Preface

This report is written as a fulfillment of my masters degree in computer science. I participate in a program called "forskerskolen", which means that the last year of my master overlaps the first year of my PhD. According to this program, my diploma thesis is supposed to be a research plan for my PhD and thus quite different from other diploma theses at my institute.

This research plan consists of three parts. The first part is a survey of methods for the discovery of DNA regulatory elements, with a particular focus on the computational models used. The second part is an article describing my contributions in the field so far. The article was written this spring, and is based on an algorithm I developed last year. The last part describes several research projects that I will be working on in my PhD. Plans always tend to change, and I therefore doubt that all of the research projects described here will be pursued. This may be due to publications from other researchers that make my research projects excessive, or due to shifting priorities. Still I expect to explore most of the proposed research projects.

Both computational and biological aspects are important for the discovery of DNA regulatory elements. As my focus is on algorithms and computational modeling, Finn Drabløs has written some paragraphs that are purely biological. More specifically, Finn has written paragraph 2-8 of section 2.1, the first two paragraphs of section 2.6.2, the two first paragraphs of section 3.2 and the first five paragraphs of section 3.4.

I want to thank Arne Halaas, Finn Drabløs, Magnus Lie Hetland, Rolv Seehuus, Øystein Lekang, Kai Trengereid and Tarjei Hveem for help and suggestions during this project.

Contents

Chapter 1

Introduction

Discovery of regulatory elements in DNA sequences is an important and still open problem. The goal is to discover short substrings in the region around a gene that binds to proteins called transcription factors. As binding of transcription factors to these substrings are important for the regulation of gene expression, the substrings are referred to as regulatory elements.

What makes computational discovery of regulatory elements possible, is that the elements are often reused for several genes in the same genome, and conserved across species. This means that motifs corresponding to real elements are often overrepresented in the genome.

Both the motif model used to represent regulatory elements, the calculation of motif overrepresentation and the algorithm used to discover the most overrepresented motifs are important. Motif models ought to accurately capture the sequence variability between elements with equal functionality. Moreover, calculations of overrepresentation should reflect the biological significance of a motif, and not just be statistical measures without biological interpretation. Finally, if exhaustive discovery of optimal motifs is prohibitive for a given motif model, the search space should be explored with efficient heuristics. To achieve high sensitivity, computational methods should incorporate as many relevant sources of information as possible.

Chapter 2 of this report is a survey of computational models for the discovery of regulatory elements in DNA. Chapter 3 describes a method for composite motif discovery. Finally, chapter 4 describes my plans for future research. Several specific research projects are described, along with some more general directions for future research.

Chapter 2

Computational models for motif discovery in DNA regulatory regions

2.1 Introduction

Understanding the regulatory networks of higher organisms is one of the main challenges of functional genomics. An important part of this is the discovery of regulatory elements. As biological verification of such elements is a tedious process, much effort has been put into the development of computational methods. Good computational methods can potentially provide high-quality prediction of binding sites and reduce the time needed for experimental verification.

The system for transcriptional regulation of the eukaryotic genome is complex. The regulatory processes are found at several hierarchical levels, in particular at the sequence level, the chromatin level and the nuclear level [136].

The sequence level includes coding regions, regulatory binding sites and sequence elements affecting the 3D fold of the chromatin fibre. In particular the binding sites for transcription factors will be discussed extensively here.

In eukaryotic cells DNA is packed as chromatin, and this affects transcriptional regulation. The basic unit consists of 150 base pairs of DNA wrapped 1.7 times around a protein octamer, consisting of histones. This unit is called the nucleosome, and it can exist in different structural and functional states.

Transitions between states are linked to gene activity. These transitions are influenced by post-translational modifications of histones, and this is often described as the histone code. Also gene silencing by DNA methylation is an important chromatin modification.

In addition to the linear (sequence) and pseudo-linear (chromatin) organisation of DNA, it is also organised in a highly folded state. This brings together genome regions that are far apart, which may affect the co-regulation of these regions. However, we lack efficient tools for studying global chromatin folding.

In particular the transcriptional regulation at the sequence level has been extensively studied, and several reviews are available, e.g. by Werner [143], Wray et al. [145] and Pedersen et al. [97]. The key regulatory region is the promoter, located upstream of the coding sequence. It is often separated into the basal (or core) promoter, where the transcriptional machinery is assembled, and the general promoter, where most of the transcription factors bind. The promoter basically integrates information about the status of the cell, and adjusts the transcription level according to this information. The transcription factors are proteins that bind to specific DNA motifs. These motifs are short. The effective length may be just 4-6 base pairs (bp) for a typical binding site, although the region affected by the transcription factor (the footprint) is longer, typically 10–20 bp. Each gene contains a large number of binding sites, 10–50 binding sites for 5–15 different transcription factors is not unusual. These transcription factor binding sites are often organised in modules consisting of several binding sites, where each module produces a discrete aspect of the total transcription profile. For many genes most of the binding sites are found within a few kb upstream of the start site. However, the variation is large, the size of the cis-regulatory region can vary by nearly three orders of magnitude from a few hundred bp to >100 kb. Regions have also been found downstream, in introns and even in exons of genes. The actual transcriptional regulation is achieved through a complex, combinatorial set of interactions between transcription factors at their binding sites [70].

What makes computational discovery of novel elements possible, is that functional elements are often reused for several genes in the same genome, and conserved across species. This means that novel regulatory elements may be discovered by searching for overrepresented motifs across regulatory regions. Searching for exact copies of short sequences is usually not adequate, because some sequence variety without change of function is common for regulatory

elements. Having an accurate model of this sequence variability is of course of utmost importance.

The basic approach to de novo computational discovery of regulatory elements is to first extract a set of sequences from the genome. This is typically fixed size upstream regions for a set of genes having similar functional annotation or gene expression. An algorithm is then used to discover the most overrepresented motifs according to some motif model and statistical measure.

Several extensions of this basic approach may increase the sensitivity of motif discovery. Regulatory elements are not distributed evenly in a fixed region upstream of a gene. Different genes will have varying degrees of similarity with the rest of the set. The context of a putative regulatory element may be important, such as other nearby regulatory elements, the presence of CpG-islands, or the position in the overall DNA structure. Finally, additional sources of information, such as regulatory regions of orthologous genes, are often available.

In principle, there is no difference between the models used for discovery of novel motifs and those used for discovery of new instances of known motifs. They are therefore treated equally in our discussion. However, since occurrences have to be determined for a very large number of putative motifs when doing de novo discovery, the models tend to be a bit simpler than those typically used in pure searching metods. Figure 2.2 shows the relation between searching and de novo discovery.

In section 2.2, we introduce an integrated model for the computational discovery of regulatory elements. In the following sections we discuss how recent methods approach various elements of the model, although no single method takes more than a few of these elements into account. More specifically, single motif models are covered in section 2.3, composite motif models in section 2.4, and the effect of motifs on gene regulation in section 2.5. Additionally, calculation of motif significance is briefly discussed in section 2.6, and some algorithmic concerns mentioned in section 2.7

2.2 An integrated model for computational discovery of regulatory elements

The regulatory machinery operates at several different levels, and this should be reflected in a computational model. A schematic view of our model is given in figure 2.1.

The lowest level consists of transcription factors (TFs) that bind to short contiguous sequence segments. These sequence segments are modeled by single motifs that give a distinct score for each sequence segment in a regulatory region. This score is based on the match between the sequence segment and a motif consensus model, and on the prior belief that any regulatory element is to occur at the given location.

At the next level are modules: clusters of TFs that bind to DNA in proximity to each other, but with a certain distance flexibility between binding sites. This is modeled by a composite motif model, consisting of a set of single motifs. Given a set of positions, one for each single motif, the score of a composite motif is calculated from the score of each single motif at its position and the inter-motif distances.

The third level concerns how the multitude of binding possibilities for a set of modules add up to determine regulation of a single gene. This is modeled by a gene score function that combines composite motif scores across the regulatory region.

The final level of our computational model concerns evaluation and ranking of discovered motifs. Significance evaluation is based on either genome-wide overrepresentation of motifs, or correspondence between motif scores and experimental data.

Table 2.1 gives an overview of how various elements of our model are approached by selected methods, including both novel and more established approaches.

2.3 Single motif model

We define a single motif model as a function $m_g(p) : \mathbb{M} = \mathbb{N} \rightarrow \mathbb{R}$ that takes as input a position in the genome, and returns a value indicating whether an occurrence of the motif starts at this position. This function is typically the product or sum of two conceptually different functions: The match model, $m^*(p)$, gives the degree of match between the substring beginning at position

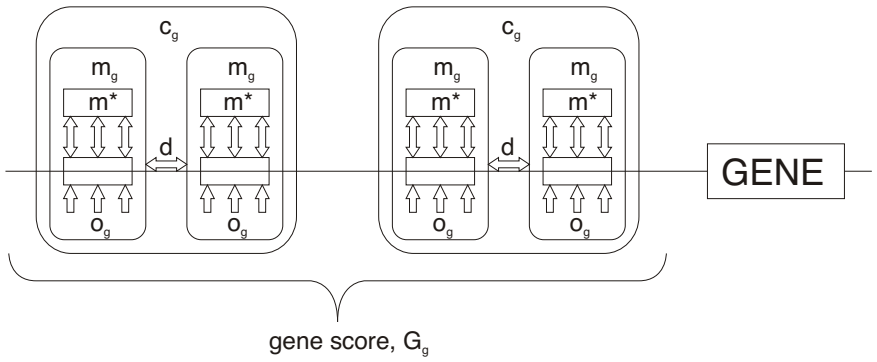


Figure 2.1: Schematic view of our integrated model. Each m_g corresponds to a single regulatory element, while each c_g corresponds to a module.

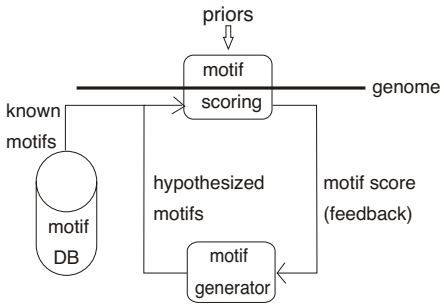


Figure 2.2: Relation between search and de novo discovery. In de novo discovery, the scores of hypothesized motifs are used to guide the search for new significant motifs

Table 2.1: Overview of methods

ALGORITHM NAME	**MATCH MODEL**	**DISTANCE FUNCTION**	**GENE SCORE**
Weeder[96]	mismatch	-	max
Dyad analysis[139]	oligos	constraint	max
MCAST[9]	PWM	gap penalty	HMM
[29]	PWM	constraint	sum
MDScan[84]	PWM	-	max
Gibbs sampler[78]	PWM	uniform	max
MEME[8]	PWM	-	sum
LOGOS[148]	DM	distribution	HMM
Motif regressor[34]	PWM	-	sum
ModuleSearcher[3]	PWM	window	max
Stubb[124]	PWM	window	HMM
GANN[13]	flexible	window	ANN
ANN-Spec[144]	PWM	-	max
[142]	PWM	window	max
CoBind[56]	PWM	window	sum
Cister[51]	PWM	distribution	HMM
SeSiMCMC [46]	PWM	-	mixture model
[88, 87]	mismatch	constraint	max
BioProspector[83]	PWM	constraint	sum
[117]	PWM	-	logistic func.
[122]	reg.exp	constraint	sum
ConsecID[118]	PWM	window	sum
SCORE[105]	IUPAC	window	sum
Gibbs recursive [132]	PWM	distribution	mixture model
[94]	PWM	-	hyperb. tan.
AlignACE[109]	PWM	-	mixture model
Improbizer[4]	PWM	-	mixture model
CisModule[153]	PWM	mixture model	mixture model
[131]	PWM	constraint	max

p and an underlying consensus model. The occurrence prior, $o_g(p)$, gives the prior belief that position p contains any regulatory element for gene g.

2.3.1 Match model

In the most general sense, the match model m^* is a function that gives a distinct score for any given substring. However, the number of free parameters has to be restricted to allow training of the model from a limited number of examples (known regulatory elements).

Numerous match models have been proposed, and they are often divided into two groups: deterministic models with binary scores and probabilistic models with weighted scores.

Probabilistic match models

The most widely used probabilistic model is without doubt the position weight matrix (PWM), also called PSSM or PSWM [125], that assumes independence between positions. The score of an aligned substring is the log-likelihood of the substring under a product multinomial distribution. PWM scores can also be described in a physical framework as the sum of binding energies for all nucleotides aligned with the PWM.

Many different extensions to the basic PWMs has been proposed in the literature. Most of these extensions concern positional dependencies within a motif. There is an ongoing discussion on the importance of such positional dependencies, see for instance [15, 152, 93].

The most direct way of incorporating dependencies within motifs, is to extend the PWM to include pairs of correlated positions [152]. Another straightforward approach is to use a mixture model in which the motif occurs as one of a limited number of stochastic prototypes [12]. Each stochastic prototype may be a traditional PWM, or any other model discussed in this section. A third extension is to model probabilistic motifs as n‘th order Markov chains [82]. It is, however, hard to find a good compromise between a high n that gives too many free parameters and a low n that misses out the dependencies of interest. If the relative importance of dependencies varies within a motif, a variable-length markov model(VLMM) [32], may be preferrable. Furthermore, if some long-range dependencies seems to be significantly stronger than dependencies between neighbouring positions, the

order of the positions in the markov chain may also be permuted before a VLMM is applied [151].

Another way to model dependencies is to use bayesian networks. Barash et al. discuss different Bayesian network models and conclude that the use of a Bayesian tree model, or possibly a mixture of trees, is a good compromise between the number of free parameters, the ability to model dependencies, and computational tractability [12]. Similarly, Ben-Gal et al. [14] argue for variable order Bayesian nets.

Instead of focusing on dependencies between specific nucleotides at different positions, Xing et al. model the distribution of conserved positions within a motif [149]. In this model there is an underlying markov chain of position prototypes. Each prototype defines a certain Dirichlet distribution on the parameters of the multinomial nucleotide distribution at that position. The underlying Markov chain favors transitions between position prototypes with similar degrees of conservation. This makes it possible to favor models where highly conserved positions are partially contiguous rather than evenly spread out in the motif.

Deterministic match models

A deterministic match model evaluates to a binary value indicating hit or no-hit. The three main kinds of deterministic match models are oligos, regular expressions and mismatch expressions.

The simplest deterministic model is the oligo model. This is a function that is 1 for a single specific substring, and 0 for all other substrings. The oligo model was commonly used in early motif discovery methods , but has also been used in recent word-counting methods [137, 64, 122] and dictionary models [28].

A regular expression model m^* returns 1 if an underlying regular expression matches the given substring. As reviewed by Brazma et al. [23], the models used in motif discovery are typically composed of exact symbols, ambigous symbols, fixed gaps and/or flexible gaps. Regular expression models are used in e.g. [139, 122, 119, 128].

Many methods use mismatch expressions as motif match models, e.g. [134, 87, 98, 96, 43, 10]. These models evaluate to 1 if the number of mismatches(Hamming distance) between a substring and the underlying consensus substring is below a given threshold. A variant is described in [81], where the threshold is on the sum of mismatches between all motif occurrences and

the underlying motif substring. A similar variant, with a threshold on mismatches between occurrences in sequences arranged in a phylogenetic tree, is described in [18].

Both oligos, regular expressions and mismatch expressions can be represented as PWMs. However, a major benefit of these models is that they allow exhaustive discovery of optimal motifs. This is discussed further in 2.7.

2.3.2 Occurrence prior

The genetic context of a regulatory element is important for its activity. Distance to transcription start site, sequence conservation in orthologous genes, DNA structure and presence of CpG-islands may all be relevant.

In our model, these context features are represented by an occurrence prior, $o_g(p)$, representing the prior belief that any regulatory element is located at a given position p.

The simplest kind of occurrence prior is a motif abundance ratio [66]. This ratio influences only the number of substrings that count as occurrences. Another simple prior is strand bias, which corresponds to an occurrence prior that is higher on one strand than on the other [124]. Similarly, Liu et al. [83] and Donaldson et al. [38] optionally constrain the search to only one of the strands, which corresponds to a binary strand bias.

Spatial distribution of binding sites

In higher organisms, regulatory elements may be located far upstream of the gene, downstream of the gene, in introns, and even in exons. Nevertheless, most elements are located immediately upstream of the transcription start site (TSS).

In general, this can be represented by a function that gives the prior belief that a regulatory element is located at a given position relative to the TSS.An occurrence prior based on the empirical distribution of element locations in E.coli has been used in [89] and [132]. Nevertheless, the by far most common approach is to only search for motifs in a fixed region upstream of TSS, which corresponds to a binary function o_g.

Conservation in orthologous sequences

The term phylogenetic footprinting is commonly used to describe phylogenetic comparisons that reveal conserved elements in regulatory regions of homologous (in particular orthologous) genes [127].

The reasoning behind phylogenetic footprinting is that since regulatory elements are functionally important and are under evolutionary selection, they should have evolved much more slowly than other non-coding sequences. Moreover, genome-wide sequence comparison and studies on individual genes have confirmed that regulatory elements are indeed conserved between related species [150].

More specifically, Krivan and Wasserman [77] reported that highly conserved regions were around 320 times more likely to contain regulatory elements than non-conserved regions, based on findings from a set of liver-specific genes.

Several methods exploit information about conservation in orthologous genes by searching for motifs only in highly conserved sequence parts (typically human-mouse orthologs) [118, 124, 10, 36]. This approach corresponds to using a binary occurrence prior that is 1 if the conservation score is above a threshold, and 0 otherwise.

Wasserman and Fickett [142] use non-binary conservation scores, but they do not incorporate these into the search as priors. Instead, they only use conservation to filter the discovered motifs.

DNA structure

The three-dimensional structure of DNA, densely packed as chromatin, inhibits transcriptional initiation in vivo [97]. The bendability of a region, as well as its position in DNA loops, may give indications on whether it contains regulatory elements.

Only a few motif discovery methods take DNA structure into consideration. Beiko and Charlebois [13] average structure scores of all k-mers in a window around a given position, independently of any particular motif. Conversely, Pudimat et al. [102] incorporate helical parameter features [100, 41] in a bayesian net that is specific for each motif.

Nucleotide distribution

Both high GC content and presence of CpG-islands may indicate that a region contains regulatory elements. Pudimat et al. [102] is one of few methods that take this into consideration when calculating motif scores.

2.4 Composite motif model

Clusters of binding sites for cooperating TFs, often called modules, are believed to be essential building blocks of the regulatory machinery. Werner [143] states that "Within a promoter module, both sequential order and distance can be crucial for function, indicating that these modules may be the critical determintants of a promoter rather than individual binding sites". The multitude of models developed for the discovery of modules is another indication of the conceived importance of modules.

It is therefore natural to define a computational motif model that represents a combination of single motifs. We define a composite motif as a function $c_g(\vec{p}) : \mathbb{C} = 2^{\mathbb{N}} \rightarrow \mathbb{R}$. A composite motif consists of a set of single motifs $\vec{m_g}$, with each single motif giving a separate score at its position. In addition, functions may be defined on the distances between single motifs. Given a set of positions, the score of a composite motif will typically be the sum or product of each single motif and distance score.

The cooperativity in modules may be homotypic, involving binding of the same TF to multiple binding sites, or it may be heterotypic, involving binding of different TFs [140]. This means that a composite motif may be composed of several instances of the same single motif, or it may be composed of distinct single motifs.

2.4.1 Distance functions

Many different models have been propsed to capture the importance of inter-motif distances within a module. Several methods put constraints on the distances between consecutive motifs, requiring either fixed distances [83, 122] , distances below thresholds [72, 62, 131], or distances within intervals (e.g. [139, 122, 29, 83, 43])

Another common way of capturing the importance of proximity, is to constrain all single motifs to be within a window of a certain length (e.g.

[142, 56, 105, 3, 124]). This corresponds to a threshold on the maximum distance between any two single motifs.

A more general approach is to define non-binary score functions on the distances between single motifs. This can simply be functions that increase linearly with distance as in [9].

The conservation of inter-motif distances across modules can also serve as basis for distance score. Wagner [140] calculates a distance score from the p-value of observing the given degree of distance conservation in a background model of poisson-distributed inter-motif distances. Similarly, Frech and Werner [48] calculate scores by comparing the distances with a histogram of distances between the same regulatory elements in other modules. Finally, a geometric distribution on inter-motif distances follows implicitly from many HMM models [51, 148].

We have implicitly assumed in this discussion that distance is the number of base pairs between two positions in the genome. It is in principle possible to measure distance in other ways. An example is to require all motifs in a module to be on the same strand [119], which corresponds to a simple binary distance function. More importantly, as our understanding of DNA folding increase, new and more complex distance measures may appear.

2.4.2 Combining single motifs

There are many ways of combining all single motif and distance scores in a single measure.

For methods using deterministic match models, and constraints on distances, all component scores are binary. Furthermore, many probabilistic methods use thresholds on single motif scores to obtain only binary values. The composite motif score is then typically the intersection of component scores (e.g. [140, 114, 118, 24].) A slight variation of this is to require that M out of N single motif scores are one [99]. Similarly, the count of binary single motif values can be used directly as composite motif score [16, 115, 122].

For methods that use non-binary single motif scores, a common approach is to calculate the sum of single motif and distance scores [9, 48]. Some methods require that all distance functions are 1, and if they are, composite motif score is the sum of single motif scores [3, 2, 56, 76]. Similarly, the method Modulescanner sums only single motif scores above a threshold, and MotifLocator sums the N highest single motif scores [3]. Another slight variation is to multiply the sum of single motif scores with a motif density

factor, calculated from the length of the window that contains all of the single motifs [72]. Finally, a few methods take the composite motif score to be the highest single motif score [96], or the lowest single motif score [67].

Many other specialized models have also been used to combine single motif and distance scores: Hidden Markov Model(HMM) [148], history-concious HMM(hcHMM) [124], Self-organizing Map (SOM) [85], and Artificial Netural Network (ANN) [13]. In all of these models, the score of several homotypic and/or heterotypic single motifs are combined in a relatively complex way.

It should be noted that composite motif scores need not be relevant although a method discovers composite motifs. Many methods discover composite motifs iteratively in a greedy way (e.g. MEME [8] and AlignACE [109]). As only the highest scoring single motif is added in each iteration, there is no need to evaluate and rank entirely different composite motifs.

2.5 Gene score

Motif scores are defined for specific positions, and indicate likely locations of regulatory elements. Additionally, we are often interested in how much influence a motif has on the regulation of a gene. This is calculated from composite motif scores, $c_g(\vec{p})$, across the whole regulatory region of gene g, and is referred to as gene score ($G_g(c) : \mathbb{C} \rightarrow \mathbb{R}$).

Gene score is often defined simply as the maximum motif score in the regulatory region of a gene [144, 84, 18, 115, 3]. This corresponds to an implicit assumption of exactly one relevant occurrence of a motif in a regulatory region.

There is, however, reason to believe that the presence of multiple binding sites for TFs plays an important biological role that should not be neglected. Many methods therefore calculate gene score as the sum of motif scores across the regulatory region. As motif scores are typically log-scores, most methods add the exponentials of motif scores (e.g. [29, 56, 141, 49]). A slight variation is to only sum motif scores above a certain threshold [9].

In addition to the before mentioned formulas, many variations have been used to calculate gene score: Caselle et al. [31], Cora et al. [35] and Cora et al. [36] calculate gene score as the p-value of the observed set of motif scores. Curran et al. [37] calculate gene scores based on logistic regression. Similarly Segal et al. [117] use a logistic function, and P. et al. [94] a hyperbolic tangent, on the sum of motif scores. Finally, Beiko and Charlebois [13] use an artificial

neural network to combines motif scores.

A special case arises with the dictionary models of Bussemaker et al. [28] and Gupta and Liu [57], which always span whole regulatory regions. In these methods, the score of all valid segmentations of the region into contiguous words from the dictionary is added together to form the gene score.

2.6 Motif significance

In de novo motif discovery, numerous motifs are typically hypothesized, while only a few correspond to real biological entities. Therefore, evaluation and ranking of motifs is essential.

Motif significance, $s(c) : \mathbb{C} \rightarrow \mathbb{R}$, is based on either the genome-wide overrepresentation of the motif, or on the correspondence between gene scores and experimental data.

2.6.1 Genome-wide overrepresentation

Computational motif discovery is possible primarily because regulatory elements are overrepresented. Many methods use this overrepresentation directly when evaluating the significance of a discovered motif. The exact way of calculating motif significance varies from method to method, but can roughly be divided into five different approaches.

The most direct approach is to determine overrepresentation by comparing observed motif scores with expected scores in a background model. More specifically, the p-value [105, 128] and z-score [134, 122] of the observed sum of gene scores has been used. The background is typically a higher order Markov model, with parameters estimated from the sequences used for motif discovery. Alternatively, shuffled control sequences can be used as background [78].

A simpler approach is to only compare the raw sum of gene scores when ranking motifs []. This is equivalent with the first approach under the assumption of equal expected scores for all motifs in the background model.

A third approach is to use a significance measure related to the IC of discovered PWMs [8]. For methods that use mixture models of log-ratio PWMs and background, the PWM with highest IC corresponds to a maximum likelihood solution of the mixture model.

A common approach in deterministic motif discovery is to calculate two separate values when evaluating motifs: one concerning the support, or coverage, of a motif, and a second concerning the unexpectedness of a motif [68, 108, 87].

The fifth approach is completely different, and focuses only on overrepresentation of motif combinations. Motif significance is based on the observed versus expected scores of *composite* motifs, given the observed score distribution of *single* motifs. The significance can, for instance, be the p-value of the observed composite motif scores in a background model where all single motif occurrences are randomly reshuffled [118].

2.6.2 Correspondence with experimental data

In recent years the development of microarray technology has revolutionised studies of regulatory processes, in particular because it can be used to identify genes that are co-regulated under specific conditions. Microarrays are used to measure relative expression levels of genes in a set of experiments. This may be e.g. time series experiments like cell cycle studies or before/after experiments like stress response studies and studies of malignant vs. normal tissue. It is a reasonable hypothesis that genes showing synchronised changes in expression levels share important aspects of transcriptional regulation, e.g. transcription factor binding sites. Sets of genes showing co-regulation may therefore be used for data mining for shared regulatory motifs [111], although it has been shown that this type of data mining is difficult and error prone [135].

Recently, genome-wide binding analysis like ChIP/chip experiments have appeared as an approach for more reliable identification of actual binding sites [107, 26]. In a ChIP/chip experiment a known transcription regulator is tagged with an antibody epitope, and the tagged regulator is expressed in a suitable system where it binds to DNA, either directly or via other proteins. The complex is then chemically crosslinked, the DNA is fragmented, and the protein/DNA complex is isolated by immunoprecipitation. The genomic position of the DNA fragment is then identified by a microarray experiment. This gives the location of binding sites for this specific regulator, although the relevance of the information may be limited by the specific set of experimental conditions used and the resolution of the experiment itself (DNA fragment size and genome resolution on the microarray chip).

Besides ChIP/chip and microarray experiments, gene groups are often

formed from conserved orthologous genes [91, 18, 104, 141], or genes with similarities in functional annotation [64, 36]. Finally, genes that make up functional pathways, genes that are homologous to regulons from a well-studied species, and groups of genes derived from conserved operons have also been used [90].

The availability of experimental data makes it possible to form initial hypotheses about co-regulation. Many methods cluster the genes based on experimental similarities, assigning each gene to a single group of putatively co-regulated genes. All genes are then treated equally during motif discovery, regardless of the degree of similarity between a gene and the rest of the group (e.g. [90, 130, 37, 131, 94]).

As genes can be co-regulated with several groups, we use fuzzy sets to represent prior grouping of genes. In our model, every gene g has a weighted membership $\mu F(g)$ in each fuzzy set F. Segal et al. [115] and Liu et al. [84] are among the few authors that have used weighted values for experimental data during motif discovery.

The correspondence between gene scores and experimental data may be used as a measure of motif significance. This can be calculated in several ways. One approach is to evaluate the fit of a logistic regression from gene scores G_g to membership values $\mu F(g)$ [142, 37]. A simplification of this approach is to compare binary gene scores with binary membership values , and calculate the mismatch ratio [94] or ROC_{50} score [9]. Alternatively, grouping of genes can be avoided altogether, and motif significance measured as the fit of a linear regression directly from gene scores to observed log-expression in micro-array experiments [29, 34, 117].

Park et al. [95] consider the problem in the opposite direction. They first discover motifs in the regulatory regions of all genes and form groups of genes that share common motifs. Motif significance is then measured as the similarity in gene expression within the group formed from the common motif.

Finally, Holmes and Bruno [60] calculate the joint likelihood of both shared motifs and expression similarity for hypothesized gene groups.

Although several methods may be configured to use different kinds of experimental data [64, 90, 36], only a few methods tries to combine different kinds of data in a single similarity measure. Takusagawa and Gifford [128] use the GRAM algorithm [11] to cluster genes based on both ChIP-data and gene expression data. Further work, incorporating more kinds of experimental data and using fuzzy set membership, could give more robust priors on co-

regulation and increase the sensitivity of motif discovery.

2.7 Some algorithmic concerns

An important trade-off in motif discovery is between representational expressibility and computational efficiency. For the case of binary priors and restricted deterministic motif models, several algorithms exist that can exhaustively discover the optimal motifs [68, 108, 44].

Probabilistic motif discovery algorithms do not guarantee returning the global optimum when applied to realistic problems. These algorithms are typically based on either iterative refinement or stochastic optimization. Expectation maximization (EM) [79, 8, 121, 101, 4] is the most widely used iterative refinement method, but variational EM [148]: have also been used. The stochastic optimization technique most widely used for motif discovery is Gibbs sampling [78, 92, 83, 132], sometimes combined with general Metropolis-Hastings [57, 66, 153]. Recently, simulated annealing has also gained some popularity [115, 151, 55]

Seed-driven algorithms have been used with success in deterministic motif discovery. They start by evaluating seeds from a very restricted class of simple motifs, and then expand promising seeds to full motifs either heuristically [58] or exhaustively [108]. A promising approach to motif discovery is to first use efficient deterministic motif discovery, and then use the highest scoring deterministic motifs as seeds for probabilistic motif discovery with expressive models. In addition, motifs may first be discovered in the sequence parts with highest priors, and then used as seeds for motif discovery in the full set of sequences. The method of Liu et al. [84] is a good example of such a strategy. Several overrepresented mismatch expressions are first discovered in upstream regions of the genes with highest group membership ($\mu F(g)$). The highest scoring mismatch expressions are then used as seeds for probabilistic motif discovery in the whole set of sequences.

2.8 Concluding remarks

The field of motif discovery brings together researchers from several disciplines, in particular from biology, statistics and informatics. Additionally, research in the field is fairly recent and moving at a fast pace. This has

resulted in a broad range of computational methods that are described with different vocabulary and different focus, making it difficult to spot similarities and differences between methods. Most articles on novel computational methods focus on the authors own biological results. Hence, the authors often put less emphasis on giving clear descriptions of precisely what an algorithm requires as input, how it evaluates motifs, and what it returns as output. This also contributes to making it harder to compare methods from their description.

When trying to compare the accuracy and computational efficiency of methods by measurement, one faces additional problems: The choice of data set, choice of performance measures and tuning of program parameters all have strong influence on the relative performance of methods [135].

Establishing a standardized framework for testing and comparing methods would be an important contribution in the field. Such a framework should include a collection of diverse data sets and a few different measures of performance. Furthermore, a consensus on what constitutes essential aspects of motif discovery methods could ease comparison of methods based on their description, making it easier to choose between or integrate different methods. The integrated model proposed in this article may be one step towards a common vocabulary and understanding of the problem.

Chapter 3

The generalized composite motif discovery (GCMD) algorithm

3.1 Abstract

This paper discusses a general algorithm for the discovery of motif combinations. From a large number of input motifs, discovered by any single motif discovery tool, our algorithm discovers sets of motifs that occur together in sequences from a positive data set. Generality is achieved by working on occurrence sets of the motifs. The output of the algorithm is a Pareto front of composite motifs with respect to both support and significance. We have used our method to discover composite motifs for the AlkB family of homologues. Some of the returned motifs confirm previously known conserved patterns, while other sets of strongly conserved patterns may characterize subfamilies of AlkB.

3.2 Introduction

Motif discovery in DNA and protein sequences is an important field in bioinformatics. Unique motifs found in a set of related sequences are often associated with the biological activity of the sequences. Motifs representing active site residues in enzymes (proteins) or transcription factor binding sites in genomes (DNA) are typical examples. Such motifs can also be used for

classification of novel sequences or sequences outside the original training set. Both probabilistic and deterministic approaches are used. Arguably, deterministic approaches give the most easily interpretable results, as they represent motifs e.g. by subsets of regular expressions that either match a given sequence or not.

There are many different algorithms for motif discovery, including manual approaches. The earliest algorithms had very limited expressibility and could only discover substrings of amino acid symbols. PROSITE [25], a database of manually annotated motifs, in many ways set the standard for expressibility of deterministic motifs for proteins. In addition to exact symbols, PROSITE patterns also consist of fixed gaps, flexible gaps and ambigous symbols. Most automated motif discovery tools are only able to discover motifs consisting of a subset of these components.

The discovery of motif combinations is an area of active research, for which both probabilistic and combinatorial approaches are used. Gibbs sampler [92] and PRINTS [6] are two well-known probabilistic approaches. Most combinatorial approaches discover spaced dyads [139, 43] or ordered sets of motifs with strong distance constraints [87]. Brazma et al. [24] are among the few methods that discover unordered sets of motifs.

A set of single motifs is a general starting point for composite motif discovery. Many advanced methods exist for the discovery of single motifs, and none are superior in all respects [135]. We have therefore chosen to develop an algorithm for the discovery of motif combinations that can use single motifs generated by any deterministic motif discovery tool.

GCMD (Generalized Composite Motif Discovery) exhaustively identifies the most significant combinations of a set of precomputed motifs. It can be set to discover both ordered and unordered motifs, with or without distance constraints. In addition to being flexible with regards to both single and composite motif model, and exhaustive in search for combinations, two properties clearly distinguish our algorithm from previous approaches: we model the problem as a two-goal optimization with the optimal Pareto front as output, and we automatically discover potential subfamilies.

GCMD is here discussed mainly in terms of protein sequence motifs. However, the tool itself is general and can also be applied to motifs from DNA sequences.

3.3 The Generalized Composite Motif Discovery tool

In broad terms, GCMD takes as input a set of single motifs and exhaustively discovers the optimal motif combinations with respect to both support and significance. This is more thoroughly explained in the following sections.

3.3.1 Vocabulary

The set of sequences that have at least one occurrence of a given motif, is called the *occurrence set* of the motif. The cardinality of the occurrence set is referred to as *support*.

We use the term *single motifs* to denote the motifs that are input to the GCMD algorithm, and *composite motifs* to denote the discovered motifs that are sets of single motifs. The term *component* is used to denote one of the single motifs that makes up a composite motif.

We also use the terms *Pareto domination* and *Pareto front* in multiple criteria optimization. A motif is Pareto dominated if there exists another motif having equal or higher values of both support and significance, where one of the values is strictly higher. Since support is a discrete value, this means that a motif is Pareto dominated if there exist another motif with equal or higher support, and strictly higher significance. The Pareto front is the set of all non-dominated motifs. In our case this is the most significant motif for each value of support.

3.3.2 Motif representations

The first step in using GCMD is to discover deterministic single motifs with a separate motif discovery tool. For tools that discover probabilistic motifs, a threshold may be used to make them deterministic. A bitstring is then constructed for each motif, where the i'th bit is 1 if the motif has an occurrence in the i'th sequence, and 0 otherwise [24]. A composite motif occurs in a sequence if, and only if, every single motif in the set occurs in the sequence. This leads to a basic representation of a composite motif as a set of indexes to its component motifs, as well as an occurrence set calculated by taking the intersection of the occurrence sets of all component motifs.

3.3.3 Significance evaluation

Significance of motifs is measured as negative log-likelihoods, using the same calculations as the motif discovery method Splash [58]. More specifically, the significance of a single motif is the negative log-likelihood of observing the motif in a random background sequence with the same amino acid distribution as the input sequences. As single motifs usually are short compared to sequence length, the log-likelihood of a composite motif is in general well approximated as the sum of log-likelihoods of its components.

3.3.4 Significance vs support

Both significance as well as support is important when evaluating motifs, and it is not easy to make the right trade-off between these properties when doing automated motif discovery. Most algorithms require a threshold on support, and this threshold is often user specified. Using a very strict value may lead to loss of significant motifs that are characteristic of subfamilies of sequences. On the other hand, a too permissive threshold may lead to searches dominated by motifs with high statistical significance in subsets of sequences, and one may lose less significant motifs representing weak commonalities characteristic of larger sequence families.

By formulating the motif discovery problem as a two goal optimization, we can explore a very large search space of interesting motifs, and return information about this in a condensed form as a Pareto front. The user gets a diverse set of motifs, and can readily see the tradeoff between significance and support as the number of sequences taken into consideration increases. This removes the need to set explisit thresholds on support or significance.

3.3.5 Pruning of search space

GCMD traverses the search space exhaustively and returns the set of Pareto optimal composite motifs. The size of the search space is $\binom{n}{c}$, where n is the number of single motifs used as input to GCMD, and c is the desired number of components in the composite motifs. Many algorithms exist for the mining of frequent item sets. Brazma et al. [24] uses the algorithm of Toivonen [133] to discover unordered sets of motifs. As this algorithm do pruning only based on support, it can not handle the large number of input motifs and low values of support that we are interested in.

We have developed a branch-and-bound algorithm, tailored to our two goal optimization problem, that is very efficient on real biological data. Since our goal is to find an optimal Pareto front with respect to support and significance, we need to determine upper bounds on both of these values. An upper bound on support is simply the minimum support of the current components of the composite motif. To introduce an upper bound on significance, we ensure that when a composite motif is expanded, the new component has a lower significance value than all other components of the motif. Note that this does not reduce the set of composite motifs we are able to discover, it only excludes all but one of the $n!$ permutations that corresponds to the same combination of n single motifs. For a given motif c_i, this leads to a straightforward upper significance bound on any expansions of c_i with n components :
$s(c_n) \leq s(c_i) + (n - i) * s(c_i(i))$, where c_i is a motif with i components, $c_i(i)$ is the i'th component of motif c_i, and $s(c)$ is the significance of motif c.

With these upper bounds in place we can make a recursive function that takes as parameter a composite motif c that is to be expanded. For each single motif s with significance lower than all current component significances of the motif, we check whether the resulting upper bounds on support and significance are dominated by the current Pareto front. If not, a new composite motif is formed from c, with the single motif s as an added component. The resulting motif is stored in the Pareto front if it has reached the desired number of components, otherwise it is again expanded recursively.

In order to reduce the number of explored composite motifs even further, we explore the expansions of a given composite motif in order of decreasing significance of single motifs. Note that the support of the composite motif before any new expansion is an upper bound on support. As the upper bounds are monotonically decreasing, we can stop exploring new expansions of a composite motif as soon as the upper bounds on support and significance are dominated by the Pareto front.

3.3.6 Automated subfamily discovery

The Pareto front of composite motifs for a family may contain significant motifs with relatively low values of support. It is natural to ask whether such a motif characterize a subfamily of the data set. One may therefore try to discover new motifs in the sequences that are not in the occurrence set of the first composite motif. Since the goal is to find motifs that are

common to as many sequences as possible, we have restricted automated subfamily discovery to only two subfamilies and also demand that one of the motifs belong to the Pareto front of the whole family. Significance values of motifs are log-likelihoods, and a two-subfamily-motif occur in a given sequence if either of the one-subfamily-motifs occur in the sequence. Therefore, the significance of a 2-subfamily-motif c is well approximated as: $s(c) = log_2(2^{s(c_a)} + 2^{s(c_b)} + 2^{s(c_a)+s(c_b)})$, where c_a and c_b are the one-subfamily-motifs.

3.4 Results and discussion

The family of AlkB homologues (ABHs) was used as a test case for composite motif discovery. The ABHs are members of the 2-oxoglutarate and Fe^{2+}-dependent (2OG-Fe(II)) oxygenase superfamily [5]. They have been shown to be involved in repair of methylation damage of DNA and RNA through a direct reversal mechanism, where the methyl group is oxidised and spontaneously released as formaldehyde [45]. Recent screening of databases using sensitive search methods has shown that ABH-like sequences are widespread in bacteria and eukaryotes, see Drabløs et al. [40] for a review.

The degree of sequence conservation in the ABH family seems to be very low, basically just a `H.D` motif, an isolated `H` and a `R.....R` motif (using single-letter amino acid symbols) is completely conserved in most ABH alignments. All except the final `R` are involved in coordination of the Fe^{2+} ion, the final `R` is probably involved in substrate binding as it seems to be relatively unique to the ABH family of this superfamily [5]. However, there may be subfamilies within the ABH family with more extensive conservation, and there may be additional conserved patterns in sequence regions that are difficult to align correctly by traditional methods. The ABH family is therefore an interesting test case with practical implications.

A set of 82 AHB-like sequences, previously investigated in [40], was used for the analysis. Teiresias [108] was used to generate 50.000 single motifs from the input sequences, and GCMD was used to identify the Pareto front for composite motifs with 2 and 4 components, using chemical equivalence sets for residue types (Fig. 3.1). The significance of the composite motifs is higher for most support values compared to single motifs. However, here GCMD is used mainly to identify interesting composite motifs and correlate this with biological significance.

Figure 3.1: Pareto front of single and composite motifs for ABH. Significance is the negative $\log_2$-likelihood of a motif

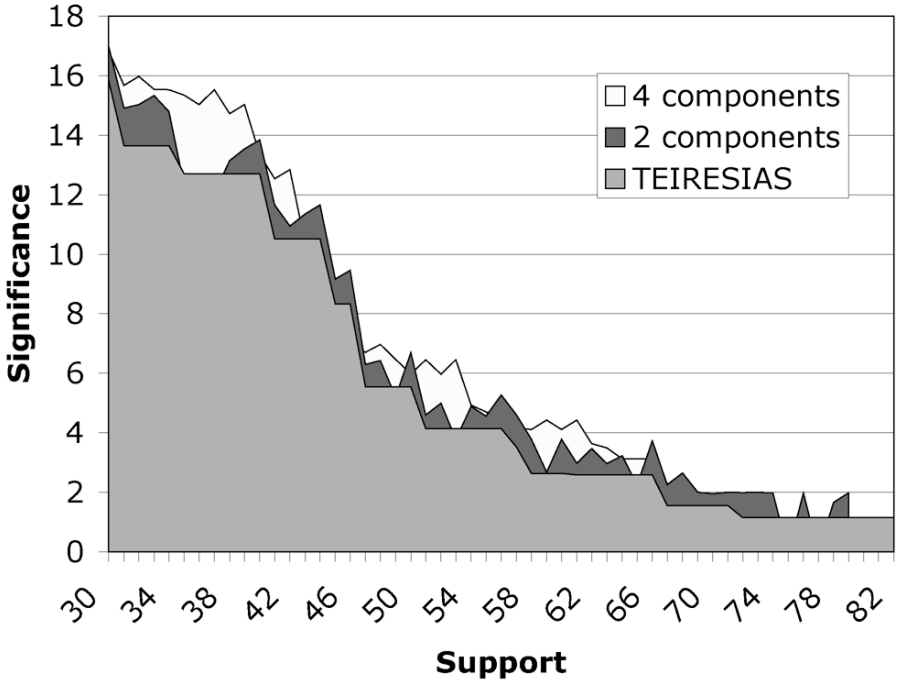
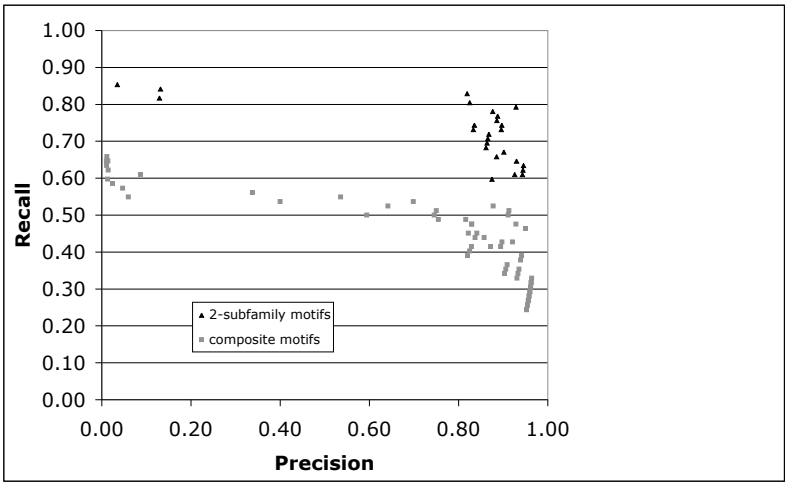


Figure 3.2: ROC of the discovered motifs for ABH. Recall is $TP/(TP+FN)$ and precision is $TP/(TP+FP)$)



The dominating single motif, which is used in most of the composite motifs, is `[ILMV]..H.[DE]`. This corresponds to the first Fe^{2+} binding motif in the ABH sequences. In particular for composite motifs with high support this is mainly combined with variants of the motif `[KR]..[ILMV]..[KR]`, which corresponds to the Fe^{2+} and possibly substrate binding `R` groups. This shows that the most interesting motifs identified by GCMD also have biological relevance, and that GCMD is able to identify such motifs from a large and complex set of input data.

However, it is evident that there are subfamilies of ABH-like sequences in the data set, and depending on the selected threshold for support several such subfamilies may be identified. One example is the composite motif `(L..G.[ILMV][ILMV].M....[QN]) & ([FY]....[DE].[ILMV]..H.D)`, which seems to be characteristic of the hABH2/hABH3 subfamily (human ABH type 2 and 3). This subfamily has been extensively studied experimentally [1]. As the detailed 3D structure of the ABH family still has not been experimentally determined, a detailed investigation of the biological relevance of these motifs probably has to be postponed until such data are available. However, this test shows that the GCMD method is able to identify biologically interesting subfamilies in a complex data set.

Although GCMD has not been developed as a classification tool, the classification performance may still serve to validate that the discovered motifs are indeed characteristic for a given family. Fig. 3.2 shows the receiver operating characteristic with respect to recall and precision when using the set of motifs in the Pareto front for classification. The introduction of subfamily motifs leads to a significant improvement in recall, and a larger fraction of the motifs have a high precision, compared to general composite motifs.

The performance of GCMD was also tested on 5 selected families from the PROSITE database. These PROSITE families are assumed to be difficult test cases, as the existing PROSITE patterns give low values for precision and recall. We used TEIRESIAS for single motif discovery. The Pareto front of composite motifs showed an average log-likelihood improvement of 20.4 compared to single motifs. The composite motifs in the Pareto front were used to classify the full set of SWISS-PROT [21] entries. For two of the five families (PS00485, PS00690) we were able to improve both precision and recall as compared to PROSITE, for two families we got comparable performance (PS00732, PS01048), and for the last family the PROSITE motif performed better (PS00187).

3.5 Conclusion

In our work we have built directly on previous work and focused on finding interesting combinations of single deterministic motifs discovered by separate motif discovery tools. Tests show that our tool is able to identify unique and biologically relevant composite motifs in very large data sets of single motifs.

Future directions of research include expanding the expressibility of deterministic motifs even further, as well as using the tool on other motif discovery problems, like for instance the discovery of transcription factor binding sites.

Acknowledgements

We want to thank Arne Halaas, Kai Trengereid, Magnus Lie Hetland, Rolv Seehuus, Tarjei Hveem and Øystein Lekang for help and suggestions. FD has been supported by FUGE/NFR (151899/150).

Chapter 4

Ph.D. research plan

This section describes my plans for future research. All the research projects described here concern motif discovery in DNA regulatory regions. As these projects require a close interplay between computer science and biology, I plan to cooperate closely with Finn Drabløs.

The main part of my research plan is the seven current research projects in section 4.1. I will start working on several of these in the coming autumn. Section 4.2 describes five directions for future research that are more general and open.

Together with Finn Drabløs, Arne Halaas and Magnus Lie Hetland, I have offered seven project tasks for students in the last year of their master studies. These tasks are based on my research projects in section 4.1, but are described more generally to make them understandable to students that are new to the field. The student tasks are described in section 4.3.

4.1 Current research projects

This section describes research projects that I will work on in the short term. Some of the research projects explore specific hypotheses, while other projects are more general and represent new computational approaches to the discovery of composite motifs.

4.1.1 Using GCMD for analysis of distance conservation in DNA composite motifs

Many authors argue that regulatory elements occur in combinations with strong restrictions on the distances between individual elements. This is typically modeled either as lower and upper constraints on the distance between consecutive single motifs, or by constraining all single motifs to occur within a window of a certain length.

In this project, I will investigate whether the most overrepresented composite motifs do indeed show strong restrictions on distance between individual elements. Moreover, I will systematically explore which of these two distance models that shows strongest conservation between occurrences in different regulatory regions.The GCMD algorithm, described in chapter 3, has functionality for analyzing inter-motif distances, and for grouping discovered motifs by distance criteria.

In this project the significance calculation of GCMD will be improved by using a higher-order Markov model as background. Furthermore, the conservation of distances will be directly incorporated in the significance calculation. Necessary adjustments will be made to ensure that GCMD can handle PWMs as single motifs, and suited methods for the generation of input motifs to GCMD will be found. Finally, results will be generated for several different sets of regulatory regions, from different genomes.

4.1.2 Comparing the ability of different sequence models to capture the variability between regulatory elements for a common factor

Both regular expressions, mismatch expressions (hamming distance), and position weight matrices have been extensively used as motif models in de novo motif discovery. As far as I know, no study has systematically compared the ability of these models to capture the sequence variability between related sites in proteins or between regulatory elements binding to the same transcription factor in DNA.

In de novo motif discovery, neither the underlying motif nor the locations of its occurrences are known. To compare motif models we look at the simpler situation with known occurrence locations. The performance of a motif model is then its ability to accurately separate a set of positive and a set of negative

substrings. By using the same set for both training and testing, we see whether it is at all possible to separate the data set with a given model. By using separate training and test sets, we get a measure of the generalization capability of a model. A systematic study of model performance as described here can show the limits of different motif models. De novo motif discovery is however a harder problem, and motifs discovered for complex models like PWMs, are in general not optimal. The conclusion of our study will therefore not be a direct answer on which model is best suited for de novo discovery.

In this project, methods have to be developed for each model to find the motif that is best able to separate a positive and negative set of substrings. Several methods have been developed that discover PWMs that best separates positive elements from a background [61]. A search in the literature may reveal efficient methods also for the discovery of regular expressions, and mismatch expressions, that best separates positive and negative substrings. If not, I will implement and use algorithms that I have sketched for the discovery of optimal motifs from each of these models.

4.1.3 Developing a customized method to discover frequent approximate item sets

Many different models have been used to represent combinations of single motifs. The by far most common approach is to model composite motifs as sets of single motifs. According to this model, a composite motif occurs in a region if, and only if, all of the component single motifs occur in a region. Recently, a method was published that can search for approximate combination matches [99]. This means that only M out of the N single motifs in a set has to occur in order for the composite motif to occur. The method can, however, not discover de novo motifs of this kind.

I have sketched an algorithm for de novo discovery of such M-of-N composite motifs. This algorithm is a variation of the algorithm used in GCMD. Much of the GCMD code can therefore be reused when implementing this new algorithm. The significance of motifs can be calculated with the same formulas as used in [99], and a Pareto front of composite motifs returned (as in GCMD). The algorithm could be applied on both protein families and DNA regulatory regions.

4.1.4 Extending the expressibility of composite motifs

A promising approach to single motif discovery is to first discover optimal motifs with a relatively simple model, and then use these as seeds for motif discovery with more expressive models [84].

I will explore a similar approach for the discovery of motif combinations. The highest scoring composite motifs from a simple combinatorial model will be used as seeds for discovery with a more expressive combinatorial model. Several different methods can be used to discover seed motifs, for instance the GCMD algorithm. The expressive combinatorial model will have weighted instead of binary scores for both single motifs and inter-motif distances. The motif model should be flexible with respect to distance conservation function. Expressive motifs will be discovered by some kind of heuristic search, locally around each seed.

4.1.5 Discovering over-represented combinations of signals

As pointed out by Sharan et al. [118], there are two basic ways of measuring overrepresentation of composite motifs in a data set D against a background B:

1. Which clusters occur more frequently in D than would be expected from their frequencies in B?

2. Which clusters occur more frequently in D than would be expected from the frequencies in D of their component (single) motifs?

A natural question is whether these two measures are highly correlated. As significant single motifs may correspond to regulatory elements, and composite motifs may correspond to combinatorially acting elements, it may be that the most significant composite motifs correspond to conserved combinations of elements that are also individually overrepresented. This would mean that the same composite motifs would get very high values of both measure (1) and (2). If the composite motifs with highest values for measure (2) do not have very high values for measure (1), this may be due to one of the following reasons:

- The composite motifs with highest values for measure (2) represent real modules, but the individual elements are not overrepresented by themselves.
- The composite motifs with highest values for measure (2) do not represent biological entities. The combination overrepresentation of the real modules are dominated by noise.

I will use GCMD to explore whether the motifs with highest score for measure (2) do indeed have high scores also for measure (1). I will use measure (1) for grouping of discovered motifs, and use measure (2) as the main significance measure. A 3-D plot will then show the Pareto front of most significant motifs for each grouping. Depending on the outcome of this study, further research questions may be posed.

4.1.6 Implementing a hardware-accelerated version of an established algorithm in bioinformatics

MEME [8] is a well-known and much used algorithm for discovering frequent patterns in biological sequences. The algorithm uses expectation maximization (EM) to discover overrepresented motifs in the form of PWMs (position weight matrices).

I will implement a version of expectation maximization that uses specialized hardware, the Pattern Matching Chip (PMC), to match motifs probabilistically against sequences. This corresponds to the expectation part of EM.

By using the PMC for pattern matching, I expect to speed up the algorithm considerably. Based on some initial calculations, I expect to process about 500 PWMs in 1000 gene regions of 5000 bp (base pairs) average length per second. This corresponds to sequentially calculating PWM scores for about 3GB of sequence data per second. The calculations are for one PCI card with 16 PMCs, and scales linearly with the number of PMCs.

The first goal of this project is to make a simple implementation of EM using the PMC, and then compare the computational performance of a PMC with that of a standard microprocessor. If the results are promising, a more advanced EM-based motif discovery algorithm may be implemented. This algorithm may take advantage of the computational efficiency to either explore a larger part of the search space, or use a more expressive motif model.

4.1.7 Using an evolutionary algorithm to discover composite motifs

A widespread hypothesis is that modules, i.e. combinations of regulatory elements, and not individual elements, are the primary determinants of gene expression. Furthermore it is believed that regulatory elements are extensively re-used across modules for diverse regulatory behaviour. The computational counterpart of this hypothesis is that single motifs are overrepresented in regulatory regions across the whole genome, while composite motifs scores are consistent with observed regulatory behaviour. Furthermore, the components (single motifs) of a significant composite motif are likely to be part of other composite motifs as well.

An algorithmic idea based on this, is to first discover a large number of single motifs that are overrepresented in regulatory regions across the whole genome. Composite motifs are then formed from combinations of these single motifs in an extensive search space of size N^M, where N is the number of distinct single motifs, and M is the number of components in a composite motif. Composite motifs with gene scores that are consistent with gene expression are considered significant. As some high scoring composite motifs are discovered, the components of these motifs should be favored by the search heuristic when forming new composite motifs.

Based on these considerations, evolutionary algorithms (EA) seems particularly suited for the discovery of composite motifs. A standard motif discovery method can be used to discover numerous single motifs that are overrepresented across all regulatory regions. The EA algorithm will then form an initial population of random combinations of these single motifs. The fitness of composite will be evaluated based on consistency between motif gene scores and gene expression (for instance as the sum of residuals of a linear regression from gene scores to gene log-expression). High scoring composite motifs will form the basis of future generations through selection, mutation and crossover. Cross-over between fit individuals will ensure that single motifs that are part of high scoring composite motifs, will more often be used as parts of new composite motifs. Mutation ensures that combinations involving unexplored single motifs are also occasionally evaluated.

An evolutionary algorithm as sketched above will be implemented, and the performance of this algorithm compared to that of GCMD with the same measure of motif significance. This will determine whether the evolutionary search strategy performs better in this situation than the branch-and-bound

algorithm GCMD.

4.2 Directions for future research

This section describes some future directions for research. These are directions I find promising, even though I do not have plans for concrete research projects. Being explicit about these directions may over time help in collecting relevant literature, establishing contacts and find inspiration about concrete research projects.

4.2.1 Iterative motif discovery at different levels of model complexity

The MDSCAN algorithm [84] first discovers the highest scoring motifs in a set of core sequences, using mismatch expressions to model motifs. These motifs are then used as seeds for the discovery of PWMs in the whole set of sequences. Somewhat similarly, Eskin et al. [42] use mismatch expressions as bounding boxes in the search space of PWMs, and thereby reduce the size of the search space of PWMs.

Especially the approach of Liu et al. [84] seems very promising. It shows an algorithm that takes advantage of additional information (in this case that some of the sequence are at the core of the set of sequences) not only to increase the sensitivity of the method, but also to increase the computational efficiency. Combining this idea with the iterative discovery of motifs at different complexity levels, could give rise to interesting research projects.

4.2.2 Developing a framework for the integration of motif discovery methods

More than a hundred different methods for motif discovery in regulatory regions have been published in recent years. As is apparent from chapter 2, the different methods extend on the basic approach in many different directions. Developing a framework that makes it possible to combine the strengths of several different methods therefore seems very promising.

The ideal approach is probably a framework that integrates several computational modules, concerning different parts of the problem. A possible division into computational modules is apparent from the integrated model in

chapter 2, including modules for priors based on gene expression, functional annotation, distance from TSS and phylogenetic footprinting, in addition to separate modules for single motif discovery, composite motif discovery, and calculation of gene score and motif significance.

Another kind of framework is a consensus model that uses many different methods independently to discover motifs, returning a consensus answer based on the answers from the different methods and possibly based on some context information. Neural networks has for instance been used in other research fields to reach a consensus decision from the decisions of many single predictors.

4.2.3 Exploring regression models from motif gene scores to gene expression

Some recent methods discover motifs that can serve as factors (dependents) of a linear regression from motif gene scores to gene expression [29, 34]. Similarly, logistic regression has been used from gene scores to binary expression values [142, 37, 117]. Bussemaker et al. [29] argues that only a single microarray experiment can be used in the regression approach, because the signal seems to disappear when expression is averaged over several experiments.

Many kinds of further exploration seems interesting. The current methods that discover motifs as factors of a linear regression, does this in an iterative, greedy fashion. It could be possible to develop algorithms that avoids the drawbacks of the greedy strategy. Moreover, regression models that are not only linear combinations of single motifs could be explored. Finally, the feasibility of more sophisticated approaches to the combination of expression values from different experiments could be explored. As long as the systematic differences between experiments is compensated for, the averaging of experiments should in principle reduce noise and thereby strengthen the signal of gene expression patterns.

4.2.4 Exploring background models for motif discovery in DNA

As overrepresentation is the key to computational discovery of regulatory elements, significance calculation is at the core of motif discovery. Many different measures are used for significance, but common to them all is that

they in some way compare the occurrence frequency in the positive set of sequences with a background model. The background is typically based on either the positive sequences, a set of (negative) sequences from other genomes, or sequences from other parts of the same genome. The background model can simply be the raw set of background sequences, but more often it is a higher order Markov model of the sequences, or a set of random shuffles of the sequences. The choice of data to use as background is of high importance, as it determines what is considered as overrepresented.

Many articles have focused on how to generate models from background data, and how to measure overrepresentation based on a background model. However, only a few methods explicitly consider what to use as negative data for motif discovery in regulatory regions. Takusagawa and Gifford [128] does this for the relatively simple organism S. cerevisiae (yeast).

Choosing, transforming and combining negative data are problems that seem both challenging and important. Improvements in these directions could improve the sensitivity of both new and existing algorithms for motif discovery in regulatory regions.

4.3 Student projects

This section describes project tasks I have offered to students writing a project in the ninth semester of their integrated master ("sivilingeniør"). These tasks are, as mentioned earlier, based on my research projects in section 4.1, but are described more generally to make them understandable to students new to the field.

The tasks are offered in cooperation with Finn Drabløs, Arne Halaas and Magnus Lie Hetland.

4.3.1 New methods for the discovery of DNA regulatory elements

One of the distinguishing features of human DNA is the complexity of the regulatory mechanism. This ensures that a proper amount of proteins are produced from a gene in different cells and at different times. An important part of this mechanism is played by transcription factors (TF) that bind to the DNA near a gene and enhance production of the gene. There is much interest in predicting binding sites (the positions where TFs can bind), and

many methods have been proposed. None of them perform well on complex organisms like humans.

The purpose of this project is to use novel criteria for what constitutes interesting binding sites. Different algorithms and machine learning methods may be applicable.

This task does not require any specific biological knowledge, but students should have a general interest in biology.

4.3.2 Developing a framework for the discovery of DNA regulatory elements

Regulatory elements are central to the behavior of a human or animal cell. More than a hundred different methods exist for the discovery of such elements in DNA. The discovery of regulatory elements consists of several aspects and subproblems. No single method is superior in all respects.

The task of this project is to develop a formal framework that makes it possible to integrate different methods that discover binding sites. Several computational modules, that correspond to the different subproblems, should be formalized. Furthermore, protocols for the flow of data and interaction between modules should be defined.

This task does not require any specific biological knowledge, but students should have a general interest in biology.

4.3.3 Using the PMC (hardware chip) for pattern discovery in DNA

The purpose of this task is to implement a version of a well-known algorithm (MEME), using a specialized hardware (PMC) to do the pattern matching.

MEME is a well-known and much used algorithm for discovering frequent patterns in biological sequences. This algorithm uses the established statistical optimization technique "Expectation Maximization" to discover position weight matrices (PWMs). A PWM is a probabilistic pattern that gives a weighted match against a sequence. By using the PMC to do the pattern matching, we expect to improve the running time of the algorithm considerably. A special interest in biology is not necessary for this task.

4.3.4 Developing algorithms for the discovery of pattern combinations

When several frequent patterns have been discovered in a text sequence, a next step is to discover overrepresented combinations of these patterns. Such pattern combinations are important in e.g. DNA sequences, but this project task can be done without any special interest or knowledge about biology.

4.3.5 Learning pattern models from examples

Several different pattern models are commonly used to classify a positive set of sequences from a negative set of sequences. One approach is to use regular expressions that match sequences in the positive set and do not match sequences in the negative set. A second approach is to use prototypic substrings that match all sequences that have at least M out of N symbols in common with the prototype. A third approach is to use weight matrices that assign a match score to each sequence, and use a threshold afterwards to determine matches.

The purpose of this task is to explore which pattern model is best suited to separate a set of real life positive sequences from negative sequences. To do this, methods have to be developed that learn the pattern that best separates positive and negative data for each kind of pattern model. The results when using these three methods on real data will then be compared to determine which pattern model works best. A special interest in biology is not necessary for this task.

4.3.6 Exploring the properties of "junk DNA"

One of the main approaches to data mining in DNA is to look for some kind of pattern that occurs unexpectedly often. The main idea is that such patterns may be frequent because they play an important biological function. Patterns that occur unexpectedly often may therefore serve as candidates for further study.

One important question is then "unexpectedly often compared to what?". The purpose of this task is to explore what can be used as negative data set. One possibility is to use parts of the genome that are believed to have no specific function, often referred to as "junk DNA". Another possibility is to

infer statistical models of the DNA sequence, typically as Markov models or by randomly shuffling the sequences.

This task does not require any specific biological knowledge, but students should have a general interest in biology.

4.3.7 Using an evolutionary algorithm for data mining in DNA

The discovery of pattern combinations in DNA is an important problem.

The purpose of this task is to use an evolutionary algorithm, for instance a genetic algorithm, to discover pattern combinations. Several different fitness functions could be tried, to see which fitness measure that gives results which are interpreted as most interesting by a biologist. Also, smart ways to do mutation and crossover will be explored.

This task does not require any specific biological knowledge, but students should have a general interest in biology.

Appendix A

An overview of motif discovery methods

This appendix shows the characteristics of 119 motif discovery methods with respect to the integrated model described in chapter 2. More specifically, table A.1 gives an overview of match models, occurrence priors and score functions on inter-motif distances. Table A.2, gives an overview of models of single motif combination, gene score functions and significance measures. As these aspects are not always described in articles presenting new methods, some fields are left blank.

Table A.1: Match model, occurrence prior and distance score for different methods

NR	ALGORITHM NAME	MATCH MODEL	OCC. PRIOR	DISTANCE FUNCTION
1	Pratt2[68]	reg.exp	–	–
2	MultiProfiler[71]	mismatch	–	–
3	Weeder[96]	mismatch	–	–
4	YMF[122, 123]	reg.exp	–	–
5	TEIRESIAS[108]	reg.exp	–	–
6	Splash[58]	reg.exp		–
7	Mitra[43]	mismatch	–	–
8	Mitra-dyad[43]	mismatch	–	constraint
9	Mot.Disc.Toolkit[10]	mismatch	–	
10	MERMAID[62]	PWM	–	constraint

Table A.1: Match model, occurrence prior and distance score for different methods

NR	ALGORITHM NAME	MATCH MODEL	OCC. PRIOR	DISTANCE FUNCTION
11	DMotifs[120]	reg.exp	–	constraint
12	Dyad analysis[139]	oligos	–	constraint
13	TFBScluster[38]	PWM	strand bias	window
14	MCAST[9]	PWM	–	gap penalty
15	GCMD[113]	flexible	–	constraint
16	[81]	mismatch	–	flexible
17	[57]	PWM	–	–
18	[151]	DM	–	–
19	[29]	PWM	–	constraint
20	MDScan[84]	PWM	chip	–
21	HMDM[146, 149]	DM	–	–
22	[12]	DM	–	–
23	Gibbs sampler[78]	PWM	–	uniform
24	MEME[8]	PWM	–	–
25	[28]	oligos	–	–
26	LOGOS[148]	DM	–	distribution
27	[27]	known sites	–	constraint
28	[74]	oligos	–	
29	MM[7]	PWM	–	–
30	Motif regressor[34]	PWM	–	–
31	SOMBERO[85]	PWM	–	–
32	MISAE[126]	mismatch	–	–
33	CENSUS[44]	mismatch	–	–
34	MScan[67]	PWM	–	
35	[119]	reg.exp	–	constraint
36	[99]		–	constraint
37	[52]		–	distribution
38	[24]	flexible	–	uniform
39	[121]		–	–
40	Oligo-analysis[137]	oligos	–	–
41	Pattern-assembly[138]		–	–
42	ModuleSearcher[3]	PWM	conservation	window

Table A.1: Match model, occurrence prior and distance score for different methods

NR	ALGORITHM NAME	MATCH MODEL	OCC. PRIOR	DISTANCE FUNCTION
43	[2]	PWM	–	window
44	COMET[53]		–	–
45	Stubb[124]	PWM	conservation	window
46	Modulescanner[3]	PWM	conservation	window
47	MotifLocator[3]	PWM	conservation	window
48	MotifSampler[129]	PWM	–	–
49	Footprinter[19]		–	–
50	GANN[13]	flexible	DNA struct.	window
51	FrameWorker[30]	PWM	–	constraint
52	[36]	oligos	conservation	–
53	[35]	oligos	–	–
54	[31]	oligos	–	–
55	[37]		–	–
56	MITRA-PSSM[42]	PWM	–	–
57	Partition-PSSM[42]	PWM	–	–
58	ModelGenerator[48]	PWM	–	distribution
59	ModelInspector[48]	PWM	–	distribution
60	GLAM[50]		–	–
61	DMS[63]	PWM	–	
62	ANN-Spec[144]	PWM	–	–
63	[142]	PWM	conservation	window
64	CoBind[56]	PWM	–	window
65	[102]	DM	–	–
66	OrthoMEME[101]	PWM	–	
67	WINNOWER[98]	mismatch	–	–
68	[73]	PWM	–	–
69	MAPPER[86]	HMM	–	
70	[64]	oligos	–	–
71	Footprinter[18, 17]	mismatch	–	–
72	[90]	PWM	–	–
73	Cister[51]	PWM	–	distribution
74	PromoterInsp.[114]	oligos	–	constraint

Table A.1: Match model, occurrence prior and distance score for different methods

NR	ALGORITHM NAME	MATCH MODEL	OCC. PRIOR	DISTANCE FUNCTION
75	[20]	PWM	–	uniform
76	SeSiMCMC [46]	PWM	–	–
77	FastM[76]	PWM	–	constraint
78	[88, 87]	mismatch	–	constraint
79	[140]	flexible	–	distribution
80	BioProspector[83]	PWM	strand bias	constraint
81	[117]	PWM	–	–
82	[122]	reg.exp	–	constraint
83	[134]	mismatch	–	–
84	ConsecID[118]	PWM	conservation	window
85	SCORE[105]	IUPAC	–	window
86	ClusterScan[72]	PWM	–	constraint
87	Gibbs recursive [132]	PWM	location	distribution
88	[95]	known sites	–	–
89	[94]	PWM	–	–
90	[14]	DM	–	–
91	Cis-analyst [16]	PWM	–	window
92	[60]	PWM	–	–
93	BioOptimizer[65]	PWM	–	constraint
94	[152]	DM	–	–
95	[115]	PWM	–	
96	[91]	PWM	–	–
97	[33]	oligos	–	–
98	Clover[49]	PWM	–	–
99	ProMapper[103]	DM	–	–
100	COOP[22]	reg.exp	–	–
101	CAGER[110]		–	–
102	AlignACE[109]	PWM	–	–
103	Consensus[59]	PWM	–	–
104	Improbizer[4]	PWM	–	–
105	QuickScore[106]	IUPAC	–	–
106	Motifprototyper[147]	DM	–	–

Table A.1: Match model, occurrence prior and distance score for different methods

NR	ALGORITHM NAME	MATCH MODEL	OCC. PRIOR	DISTANCE FUNCTION
107	CisModule[153]	PWM	–	mixture model
108	[104]	PWM	–	–
109	NONPAR [75]	Mixture	–	–
110	[47]	alignment	–	–
111	NestedMICA[39]	PWM	–	mixture model
112	[128]	reg.exp	–	–
113	Motif sampler[130]	PWM	–	–
114	[69]	PWM	–	uniform
115	[131]	PWM	conservation	constraint
116	ConSite[112, 80]	PWM	conservation	–
117	PhyloCon[141]	PWM	–	–
118	[54]	PWM	–	–
119	[116]	PWM	–	uniform

Table A.2: Composite motif model, gene score and significance evalution for different methods

NR	ALGORITHM NAME	MOTIF COMB.	GENE SCORE	SIGNIFI-CANCE
1	Pratt2[68]	–		
2	MultiProfiler[71]	–		
3	Weeder[96]	–	max	sum
4	YMF[122, 123]	–		
5	TEIRESIAS[108]	–		
6	Splash[58]	–	max	sum
7	Mitra[43]	–		
8	Mitra-dyad[43]	dyad		
9	Mot.Disc.Toolkit[10]	intersection		
10	MERMAID[62]	dyad		
11	DMotifs[120]	dyad		
12	Dyad analysis[139]	dyad	max	
13	TFBScluster[38]	intersection	sum	–

Table A.2: Composite motif model, gene score and significance evalution for different methods

NR	ALGORITHM NAME	MOTIF COMB.	GENE SCORE	SIGNIFICANCE
14	MCAST[9]	sum	HMM	classification
15	GCMD[113]	intersection	max	sum
16	[81]			
17	[57]	dictionary	sum	
18	[151]	–		
19	[29]	dyad	sum	regression
20	MDScan[84]	–	max	MAP
21	HMDM[146, 149]	–		
22	[12]	–		
23	Gibbs sampler[78]	intersection	max	p-value
24	MEME[8]	–	sum	IC of PWM
25	[28]	dictionary	special	
26	LOGOS[148]	HMM	HMM	
27	[27]	dyad		
28	[74]	sum	max	
29	MM[7]	–		
30	Motif regressor[34]	–	sum	regression
31	SOMBERO[85]	SOM		
32	MISAE[126]	–		
33	CENSUS[44]	–		
34	MScan[67]	min comp.score	max	
35	[119]	intersection	constraint	
36	[99]	mismatch	max	
37	[52]			
38	[24]	intersection	max	sum
39	[121]	–		
40	Oligo-analysis[137]	–	sum	sum
41	Pattern-assembly[138]	–		
42	ModuleSearcher[3]	sum	max	sum
43	[2]	sum	max	
44	COMET[53]	–		
45	Stubb[124]	HMM	HMM	–

Table A.2: Composite motif model, gene score and significance evalution for different methods

NR	ALGORITHM NAME	MOTIF COMB.	GENE SCORE	SIGNIFICANCE
46	Modulescanner[3]	sum	max	sum
47	MotifLocator[3]	sum	max	
48	MotifSampler[129]	–		
49	Footprinter[19]	–		
50	GANN[13]	ANN	ANN	
51	FrameWorker[30]	intersection	max	min
52	[36]	–	p-value	–
53	[35]	single motif	p-value	–
54	[31]	single motif	p-value	–
55	[37]	single motif		regression
56	MITRA-PSSM[42]	–	max	Discrete IC
57	Partition-PSSM[42]	–	max	Discrete IC
58	ModelGenerator[48]	sum		min
59	ModelInspector[48]	sum	max	min
60	GLAM[50]	–		
61	DMS[63]			sum
62	ANN-Spec[144]	–	max	IC of PWM
63	[142]	Logistic regr.	max	regression
64	CoBind[56]	sum	sum	sum
65	[102]	–	–	–
66	OrthoMEME[101]			sum
67	WINNOWER[98]	–	max	
68	[73]	–	max	sum
69	MAPPER[86]			
70	[64]	–	max	
71	Footprinter[18, 17]	–	max	sum
72	[90]	–		
73	Cister[51]	HMM	HMM	
74	PromoterInsp.[114]	intersection		
75	[20]	mixture model	mixture model	
76	SeSiMCMC [46]	–	mixture model	
77	FastM[76]	sum	max	

Table A.2: Composite motif model, gene score and significance evalution for different methods

NR	ALGORITHM NAME	MOTIF COMB.	GENE SCORE	SIGNIFICANCE
78	[88, 87]	intersection	max	sum
79	[140]	intersection		
80	BioProspector[83]	sum	sum	z-score
81	[117]	–	logistic func.	regression
82	[122]	dyad	sum	z-value
83	[134]	–	max	z-value
84	ConsecID[118]	intersection	sum	p-value
85	SCORE[105]	intersection	sum	p-value
86	ClusterScan[72]	sum	sum	
87	Gibbs recursive [132]	mixture model	mixture model	
88	[95]	special	special	special
89	[94]	–	hyperb. tan.	classification
90	[14]	–	–	classification
91	Cis-analyst [16]	sum	–	
92	[60]	–	max	special
93	BioOptimizer[65]	dyad	sum	sum
94	[152]	–		
95	[115]	sum	max	
96	[91]	–		
97	[33]	–	–	special
98	Clover[49]	–	sum	special
99	ProMapper[103]	–		
100	COOP[22]	–	–	
101	CAGER[110]	–		
102	AlignACE[109]	–	mixture model	p-value
103	Consensus[59]	–		IC of PWM
104	Improbizer[4]	–	mixture model	mixture model
105	QuickScore[106]	–		
106	Motifprototyper[147]	–		
107	CisModule[153]	mixture model	mixture model	
108	[104]	–	mixture model	
109	NONPAR [75]	–	–	

Table A.2: Composite motif model, gene score and significance evalution for different methods

NR	ALGORITHM NAME	MOTIF COMB.	GENE SCORE	SIGNIFICANCE
110	[47]	–	max	
111	NestedMICA[39]	mixture model	mixture model	
112	[128]	–	max	p-value
113	Motif sampler[130]	–	distribution	IC of PWM
114	[69]	intersection	max	expr. similarity
115	[131]	Markov model	max	
116	ConSite[112, 80]	–	–	–
117	PhyloCon[141]	–	sum	sum
118	[54]	–	–	special
119	[116]	dyad	max	p-valus

Appendix B

Some less prioritized research projects

In this appendix I have collected a few ideas for research projects that I currently do not intend to pursue. Still I want to have them documented in my research plan in case future insights make them more interesting either as research projects directly or as inspiration for other projects.

B.1 Comparing binding site variability in different genomes

The sequence variability of regulatory elements binding to the same transcription factor can be modeled by regular expressions, expressions allowing mismatches, standard PWMs, or by weight matrices that can also represent dependencies between motif positions. One important and much discussed question is whether the increased expressibility of expressive models is really needed. There are at least three different aspects of the variability that can legitimize complex models: the degree of dependency between positions, the difference in importance between positions in a motif, and the deviance of position distributions from those distributions possible to represent by the deterministic models.

An analysis of this for several genomes at different levels of organism complexity, could reveal insights about the complexity of regulatory mechanisms in different organisms. Morover, the relative performance of different motif models for organisms of different complexity could be determined.

B.2 Exploiting known DNA regulatory elements to discover new putative elements

Binding sites for transcription factors are believed to be organized in modules with restrictions on inter-motif distances. This combinatorial nature of binding sites can be used to filter discovered putative single motifs. If a new single motif, discovered based only on overrepresentation, occurs consistently in combination with occurrences of a known binding site, this strengthens the confidence in the new motif. In this way, known binding sites and knowledge about the combinatorial nature of regulatory elements, is used to evaluate biological significance of novel single motifs. This process can be repeated iteratively. The motifs discovered in one iteration can be considered as known motifs in the next iteration, and can thereby serve to strengthen the confidence in new motifs.

B.3 Motif discovery in mutation experiments

As described in chapter 2, computational motif discovery is typically based on overrepresentation of regulatory elements through reuse and conservation.

A different situation arises in mutation experiments with bacterias. Here, the behaviour of several genetically modified variants of a bacteria is observed. Such behaviour can for instance be the level of production of certain substances. The challenge is to extract sequence features that determine behaviour, and to predict the behaviour of hypothesized sequences in silico. In many cases the observed behaviour can be assigned a binary value, thus making it a classification problem (with a positive and negative set of (bacteria) sequences).

One approach to this problem is to discover a combination of sequence motifs that can explain the differences in behaviour of the bacterias. Algorithms could then be devised that discover sequence motifs based on the differences between sequences in the positive and negative data set. Efficient algorithms especially suited for the problem, possibly based on smart preprocessing of the data, will be necessary to find the combination of motifs that best separates the positive and negative data. Moreover, incorporating information about conservation between the original bacteria and its related species, could increase the sensitivity of the method.

Bibliography

[1] P. A. Aas, M. Otterlei, P. O. Falnes, C. B. Vaagboe, F. Skorpen, M. Akbari, O. Sundheim, M. Bjoras, G. Slupphaug, E. Seeberg, and H. E. Krokan. Human and bacterial oxidative demethylases repair alkylation damage in both RNA and DNA. *Nature*, 421:859–863, 2003.

[2] S. Aerts, P. Van Loo, Y. Moreau, and B. De Moor. A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes. *Bioinformatics*, 20(12):1974–6, 2004.

[3] S. Aerts, P. Van Loo, G. Thijs, Y. Moreau, and B. De Moor. Computational detection of cis -regulatory modules. *Bioinformatics*, 19 Suppl 2(1367-4803):II5–II14, 2003.

[4] W. Ao, J. Gaudet, W. J. Kent, S. Muttumu, and S. E. Mango. Environmentally induced foregut remodeling by pha-4/foxa and daf-12/nhr. *Science*, 305(5691):1743–6, 2004.

[5] L. Aravind and E. V. Koonin. The DNA-repair protein AlkB, EGL-9, and leprecan define new families of 2-oxoglutarate- and iron-dependent dioxygenases. *Genome Biol.*, 2(research0007.1–0007.8), 2001.

[6] T. K. Attwood, M. E. Beck, A. J. Bleasby, and D. J. Parry-Smith. PRINTS - a database of protein motif fingerprints. *Nucleic Acids Research*, 22(17):3590–3596, 1994.

[7] T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2(1553-0833):28–36, 1994.

[8] T. L. Bailey and C. Elkan. The value of prior knowledge in discovering motifs with meme. *Proc Int Conf Intell Syst Mol Biol*, 3(1553-0833): 21–9, 1995.

[9] T. L. Bailey and W. S. Noble. Searching for statistically significant regulatory modules. *Bioinformatics*, 19 Suppl 2(1367-4803):II16–II25, 2003.

[10] N. E. Baldwin, R. L. Collins, M. A. Langston, M. R. Leuze, C. T. Symons, and B. H. Voy. High performance computational tools for motif discovery. In *IEEE International Workshop on High Performance Computational Biology (HiCOMB)*, 2004.

[11] Z. Bar-Joseph, G. K. Gerber, T. I. Lee, N. J. Rinaldi, J. Y. Yoo, F. Robert, D. B. Gordon, E. Fraenkel, T. S. Jaakkola, R. A. Young, and D. K. Gifford. Computational discovery of gene modules and regulatory networks. *Nat Biotechnol*, 21(11):1337–42, 2003.

[12] Y. Barash, G. Elidan, N. Friedman, and T. Kaplan. Modeling dependencies in protein-dna binding sites. In *RECOMB '03: Proceedings of the seventh annual international conference on Computational molecular biology*, 2003.

[13] R. G. Beiko and R. L. Charlebois. Gann: genetic algorithm neural networks for the detection of conserved combinations of features in dna. *BMC Bioinformatics*, 6(1):36, 2005.

[14] I. Ben-Gal, A. Shani, A. Gohr, J. Grau, S. Arviv, A. Shmilovici, S. Posch, and I. Grosse. Identification of transcription factor binding sites with variable-order bayesian networks. *Bioinformatics*, pages 1367–4803, 2005.

[15] P. V. Benos, M. L. Bulyk, and G. D. Stormo. Additivity in protein-dna interactions: how good an approximation is it? *Nucleic Acids Res*, 30 (20):4442–51, 2002.

[16] B. P. Berman, Y. Nibu, B. D. Pfeiffer, P. Tomancak, S. E. Celniker, M. Levine, G. M. Rubin, and M. B. Eisen. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the drosophila genome. *Proc Natl Acad Sci U S A*, 99(2):757–62, 2002.

[17] M. Blanchette, B. Schwikowski, and M. Tompa. An exact algorithm to identify motifs in orthologous sequences from multiple species. *Proc Int Conf Intell Syst Mol Biol*, 8(1553-0833):37–45, 2000.

[18] M. Blanchette and M. Tompa. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res*, 12 (5):739–48, 2002.

[19] M. Blanchette and M. Tompa. Footprinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res*, 31(13):3840–2, 2003.

[20] K. Blekas, D. I. Fotiadis, and A. Likas. Greedy mixture learning for multiple motif discovery in biological sequences., 2003.

[21] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Res*, 31(1):365–70, 2003.

[22] S. Bortoluzzi, A. Coppe, A. Bisognin, C. Pizzi, and G. Danieli. A multistep bioinformatic approach detects putative regulatory elements in gene promoters. *BMC Bioinformatics*, 6(1):121, 2005.

[23] A. Brazma, I. Jonassen, I. Eidhammer, and D. Gilbert. Approaches to the automatic discovery of patterns in biosequences. *J Comput Biol*, 5 (2):279–305, 1998.

[24] A. Brazma, J. Vilo, E. Ukkonen, and K. Valtonen. Data mining for regulatory elements in yeast genome. *Proc Int Conf Intell Syst Mol Biol*, 5(1553-0833):65–74, 1997.

[25] P. Bucher and A. Bairoch. A generalized profile syntax for biomolecular sequence motifs and its fuction in automatic sequence interpretation. In *Proc Int Conf Intell Syst Mol Biol*, volume 2, pages 53–61, 1994.

[26] M. J. Buck and J. D. Lieb. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83(3):349–60, 2004.

[27] M. L. Bulyk, A. M. McGuire, N. Masuda, and G. M. Church. A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in escherichia coli. *Genome Res*, 14(2):201–8, 2004.

[28] H. J. Bussemaker, H. Li, and E. D. Siggia. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc Natl Acad Sci U S A*, 97(18):10096–100, 2000.

[29] H. J. Bussemaker, H. Li, and E. D. Siggia. Regulatory element detection using correlation with expression. *Nat Genet*, 27(2):167–71, 2001.

[30] K. Cartharius, K. Frech, K. Grote, B. Klocke, M. Haltmeier, A. Klingenhoff, M. Frisch, M. Bayerlein, and T. Werner. Matinspector and beyond: promoter analysis based on transcription factor binding sites., 2005.

[31] M. Caselle, F. Di Cunto, and P. Provero. Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes. *BMC Bioinformatics*, 3(1):7, 2002.

[32] S. Cawley. *Statistical models for DNA sequencing and analysisspliceosome: motors, clocks, springs, and things. Cel l, Statistical models for DNA sequencing and analysis*. PhD thesis, University of California at Berkely, Berkely, CA, 2000.

[33] P. Cliften, P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, J. Majors, R. Waterston, B. A. Cohen, and M. Johnston. Finding functional features in saccharomyces genomes by phylogenetic footprinting. *Science*, 301(5629):71–6, 2003.

[34] E. M. Conlon, X. S. Liu, J. D. Lieb, and J. S. Liu. Integrating regulatory motif discovery and genome-wide expression analysis. *Proc Natl Acad Sci U S A*, 100(6):3339–44, 2003.

[35] D. Cora, F. Di Cunto, P. Provero, L. Silengo, and M. Caselle. Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrepresented upstream motifs. *BMC Bioinformatics*, 5(1):57, 2004.

[36] D. Cora, C. Herrmann, C. Dieterich, F. Di Cunto, P. Provero, and M. Caselle. Ab initio identification of putative human transcription factor binding sites by comparative genomics. *BMC Bioinformatics*, 6 (1):110, 2005.

[37] M. D. Curran, H. Liu, F. Long, and N. Ge. Statistical methods for joint data mining of gene expression and dna sequence database. *SIGKDD Explor Newsl*, 5(2):122–129, 2003.

[38] I. J. Donaldson, M. Chapman, and B. Gottgens. TFBScluster: a resource for the characterisation of transcriptional regulatory networks. *Bioinformatics*, pages 1367–4803, 2005.

[39] T. A. Down and T. J. P. Hubbard. Nestedmica: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Res*, 33 (5):1445–53, 2005.

[40] F. Drabløs, E. Feyzi, P. A. Aas, C. B. Vaagboe, B. Kavli, M. S. Bratlie, J. Peña-Diaz, M. Otterlei, G. Slupphaug, and H. E. Krokan. Alkylation damage in DNA and RNA–repair mechanisms and medical significance. *DNA Repair*, 3:1389–1407, 2004.

[41] M. A. El Hassan and C. R. Calladine. Conformational characteristics of dna: empirical classifications and a hypothesis for the conformational behaviour of dinucleotide steps. *Roy Soc of London Phil Tr A*, 355 (1722):43–100, 1997.

[42] E. Eskin, W. Noble, Y. Singer, and S. Snir. A unified approach for sequence prediction using sparse sequence models. Technical report, Hebrew University, 2003.

[43] E. Eskin and P. A. Pevzner. Finding composite regulatory patterns in dna sequences., 2002.

[44] P. A. Evans and A. D. Smith. Toward optimal motif enumeration. In *Proceedings of Workshop on Algorithms and Data Structures (WADS 2003)*, volume 2751 of *LNCS*, pages 47–58. Springer-Verlag, 2003.

[45] P. O. Falnes, R. F. Johansen, and E. Seeberg. AlkB-mediated oxidative demethylation reverses DNA damage in Escherichia Coli. *Nature*, 419: 178–182, 2002.

[46] A. V. Favorov, M. S. Gelfand, A. V. Gerasimova, D. A. Ravcheev, A. A. Mironov, and V. J. Makeev. A gibbs sampler for identification of symmetrically structured, spaced dna motifs with improved estimation of the signal length. *Bioinformatics*, 21(10):2240–2245, 2005.

[47] G. B. Fogel, D. G. Weekes, G. Varga, E. R. Dow, H. B. Harlow, J. E. Onyia, and C. Su. Discovery of sequence motifs related to coexpression of genes using evolutionary computation. *Nucleic Acids Res*, 32(13): 3826–35, 2004.

[48] K. Frech and T. Werner. Specific modelling of regulatory units in dna sequences. *Pac Symp Biocomput*, pages 151–62, 1997.

[49] M. C. Frith, Y. Fu, L. Yu, J.-F. Chen, U. Hansen, and Z. Weng. Detection of functional dna motifs via statistical over-representation. *Nucleic Acids Res*, 32(4):1372–81, 2004.

[50] M. C. Frith, U. Hansen, J. L. Spouge, and Z. Weng. Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res*, 32 (1):189–200, 2004.

[51] M. C. Frith, U. Hansen, and Z. Weng. Detection of cis-element clusters in higher eukaryotic dna. *Bioinformatics*, 17(10):878–89, 2001.

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

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

[54] N. I. Gershenzon, G. D. Stormo, and I. P. Ioshikhes. Computational technique for improvement of the position-weight matrices for the dna/protein binding sites. *Nucleic Acids Res*, 33(7):2290–301, 2005.

[55] Y. H. Grad, F. P. Roth, M. S. Halfon, and G. M. Church. Prediction of similarly-acting cis-regulatory modules by subsequence profiling and comparative genomics in D. melanogaster and D. pseudoobscura. *Bioinformatics*, 20(16):2738–2750, May 4 2004.

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

[57] M. Gupta and J. S. Liu. Discovery of conserved sequence patterns using a stochastic dictionary model. *Journal of the American Statistical Association*, 98:55–66, 2003.

[58] R. K. Hart, A. K. Royyuru, G. Stolovitzky, and A. Califano. Systematic and fully automated identification of protein sequence patterns. *J Comput Biol*, 7(3-4):585–600, 2000.

[59] G. Z. Hertz and G. D. Stormo. Identifying dna and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7-8):563–77, 1999.

[60] I. Holmes and W. J. Bruno. Finding regulatory elements using joint likelihoods for sequence and expression profile data. *Proc Int Conf Intell Syst Mol Biol*, 8(1553-0833):202–10, 2000.

[61] P. B. Horton and M. Kanehisa. An assessment of neural network and statistical approaches for prediction of e. coli promoter sites. *Nucleic Acids Res*, 20(16):4331–8, 1992.

[62] Y.-J. Hu. Finding subtle motifs with variable gaps in unaligned dna sequences. *Comput Methods Programs Biomed*, 70(1):11–20, 2003.

[63] Y. J. Hu, S. Sandmeyer, C. McLaughlin, and D. Kibler. Combinatorial motif analysis and hypothesis generation on a genomic scale. *Bioinformatics*, 16(3):222–32, 2000.

[64] L. J. Jensen and S. Knudsen. Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation. *Bioinformatics*, 16(4):326–33, 2000.

[65] S. T. Jensen and J. S. Liu. Biooptimizer: a bayesian scoring function approach to motif discovery. *Bioinformatics*, 20(10):1557–64, 2004.

[66] S. T. Jensen, X. S. Liu, J. S. Liu, and Q. Zhou. Computational discovery of gene regulatory binding motifs: A bayesian perspective. *Statist Sci*, 19(1):188–204, 2004.

[67] O. Johansson, W. Alkema, W. W. Wasserman, and J. Lagergren. Identification of functional clusters of transcription factor binding motifs in genome sequences: the mscan algorithm. *Bioinformatics*, 19 Suppl 1 (1367-4803):i169–76, 2003.

[68] I. Jonassen. Efficient discovery of conserved patterns using a pattern graph. *Comput Appl Biosci*, 13(5):509–22, 1997.

[69] J.-G. Joung, S. J. Oh, and B.-T. Zhang. Searching transcriptional modules using evolutionary algorithms. In *Parallel Problem Solving from Nature - PPSN VIII*, volume 3242 of *Lecture Notes in Computer Science*, pages 532–540, Berlin, 2004. Springer-Verlag.

[70] M. Kato, N. Hata, N. Banerjee, B. Futcher, and M. Q. Zhang. Identifying combinatorial regulation of transcription factors and binding motifs. *Genome Biol*, 5(8):R56, 2004.

[71] U. Keich and P. A. Pevzner. Finding motifs in the twilight zone., 2002.

[72] A. Kel, O. Kel-Margoulis, T. Ivanova, and E. Wingender. Clusterscan: A tool for automatic annotation of genomic regulatory sequences by searching for composite clusters. In *German conference on bioinformatics*, 2001.

[73] A. Kel, Y. Tikunov, N. Voss, and E. Wingender. Recognition of multiple patterns in unaligned sets of sequences: comparison of kernel clustering method with other methods., 2004.

[74] S. Keles, M. van der Laan, and M. B. Eisen. Identification of regulatory elements using a feature selection method., 2002.

[75] O. D. King and F. P. Roth. A non-parametric model for transcription factor binding sites. *Nucleic Acids Res*, 31(19):e116, 2003.

[76] A. Klingenhoff, K. Frech, K. Quandt, and T. Werner. Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics*, 15(3):180–6, 1999.

[77] W. Krivan and W. W. Wasserman. A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res*, 11(9): 1559–66, 2001.

[78] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–14, 1993.

[79] C. E. Lawrence and A. A. Reilly. An expectation maximization (em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, 7(1):41–51, 1990.

[80] B. Lenhard, A. Sandelin, L. Mendoza, P. Engstrom, N. Jareborg, and W. W. Wasserman. Identification of conserved regulatory elements by comparative genome analysis. *J Biol*, 2(2):13, 2003.

[81] H.-L. Li and C.-J. Fu. A linear programming approach for identifying a consensus sequence on dna sequences. *Bioinformatics*, 21(9):1838–45, 2005.

[82] L. P. Lim and C. B. Burge. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A*, 98(20):11193–8, 2001.

[83] X. Liu, D. L. Brutlag, and J. S. Liu. Bioprospector: discovering conserved dna motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*, pages 127–38, 2001.

[84] X. S. Liu, D. L. Brutlag, and J. S. Liu. An algorithm for finding protein-dna binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol*, 20(8):835–9, 2002.

[85] S. Mahony, D. Hendrix, A. Golden, T. J. Smith, and D. S. Rokhsar. Transcription factor binding site identification using the self-organizing map. *Bioinformatics*, 21(9):1807–14, 2005.

[86] V. D. Marinescu, I. S. Kohane, and A. Riva. Mapper: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. *BMC Bioinformatics*, 6(1):79, 2005.

[87] L. Marsan and M. F. Sagot. Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *J Comput Biol*, 7(3-4):345–62, 2000.

[88] L. Marsan and M.-F. Sagot. Extracting structured motifs using a suffix treealgorithms and application to promoter consensus identification. In

RECOMB '00: Proceedings of the fourth annual international conference on Computational molecular biology, pages 210–219, New York, NY, USA, 2000. ACM Press. ISBN 1-58113-186-0.

[89] L. McCue, W. Thompson, C. Carmack, M. P. Ryan, J. S. Liu, V. Derbyshire, and C. E. Lawrence. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res*, 29 (3):774–82, 2001.

[90] A. M. McGuire, J. D. Hughes, and G. M. Church. Conservation of dna regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res*, 10(6):744–57, 2000.

[91] A. A. Mironov, E. V. Koonin, M. A. Roytberg, and M. S. Gelfand. Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucleic Acids Res*, 27(14):2981–9, 1999.

[92] A. F. Neuwald, J. S. Liu, and C. E. Lawrence. Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci*, 4 (8):1618–32, 1995.

[93] R. A. O'Flanagan, G. Paillard, R. Lavery, and A. M. Sengupta. Non-additivity in protein-dna binding. *Bioinformatics*, 21(10):2254–63, 2005.

[94] H. P., L. X.S., Z. Q., L. X., L. J. S., and W. W. H. A boosting approach for motif modeling using chip-chip data. *Bioinformatics*, pages 1367–4803, 2005.

[95] P. J. Park, A. J. Butte, and I. S. Kohane. Comparing expression profiles of genes with similar promoter regions. *Bioinformatics*, 18(12):1576–84, 2002.

[96] G. Pavesi, G. Mauri, and G. Pesole. An algorithm for finding signals of unknown length in dna sequences. *Bioinformatics*, 17 Suppl 1(1367-4803):S207–14, 2001.

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

[98] P. A. Pevzner and S. H. Sze. Combinatorial approaches to finding subtle signals in dna sequences. *Proc Int Conf Intell Syst Mol Biol*, 8 (1553-0833):269–78, 2000.

[99] A. Policriti, N. Vitacolonna, M. Morgante, and A. Zuccolo. Structured motifs search. In *RECOMB '04: Proceedings of the eighth annual international conference on Computational molecular biology*, pages 133–139, New York, NY, USA, 2004. ACM Press. ISBN 1-58113-755-9.

[100] J. V. Ponomarenko, M. P. Ponomarenko, A. S. Frolov, D. G. Vorobyev, G. C. Overton, and N. A. Kolchanov. Conformational and physicochemical dna features specific for transcription factor binding sites. *Bioinformatics*, 15(7-8):654–68, 1999.

[101] A. Prakash, M. Blanchette, S. Sinha, and M. Tompa. Motif discovery in heterogeneous sequence data. *Pac Symp Biocomput*, pages 348–59, 2004.

[102] R. Pudimat, E. G. Schukat-Talamazzini, and R. Backofen. Feature based representation and detection of transcription factor binding sites. In *German Conference on Bioinformatics*, pages 43–52, 2004.

[103] R. Pudimat, E. G. Schukat-Talamazzini, and R. Backofen. A multiple-feature framework for modelling and predicting transcription factor binding sites. *Bioinformatics*, (1367-4803), 2005.

[104] Z. S. Qin, L. A. McCue, W. Thompson, L. Mayerhofer, C. E. Lawrence, and J. S. Liu. Identification of co-regulated genes through bayesian clustering of predicted regulatory binding sites. *Nat Biotechnol*, 21(4): 435–9, 2003.

[105] M. Rebeiz, N. L. Reeves, and J. W. Posakony. Score: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. site clustering over random expectation. *Proc Natl Acad Sci U S A*, 99(15):9888–93, 2002.

[106] M. Régnier and A. Denise. Rare events and conditional events on random strings. *Discrete Math. Theor. Comput. Sci.*, 6:191–214, 2004.

[107] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert,

C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of dna binding proteins. *Science*, 290(5500):2306–9, 2000.

[108] I. Rigoutsos and A. Floratos. Combinatorial pattern discovery in biological sequences: The teiresias algorithm. *Bioinformatics*, 14(1):55–67, 1998.

[109] F. P. Roth, J. D. Hughes, P. W. Estep, and G. M. Church. Finding dna regulatory motifs within unaligned noncoding sequences clustered by whole-genome mrna quantitation. *Nat Biotechnol*, 16(10):939–45, 1998.

[110] J. Ruan and W. Zhang. Cager: classification analysis of gene expression regulation using multiple information sources. *BMC Bioinformatics*, 6 (1):114, 2005.

[111] G. Rustici, J. Mata, K. Kivinen, P. Lio, C. J. Penkett, G. Burns, J. Hayles, A. Brazma, P. Nurse, and J. Bahler. Periodic gene expression program of the fission yeast cell cycle. *Nat Genet*, 36(8):809–17, 2004.

[112] A. Sandelin, W. W. Wasserman, and B. Lenhard. Consite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res*, 32(Web Server issue):W249–52, 2004.

[113] G. K. Sandve and F. Drabløs. Generalized composite motif discovery. In *7th Int. Conf. on Knowledge-Based Intelligent Information and Engineering Systems, KES*, volume In press of *LNCS/LNAI*. Springer-Verlag, 2005.

[114] M. Scherf, A. Klingenhoff, and T. Werner. Highly specific localization of promoter regions in large genomic sequences by promoterinspector: a novel context analysis approach. *J Mol Biol*, 297(3):599–606, 2000.

[115] E. Segal, Y. Barash, I. Simon, N. Friedman, and D. Koller. From promoter sequence to expression: a probabilistic framework. In *RECOMB '02: Proceedings of the sixth annual international conference on Computational biology*, pages 263–272, New York, NY, USA, 2002. ACM Press. ISBN 1-58113-498-3.

[116] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules

and their condition-specific regulators from gene expression data. *Nat Genet*, 34(2):166–76, 2003.

[117] E. Segal, R. Yelensky, and D. Koller. Genome-wide discovery of transcriptional modules from dna sequence and gene expression. *Bioinformatics*, 19 Suppl 1(1367-4803):i273–82, 2003.

[118] R. Sharan, I. Ovcharenko, A. Ben-Hur, and R. M. Karp. Creme: a framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics*, 19 Suppl 1(1367-4803):i283–91, 2003.

[119] D. Shinozaki and O. Maruyama. A method for the best model selection for single and paired motifs. In *Genome Informatics*, volume 13, pages 432–433. Universal Academy Press, 2002.

[120] S. Sinha. Discriminative motifs. *J Comput Biol*, 10(3-4):599–615, 2003.

[121] S. Sinha, M. Blanchette, and M. Tompa. Phyme: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, 5(1):170, 2004.

[122] S. Sinha and M. Tompa. A statistical method for finding transcription factor binding sites. *Proc Int Conf Intell Syst Mol Biol*, 8(1553-0833): 344–54, 2000.

[123] S. Sinha and M. Tompa. Ymf: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res*, 31(13):3586–8, 2003.

[124] S. Sinha, E. van Nimwegen, and E. D. Siggia. A probabilistic method to detect regulatory modules. *Bioinformatics*, 19 Suppl 1(1367-4803): i292–301, 2003.

[125] G. D. Stormo, T. D. Schneider, L. Gold, and A. Ehrenfeucht. Use of the 'perceptron' algorithm to distinguish translational initiation sites in e. coli. *Nucleic Acids Res*, 10(9):2997–3011, 1982.

[126] Z. Sun, J. Yang, and J. S. Deogun. Misae: A new approach for regulatory motif extraction. In *CSB '04: Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB'04)*, pages 173–181, Washington, DC, USA, 2004. IEEE Computer Society. ISBN 0-7695-2194-0.

[127] D. A. Tagle, B. F. Koop, M. Goodman, J. L. Slightom, D. L. Hess, and R. T. Jones. Embryonic epsilon and gamma globin genes of a prosimian primate (galago crassicaudatus). nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol*, 203 (2):439–55, 1988.

[128] K. T. Takusagawa and D. K. Gifford. Negative information for motif discovery. *Pac Symp Biocomput*, pages 360–71, 2004.

[129] G. Thijs, M. Lescot, K. Marchal, S. Rombauts, B. De Moor, P. Rouze, and Y. Moreau. A higher-order background model improves the detection of promoter regulatory elements by gibbs sampling. *Bioinformatics*, 17(12):1113–22, 2001.

[130] G. Thijs, K. Marchal, M. Lescot, S. Rombauts, B. De Moor, P. Rouze, and Y. Moreau. A gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J Comput Biol*, 9 (2):447–64, 2002.

[131] W. Thompson, M. J. Palumbo, W. W. Wasserman, J. S. Liu, and C. E. Lawrence. Decoding human regulatory circuits. *Genome Res*, 14(10A): 1967–74, 2004.

[132] W. Thompson, E. C. Rouchka, and C. E. Lawrence. Gibbs recursive sampler: finding transcription factor binding sites. *Nucleic Acids Res*, 31(13):3580–5, 2003.

[133] H. Toivonen. *Discovery of Frequent Patterns in Large Data Collections*. PhD thesis, University of Helsinki, 1996.

[134] M. Tompa. An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. In *Proc Int Conf Intell Syst Mol Biol*, number 1553-0833, pages 262–71, Heidelberg, Germany, August 1999.

[135] M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. De Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. Assessing computational tools for the

discovery of transcription factor binding sites. *Nat Biotechnol*, 23(1): 137–44, 2005.

[136] R. van Driel, P. F. Fransz, and P. J. Verschure. The eukaryotic genome: a system regulated at different hierarchical levels., 2003.

[137] J. van Helden, B. Andre, and J. Collado-Vides. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol*, 281(5):827–42, 1998.

[138] J. van Helden, B. Andre, and J. Collado-Vides. A web site for the computational analysis of yeast regulatory sequences. *Yeast*, 16(2): 177–87, 2000.

[139] J. van Helden, A. F. Rios, and J. Collado-Vides. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res*, 28(8):1808–18, 2000.

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

[141] T. Wang and G. D. Stormo. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, 19(18): 2369–80, 2003.

[142] W. W. Wasserman and J. W. Fickett. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol*, 278(1): 167–81, 1998.

[143] T. Werner. Models for prediction and recognition of eukaryotic promoters. *Mamm Genome*, 10(2):168–75, 1999.

[144] C. T. Workman and G. D. Stormo. ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac Symp Biocomput*, pages 467–78, 2000.

[145] G. A. Wray, M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman, and L. A. Romano. The evolution of transcriptional regulation in eukaryotes., 2003.

[146] E. Xing, M. Jordan, R. Karp, and S. Russell. A hierarchical bayesian markovian model for motifs in biopolymer sequences. In S. Becker, S. andThrun and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, Cambridge, MA, 2002.

[147] E. P. Xing and R. M. Karp. Motifprototyper: a bayesian profile model for motif families. *Proc Natl Acad Sci U S A*, 101(29):10523–8, 2004.

[148] E. P. Xing, W. Wu, M. I. Jordan, and R. M. Karp. Logos: a modular bayesian model for de novo motif detection. *J Bioinform Comput Biol*, 2(1):127–54, 2004.

[149] E. P. Xing, W. Wu, and R. M. Karp. Capturing characteristic structural features for motif detection using a hierarchical bayesian markovian model. *Genome Biol*, 2003.

[150] Z. Zhang and M. Gerstein. Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. *J Biol*, 2(2):11, 2003.

[151] X. Zhao, H. Huang, and T. P. Speed. Finding short dna motifs using permuted markov models. In *RECOMB '04: Proceedings of the eighth annual international conference on Computational molecular biology*, pages 68–75, New York, NY, USA, 2004. ACM Press. ISBN 1-58113-755-9.

[152] Q. Zhou and J. S. Liu. Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, 20(6):909–16, 2004.

[153] Q. Zhou and W. H. Wong. Cismodule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc Natl Acad Sci U S A*, 101(33):12114–9, 2004.