 1)$	$\beta_{3',k}$	see Definition (3.43)-(3.44)
$P(Y_k = 1, \mathbf{Y}_{[1:k-1]} = \mathbf{0} \mid Y'_0 = 1)$	$\beta_{5',k}$	see Definition (3.45)
$P(\mathbf{Y}_{[1:k]} = \mathbf{0}, Y'_0 = 0 \mid Y_0 = 1)$	δ_k	see Definition (3.51)
$P(\mathbf{Y}_{[1:k]} = \mathbf{0} \mid Y'_0 = 1)$	δ'_k	see Definition (3.52)

6.2.2 Efficient summation over combinations of placing x motif hits

In order to evaluate the probability of obtaining x hits in the Bernoulli process $\mathbf{Y}_{[1:N-M+1]}$, we extend the *combinatorial model* from the previous section, where a single DNA strand was scanned for motif hits, to the case where both DNA strands are scanned. To this end, in addition to overlapping hits (on the same strand), we also need to take overlapping hits due to base pair complementarity into account. That is, a forward strand hit may also be overlapped by reverse strand hit and vice versa.

We start by defining the random variables $X^a_{[1:j]}$ and $X'^a_{[1:j]}$, which, similarly to the previous section, denote the number of hits in the segment $\mathbf{Y}_{[1:j]}$ of the Bernoulli process. The indicator a has the same interpretation as described above. Namely, for $a < M$ the last hit in the segment occurred at position $j - a + 1$ and for $a = M$, the last hit occurred at least $M - 1$ position upstream of j. The additional refinement concerns the absence or presence of the prime: $X^a_{1:j]}$ denotes that the last hit occurred on the forward strand, while, $X'^a_{[1:j]}$ denotes that the last hit occurred on the reverse strand. Therefore, using $X^a_{1:j]}$ and $X'^a_{[1:j]}$ we can properly account the strandedness of hits. An example the relationship between realizations of $\mathbf{Y}_{[1:4]}$ as well as the corresponding indicators a and

Table 6.3: Illustration of the use of the indicator variable a and the presence or abscence of the prime. On the left, different realizations of $\mathbf{Y}_{[1:4]}$ are shown. The corresponding indicator variable a and prime are shown in column two and three.

$\mathbf{Y}_{[1:4]}$	indicator a	prime present?
$\begin{pmatrix} Y'_1=0 & Y'_2=0 & Y'_3=0 & Y'_4=1 \\ Y_1=0 & Y_2=0 & Y_3=0 & Y_4=0 \end{pmatrix}$	1	yes
$\begin{pmatrix} Y'_1=0 & Y'_2=0 & Y'_3=1 & Y'_4=0 \\ Y_1=0 & Y_2=0 & Y_3=0 & Y_4=0 \end{pmatrix}$	2	yes
$\begin{pmatrix} Y'_1=0 & Y'_2=1 & Y'_3=0 & Y'_4=0 \\ Y_1=0 & Y_2=0 & Y_3=0 & Y_4=0 \end{pmatrix}$	3	yes
$\begin{pmatrix} Y'_1=1 & Y'_2=0 & Y'_3=0 & Y'_4=0 \\ Y_1=0 & Y_2=0 & Y_3=0 & Y_4=0 \end{pmatrix}$	3	yes
$\begin{pmatrix} Y'_1=0 & Y'_2=0 & Y'_3=0 & Y'_4=0 \\ Y_1=0 & Y_2=0 & Y_3=0 & Y_4=1 \end{pmatrix}$	1	no
$\begin{pmatrix} Y'_1=0 & Y'_2=0 & Y'_3=0 & Y'_4=0 \\ Y_1=0 & Y_2=0 & Y_3=1 & Y_4=0 \end{pmatrix}$	2	no
$\begin{pmatrix} Y'_1=0 & Y'_2=0 & Y'_3=0 & Y'_4=0 \\ Y_1=0 & Y_2=1 & Y_3=0 & Y_4=0 \end{pmatrix}$	3	no
$\begin{pmatrix} Y'_1=0 & Y'_2=0 & Y'_3=0 & Y'_4=0 \\ Y_1=1 & Y_2=0 & Y_3=0 & Y_4=0 \end{pmatrix}$	3	no

the prime symbol is presented in Table 6.3.

Since, each realization of $\mathbf{Y}_{[1:j]}$ is uniquely maps to a single event $X^a_{[1:j]}$ and $X'^a_{[1:j]}$ (with a specific a and prime), $X^a_{[1:j]}$ and $X'^a_{[1:j]}$ represent mutually disjoint events. Thus, we obtain the total probability of the number of hits in the segment $\mathbf{Y}_{[1:j]}$ by computing

$$P(X_{[1:j]}) = \sum_{a=1}^{M} P(X^a_{[1:j]}) + P(X'^a_{[1:j]}). \tag{6.19}$$

Next, we discuss the dynamic programming algorithm for computing the probability of obtaining x hits in the process $\mathbf{Y}_{[1:j]}$ when both strands of the DNA are scanned for motif hits.

We initialize the algorithm by computing the probability of obtaining $x = 1$ hit for each

a and for $1 \leq j \leq N - M + 1$ according to

$$P(X^a_{[1:j]} = 1) = \begin{cases} (j - M + 1) \times (1 - \alpha'(2 - \beta_{3',0}))^{j-M} \alpha' \delta_{M-1} & \text{for } a = M \\ (1 - \alpha'(2 - \beta_{3',0}))^{j-a} \alpha' & \text{ow.} \end{cases}$$

$$P(X'^a_{[1:j]} = 1) = \begin{cases} (j - M + 1) \times (1 - \alpha'(2 - \beta_{3',0}))^{j-M} \alpha'(1 - \beta_{3',0}) \delta'_{M-1} & \text{for } a = M \\ (1 - \alpha'(2 - \beta_{3',0}))^{j-a} \alpha'(1 - \beta_{3',0}) & \text{ow.} \end{cases}$$

where we made use of the lines 1, 2, 3, 7 and 8 in Table 6.2.

Subsequently, we proceed by computing the probabilities $P(X^a_{[1:j]} = x+1)$ and $P(X'^a_{[1:j]} = x + 1)$ by utilizing the quantities $P(X^a_{[1:i]} = x)$ and $P(X'^a_{[1:i]} = x)$ for $i < j$, which were determined in the previously. We append another motif hit at position $j+1$ (one on each strand, separately) and fill the remaining positions $j - i - 1$ with *zeros*. This leads us to the following dynamic programming equation for $x + 1$:

$$\begin{aligned} P(X^b_{[1:j]} = x + 1) = &\sum_{i=1}^{j-1} \sum_{a=1}^{M} P(X^a_{[1:i]} = x) \cdot h(a) \cdot nh(j - i) \\ &+ \sum_{i=1}^{j-1} \sum_{a=1}^{M} P(X'^a_{[1:i]} = x) \cdot h_{5'}(a) \cdot nh(j - i) \qquad (6.20) \\ P(X'^b_{[1:j]} = x + 1) = &\sum_{i=1}^{j-1} \sum_{a=1}^{M} P(X^a_{[1:i]} = x) \cdot h_{3'}(a) \cdot nh'(j - i) \\ &+ \sum_{i=1}^{j-1} \sum_{a=1}^{M} P(X'^a_{[1:i]} = x) \cdot h'(a) \cdot nh'(j - i) \\ &+ \sum_{i=1}^{j} P(X^1_{[1:i]} = x) \cdot h_{3'}(0) \cdot nh'(j - i + 1) \qquad (6.21) \end{aligned}$$

where we iterate in an ordered fashion from $x = 1$ to $x = x_{max}$ and determine the equations for $1 \leq j \leq N - M + 1$. Equation (6.20) and (6.21) make use of the following

auxiliary terms

$$b := \begin{cases} j - i & \text{if } j - i < M \\ M & \text{o.w.} \end{cases} \tag{6.22}$$

$$h(a) := \begin{cases} \beta_a & \text{if } a < M \\ \alpha' & \text{o.w.} \end{cases} \tag{6.23}$$

$$h'(a) := \begin{cases} \beta_a & \text{if } a < M \\ \alpha'(1 - \beta_{3',0}) & \text{o.w.} \end{cases} \tag{6.24}$$

$$h_{3'}(a) := \begin{cases} \beta_{3',a} & \text{if } a < M \\ \alpha' \cdot (1 - \beta_{3',0}) & \text{o.w.} \end{cases} \tag{6.25}$$

$$h_{5'}(a) := \begin{cases} \beta_{5',a} & \text{if } a < M \\ \alpha' & \text{o.w.} \end{cases} \tag{6.26}$$

$$nh(a) := \begin{cases} 1 & \text{if } a < M \\ \delta_{M-1} \cdot \Omega^{a-M} & \text{o.w.} \end{cases} \tag{6.27}$$

$$nh'(a) := \begin{cases} 1 & \text{if } a < M \\ \delta'_{M-1} \cdot \Omega^{a-M} & \text{o.w.} \end{cases} \tag{6.28}$$

that are defined in terms of the factors summarized in Table 6.2.

The first and second summation on the right hand side of Equation (6.20) account for a previous forward and reverse strand hit, respectively, that lead to the a forward strand hit $Y_{i+1} = 1$. Likewise, the first and second summation on the right hand side of Equation (6.21) account for a previous forward and reverse strand hit, respectively, that is followed by a reverse strand hit $Y'_{i+1} = 1$. Additionally, the third summation on the right hand side of Equation (6.21) accounts for palindromic motif hits.

The dynamic programming algorithm terminates by computing

$$P(X_{[1:N-M+1]} = x) = P(X^{M}_{[1:N-M+1]} = x) + P(X'^{M}_{[1:N-M+1]} = x) + \sum_{a=1}^{M-1} P(X^{a}_{[1:N-M+1]} = x)\delta_{a-1} + \sum_{a=1}^{M-1} P(X'^{a}_{[1:N-M+1]} = x)\delta'_{a-1} \quad (6.29)$$

Finally, the probability of observing only *zeros* is given by $P(X_{[1:N-M+1]} = 0) = (1 - \alpha'(2 - \beta_{3',0}))^{N-M+1}$. This concludes our derivation of the *combinatorial model* for the case of scanning both DNA sequences for motif hits and establishes the desired approximative distribution $P(X_{[1:N-M+1]})$.

6.3 Runtime of the combinatorial model

The asymptotic runtime of the combinatorial model is given by $O(x_{max}(N - M + 1)^2 M)$, where x_{max} denotes the maximum number of hits after which the distribution is truncated and N denotes the length of the DNA sequence and M denotes the length of the TF motif.

Typical values for N, M and x_{max} are $N = 200$, $M = 15$ and $x_{max} = 30$. Thus, since $N \gg M$ and $N \gg x_{max}$, in practice, the primary determinant of the runtime is the DNA sequence length N.

6.4 Combinatorial model across multiple distinct DNA sequences

In many cases, it is of interest to determine the distribution of the number of motif hits across multiple distinct pieces of DNA. For instance, the number of motif hits across a set of non-overlapping enhancer sequences.

Assuming that the individual DNA sequences are of equal length N and mutually non-overlapping, we can determine the distribution of the number of hits across S sequences of length N. For this purpose, we could determine $P(X_{[1:N-M+1]})$ by means of the dynamic programming procedure once and repeatedly employ the convolution operation in order to add up the numbers of motif hits across S sequences using

$$P^S(X) = P_1(X_{[1:N-M+1]}) * P_2(X_{[1:N-M+1]}) * \cdots P_S(X_{[1:N-M+1]}) \tag{6.30}$$

where $P_i(X_{[1:N-M+1]}) = P(X_{[1:N-M+1]})$ for each $1 \leq i \leq S$ and the superscript S denotes the distribution of the number of hits across S sequences.

We can further optimize the runtime for computing $P^S(X)$ by reusing intermediately derived distributions. To this end, recursively compute $P(X)^S$ where in each recursive step one of three operations are performed: 1) if $i > 1$ and $P(X)^i$ was not evaluated previously, the following operation is carried out

$$P(X)^i = P(X)^{\lfloor i/2 \rfloor} * P(X)^{\lceil i/2 \rceil} \tag{6.31}$$

where $*$ denotes the convolution operation, $\lfloor\rfloor$ rounds down to the closest integer and

$\lceil\rceil$ rounds up to the closest integer. The resulting distribution $P^i(X)$ is then stored for potential reuse. 2) if $P(X)^i$ was already determined previously, it is directly returned. 3) if $i = 1$, the base case $P^1(X) = P(X_{[1:N-M+1]})$ is returned.

The asymptotic runtime for this recursive aggregation procedure is given by $O(x_{max} \log(S))$.

6.5 Results

In this Section, we compare the combinatorial model for the distribution of the number of motif hits on both strands of the DNA against the novel compound Poisson model that we have discussed in Chapter 4 and a simple binomial model, which we denote by $P_{DP}(X)$, $P^N_{CP}(X)$ and $P_{Bin}(X)$, respectively. An empirical motif hits count distribution, denoted by $P_E(X)$, is determined according to the description in Chapter 4, which serves as a reference for the comparison. We compared the models for a range of different scenarios, including different background model orders, different score thresholds, different numbers and lengths of the individual sequences, and different motifs (see Figure 2.5 and Table 6.4). We estimated the background models on a subset of DNase-I-hypersensitive site that were detected in the ENCODE project [50].

In order to measure the discrepancy between $P_E(X)$ and $P_{DP}(X)$, $P^N_{CP}(X)$ and $P_{Bin}(X)$, we evaluate the total variation distance between the two distributions across the entire range of motif hits counts x as well as only for counts that are beyond the 95%-tile with respect to $P_E(X)$ (see Equation (4.35) and (4.36)). The latter measure focuses on potential biases on the right tail of the distribution which may affect the *motif hit enrichment* test.

Table 6.4: Summary of settings across which the combinatorial model, the compound Poisson model and the binomial model were compared against. In addition to that we tested each setting for order-0 and order-1 background and several motifs (see Figure 2.5).

Seq. length	Num. of seqs.	α
100	2	0.05
100	10	0.01
100	100	0.001
100	1000	0.0001
500	2	0.01
500	20	0.001
500	200	0.0001

As described in the results of Chapter 4, the motif structure dictates the shape of the motif hits count distribution. For instance, palindromic motifs give rise to a motif hits count distribution in which odd numbers of motif hits are highly inprobable (or even impossible), repetitive motifs exhibit an increased variance due to the clumping effect and non-self-overlapping motifs approximately bring about normally distributed motif hit counts. As expected, both $P_{DP}(X)$ and $P^{N}_{CP}(X)$ in principle support all motif structure types, because both account for self-overlapping motif hits, while $P_{Bin}(X)$ only represents non-self-overlapping motifs adequately (see Figures 6.5 - 6.12). In particular, for stringent significance levels, e.g. $\alpha = 10^{-4}$, we observe a high agreement between the results computed according to $P^{N}_{CP}(X)$ and $P_{DP}(X)$ across all motif structures (see Tables 6.5 for $\alpha = 10^{-4}$). As the score threshold t_s for calling motif hits is relaxed (e.g. to $\alpha = 0.05$), the compound Poisson model incurs biases due to the violation of the "rare hit" assumption, which is expected. This phenomenon is evident across all motif structure types and leads to an overestimation of the variance of $P^{N}_{CP}(X)$ with respect to $P_{E}(X)$ (see Tables 6.5 for $\alpha = 0.05$).

By contrast, for low-stringency score threshold t_s, $P_{DP}(X)$ still yields highly accurate

results across all motif structures, because the combinatorial model does not explicitly rely on the Poisson assumption (see Tables 6.5-6.6; compare $P_{DP}(X)$, $P^{N}_{CP}(X)$ and $P_{Bin}(X)$ for $\alpha = 0.05$ and $\alpha = 0.01$). Therefore, the combinatorial model seems to be particularly well suited for low-stringency score thresholds.

Next, we focus on the accuracy of $P_{DP}(X)$ across different background model orders $d \in \{0, 1\}$ to exclude the possibility that biases are incurred due to the simplifying assumptions that were made for dealing with higher-order background models (see Section 1.2.4). We found that the combinatorial model obtains accurate results across different background model orders, suggesting that biases due to the simplifying assumptions are negligible (see Table 6.5 and 6.6; compare different d in column 2).

We also measured the accuracy for different DNA sequence lengths (100 bp and 500 bp) to rule out any biases that are associated with the finite sequence length (see Table 6.4). We found that the differences in the sequence lengths do not diminish the accuracy of the combinatorial model significantly (see Table 6.5).

Next, we measured the discrepancy between the $P_{DP}(X), P^{N}_{CP}(X), P_{Bin}(X)$ and $P_{E}(X)$ using Equation (4.36) (as an alternative to using Equation (4.35)) to assess any biases that are attributable to the right tail of the distribution, because, even though the entire distribution might be fairly accurate, it is still possible that it is inaccurate with respect to the right tail of the distribution, which, however, affects the *motif hit enrichment* test. We found that both measures defined by (4.35) and (4.36) yield highly concordant results for all approximative distributions (see Table 6.5 and 6.6). That is, when the entire distribution is accurate, the right tail is also captured accurately.

Finally, we want to stress that in most cases $P_{DP}(X)$ obtains more accurate results com-

pared to $P^N_{CP}(X)$ as well as $P_{Bin}(X)$ (see bold entries in Tables 6.5 and 6.6). For cases, for which either $P_{CP}(X)$ or $P_{Bin}(X)$ achieved advantageous results, $P_{DP}(X)$ produces highly accurate results as well.

Table 6.5: Distances between the empirical distribution and the analytical models for the number of motif hits over the entire distribution measured by Equation (4.35). Bold values mark the most accurate result in each row.

Motif	d	α	Seqlen	$d(P_E, P^N_{CP})$	$d(P_E, P_{DP})$	$d(P_E, P_{Bin})$
E47	0	0.05	100	0.848	**0.151**	0.408
E47	0	0.01	100	0.224	**0.0318**	0.403
E47	0	0.001	100	0.0244	**0.0105**	0.403
E47	0	1×10^{-4}	100	0.00987	**0.00922**	0.402
E47	0	0.01	500	0.221	**0.0356**	0.116
E47	0	0.001	500	0.0301	**0.01**	0.0879
E47	0	1×10^{-4}	500	0.012	**0.0112**	0.0766
E47	1	0.05	100	0.864	**0.146**	0.382
E47	1	0.01	100	0.229	**0.019**	0.385
E47	1	0.001	100	0.0336	**0.0175**	0.382
E47	1	0.01	500	0.226	**0.0166**	0.0912
E47	1	0.001	500	0.0396	**0.0235**	0.0663
SP1SP3	0	0.05	100	0.619	**0.141**	0.26
SP1SP3	0	0.01	100	0.16	**0.0299**	0.227
SP1SP3	0	0.001	100	0.0266	**0.013**	0.216
SP1SP3	0	1×10^{-4}	100	0.012	**0.0119**	0.215
SP1SP3	0	0.01	500	0.162	**0.0321**	0.0803
SP1SP3	0	0.001	500	0.0231	**0.00938**	0.0587

Table 6.5: (continued)

SP1SP3	0	1×10^{-4}	500	0.0128	**0.0126**	0.0395
SP1SP3	1	0.05	100	0.59	**0.121**	0.268
SP1SP3	1	0.01	100	0.151	**0.0264**	0.233
SP1SP3	1	0.001	100	0.0229	**0.0129**	0.214
SP1SP3	1	0.01	500	0.148	**0.0303**	0.119
SP1SP3	1	0.001	500	0.016	**0.0153**	0.0599
Pal	0	0.05	100	0.447	**0.0609**	1
Pal	0	0.01	100	0.112	**0.0154**	1
Pal	0	0.001	100	0.0145	**0.0125**	1
Pal	0	0.01	500	0.0984	**0.0118**	1
Pal	0	0.001	500	0.0119	**0.00827**	1
Pal	1	0.05	100	0.471	**0.0715**	1
Pal	1	0.01	100	0.108	**0.0127**	1
Pal	1	0.001	100	0.0179	**0.0157**	1
Pal	1	0.01	500	0.0991	**0.0168**	1
Pal	1	0.001	500	0.013	**0.00765**	1
Rep	0	0.05	100	0.319	**0.0308**	0.835
Rep	0	0.01	100	0.0916	**0.0333**	0.76
Rep	0	0.001	100	**0.0211**	0.023	0.628
Rep	0	0.01	500	0.0794	**0.0206**	0.749
Rep	0	0.001	500	**0.0181**	**0.0181**	0.613
Rep	1	0.05	100	0.325	**0.0286**	0.836
Rep	1	0.01	100	0.0893	**0.0332**	0.748

Table 6.5: (continued)

Rep	1	0.001	100	**0.0314**	0.0356	0.643
Rep	1	0.01	500	0.0829	**0.0172**	0.73
Rep	1	0.001	500	0.0231	**0.0179**	0.611

Table 6.6: Distances between the empirical distribution and the analytical models for the number of motif hits over the 5% significance region of the distribution measured by Equation (4.36). Bold values mark the most accurate result in each row.

Motif	d	α	Seqlen	$d_{5\%}(P_E, P^N_{CP})$	$d_{5\%}(P_E, P_{DP})$	$d_{5\%}(P_E, P_{Bin})$
E47	0	0.05	100	0.125	**0.0279**	0.0449
E47	0	0.01	100	0.0353	**0.00432**	0.0629
E47	0	0.001	100	0.00435	**0.00207**	0.0871
E47	0	1×10^{-4}	100	0.00145	**0.00138**	0.0876
E47	0	0.01	500	0.0341	**0.00388**	0.00587
E47	0	0.001	500	0.0046	**0.00233**	0.00574
E47	0	1×10^{-4}	500	**0.00157**	**0.00157**	0.00793
E47	1	0.05	100	0.126	**0.028**	0.0494
E47	1	0.01	100	0.0306	**0.00131**	0.0544
E47	1	0.001	100	**0.00212**	0.00279	0.0797
E47	1	1×10^{-4}	100	**0.00209**	0.00235	0.0928
E47	1	0.01	500	0.0323	**0.00226**	0.0058
E47	1	0.001	500	**0.00168**	0.00273	0.00315
E47	1	1×10^{-4}	500	**0.00193**	0.00215	0.00764
SP1SP3	0	0.05	100	0.0969	0.0286	**0.0119**
SP1SP3	0	0.01	100	0.0248	**0.00488**	0.0209

Table 6.6: (continued)

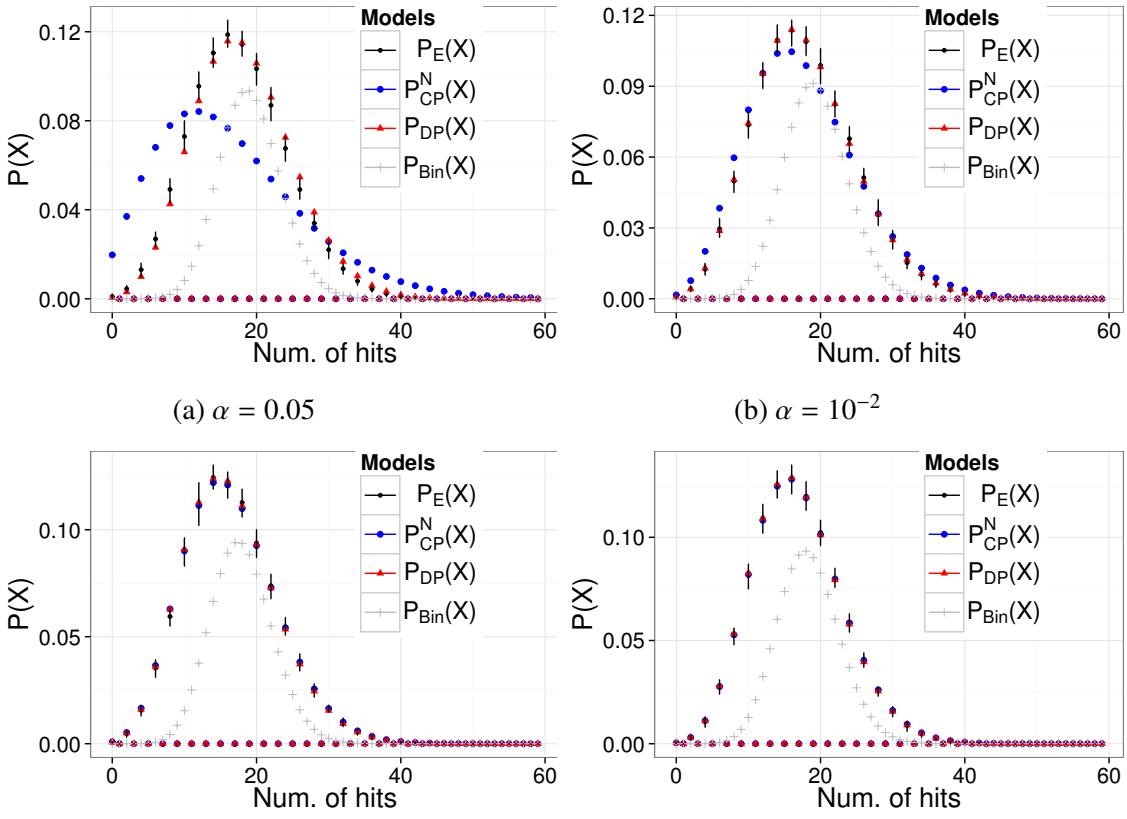
SP1SP3	0	0.001	100	0.00272	**0.000861**	0.0283
SP1SP3	0	1×10^{-4}	100	**0.00191**	0.00198	0.0445
SP1SP3	0	0.01	500	0.0269	0.00704	**0.00632**
SP1SP3	0	0.001	500	0.00376	**0.00205**	0.0041
SP1SP3	0	1×10^{-4}	500	**0.00155**	0.00159	0.00553
SP1SP3	1	0.05	100	0.0911	0.0257	**0.00975**
SP1SP3	1	0.01	100	0.0225	**0.00421**	0.0119
SP1SP3	1	0.001	100	0.00389	**0.0022**	0.0384
SP1SP3	1	1×10^{-4}	100	0.00775	**0.00746**	0.0552
SP1SP3	1	0.01	500	0.0244	**0.00636**	0.0143
SP1SP3	1	0.001	500	0.00284	**0.00192**	0.00282
SP1SP3	1	1×10^{-4}	500	0.00811	**0.00786**	0.0148
Pal	0	0.05	100	0.0583	**0.0124**	0.0572
Pal	0	0.01	100	0.0152	**0.00273**	0.0486
Pal	0	0.001	100	0.00287	**0.00271**	0.0756
Pal	0	1×10^{-4}	100	**0.0012**	0.00209	0.054
Pal	0	0.01	500	0.0129	**0.00342**	0.0551
Pal	0	0.001	500	0.00241	**0.0013**	0.0711
Pal	0	1×10^{-4}	500	**0.00129**	0.00132	0.0532
Pal	1	0.05	100	0.0627	**0.0135**	0.0496
Pal	1	0.01	100	0.0168	**0.00333**	0.0621
Pal	1	0.001	100	0.00244	**0.0021**	0.0591
Pal	1	1×10^{-4}	100	0.00246	**0.00128**	0.0553

Table 6.6: (continued)

Pal	1	0.01	500	0.0143	**0.0044**	0.0707
Pal	1	0.001	500	0.00291	**0.00211**	0.0637
Pal	1	1×10^{-4}	500	0.00204	**0.0018**	0.0603
Rep	0	0.05	100	0.032	**0.00745**	0.0556
Rep	0	0.01	100	0.0124	**0.00269**	0.0532
Rep	0	0.001	100	0.0039	**0.00214**	0.0495
Rep	0	1×10^{-4}	100	0.00265	**0.00185**	0.0497
Rep	0	0.01	500	0.00871	**0.00403**	0.0512
Rep	0	0.001	500	0.00374	**0.00355**	0.0571
Rep	0	1×10^{-4}	500	0.00244	**0.00222**	0.0481
Rep	1	0.05	100	0.0306	**0.00618**	0.0564
Rep	1	0.01	100	0.0107	**0.00346**	0.0564
Rep	1	0.001	100	0.0036	**0.00323**	0.0513
Rep	1	0.01	500	0.00823	**0.00306**	0.0535
Rep	1	0.001	500	0.00364	**0.00247**	0.0572
Rep	1	1×10^{-4}	500	0.00289	**0.0026**	0.0472

6.6 Discussion

In this Chapter, we have introduced a novel analytic model to describe the number of motif hits, which we termed the *combinatorial model*. The model builds on the results of Chapter 3 and 5 in order to compute the probability of obtaining x motif hits in a



(a) $\alpha = 0.05$

(b) $\alpha = 10^{-2}$

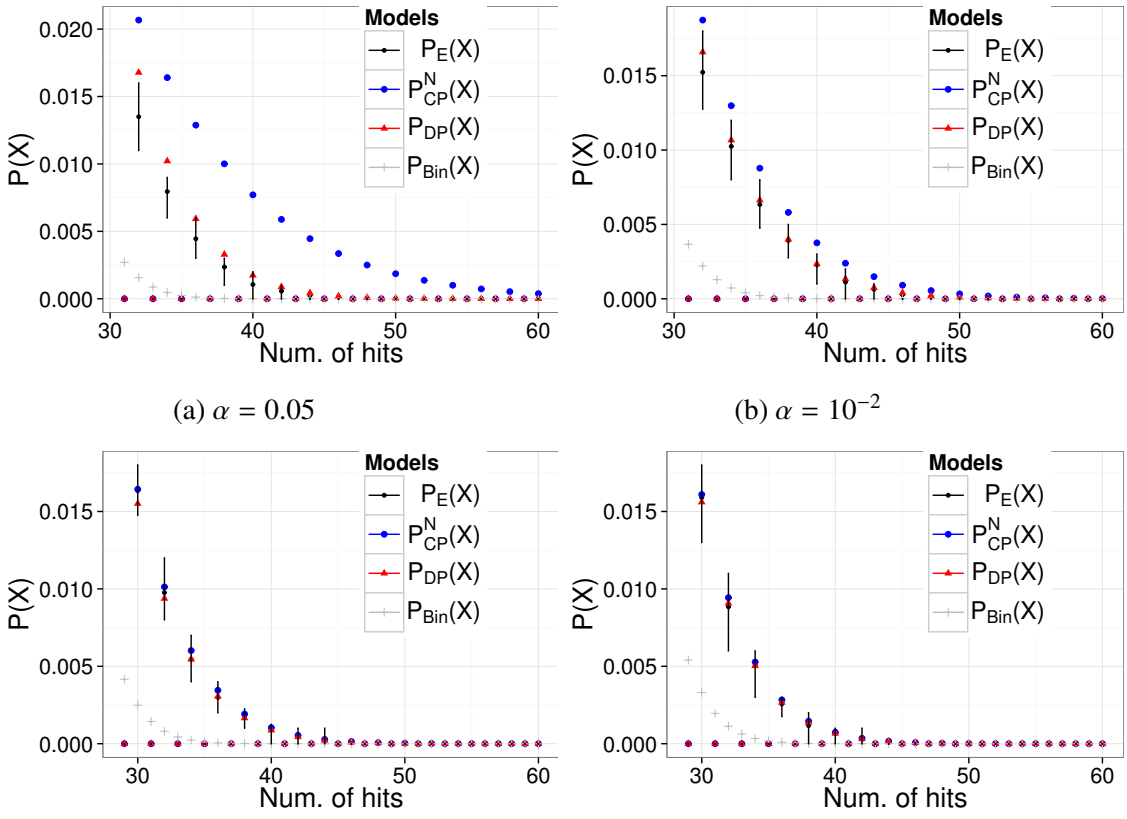
(c) $\alpha = 10^{-3}$

(d) $\alpha = 10^{-4}$

Figure 6.5: Motif hits count distribution for the palindromic motif (see Figure (2.5a)). For each panel, we used an order-1 background model.

finite-length DNA sequence of length N. To this end, we sum up the probabilities of all realization of $\mathbf{Y}_{[1:N-M+1]}$ that emit x hits using dynamic programming. We developed the *combinatorial model* for didactical reasons by first assuming i.i.d. events in $\mathbf{Y}_{[1:N-M+1]}$ when a single DNA strand is scanned for motif hits. Then, we extended the algorithm such that overlapping motif hits are taken into account. Finally, we extended the algorithm to take overlapping hits on both strands of the DNA sequence into account.

The key advantage of the *combinatorial model* is that it does not rely on the 'rare hit' assumption as is the case for the compound Poisson model. Therefore, the combi-



(a) $\alpha = 0.05$

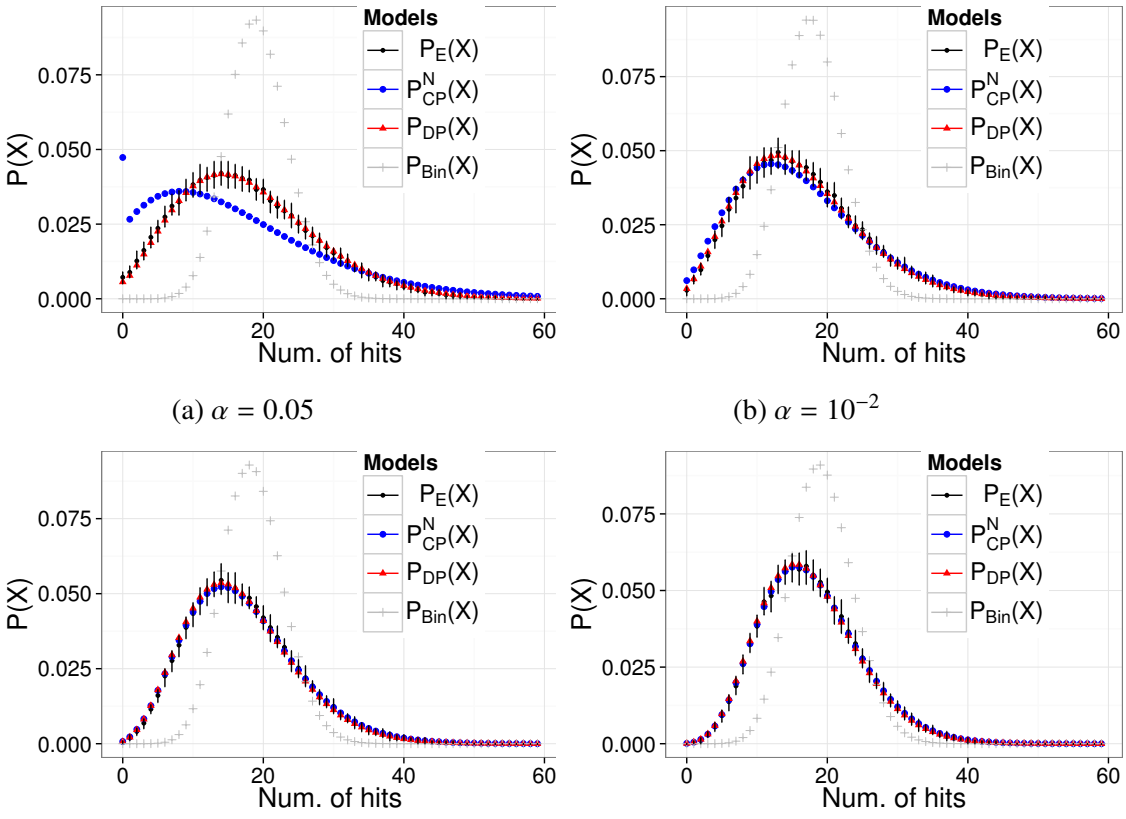
(b) $\alpha = 10^{-2}$

(c) $\alpha = 10^{-3}$

(d) $\alpha = 10^{-4}$

Figure 6.6: 5% significance region of the motif hits count distribution for the palindromic motif (see Figure (2.5a)). For each panel, we used an order-1 background model.

natorial model can be used with low-stringency score thresholds (e.g. $\alpha = 0.05$) for which the compound Poisson model becomes significantly biased. While, the combinatorial model is based on the context-specific independence Assumptions (3.3) - (3.8) which are violated for low stringency score thresholds as well, we found that the Assumptions (3.3) - (3.8) induce only insignificant biases for reasonable choices of α (e.g. $\alpha \leq 0.05$). Only for extremely low score thresholds (e.g. $\alpha > 0.1$), we indeed observe significant biases due to that violated assumption. However, this regime is not of interest for us as we expect motif hits to be scarcely distributed across the regulatory regions,



(a) $\alpha = 0.05$

(b) $\alpha = 10^{-2}$

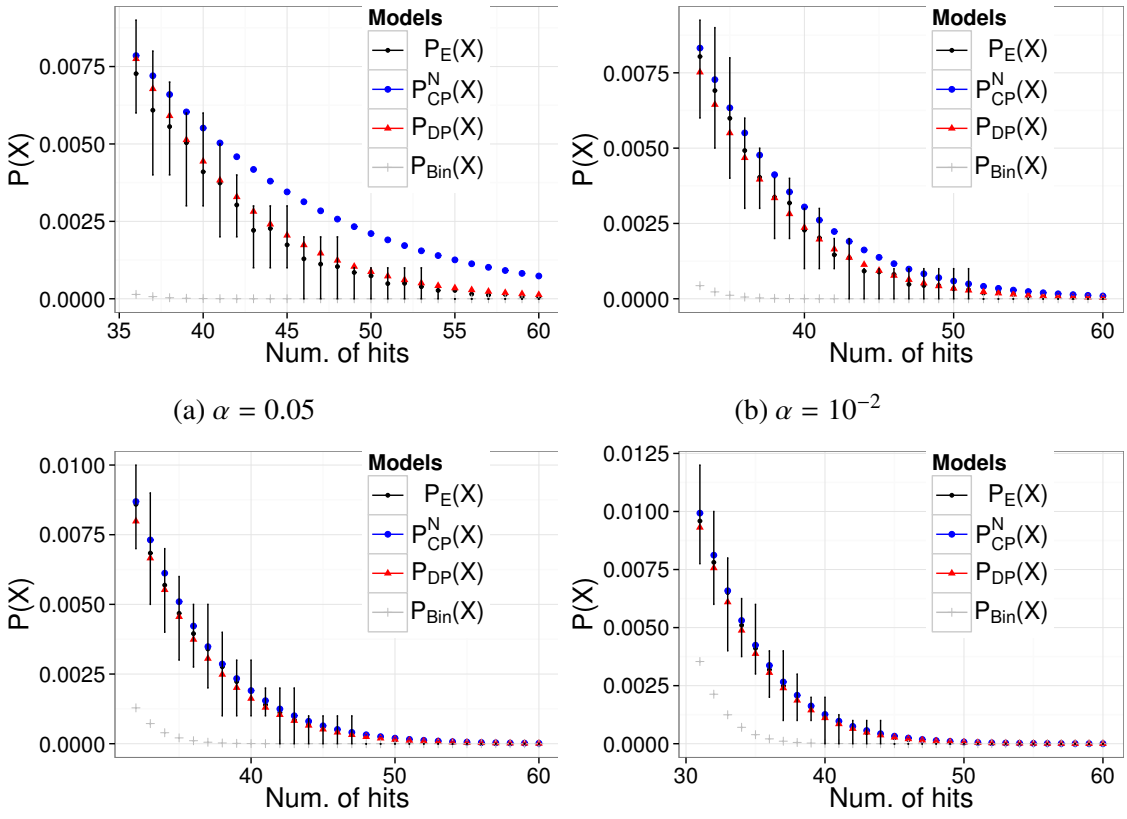
(c) $\alpha = 10^{-3}$

(d) $\alpha = 10^{-4}$

Figure 6.7: Motif hits count distribution for the repetitive motif (see Figure (2.5b)). For each panel, we used an order-1 background model.

but with $\alpha = 0.1$ every tenth base would be called a putative TFBS on a single DNA strand, and every approximately every fifth base would be called a TFBS when both DNA strands are considered. Therefore, essentially every nucleotide would be bound by a TF, which is biologically implausible.

Besides achieving accurate results for low-stringency score thresholds, the combinatorial model also yields accurate results for high-stringency score thresholds, e.g. $\alpha = \{0.001, 0.0001\}$, across a range of motifs (see Tables 6.5 and 6.6). In this regime, the combinatorial model and the compound Poisson model yield highly concordant results



(a) $\alpha = 0.05$

(b) $\alpha = 10^{-2}$

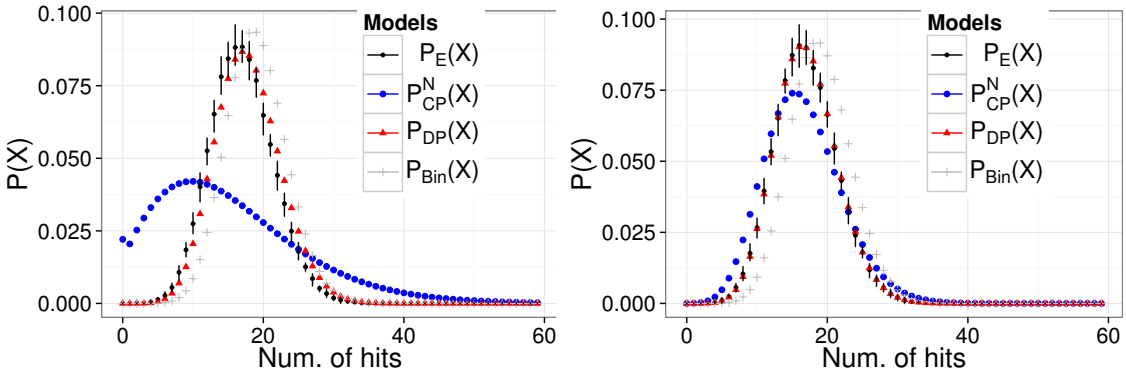
(c) $\alpha = 10^{-3}$

(d) $\alpha = 10^{-4}$

Figure 6.8: 5% significance region of the motif hits count distribution for the repetitive motif (see Figure (2.5b)). For each panel, we used an order-1 background model.

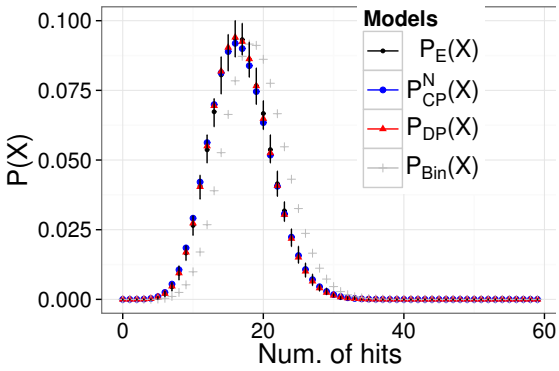
across the motif structures.

While the combinatorial model yields highly accurate results across a wider range of score thresholds and motif structures, its evaluation is tied to a higher computational cost compared to the compound Poisson model, because the runtime depends quadratically on the DNA sequence length which is scanned for motif hits. Therefore, the compound Poisson model is still advantageous when scanning long contiguous stretches of DNA (e.g. $>$ 1kb) for motif hits, in this regime, the compound Poisson model produces accurate results with less computational efford.

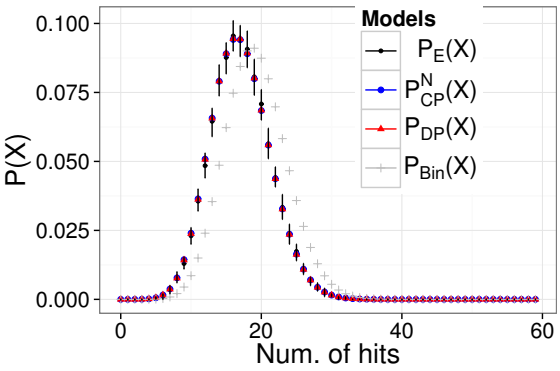


(a) $\alpha = 0.05$

(b) $\alpha = 10^{-2}$

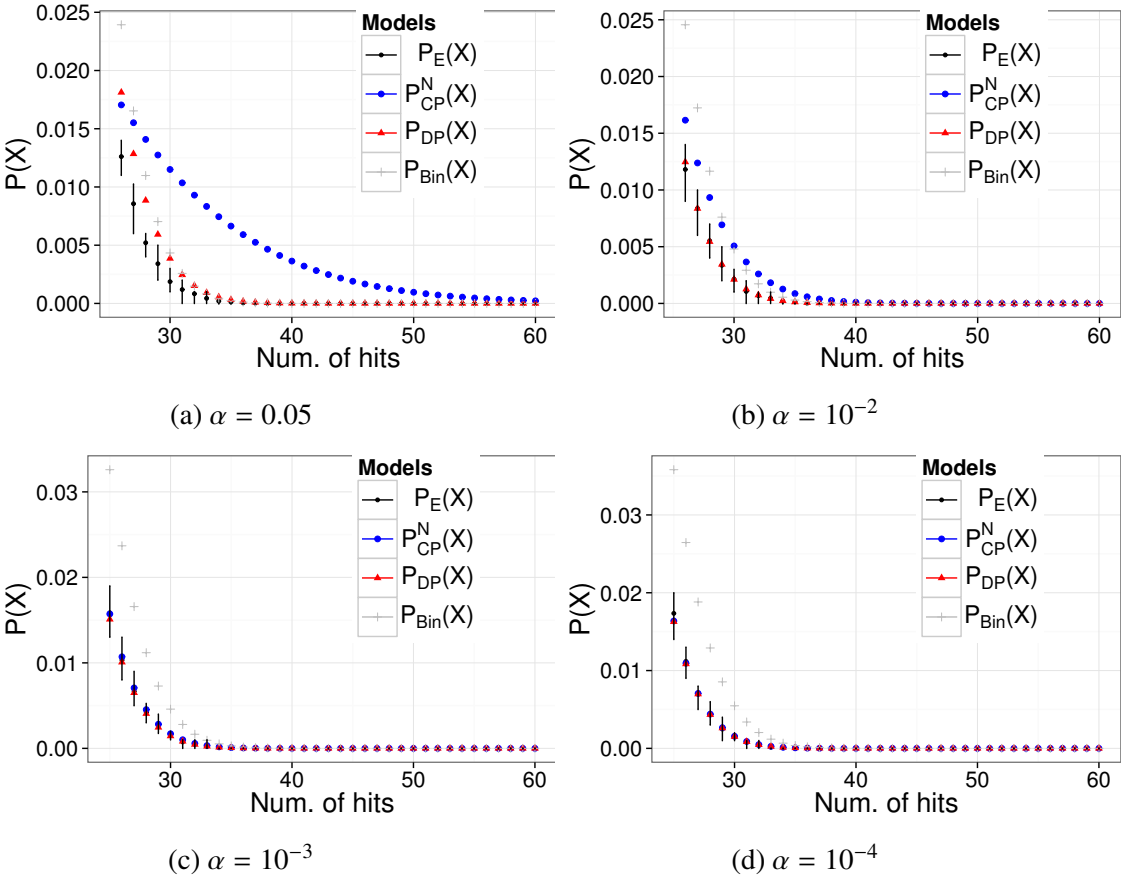


(c) $\alpha = 10^{-3}$



(d) $\alpha = 10^{-4}$

Figure 6.9: Motif hits count distribution for *E47* (see Figure (2.5c)). For each panel, we used an order-1 background model.



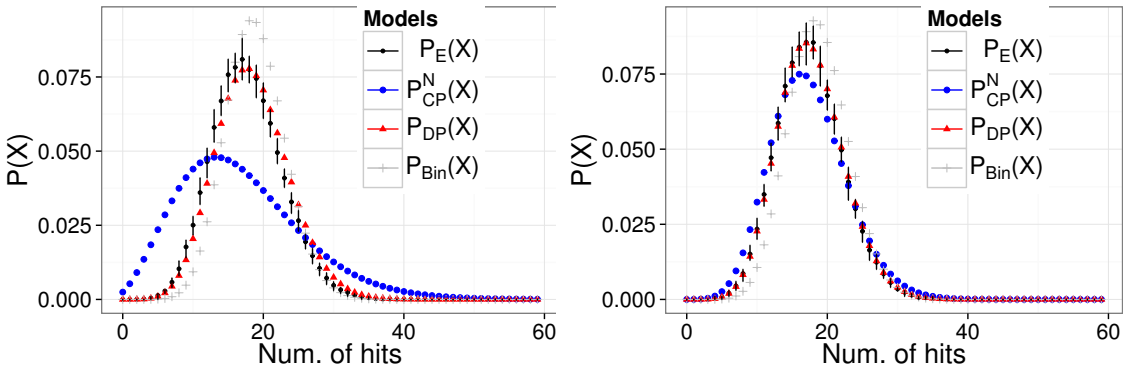
(a) $\alpha = 0.05$

(b) $\alpha = 10^{-2}$

(c) $\alpha = 10^{-3}$

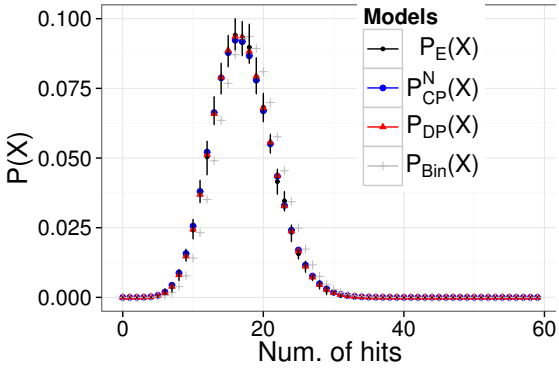
(d) $\alpha = 10^{-4}$

Figure 6.10: 5% significance region of the motif hits count distribution for *E47* (see Figure (2.5c)). For each panel, we used an order-1 background model.

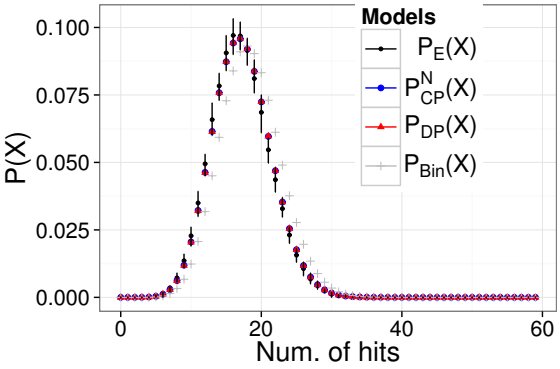


(a) $\alpha = 0.05$

(b) $\alpha = 10^{-2}$

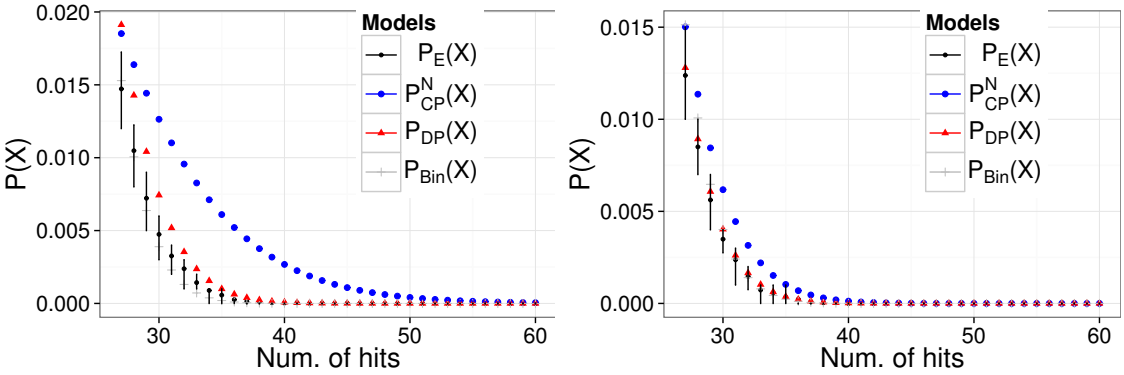


(c) $\alpha = 10^{-3}$



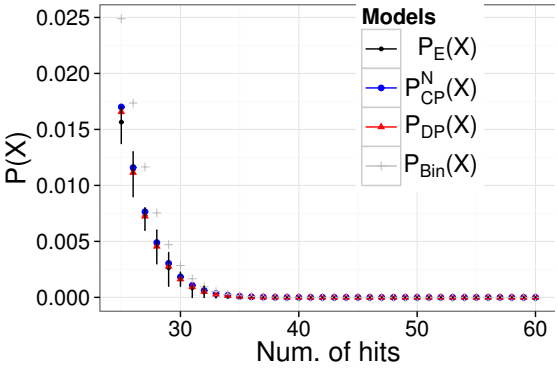
(d) $\alpha = 10^{-4}$

Figure 6.11: Motif hits count distribution for *SP1SP3* (see Figure (2.5d)). For each panel, we used an order-1 background model.

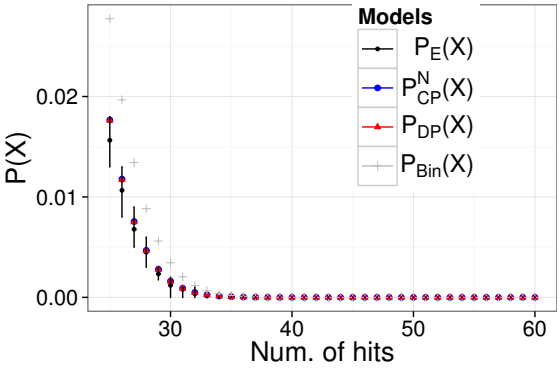


(a) $\alpha = 0.05$

(b) $\alpha = 10^{-2}$



(c) $\alpha = 10^{-3}$



(d) $\alpha = 10^{-4}$

Figure 6.12: 5% significance region of the motif hits count distribution for *SP1SP3* (see Figure (2.5d)). For each panel, we used an order-1 background model.

Zusammenfassung

In dieser Dissertation beschäftigen wir uns mit der statistischen Analyse von nichtkodierenden Segmenten des Genoms. Insbesondere betrachten wir Verfahren zur Identifikation und Anreicherungsanalyse von Transkriptionsfaktorbindungsstellen (TFBSen) in DNA Segmenten (z.B. in Promotoren), da das Binden von Transkriptionsfaktoren regulatorisch auf die Geneexpression benachbarter Gene wirkt. Die Identifikation von TFBSen basiert auf dem *Log-likelihood* Verhältnis zwischen einem bekanntem Transkriptionsfaktormotiv, welches die DNA Bindungsaffinität des Transkriptionsfaktors beschreibt, und einem Hintergrundmodel, z.B. einem Markov Model der Ordnung d, unter Verwendung eines festgelegten Schwellwerts. Kapitel 2 beschreibt die Berechnung der falsch-positiv Wahrscheinlichkeit für den gewählten Schwellwert. Da die Identifikation von TFBSen zu selbst-überlappenden Vorhersagen führen kann, welche die Anreicherungsanalyse beeinflussen, behandeln wir in Kapitel 3 die Quantifizierung der selbst-überlappenden Vorhersagen. In Kapitel 4 behandeln wir die *Compound Poisson* Verteilung als analytisches Model für die Anreicherungsanalyse welche selbst-überlappende TFBSen auf beiden DNA-Strängen berücksichtigt und eine direkte Weiterentwicklung von Pape *et al.* [36] darstellt. Der zentrale Fortschritt in diesem Kapitel wurde durch die Verwendung der neuer Überlappwahrscheinlichkeiten und durch die Verwendung von DNA-Hintergrundmodellen höherer Ordnung geleistet. In Kapitel 5 führen wir eine Markov Model ein, welches die Wahrscheinlichkeit einer TFBS, die nicht von einer vorhergehenden TFBSen überlappt wird, modelliert. Jene TFBSen markieren immer den Beginn eines oder mehrerer selbst-überlappender TFBS Vorhersagen (auch *motif clumps* benannt). Das Ergebnis von Kapitel 5 spielt eine wichtige Hilfsrolle für das darauffolgende Kapitel 6. Schließlich stellen wir ein neues kombinatorischen Model für die Anreicherungsanalyse in Kapitel 6 vor, welches effizient die Wahrscheinlichkeiten aller möglich Kombinationen x TFBSen in einer endlichen Sequenz der Länge N zu platzieren aufsummiert. Vergleiche mit dem *Compound Poisson* Model zeigten, dass das kombinatorische Model insbesondere für niedrige *Log-likelihood*-Schwellwerte wesentlich genauere Ergebnisse erzielt. Eine Implementierung der diskutierten Methoden ist als R Paket unter `https://github.com/wkopp/mdist` verfügbar.

Summary

In this thesis, we discuss methods for analyzing the non-coding sequence of the genome (e.g promoters) with respect to the identification and enrichment of transcription factor binding sites (TFBSs), as they are related to gene regulation. The identification of putative TFBSs is based on the log-likelihood ratio between a TF motif, which describes the binding affinity of a TF towards the DNA, and a background model, which is implemented by an order-d Markov models with $d \geq 0$, in conjunction with a pre-defined log-likelihood ratio threshold. Chapter 2 reviews algorithms for computing the false positive probability of calling motif hits for a given threshold. As putative TFBSs can self-overlap one another, which affects the enrichment test of the number of TFBSs, we discuss the quantification of overlapping TFBS predictions in Chapter 3. In Chapter 4, we discuss a compound Poisson model for modeling the distribution of the number of TFBSs in both strands of the DNA sequence, which represents an extension of Pape *et al.* [36]. The main advance of our model regards the use of newly derived *principal overlapping hit probabilities*, which are motivated by the discussion of *principal periods* in Reinert *et al.* [41], as well as by facilitating the use higher-order Markov models for the background. In Chapter 5 we discuss a novel Markov model which is utilized to determine the probability of a TFBS occurrence that does not overlap a previous TFBS occurrences, termed *clump start probability*, which mark the beginning of a clump. The resulting *clump start probability* then serves as an important building block for the subsequent Chapter 6. Finally, in Chapter 6 we present a novel combinatorial model for the distribution of the number of motif hit. To that end, we efficiently sum up the probabilities of all realizations of placing x TFBSs in a finite-length sequence of length N. We systematically compared the accuracy of the combinatorial model, the compound Poisson model and the binomial model. An implementation of the algorithms that were discussed in this thesis is provided as an R package that is available at `https://github.com/wkopp/mdist`.

Software implementation

The algorithms that were discussed throughout Chapter 2-6 were implemented in C using openMP to parallelize the routines. Additionally, we created an interface to R [39] which lead to an R package named 'mdist'. The R package is freely available on github: https://github.com/wkopp/mdist. Further details on the functionality of the package are available in the help of the package. All plots and tables were generated with functionality from the package.

Bibliography

[1] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell, Fourth Edition*. Garland Science, 4 edition, 2002.

[2] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 2015.

[3] Timothy L Bailey, Charles Elkan, et al. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. 1994.

[4] Yoseph Barash, John A Calarco, Weijun Gao, Qun Pan, Xinchen Wang, Ofer Shai, Benjamin J Blencowe, and Brendan J Frey. Deciphering the splicing code. *Nature*, 465(7294):53–59, 2010.

[5] Andrew D Barbour and O Chryssaphinou. Compound poisson approximation: a user's guide. *Annals of Applied Probability*, pages 964–1002, 2001.

[6] Michael Beckstette, Robert Homann, Robert Giegerich, and Stefan Kurtz. Fast index based algorithms and software for matching position specific scoring matrices. *BMC bioinformatics*, 7(1):389, 2006.

[7] Bradley E Bernstein, John A Stamatoyannopoulos, Joseph F Costello, Bing Ren, Aleksandar Milosavljevic, Alexander Meissner, Manolis Kellis, Marco A Marra, Arthur L Beaudet, Joseph R Ecker, et al. The nih roadmap epigenomics mapping consortium. *Nature biotechnology*, 28(10):1045–1048, 2010.

[8] Frederick R Blattner, Guy Plunkett, Craig A Bloch, Nicole T Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, Jeremy D Glasner, Christopher K Rode, George F Mayhew, et al. The complete genome sequence of escherichia coli k-12. *science*, 277(5331):1453–1462, 1997.

[9] K Cartharius, Kornelie Frech, Korbinian Grote, Bernward Klocke, Manuela Haltmeier, Andreas Klingenhoff, Matthias Frisch, M Bayerlein, and Thomas Werner.

Matinspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, 21(13):2933–2942, 2005.

[10] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[11] International Human Genome Sequencing Consortium et al. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, 2004.

[12] Francis Crick et al. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[13] Justin Crocker, Namiko Abe, Lucrezia Rinaldi, Alistair P McGregor, Nicolás Frankel, Shu Wang, Ahmad Alsawadi, Philippe Valenti, Serge Plaza, François Payre, et al. Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell*, 160(1):191–203, 2015.

[14] Charles Darwin. On the origins of species by means of natural selection. *London: Murray*, page 247, 1859.

[15] Patrik D'haeseleer. How does dna sequence motif discovery work? *Nature biotechnology*, 24(8):959–961, 2006.

[16] Jason Ernst and Manolis Kellis. Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types. *Genome research*, 23(7):1142–1154, 2013.

[17] Brent Ewing and Phil Green. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature genetics*, 25(2):232–234, 2000.

[18] Walter Fiers, Roland Contreras, Fred Duerinck, Guy Haegeman, Dirk Iserentant, Jozef Merregaert, Willy Min Jou, Francis Molemans, Alex Raeymaekers, A Van den Berghe, et al. Complete nucleotide sequence of bacteriophage ms2 rna: primary and secondary structure of the replicase gene. *Nature*, 260(5551):500–507, 1976.

[19] Brandon Franzke and Bart Kosko. Noise can speed convergence in markov chains. *Physical Review E*, 84(4):041112, 2011.

[20] André Goffeau, BG Barrell, Howard Bussey, RW Davis, Bernard Dujon, Heinz Feldmann, Francis Galibert, JD Hoheisel, Cr Jacq, Michael Johnston, et al. Life with 6000 genes. *Science*, 274(5287):546–567, 1996.

[21] Charles E Grant, Timothy L Bailey, and William Stafford Noble. Fimo: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, 2011.

[22] Charles Miller Grinstead and James Laurie Snell. *Introduction to probability*. American Mathematical Soc., 2012.

[23] Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, et al. Gencode: the reference human genome annotation for the encode project. *Genome research*, 22(9):1760–1774, 2012.

[24] David Haussler, Stephen J O'Brien, Oliver A Ryder, F Keith Barker, Michele Clamp, Andrew J Crawford, Robert Hanner, Olivier Hanotte, Warren E Johnson, Jimmy A McGuire, et al. Genome 10k: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *Journal of Heredity*, 100(6):659–674, 2009.

[25] Nathaniel D Heintzman, Gary C Hon, R David Hawkins, Pouya Kheradpour, Alexander Stark, Lindsey F Harp, Zhen Ye, Leonard K Lee, Rhona K Stuart, Christina W Ching, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459(7243):108–112, 2009.

[26] Samuel Karlin and Howard E Taylor. *A second course in stochastic processes*. Elsevier, 1981.

[27] CD Kemp. " stuttering-poisson" distributions. 21:151–157, 1967.

[28] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[29] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[30] Charles E Lawrence, Stephen F Altschul, Mark S Boguski, Jun S Liu, Andrew F Neuwald, and John C Wootton. Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *science*, 262(5131):208–214, 1993.

[31] Jun S. Liu and Charles E. Lawrence. Bayesian inference on biopolymer models. *Bioinformatics*, 15(1):38–52, 1999.

[32] Tobias Marschall. *Algorithms and statistical methods for exact motif discovery*. PhD thesis, Technische Universität Dortmund, 2011.

[33] Tobias Marschall, Inke Herms, Hans-Michael Kaltenbach, and Sven Rahmann. Probabilistic arithmetic automata and their applications. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(6):1737–1750, 2012.

[34] Timothy W Nilsen and Brenton R Graveley. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–463, 2010.

[35] Carl O Pabo and Robert T Sauer. Transcription factors: structural families and principles of dna recognition. *Annual review of biochemistry*, 61(1):1053–1095, 1992.

[36] Utz J Pape, Sven Rahmann, Fengzhu Sun, and Martin Vingron. Compound poisson approximation of the number of occurrences of a position frequency matrix (pfm) on both strands. *Journal of Computational Biology*, 15(6):547–564, 2008.

[37] Ezra Peisach and Carl O Pabo. Constraints for zinc finger linker design as inferred from x-ray crystal structure of tandem zif268–dna complexes. *Journal of molecular biology*, 330(1):1–7, 2003.

[38] Dan S Prestridge. Signal scan: a computer program that scans dna sequences for eukaryotic transcriptional elements. *Computer applications in the biosciences: CABIOS*, 7(2):203–206, 1991.

[39] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[40] Sven Rahmann, Tobias Müller, and Martin Vingron. On the power of profiles for transcription factor binding site detection. *Statistical Applications in Genetics and Molecular Biology*, 2(1), 2003.

[41] Gesine Reinert, Sophie Schbath, and Michael S Waterman. Probabilistic and statistical properties of words: an overview. *Journal of Computational Biology*, 7(1-2):1–46, 2000.

[42] Remo Rohs, Sean M West, Alona Sosinsky, Peng Liu, Richard S Mann, and Barry Honig. The role of dna shape in protein–dna recognition. *Nature*, 461(7268):1248–1253, 2009.

[43] Albin Sandelin, Wynand Alkema, Pär Engström, Wyeth W Wasserman, and Boris Lenhard. Jaspar: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, 32(suppl 1):D91–D94, 2004.

[44] Serge Saxonov, Paul Berg, and Douglas L Brutlag. A genome-wide analysis of cpg dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences*, 103(5):1412–1417, 2006.

[45] Yin Shen, Feng Yue, David F McCleary, Zhen Ye, Lee Edsall, Samantha Kuan, Ulrich Wagner, Jesse Dixon, Leonard Lee, Victor V Lobanenkov, et al. A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488(7409):116–120, 2012.

[46] François Spitz and Eileen EM Furlong. Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, 13(9):613–626, 2012.

[47] Gary D Stormo. Dna binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, 2000.

[48] Xiao-Hong Sun and David Baltimore. An inhibitory domain of e12 transcription factor prevents dna binding in e12 homodimers but not in e12 heterodimers. *Cell*, 64(2):459–470, 1991.

[49] Morgane Thomas-Chollier, Olivier Sand, Jean-Valéry Turatsinze, Matthieu Defrance, Eric Vervisch, Sylvain Brohée, Jacques van Helden, et al. Rsat: regulatory sequence analysis tools. *Nucleic acids research*, 36(suppl 2):W119–W127, 2008.

[50] Robert E Thurman, Eric Rynes, Richard Humbert, Jeff Vierstra, Matthew T Maurano, Eric Haugen, Nathan C Sheffield, Andrew B Stergachis, Hao Wang, Benjamin Vernot, et al. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, 2012.

[51] Hélène Touzet, Jean-Stéphane Varré, et al. Efficient and accurate p-value computation for position weight matrices. *Algorithms Mol Biol*, 2(1510.1186):1748–7188, 2007.

[52] Anatoly V Ulyanov and Gary D Stormo. Multi-alphabet consensus algorithm for identification of low specificity protein-dna interactions. *Nucleic acids research*, 23(8):1434–1440, 1995.

[53] Nicolaas Godfried Van Kampen. *Stochastic processes in physics and chemistry*, volume 1. Elsevier, 1992.

[54] Juan M Vaquerizas, Sarah K Kummerfeld, Sarah A Teichmann, and Nicholas M Luscombe. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, 10(4):252–263, 2009.

[55] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

[56] Michael S Waterman. *Introduction to computational biology: maps, sequences and genomes*. CRC Press, 1995.

[57] James D Watson, Francis HC Crick, et al. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[58] Edgar Wingender, Peter Dietze, Holger Karas, and Rainer Knüppel. Transfac: a database on transcription factors and their dna binding sites. *Nucleic acids research*, 24(1):238–241, 1996.

[59] Jing Zhang, Bo Jiang, Ming Li, John Tromp, Xuegong Zhang, and Michael Q Zhang. Computing exact p-values for dna motifs. *Bioinformatics*, 23(5):531–537, 2007.