Potentials and limitations of motif-based binding site prediction in DNA

Geir Kjetil Ferkingstad Sandve

Doctoral Thesis

Submitted for the Partial Fulfillment of the Requirements for the Degree of

philosophiae doctor

Department of Computer and Information Science
Faculty of Information Technology, Mathematics and Electrical Engineering
Norwegian University of Science and Technology

May 16, 2008

Abstract

As the full genomic DNA sequence is now available for several organisms, a major next challenge is determining the function of DNA elements. This task is often referred to as functional genomics. An important part of functional genomics is gene regulation, and particularly the binding of specific proteins called Transcription Factors (TFs) to DNA. This TF binding regulates the production of mRNA, and thereby eventually proteins, from genes. As experimental determination of TF binding sites in DNA is a very laborious process, there is great interest in computational prediction methods.

The basic idea behind computational binding site prediction is to use motifs (sequence patterns) to capture sequence similarity between separate binding sites for a given TF. Based on a set of known binding site examples, the sequence similarity can be exploited for prediction of additional binding sites for a given TF. As motifs representing TF binding sites should occur more frequently than expected by chance alone in co-regulated DNA sequences, computational methods can even be used to discover novel TF binding site motifs and associated binding sites using only un-annotated target DNA sequences as input.

The focus of this thesis is on the computational prediction of TF binding sites, and specifically on understanding the current limitations and potential for improvement of binding site prediction. Two of the papers in the thesis relate to the assessment of computational predictions. The data sets used in a recent benchmark of prediction methods is analyzed in relation to three commonly used motif models, showing some fundamental performance limitations that should be attributed either to the motif models or to the benchmark data sets themselves. A first broad benchmark of methods predicting higher-order organization of TF binding sites is also part of this thesis. The benchmark showed some differences in prediction accuracy between methods, and more generally that a moderate level of prediction accuracy can be expected in the considered scenario.

Two novel motif discovery methods are also presented in the thesis. Both of the methods consider the problem of predicting higher-order organization of binding sites, given motifs representing binding of individual TFs as input. One method takes a Bayesian probabilistic approach to binding site modeling, while the other method uses a discrete approach. Both methods use highly expressive models and show good quantitative performance in relation to existing methods. Each method also introduces some additional elements that may bring qualitative advantages.

A third and final direction of research in this thesis concerns the extended process of motif discovery in DNA. Topics considered include how data is compiled before binding site prediction is performed, how prediction results can be interpreted in a multiple-testing scenario, and how prediction can be accelerated by the use of parallel hardware.

Contents

List of Figures

Preface

This thesis was submitted to the Norwegian University of Science and Technology (NTNU) in partial fullment of the requirements for the degree of philosophiae doctor (PhD). It is organised as a collection of papers, with a research overview, consisting of an introduction, a research description, and conclusions, given in the first part. The articles, following the originally published text, can be found in the second part.

This doctoral work has been performed at the Department of Computer and Information Science, NTNU, with Professor Arne Halaas as main supervisor and with co-supervisor Professor Finn Drabløs (Department of Cancer Research and Molecular Medicine, NTNU). The work was funded by the Faculty of Information Technology, Mathematics and Electrical Engineering, NTNU.

On joint authorship

All papers in this thesis have been made in collaboration with others. For several of the included papers, a description of author contributions was part of the original paper. In these papers my own contribution has been extracted from the original description and only slightly reformulated. For the other papers, a brief description of my own contributions has been written for the purpose of this thesis.

Paper 1: Geir Kjetil Sandve developed the basic analysis perspective and drafted the main parts of the manuscript (all sections except "Biological background" and "Correspondence with experimental data").

Paper 2: Geir Kjetil Sandve contributed to experimentation and evaluation of the method, drafted parts of the manuscript, and took part in writing

the full paper.

Paper 3: Geir Kjetil Sandve conceived the initial idea, devised the algorithms, implemented the method and drafted the main parts of the manuscript.

Paper 4: Geir Kjetil Sandve conceived the initial idea together with Finn Drabløs, devised the discrimination algorithms and drafted the manuscript.

Paper 5: Geir Kjetil Sandve and Osman Abul conceived of the study. All authors participated in the design of the study. Kjetil Klepper and Geir Kjetil Sandve assembled the datasets. All authors helped revise and approved the final manuscript.

Paper 6: Geir Kjetil Sandve took part in scientific discussions on the study and writing of the manuscript.

Paper 7: Geir Kjetil Sandve took part in scientific discussions on the study and writing of the manuscript.

Paper 8: Geir Kjetil Sandve conceived the initial idea, developed the main algorithms and drafted the manuscript.

Acknowledgements

With my PhD now finished, there are a lot of people that have my deepest gratitude — not for just pushing me through, but for actually making it very enjoyable.

First and foremost, I would like to thank my two supervisors: Arne Halaas for giving me responsibilities and opportunities beyond what I would have dared to ask for, and for showing an impressive consistency during the many years we have worked together; Finn Drabløs for showing me how to approach biology from a computational side, and for always inspiring and challenging me in our many discussions.

Thanks to all my colleagues at IDI for support, ideas and discussions. And to everyone in our lunch group: I will never forget all our interesting and crazy discussions. A special thanks to Magnus Lie Hetland for being a solid rock and always providing rational answers for the full spectrum of issues that I met in the different stages of my PhD work. Also, thanks to Haakon Dybdahl for our many sessions of letting out PhD frustrations, either in the corridors or out fishing. Thanks to Kjell Bratbergsengen for giving me many interesting challenges and always giving me very decent treatment. Thanks to my project and master students — Kai, Tarjei, Øystein, Kristoffer, Vetle, Lars Andreas, Øyvind, Vegard and Magne — for giving me both frustration and inspiration. Thanks to Magnar Nedland for challenging many of my habits during our co-supervision.

Thanks to everyone at the Laboratory Center for including me in discussions and social settings although I never formally belonged there. A special thanks to Osman Abul, Kjetil Klepper and Jostein Johansen for our many interesting discussions and cooperation on articles.

Thanks to Hans Petter Ulven and Terje Rydland for very rewarding cooperation on the computer science for teachers course, and to the NKUL program committee members for making the most lively meetings I have ever experienced.

Thanks to everyone at Radiumhospitalet and the joint centre for bioinformatics in Oslo for welcoming me and including me in ongoing research during my guest stay in Oslo. A special thanks to Eivind Hovig for giving me new challenges and inspiration, and still accepting that most of my effort was spent on my existing projects.

Thanks to Eric P. Xing for inspirational scientific papers, and for encouragement and good cooperation during my stay at the Carnegie Mellon University. Also, thanks to my other colleagues in Pittsburgh, and Pradipta Ray in particular, for a very fruitful and enjoyable cooperation.

Thanks to all my friends and family. Thanks to my parents for never putting pressure on me, and for encouraging my curiosity throughout all these years. Thanks to Egil for always being one year ahead and giving me inspiration for what is to come.

And finally, thanks to my wife Christin and my son Audun, which I consider to be the real achievements of my life.

List of papers

Paper 1 G. K. Sandve and F. Drabløs. A survey of motif discovery methods in an integrated framework. *Biol Direct.* 2006;1:11.

Paper 2 T. Lin, P. Ray, G. K. Sandve, S. Uguroglu and E. P. Xing. Baycis: a bayesian hierarchical HMM for cis-regulatory module decoding in metazoan genomes. In *Proceedings of the Twelfth Annual International Conference on Research in Computational Molecular Biology (RECOMB).* 2008;66–81.

Paper 3 G. K. Sandve, O. Abul and F. Drabløs. Compo: composite motif discovery using discrete models. *BMC Bioinformatics* (submitted).

Paper 4 G. K. Sandve, O. Abul, V. Walseng and F. Drabløs. Improved benchmarks for computational motif discovery. *BMC Bioinformatics.* 2007;8:193.

Paper 5 K. Klepper, G. Sandve, O. Abul, J. Johansen and F. Drabløs. Assessment of composite motif discovery methods. *BMC Bioinformatics.* 2008;9:123.

Paper 6 O. Abul, F. Drabløs and G. K. Sandve. A methodology for motif discovery employing iterated cluster re-assignment. *Series on Advances in Bioinformatics and Computational Biology.* 2006;4:257–268.

Paper 7 O. Abul, G. K. Sandve and F. Drabløs. False discovery rates in identifying functional DNA motifs. In *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering (BIBE).* 2007;387–394.

Paper 8 G. Sandve, M. Nedland, Ø. Syrstad, L. Eidsheim, O. Abul and F. Drabløs. Accelerating motif discovery: motif matching on parallel hardware. *Lecture Notes in Computer Science*. 2006;4175:197–206.

Other publications

Publications made in the PhD period that are not included as main papers in this thesis:

G. Sandve, and F. Drabløs. Generalized composite motif discovery. *Lecture Notes in Computer Science*. 2005;3683:763–769.

O. Abul, G. Sandve, and F. Drabløs. TScan: A two-step *de novo* motif discovery method. *Third Annual RECOMB Satellite Workshop on Regulatory Genomics*.

Chapter 1

Introduction

Motif discovery is, in the context of this thesis, the discovery of sequence patterns that occur frequently in a set of target DNA sequences [1]. The discovery of sequence motifs in DNA is an established field that has developed from a few basic methods in the early 80s to a large variety of complex models and algorithms today [4, 5]. This chapter will give a brief and general introduction to DNA motif discovery. It will present important principles of transcriptional regulation and the basics of the computational problem. A more detailed presentation of motif models, motif discovery algorithms and other aspects that are of particular relevance to this thesis are given in the state-of-the-art chapter.

In DNA motif discovery, the DNA is basically viewed as a sequence of the letters a,c,g and t. Motifs (sequence patterns) occurring frequently in selected subsets of DNA are often thought to represent functional elements in DNA, typically regulatory elements which are the main focus of this thesis. The full DNA motif discovery process involves compiling a set of DNA target sequences of interest, predicting functional elements by running a motif discovery method on the target sequences, interpreting the results, and possibly proceeding with experimental validation of the computational predictions.

The accuracy of computational predictions is a main concern in the motif discovery process, but several other aspects are also important for a successful outcome. Appropriate selection of target sequences is important to ensure that a strong motif exist for the computational methods to discover. Execution speed of the computational methods may put limitations both on the maximum size of target sequences and the possibility for iterating between motif discovery and other parts of the process. Precise interpretation

of results is important to handle the uncertainty involved in the computational predictions and allow appropriate decisions on further use on the motif discovery results.

1.1 What are regulatory elements

It is commonly known that the fundamental instructions of living organisms are contained in a double helix of DeoxyriboNucleic Acid (DNA). Each strand of this double helix contains a sequence of simple subunits called nucleotides. There are only four different nucleotides in DNA, and the DNA can for computational purposes be represented as a long text string of four different letters. Today, relatively complete drafts are available of the DNA sequence of several organisms, including humans. For many organisms, most of the genes are also known. Genes are relatively short stretches of DNA that are direct recipes for proteins, which again perform most of the biological functions of an organism.

The DNA and genes do, however, only give a static and general view of the genome. The body of an advanced organism like human is composed of several very different tissues, consisting of cells that are also dynamic and changing over time, although the basic DNA is the same across the body and across time. The dynamics of organisms are handled by the gene regulatory mechanisms. A main part of regulation is performed by specific proteins called transcription factors (TFs) that regulate the production of RNA and proteins from genes. This regulation is achieved by the TFs binding to DNA near genes, thus influencing the recruitment of RNA polymerase. RNA polymerase is a protein that performs the translation of genes into RNA, the first step in translating genes to proteins. The regions where TFs bind are often called regulatory regions. The region just before the gene, called upstream region, is the most basic regulatory region, but TFs can also bind in regulatory region that are situated after the gene (downstream), within the gene (introns), or further upstream.

Gene regulation is a finely orchestrated mechanism, and the TFs do not attach randomly to the DNA. As both the TFs and the DNA are molecules containing a structured organization of positive and negative charges, binding of a TF to DNA will depend on whether these charges can be aligned in a complementary way that forms strong physical bonds between the molecules. Because of this, each TF will have its own sequence-specific requirement for binding to DNA. This will not be a strict requirement for an exact DNA

sequence, however, but rather a preference for binding to short stretches of DNA that share some sequence similarity.

Determining where in the DNA each TF can bind is important for several reasons. The regulation of genes by TFs is a basic component of a very complex system of interactions between genes (mediated by proteins). Accurate understanding of TF regulation is thus essential for understanding a range of different processes that occur within cells, including differentiation and cell-type specific regulation. Knowing the exact locations where TFs can bind is an important step towards determining how genes are regulated by a given TF, and it may also explain how slight sequence variations between individuals in the regulatory regions may influence for instance the risk for a specific disease.

Experimental determination of where TFs bind to DNA is still a very tedious process, and therefore our current knowledge of the details of regulation is very limited, even for the most well-studied organisms. Computational methods for predicting binding sites is an important alternative, but after more than 20 years of development, the problem is still very challenging.

1.2 Basic problem formulation

The basic idea behind computational binding site prediction is to capture the sequence similarity of related binding sites by a motif (sequence pattern), and apply this motif on a target sequence. Three basic computational problem formulations can be envisioned regarding transcription factor binding sites (TFBS) and motifs: Motif inference, motif scanning and *de novo* motif discovery.

1.2.1 Motif inference

In the motif inference problem, a set of binding sites for a specific TF are known, and the task is to infer a motif representation that captures the binding specificity of the TF.

As the DNA sequence of related binding sites shows some variation, it is usually not sufficient to describe the binding specificity of a TF by an exact sequence. In computer science, regular expressions are commonly used to represent sequence patterns. Simplified regular expressions, consisting of

character sets at each position and sometimes also variable gaps, have been used for describing patterns in proteins, but have not seen much use for describing patterns in DNA. A discrete Hamming distance string model has seen more use. This model allows instances to have a certain number of mismatches compared to a base string. This allows thresholds on variation to be applied over the whole instance, instead of being specified independently on each position as in the simplified regular expressions. Both simplified regular expressions and Hamming distance models can be constructed from a set of observed sites by enumerating motifs, using a scoring function to determine how well candidate motifs capture the observed sequences, and then selecting the highest scoring motif.

The most widespread model for capturing the binding specificity of a TF is the position weight matrix (PWM), sometimes also referred to as a position-specific scoring matrix (PSSM). This is technically a product-multinomial probability distribution describing the observed sequences. At each position, the PWM gives the probability of observing each of the four nucleotides. The probability of observing a specific sequence given a PWM model is then the product of the probabilities of the observed symbols over all motif positions. A PWM is constructed simply by aligning the observed sequences (which should be equal length), counting the observed frequency of each symbol at each position, and then setting the underlying probabilities equal to the observed frequencies (see Figure 1.1). This corresponds to a maximum likelihood solution for seeing the observed sequences given the product-multinomial probability distribution represented by the PWM. To avoid estimating zero- or close-to-zero-probabilities based on few examples, a small number of pseudo counts are typically added to the real counts, a process which corresponds to enforcing a soft prior on the multinomial probabilities represented by the PWM.

A main property, and potential limitation, of the PWM model is that the probability distributions for individual positions are independent. This means that the PWM is not able to capture any higher-order nucleotide dependencies. The importance of this limitation is, however, still a matter of debate. Sometimes a model of how nucleotides are distributed in background is also incorporated into a PWM, with the PWM then representing the odds of a sequence belonging to motif versus background.

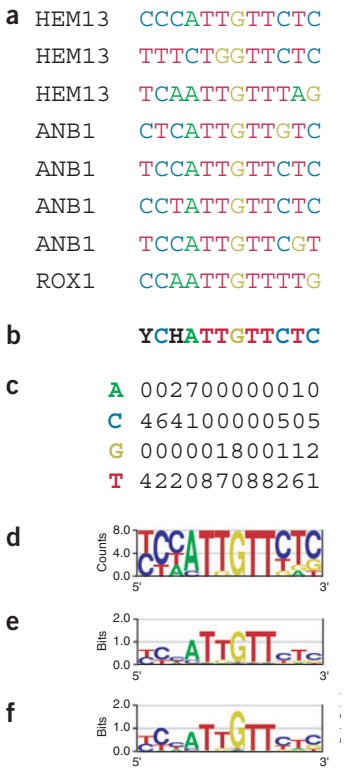


Figure 1.1: Constructing a PWM from aligned binding sites. a) shows aligned binding site sequences for a common TF. b) shows a consensus string representation of the binding sites, with e.g. symbol Y representing either C or T at a position. c) shows the counts of a,c,g and t at each position. d) visually shows the proportion of counts at each position. e) shows a sequence logo, with total height of a column representing the total information content at a position, and each symbol scaled by its relative frequency at the position. f) Same as e, but with information in relation to a non-uniform background distribution. From D'haeseleer [1].

1.2.2 Motif scanning

The second problem formulation can be seen as the opposite of the first problem. The binding specificity of a TF is known in the form of a motif, but the binding sites are unknown. In other words, this corresponds to scanning for instances of a motif in a target sequence. This is shown schematically for scanning of PWM model against sequence in Figure 1.2.

With discrete motif models, determining instances corresponds to trivial pattern scanning. As DNA motif models are almost always of fixed length, every substring of the target sequence of the required length is matched against the motif model (pattern). Every substring that gives a match are considered binding sites for the TF.

As the PWM gives a continuous match score for each substring of the target sequence, a score threshold has to be applied in order to determine hits and non-hits. This threshold can be determined in several ways. Since the PWM describes a probability distribution, a threshold can be set directly based on probability considerations. Alternatively, scores can be computed for every substring of the target sequence, and a threshold specified afterwards that will lead to a desired number or frequency of hits.

Although the matches of the common motif models are very simple to compute, the accuracy of motif scanning is often quite low in realistic settings. In addition to the annotated binding sites, a lot of locations without any known binding activity will typically be predicted as being binding sites. In some cases the reason for the discrepancy may simply be limited data on binding activity, but in general it reflects a prediction weakness. The reason for the many false predictions is that binding sites represented by TF motifs are often quite short, with a relatively large degree of sequence variation. Although several locations may then appear to be binding sites according to a simple computational model, there are a variety of reasons for why such locations may not constitute real biological binding sites. The suitability for binding may be more complex than what is captured by simple motif models, the surrounding DNA sequence may influence binding, and finally e.g. the three-dimensional structure of DNA may make parts of the sequence less available for binding.

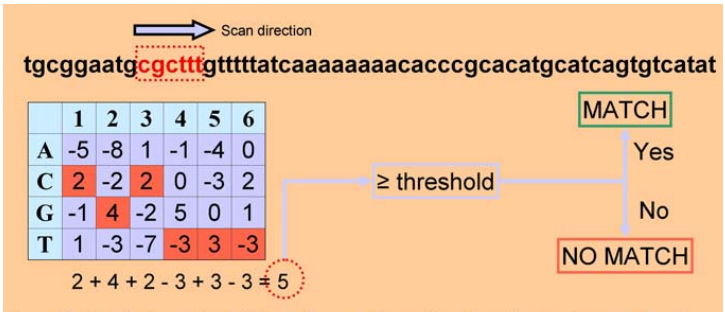


Figure 1.2: Scanning target sequence for discrete PWM hits. The PWM is matched against a sliding window on the target sequence, and positions with scores above a given threshold are reported. From MacIsaac and Fraenkel [2].

1.2.3 *De novo* motif discovery

A third and very challenging problem formulation is the discovery of both motif and binding sites when neither are known. The only thing that is known (or assumed) is that a TF has multiple binding sites with a certain sequence similarity within one or more target sequences.

The basic assumption behind *de novo* motif discovery is that since there exists a set of binding sites with shared sequence similarity that could be represented by a motif, this unknown motif will have more instances than expected in the target sequence. In other words, an unknown motif representing the binding specificity of the TF should be over-represented in the target sequence.

In addition to the motif model chosen to represent the TF binding specificity, a central aspect of a *de novo* motif discovery method is the score function used to measure the over-representation of a motif in the target sequences. Several different score function are presented in the state-of-the-art section.

As both instance locations and motif representation are unknown, there are two basic ways to approach the problem from the algorithmic side. The first approach is to enumerate motif candidates, scan for instances, score each candidate, and select the best scoring motif. This works well for relatively short and simple motif models that can be enumerated efficiently. The other basic approach is to instead enumerate instance locations, infer a motif from these instances, score the inferred motif, and select the best scoring motif.

As there are too many possible combinations of instance locations, these can not be enumerated exhaustively, but have to be estimated heuristically, e.g. by MCMC methods. Several algorithmic approaches for motif discovery are presented in the state-of-the-art section.

1.3 Higher-order organization

1.3.1 Cis-regulatory modules

When TFs bind to DNA and influence gene expression, the TFs do not perform this function in isolation. Instead, several TFs, and often also several other proteins called co-activators, form a complex of molecules. It is this complex as a whole that attracts the transcriptional machinery and thus influences gene expression. This structure of TFs is also reflected in the binding sites for TFs on DNA, meaning that TFBS are not distributed randomly in the regulatory regions. As the TFs form complexes by physical contact, the DNA binding sites for interacting TFs will also have to be in physical proximity to each other. As the DNA may form loops, the binding sites are not necessarily close in DNA sequence, but they are generally thought to be so. This observation has led to the notion of a cis-regulatory module (CRM) — a cluster of binding sites for interacting TFs. Sets of TFs are also thought to interact similarly in the regulation of several genes, meaning that the same combination of binding site motifs would be co-occuring in several regulatory regions. The higher-order organization of TF binding is shown schematically in Figure 1.3.

1.3.2 Composite motif discovery

Composite motif models — models composed of several short contiguous sequence motifs — have been introduced in many varieties to model the CRMs directly (see Figure 1.4). The rationale for this has been that the clustering and co-occurrence of binding sites for interacting TFs can be exploited to increase the accuracy of CRM detection. In the same way as for individual binding sites, a distinction can be drawn between scanning and *de novo* discovery of CRMs. Pure scanning for CRM instances would mean that all parameters of the CRM model are fixed beforehand, while *de novo* discovery would mean that all parameters of the CRM model are inferred from the data.

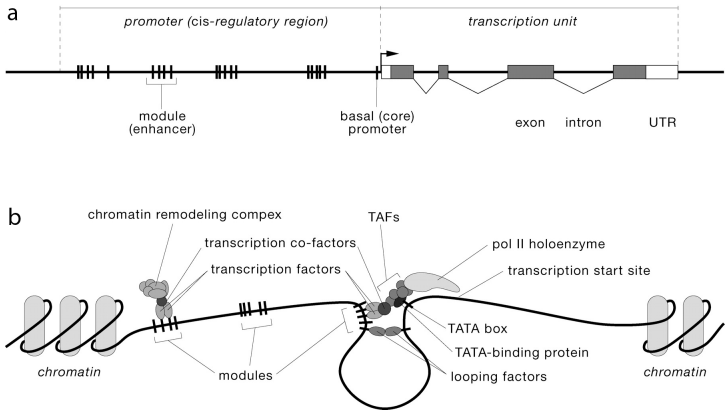


Figure 1.3: Higher order organization of transcriptional regulation. a) shows how cis-regulatory modules, consisting of several motif instances, are distributed at different locations upstream of the gene (transcription unit). b) shows how the DNA sequence may form loops that allow TFs binding to different CRMs to form complexes based on physical contact. From Wray et al. [3].

Something between pure scanning and pure *de novo* discovery is a model where a list of potential short sequence models are pre-defined, while the selection of, and relation between, these short motifs are inferred from the data. This variant is here referred to as supervised composite motif discovery, and is a realistic and practical variant, since potentially relevant TF motifs are often available in public motif libraries. Computational methods can then be used to select from the list of potential TF motifs and to infer the structure between binding sites for these factors.

***De novo* composite motif discovery** The most common argument for doing *de novo* discovery of composite motifs directly, instead of discovering motifs for single TFs individually, is that considering CRMs can help in discovering binding site motifs that are too weak to be detected in isolation, but that are significantly different from background when viewed in relation to other TFs. *De novo* inference of composite motifs is, however, more computationally demanding than inference of single motifs, and it remains to be shown on sufficiently large collections of data that the theoretical advantage of directly discovering CRMs is realized in practice.

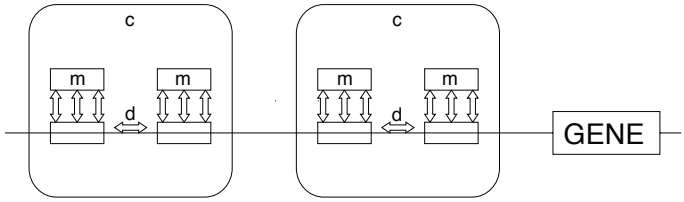


Figure 1.4: A computational view of the higher order organization of transcriptional regulation. The main components of this computational view are single motifs (denoted as m), distances between single motif instances (denoted as d), and composite motifs (denoted as c).

Supervised composite motif discovery The main rationale for doing supervised composite motif discovery, compared to just scanning for instances of the pre-defined PWMs in isolation, is to reduce the number of false positive predictions. In order to detect a reasonable number of the annotated binding sites for a TF, the threshold on PWM score has to be set at a level that will also predict a large number of binding sites at un-annotated positions across the genome. By exploiting the clustering property and co-occurrence of binding sites for interacting TFs, it is possible to filter out a lot of false positives, while not loosing out too many annotated binding sites.

In addition to being an advanced filter for PWM matches, the inference of CRM structure and delineation of CRM instances can also be of value in itself. As CRMs are thought to be the basic units of gene regulation, knowledge of CRM instances, rather than isolated binding sites, may be fundamental for understanding and manipulation of gene regulation. Furthermore, knowledge of which TFs are interacting and how this is reflected on the DNA (binding site) level is important, both for general understanding of how different TFs act together in transcriptional regulation, and for determining which genes are co-regulated.

1.4 Measuring prediction performance

There is no unanimously accepted way of evaluating the prediction performance of a CRM discovery method against annotations. First, as CRM discovery methods typically predict both the locations of individual motifs and also predict a grouping of motif locations into CRMs, either the predicted

motif intervals or the predicted CRM intervals can be compared against annotations.

Second, the way of evaluating single predictions and determining the number of true/false positives/negatives (TP,FP,TN,FN) has to be decided. True/false here refers to whether a prediction is correct or not. Positive/negative refers to annotated binding sites and non-binding respectively (ground truth). Only considering exact correspondence between motif start positions is for instance generally not a reasonable choice, as the experimentally based annotation of binding sites is not an exact and unambiguous procedure. One reasonable choice is to consider a motif/CRM prediction as a true positive if it is less than a fixed number of bases off an annotated position, or similarly if at least a fixed proportion of the prediction overlaps with annotation (site-level). A problem with the site-level is that one has no trivial measure of true negatives, as there is some tolerance in predicted locations. Another possibility is to operate on the single nucleotide level, counting each nucleotide that is encompassed both by a predicted and annotated motif/CRM as a true positive, and FP,TN,FN counted similarly (nucleotide-level).

Third, some summarizing statistics should be computed and used for comparing the performance of different methods. The precision (TP/(TP+FP)) and recall (TP/(TP+FN)) together gives a good view of the performance of a method. However, as different methods may get high score on different measures, a simple ranking of method performance is generally not possible. It is difficult to compare the performance of methods that achieve a very different balance between precision and recall. Alternatively, a single measure such as the phi coefficient of correlation (CC) can be used to achieve a single performance value that can be compared and used to rank methods.

Although producing a single performance value is often convenient, is may also hide important aspects of the performance of methods. Another option for comparing performance is therefore to go the other way, and instead produce a richer representation of performance in the form of a precision-recall (PR) curve. The PR-curve shows the relationship between precision and recall for a given method. This gives a more unbiased comparison between different methods.

When using several datasets, summarizing statistics could either be computed for each dataset and then averaged, or statistics could be computed from the combined counts across datasets. As the confidence of predictions are not necessarily comparable across datasets, PR-curves are generally specific to a single dataset.

Figure 1.5 shows the basic way of measuring prediction performance on the single nucleotide level.

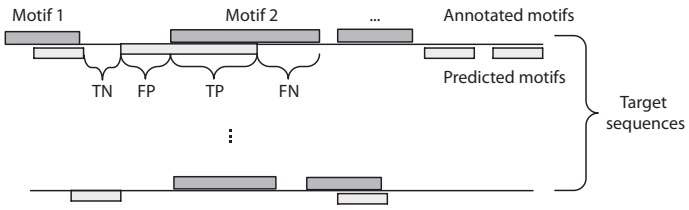


Figure 1.5: Counting the number of true/false positives/negatives between predictions and annotations at the nucleotide level.

1.5 Developments in the motif discovery field

The first computer programs for computational prediction of binding sites for transcription factors appeared in the late 70s. In the first articles, such as that by Korn *et al.* [6], the idea of using a computer and the software itself was of primary interest. These methods used very simple models of sequence similarity and searched for binding site in the core promoter region of genes — short stretches of sequence immediately upstream of transcription start. The sequence similarity model used by Staden was a Hamming distance model. Already a few years later the position weight matrix (PWM), which still is the most common motif model, was introduced by Stormo [7] and Staden [8].

Around ten years later, a new burst of development occurred, with Gibbs Sampling and Expectation Maximization techniques successfully applied to the motif discovery problem. The Gibbs Sampler [9] and MEME [10], both using PWM as motif model, are still among the most widely used methods. During the next ten years, more than a hundred motif discovery methods were proposed, employing a wide variety of models and algorithms. In the latter half of the 90s, it became widely acknowledged that TFs act in combination to regulate genes, and that this combinatorial nature was important also for computational motif discovery. Several methods therefore tried to discover composite motifs, either directly from regulatory regions alone or by using a list of single motifs as additional input. Alongside the focus on composite motifs, there was in the beginning of 2000 also strong focus on integrative motif discovery methods. Different methods integrated orthologous

sequences, gene expression and additional properties of the DNA sequence into the discovery of motifs.

After the DNA sequence of the human genomebecame available around the turn of the century, there has been an explosion of studies making use of genomics, either to understand the dynamics of a normal cell, or to pinpoint the genomic causes of diseases. As gene regulation is central to such studies, a need has risen for more knowledge about TF regulation. As experimentally determined binding site locations are yet very sparse and incomplete even for the most well-studied organisms, there is a clear need for computational predictions to complement the experimentally verified sites. This wide need for TFBS predictions has led to much focus on increasing usability of computational TFBS prediction. Web page versions has been published for several motif discovery methods, making them more easily available for use by non-specialists. Recently, several web systems have also been published, integrating different motif discovery methods in a uniform service that supports the process from extraction of regulatory sequences, via running prediction methods, and finally to visualization of results. A third line of work that has made computational TFBS predictions more accessible, are databases of genome-wide predicted TFBS locations for specific organisms. This even alleviate the need of running computational tools in order to obtain TFBS predictions.

1.6 Current challenges

After more than 20 years of development, the field still faces some fundamental questions. A recent and broad assessment of motif discovery methods revealed a low absolute level of correctness for the prediction of DNA binding sites [11]. The notorious difficulty of the problem makes it suited as a challenge for novel methodological development in the machine learning field, and the wide use of computational TFBS predictions ensures clear biological usefulness of improvements in computational methods. This suggests that the motif discovery field will see new development for many years to come.

Although it has become clear that the discovery of TFBS is a very difficult problem, there is large uncertainty regarding the performance level that realistically can be expected from motif-based computational approaches. Alhtough numerous methods have been proposed for the problem, it is still not clear which methods, or even which fundamental approaches that perform best.

A complicating factor both for determination of a realistic performance level, and for comparing the performance of different methods, is that the ground truth is not accurately known. Even for the best studied sequences there may be many un-annotated TFBSs turning up as false positives instead of true positives when evaluating methods. Furthermore, there may be strong biases regarding which DNA regions have been studied experimentally, meaning that the modeling assumptions that work well on data sets used in testing may not be representative of binding sites in general.

In spite of such challenges, increasing the quality of method evaluations, increasing understanding of how different aspects of methods influence and limits prediction performance, and increasing the understanding of how limited experimental data may influence the results of evaluations should be a worthwhile and interesting endeavor. The limited performance level of current methods also suggests that there still is a clear potential for improved algorithmic approaches to the problem.

Chapter 2

Research outline

2.1 Research questions

Four basic questions regarding the potential for improvement in the motif discovery field have formed a basis for studies presented in this thesis. Several other relevant questions may also be envisioned, but could not be pursued within the time frame of this PhD project. Particularly, the potential for improving prediction accuracy by incorporating information beyond pure sequence, such as additional DNA properties or different experimental values, would have been interesting to explore in relation to several of the studies covered in this thesis. This is however a complex issue requiring one or more separate projects.

The research questions explored in this thesis are as follows:

RQ1 — Sequence based methods

The most basic version of the motif discovery problem in DNA is to predict a motif representing related regulatory elements, based on a set of target sequences only. No properties except the pure base pair sequence is used, and nothing is known about the motif *a priori*.

What are the limitations on prediction accuracy for pure sequence-based motif discovery? Is there a real potential for improving the prediction accuracy beyond the level achieved by current methods based on regulatory sequence only?

RQ2 — Composite motifs

As the goal of composite motif discovery in DNA is to discover motifs representing a set of interacting regulatory elements (CRM), two subversions of the problem have in general been considered.

- De novo: discover single motifs and their relationships simultaneously from a set of target sequences.
- Supervised: discover relationships between motifs, often within a single target sequence, with single motifs given as input.

What performance level can be achieved when searching for CRMs based on precompiled single motifs, compared to de novo *motif discovery? Furthermore, how much can prediction performance for supervised composite motif discovery be increased by considering sets of co-regulated sequences, instead of considering sequences in isolation?*

RQ3 — Background models

As motif discovery is based on the principle of overrepresentation, a background model to contrast observations against is an important element in evaluation of candidate motifs. For some methods, contrasting against background is done explicitly, while in other methods it is implicit in the way target sequence is assigned to hidden motif and background states.

What background models perform well as contrast to TFBSs in motif discovery? Can the use of real negative background sequence as opposed to random background models increase prediction performance?

RQ4 — Motif models

A fundamental difference between methods for motif discovery is whether they use probabilistic or discrete motif models. The probabilistic and discrete models are generally also tied to distinct inference procedures and typically have their roots in different fields of computer science or statistics.

What are the advantages and disadvantages of discrete versus probabilistic motif models? Is the recent trend towards probabilistic composite motif models

(HMMs) warranted by an increase in prediction performance? What other advantages or disadvantages do each of the two approaches have?

2.2 Research context

The contributions in this thesis have been made within three subareas of motif discovery: Development of prediction methods, assessment of prediction accuracy, and improvements in the extended motif discovery process.

RC1 — Methods

The first research context concerns development of *de novo* novel motif discovery methods for the discovery of CRMs from a set of regulatory regions.

RC2 — Assessment

The second research context concerns analysis of the difficulty of TFBS motif discovery and the performance of current methods.

RC3 — Process

The third research context concerns improvements in the extended DNA motif discovery process, from compilation of input data to interpreting predictions made by a method.

Chapter 3

State-of-the-art

More than a hundred methods have been proposed for motif discovery in recent years, representing a large variation with respect to both algorithmic approaches as well as the underlying models of regulatory regions. There is also large variation regarding how methods are described and tested, making it even harder to get a good overview of the field. Many reviews of motif discovery methods have therefore been written, with varying focus and intended audience. A recent review by Pavesi *et al.* [12] gives a very accessible and broad introduction to the field. It divides methods into consensus- and alignment-based, and surveys the most established methods one at a time. It also discusses background modeling, evaluation of motifs and the practicalities of using these methods. The review by Wasserman and Krivan [13] has a stronger focus on the underlying biology of motif discovery in regulatory regions. It also goes a bit more into the combinatorial nature of binding sites, and touches upon issues such as phylogenetic footprinting, CpG-islands and chromatin structure. Finally, some reviews focus on specific techniques such as phylogenetic footprinting [14], or on specific genomes [15].

3.1 A mathematical perspective

As motif discovery methods can be very complex, with many possible differences, several authors have proposed frameworks for classifying motif discovery methods. Brazma *et al.* [16] categorize motif discovery methods with respect to whether they use explicit negative sequence sets or not, expressiveness of the pattern models, whether patterns are deterministic or statistical, and whether the algorithms are pattern driven or sequence driven. In a later

paper Brazma *et al.* [17] define a three step paradigm consisting of choosing a class of grammars (motif model), designing a rating function (motif score), and developing an algorithm.

Here a mathematical perspective is used for describing motif discovery methods, with a special emphasis on the hierarchical modeling of regulatory regions[1]. Motif discovery methods are described at three levels:

The most basic level (Level 1) represents the binding of individual transcription factors (TFs) to short contiguous sequence segments. These sequence segments are modeled by single motif models that give a distinct score for each sequence segment in a regulatory region.

The next level (Level 2) represents CRMs: clusters of TFs that bind to DNA in proximity to each other, but with a certain flexibility regarding distance 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 can be calculated from the score of single motifs at given positions as well as inter-motif distances.

The final level (Level 3) represents how a single or composite motif is overrepresented across several regulatory regions. This overrepresentation is typically used to rank candidate motifs in order to return the most interesting motifs to the user of a motif discovery tool.

3.2 Single motif models (Level 1)

Transcription factors bind to specific short segments of DNA, transcription factor binding sites. This is the most basic element of the regulatory system, and can be modeled using single motif models.

A single motif model is defined as a function $\mathbb{N} \rightarrow \mathbb{R}$ *that maps a sequence position as a non-negative integer to a real numbered motif score.*

The single motif function gives the degree of match between the substring beginning at a specific position and an underlying consensus model. In the most general sense, the single motif function gives a distinct score for any given substring. However, the number of free parameters has to be restricted

[1]This part is based on Paper 1, but uses a simplified mathematical framework. The full framework is given in the original paper.

to allow training of the model from a limited number of examples (e.g. 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.

3.2.1 Probabilistic match models

The most widely used probabilistic model is without doubt the position weight matrix (PWM), also known as position specific scoring matrix (PSSM), that assumes independence between positions [7]. 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 [18].

Many different extensions to the basic PWMs have 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 [19, 20, 21].

The most direct way of incorporating dependencies within motifs is to extend the PWM to include pairs of correlated positions [22, 20]. Another straightforward approach is to use a mixture model in which the motif occurs as one of a limited number of stochastic prototypes [23]. 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 [24]. However, it is hard to find a good compromise between a high n that may give too many free parameters and a low n that may miss out the dependencies of interest. If the relative importance of dependencies varies within a motif, a variable-length Markov model (VLMM) [25] may be preferable. Furthermore, if some long-range dependencies seem to be significantly stronger than dependencies between neighboring positions, the order of the positions in the Markov chain may also be permuted before a VLMM is applied [26].

Another way to model dependencies is to use Bayesian networks. Barash *et al.* [23] 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. Similarly, Ben-Gal *et al.* [27] argue for variable order Bayesian nets.

Instead of focusing on dependencies between specific nucleotides at different

positions, Xing *et al.* [28] model the distribution of conserved positions within a motif. 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. The work of Kechris *et al.* [29] achieves similar properties by assigning conservation types (strong, moderate or low) to blocks of motif positions.

3.2.2 Deterministic match models

A deterministic match model evaluates to a binary value indicating either 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 [30, 31, 32] and dictionary models [33].

A regular expression model returns 1 if the given substring is matched by an underlying regular expression. As reviewed by Brazma *et al.* [16], the models used in motif discovery are typically composed of exact symbols, ambiguous symbols, fixed gaps and/or flexible gaps. Regular expression models are used in e.g. [34, 32, 35, 36, 37].

Many methods use mismatch expressions as motif match models, e.g. [38, 39, 40, 41, 42, 43]. 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 [44], 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 [45].

The probabilistic models are much more expressive than the deterministic models. In fact, all oligos, regular expressions and mismatch expressions can be represented as PWMs. However, a major benefit of the deterministic

models is that they often allow exhaustive discovery of optimal motifs.

3.3 Composite motif models (Level 2)

Clusters of binding sites for cooperating TFs, often called CRMs, are believed to be essential building blocks of the regulatory machinery. Werner [46] states that "Within a promoter module, both sequential order and distance can be crucial for function, indicating that these CRMs may be the critical determinants of a promoter rather than individual binding sites". The multitude of models developed for the discovery of CRMs is another indication of the conceived importance of this. It is therefore natural to define a computational motif model that represents a combination of single motifs.

A composite motif model is defined as a function $2^{\mathbb{N}} \rightarrow \mathbb{R}$ *that maps a set of single motif sequence positions as non-negative integers to a real numbered composite motif score.*

The composite motif function consists of a set of (generally different) single motifs, with each single motif contributing with 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 individual single motif and distance scores.

Distance functions

Many different models have been proposed to capture the importance of inter-motif distances within a CRM. Several methods put constraints on the distances between consecutive motifs, requiring either fixed distances [47, 32], distances below thresholds [48, 49, 50], or distances within intervals (e.g. [34, 32, 51, 47, 42]).

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. [52, 53, 54, 55, 56]). 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 [57]. Similarly, a

geometric distribution on inter-motif distances follows implicitly from many HMM models [58, 59], and is assumed explicitly in Gupta and Liu [60].

The conservation of inter-motif distances across CRMs can also serve as a basis for distance score functions. Wagner [61] 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 [62] calculate scores by comparing the distances with a histogram of distances between the same regulatory elements in other CRMs.

Combining single motifs

There are many ways in which a set of single motif and distance scores can be combined into 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. [61, 63, 64, 65]). A variation of this is to require that M out of N single motif scores are 1 [66]. Similarly, the count of binary single motif values can be used directly as a composite motif score [67, 68, 32].

For methods that use non-binary single motif scores, a common approach is to calculate the sum of single motif and distance scores [57, 62]. Some methods require that all distance functions are 1, and if they are, composite motif score is the sum of single motif scores [55, 69, 53, 70]. Similarly, the method ModuleScanner sums only single motif scores above a threshold, and MotifLocator sums the N highest single motif scores [55]. Another 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 the single motifs [48]. Finally, a few methods take the composite motif score to be the highest single motif score [41], or the lowest single motif score [71].

Many specialized models have also been used to combine single motif and distance scores, e.g. the hidden Markov model (HMM) [59], history-conscious HMM (hcHMM) [56], self-organizing map (SOM) [72], and artificial neural network (ANN) [73]. In all of these models, the score of several homotypic and/or heterotypic single motifs are combined in a relatively complex way.

3.4 Motif overrepresentation (Level 3)

Computational motif discovery is possible primarily because motifs representing 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 from a background model. More specifically, the p-value [54, 36] and z-score [38, 32] 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. Shuffled control sequences may also be used as background [9].

A simpler approach is to compare only the raw sum of gene scores when ranking motifs. This is equivalent to 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 information content (IC) of discovered PWMs [74]. 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 [75, 76, 39].

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 [64].

3.5 Background models

As the goal of motif discovery is to find a set of substrings with a significantly high level of sequence similarity, the notion of a background to compare against becomes important. In the common mixture model view of sequences, which entails e.g. HMM-based composite motif models, the background is

only implicit, as all non-motif parts of the target sequence will take the role of background. In this perspective, the goal is to find a motif that can explain as much of the target sequence as possible, reducing the entropy of the remaining sequence.

In several other approaches, the background is explicit. One kind of explicit background is a random model that allows analytical computation of match probabilities, typically as a (higher-order) Markov model that is estimated from either the target sequences or from other DNA sequence. This model can also be directly integrated into scanning of target sequences, giving scores that contrast motifs against background. Alternatively, background can be empirical in the form of real negative sequence that motifs are scanned against to compare with scores in the target sequence. By e.g. making normality assumptions, scores can be given as easily interpretable z-scores.

3.6 Algorithmic approaches

An important trade-off in motif discovery is between representational expressibility and computational efficiency. For the case of restricted deterministic motif models, several algorithms exist that can exhaustively discover the optimal motifs [75, 76, 77].

However, probabilistic motif discovery algorithms do not guarantee returning the global optimum when applied to realistic problems. These algorithms are typically based either on iterative refinement or stochastic optimization. Expectation maximization (EM) [78, 74, 79, 80, 81] is the most widely used iterative refinement method, but variational EM [59] has also been used. The stochastic optimization technique most widely used for motif discovery is Gibbs sampling [9, 82, 47, 83], sometimes combined with general Metropolis-Hastings [84, 85, 86]. Recently, genetic algorithms [69], evolutionary Monte Carlo [60] and simulated annealing [68, 26, 87] have also gained some popularity.

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 [88] or exhaustively [76]. A promising approach to motif discovery is first to 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 be used as seeds for motif discovery in the full set of sequences. The method of Liu *et al.* [89] 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. The highest scoring mismatch expressions are then used as seeds for probabilistic motif discovery in the whole set of sequences.

Chapter 4

Advancing the state-of-the-art

This chapter describes the contributions made within the three research contexts described in the research outline. Figure 4.1 shows a categorization of the research articles included as main papers in this thesis. Within the first research context, Paper 2 and 3 use probabilistic and discrete composite motif models, respectively. In RC2, Paper 4 makes considerations relevant for both scanning and discovery of single motifs, while Paper 5 assesses both probabilistic and discrete composite motif discovery methods. In the third reserach context, Paper 6 considers false discovery rate when scanning for single motif hits, the PAMM framework in Paper 7 can be used for acceleration of both scanning and discovery of single motifs, and the iterative methodology in Paper 8 can be used together with any motif model, and with either scanning or discovery.

<table>
<tr><th colspan="2"></th><th>RC1
Development</th><th>RC2
Assessment</th><th colspan="3">RC3
Usability</th></tr>
<tr><td rowspan="2">Single
motifs</td><td>scanning</td><td rowspan="2"></td><td rowspan="2">Paper 4</td><td>Paper 6</td><td>Paper 7</td><td rowspan="3">Paper 8</td></tr>
<tr><td>discovery</td><td></td><td></td></tr>
<tr><td rowspan="2">Composite
motifs</td><td>probabilistic</td><td>Paper 2</td><td rowspan="2">Paper 5</td><td></td><td></td></tr>
<tr><td>discrete</td><td>Paper 3</td><td></td><td></td></tr>
</table>

Figure 4.1: Research papers included in the thesis, categorized according to research context and motif model/computational problem.

4.1 Development of methods (RC1)

With the important role TF regulation plays in many settings, the low availability of experimentally verified binding sites, and the low prediction accuracy of current computational methods ([11]), there is a clear need for further development in the motif discovery field. Also, the large variety of regulation *in vivo*, and the large variety of contexts in which motif discovery methods are used, means that a variation of available approaches might be of value in itself. In addition to the obvious contribution to biology improved methods represent, the motif discovery problem may also play a role in computer science as a challenge for novel methodological development. In this thesis, two very different methods are presented: one method uses hierarchical hidden Markov models for rich probabilistic modeling of CRMs, while another method learns expressive discrete CRM models by traversing an implicit discrete search space.

4.1.1 BayCis

Several recent methods have used Hidden Markov Models (HMMs) for representing CRMs, with states representing different motifs and background, and transition probabilities capturing the clustering bias of CRMs. An HMM can be trained from data and it can be scanned against data. In training, the parameters of the HMM are estimated from the observed sequence, while in scanning only the probabilities of being in the different hidden states at each position in the sequence are estimated. Most HMMs proposed for CRM discovery are actually quite simple methods that use fixed HMM parameters, and only perform scanning to classify sequence positions to motif and background states (e.g. [58, 90, 57]).

The BayCis method presented in Paper 2 of this thesis takes a considerably more sophisticated approach to composite motif discovery. It uses a hierarchical Hidden Markov Model, with a much richer state space than other models. It is able to capture several different kinds of non-motif states and is also able to capture preferences for specific transitions between different motif and non-motif states. Instead of enforcing fixed HMM parameters, only soft priors are specified, allowing parameters to be estimated from the data based on Bayesian posterior inference. This makes the approach more robust with respect to assumptions, and allows models to be automatically tailored to each data set. The disadvantage of the rich model is a more computationally intensive inference and a need for more data to ensure reliable

inference.

4.1.2 Compo

Using discrete models of CRMs has some obvious advantages regarding simplicity, flexibility, and inference, as well as interpretation of results. A main concern with the discrete approach has been that it is too rigid in determining what constitutes binding sites, and in enforcing distance constraints. With the large degree of uncertainty and irregularity in biology, making hard decisions on for example what constitutes motif hits and whether a distance constraint is met or broken, makes the approach too dependent on thresholds. Discrete models have therefore seen less use in the last years. The Compo method presented in Paper 3 revisits the discrete approach to composite motif discovery, introducing several novel elements.

The Compo method tries to address the limitations of discrete models by applying multiple thresholds instead of a single one for each discrete decision, and for each data set automatically select the most appropriate thresholds. More specifically, several thresholds are used to determine hits and non-hits from PWM scores, and several values are used as distance constraint of CRMs. Additionally, instead of requiring that all components of a composite motif occurs in every CRM, Compo optionally allows some of the component motifs to be absent from CRM instances. Also, several values are used for the number of components of composite motifs, and for the number of component motifs (if any) that are allowed to be missing from CRM instances. While using multiple thresholds is trivial in itself, it is more challenging to support automatic selection of the most interesting motifs arising from different threshold values. Without such selection across threshold values, the user would be flooded with a large number of similar motif variants and a very difficult manual selection process.

Compo uses "unexpectedness" of observed composite motif support as significance measure, and computes p-values of getting the observed support in a partly random and partly empirical background. These p-values can be compared across threshold values, thus allowing automatic selection among motifs that employ different threshold values for delineating motif instances. Other novel elements of the Compo approach is the use of a background that is a mix of empirical data and algebraic model, and the introduction of multi-objective optimization to the motif discovery field. When applied on sets of co-regulated sequences, which allows estimation of CRM structure

based on several instances, the performance of Compo is superior to several previously published methods. This is particularly the case when the list of input motifs contains many non-relevant motifs. The use of support across sequences allows Compo to be more robust against non-relevant (noise) motifs compared to methods that use fixed parameters and consider each single sequence individually.

While the strength of Compo lies in the use of the general unexpectedness measure (probability in background) as a uniform significance measure, the reliance on this measure for motif selection is also a potential limitation on further development. It may be difficult to combine this general unexpectedness measure with prior biological knowledge. If knowledge is accumulated on how certain features of DNA influence the tendency of a sub-sequence to act as binding site for transcription factors, it is not obvious how to combine such biological priors on target sequence with a significance measure based on unexpectedness in background.

4.2 Assessment of motif discovery (RC2)

Due to the large number of proposed methods, assessing the prediction performance becomes important. Assessment is important for determining the general performance level that can be expected from computational methods, for providing guidance in selecting among methods in practical use settings and for giving directions on which approaches to pursue in further algorithmic development. Assessment is challenging for several reasons. It is for example difficult to find reliable and experimentally verified binding site data. There are challenges regarding how to select and process binding site data to serve as answers in evaluations, how to present input data and parameters to evaluated methods, and how to objectively measure the prediction performance of different methods. Understanding data set properties and how they should be used for performing large scale benchmarking thus becomes very important. As the assessment of CRM discovery is even more challenging than for single motif discovery, these questions become even more important there.

4.2.1 Assessment of single motif discovery

There has in recent years been clear progress on assessment of single motif discovery methods. The seminal paper by Tompa *et al.* ([11]) presented a

first really broad and neutral benchmarking effort in the TFBS prediction field, assessing 13 different methods on a reasonably large collection of data sets. Unfortunately, the study was not able to draw clear conclusions regarding preferred methods or regarding preferred approaches at a more general level. It has not acquired status as a standard benchmark for assessing newly proposed methods. Only a few papers have since used the Tompa benchmark for assessment (e.g. [91, 92, 93]).

Paper 4 of this thesis showed that binding sites in several of the data sets used in the Tompa benchmark could not be represented by any of the common motif models. Even when motif models were inferred directly from the known binding sites, the motif models were not able to reliably delineate the same binding sites during scanning of the target DNA sequence.

This might be interpreted as a limitation of the common motif models, or more generally as a limitation of the pure sequence based approach to motif discovery. This might, however, also be partly due to inaccuracies in the exact definition of binding site intervals from experimental verification, and partly due to undetected binding sites turning up as false positives instead of true positives when predicted by computational methods. To distinguish between limitations of the discovery methods on one hand, and the motif models on the other hand, benchmarking data sets were processed and filtered in Paper 4. Two different benchmarks suites were proposed, one that allows good discrimination of binding sites with common motif models, and one that does not allow good discrimination. Both of these suites also have binding sites of fixed length for a given TF, which corresponds to a restriction that is made by the large majority of motif discovery methods.

Although such processed and filtered data sets might provide for more accurate comparisons between methods, the processing might make them less suited for estimating a realistic performance level in practical use. Still, as they are in a form that more closely reflects what can be achieved by the computational approach, we think they might more easily be accepted by developers of new methods. In order for a benchmark suite to become a real *de facto* standard in the field, a stronger consensus on details of TF binding and function *in vivo* will probably have to reached.

4.2.2 Assessment of composite motif discovery

Assessment of composite motif discovery is even more challenging than for single motif discovery. The difficulties due to incomplete and inaccurate

binding site annotations carry directly over to CRM discovery. There is an even larger variety of approaches and even less consensus on the nature of CRMs compared to single TFBS. Composite motif discovery methods pose very different requirements and assumptions, making it difficult to do unbiased comparisons of prediction accuracy. While single motif discovery is based mainly on overrepresentation across sequences, CRM discovery is also based on clustering of binding sites within limited stretches of sequence. Some methods are even based solely on this clustering property. There is, however, currently not a clear consensus on how large and densely populated CRMs in general are.

Due to the difficulty of quantitative assessment, most proposed composite motif discovery methods have resorted to qualitative assessment on small sets of selected CRMs. This does make it very difficult to compare the performance of different methods, and with the numerous proposed methods it becomes difficult to navigate in the field. Paper 5 in this thesis is the first broad and neutral benchmarking effort for composite motif discovery methods. The paper compares eight methods on a relatively large and diverse collection of sequences containing binding sites for interacting TFs. The main focus is on supervised composite motif discovery, with a list of potential regulators given as input. One motif discovery method is also included, but its performance is close to random. Prediction accuracy is tested with different levels of noise in the list of input motifs. Although the benchmark is not able to identify a single preferred method, it still shows clear differences in prediction accuracy between methods. It may also pave the the way for future benchmarking efforts, which could add even more variation regarding CRM size, single motif density and support across sequences. Also, future assessments of CRM discovery would be able to make use of a continuously increasing number of experimentally determined CRMs.

4.3 The extended motif discovery process (RC3)

In addition to improving prediction quality by algorithmic development and improving evaluation of predictions by benchmark development, it is also important to consider possibilities for improving the extended process of DNA motif discovery. The first step of this process is to compile a set of regulatory regions to serve as input data for motif discovery. It is essential that binding sites for a common TF are highly enriched in the input data, if the use of a motif discovery method in a next step is to be successful. In motif

discovery, the main concern will generally be prediction accuracy. However, in certain settings, the running time of a method may also be important. Improvements in running time may allow larger data sets to be considered, and may allow running motif discovery in a more interactive manner, where input data and parameters are iteratively modified based on previous runs. A final issue considered in this research context is the uncertainty of predictions made by motif-based methods. Improved approaches for assigning confidence to predictions is important to interpretation and further use of predictions.

4.3.1 Pre-processing of input data

A common use of motif discovery methods is to search for binding sites in sets of regulatory regions arising from micro-array experiments. A set of genes show similar expression patterns in the experiments, and the question asked is whether some of these co-expressed genes are also co-regulated, i.e. if the genes have binding sites for the same TFs in their regulatory regions.

In order for a binding site motif to stand out from noise, a TF should have binding sites in most of the collected regulatory regions. This is, however, not ensured by the gene expression clustering methods, which only consider similarity of expression when forming groups of genes. Paper 6 describes an iterative approach to gene clustering and motif discovery. The gene grouping given purely by similarity of expression is modified by similarity of motif predictions in the regulatory regions of the genes. The gene grouping is iteratively updated until converging on a set of genes that show high similarity of binding site predictions.

This iterated methodology allows expression similarity and sequence similarity to be considered together, similar to what can be achieved by methods based on regression from motif scores to expression [52, 51, 94, 95, 96], or methods based on joint probabilistic modeling of sequence similarity and expression [97]. The iterative nature is both its strength and weakness: it is not as expressive as a joint probabilistic model, but the simpler interaction between expression data and sequence data allows more efficient computational inference and allows standard methods for gene expression clustering and sequence motif discovery to be used.

4.3.2 Significance of motifs

After using a motif-based prediction method, a user is typically faced with a large number of predictions that should be viewed as a mixture of correct and spurious predictions. Such predictions can arise at several levels: predictions of exact binding site locations for a given motif, predictions of which genes are regulated by a given motif (TF), or predictions of which motifs (TFs) are regulators for a group of genes.

Several motif-based methods compute p-values for predictions. These p-values should, however, be interpreted with caution. One issue is that they are computed based on simplified randomness assumptions that do not accurately reflect properties of real DNA. As a large number of predictions are typically made simultaneously, another issue is multiple testing.

In Paper 7 we develop an approach based on false discovery rates to support decisions when scanning a large number of library motifs against sets of regulatory regions. The task is to determine which motifs (TFs) are functional regulators for a given gene group (set of regulatory regions). Two different approaches to computing p-values were used. Based on the collection of observed p-values, a false discovery rate is estimated for different p-value thresholds. In principle this allows a researcher to control the expected proportion of false positive predictions among the TFs (motifs) accepted as regulators after a threshold is applied on p-values. This will not correct inaccuracies in the original p-value calculation, however, so the simplified DNA assumptions might still give overly optimistic results. Still, the ability to control the number of accepted findings based on FDR rate instead of raw p-values, is an important step forward.

4.3.3 Acceleration of motif scanning

Although prediction accuracy is usually the main concern in motif discovery, running time may also be of importance. Substantial reductions in running time can significantly increase the usability and effectiveness of motif discovery, especially in explorative settings when several conditions are tried. Also, faster methods can increase the maximum size of data sets that are practical to pursue in a motif discovery setting, allowing discovery in regulatory regions of much larger sets of genes. In Paper 8, a framework is developed for acceleration of motif scanning and motif discovery on parallel hardware. The potential of the approach is shown by an implementation that runs a

widely used motif discovery method on special purpose search hardware. The speed-up of this implementation is close to a factor of one hundred compared to the standard software version, showing that substantial improvement of running time is possible through the use of special purpose hardware. Other hardware alternatives for use in the proposed PAMM framework are graphical processors (GPUs) in standard PC graphics cards, or field-programmable gate arrays (FPGAs). We have ourselves pursued the line of FPGA acceleration, although we do not yet have a fully working implementation. We are also aware of other groups pursuing GPU acceleration, but do not yet know whether this has been successful.

Chapter 5

Concluding remarks

The papers included in this thesis can all be said to have contributed towards understanding the potential for further improvement of DNA motif discovery.

First, the work done on analyzing data sets and assessing current methods contributes towards understanding the current status of the field, as well as some fundamental limitations of the motif discovery approach. Second, the work done on developing new methods shows that significant quantitative improvement can be achieved for certain motif discovery problems. Finally, the work done on improving the usability of motif discovery shows potential improvements in the directions of data pre-processing, interpretation of results, and running time.

Although none of the research questions proposed in the introduction have been answered once and for all by the work presented in this thesis, significant contributions have been made in relation to all of them.

Regarding limitations of sequence-based motif discovery (RQ1), Paper 4 showed that binding sites could only to a limited degree be discriminated from surrounding sequence based on sequence motifs, using data sets taken from the recent benchmark study by Tompa et al [11].

Regarding prediction accuracy in different motif discovery settings (RQ2), Paper 5 showed clearly superior prediction accuracy with library PWMs supplied as input, compared to the *de novo* setting, for the discovery of composite regulatory elements from the Transfac database. Additionally, Paper 3 showed clearly higher performance of the method Compo when sets of co-regulated sequences were considered together.

Regarding background models for motif discovery (RQ3), Paper 3 showed

that Compo performed clearly better using a partly empirical background (mix of real DNA and random assumptions), compared to a pure random background. The use of randomly selected upstream regions and highly conserved non-coding regions in the empirical background gave similar performance.

Regarding probabilistic versus discrete composite motif models (RQ4), the expressive models presented in this thesis show advantages of both approaches. The probabilistic method BayCis, presented in Paper 2, allows expressiveness by inferring continuous parameters, allows weak prior knowledge to be specified as soft priors, and allows rich interpretation of results by posterior probabilities. The discrete method Compo, presented in Paper 3, allows expressiveness through a flexible discrete model, robustness by exhaustively finding the highest-scoring motifs, accurate DNA modeling by supporting a partly empirical background, inference of CRM structure by considering sets of co-regulated sequences, and rich interpretation of results either by uniform ranking criteria or a multi-objective view of motif discovery.

Many directions of further research can be envisioned from the work presented in this thesis. The assessment studies raised some novel and important questions that are still far from resolved, and thus opens some interesting directions of further research. The analysis of benchmark sets could be strengthened by considering larger collections of data sets, and could be complemented by studying accuracy levels in different scenarios: when motifs are used to detect the same TFBS they were compiled from, when motifs are compiled from some TFBS and applied on independent binding sites for the same TFs, and when motifs are discovered *de novo* from target sequences. Composite motif discovery methods could be systematically assessed on additional data sets, with controlled variation along several additional dimensions, e.g. variation of binding sites density in regulatory regions, variation of CRM complexity, and variation in availability of additional information beyond pure DNA sequence.

Further improvement of motif discovery methods is clearly also important, and the benchmarks suggest that there is still much room for future improvement of prediction accuracy. Future methods could borrow ideas from both of the methods presented in this thesis, and combine them with other novel elements into new methods that provide more accurate predictions, in a larger variety of settings. Acceleration of motif matching may play a role for improved usability of motif discovery. Future solutions for motif acceleration on commonly available hardware could allow highly interactive motif scanning and motif discovery, or motif discovery on much larger data sets. The

false discovery rate approach could be extended from motif scanning to also include motif discovery, although the large degree of correlations between motif candidates would bring methodological challenges. Finally, the idea of iterative data set collection as input to motif discovery could be extended to include iterations between more types of information relevant for enrichment of motif discovery data sets.

Bibliography

[1] D'haeseleer P: **How does DNA sequence motif discovery work?** *Nature biotechnology* 2006, **24**(8):959–61.

[2] MacIsaac KD, Fraenkel E: **Practical strategies for discovering regulatory DNA sequence motifs.** *PLoS Comput Biol* 2006, **2**(4):e36.

[3] Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA: **The evolution of transcriptional regulation in eukaryotes.** *Mol Biol Evol* 2003, **20**(9):1377–419.

[4] Stormo GD: **DNA binding sites: representation and discovery.** *Bioinformatics* 2000, **16**:16–23.

[5] Das MK, Dai HK: **A survey of DNA motif finding algorithms**. *BMC Bioinformatics* 2007, **8 Suppl 7**.

[6] Korn LJ, Queen CL, Wegman MN: **Computer analysis of nucleic acid regulatory sequences.** *Proc Natl Acad Sci U S A* 1977, **74**(10):4401–5.

[7] Stormo GD, Schneider TD, Gold L, Ehrenfeucht A: **Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli.** *Nucleic Acids Res* 1982, **10**(9):2997–3011.

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

[9] Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.** *Science* 1993, **262**(5131):208–14.

[10] Bailey TL, Elkan CE: **Unsupervised Learning of Multiple Motifs in Unsupervised Learning of Multiple Motifs in Biopolymers**

Using Expectation Maximization. *Machine Learning* 1995, **21**:51–80.

[11] Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z: **Assessing computational tools for the discovery of transcription factor binding sites.** *Nat Biotechnol* 2005, **23**:137–44.

[12] Pavesi G, Mauri G, Pesole G: **In silico representation and discovery of transcription factor binding sites.** *Brief Bioinform* 2004, **5**(3):217–36.

[13] Wasserman WW, Krivan W: **In silico identification of metazoan transcriptional regulatory regions.** *Naturwissenschaften* 2003, **90**(4):156–66.

[14] Bulyk ML: **Computational prediction of transcription-factor binding site locations.** *Genome Biol* 2003, **5**:201.

[15] Hannenhalli S, Levy S: **Promoter prediction in the human genome.** *Bioinformatics* 2001, **17 Suppl 1**:S90–6.

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

[17] Brazma A, Jonassen I, Vilo J, Ukkonen E: **Pattern Discovery in Biosequences**. In *ICGI '98: Proceedings of the 4th International Colloquium on Grammatical Inference*, London, UK: Springer-Verlag 1998:257–270.

[18] Berg OG, von Hippel PH: **Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters.** *J Mol Biol* 1987, **193**(4):723–50.

[19] Benos PV, Bulyk ML, Stormo GD: **Additivity in protein-DNA interactions: how good an approximation is it?** *Nucleic Acids Res* 2002, **30**(20):4442–51.

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

[21] O'Flanagan RA, Paillard G, Lavery R, Sengupta AM: **Non-additivity in protein-DNA binding.** *Bioinformatics* 2005, **21**(10):2254–2263.

[22] Stormo GD, Schneider TD, Gold L: **Quantitative analysis of the relationship between nucleotide sequence and functional activity.** *Nucleic Acids Res* 1986, **14**(16):6661–79.

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

[24] Lim LP, Burge CB: **A computational analysis of sequence features involved in recognition of short introns.** *Proc Natl Acad Sci U S A* 2001, **98**(20):11193–8.

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

[26] Zhao X, Huang H, Speed TP: **Finding short DNA motifs using permuted markov models**. In *RECOMB '04: Proceedings of the eighth annual international conference on Computational molecular biology*, New York, NY, USA: ACM Press 2004:68–75.

[27] Ben-Gal I, Shani A, Gohr A, Grau J, Arviv S, Shmilovici A, Posch S, Grosse I: **Identification of transcription factor binding sites with variable-order Bayesian networks.** *Bioinformatics* 2005, **21**(11):1367–4803.

[28] Xing EP, Jordan MI, Karp RM, Russell S: **A hierarchical bayesian markovian model for motifs in biopolymer sequences**. In *Advances in Neural Information Processing Systems*, Volume 16. Edited by Becker S, Thrun S, Obermayer K, MIT Press, Cambridge, MA 2002.

[29] Kechris KJ, van Zwet E, Bickel PJ, Eisen MB: **Detecting DNA regulatory motifs by incorporating positional trends in information content.** *Genome Biol* 2004, **5**(7):R50.

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

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

[32] Sinha S, Tompa M: **A statistical method for finding transcription factor binding sites.** *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:344–54.

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

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

[35] Shinozaki D, Maruyama O: **A Method for the Best Model Selection for Single and Paired Motifs**. In *Genome Informatics*, Volume 13, Universal Academy Press 2002:432–433.

[36] Takusagawa KT, Gifford DK: **Negative information for motif discovery.** In *Pac Symp Biocomput* 2004:360–71.

[37] Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M: **Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.** *Nature* 2005, **434**(7031):338–45.

[38] Tompa M: **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*, Heidelberg, Germany 1999:262–71.

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

[40] Pevzner PA, Sze SH: **Combinatorial approaches to finding subtle signals in DNA sequences.** *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:269–78.

[41] Pavesi G, Mauri G, Pesole G: **An algorithm for finding signals of unknown length in DNA sequences.** *Bioinformatics* 2001, **17 Suppl 1**:S207–14.

[42] Eskin E, Pevzner PA: **Finding composite regulatory patterns in DNA sequences.** *Bioinformatics* 2002, **18 Suppl 1**:S354–63.

[43] Baldwin NE, Collins RL, Langston MA, Leuze MR, Symons CT, Voy BH: **High performance computational tools for motif discovery**. In *18th International Parallel and Distributed Processing Symposium (IPDPS'04) - Workshop 9* 2004:p. 192a.

[44] Li HL, Fu CJ: **A linear programming approach for identifying a consensus sequence on DNA sequences.** *Bioinformatics* 2005, **21**(19):1838–1845.

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

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

[47] Liu X, Brutlag DL, Liu JS: **BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes.** In *Pac Symp Biocomput* 2001:127–38.

[48] Kel A, Kel-Margoulis O, Ivanova T, Wingender E: **ClusterScan: A Tool for Automatic Annotation of Genomic Regulatory Sequences by Searching for Composite Clusters**. In *Proceedings of the German Conference on Bioinformatics* 2001:96–101.

[49] Hu YJ: **Finding subtle motifs with variable gaps in unaligned DNA sequences.** *Comput Methods Programs Biomed* 2003, **70**:11–20.

[50] Thompson W, Palumbo MJ, Wasserman WW, Liu JS, Lawrence CE: **Decoding human regulatory circuits.** *Genome Res* 2004, **14**(10A):1967–74.

[51] Bussemaker HJ, Li H, Siggia ED: **Regulatory element detection using correlation with expression.** *Nat Genet* 2001, **27**(2):167–71.

[52] Wasserman WW, Fickett JW: **Identification of regulatory regions which confer muscle-specific gene expression.** *J Mol Biol* 1998, **278**:167–81.

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

[54] Rebeiz M, Reeves NL, Posakony JW: **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* 2002, **99**(15):9888–93.

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

[56] Sinha S, van Nimwegen E, Siggia ED: **A probabilistic method to detect regulatory modules.** *Bioinformatics* 2003, **19 Suppl 1**:i292–301.

[57] Bailey TL, Noble WS: **Searching for statistically significant regulatory modules.** *Bioinformatics* 2003, **19 Suppl 2**:II16–II25.

[58] Frith MC, Hansen U, Weng Z: **Detection of cis-element clusters in higher eukaryotic DNA.** *Bioinformatics* 2001, **17**(10):878–89.

[59] Xing EP, Wu W, Jordan MI, Karp RM: **Logos: a modular bayesian model for de novo motif detection.** *J Bioinform Comput Biol* 2004, **2**:127–54.

[60] Gupta M, Liu JS: **De novo cis-regulatory module elicitation for eukaryotic genomes.** *Proc Natl Acad Sci U S A* 2005, **102**(20):7079–84.

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

[62] Frech K, Werner T: **Specific modelling of regulatory units in DNA sequences.** In *Pac Symp Biocomput* 1997:151–62.

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

[64] Sharan R, Ovcharenko I, Ben-Hur A, Karp RM: **CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments.** *Bioinformatics* 2003, **19 Suppl 1**:i283–91.

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

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

[67] Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB: **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* 2002, **99**(2):757–62.

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

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

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

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

[72] Mahony S, Hendrix D, Golden A, Smith TJ, Rokhsar DS: **Transcription factor binding site identification using the self-organizing map.** *Bioinformatics* 2005, **21**(9):1807–1814.

[73] Beiko RG, Charlebois RL: **GANN: genetic algorithm neural networks for the detection of conserved combinations of features in DNA.** *BMC Bioinformatics* 2005, **6**:36.

[74] Bailey TL, Elkan C: **The value of prior knowledge in discovering motifs with MEME.** *Proc Int Conf Intell Syst Mol Biol* 1995, **3**:21–9.

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

[76] Rigoutsos I, Floratos A: **Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm.** *Bioinformatics* 1998, **14**:55–67.

[77] Evans PA, Smith AD: **Toward optimal motif enumeration**. In *Proceedings of Workshop on Algorithms and Data Structures (WADS 2003)*, Volume 2751 of *LNCS*, Springer-Verlag 2003:47–58.

[78] Lawrence CE, Reilly AA: **An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences.** *Proteins* 1990, **7**:41–51.

[79] Sinha S, Blanchette M, Tompa M: **PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences.** *BMC Bioinformatics* 2004, **5**:170.

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

[81] Ao W, Gaudet J, Kent WJ, Muttumu S, Mango SE: **Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR.** *Science* 2004, **305**(5691):1743–6.

[82] Neuwald AF, Liu JS, Lawrence CE: **Gibbs motif sampling: detection of bacterial outer membrane protein repeats.** *Protein Sci* 1995, **4**(8):1618–32.

[83] Thompson W, Rouchka EC, Lawrence CE: **Gibbs Recursive Sampler: finding transcription factor binding sites.** *Nucleic Acids Res* 2003, **31**(13):3580–5.

[84] Gupta M, Liu JS: **Discovery of Conserved Sequence Patterns Using a Stochastic Dictionary Model**. *Journal of the American Statistical Association* 2003, **98**:55–66.

[85] Jensen ST, Liu XS, Liu JS, Zhou Q: **Computational Discovery of Gene Regulatory Binding Motifs: A Bayesian Perspective**. *Statist Sci* 2004, **19**:188–204.

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

[87] Grad YH, Roth FP, Halfon MS, Church GM: **Prediction of similarly-acting cis-regulatory modules by subsequence profiling and comparative genomics in D. melanogaster and D. pseudoobscura**. *Bioinformatics* 2004, **20**(16):2738–2750.

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

[89] Liu XS, Brutlag DL, Liu JS: **An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments.** *Nat Biotechnol* 2002, **20**(8):835–9.

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

[91] Bortoluzzi S, Coppe A, Bisognin A, Pizzi C, Danieli G: **A Multistep Bioinformatic Approach Detects Putative Regulatory Elements In Gene Promoters.** *BMC Bioinformatics* 2005, **6**:121.

[92] Pizzi C, Bortoluzzi S, Bisognin A, Coppe A, Danieli GA: **Detecting seeded motifs in DNA sequences.** *Nucleic Acids Res* 2005, **33**(15):e135.

[93] Hon LS, Jain AN: **A deterministic motif finding algorithm with application to the human genome.** *Bioinformatics* 2006, **22**(9):1047–1054.

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

[95] Conlon EM, Liu XS, Lieb JD, Liu JS: **Integrating regulatory motif discovery and genome-wide expression analysis.** *Proc Natl Acad Sci U S A* 2003, **100**(6):3339–44.

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

[97] Holmes I, Bruno WJ: **Finding regulatory elements using joint likelihoods for sequence and expression profile data.** *Proc Int Conf Intell Syst Mol Biol* 2000, **8**:202–10.

Biology Direct

BioMed Central

Review

Open Access

A survey of motif discovery methods in an integrated framework

Geir Kjetil Sandve*[1] and Finn Drabløs[2]

Address: [1]Department of Computer and Information Science, NTNU – Norwegian University of Science and Technology, N-7052, Trondheim, Norway and [2]Department of Cancer Research and Molecular Medicine, NTNU – Norwegian University of Science and Technology, N-7006, Trondheim, Norway

Email: Geir Kjetil Sandve* - sandve@idi.ntnu.no; Finn Drabløs - finn.drablos@ntnu.no

* Corresponding author

Published: 06 April 2006

Biology Direct 2006, **1**:11 doi:10.1186/1745-6150-1-11

Received: 29 March 2006
Accepted: 06 April 2006

This article is available from: http://www.biology-direct.com/content/1/1/11

Abstract

Background: There has been a growing interest in computational discovery of regulatory elements, and a multitude of motif discovery methods have been proposed. Computational motif discovery has been used with some success in simple organisms like yeast. However, as we move to higher organisms with more complex genomes, more sensitive methods are needed. Several recent methods try to integrate additional sources of information, including microarray experiments (gene expression and ChIP-chip). There is also a growing awareness that regulatory elements work in combination, and that this combinatorial behavior must be modeled for successful motif discovery. However, the multitude of methods and approaches makes it difficult to get a good understanding of the current status of the field.

Results: This paper presents a survey of methods for motif discovery in DNA, based on a structured and well defined framework that integrates all relevant elements. Existing methods are discussed according to this framework.

Conclusion: The survey shows that although no single method takes all relevant elements into consideration, a very large number of different models treating the various elements separately have been tried. Very often the choices that have been made are not explicitly stated, making it difficult to compare different implementations. Also, the tests that have been used are often not comparable. Therefore, a stringent framework and improved test methods are needed to evaluate the different approaches in order to conclude which ones are most promising.

Reviewers: This article was reviewed by Eugene V. Koonin, Philipp Bucher (nominated by Mikhail Gelfand) and Frank Eisenhaber.

Open peer review

Reviewed by Eugene V. Koonin, Philipp Bucher (nominated by Mikhail Gelfand) and Frank Eisenhaber. For the full reviews, please go to the Reviewers' comments section.

Introduction

Understanding the regulatory networks of higher organisms is one of the main challenges of functional genomics. Gene expression is regulated by transcription factors (TF) binding to specific transcription factor binding sites (TFBS) in regulatory regions associated with genes or gene clusters. Identification of regulatory regions and binding

sites is a prerequisite for understanding gene regulation, and as experimental identification and verification of such elements is challenging, much effort has been put into the development of computational approaches. Good computational methods can potentially provide high-quality prediction of binding sites and reduce the time needed for experimental verification. However, the computational approach has turned out to be at least as challenging as the experimental one, and a very large number of different methods have been developed.

Computational discovery of regulatory elements is mainly possible because they occur several times in the same genome, and because they may be evolutionary conserved. This means that novel regulatory elements may be discovered by searching for overrepresented motifs across regulatory regions. However, this apparently simple approach is complicated by the fact that most binding site motifs are short, and they may also show some sequence variation without loss of function. Therefore most motifs are also found as random hits throughout the genome, and it is a challenging problem to distinguish between these false positive hits and true binding sites.

One of the early origins of DNA motif discovery is the computer program written in 1977 by Korn *et al.* [1] that was able to discover sequence similarities in regions immediately upstream of TSS. Both mismatches and flexible gaps were accounted for, but using only pairwise comparisons. This approach was further developed by Queen *et al.* [2], comparing multiple sequences simultaneously. In this work, the exact requirements of a motif was also defined clearly, with quorum constraints on sequence support, max number of mismatches in occurrences, and max distances between occurrence positions in the different sequences. In the same year, Stormo *et al.* [3] introduced a Perceptron algorithm that calculated the sum of independent weighted match scores for each position of a motif aligned with a sequence. Similar to this, Staden [4] introduced a position weight matrix with weights corresponding to log-frequencies of nucleotides in aligned motif occurrences. A very nice historical account of the early development of motif models is given in [5].

The most common approach to *de novo* computational discovery of regulatory elements is to extract a set of sequences from the genome, typically fixed size upstream regions for a set of genes having e.g. 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 to this basic approach may be used to increase its sensitivity, by including additional prior knowledge about gene regulation. Regulatory elements are not randomly distributed, but tend to form clusters of regulatory modules. The context of putative regulatory elements may also be important, such as other nearby elements, the presence of CpG-islands, or the position in the overall DNA structure. Individual genes in a gene set may show different levels of co-regulation e.g. in a microarray experiment, and this may be used as a weight function to increase the influence from potentially important genes. Finally, additional sources of information, such as regulatory regions of orthologous genes, will often be available.

More than a hundred methods have been proposed for motif discovery in recent years, representing a large variation with respect to both algorithmic approaches as well as the underlying models of regulatory regions. There is also large variation regarding how methods are described and tested, making it even harder to get a good overview of the field. Many reviews of motif discovery methods have therefore been written, with varying focus and intended audience. The recent review by Pavesi *et al.* [6] is a very accessible and broad introduction to the field. It divides methods into consensus- and alignment-based, and surveys the most established methods one at a time. It also discusses background modeling, evaluation of motifs and the practicalities of using these methods. The review by Wasserman and Krivan [7] has a stronger focus on the underlying biology of motif discovery in regulatory regions. It also goes a bit more into the combinatorial nature of binding sites, and touches upon issues such as phylogenetic footprinting, CpG-islands and chromatin structure. Finally, some reviews focus on specific techniques such as phylogenetic footprinting [8], or on specific genomes [9].

Here we present a structured framework for describing motif discovery methods, where we focus on the modeling of regulatory regions, in particular in eukaryote genomes, and with a finer level of detail compared to previous surveys. The emphasis is on how the multiple binding sites for modules of combinatorially acting regulatory elements can be modeled, and how additional data sources may be integrated into such models.

Our framework allows for a systematic and quite exhaustive survey of recent methods. Here we survey methods with respect to individual elements of our model, which makes it easier to spot important differences and similarities between methods. Furthermore, this approach reveals important differences between methods on aspects that in most papers are not discussed as deliberate choices. Relevant examples are how matching scores of several motifs in a module are combined, and how the score of multiple binding sites for the same factor is calculated.

As discussed e.g. by Tompa *et al*. [10] it is very difficult to compare the performance of methods, in particular on complex genomes like the human. Furthermore, methods will also differ in aspects like average running time, need for manual parameter-tuning, exhaustiveness of results, general usability and so on. Individual methods may also perform better on one type of genomes compared to others, making it difficult to compare performance on a general scale. We have therefore to a large extent deliberately avoided comparing relative performance of individual methods. We mainly indicate important elements of the problem, and show the breadth of possible solutions that have been tested, both when it comes to established elements of motif discovery, such as single motif models, as well as less common approaches, such as the incorporation of DNA structure. However, there is a definite need for more standardized routines for testing and comparing alternative approaches to motif discovery, and the work by Tompa *et al*. [10] is an important step in that direction.

Biological background

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 [11]. The sequence level includes coding regions, regulatory binding sites and sequence elements affecting the 3-dimensional fold of the chromatin fiber. It is mainly the binding sites for transcription factors that will be discussed 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) organization of DNA, it is also organized 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 [12], Wray *et al*. [13] and Pedersen *et al*. [14]. The key regulatory region is the promoter region, 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 organized 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 region where cis-regulatory elements are found can vary by nearly three orders of magnitude from a few hundred bp to >100 kb. Regulatory 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 [15].

An integrated framework

As motif discovery methods can be very complex, with many possible differences, several authors have proposed frameworks for classifying motif discovery methods. Brazma *et al*. [16] categorize motif discovery methods with respect to whether they use explicit negative sequence sets or not, expressiveness of the pattern models, whether patterns are deterministic or statistical, and whether the algorithms are pattern driven or sequence driven. In a later paper Brazma *et al*. [17] define a three step paradigm consisting of choosing a class of grammars (motif model), designing a rating function (motif score), and developing an algorithm. However, the major recent advances in the field have been on modeling of regulatory regions, rather than individual sites, and on integration of additional data. The frameworks mentioned above are not well suited to highlight developments in these directions. We therefore use an extended, integrated framework for the description of motif discovery methods, where both the representation of the transcription factor based regulatory system itself, as well as additional sources of information, can be represented.

The most basic level of our framework (Level 1) represents the binding of transcription factors (TFs) to short contiguous sequence segments. These sequence segments are modeled by single motif models 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 may occur at the given location.

The next level of our framework (Level 2) represents modules: clusters of TFs that bind to DNA in proximity to each other, but with a certain flexibility regarding distance 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 can be calculated from the score of single motifs at given positions as well as inter-motif distances.

The third level of the framework (Level 3) represents how several modules may act together, possibly in a combinatorial manner, to determine the regulation of a single gene. This is modeled by a gene score function that combines composite motif scores across the regulatory reglon(s).

The final level of our framework (Level 4) represents several sets of modules acting on sets of genes, e.g. at the genome level. Scores at this level are mostly used for evaluation and ranking of *de novo* discovered motifs. The evaluation is based either on overrepresentation of motifs, or on correspondence between motif scores and experimental data.

A schematic view of our framework, reflecting the different levels of regulatory processes, is given in Figure 1. The different elements of this figure will be described in more detail in the following sections.

We will now use this framework to categorize a large number of existing methods for motif discovery. Table 1 gives an overview of how various elements of our framework are approached by selected methods, including both novel and more established approaches. A larger table, which includes most current methods, is available as supplementary material [18].

Single motif models (Level 1)

Transcription factors bind to specific short segments of DNA, transcription factor binding sites. This is the most basic element of the regulatory system, and can be modeled using single motif models. *A single motif model is defined as a function* $m_g : \mathbb{N} \rightarrow \mathbb{R}$ *that maps a sequence position* p *as a non-negative integer to a real numbered motif score* $m_g(p)$. *It consists of a match score* $m^*(p)$ *and an occurrence prior* $o_g(p)$.

The function $m_g(p)$ returns a value indicating whether an occurrence of the motif is found at position p. 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 p and an underlying consensus model. The occurrence prior, $o_g(p)$, gives the prior belief that position p represents a regulatory element for gene g.

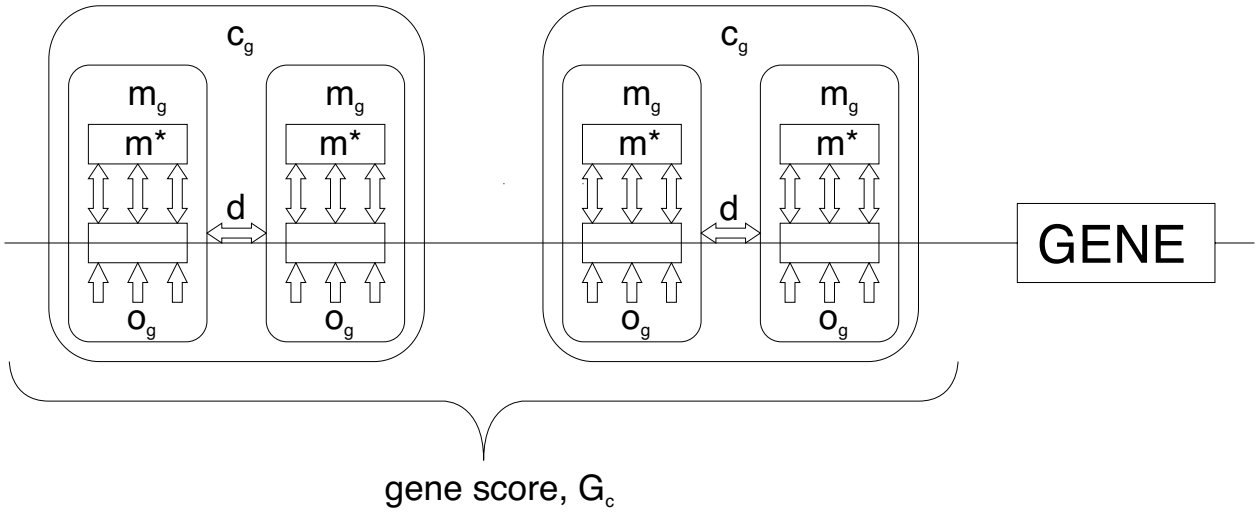


Figure 1

A schematic view of the integrated framework. A single motif, denoted by m_g, consists of two parts, m_g is how well the sequence matches a consensus, while o_g is a prior on whether any regulatory element is to occur at that position. A set of single motifs, together with inter-motif distance restrictions (d), then forms a composite motif (c_g). Finally, multiple occurrences of a composite motif in the regulatory regions of a gene is represented by a gene score G_c.

Table 1: Overview of methods. The match model is the consensus representation of a single motif, motif combination is how the component scores of a composite motif are combined, and distance score is how the conservation of inter-motif distances within a composite motif is modeled.

ALGORITHM NAME	MATCH MODEL	MOTIF COMBINATION	DISTANCE SCORE
Weeder [42]	mismatch	-	-
Dyad analysis [35]	oligos	dyad[1]	constraint
MCAST [71]	PWM	sum	gap penalty
REDUCE [67]	PWM	dyad	constraint[2]
MDScan [87]	PWM	-	-
Gibbs sampler [97]	PWM	intersection[3]	uniform
MEME [98]	PWM	-	-
LOGOS [73]	DM	HMM	distribution
Motif regressor [89]	PWM	-	-
ModuleSearcher [70]	PWM	sum	window[4]
Stubb [48]	PWM	HMM	window
GANN [60]	flexible	ANN[5]	window
ANN-Spec [86]	PWM	-	-
(Wasserman) [58]	PWM	Logistic regr.	window
CoBind [68]	PWM	sum	window
Cister [72]	PWM	HMM	distribution
SeSiMCMC [122]	PWM	-	-
SMILE [40, 123]	mismatch	intersection	constraint
BioProspector [49]	PWM	sum	constraint
(Segal) [94]	PWM	-	-
(Sinha) [33]	reg.exp	dyad	constraint
ConsecID [56]	PWM	intersection	window
SCORE [69]	IUPAC	intersection	window
Gibbs recursive [52]	PWM	mixture model	distribution
(Hong) [95]	PWM	-	-
AlignACE [124]	PWM	-	-
Improbizer [117]	PWM	-	-
CisModule [119]	PWM	mixture model	mixture model
(Thompson) [66]	PWM	Markov model	constraint

[1]Two single motifs that both have to occur
[2]Separate constraints on each inter-motif distance
[3]Several single motifs that all have to occur
[4]All single motifs have to occur within a sequence window of restricted length
[5]Artificial neural network

Match models

In the most general sense, the match model $m^*(p)$ 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 (e.g. 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 known as position specific scoring matrix (PSSM), that assumes independence between positions [3]. 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 [19].

Many different extensions to the basic PWMs have 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 [20-22].

The most direct way of incorporating dependencies within motifs is to extend the PWM to include pairs of correlated positions [21,23]. Another straightforward approach is to use a mixture model in which the motif occurs as one of a limited number of stochastic prototypes [24]. 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

[25]. However, it is hard to find a good compromise between a high n that may give too many free parameters and a low n that may miss out the dependencies of interest. If the relative importance of dependencies varies within a motif, a variable-length Markov model (VLMM) [26] may be preferable. Furthermore, if some long-range dependencies seem to be significantly stronger than dependencies between neighboring positions, the order of the positions in the Markov chain may also be permuted before a VLMM is applied [27].

Another way to model dependencies is to use Bayesian networks. Barash *et al*. [24] 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. Similarly, Ben-Gal *et al*. [28] argue for variable order Bayesian nets.

Instead of focusing on dependencies between specific nucleotides at different positions, Xing *et al*. [29] model the distribution of conserved positions within a motif. 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. The work of Kechris *et al*. [30] achieves similar properties by assigning conservation types (strong, moderate or low) to blocks of motif positions.

Deterministic match models

A deterministic match model evaluates to a binary value indicating either 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 [31-33] and dictionary models [34].

A regular expression model $m^*(p)$ returns 1 if the given substring is matched by an underlying regular expression. As reviewed by Brazma *et al*. [16], the models used in motif discovery are typically composed of exact symbols, ambiguous symbols, fixed gaps and/or flexible gaps. Regular expression models are used in e.g. [33,35-38].

Many methods use mismatch expressions as motif match models, e.g. [39-44]. 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 [45], 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 [46].

The probabilistic models are much more expressive than the deterministic models. In fact, all oligos, regular expressions and mismatch expressions can be represented as PWMs. However, a major benefit of the deterministic models is that they often allow exhaustive discovery of optimal motifs.

Occurrence priors

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

The simplest kind of occurrence prior is a motif abundance ratio [47]. 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 [48]. Several methods including Bioprospector [49] and TFBScluster [50] 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 known elements are located immediately upstream of the transcription start site (TSS). In general, this can be represented by a function giving 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 [51] and [52]. 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 for $o_g(p)$.

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 [53].

The reasoning behind phylogenetic footprinting is that since regulatory elements are functionally important and are under evolutionary selection, they should evolve much more slowly than other non-coding sequences. Moreover, genome-wide sequence comparisons and studies of individual genes have confirmed that regulatory elements are indeed conserved between related species [54]. More specifically, Krivan and Wasserman [55] 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 gene regulatory regions by searching for motifs only in highly conserved sequence parts (typically human-mouse orthologs) [44,48,56,57]. This approach corresponds to using a binary occurrence prior that is 1 if the conservation score is above a given threshold and 0 otherwise. Wasserman and Fickett [58] use non-binary conservation scores, but they do not incorporate these into the search as priors. Instead, they use conservation to filter the discovered motifs. Similarly, Xie *et al*. [38] calculates the proportion of motif occurrences that are conserved in related species, and uses this in the evaluation of motif significance. Finally, Wang and Stormo [59] constructs phylogenetic profiles, representing the frequency of nucleotides in each position based on multiple alignment of promoters in related species.

DNA structure

The three-dimensional structure of DNA, densely packed as chromatin, inhibits transcriptional initiation *in vivo* [14]. The bendability of a region, as well as its position in DNA loops, may indicate whether it contains regulatory elements or not. Only a few motif discovery methods take DNA structure into consideration. Beiko and Charlebois [60] average structure scores of all k-mers in a window around a given position, independently of any particular motif. Conversely, Pudimat *et al*. [61] incorporate helical parameter features [62,63] in a Bayesian net that is specific for each motif.

Nudeotide distribution

Both high GC content and presence of CpG-islands may indicate that a region contains regulatory elements. The method of Pudimat *et al*. [61] is one of a few methods that take GC content and CpG-islands into consideration when calculating motif scores.

Composite motif models (Level 2)

Clusters of binding sites for cooperating TFs, often called modules, are believed to be essential building blocks of the regulatory machinery. Werner [12] states that "Within a promoter module, both sequential order and distance can be crucial for function, indicating that these modules may be the critical determinants 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 this. It is therefore natural to define a computational motif model that represents a combination of single motifs.

A composite motif model is defined as a function c_g: $2^N \rightarrow \mathbb{R}$ *that maps a set of single motif sequence positions* $\vec{p}$ *as non-negative integers to a real numbered composite motif score* $c_g(\vec{p})$. *It consists of single motifs* $\vec{m}_g$.

The function $c_g(\vec{p})$ consists of a set of (generally different) single motifs $\vec{m}_g$, with each single motif contributing with 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 individual single motif and distance scores.

Distance functions

Many different models have been proposed 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 [33,49], distances below thresholds [64-66], or distances within intervals (e.g. [33,35,43,49,67]).

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. [48,58,68-70]). 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 [71]. Similarly, a geometric distribution on inter-motif distances follows implicitly from many HMM models [72,73], and is assumed explicitly in Gupta and Liu [74].

The conservation of inter-motif distances across modules can also serve as a basis for distance score functions. Wagner [75] 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 dis-

tances. Similarly, Frech and Werner [76] calculate scores by comparing the distances with a histogram of distances between the same regulatory elements in other modules.

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 [36], which corresponds to a simple binary distance function. More importantly, as our understanding of DNA folding increases, new and more complex distance measures may appear.

Combining single motifs

There are many ways in which a set of single motif and distance scores can be combined into 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. [56,75,77,78]). A variation of this is to require that *M* out of *N* single motif scores are 1 [79]. Similarly, the count of binary single motif values can be used directly as a composite motif score [33,80,81].

For methods that use non-binary single motif scores, a common approach is to calculate the sum of single motif and distance scores [71,76]. Some methods require that all distance functions are 1, and if they are, composite motif score is the sum of single motif scores [68,70,82,83].Similarly, the method Modulescanner sums only single motif scores above a threshold, and MotifLocator sums the *N* highest single motif scores [70]. Another 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 the single motifs [64]. Finally, a few methods take the composite motif score to be the highest single motif score [42], or the lowest single motif score [84].

Many specialized models have also been used to combine single motif and distance scores, e.g. the hidden Markov model (HMM) [73], history-conscious HMM (hcHMM) [48], self-organizing map (SOM) [85], and artificial neural network (ANN) [60]. In all of these models, the score of several homotypic and/or heterotypic single motifs are combined in a relatively complex way.

Gene level models (Level 3)

In addition to the motif scores, which are defined for specific positions, we may also be interested in the presence of motifs across the regulatory regions of a gene. The possibility of multiple binding sites for TFs is often not discussed explicitly in articles presenting motif discovery methods. Scores at this level may, however, be relevant both when predicting which genes are regulated by a TF or module, and when evaluating the significance of a *de novo* discovered motif.

A gene score model is defined as a function $G_c: \mathbb{N} \to \mathbb{R}$ *that maps a gene index g as a non-negative integer to a real numbered gene score* $G_c(g)$. *It consists of composite motif models* $c_g(\vec{p})$.

The gene level score is calculated from composite motif scores, $c_g(\vec{p})$, across the regulatory region of gene *g*, and is referred to as gene score. For methods that only discover binding sites for single TFs, the composite motif score is simply the single motif score.

Multiple binding sites

The gene level score is often defined simply as the maximum motif score in the regulatory region(s) of a gene [46,70,81,86,87]. This corresponds to an implicit assumption of exactly one relevant occurrence of a motif in the regulatory reglon(s).

It is, however, reasonable to assume 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 from all motif scores across the regulatory region(s) of a gene. As motif scores are typically log-scores, most methods add the exponentials of motif scores (e.g. [67,68,88-90]). A slight variation is to only sum motif scores above a certain threshold [71].

In addition to these approaches, many variations have been used to calculate gene score. Caselle *et al*. [91] and Cora *et al*. [57,92] calculate gene score as the *p*-value of the observed set of motif scores. Curran *et al*. [93] calculate gene scores based on logistic regression. Similarly Segal *et al*. [94] use a logistic function, and Hong *et al*. [95] a hyperbolic tangent, on the sum of motif scores. Finally, Beiko *et al*. [60] use an artificial neural network to combine motif scores.

The dictionary models of Bussemaker *et al*. [34] and Gupta and Liu [96] represent a special case, as they 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.

Multiple modules

In addition to multiple binding sites for the same module, a set of different modules may also be introduced at the gene level. A gene may be seen as having several regulatory

regions, with tight distance constraints between binding sites within a regulatory region (module), and larger and more variable distances between different regulatory regions. Xing *et al*. [73] define an HMM that can represent different modules of binding sites with different implicit geometric distributions within and between modules. This model can also represent different intra-module background distributions in addition to the global inter-module background distribution. This corresponds to a gene score that is calculated from the scores of several different composite motifs across the regulatory regions of a gene.

Genome level models (Level 4)

Motif scores at the genome level are generally used for significance evaluation of *de novo* motifs, although it may in some situations also be relevant to look at the presence of motifs (TFs or modules) in different genomes. Here we focus on the first situation, evaluation of motif significance at the genome level. In most cases the genome level score is based on just the (assumed) regulatory regions for a selected subset of the genes.

A genome score model is defined as a function $s_{c,F}: \mathbb{N} \rightarrow \mathbb{R}$ *that maps a genome index i as a non-negative integer to a real numbered genome score* $s_{c,F}(i)$. *It consists of a gene score model* $G_c(g)$ *and a gene membership function* $\mu F(g)$.

Genome score (motif significance) is typically based on either the genome level overrepresentation of the motif, or on the correspondence between gene scores and experimental data.

Motif overrepresentation

Computational motif discovery is possible primarily because motifs representing regulatory motifs 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 from a background model. More specifically, the *p*-value [37,69] and *z*-score [33,39] 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. Shuffled control sequences may also be used as background [97].

A simpler approach is to compare only the raw sum of gene scores when ranking motifs. This is equivalent to 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 information content (IC) of discovered PWMs [98]. 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 [40,99,100].

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 [56].

Correspondence with experimental data

In recent years, the development of microarray technology has revolutionized 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 synchronized 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 [101], although it has been shown that this type of data mining is difficult and error prone [10]. A variant of this approach is to cluster genes based on expression similarity with specific transcription factors [102,103].

Recently, genome-wide binding analysis like ChIP/chip experiments have appeared as an approach for more reliable identification of actual binding site regions [104,105]. 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 [46,88,106,107], or genes with similarities in functional annotation [32,57]. 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 [108].

Many methods cluster 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. [66,93,95,108,109]). However, as a gene may be co-regulated with several groups of genes, depending on conditions, it may make sense to 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*. [81] and Liu *et al*. [87] are among the few authors that have used weighted values for set membership during motif discovery.

The correspondence between gene level 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 $G_c(g)$ to membership values $\mu F(g)$ [58,93]. A simplification of this approach is to compare binary gene scores with binary membership values, and calculate the mismatch ratio [95] or ROC_{50} score [71]. Alternatively, grouping of genes can be avoided altogether, and motif significance can be measured as the fit of a linear regression directly from gene scores to observed log-expression in microarray experiments [67,89,94].

Park *et al*. [110] 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 [111] 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 [32,57,108], only a few methods try to combine different kinds of data in a single similarity measure. Takusagawa and Gifford [37] use the GRAM algorithm [112] 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.

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 [99,100,113].

However, probabilistic motif discovery algorithms do not guarantee returning the global optimum when applied to realistic problems. These algorithms are typically based either on iterative refinement or stochastic optimization. Expectation maximization (EM) [98,114-117] is the most widely used iterative refinement method, but variational EM [73] has also been used. The stochastic optimization technique most widely used for motif discovery is Gibbs sampling [49,52,97,118], sometimes combined with general Metropolis-Hastings [47,96,119]. Recently, genetic algorithms [82], evolutionary Monte Carlo [74] and simulated annealing [27,81,120] has also gained some popularity.

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 [121] or exhaustively [100]. A promising approach to motif discovery is first to 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 be used as seeds for motif discovery in the full set of sequences. The method of Liu *et al*. [87] 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.

Comparison of methods

Given the very large number of different methods for motif discovery, it is obviously crucial to have good test methods in order to identify the most promising approaches. However, this has turned out to be a challenging problem by itself.

It is difficult to identify optimal test sets for benchmarking. When comparing the performance of methods the output has to be compared against some biological truth.

Even though biological sequences with experimentally verified binding sites are available, they may contain additional (yet unidentified) binding sites that may show up as false positives in motif discovery. Using implanted motifs in synthetic background sequences may avoid this problem, but creates new problems with respect to realistic background sequences and motif distributions, in particular for composite motifs. It may also be difficult to get enough data to get a good representation of the diversity of regulatory regions.

It is also difficult to know whether a test result actually reflects the assumed methodological difference between alternative approaches. Many methods will require different degrees of parameter tuning. This may introduce bias in test results, and makes automatic testing difficult. Typical examples of tunable parameters may be motif length, expected number of motif occurrences, and inter-motif distances. Also, many methods make use of additional data, in addition to the actual sequences, in order to increase performance. For instance, several methods include phylogenetic footprinting using related organisms. Finally, different implementations may have been optimized and fine tuned to different degree. This makes it difficult to distinguish between the performance of underlying algorithmic approaches and the effect of several years of tweaking on a specific implementation. If radically different and possibly better performing approaches are to be identified, it is essential that novel algorithmic approaches are tested against existing methods in comparable frameworks and implementations.

These challenges make it difficult to actively compare the performance of alternative approaches and use this as a basis for recommendations. The seminal benchmark of single motif discovery methods by Tompa *et al*. [10] mainly concludes that biologists are advised to use a few complementary tools in combination rather than relying on a single one, and to pursue the top few predicted motifs of each rather than the single most significant motif of any given method. Some of the most established methods, such as MEME, AlignACE and ANN-Spec, performed reasonably well, at least on simple data (e.g. yeast). However, the best method overall on these datasets was the more recent method Weeder. Only single motif discovery was tested in this work. No other study of comparable breadth has tested composite motif discovery methods, probably because it is even more challenging to find suitable test sets and to evaluate alternative methods for composite motifs.

However, on a more general basis we believe that some recent developments on expressive models for combination of motifs are particularly interesting. The method "motif regressor" represents a relatively simple, yet promising approach [89]. First it uses the MDScan algorithm [87] to discover single motifs based on CHiP-chip data. Motifs that are too similar to the background distribution are filtered out, and the remaining motifs are used as features in a multiple regression from gene level scores of motifs to gene expression levels. In this way, only motifs that serve (independent) explanatory roles on gene expression are retained. Another interesting approach is the LOGOS method [73] that uses a hidden Markov model (HMM) to model the combinatorial nature of binding sites. Furthermore, single motifs are modeled by a HMDM model [29] that promotes binding sites with certain spatial distributions on single nucleotide conservation. All of this is combined using a coherent probabilistic model.

Conclusion

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 as well as differences between methods. Most papers on novel computational methods tend to focus on the authors' own data sets and scientific problems. Hence, the authors often put less emphasis on giving a clear description of the algorithm itself, e.g. precisely what it requires as input, how it evaluates motifs, and what it returns as output. This makes it harder to compare methods based on their descriptions.

When trying to compare the accuracy and computational efficiency of methods by measurement, there are 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 [10].

Establishing a standardized framework for testing would be an important contribution to the field. Such a framework should include a collection of diverse data sets and several complementary measures of performance. Furthermore, a consensus on what constitutes essential aspects of motif discovery methods could ease the comparison of methods, making it easier to choose between or integrate different approaches. This could also make it easier for researchers to identify the choices that have to be made when a new model or approach is being developed, as well potential previous models where these choices already have been evaluated. The integrated model described in this paper may be one step towards a common vocabulary and framework for this problem.

When surveying recent literature we have made several interesting observations. One is the sheer breadth of approaches used in the field when it comes to how motifs are modeled and how experimental information is integrated. A somewhat related observation is the great variation between motif models, even when it comes to aspects that are typically not discussed explicitly in papers, e.g. how the gene level score is calculated. In other words, some papers implicitly treat the chosen model as obvious and the only possible solution, whereas comparison to similar methods shows that there indeed are several possible approaches that should have been evaluated.

A third observation is that even though there are many aspects of a basic motif model that can be improved, each article typically considers only one of them. If we add together the possible enhancements to different parts of the models for regulatory regions, and the different kinds of additional data that have been incorporated, based on all papers in the field, wee see a much more complex and enhanced model. Although such a model may be too complex for a full implementation, one should at least make deliberate choices with respect to which elements are included in a given approach. Hopefully the integration of techniques and experiences across existing approaches will give rise to refined and advanced methods with higher sensitivity than what we have seen so far.

Reviewers' comments

Reviewer's report 1

Eugene V. Koonin, National Institutes of Health, Bethesda, MD, USA

This is a detailed and useful survey of the computational approaches used for discovery of sequence motifs in DNA, with an emphasis on transcription-factor-binding sites. The paper is well-structured and properly referenced. I believe that many researchers will find it helpful.

Reviewer's report 2

Philipp Bucher, Swiss Institute of Bioinformatics and Swiss Institute for Experimental Cancer Research, Switzerland (nominated by Mikhail Gelfand, Institute of Information Transfer Problems, Moscow, Russia)

This article clearly responds to a need. The literature on motif discovery methods has grown vast, confronting the reader with a bewildering variety of methods and concepts. The authors rightly point out that the different methods are not always appropriately described in the scientific articles. Underlying assumptions are often not explicitly stated, and methodological choices are not mentioned as they may appear self-explanatory to the developers.

This comprehensive review makes and attempt to consolidate the field by providing a framework for categorizing the large number of existing motif discovery methods. The various methods are classified according to four hierarchical levels of genome organization: Individual motifs, composite elements, genes, and genomes. This framework is useful from a biological perspective as it allows for joint presentation and comparison of methods that address similar questions. A potential drawback is that technical issues may be arbitrarily spread over different parts of the manuscript. For instance, it is debatable whether the significance measure related to the information content of a PWM, which is used by MEME, should be presented under the heading " genome level models".

What is lacking in this review is a historical perspective. The manuscript focuses on recent work disregarding largely how current concepts have evolved over time. I would propose to add some of the earlier landmark papers to the bibliography, for instance:

Korn LJ, Queen CL, Wegman MN. (1977) Computer analysis of nucleic acid regulatory sequences. Proc Natl Acad Sci USA. 10:4401–4405. This is perhaps the first paper describing a computer algorithm that helps to find an over-represented sequence motif.

Queen C, Wegman MN, Korn LJ. (1982) Improvements to a program for DNA analysis: a procedure to find homologies among many sequences. Nucleic Acids Res. 10:449–456. Perhaps the first paper implicitly using a mismatch model for motif discovery. It also presents an efficient algorithm to find optimal motifs of this type.

Staden R. (1984) Computer methods to locate signals in nucleic acid sequences. Nucleic Acids Res. 12:505–19. First paper proposing PWMs with weights proportional to the logarithms of the observed base frequencies.

Brendel V, Trifonov EN. (1984) A computer algorithm for testing potential prokaryotic terminators. Nucleic Acids Res. 12:4411–4427. This work extends position independent weight matrices to dinucleotide matrices, thereby accounting for nearest-neighbor dependencies.

Galas DJ, Eggert M, Waterman MS. (1985) Rigorous pattern-recognition methods for DNA sequence sequence analysis of promoter sequences from Escherichia coli. J. Mol. Biol. 186:117–128. An early paper presenting a method that takes into account a motif's distance to the transcription start site.

Berg OG, von Hippel PH (1987) Selection of DNA binding sites by regulatory proteins, statistical-mechanical theory and application to operators and promoters. J. Mol.

Biol. 193: 723–750. Provides a physical (thermodynamic) interpretation of PWMs.

Author response: *We have added a brief historical overview to the introduction, including most of the references mentioned here.*

Regarding present-day genome-wide approaches, the following two papers may be worthwhile to mention: Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature. 434:338–345.

Wang T, Stormo GD. (2005) Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. Proc Natl Acad Sci USA. 102:17400–17405. Epub 2005 Nov 21.

Author response: *These references have been added to the article.*

Reviewer's report 3

Frank Eisenhaber, Institute of Molecular Pathology, Vienna, Austria

The question on how to determine the occurrence of regulatory elements in nucleic acid sequences is in the center of biomolecular sequence analysis since many decades. The literature has become large, it is not easy to oversee and to evaluate. Thus, a review in this area is appropriate.

The present revised MS of Sandve and Drablos has an acceptable style and language, the article is well structured and easy to read.

The authors wish to present their quite formalized, integrated framework (level 1 – small motif binding sites, level 2 – clusters of sites in close proximity (= modules), level 3 – combinations of modules in the regulatory region of a gene, level 4 – sets of modules in regulatory regions of sets of genes) for organizing the vast literature and for delineating the elementary recognition tasks in the prediction of regulatory elements.

From the very beginning (last paragraph in the introduction), the authors refrain from a comparison of various methods with respect to their performance. Moreover, there is no quantitative assessment in the manuscript that allows to estimate what can be expected from the group of methods described in this review in general. It is the pity reality that prediction of regulatory regions is pretty unreliable with both false-positive and false-negative prediction rivalling the number of true predictions.

The following manuscript text is merely a compilation of the variations in mathematical formulations used in the different methods in the literature. For assessing the relative merit of the various approaches, the authors do not have appropriate criteria. Although a performance comparison is difficult and gold standard test sets are not readily available, it would nevertheless give some hint on the reliability of methods and their relative accuracy. The comparative work of Bajic VB, Tan SL, Suzuki Y, Sugano S. (Promoter prediction analysis on the whole human genome. Nat Biotechnol. 2004 Nov;22(11):1467–73) is focused on a very specifc type of a regulatory region but it is at least a beginning of a large-scale performance evaluation. If the authors do not wish to get involved in such a comparative study, they should at least provide a review of published data. To a certain extent, this has been provided in an additional section in the revised version but the wording appears very polite and a quantification of performance is not provided. To emphasize the view of a practitioner, this is what matters.

Author response: *We acknowledge the concern about evaluation of methods, which is why we have included an expanded section in the revised version discussing comparison of motif discovery methods. However, we do not feel that it is currently possible to give clear recommendations on the issues considered in our survey. We have elaborated more on the reasons for this in our revised manuscript. As our focus is on the recent development of methods taking combinatorial mechanisms and additional data into consideration, the benchmark of Tompa et al. (2005) could only give limited guidance. The recent article of Bajic et al. is also very interesting, but it considers methods for promoter prediction and in particular prediction of transcription start sites* (*TSS*). *These methods are related to, but still somewhat different from the methods considered in our survey that predict locations of binding sites.*

It would be another way to assess methods by their implementation of true biological mechanisms into their formal approaches. I wonder that biological literature on transcription regulation is not considered in this review. A comprehensive survey is not indicated for this review. But for the purpose of gussing future ways out of the difficulties, one might analyze the experimental data available for a few well-studied transcription complexes and genes regulated by them. Even if a method yet fails to perform in a large-scale test, it might be a good start for further development if its mathematical/analytical formulations captures major mechanistic aspects of the biological process of recognizing regulatory sequences. Another mathematical reformulation of existing approaches will certainly not change the status of the field.

Author response: *We completely agree that it would be beneficial to have access to a good state of the art overview over the*

biological aspects of transcription regulation, from the point of view of motif discovery. However, we feel that such an overview will be outside the scope of this review, and probably more suited as a separate review paper.

The increasing availability of data from high-throuput methodologies (e.g., microarray (ChIP) data) for certain DNA-binding protein complexes will possibly change the situation for developing prediction tools in the near future.

In its present form, the review can be useful for people in the field since some part of the vast literature is organized in a reasonable way. At the same time, the review does not give guidance to the reader, which lines of prediction tool development are most promising and what conditions must be fulfilled to move the field out of its apparent stagnation.

Author response: *Our strong focus on methods using different types of data in an integrated analysis, combined with a critical attention to implementation details, should be read as a guidance to the reader.*

Acknowledgements

Finn Drabløs was supported by The National Programme for Research in Functional Genomics in Norway (FUGE) in The Research Council of Norway and by The Svanhild and Arne Must Fund for Medical Research. We want to thank Magnus Lie Hetland for helpful discussions and the referees for useful input.

References

1. Korn LJ, Queen CL, Wegman MN: **Computer analysis of nucleic acid regulatory sequences.** *Proc Natl Acad Sci U S A* 1977, **74(10):**4401-5.
2. Queen C, Wegman MN, Korn LJ: **Improvements to a program for DNA analysis: a procedure to find homologies among many sequences.** *Nucleic Acids Res* 1982, **10:**449-56.
3. Stormo GD, Schneider TD, Gold L, Ehrenfeucht A: **of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli.** *Nucleic Acids Res* 1982, **10(9):**2997-3011.
4. Staden R: **Computer methods to locate signals in nucleic acid sequences.** *Nucleic Acids Res* 1984, **12(1 Pt 2):**505-19.
5. Stormo GD: **DNA binding sites: representation and discovery.** *Bioinformatics* 2000, **16:**16-23.
6. Pavesi G, Mauri G, Pesole G: **In silico representation and discovery of transcription factor binding sites.** *Brief Bioinform* 2004, **5(3):**217-36.
7. Wasserman WW, Krivan W: **In silico identification of metazoan transcriptional regulatory regions.** *Naturwissenschaften* 2003, **90(4):**156-66.
8. Bulyk ML: **Computational prediction of transcription-factor binding site locations.** *Genome Biol* 2003, **5:**201.
9. Hannenhalli S, Levy S: **Promoter prediction in the human genome.** *Bioinformatics* 2001, **17(Suppl 1):**S90-6.
10. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z: **Assessing computational tools for the discovery of transcription factor binding sites.** *Nat Biotechnol* 2005, **23:**137-44.
11. van Driel R, Fransz PF, Verschure PJ: **The eukaryotic genome: a system regulated at different hierarchical levels.** *J Cell Sci* 2003, **116(Pt 20):**4067-75.
12. Werner T: **Models for prediction and recognition of eukaryotic promoters.** *Mamm Genome* 1999, **10(2):**168-75.
13. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA: **The evolution of transcriptional regulation in eukaryotes.** *Mol Biol Evol* 2003, **20(9):**1377-419.
14. Pedersen AG, Baldi P, Chauvin Y, Brunak S: **The biology of eukaryotic promoter prediction–areview.** *Comput Chem* 1999, **23(3–4):**191-207.
15. Kato M, Hata N, Banerjee N, Futcher B, Zhang MQ: **Identifying combinatorial regulation of transcription factors and binding motifs.** *Genome Biol* 2004, **5(8):**R56.
16. Brazma A, Jonassen I, Eidhammer I, Gilbert D: **Approaches to the automatic discovery of patterns in biosequences.** *J Comput Biol* 1998, **5(2):**279-305.
17. Brazma A, Jonassen I, Vilo J, Ukkonen E: **Pattern Discovery in Biosequences.** In *ICGI '98:Proceedings of the 4th International Colloquium on Grammatical Inference* London, UK: Springer-Verlag; 1998:257-270.
18. **Table of motif discovery tools** [http://www.ntnu.no/~drablos/motif/discovery_tools.html]
19. Berg OG, von Hippel PH: **Selection of DNA binding sites by regulatory proteins.Statistical-mechanical theory and application to operators and promoters.** *J Mol Bio l* 1987, **193(4):**723-50.
20. Benos PV, Bulyk ML, Stormo GD: **Additivity in protein-DNA interactions: how good an approximation is it?** *Nucleic Acids Res* 2002, **30(20):**4442-51.
21. Zhou Q, Liu JS: **Modeling within-motif dependence for transcription factor binding site predictions.** *Bioinformatics* 2004, **20(6):**909-16.
22. O'Flanagan RA, Paillard G, Lavery R, Sengupta AM: **Non-additivity in protein-DNA binding.** *Bioinformatics* 2005, **21(10):**2254-2263.
23. Stormo GD, Schneider TD, Gold L: **Quantitative analysis of the relationship between nucleotide sequence and functional activity.** *Nucleic Acids Res* 1986, **14(16):**6661-79.
24. Barash Y, Elidan G, Friedman N, Kaplan T: **Modeling dependencies in protein-DNA binding sites.** In *RECOMB '03: Proceedings of the seventh annual international conference on Computational molecular biology* New York, NY, USA: ACM Press; 2003:28-37.
25. Lim LP, Burge CB: **A computational analysis of sequence features involved in recognition of short introns.** *Proc Natl Acad Sci USA* 2001, **98(20):**11193-8.
26. Cawley S: **Statistical models for DNA sequencing and analysisspliceosome: motors, clocks, springs, and things. Cel l, Statistical models for DNA sequencing and analysis.** In *PhD thesis* University of California at Berkely, Berkely, CA; 2000.
27. Zhao X, Huang H, Speed TP: **Finding short DNA motifs using permuted markov models. In.** In *RECOMB '04-' Proceedings of the eighth annual international conference on Computational molecular biology* New York, NY, USA: ACM Press; 2004:68-75.
28. Ben-Gal I, Shani A, Gohr A, Grau J, Arviv S, Shmilovici A, Posch S, Grosse I: **Identification of transcription factor binding sites with variable-order Bayesian networks.** *Bioinformatics* 2005, **21(11):**1367-4803.
29. Xing EP, Jordan MI, Karp RM, Russell S: **A hierarchical bayesian markovian model for motifs in biopolymer sequences.** In *Advances in Neural Information Processing Systems Volume 16*. Edited by: Becker S, Thrun S, Obermayer K. MIT Press, Cambridge, MA; 2002.
30. Kechris KJ, van Zwet E, Bickel PJ, Eisen MB: **Detecting DNA regulatory motifs by incorporating positional trends in information content.** *Genome Biol* 2004, **5(7):**R50.
31. van Helden J, Andre B, Collado-Vides J: **Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies.** *J Mol Biol* 1998, **281(5):**827-42.
32. Jensen LJ, Knudsen S: **Automatic discovery of regulatory patterns in promoter regions based on whole cell expression data and functional annotation.** *Bioinformatics* 2000, **16(4):**326-33.
33. Sinha S, Tompa M: **A statistical method for finding transcription factor binding sites.** *Proc Int Conf Intell Syst Mol Biol* 2000, **8:**344-54.
34. Bussemaker HJ, Li H, Siggia ED: **Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis.** *Proc Natl Acad Sci USA* 2000, **97(18):**10096-100.

35. van Helden J, Rios AF, Collado-Vides J: **Discovering regulatory elements in non-coding sequences by analysis of spaced dyads.** *Nucleic Acids Res* 2000, **28(8):**1808-18.
36. Shinozaki D, Maruyama O: **A Method for the Best Model Selection for Single and Paired Motifs.** In *Genome Informatics Volume 13*. Universal Academy Press; 2002:432-433.
37. Takusagawa KT, Gifford DK: **Negative information for motif discovery.** *Pac Symp Biocomput* 2004:360-71.
38. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M: **Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.** *Nature* 2005, **434(7031):**338-45.
39. Tompa M: **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* Heidelberg, Germany; 1999:262-71.
40. Marsan L, Sagot MF: **Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification.** *J Comput Biol* 2000, **7(3–4):**345-62.
41. Pevzner PA, Sze SH: **Combinatorial approaches to finding subtle signals in DNA sequences.** *Proc Int Conf Intell Syst Mol Biol* 2000, **8:**269-78.
42. Pavesi G, Mauri G, Pesole G: **An algorithm for finding signals of unknown length in DNA sequences.** *Bioinformatics* 2001, **17(Suppl 1):**S207-14.
43. Eskin E, Pevzner PA: **Finding composite regulatory patterns in DNA sequences.** *Bioinformatics* 2002, **18(Suppl 1):**S354-63.
44. Baldwin NE, Collins RL, Langston MA, Leuze MR, Symons CT, Voy BH: **High performance computational tools for motif discovery.** *18th International Parallel and Distributed Processing Symposium (IPDPS'04) – Workshop 9* 2004:192a.
45. Li HL, Fu CJ: **A linear programming approach for identifying a consensus sequence on DNA sequences.** *Bioinformatics* 2005, **21(19):**1838-1845.
46. Blanchette M, Tompa M: **Discovery of regulatory elements by a computational method for phylogenetic footprinting.** *Genome Res* 2002, **12(5):**739-48.
47. Jensen ST, Liu XS, Liu JS, Zhou Q: **Computational Discovery of Gene Regulatory Binding Motifs: A Bayesian Perspective.** *Statist Sci* 2004, **19:**188-204.
48. Sinha S, van Nimwegen E, Siggia ED: **A probabilistic method to detect regulatory modules.** *Bioinformatics* 2003, **19(Suppl 1):**i292-301.
49. Liu X, Brutlag DL, Liu JS: **BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes.** *Pac Symp Biocomput* 2001:127-38.
50. Donaldson IJ, Chapman M, Gottgens B: **TFBScluster: a resource for the characterisation of transcriptional regulatory networks.** *Bioinformatics* 2005, **21(13):**1367-4803.
51. McCue L, Thompson W, Carmack C, Ryan MP, Liu JS, Derbyshire V, Lawrence CE: **Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes.** *Nucleic Acids Res* 2001, **29(3):**774-82.
52. Thompson W, Rouchka EC, Lawrence CE: **Gibbs Recursive Sampler: finding transcription factor binding sites.** *Nucleic Acids Res* 2003, **31(13):**3580-5.
53. Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT: **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* 1988, **203(2):**439-55.
54. Zhang Z, Gerstein M: **Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements.** *J Biol* 2003, **2(2):**11.
55. Krivan W, Wasserman WW: **A predictive model for regulatory sequences directing liver-specific transcription.** *Genome Res* 2001, **11(9):**1559-66.
56. Sharan R, Ovcharenko I, Ben-Hur A, Karp RM: **CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments.** *Bioinformatics* 2003, **19(Suppl 1):**i283-91.
57. Cora D, Herrmann C, Dieterich C, Di Cunto F, Provero P, Caselle M: **Ab initio identification of putative human transcription factor binding sites by comparative genomics.** *BMC Bioinformatics* 2005, **6:**110.
58. Wasserman WW, Fickett JW: **Identification of regulatory regions which confer muscle-specific gene expression.** *J Mol Biol* 1998, **278:**167-81.
59. Wang T, Stormo GD: **Identifying the conserved network of cis-regulatory sites of a eukaryotic genome.** *Proc Natl Acad Sci USA* 2005, **102(48):**17400-5.
60. Beiko RG, Charlebois RL: **GANN: genetic algorithm neural networks for the detection of conserved combinations of features in DNA.** *BMC Bioinformatics* 2005, **6:**36.
61. Pudimat R, Schukat-Talamazzini EG, Backofen R: **Feature Based Representation and Detection of Transcription Factor Binding Sites.** *Proceedings of the German Conference on Bioinformatics* 2004:43-52.
62. Ponomarenko JV, Ponomarenko MP, Frolov AS, Vorobyev DG, Overton GC, Kolchanov NA: **Conformational and physicochemical DNA features specific for transcription factor binding sites.** *Bioinformatics* 1999, **15(7–8):**654-68.
63. El Hassan MA, Calladine CR: **Conformational characteristics of DNA: empirical classifications and a hypothesis for the conformational behaviour of dinucleotide steps.** *Roy Soc of London Phil Tr A* 1997, **355(1722):**43-100.
64. Kel A, Kel-Margoulis O, Ivanova T, Wingender E: **ClusterScan: A Tool for Automatic Annotation of Genomic Regulatory Sequences by Searching for Composite Clusters.** *Proceedings of the German Conference on Bioinformatics* 2001:96-101.
65. Hu YJ: **Finding subtle motifs with variable gaps in unaligned DNA sequences.** *Comput Methods Programs Biomed* 2003, **70:**11-20.
66. Thompson W, Palumbo MJ, Wasserman WW, Liu JS, Lawrence CE: **Decoding human regulatory circuits.** *Genome Res* 2004, **14(10A):**1967-74.
67. Bussemaker HJ, Li H, Siggia ED: **Regulatory element detection using correlation with expression.** *Nat Genet* 2001, **27(2):**167-71.
68. GimaThakurta D, Stormo GD: **Identifying target sites for cooperatively binding factors.** *Bioinformatics* 2001, **17(7):**608-21.
69. Rebeiz M, Reeves NL, Posakony JW: **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 USA* 2002, **99(15):**9888-93.
70. Aerts S, Van Loo P, Thijs G, Moreau Y, De Moor B: **Computational detection of cis -regulatory modules.** *Bioinformatics* 2003, **19(Suppl 2):**II5-II14.
71. Bailey TL, Noble WS: **Searching for statistically significant regulatory modules.** *Bioinformatics* 2003, **19(Suppl 2):**II16-II25.
72. Frith MC, Hansen U, Weng Z: **Detection of cis-element clusters in higher eukaryotic DNA.** *Bioinformatics* 2001, **17(10):**878-89.
73. Xing EP, Wu W, Jordan MI, Karp RM: **Logos: a modular bayesian model for de novo motif detection.** *J Bioinform Comput Biol* 2004, **2:**127-54.
74. Gupta M, Liu JS: **De novo cis-regulatory module elicitation for eukaryotic genomes.** *Proc Natl Acad Sci USA* 2005, **102(20):**7079-84.
75. Wagner A: **Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes.** *Bioinformatics* 1999, **15(10):**776-84.
76. Frech K, Werner T: **Specific modelling of regulatory units in DNA sequences.** *Pac Symp Biocomput* 1997:151-62.
77. Scherf M, Klingenhoff A, Werner T: **Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach.** *J Mol Biol* 2000, **297(3):**599-606.
78. Brazma A, Vilo J, Ukkonen E, Valtonen K: **Data mining for regulatory elements in yeast genome.** *Proc Int Conf Intell Syst Mol Biol* 1997, **5:**65-74.
79. Policriti A, Vitacolonna N, Morgante M, Zuccolo A: **Structured motifs search.** In *RECOMB '04: Proceedings of the eighth annual international conference on Computational molecular biology* New York, NY, USA: ACM Press; 2004:133-139.
80. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB: **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome.** *Proc Natl Acad Sci USA* 2002, **99(2):**757-62.
81. Segal E, Barash Y, Simon I, Friedman N, Koller D: **From promoter sequence to expression: a probabilistic framework.** In

RECOMB '02: Proceedings of the sixth annual international conference on Computational biology New York, NY, USA: ACM Press; 2002:263-272.
82. Aerts S, Van Loo P, Moreau Y, De Moor B: **A genetic algorithm for the detection of new cis-regulatory modules in sets of coregulated genes.** *Bioinformatics* 2004, **20(12):**1974-6.
83. Klingenhoff A, Freeh K, Quandt K, Werner T: **Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity.** *Bioinformatics* 1999, **15(3):**180-6.
84. Johansson O, Alkema W, Wasserman WW, Lagergren J: **Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm.** *Bioinformatics* 2003, **19(Suppl 1):**i169-76.
85. Mahony S, Hendrix D, Golden A, Smith TJ, Rokhsar DS: **Transcription factor binding site identification using the self-organizing map.** *Bioinformatics* 2005, **21(9):**1807-1814.
86. Workman CT, Stormo GD: **a method for discovering transcription factor binding sites with improved specificity.** *Pac Symp Biocomput* 2000:467-78.
87. Liu XS, Brutlag DL, Liu JS: **An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments.** *Nat Biotechnol* 2002, **20(8):**835-9.
88. Wang T, Stormo GD: **Combining phylogenetic data with co-regulated genes to identify regulatory motifs.** *Bioinformatics* 2003, **19(18):**2369-80.
89. Conlon EM, Liu XS, Lieb JD, Liu JS: **Integrating regulatory motif discovery and genome-wide expression analysis.** *Proc Natl Acad Sci USA* 2003, **100(6):**3339-44.
90. Frith MC, Fu Y, Yu L, Chen JF, Hansen U, Weng Z: **Detection of functional DNA motifs via statistical over-representation.** *Nucleic Acids Res* 2004, **32(4):**1372-81.
91. Caselle M, Di Cunto F, Provero P: **Correlating overrepresented upstream motifs to gene expression: a computational approach to regulatory element discovery in eukaryotes.** *BMC Bioinformatics* 2002, **3:**7.
92. Cora D, Di Cunto F, Provero P, Silengo L, Caselle M: **Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrepresented upstream motifs.** *BMC Bioinformatics* 2004, **5:**57.
93. Curran MD, Liu H, Long F, Ge N: **Statistical methods for joint data mining of gene expression and DNA sequence database.** *SIGKDD Explor Newsl* 2003, **5(2):**122-129.
94. Segal E, Yelensky R, Koller D: **Genome-wide discovery of transcriptional modules from DNA sequence and gene expression.** *Bioinformatics* 2003, **19(Suppl 1):**i273-82.
95. Hong P, Liu X, Zhou Q, Lu X, Liu JS, Wong WH: **A boosting approach for motif modeling using ChIP-chip data.** *Bioinformatics* 2005, **21(11):**2636-2643.
96. Gupta M, Liu JS: **Discovery of Conserved Sequence Patterns Using a Stochastic Dictionary Model.** *Journal of the American Statistical Association* 2003, **98:**55-66.
97. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.** *Science* 1993, **262(5131):**208-14.
98. Bailey TL, Elkan C: **The value of prior knowledge in discovering motifs with.** *Proc Int Conf Intell Syst Mol Biol* 1995, **3:**21-9.
99. Jonassen I: **Efficient discovery of conserved patterns using a pattern graph.** *Comput Applbiosci* 1997, **13(5):**509-22.
100. Rigoutsos I, Floratos A: **Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm.** *Bioinformatics* 1998, **14:**55-67.
101. Rustici G, Mata J, Kivinen K, Lio P, Penkett CJ, Burns G, Hayles J, Brazma A, Nurse P, Bahler J: **Periodic gene expression program of the fission yeast cell cycle.** *Nat Genet* 2004, **36(8):**809-17.
102. Birnbaum K, Benfey PN, Shasha DE: **cis element/transcription factor analysis (cis/TF): a method for discovering transcription factor/cis element relationships.** *Genome Res* 2001, **11(9):**1567-73.
103. Zhu Z, Pilpel Y, Church GM: **Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm.** *J Mol Biol* 2002, **318:**71-81.
104. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA: **Genome-wide location and function of DNA binding proteins.** *Science* 2000, **290(5500):**2306-9.
105. Buck MJ, Lieb JD: **ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments.** *Genomics* 2004, **83(3):**349-60.
106. Mironov AA, Koonin EV, Roytberg MA, Gelfand MS: **Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes.** *Nucleic Acids Res* 1999, **27(14):**2981-9.
107. Qin ZS, McCue LA, Thompson W, Mayerhofer L, Lawrence CE, Liu JS: **Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites.** *Nat Biotechnol* 2003, **21(4):**435-9.
108. McGuire AM, Hughes JD, Church GM: **Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes.** *Genome Res* 2000, **10(6):**744-57.
109. Thijs G, Marchal K, Lescot M, Rombauts S, De Moor B, Rouze P, Moreau Y: **A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes.** *J Comput Biol* 2002, **9(2):**447-64.
110. Park PJ, Butte AJ, Kohane IS: **Comparing expression profiles of genes with similar promoter regions.** *Bioinformatics* 2002, **18(12):**1576-84.
111. Holmes I, Bruno WJ: **Finding regulatory elements using joint likelihoods for sequence and expression profile data.** *Proc Int Conf Intell Syst Mol Biol* 2000, **8:**202-10.
112. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK: **Computational discovery of gene modules and regulatory networks.** *Nat Biotechnol* 2003, **21(11):**1337-42.
113. Evans PA, Smith AD: **Toward optimal motif enumeration.** In *Proceedings of Workshop on Algorithms and Data Structures (WADS 2003) Volume 2751*. Springer-Verlag; 2003:47-58.
114. Lawrence CE, Reilly AA: **An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences.** *Proteins* 1990, **7:**41-51.
115. Sinha S, Blanchette M, Tompa M: **PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences.** *BMC Bioinformatics* 2004, **5:**170.
116. Prakash A, Blanchette M, Sinha S, Tompa M: **Motif discovery in heterogeneous sequence data.** *Pac Symp Biocomput* 2004:348-59.
117. Ao W, Gaudet J, Kent WJ, Muttumu S, Mango SE: **Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR.** *Science* 2004, **305(5691):**1743-6.
118. Neuwald AF, Liu JS, Lawrence CE: **Gibbs motif sampling: detection of bacterial outer membrane protein repeats.** *Protein Sci* 1995, **4(8):**1618-32.
119. Zhou Q, Wong WH: **CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling.** *Proc Natl Acad Sci USA* 2004, **101(33):**12114-9.
120. Grad YH, Roth FP, Halfon MS, Church GM: **Prediction of similarly-acting cis-regulatory modules by subsequence profiling and comparative genomics in D. melanogaster and D. pseudoobscura.** *Bioinformatics* 2004, **20(16):**2738-2750.
121. Hart RK, Royyuru AK, Stolovitzky G, Califano A: **Systematic and fully automated identification of protein sequence patterns.** *J Comput Biol* 2000, **7(3–4):**585-600.
122. Favorov AV, Gelfand MS, Gerasimova AV, Ravcheev DA, Mironov AA, Makeev VJ: **A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length.** *Bioinformatics* 2005, **21(10):**2240-2245.
123. Marsan L, Sagot MF: **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* New York, NY, USA: ACM Press; 2000:210-219.
124. Roth FP, Hughes JD, Estep PW, Church GM: **Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation.** *Nat Biotechnol* 1998, **16(10):**939-45.

Paper 2 T. Lin, P. Ray, G. K. Sandve, S. Uguroglu and E. P. Xing. Baycis: a bayesian hierarchical HMM for cis-regulatory module decoding in metazoan genomes. In *Proceedings of the Twelfth Annual International Conference on Research in Computational Molecular Biology (RECOMB)*. 2008;66–81.

Paper 3 G. K. Sandve, O. Abul and F. Drabløs. Compo: composite motif discovery using discrete models. *BMC Bioinformatics* (submitted).

Are not included due to copyright

BMC Bioinformatics

Research article

Improved benchmarks for computational motif discovery

Geir Kjetil Sandve*[1], Osman Abul[2], Vegard Walseng[1] and Finn Drabløs[3]

Address: [1]Department of Computer and Information Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, [2]Department of Computer Engineering, TOBB University of Economics and Technology, Ankara, Turkey and [3]Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

Email: Geir Kjetil Sandve* - sandve@ntnu.no; Osman Abul - osmanabul@etu.edu.tr; Vegard Walseng - walseng@stud.ntnu.no; Finn Drabløs - finn.drablos@ntnu.no

* Corresponding author

Published: 8 June 2007

BMC Bioinformatics 2007, **8**:193 doi:10.1186/1471-2105-8-193

Received: 9 November 2006
Accepted: 8 June 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/193

Abstract

Background: An important step in annotation of sequenced genomes is the identification of transcription factor binding sites. More than a hundred different computational methods have been proposed, and it is difficult to make an informed choice. Therefore, robust assessment of motif discovery methods becomes important, both for validation of existing tools and for identification of promising directions for future research.

Results: We use a machine learning perspective to analyze collections of transcription factors with known binding sites. Algorithms are presented for finding position weight matrices (PWMs), IUPAC-type motifs and mismatch motifs with optimal discrimination of binding sites from remaining sequence. We show that for many data sets in a recently proposed benchmark suite for motif discovery, none of the common motif models can accurately discriminate the binding sites from remaining sequence. This may obscure the distinction between the potential performance of the motif discovery tool itself versus the intrinsic complexity of the problem we are trying to solve. Synthetic data sets may avoid this problem, but we show on some previously proposed benchmarks that there may be a strong bias towards a presupposed motif model. We also propose a new approach to benchmark data set construction. This approach is based on collections of binding site fragments that are ranked according to the optimal level of discrimination achieved with our algorithms. This allows us to select subsets with specific properties. We present one benchmark suite with data sets that allow good discrimination between positive and negative instances with the common motif models. These data sets are suitable for evaluating algorithms for motif discovery that rely on these models. We present another benchmark suite where PWM, IUPAC and mismatch motif models are not able to discriminate reliably between positive and negative instances. This suite could be used for evaluating more powerful motif models.

Conclusion: Our improved benchmark suites have been designed to differentiate between the performance of motif discovery algorithms and the power of motif models. We provide a web server where users can download our benchmark suites, submit predictions and visualize scores on the benchmarks.

Background

Computational discovery of motifs in biological sequences is an important challenge. It has in recent years attracted much research interest, resulting in more than a hundred different tools for motif discovery [1]. A motif discovery method has three important elements: a motif model that can capture the similarities of a diverse set of binding sites for the same transcription factor, an objective function defining the ranking of potential motifs and a search strategy for parameterisation of the motif model. The first two elements can be given an abstract representation, but should probably be designed to utilize and enhance biologically relevant information. The most commonly used motif models are position weight matrices (PWMs) [2,3], mismatch strings (MMs) [4,5] (consensus string allowing some mismatches) and IUPAC strings (IUPACs) [6,7] (consensus string with degenerate symbols).

Due to the large number of available tools, robust assessment of motif discovery methods becomes important, not only for validation of existing tools, but also for pointing out the most promising directions for future research in the field. A major difficulty is our limited knowledge about the biological mechanisms of gene regulation at a detailed level. Although collections of experimentally determined transcription factor binding sites (TFBS) are available, these collections do have inaccuracies and biases. This has been shown e.g. by Fogel *et al*. in their analysis of the TRANSFAC database [8], and by Bergman *et al*. in their study of *Drosophila* gene regulation [9].

A recent article by Tompa *et al*. [10] used experimental collections of TFBS to benchmark a large number of motif discovery tools. This was an important and timely contribution to the field, and it gave good guidance to biologists regarding the level of performance that can be expected with current tools. However, it gave less guidance to the motif discovery field itself. That is, although the study clearly showed a lack of correspondence between *in silico* predictions and *in vivo* experiments, the authors were not able to give much guidance with respect to how we can identify the most promising motif discovery approaches. Furthermore, due to the inherent complexities of the data set, it was hard to distinguish between clever preprocessing and method parameterization done by the expert user on one hand, and the performance of the motif discovery algorithms themselves on the other hand. We note that one of the few clear differences that can be spotted from the generally low performance values – the relatively high score of Weeder – is in the paper partly attributed to judicious choices regarding when to make predictions, while nothing is concluded regarding any superiority of the algorithm itself.

Synthetic data sets may avoid many of these problems. By ensuring that high motif discovery performance is at least theoretically possible, the performance differences between tools may be clearer and more consistent, thus giving more guidance to developers. On the other hand, the coupling may be too loose between the synthetic data sets and the biological reality, introducing an artificial bias. This bias may favor specific classes of tools in a way that lacks biological relevance.

The performance of any motif discovery algorithm can be measured by how well it is able to identify true binding sites in a data set. However, the optimal performance that can be achieved will depend upon the complexity of the data set itself. Here we use a machine learning perspective to analyse collections of TFBS with known binding site locations, in order to estimate an upper bound to the motif discovery performance that can be expected for a given data set. We formulate the problem as a binary classification problem where all sequence windows corresponding to binding sites are termed positive samples, and all other windows are negative samples. Algorithms are given for finding MM, IUPAC and PWM models with optimal discrimination between positive and negative samples.

We use this approach to analyze the experimentally based benchmark data sets used in the recent assessment of motif discovery tools by Tompa *et al*. We also analyze some synthetic benchmark data sets proposed by Pevzner *et al*. [11] and compare the results to those for the experimental collections. Finally we show how the same approach can be used to construct benchmark data sets that combine advantageous properties of both experimentally based and synthetic benchmarks. Data sets are ranked according to the best possible discrimination score as computed by our discrimination approach, and this ranking is used to select subsets with specific properties. We present one benchmark suite with data sets that allow good discrimination between positive and negative instances. This suite, the algorithm benchmark, is useful for evaluating algorithms for motif discovery that rely on the common motif models, as we know that it should be possible to achieve good discrimination with these models. We present another benchmark suite for evaluating motif models, the model benchmark. The data sets in this suite are selected so that none of the common motif models are able to discriminate between positive and negative instances in a reliable way. This suite is useful for evaluating novel and more expressive motif models, as we know that it is not possible to achieve good discrimination with the standard models.

Results and discussion

We have used the discrimination algorithms described in Methods to analyze motif occurrences in both experimentally based and synthetic benchmark data sets. We present an alternative way of constructing benchmark data sets that uses the discrimination algorithms as a key component.

Discrimination algorithms

We view a collection of binding sites in a machine learning perspective, where the goal is to find motifs that achieve optimal discrimination of binding sites (at known positions) from remaining sequence. Binding sites are assumed to be of equal length, which may require some alignment and truncation of related sites. Sequence windows corresponding to binding sites are considered positive samples, and all other sequence windows are considered negative samples. For each of the three common motif models, MM, IUPAC and PWM, algorithms have been developed that find the motif that best discriminates between the known positive and negative samples. Discrimination is here defined as finding the single motif that best separates true from false sites, and the discrimination score is the nucleotide-level correlation coefficient (nCC) for this separation, using Formula 1 according to Tompa *et al*. [10]. Details on the problem definition and the individual algorithms are given in Methods and in supplementary material (see additional file 1: IUPAC_details.pdf).

Analysis of existing benchmark data

We used our discrimination approach to analyze the benchmark suite of Tompa *et al*. For each data set we computed the best possible discrimination between binding sites and remaining sequence using the three motif models. As the binding sites are unaligned and of different length within each individual data set, we had to align and possibly truncate each set of binding sites as a preprocessing step using a gapless alignment [12]. The resulting set of consensus-aligned, equal-length binding site fragments is representative of what can be discovered by standard motif discovery methods.

Figure 1 shows to what extent it is possible to discriminate the set of binding sites from remaining sequence in each of the 50 data sets with a given motif model. We see that this varies a lot, some data sets allow a discrimination score (nCC) of more than 0.8, while other data sets do not allow discrimination score above 0.2 with any of the models. These results are from the "real" data sets from Tompa *et al*. (actual promoter regions), but the scores were similar in the "generic" (binding sites implanted in randomly selected promoter regions) and "Markov" (binding sites implanted in Markov model backgrounds) data sets (see Figure 2).

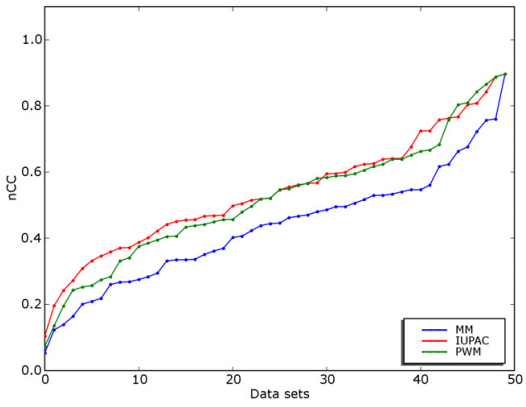


Figure 1
Discrimination on data sets by Tompa et al. Nucleotide-level CC-score for discrimination between binding sites and remaining sequence on data sets from Tompa *et al*. Data sets (x-axis) are sorted individually for each model in order of increasing nCC, making it easier to compare the overall distributions of discrimination scores.

The IUPAC model had the highest average score, followed by PWM and MM. The score differences between models were statistically significant using paired t-test with 95% confidence level. However, the difference between IUPAC and PWM was very small, and probably not of practical relevance. On the other hand, the score for MM was considerably lower than the others.

Although PWMs are more expressive than IUPAC models, IUPAC scored slightly higher in our tests. PWMs were restricted to either contain log-likelihoods based on aligned binding sites, or to contain log-odds values taking negative data into consideration through a Markov model. All established PWM-based methods use log-likelihood or log-odds matrices, we therefore see this restriction as a reasonable choice. We tried different pseudocount values and backgrounds with different Markov order, and chose the values that gave best overall score. On the other hand, the algorithms for the IUPAC and mismatch models take negative data directly into consideration, and this leads to slightly better classification performance under certain conditions.

Although the discrimination algorithms return optimal discrimination results on the data they are given, the initial alignment of binding sites in our pre-processing step may be sub-optimal. Multiple alignment algorithms are heuristic, and cannot guarantee optimal solutions. Also, the criteria for optimality of an alignment may not ensure a motif representation that is optimal for classification. As

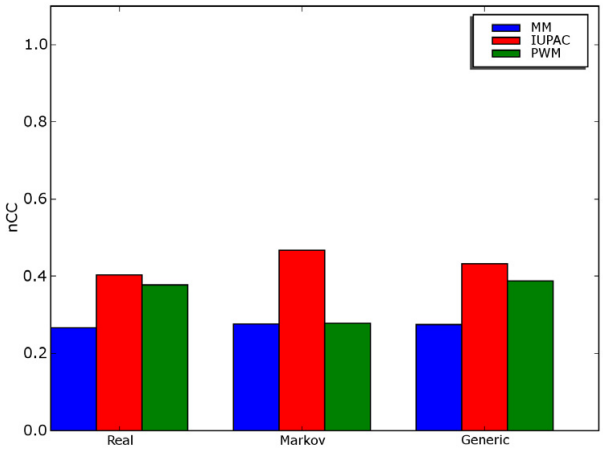


Figure 2
Discrimination on different data set versions by Tompa et al. Nucleotide-level CC-score for discrimination between binding sites and remaining sequence for the three motif models on real, generic and Markov versions of data sets.

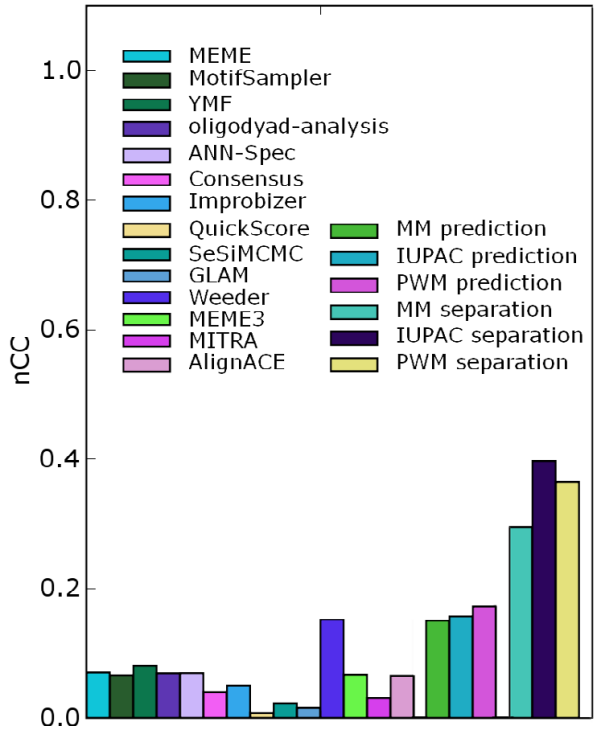


Figure 3
Motif discovery scores from Tompa et al. nCC-scores of 14 motif discovery methods given in the Tompa assessment, compared to prediction and discrimination scores with the three main motif models.

the benchmarked motif discovery methods do not depend on this initial alignment, they may in some cases achieve a somewhat higher nCC-score than what we estimate in the discrimination case (if they can find a better alignment). However, from our experience this is a relatively rare situation, and heuristic ungapped alignment was in general found to perform well on the data sets analyzed here.

Cross-validation performance

Averaged prediction scores for the three motif models in a leave-one-out cross-validation experiment on the benchmark data sets of Tompa *et al*. is given together with discrimination and motif discovery scores in Figure 3. We counted the sum of TP, TN, FP and FN for the test sets across all folds, and calculated the nCC from these accumulated numbers.

As expected, for all models the scores are much lower for cross-validation based prediction than for discrimination. With nCC-scores below 0.2, it shows that even when most binding sites for a TF are known, it is still difficult to predict the location of unseen related binding sites (i.e. it is difficult to generalize from training set to independent test set). Using some strategy to avoid overfitting, e.g. adding regularization terms, could improve the prediction performance somewhat. Still, this means that even if better objective functions [13] could bridge the gap between unsupervised and supervised motif discovery, it would only amount to a limited increase in prediction accuracy on the Tompa benchmark suite. Representation of the sequence similarity between related binding sites seems to be a strong limiting factor. We also see that the IUPAC scores are lower than PWM scores in the cross-validation, confirming that the high IUPAC scores for the discrimination case were partly due to overfitting. Still, the difference in prediction performance between the motif models is very low. Our results thus indicate that the choice of motif model should not be a major limiting factor on motif discovery performance on the benchmark suite of Tompa *et al*. This fits well with the observation that Weeder, which internally uses the simple mismatch model during motif discovery, is able to outperform the many PWM-based methods on this benchmark.

Comparison of motif discovery methods

Figure 3 also shows the scores of different *de novo* motif discovery methods on the benchmark suite of Tompa *et al*., in addition to the average discrimination and prediction scores for each of the three motif models. Although the limited possibility for discrimination between binding sites and remaining sequence puts an upper bound on motif discovery performance on the data sets, the bound is still clearly above the actual scores of these *de novo* motif

discovery methods. The discrimination score suggests that motif discovery could be particularly difficult on many of these data sets. We therefore looked at how the maximum score across all motif discovery methods reported in Tompa *et al.* correlated with the discrimination scores on the different data sets. The scatter plot in figure 4 shows that the discrimination score generally represents an upper bound on the motif discovery score, with maximum motif discovery score for data sets typically distributed between zero and the bound given by discrimination score. Only in rare cases may the bound be exceeded due to suboptimal alignments, as already discussed. Most of the motif discovery score values are well below the estimated discrimination score, even though the motif discovery scores we are looking at are maximums over the 14 methods considered in Tompa *et al.*

We also looked at how the total score of a typical motif discovery method would change if data sets were removed according to the discrimination score (Figure 5). We used MEME as example, as it is a well-known method with reasonable performance in the assessment by Tompa *et al.* If only the 13 data sets with lowest discrimination score had been included in the benchmark suite, the nCC-score for MEME would have been just 0.004, compared to a nCC-score of 0.33 if only the 6 data sets with highest discrimination score were used. The nCC-score for MEME on the full benchmark suite was 0.07. We also wanted to explore the remark by Tompa *et al.* that one reason for the good performance of Weeder in the assessment was that the Weeder group was conservative about making predictions. The possible level of discrimination is of course only one of several factors that could influence such a decision, but we wanted to see whether canceling predictions based on discrimination scores alone could have increased the score of MEME on this benchmark suite. We found that the total score of MEME could indeed have been increased slightly by not making any predictions on the data sets with low discrimination score. If no predictions were made on the 14 data sets with lowest discrimination scores, the nCC-score of MEME on the full benchmark suite would have increased by 30%, from 0.076 to 0.099. Actually, because of the generally low performance, MEME would have gotten higher total scores in the assessment (when judged by nCC-score) even if they had submitted blank predictions on all but the 6 data sets with highest discrimination scores.

Analysis of synthetic benchmark data

Synthetic benchmark data sets avoid many of the problems associated with binding site collections, as the precise locations of synthetic binding sites are known and consistent with the location of sequence consensus. Furthermore, the level of discrimination that is possible to achieve with a given motif model can be controlled.

The problem with synthetic benchmark data is that the generation of synthetic binding sites must necessarily presuppose a model of sequence variability between related sites, for example in the way instances of a base consensus sequence are "mutated" before being implanted in the benchmark sequences. As different motif discovery methods rely on different models of sequence conservation,

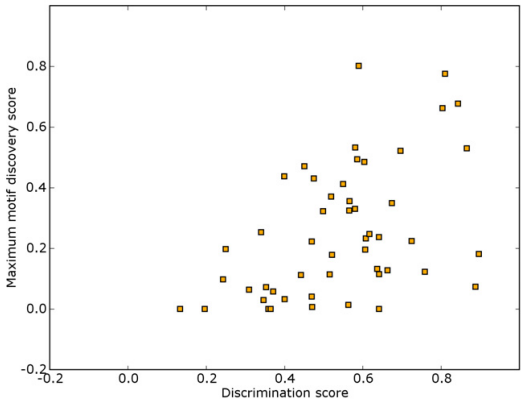


Figure 4
Motif discovery versus discrimination. Scatter plot of maximum motif discovery score versus discrimination score for the 50 data sets in the suite by Tompa *et al.*

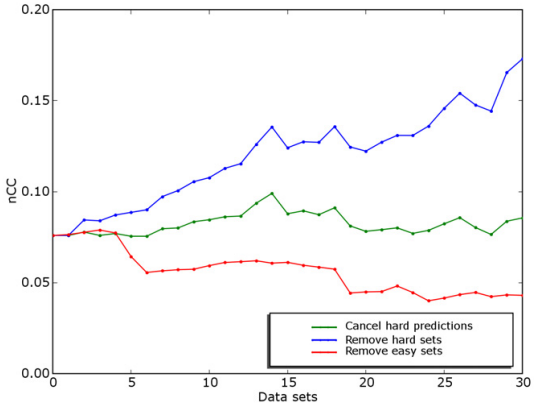


Figure 5
MEME scores after removals or erasures. Total MEME score if the data sets with highest or lowest discrimination scores, respectively, had been incrementally removed from the Tompa benchmark, as well as total MEME score if predictions on the data sets with lowest discrimination scores had been incrementally erased.

this will incur a bias towards methods using models similar to the one used when generating data sets. Synthetic benchmark data sets may therefore be suitable for comparing motif discovery methods using the same motif model, but will not give a fair comparison between methods using different motif models.

Pevnzer and Sze [11] proposed the Challenge Problem for motif discovery. A data set is constructed by implanting one motif instance in each of 20 sequences, 600 bp long. In the (15,4)-FM version (fixed number of mutations), each motif instance is made by mutating 4 random positions of a 15 bp motif consensus. In the (15,4)-VM version (variable number of mutations), each position of the motif consensus is mutated with a probability of 4/15 when forming a motif instance. Both versions assume that all positions are equally likely to be mutated, and that every nucleotide is equally likely to be the result of a mutation. These are the same assumptions as in the mismatch model. A slight variation to the Challenge Problem is proposed in Styczynski *et al*. [14], where experiments are done on data sets with motif instances in only 15 out of 20 sequences. Figure 6 shows the discrimination scores of the three common motif models, averaged over 10 data sets of 20 sequences randomly constructed according to the three variants of the Challenge Problem. Contrary to the results on annotated binding site collections, the MM model gets very competitive discrimination scores on the Challenge Problem data sets, only slightly lower than PWM scores. The IUPAC model, which had the highest average discrimination score on the data sets from Tompa *et al*., gets the lowest score on the synthetic data sets. The IUPAC model is the model that most clearly relies on asymmetries in positional conservation and skewed positional nucleotide distributions, properties not present in these synthetic data sets, although they are assumed to be biologically relevant. Both the high empirical scores of the mismatch model, and the low scores of the IUPAC model, support the intuition that synthetic data sets may introduce a bias towards a presupposed model.

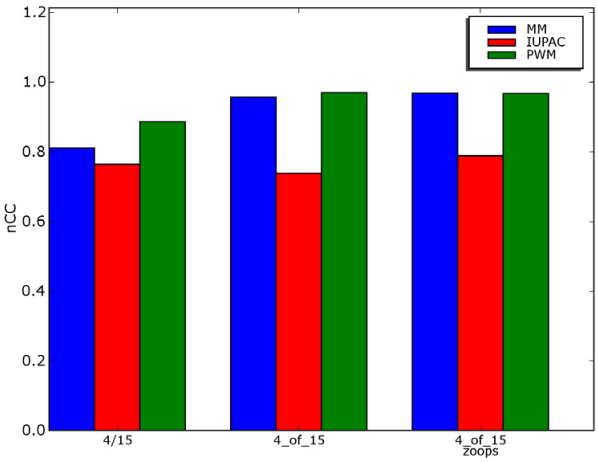


Figure 6
Discrimination on synthetic data sets. Discrimination nCC-scores for motif models on three variants of synthetic data sets: variable mutations (4/15), fixed mutations (4_of_15) and fixed mutations with instances in 75% of the sequences (4_of_15 zoops). For each variant, the scores of each model are averaged over 10 randomly generated data sets.

Generation of improved benchmark data

Based on our analysis of existing benchmark data we propose a new strategy for the generation of benchmark suites. Details are given in Methods. Basically binding site fragments corresponding to known binding sites were extracted from a suitable database (TRANSFAC) and represented either as real sequences (i.e. binding sites in their original genomic context) or Markov sequences (binding sites implanted in sequences generated with a third order Markov model). Figure 7 shows the distribution of binding sites. The best sequence-based discrimination between binding sites and remaining sequence was computed, as shown in Figure 8. Based on the discrimination score two subsets were generated, an algorithm benchmark suite and a model benchmark suite.

The algorithm benchmark suite

For our algorithm benchmark suite we selected all data sets with discrimination score higher than 0.79 for the real version and higher than 0.87 for the Markov version, giving 50 data sets of each version. Figure 9 compares the distribution of discrimination scores for this suite to the suite by Tompa *et al*., showing that the binding sites are standing out from background much more clearly in our algorithm benchmark suite.

This gives a benchmark suite where we know that it is possible to achieve good discrimination with standard motif models. This suite will therefore mainly evaluate the performance of the algorithms for motif discovery, as lack of performance has to be caused by failure to find optimal motifs, and not the motif model itself.

The model benchmark suite

The field would also gain from more powerful motif models that can better capture the variability between binding sites and discriminate these from background. This will be even more relevant as more examples of related binding sites become available.

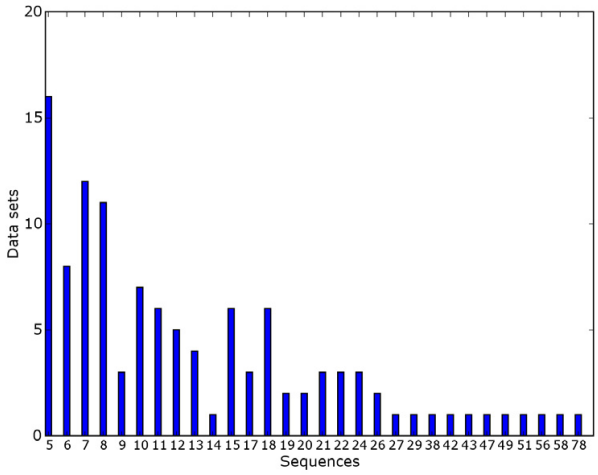


Figure 7
Sequences per data set. Distribution of number of sequences per data set.

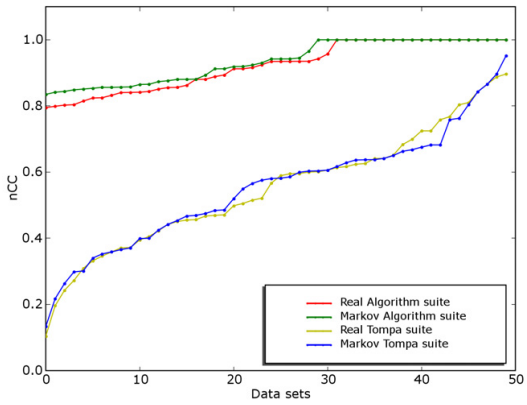


Figure 9
Discrimination on algorithm benchmark suite. Nucleotide-level CC-score for discrimination between binding sites and remaining sequence. Results are given for our algorithm benchmark suite and the suite by Tompa *et al.*, for both real and Markov versions. For each data set, the highest discrimination score achieved by any of the three motif models is selected. The distribution of scores are in sorted order for all versions independently.

For benchmarking of novel powerful motif models, we propose a model benchmark suite with binding sites that are hard to discriminate from background. The construction was similar to the preceding suite, except that for this suite data sets were selected that only allow a low level of discrimination with the common motif models. As powerful models typically require the estimation of more parameters, we also filtered out data sets with few binding sites. We selected 25 data sets with at least 18 binding sites in each data set, and with discrimination score below 0.72 for the three common motif models. Figure 10 shows the distribution of the number of binding sites and the maximum discrimination score with common models for each data set in the model benchmark suite. Table 1 shows the aggregated results in comparison to algorithm benchmark suites. As more experimentally determined binding sites become available in the future, the same methodology could give benchmark suites with a larger number of binding sites per data set, and even lower maximum discrimination scores when using the common models.

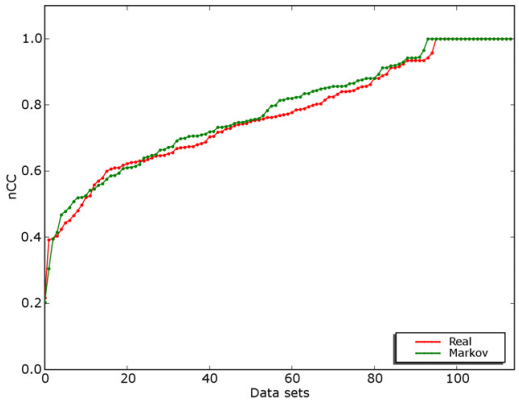


Figure 8
Discrimination on all TRANSFAC-based data sets. Nucleotide-level CC-score for discrimination between binding sites and remaining sequence on real and Markov version of TRANSFAC-based data sets. For each data set, the highest discrimination score achieved by any of the three motif models is selected. The distribution of scores are in sorted order for real and Markov versions independently.

For several data sets, some of the substrings marked as binding site also had an exact unannotated duplicate in another sequence. This means that without working with longer motif length, or operating with a motif context based on flanking sequence, it is not possible to achieve perfect discrimination with any model. The distribution of maximum discrimination scores possible with any model without taking such measures, as well as the maximum discrimination possible with the currently common motif models, is given in Figure 11.

Examples of benchmark runs

We ran MEME and Weeder on our proposed benchmark suite to indicate the level of motif discovery performance

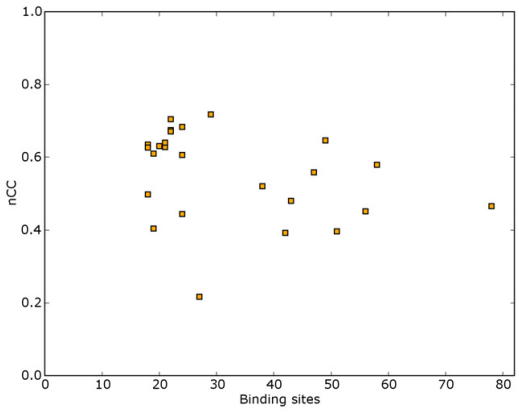


Figure 10
Discrimination score and number of binding sites. Distribution of discrimination scores (nCC) and number of binding sites for each of the 25 data sets in the model benchmark suite.

that can be expected. Table 2 compares the scores of MEME and Weeder with the discrimination scores of the PWM model. As expected, the *de novo* motif discovery scores are much lower than the upper bound given by the discrimination score. Note that all motif discovery results given on our benchmark suites have been achieved with default parameters. Slightly higher scores might be achieved by tweaking of parameters and clever post-processing of results.

The average score of MEME is higher on the real Algorithm suite than on the remaining real data sets. For Weeder this difference was less clear. While MEME achieves slightly higher scores on Markov version compared to real version of Algorithm suite, Weeder performs better on the real version. This might possibly be reflecting the different approaches to estimation of background distribution in MEME and Weeder.

Although the performance of both MEME and Weeder is better than random even with default parameters on real sequences, the performance is still much lower than the bounds given by the discrimination scores, leaving much room for improvement in the development of objective functions and search heuristics for motif discovery.

Conclusion

We have developed discrimination algorithms for the common motif models and used these algorithms both for analyzing an existing benchmark suite and for constructing new benchmark suites. The work has highlighted several important points:

• Considering discrimination of known binding sites from background separates the limitations of motif models from the limitations of objective functions and search heuristics. Discrimination algorithms for common motif models may be used to evaluate properties of data sets, for instance in a filtering step when constructing benchmark data sets.

• Motif discovery is very difficult on the data sets used in the recent benchmark of Tompa *et al*. Algorithms reveal large difficulties even with the basic task of discriminating a set of known binding sites from remaining sequence.

• Improved benchmark data sets with controlled properties can be constructed from motif databases, e.g. TRANSFAC matrix alignments, using discrimination algorithms for filtering. Using this approach, we propose one benchmark suite for evaluating the motif discovery process itself with current models, and another benchmark suite with data sets that could profit from more expressive motif models.

Our main focus has been on the level of discrimination that is possible for a given data set, and we have used the maximum score across the three models to avoid introducing a bias towards a specific model during the evaluation and filtering of benchmark data sets. Still, we have observed some consistent differences between the discrimination power of the common models: The IUPAC model achieves the highest level of discrimination, slightly above the PWM model, with the mismatch model at a clearly lower level. On the other hand, synthetic benchmark data sets rely on a chosen computational method for generating variability among implanted bind-

Table 1: Discrimination scores on model benchmark suite. Average nCC-scores of three motif models on our proposed model benchmark suites; real and Markov algorithm suite, as well as real model suite.

	Algorithm suite (Real)	Algorithm suite (Markov)	Model suite (Real)
PWMs	0.89	0.90	0.48
IUPACs	0.87	0.87	0.50
MMs	0.67	0.64	0.33

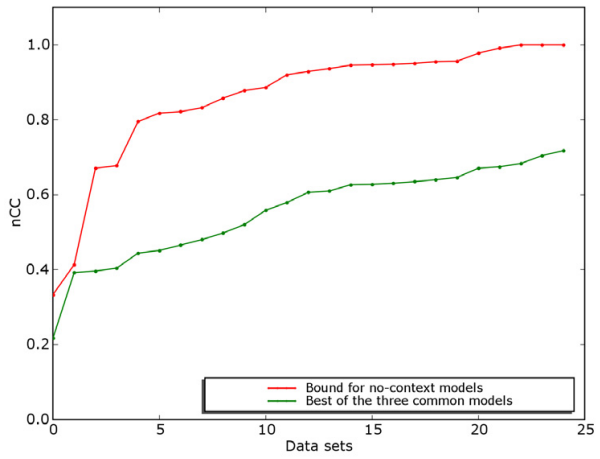


Figure 11
Discrimination score on model benchmark suite. Distribution of discrimination scores (nCC) for the 25 data sets in the model benchmark suite. One curve shows the best score of the three common motif models on the data set, while the other curve shows the score possible with a more expressive model that still do not consider the context of binding sites.

ing sites. As expected, the motif models that are more compatible with the generation model achieved better discrimination scores on three versions of synthetic data sets that were considered.

A main line of future work would be to increase the size and quality of benchmark data sets by using our proposed methodology on additional binding site collections. Also, as time goes, more data of higher quality will be available in the TRANSFAC database used in this work as well as in other similar databases. A different line of research would be to use a supervised learning approach as a first step in exploring novel and more expressive motif models. After the power of a new motif model has been determined by its discrimination scores on training sets, and its generalization ability has been determined by its prediction scores on independent test sets, the more complex task of developing efficent methods for *de novo* discovery could be commenced. Supervised learning algorithms could be developed for entirely new models, or for exploring already proposed expressive models such as HMDM [15], Bayesian nets [16,17], Markov-model motifs [18,19], dinucleotide matrices [20,21] and SPSP [22].

Methods

Motif models

The most common models of motifs in DNA sequences are PWM, IUPAC codes and mismatch strings. These are considered as three different hypothesis spaces in our work. Deciding on the hypothesis space is central to machine learning [23]. A good hypothesis space for a domain should be as small as possible while still containing a good hypothesis. The main motivation in this work is to find the best hypothesis in the respective hypothesis space of motif models. We have developed exhaustive search algorithms to avoid any search bias. Since for large model sizes exploring the whole search space becomes impractical, the algorithms developed are optimized as much as possible so as to scale well for moderate sizes.

Problem formulation

We assume that a number of upstream DNA sequences with binding site locations are given. The locations are positive examples while other oligos with the same length in the same sequence set form the negative examples (Figure 12).

Let E be a set of N TFs, *i.e.*, $E = \{TF_1, TF_2,...,TF_N\}$. Associated with each TF is its binding site length $k : E \rightarrow$, usually ranging between 6 and 20 and assumed to be known. The input space for TF_i is $X(TF_i) = \{A,C,G,T\}^{k(TF_i)}$, $i \in \{1, 2,...,N\}$. The output space $Y = \{0, 1\}$, indicating negative/positive examples.

For TF_i, let the learner be a function $A_{TF_i} : \{A,C,G,T\}^{k(TF_i)} \rightarrow \mathcal{H}$, for a predefined hypothesis space $\mathcal{H}$.

We restrict our hypothesis space set to $= \{\mathcal{H}_{PWM}, \mathcal{H}_{IUPAC}, \mathcal{H}_{MM}\}$ representing PWM, IUPAC and mismatch string models.

Table 2: Discrimination and motif discovery scores on algorithm benchmark suite. Average nCC-scores for *de novo* motif discovery with MEME and Weeder compared to best discrimination scores on our proposed algorithm benchmarks and remaining 64 datasets.

	Algorithm suite (Real)	Remaining data sets (Real)	Algorithm suite (Markov)
MEME	0.068	0.029	0.082
Weeder	0.11	0.10	0.052
Disc.	0.92	0.64	0.92

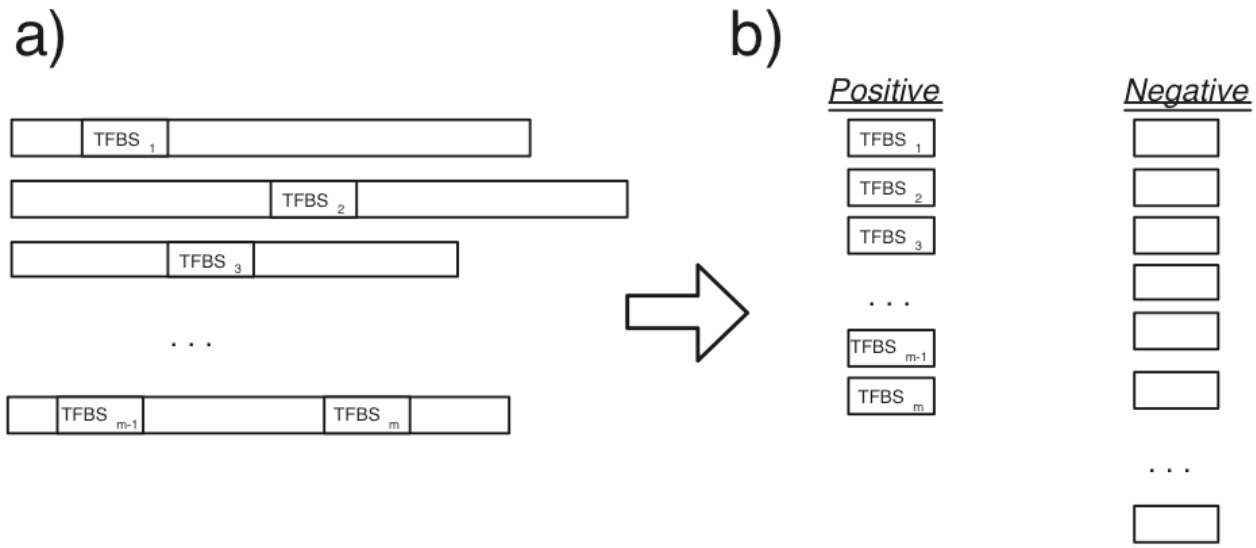


Figure 12
Generating positive and negative examples. A set of upstream DNA sequences for a transcription factor where a) *m* binding locations are identified, b) generating positive and negative examples.

We use correlation coefficient (CC) as our performance metric. In the optimization, CC is calculated at the oligo (sequence window) level, using oligos as individual samples as explained previously in this section. This differs slightly from the CC at the nucleotide level (nCC), which is the measure used in the results section to ensure consistency with the results of Tompa *et al*.

$$nCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (FP + TN) \cdot (TN + FN) \cdot (FN + TP)}} \quad (1)$$

Algorithms

Mismatch string model

The motif model for mismatch strings is a tuple, M_{MM} = <*cs*, *d* > where $cs \in \{A, C, G, T\}^n$ is a consensus string of length *n* and $d \in \{0, 1,..., n\}$ is the maximum Hamming distance from *cs*. Typical values for *n* is 6 to 20 and that of *d* is 1 to 4. Bounded values for *n* and *d* clearly suggests that hypothesis space is finite, although large when *n* gets bigger.

For mismatch strings we have developed an algorithm inspired by [24]. A main difference is that in our case, the motif locations are assumed known, *i.e.*, a supervised case. The main idea is to enumerate every substring *s* within a given Hamming distance *d* of each positive substring in the data set. For each such substring *s*, matches are determined as every substring *s*' of the sequences at a Hamming distance of at most *d* from *s*.

The method described above clearly does not consider all the hypothesis space explicitly, but the subset considered is actually enough to find the best hypothesis among all. Since the best hypothesis must cover at least one positive instance, the algorithm is guaranteed to find the best hypothesis even though not all hypotheses are explicitly enumerated. Thus, it suffices to evaluate the score of this subset of hypotheses. As this still involves scanning very many different motifs against the same sequences, a *q-gram* of the sequences is used to further accelerate matching of short motifs (length < 7) against sequences, and the algorithm of Yates *et al*. [25] for longer motifs.

IUPAC model

The motif model for IUPAC, M_{IUPAC}, is a degenerate string *ds* of length *n* where each position is a non-empty subset of {*A, C, G, T*}. These subsets correspond to the IUPAC symbols for DNA sequences. For finite *n*, the hypothesis space is finite but grows exponentially with *n*. A candidate string *s* is said to be a hit (match) against *ds* if every position of *s* is a subset of respective position in *ds*, otherwise it is a non-hit (non-match).

Finding a IUPAC expression that perfectly separates positive and negative substrings of equal length is indeed

straightforward if at all possible for a data set. IUPAC expressions are a subset of regular expressions, and induction of regular expressions from sequence examples is well studied. Each position of the motif is set to the union of characters occurring at the respective position of the positive instances. To see that this is the only solution, note that leaving out any symbol that occurs in a positive instance at a given position leads to the motif not covering the instance. Additionally, adding symbols not occurring in any positive instance at that position may only introduce hits in negative instances.

Perfect classification is generally impossible for the problem we consider. In our case, all degenerate strings should be generated exhaustively and evaluated against a scoring function. We have therefore developed an efficient algorithm that avoids unnecessary exploration of the hypothesis space. Our algorithm bears some similarities with the SPEXS algorithm for *de novo* motif discovery [26], but differs in that it uses bit-strings and pre-computation for optimization, calculates bounds and prunes subtrees, and of course that it solves a classification problem with known positive and negative instances instead of an unsupervised data mining problem. The details of the algorithm can be found in the supplementary material.

PWM model

The motif model for PWM is a tuple $M_{PWM} = <M, t>$ where M is a matrix of $4 \times n$ where each column is a probability distribution of the nucleotide vector $<A, C, G, T>$ and n is the length of the motif. A candidate string is considered to be a hit if the sum of probabilities in respective rows are greater than the threshold t, otherwise a non-hit. The hypothesis space is infinite regardless of n.

PWMs used for motif discovery are not just arbitrary matrices that best separates the motif occurrences from the remaining sequences. On the contrary, a PWM has a clear interpretation as a product multinomial probability distribution, or as containing log-odds values of motif versus background. In the supervised case we calculate the PWM from symbol frequencies in known motif locations for log-probability matrix. Additionally, background distributions are taken into account for log-odds PWM matrix. As the PWM thus is a direct function of the positive (and negative) instances of the data set, it is calculated easily and efficiently even for large data sets. We used the highest scoring PWM version for discrimination score. In motif discovery, the hypothesized motif locations used for constructing a PWM can in general be any probability distribution over all sequence locations. If the hypothesized motif locations exactly match the annotated sites, it corresponds to the solution in the supervised case.

Although the PWM itself is calculated directly from sequence data, there is more flexibility when it comes to determining a PWM score threshold to be used when determining binary hits of the PWM. Such score thresholds are commonly used to get a list of motif locations, instead of just a distribution on motif locations across the whole sequence data. As there are many ways of determining score thresholds, we exhaustively find the threshold that optimize the score of a given PWM.

We do this by exploiting the fact that the optimal threshold must be equal to the PWM match score of a positive instance. We therefore compute the classification score of the PWM with each of these thresholds and choose the threshold giving highest classification score. To see why this is optimal, consider a threshold t that is not equal to the PWM score of any positive instance. Increasing this threshold to the PWM score of the positive instance with least margin to the threshold (t') will give the same number of TP. As the threshold is more stringent the number of TN must be equal or higher. Thus, there exists a threshold t', corresponding to the PWM score of a positive instance, with at least as high score as the threshold t.

Dataset generation

We extracted sets of binding site fragments for 213 different TF matrices from the TRANSFAC database, version 9.4 [27]. A binding site fragment is the binding site region that is used in the construction of a matrix in the TRANSFAC alignment. Both real and Markov data set versions were constructed from the same fragment sets. For the real version, binding sites were kept in their original genomic sequence, which was truncated to a maximum length of 2000 bp. To make the data sets more coherent, we removed binding site fragments that contained degenerate bases, that had gaps in the TRANSFAC alignment, that were not located within the 2000 bp upstream of transcription start site in the sequence linked to by TRANSFAC, or that had two or more occurrences in the 2000 bp region. The binding sites used in a TRANSFAC matrix alignment may occur on opposite strands. To simplify the process of using these data sets we took the reverse complement of linked sequences when the binding site appeared on the negative strand. For the Markov version, binding sites were implanted in sequences generated from a third order Markov model inferred from all sequences of the corresponding real data set. Both the lengths of the Markov version sequences and the positions of the implanted binding sites were kept equal to the corresponding real sequences. Data sets with fewer than five binding sites were removed, leaving us with 114 real and 114 Markov data sets. While most data sets had from 5 to 25 sequences, there were data sets with up to 78 sequences. We then computed the best possible discrimi-

nation score, and used that for selecting the algorithm and model suites, as described in the main text.

Parameter settings

For all our runs of MEME we used version 3.5.3, downloaded from [28]. To avoid incurring biases in our results, we ran MEME with default DNA parameter values and without any manual curation of output data. We think this is realistic with regard to common usage of motif discovery methods, although performance could probably have been improved by tweaking parameters, pre-processing data sets and post-processing output data. For all runs of Weeder we used version 1.3 downloaded from [29]. We also ran Weeder with default parameters and without any manual curation. We used the large setting and the option telling Weeder that each sequence should contain at least one binding site. As Weeder requires the specification of organism, we supplied for each data set the most frequent organism.

Availability and requirements

Our proposed algorithm and model benchmark suites are available for download at http://tare.medisin.ntnu.no/. We have also implemented a web service for evaluating predictions and visualizing benchmark results. The implementation of the discrimination algorithms for the common motif models is freely available as Python source code at the same address.

Abbreviations

PWM: position weight matrix;

IUPAC: nomenclature for degenerate symbols as defined by the International Union of Pure and Applied Chemistry;

MM: mismatch motif model;

TFBS: transcription factor binding site;

TP, TN, FP, FN: true/false positives/negatives;

nCC: nucleotide-level Pearson's correlation coefficient (Formula 1)

Authors' contributions

GKS conceived the initial idea together with FD, devised the discrimination algorithms and drafted the manuscript. OA contributed to the scientific content of the paper, formalized the machine learning perspective and took part in writing the manuscript. VW revised and implemented the algorithms. FD supervised and took part in all stages of the project.

Additional material

Additional file 1

Details on discrimination algorithm for IUPAC model. *The file IUPAC_details.pdf describes the discrimination algorithm for IUPAC model in more detail, as well as several optimizations of computational efficiency.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-8-193-S1.pdf]

Acknowledgements

We want to thank Kjetil Klepper for help with extracting binding site fragments from TRANSFAC matrix alignments and Jostein Johansen for extensive help with benchmark runs and web site development. Osman Abul has been fully supported by an ERCIM fellowship. Finn Drabløs has been supported by The National Programme for Research in Functional Genomics in Norway (FUGE) in The Research Council of Norway and by The Svanhild and Arne Must Fund for Medical Research. The authors want to thank the anonymous reviwers for valuable input.

References

1. Sandve GK, Drabløs F: **A survey of motif discovery methods in an integrated framework.** *Biol Direct* 2006, **1(11):**.
2. Hughes JD, Estep PW, Tavazoie S, Church GM: **Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae.** *J Mol Biol* 2000, **296(5):**1205-14.
3. Bailey TL, Elkan C: **The value of prior knowledge in discovering motifs with MEME.** *Proc Int Conf Intell Syst Mol Biol* 1995, **3:**21-9.
4. Marsan L, Sagot MF: **Extracting structured motifs using a suffix tree-algorithms and application to promoter consensus identification.** In *RECOMB '00: Proceedings of the fourth annual international conference on Computational molecular biology* New York, NY, USA: ACM Press; 2000:210-219.
5. Blanchette M, Tompa M: **Discovery of regulatory elements by a computational method for phylogenetic footprinting.** *Genome Res* 2002, **12(5):**739-48.
6. Sinha S, Tompa M: **YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation.** *Nucleic Acids Res* 2003, **31(13):**3586-8.
7. Bortoluzzi S, Coppe A, Bisognin A, Pizzi C, Danieli G: **A Multistep Bioinformatic Approach Detects Putative Regulatory Elements In Gene Promoters.** *BMC Bioinformatics* 2005, **6:**121.
8. Fogel GB, Weekes DG, Varga G, Dow ER, Craven AM, Harlow HB, Su EW, Onyia JE, Su C: **A statistical analysis of the TRANSFAC database.** *Biosystems* 2005, **81(2):**137-54.
9. Bergman CM, Carlson JW, Celniker SE: **Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster.** *Bioinformatics* 2005, **21(8):**1747-9.
10. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z: **Assessing computational tools for the discovery of transcription factor binding sites.** *Nat Biotechnol* 2005, **23:**137-44.
11. Pevzner PA, Sze SH: **Combinatorial approaches to finding subtle signals in DNA sequences.** *Proc Int Conf Intell Syst Mol Biol* 2000, **8:**269-78.
12. Henikoff S, Henikoff JG, Alford WJ, Pietrokovski S: **Automated construction and graphical presentation of protein blocks from unaligned sequences.** *Gene* 1995, **163(2):**GC17-26.
13. Li N, Tompa M: **Analysis of computational approaches for motif discovery.** *Algorithms Mol Biol* 2006, **1(8):**.

14. Styczynski MP, Jensen KL, Rigoutsos I, Stephanopoulos GN: **An extension and novel solution to the (l,d)-motif challenge problem.** *Genome Inform* 2004, **15(2):**63-71.
15. Xing EP, Jordan MI, Karp RM, Russell S: **A hierarchical bayesian markovian model for motifs in biopolymer sequences.** In *Advances in Neural Information Processing Systems Volume 16*. Edited by: Becker S, Thrun S, Obermayer K. MIT Press, Cambridge, MA; 2002.
16. Barash Y, Elidan G, Friedman N, Kaplan T: **Modeling dependencies in protein-DNA binding sites.** In *RECOMB '03: Proceedings of the seventh annual international conference on Computational molecular biology* New York, NY, USA: ACM Press; 2003:28-37.
17. Ben-Gal I, Shani A, Gohr A, Grau J, Arviv S, Shmilovici A, Posch S, Grosse I: **Identification of transcription factor binding sites with variable-order Bayesian networks.** *Bioinformatics* 2005, **21(11):**1367-4803.
18. Lim LP, Burge CB: **A computational analysis of sequence features involved in recognition of short introns.** *Proc Natl Acad Sci USA* 2001, **98(20):**11193-8.
19. Zhao X, Huang H, Speed TP: **Finding short DNA motifs using permuted markov models.** In *RECOMB '04: Proceedings of the eighth annual international conference on Computational molecular biology* New York, NY, USA: ACM Press; 2004:68-75.
20. Stormo GD, Schneider TD, Gold L: **Quantitative analysis of the relationship between nucleotide sequence and functional activity.** *Nucleic Acids Res* 1986, **14(16):**6661-79.
21. Zhou Q, Liu JS: **Modeling within-motif dependence for transcription factor binding site predictions.** *Bioinformatics* 2004, **20(6):**909-16.
22. Leung HC, Chin FY: **Discovering DNA Motifs with Nucleotide Dependency.** *Sixth IEEE Symposium on Bioinformatics and Bioengineering (BIBE), IEEE Computer Society* 2006:70-77.
23. Mitchell TM: *Machine Learning* McGraw-Hill; 1997.
24. Keich U, Pevzner PA: **Finding motifs in the twilight zone.** *Bioinformatics* 2002, **18(10):**1374-81.
25. Baeza-Yates RA, Perleberg CH: **Fast and Practical Approximate String Matching.** In *CPM '92: Proceedings of the Third Annual Symposium on Combinatorial Pattern Matching* London, UK: Springer-Verlag; 1992:185-192.
26. Vilo J: **Discovering Frequent Patterns from Strings.** In *Tech. Rep. C-1998-9* Department of Computer Science, University of Helsinki; 1998.
27. Wingender E, Dietze P, Karas H, Knuppel R: **TRANSFAC: a database on transcription factors and their DNA binding sites.** *Nucleic Acids Res* 1996, **24:**238-41.
28. **MEME** [http://meme.nbcr.net/downloads/]
29. **Weeder** [http://159.149.109.16:8080/weederWeb/]

BMC Bioinformatics

Research article

Assessment of composite motif discovery methods

Kjetil Klepper*[1], Geir K Sandve[2], Osman Abul[3], Jostein Johansen[1] and Finn Drablos[1]

Address: [1]Department of Cancer Reasearch and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway, [2]Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway and [3]Department of Computer Engineering, TOBB University of Economics and Technology, Ankara, Turkey

Email: Kjetil Klepper* - kjetil.klepper@ntnu.no; Geir K Sandve - sandve@idi.ntnu.no; Osman Abul - osmanabul@etu.edu.tr; Jostein Johansen - j.johansen@ntnu.no; Finn Drablos - finn.drablos@ntnu.no

* Corresponding author

Published: 26 February 2008

BMC Bioinformatics 2008, **9**:123 doi:10.1186/1471-2105-9-123

Received: 26 July 2007
Accepted: 26 February 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/123

Abstract

Background: Computational discovery of regulatory elements is an important area of bioinformatics research and more than a hundred motif discovery methods have been published. Traditionally, most of these methods have addressed the problem of *single motif discovery* – discovering binding motifs for individual transcription factors. In higher organisms, however, transcription factors usually act in combination with nearby bound factors to induce specific regulatory behaviours. Hence, recent focus has shifted from single motifs to the discovery of sets of motifs bound by multiple cooperating transcription factors, so called *composite motifs* or *cis-regulatory modules*. Given the large number and diversity of methods available, independent assessment of methods becomes important. Although there have been several benchmark studies of single motif discovery, no similar studies have previously been conducted concerning composite motif discovery.

Results: We have developed a benchmarking framework for composite motif discovery and used it to evaluate the performance of eight published module discovery tools. Benchmark datasets were constructed based on real genomic sequences containing experimentally verified regulatory modules, and the module discovery programs were asked to predict both the locations of these modules and to specify the single motifs involved. To aid the programs in their search, we provided position weight matrices corresponding to the binding motifs of the transcription factors involved. In addition, selections of decoy matrices were mixed with the genuine matrices on one dataset to test the response of programs to varying levels of noise.

Conclusion: Although some of the methods tested tended to score somewhat better than others overall, there were still large variations between individual datasets and no single method performed consistently better than the rest in all situations. The variation in performance on individual datasets also shows that the new benchmark datasets represents a suitable variety of challenges to most methods for module discovery.

Background

A key step in the process of gene regulation is the binding of transcription factors to specific *cis*-regulatory regions of the genome, usually located in the proximal promoter upstream of target genes or in distal enhancer regions [1,2]. Each transcription factor recognizes and binds to a more or less distinct nucleotide pattern – a *motif* – thereby regulating the expression of the nearby gene. Determining the location and specificity of each transcription factor binding site in the genome is thus an important prerequisite for reconstructing the gene regulatory network of an organism.

Since establishing these binding sites experimentally is a rather laborious process, much effort has been made to develop methods that can automatically discover such binding sites and motifs directly from genomic sequence data. More than a hundred methods have already been proposed [3], and new methods are published nearly every month. There is a large diversity in the algorithms and models used, and the field has not yet reached agreement on the optimal approach. Most methods search for short, statistically overrepresented patterns in a set of sequences believed to be enriched in binding sites for particular transcription factors, such as promoter sequences from coregulated genes or orthologous genes in distantly related species.

In higher organism, however, transcription factors seldom function in isolation, but act in concert with nearby bound factors in a combinatorial manner to induce specific regulatory behaviours. A set of binding motifs associated with a cooperating set of transcription factors is called a *composite motif* or *cis-regulatory module*. In recent years, the field of computational motif discovery has therefore shifted from the detection of single motifs towards the discovery of entire regulatory modules.

The diversity of approaches to module discovery is even greater than for single motif discovery, and methods vary widely in what they expect as input and what they provide as output. For instance, methods like Co-Bind [4], LOGOS [5] and CisModule [6] expect only a set of coregulated or orthologous promoter sequences as input and are able to infer both the location and the structure of modules with few prior assumptions regarding their nature. These programs infer an internal model that includes a representation of each individual transcription factor binding motif as well as constraints on the distances between them. On the other hand, programs such as LRA [7] and Hexdiff [8] demand as input a collection of already known module sites to serve as training data. The known positive sites are used along with negative sequence examples to build a model representation which can then be compared to new sequences in order to identify novel module instances. Searching for new matches to a previously defined model might be considered a special case of module discovery and is often referred to as module *scanning*. Programs that specialize in searching for modules this way without inferring the models themselves include ModuleInspector [9] and ModuleScanner [10]. The general problem of module discovery, however, usually involves inferring both a model representation of the modules and to find their locations in the sequences.

Most module discovery methods require users to supply a set of candidate single motif models in the form of IUPAC consensus strings or position weight matrices (PWM) [11]. These are used to discover putative transcription factor binding sites in the sequences, and the programs then search for significant combinations of such binding sites to report as modules.

What constitutes a significant combination varies between methods. MSCAN [12], for instance, searches for regions within sequences that have unusually high densities of binding sites, more so than would be expected from chance alone. The types of the binding motifs are irrelevant, however, and each potential module instance is analyzed independently from the rest. Other tools, like ModuleSearcher [10], Composite Module Analyst [13] and CREME [14], search for specific combinations of motifs that co-occur multiple times in regulatory regions of related genes.

With an increasing number of programs available, both for single and composite motif discovery, there is a growing need among end users for reliable and unbiased information regarding the comparative merits of different approaches. A few independent investigations have been undertaken to assess the performance of selected single motif discovery methods, for instance by Sze *et al*. [15] and Hu *et al*. [16]. The most comprehensive benchmark study to date was carried out by Tompa *et al*. and included thirteen of the most popular single motif discovery methods [17]. The authors of this study also provided a web service to enable new methods to be assessed and compared to the original methods using the same datasets.

However, in spite of the increased interest in regulatory modules, we are not aware of any similar independent benchmarking efforts that have been undertaken with respect to composite motif discovery.

Results

We have developed a framework for assessing and comparing the performance of methods for the discovery of composite motifs. Sequence sets containing real, experimentally verified modules are made available for download through our web service, and users can test programs

of their own choice on these datasets and submit the results back to the web service to get the predictions evaluated. Results are presented both as tabulated values and in graphical format, and performances of different methods can be compared. Since most module discovery tools require users to input candidate motifs, each sequence dataset is supplemented by a set of PWMs capable of detecting the binding sites involved in the modules. To test how programs respond to varying levels of noise in the PWM sets, we created extended PWM sets for one of our datasets where the genuine matrices were mixed with various decoy matrices.

Scoring predictions

We adopted a simple and general definition of a module: a *module* is a *cis*-regulatory element consisting of a collection of single binding sites for transcription factors. A module is thus characterized by only two aspects in our framework: its *location* in a sequence and its *composition*, that is, the set of transcription factor binding motifs involved. A module's location is further defined as the smallest contiguous sequence segment encompassing all the single binding sites in the module, including also the intervening bases. For our purpose, the composition of a module is represented by a set of PWM identifiers. Different modules that share the same composition are said to belong to the same *module class*. Module class definitions may also be limited by structural *constraints*. These are rules governing, among others, the strand bias, order and distances between the transcription factor binding sites of modules of the same class. Since it requires a substantial effort to determine these constraints experimentally, this kind of information is available for a very limited number of classes. Few methods also report such module constraints explicitly. Consequently, we have chosen not to consider this aspect of modules further in our framework, at least for the time being.

Module discovery programs are requested to predict both the location of modules and to identify the motifs involved by naming the proper PWMs. However, not all programs are able to perform both these tasks. The MCAST program [18], for instance, only reports the location of predicted modules, even though it uses a set of PWMs to detect single binding sites internally. On the other hand, programs that discover single motifs *de novo* without relying on pre-constructed matrices have, of course, no way of correctly naming the motifs involved. Methods like that of Perco *et al*. [19] and GCMD [20] identify modules by looking for groups of PWMs whose binding sites consistently appear together in multiple sequences, but disregard any further information about the precise position of these sites. Hence, such programs only report the composition of modules but not their location. By assessing the location and composition aspects of modules separately, our framework can equally well be used with programs that predict only one or the other.

To measure prediction accuracy of methods with respect to module location, we have used the *nucleotide-level correlation coefficient* (*nCC*). This statistic has been widely used before, among others, for coding region identification and gene structure prediction [21]. It was also adopted by Tompa *et al*. to evaluate binding site predictions in their single motif discovery benchmark study. The value of *nCC* lies in the range -1 to +1. A score of +1 indicates that a prediction is coincident with the correct answer; whereas a score of -1 means that the prediction is exactly the inverse of the correct answer. Random predictions will generally result in *nCC*-values close to zero.

$$nCC = \frac{TP \cdot TN - FN \cdot FP}{\sqrt{(TP+FN)(TN+FP)(TP+FP)(TN+FN)}}$$

Here, *TP* is the number of nucleotides in a sequence that are correctly predicted by a program as belonging to a module, while *TN* is the number of nucleotides correctly identified as background. *FN* is the number of true module nucleotides incorrectly classified as background, and *FP* is the number of background nucleotides incorrectly classified as belonging to a module.

A similar statistic, the *motif-level correlation coefficient* (*mCC*), was used to evaluate prediction accuracy with respect to module composition. The definition of *mCC* follows that of *nCC*, except that instead of counting the number of nucleotides, we count the number of single motifs (or PWMs) correctly or incorrectly classified as being part of a module or not. Hence, for *mCC*, *TP* is the number of PWMs correctly identified as constituents of the module, while *FP* is the number of PWMs incorrectly predicted as being part of a module. Note that the correlation statistics, as defined here, are only applicable when both the datasets and the predictions made by a program contain a combination of module and non-module instances, if not, the divisor will be zero and the value of the statistic will be undefined. Consequently, the *mCC*-score is only informative when the set of PWMs supplied to a module discovery program contains false positives, i.e. additional matrices besides those that are actually involved in the modules. Final scores for each dataset are obtained by summing up *TP*, *FP*, *TN* and *FN* over all sequences before calculating the correlation scores. If no module predictions are made on a set of sequences, the resulting scores for *nCC* and *mCC* are assigned a value of zero rather than being left undefined. In addition to *CC* scores, several other statistics mentioned in [17] such as *sensitivity*, *specificity*, *positive predictive value*, *performance*

coefficient (phi-score) and *average site performance* are calculated for both nucleotide- and motif-level.

Datasets

We compiled three datasets from sequences containing experimentally verified regulatory modules. The first and the last two datasets have different characteristics and were chosen to complement each other to test methods under different conditions.

Our main dataset was based on annotated composite motifs from the TRANSCompel database [22]. The modules selected for this dataset are small, each consisting of exactly two single binding sites for different transcription factors (TFs), but we specifically chose modules that had multiple similar instances in several sequences. Sequences containing modules from the same class were grouped together producing ten sequence sets named after their constituent single motifs as shown in Table 1. Each of the sequences in a set contained at least one copy of the module with the same two motifs, but the order, orientation and distance between the TFBS could vary between sequences. Separate PWM collections, with matrices for the two single motifs involved, were constructed for each of the sequence sets. All in all there were eleven distinct single TF binding motifs in our full TRANSCompel dataset, and PWMs representing these motifs were collected from the companion TRANSFAC database [22]. Since TRANSFAC often contains several different PWMs for each motif, we grouped all the matrices corresponding to a particular motif into an equivalence set, essentially treating these PWMs as if they were one and the same with respect to prediction and scoring. In addition to the TRANSFAC matrix sets, we also constructed eleven custom matrices that were specifically tailored to the particular motifs and binding sites present in the sequences (see Methods). Assessment of module discovery programs on the TRANSCompel dataset was conducted using both the TRANSFAC sets and the customized PWM sets independently. The motivation for using two different PWM sets was to test the stability of methods and examine how the specific representations used for single motifs might influence the ability of methods to find the correct modules.

The two last datasets were based on combinations of TFBS found in the regulatory regions of genes specifically expressed in liver [23] and muscle [7] cells. The modules here are usually larger compared to the TRANSCompel modules, containing up to nine binding sites for four different motifs in the liver regulatory regions and up to eight sites for five motifs in the muscle regions. PWMs for these motifs were taken from the respective publications. The composition of the modules in these two datasets is variable; modules can contain multiple binding sites for the same motifs and not all motifs are present in every module.

While most programs require candidate PWMs to be entered, this can pose a problem for users who might not always know in advance the kind of modules that should be present in a sequence or which transcription factors that might bind. It could be the case, for instance, that a researcher has only a set of promoters from a coregulated set of genes and is interested in identifying the hitherto unknown module that controls the common expression of these genes. A popular strategy then is to employ an excessive set of PWMs which, hopefully, also includes the appropriate matrices. An extreme, but not unlikely, scenario would be to use all the matrices available from a published compilation like TRANSFAC (774 matrices in release 9.4) or Jaspar [24] (123 core matrices). Although this approach will inevitably lead to lots of false positive PWM matches that might thwart the module discovery process, good module discovery tools should nonetheless be able to report the true module instances without simultaneously predicting too many spurious occurrences.

Table 1: Datasets

Sequence set	Sequences	Modules	Total size (bp)	Module size, min-max (avg)
AP1-Ets	16	17	14860	14 – 99 (27)
AP1-NFAT	8	11	6893	14 – 19 (16)
AP1-NFκB	7	8	6532	18 – 135 (53)
CEBP-NFκB	8	8	7308	44 – 118 (84)
Ebox-Ets	4	6	3489	16 – 50 (25)
Ets-AML	5	5	4053	13 – 30 (19)
IRF-NFκB	6	6	5344	23 – 71 (43)
NFκB-HMGIY	6	7	5393	10 – 32 (13)
PU1-IRF	5	5	4530	12 – 14 (13)
Sp1-Ets	7	8	5787	16 – 117 (37)
Liver	12	14	11943	26 – 176 (112)
Muscle	24	24	20427	14 – 294 (120)

A brief overview of the ten TRANSCompel sequence sets and the liver and muscle datasets used in the assessment. Further information can be found in Additional File 1.

To simulate these conditions and test methods' response to noisy PWM sets, each PWM set under the TRANSCompel dataset was issued in multiple versions with progressively more decoy matrices added to the set of true annotated motifs. Decoy matrices were randomly sampled from the complete TRANSFAC compilation after removing the matrices corresponding to the true motifs for a sequence set. Decoy sets are available at 50%, 75%, 90%, 95% and 99% levels, where the percentage number relates the amount of decoy matrices in the set. Thus, a custom PWM set at the 90% level includes 2 genuine matrices and 18 decoy matrices. The number of decoy matrices in the TRANSFAC PWM sets varies with each module class but is always higher than for the custom sets at the same percentage level. Information on the exact number of PWMs in each set is available in Additional File 1. The 99% sets include as decoys all of the matrices from TRANSFAC which do not correspond to the correct motifs. They are called "99%" for consistency, although the actual percentage of decoys ranges between 95% and 99% depending on the module class. To avert artefacts stemming from possibly biased selections of decoys, all decoy sets (except at the 99% level) consist of ten independently sampled decoy collections, and the final correlation statistics for a decoy level are calculated by averaging prediction scores made from using each collection in turn. This also means that variation due to any stochastic nature of algorithms will be averaged over ten independent runs.

Benchmark of module discovery methods

Using our assessment framework, we benchmarked eight published methods for module discovery: *CisModule* [6], *Cister* [25], *Cluster-Buster* [26], *Composite Module Analyst (CMA)* [13], *MCAST* [18], *ModuleSearcher* [10], *MSCAN* [12] and *Stubb* [27]. See Table 2 for brief descriptions of each of these methods. CisModule, CMA and ModuleSearcher process all the sequences in a dataset simultaneously and look for instances of similar modules across multiple sequences. The other methods examine the sequences individually, although Stubb considers multiple instances of similar modules within the same sequence. Except for MCAST, which does not report module composition, all the programs report both the location and composition of modules. CisModule, however, predicts modules *de novo* without relying on supplied PWM sets and so does not name the single motifs involved the way we require. Hence, motif-level scores were not calculated for MCAST and CisModule. Cluster-Buster and MCAST report the full module segments, while the rest of the methods list the positions of the PWM hits in the modules. In these cases we extracted the start position of the first reported binding site and the end position of the last binding site and used these as the boundaries of a module prediction.

We generally relied on default parameter settings for all programs. However, since choosing the proper parameter values can sometimes prove crucial for a method's performance, we decided to provide the programs with a few general clues where applicable; specifically, that the size of modules should not exceed 200 bp (300 bp in the muscle dataset) and that the modules should consist of exactly two single binding sites for different TFs in the TRANSCompel dataset but possibly up to ten binding sites for four and five different TFs on the liver and muscle sets respectively. Furthermore, binding sites could potentially overlap and the composition of the modules in liver and muscle sets should be allowed to vary between sequences.

Figures 1a and 1b show the resulting nucleotide-level correlation scores on each sequence set in the TRANSCompel dataset when methods were supplied with TRANSFAC matrices and custom matrices respectively. The scores vary widely between individual sequence sets but are generally fairly well correlated between methods, so that most methods tend to get high (or low) scores on the same sets. The notable exception is CisModule which performs poorly on all sequence sets. The correlation suggests that some sequence sets are inherently more easy (or difficult) to tackle than others. Scores for CEBP-NFκB and IRF-NFκB are the highest overall. The reasons why these sets are generally easy to predict might be that their modules are quite long and the matrices representing the single binding motifs have high information content (see Table 3 and Additional File 1). Conversely, the short size of the modules and the low information content of PWMs for AP1-NFAT would make this a hard sequence set. We also calculated combined scores for the whole TRANSCompel dataset which are shown in the inset legends of Figure 1 and graphically in Figure 2. These combined scores were obtained by summing up TP, TN, FP, FN over all sequence sets when calculating the score measures. The highest combined *nCC* scores achieved were 0.388 with the TRANSFAC matrices (MSCAN) and 0.38 with custom matrices (MCAST). The average performances across all methods were also about the same with the two PWM sets. Some methods performed quite differently depending on the PWMs, however. For instance, MCAST scored much better using custom matrices than with TRANSFAC matrices, while MSCAN and Cluster-Buster did a better with job with TRANSFAC. The rank order of methods is thus somewhat altered between the two cases. Still, some tendencies remain: CMA, Cluster-Buster, MCAST, ModuleSearcher and MSCAN occupy the top five positions in both cases, followed by Cister and Stubb and then finally CisModule which consistently scored lowest.

Figure 3 shows the results of mixing the PWM sets with an equal proportion of decoy matrices. The addition of decoy PWMs leads to a drop in score values for almost all meth-

Table 2: Description of module discovery tools

CisModule	CisModule models the structure of sequences with a two-level hierarchical mixture-model and uses a Bayesian approach with Gibbs sampling to simultaneously infer the modules, TFBSs and PWMs based on their joint posterior distribution, which is the probability of a model given the input sequence set. At the first level, sequences are viewed as a mixture of module instances and background. At the second level, modules are modelled as a mixture of motifs and inter-module background. Parameters of the model include the widths and representations (PWMs) of single motifs and parameters related to distances between modules and between TFBS within modules. From a random initialization, CisModule iteratively cycles through steps of parameter update and module-motif detection. New parameter values are sampled from their conditional posterior distributions based on the currently predicted modules and motifs, and new predictions of modules and TFBSs are then sampled based on these updated parameter values. Positions in the sequences where the marginal posterior probability of being sampled within modules was greater than 0.5 were output as module predictions.
Cister	Given a set of PWMs and parameters specifying the expected number of motifs in modules, the expected distances between motifs in modules and the expected distance between modules, *Cister* builds a Hidden Markov Model (HMM) with three basic states: *motif*, *intra-module background* and *inter-module background*. Transition probabilities between these states follow geometric distributions according to the expected values input by the user. In the motif state, one of the PWMs is chosen uniformly at random and used to decide the probabilities of outputting nucleotides. Background-state emission probabilities are estimated from a sliding window centered on the current base in the query sequence. From this HMM, the posterior probability that each base in the input sequence was generated from a module state as opposed to the inter-module state can be calculated. Predicted modules are defined to occur at local maxima of this posterior probability curve where the value is at least 0.5 and no larger value is observed within 1200 bp.
Cluster-Buster	Cluster-Buster is developed by the same group that made *Cister* and is designed to search for clusters of pre-specified motifs in nucleotide sequences. Like Cister, Cluster-Buster constructs a HMM-model based on the user-supplied PWMs, an expected distance between motifs in clusters and background distributions estimated from the input sequence over sliding windows. Log likelihood ratios are used to determine whether a sequence is more likely to be generated by a "cluster-model" or a "background-model". Cluster-Buster uses a linear time heuristic to rapidly estimate log likelihood ratios for all subsequences of the input sequence and outputs those subsequences with ratios above a specified threshold that do not overlap with other higher scoring subsequences.
Composite Module Analyst (CMA)	The promoter model in CMA is expressed as a Boolean combination of one or more *composite module*s (CM), each of which consist of a set of single, independent motifs as well as pairs of motifs that must obey certain constraints on distance and orientation. Given a candidate promoter model, the method searches for potential matches to the CMs in the sequences, and a final promoter score is calculated after the presence or absence of each CM is established. CMA employs a Genetic Algorithm to search for the promoter model which best discriminates between a set of positive (co-regulated) and a set of negative sequences. The fitness function is based on a linear combination of several properties of the distribution of the promoter scores and of the individual CM scores in the two sequence sets.
MCAST	MCAST builds a HMM-model consisting of an intra-module state, an inter-module state and motif-states based on the supplied PWMs. The score for a motif-state is called a *p*-score and is the negative logarithm of the *p*-value of a log-odds score based on the probability of a segment in the target sequence being generated either by the PWM or a fixed, user-specified zero-order Markov background model. MCAST forbids transitions into motif-states that result in *p*-scores lower than some chosen threshold. Some state transitions are associated with certain costs. For instance, entering the inter-module state from a motif-state incurs a large one-time penalty while cycling through the intra-module state incurs smaller penalties for each nucleotide emitted. The Viterbi algorithm is used to find the highest scoring path through the HMM with respect to the input sequence, classifying each position in the sequence as either belonging to a module or to the background. Potential module segments are scored according to the number of motifs in the module and the *p*-scores of these motifs and are penalized by the number of intra-module background bases. Finally, modules are ranked according to the estimated *E*-values of these scores.
ModuleSearcher	Given a list of PWM hits with match scores for putative TFBSs in a sequence set, ModuleSearcher finds the module model (set of *k* PWMs) that best fits the sequences. The score of a module model is calculated as the sum of scores over all sequences, and the score function for a single sequence is based on the best scoring set of TFBSs in the sequence that corresponds to the PWMs in the module model. To be considered a valid TFBS set the binding sites must all lie within a short window, and the user can choose to ignore TFBS sets with overlapping binding sites or penalize sets that lack sites for some PWMs. An A*-algorithm (or alternatively a Genetic Algorithm) is employed to search the space of possible subsets of *k* motifs from the full PWM library in order to find the highest scoring module model.
MSCAN	MSCAN discovers modules by evaluating the combined statistical significance of sets of potential non-overlapping TF binding sites in a sliding window along the input sequence. PWMs are compared against each position within the window to obtain match scores, and *p*-values are calculated as the probability of obtaining similar or higher scores at a specific position in a random sequence with nucleotide distribution similar to the distribution in the window. MSCAN proceeds by calculating significance scores for all combinations of up to *k* binding sites in the window and then selects the optimal combination (the one with the lowest score). A prediction is output if a final *p*-value computed from this score is less than some user-specified threshold.

Table 2: Description of module discovery tools *(Continued)*

Stubb	The HMM used by Stubb consists of motif states based on supplied PWMs and a single background state based on a *k*th-order Markov model with probability distribution estimated from a sliding window. The scoring function is the log likelihood ratio that the sequence within a limited window was more likely generated by the full model than with a HMM consisting of only the background state. Unlike the other HMM methods presented here, the transition probabilities between states in Stubb are not based on user-input expectancies, but are estimated from the sequence using the Baum-Welch algorithm. This procedure finds the set of transition probabilities that maximizes the scoring function. If Stubb finds that some motifs are highly correlated with respect to order, it can make use of *correlated transition probabilities*. This means that the probability of entering a specific motif state will dependent on which previous motif was output. Stubb can also utilize phylogenetic comparisons between sequences from multiple species to highlight potentially regulatory modules.

The table contains short descriptions of the eight methods included in the assessment. All methods except for CisModule rely on supplied PWMs and consider matches on both strands, usually with equal probability (however, Stubb can estimate strand biases for all PWMs in a preprocessing step). Not all methods are able to consider overlapping single binding sites, which do occur in a few modules.

ods. The drop is greater for the TRANSFAC PWMs, presumably because these sets contain more genuine matrices and therefore also more decoys. Contrary to expectation, some methods actually score slightly better on certain sequence sets when decoys are in use. Examples are Cister on Ets-AML and Stubb on Ebox-Ets with custom matrices. One explanation for this could be that these methods make use of decoy motifs that just happen to have a high degree of overlap with genuine modules. To examine whether the modules are predicted with the correct motifs or not, we can look at the corresponding motif-level correlation scores as shown in Figure 4. The generally high mCC scores obtained for IRF-NFκB support the notion that this is an easy sequence set, while the difficulty for most methods in selecting the correct motifs for CEBP-NFκB explains the higher drop seen in nCC for this set when decoys were added. CMA and ModuleSearcher are by far the best methods at predicting the correct composition of modules with both TRANSFAC and custom PWMs as input, although CMA does perform notably poor on two specific sequence sets. The mCC score for the third best method, Cluster-Buster, is less than half of that of ModuleSearcher.

Figures 5 and 6 show score tendencies as increasingly more decoys are added to the PWM sets. The nucleotide-level performances of CMA and ModuleSearcher are only slightly affected by the larger amounts of decoys, whereas the scores for the other methods steadily decline. At the motif-level we clearly see a division into two groups with

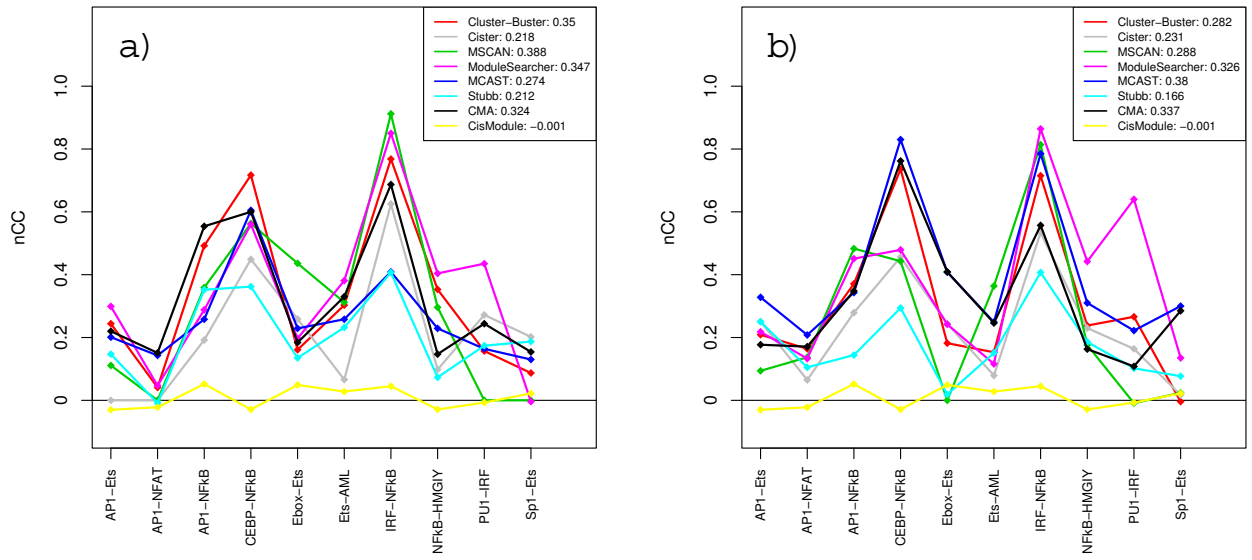


Figure 1
Nucleotide-level correlation scores on the TRANSCompel dataset. The graphs show *nCC* scores for each of the ten sequence sets in the TRANSCompel dataset when methods are supplied with TRANSFAC PWM sets (a) and custom matrices (b).

Table 3: Correlations between dataset properties and *nCC* scores

	TRANSFAC PWMs		Custom PWMs	
	Average *nCC*	Highest *nCC*	Average *nCC*	Highest *nCC*
Number of sequences	-0.23	-0.16	-0.23	-0.05
Length of shortest sequence	0.30	0.18	0.30	0.13
Average sequence length	0.40	0.33	0.42	0.43
Total sequence set length	-0.19	-0.12	-0.18	-0.02
Number of module instances	-0.38	-0.32	-0.40	-0.19
Size of smallest module	**0.61**	**0.69**	**0.67**	**0.73**
Size of largest module	0.26	0.34	0.19	0.35
Average module size	**0.60**	**0.68**	0.59	**0.70**
Module size standard deviation	0.23	0.29	0.13	0.29
IC-content (lowest)	0.46	0.45	**0.73**	0.47
IC-content (total)	**0.75**	**0.73**	**0.78**	0.54
Module/background-ratio	0.53	0.61	0.51	**0.63**

We conducted a simple correlation analysis to examine which properties of the TRANSCompel sequence sets and PWMs correlated best with the highest and average *nCC* scores obtained by the methods on these sets. "IC-content (lowest)" is the *information content* (IC) of the PWM with the lowest IC of the two involved in each sequence set. The information content of a PWM is inversely related to the amount of variability in the binding patterns from which the PWM is constructed [38]. PWMs with higher information content are more specific and match only sites with a high degree of similarity to the consensus motif. "IC-content (total)" is the sum of IC-contents for the two motifs (for TRANSFAC PWMs we used the PWM with the highest IC in each equivalence set to represent the motif). The three highest values are highlighted in each column. The properties that seem to correlate best with methods' performances are the minimum and average size of modules (in basepairs) and the total IC-content, which would imply that module discovery is harder for datasets containing short and degenerate modules.

CMA and ModuleSearcher performing significantly better than the rest. Additional graphs detailing the effects of added noise with respect to each individual sequence set and the variations due to different decoy selections can be found at our web site.

Results for the liver and muscle datasets are shown in Figures 7 and 8. For these datasets we supplied only four liver- and five muscle-PWMs respectively, and no decoy matrices were used. Since the modules in these datasets do not necessarily include binding sites for all of these motifs however, we could calculate motif-level scores by treating the PWMs for the missing motifs as false instances. All methods, except CisModule, did a better job on locating the modules in the liver dataset than in the TRANSCompel dataset. Cluster-Buster scored highest, but Stubb

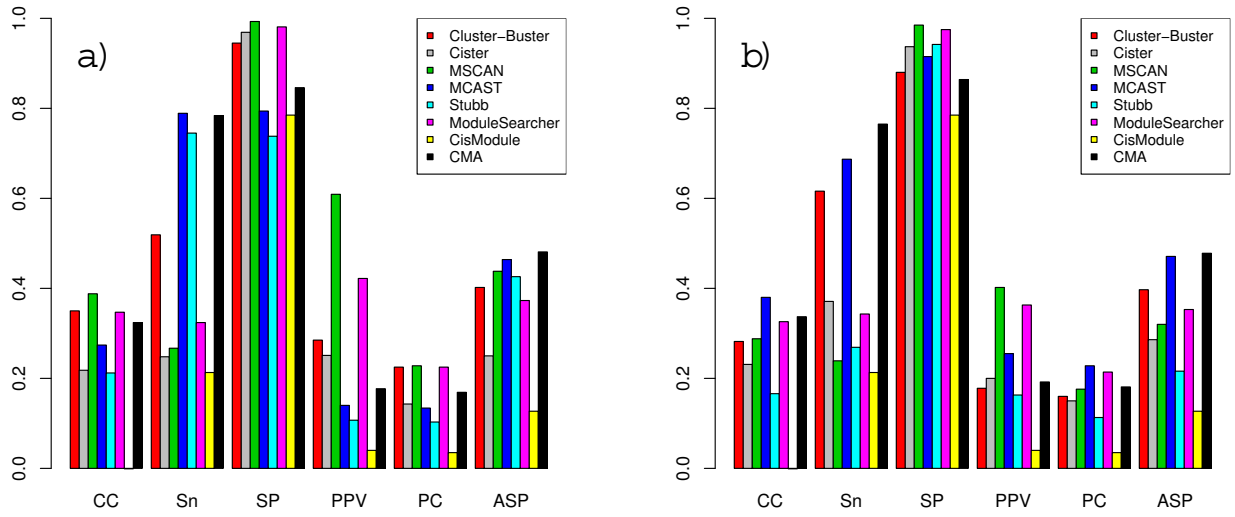


Figure 2
Combined performance scores on the full TRANSCompel dataset. Combined nucleotide-level scores obtained for different performance measures when using TRANSFAC PWM sets (a) and custom matrices (b).

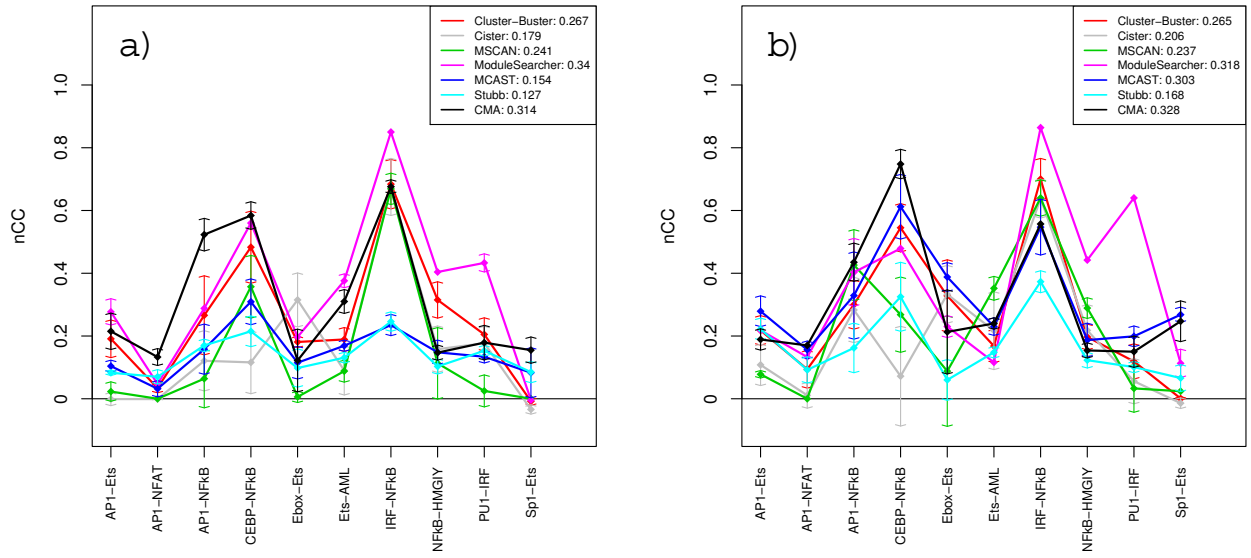


Figure 3
Nucleotide-level correlation scores with 50% noise in the PWM sets. The graphs show *nCC* scores when using TRANSFAC PWM sets (a) and custom matrices (b) with an equal proportion of decoy matrices added. Each value represents the average score over ten runs with different decoy selections.

showed the largest improvement in *nCC* score. The motif-level scores, on the other hand, were not very high, which can most likely be attributed to overprediction of motifs in the case of CMA and underprediction for MSCAN. Results on the muscle dataset display the same main tendencies as the other two datasets, but for the first time,

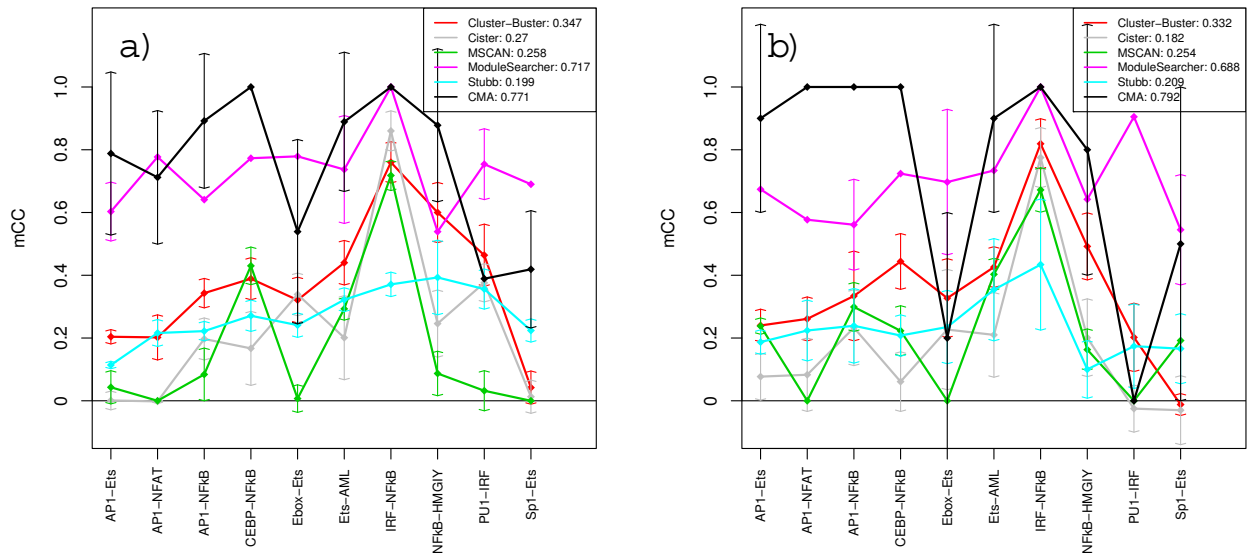


Figure 4
Motif-level correlation scores with 50% noise in the PWM sets. The graphs show *mCC* scores when using TRANSFAC PWM sets (a) and custom matrices (b) with an equal proportion of decoy matrices added. Each value represents the average score over ten runs with different decoy selections.

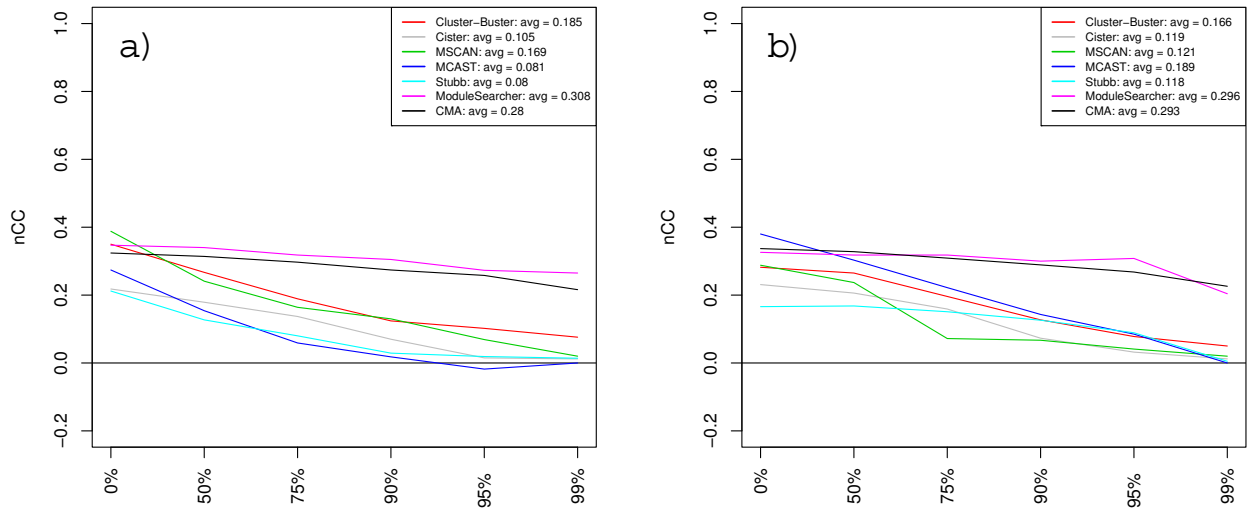


Figure 5
Nucleotide-level correlation scores at different noise levels. Plot of *nCC* scores at increasing noise levels when methods are supplied with TRANSFAC PWM sets (a) and custom matrices (b). Scores shown are averages over all sequence sets and decoy selections at each noise level. MCAST was unable to function properly with very large PWM sets and was therefore assigned a score of zero at the 99% level.

CisModule obtains an *nCC* score above zero and actually bypasses one the other methods.

Discussion

Objective benchmarking efforts are important for providing unbiased reviews of published methods and for establishing the methodological frontier with respect to bioinformatics techniques. In this study we wanted to explore benchmarking in the context of module discovery and to investigate related design issues such as dataset construction and performance evaluation.

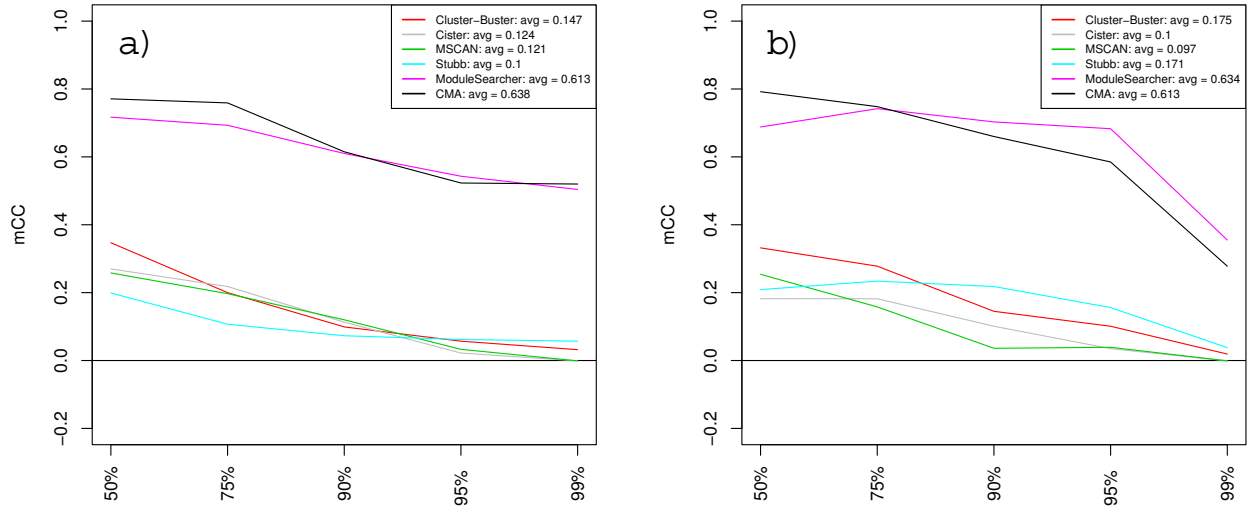


Figure 6
Motif-level correlation scores at different noise levels. Plot of *mCC* scores at increasing noise levels when methods are supplied with TRANSFAC PWM sets (a) and custom matrices (b). Scores shown are averages over all sequence sets and decoy selections at each noise level.

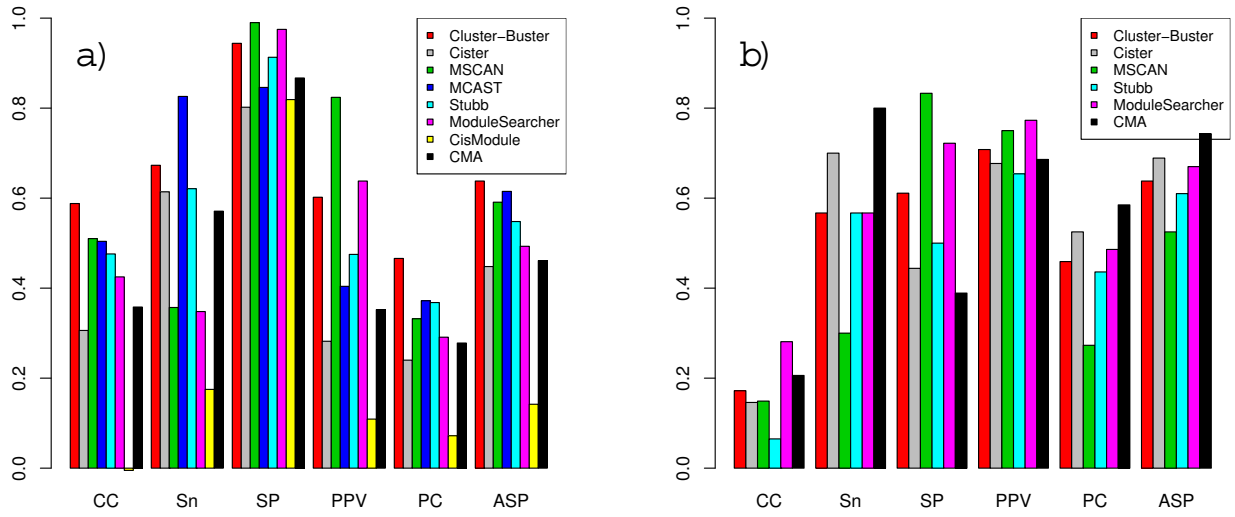


Figure 7
Performances on the liver dataset. Scores obtained on the liver dataset for different performance measures at nucleotide-level (a) and motif-level (b).

Benchmarking of tools for composite motif discovery is harder than benchmarking of single motif discovery tools, since the former methods are more diverse with respect to input requirements and the type of predictions they make. We have aimed at creating a simple and general framework that can be used with a wide range of methods. Nevertheless, we do not provide every kind of information that programs might ask for, and not all module discovery tools can be fairly assessed with our system.

To construct the benchmark datasets we relied on real genomic sequences containing experimentally verified modules, rather than creating synthetic datasets with fabricated and planted modules. The motivation for only

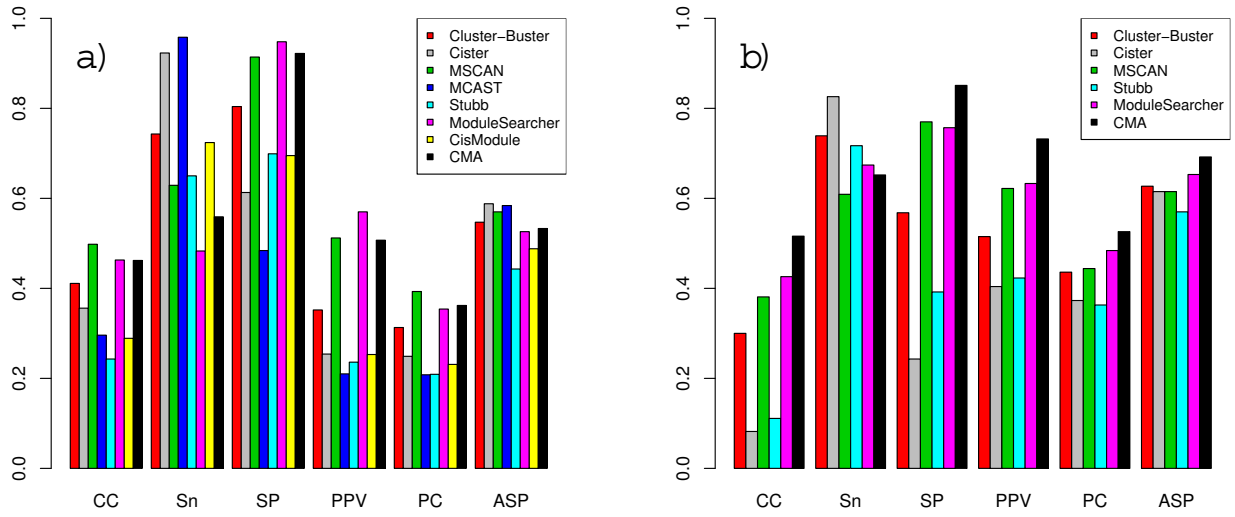


Figure 8
Performances on the muscle dataset. Scores obtained on the muscle dataset for different performance measures at nucleotide-level (a) and motif-level (b).

using real data was to avoid introducing artificial bias related to the composition and constraints of modules. Good benchmark datasets should be diverse enough to discriminate the behaviour of different methods, when possible, and provide them with a wide range of realistic challenges. For module discovery these challenges could include discovering modules with few or many single motifs, tightly clustered or widely spaced motifs and modules with highly conserved or degenerate binding sites. Ideally, benchmark datasets should also be novel to the methods tested. Currently the amount of experimental data available is too limited to achieve all of these goals. The particular dataset we have constructed based on TRANSCompel data is novel in terms of performance testing. The modules in TRANSCompel are short, however, and to include larger modules we were forced to rely on a few well-known datasets from liver and muscle regulatory regions that have been used extensively in the past for testing and possibly for designing and developing module discovery methods. Some methods might therefore be intrinsically biased to perform well on these sets. It is conspicuous, for instance, that CisModule – which was tested with muscle data in its original publication – scored comparably well to the other methods on our muscle set, yet close to zero on both the TRANSCompel and liver datasets.

We chose the *correlation coefficient* as our main statistic for evaluating and comparing module discovery methods because it captures aspects of two of the most commonly used performance measures – *sensitivity* and *specificity* – into a single score value. However, since different statistics often favour different methods, it is prudent to consider several measures to get a better comprehension of each method's qualities. The sensitivity measure (Sn), for instance, tells us to what extent a method's predictions include the true module instances. At the nucleotide level, MCAST seems overall to be the most sensitive method among those tested here, while CMA shows high sensitivity on the TRANSCompel dataset. Yet, to achieve these high sensitivity scores the methods at the same time make a lot of false positive predictions, as can be seen from the lower *positive predictive values* (PPV). MSCAN and ModuleSearcher, on the other hand, generally have the highest nucleotide-level PPV scores, which tells us that the positive predictions made by these two programs are more trustworthy than predictions made by the other programs.

PWMs from the TRANSFAC database were used to represent both the true motifs and the decoys for the TRANSCompel dataset. A potential problem when using TRANSFAC is that many of the matrices are quite similar to each other [28]. This is partly due to some TFs being represented by several PWMs, but also because different TFs might bind to similar-looking motifs. As a result, module discovery programs can be unduly penalized for selecting an incorrect PWM at the motif level, even though the predicted PWM is very similar to the target. We have tried to remedy this situation by grouping together PWMs that correspond to the same TFs and consider these as the same motif with respect to scoring. However, there might still be other matrices in the decoy sets that can match with the annotated binding sites.

Since we are using real genomic sequences, some of the predicted modules that we label as false positives can in fact represent unannotated true positives, and so the actual performance of methods might very well be better than indicated, especially at high noise levels.

It should be noted that while the annotated length of a TF motif may vary from binding site to binding site, the length of a standard PWM is fixed, and PWMs do not always match the locations of their corresponding binding sites precisely. Perfect *nCC* scores can therefore be difficult or even impossible to obtain. The *nCC* score also drops fast if a method predicts a larger module region than what is annotated, even though the target module is correctly covered by the predicted region. This can severely penalize methods that tend to predict large module regions, especially on the TRANSCompel dataset where most modules are rather short.

Some programs can utilize additional information to strengthen confidence in predictions and improve their performance. For instance, Stubb is a sensitive method and the predictions it makes usually include the correct modules, especially when using large PWM sets; yet, its *CC*-scores are generally low because it simultaneously predicts a lot of false positives. Stubb can employ a phylogenetic footprinting [29] strategy to filter out many of these false predictions, but it requires that orthologous sequences from related species are supplied along with the regular sequences. However, in order to make the tests as comparable as possible, we have not made such additional information available to the programs in our benchmark test, unless the type of information can be expected to be readily obtained for any dataset.

Caution should thus always be taken when interpreting score values, since the reported scores might not accurately reflect the optimal capabilities of the methods. Also, we have run the programs using mostly their default parameter settings. We are fully aware that adjusting the parameters can greatly affect the performance of a program, however, selecting the most appropriate parameter values be can be tricky and running methods with default settings is probably closer to typical usage.

It is inherently difficult to conduct an assessment that is fair to all methods. Even the most minute design choice can influence the outcome if it unintentionally favours some methods over others. For instance, limiting the size of input sequences will be beneficial for most module discovery tools since it improves the signal-to-noise ratio. On the other hand, using too short sequences can disadvantage methods that require substantial amounts of data in order to derive elaborate background models. The best solution, then, is to try to balance the scales by subjecting methods to several different situations with datasets exhibiting a range of characteristics. This will make it harder still to declare a winner, since it will inevitably lead to even greater variation in the results. Then again, the purpose of benchmarks tests need not be to identify a single program that can be recommended for all needs, but rather to determine how different methods behave under different conditions, thus enabling us to select the most appropriate tool to use in specific situations.

The results from our assessment of eight published module discovery tools show that the top scoring method does vary a lot between datasets. On the TRANSCompel dataset, for instance, all methods save Stubb and CisModule score better than the others on at least one sequence set. But there is also a tendency for some methods to perform consistently better or worse across several datasets. CisModule performed poorly on most sequence sets, Cister and Stubb usually scored somewhere in the middle, while CMA, ModuleSearcher, MSCAN and Cluster-Buster were often found among the top scoring methods on each set. CMA and ModuleSearcher were clearly best at identifying the correct motif types involved in the modules, and they were also the only methods capable of coping with large and noisy PWM sets. The other PWM-reliant methods appear to be more suited for detecting modules with some prior expected composition than for discovering completely new and uncharacterized modules.

There was some variation when using custom PWMs as opposed to TRANSFAC PWM sets. The average performance over all methods on the whole TRANSCompel dataset was about the same in both cases, but there were a lot of local differences between sequence sets. The most extreme example can be seen on the Ebox-Ets sequence set where MSCAN scores highest of all with TRANSFAC matrices, yet completely fails to find any true modules with custom matrices. The average deviation in scores when using either PWM set was about 0.11 and the effect could go both ways. MCAST was the only method which almost consistently scored better with one set, namely custom matrices.

Conclusion

While improvements can still be made to our systems, we have taken a first step towards developing a comprehensive testing workbench for composite motif discovery tools. The assessment system is based on two established datasets for module discovery plus a novel dataset we constructed from TRANSCompel module annotations. The performance of methods on our novel set is comparable to the previous two, demonstrating its utility as a benchmark set. Together these datasets challenge methods to discover modules with different characteristics and varying levels of difficulty.

Not surprisingly, trying to discover composite motifs *de novo* proves to be much more challenging than relying on PWMs as an aid to detect potential single binding sites. With large and noisy PWM sets, however, it becomes crucial to consider multiple instances of conserved motif combinations in order to identify true modules. In general, our study shows that there are still advances to be made in computational module discovery.

Methods

TRANSCompel dataset

Our main dataset was based on modules annotated in the TRANSCompel database [22], which is one of very few databases that contain entries for composite elements whose combinatorial binding effects have been verified through biological experiments. It comes in both a professional licensed version and a smaller public version. Our dataset was selected from TRANSCompel Professional version 9.4 which contains 421 annotated module sites from 152 different module classes. The largest modules registered in TRANSCompel are triplets (34 entries) with the remaining being pairs of binding sites (387 entries). To ensure a minimum of support for each module class, we considered only classes that had at least five annotated module sites. Unfortunately, this requirement excluded all triplets and left us with only pairs. After further discarding a few modules that were too weak to be detected with stringent PWM-thresholds, we ended up with ten sequence sets encompassing 81 module binding sites in 63 different sequences. The longest module spanned 135 bp with the average being 33 bp. The binding sequences of modules are specified in TRANSCompel by using uppercase letters to indicate bases of the constituent single motifs and lowercase letters for the intra-module background. We used the supplied references to the EMBL database [30] to obtain additional sequence bases flanking these module sites but set an upper limit of 1000 bp on the length of the sequences used. Most of the sequences were from human or mouse but also some other mammalian and a few viral sequences were included. Each sequence set was constructed around modules of one particular class made up of two single motifs

from the following set of eleven: AML, AP-1, C/EBP, E-box, Ets, IRF, HMGIY, NF-AT, NF-κB, Sp1 and PU.1. The sequence sets contained between 4 and 16 sequences and the sequences themselves ranged in length from 294 to 1000 bp (average 884 bp). All sequences contained at least one module instance, but sometimes up to three, of the designated class. Some sequences also included annotated modules of other classes. This will usually not be a problem at low noise-levels, because the other modules will only interfere if the set of PWMs supplied to a program contains decoy matrices corresponding to the motifs involved in these modules. As the noise-level approaches 99%, however, this will inevitably happen because the PWM sets then include the complete TRANSFAC collection. Since we use real genomic data, there is also always a possibility that additional unknown modules are present in the sequences. Even so, for a particular sequence set, only module sites corresponding to the designated class of that set were considered true positives.

Although the TRANSCompel database itself does not provide matrix representations for the motifs involved in modules, its companion database TRANSFAC does [22]. Unfortunately, there is not a one-to-one correspondence between transcription factors and matrices in TRANSFAC, and a single factor (or family of factors that recognize the same motif) can be represented by several different PWMs. Instead of selecting just one canonical PWM to use for each motif, we collected all matrices related to a specific motif and treated the whole set as an equivalence class. Thus, a motif can be represented by either one of the PWMs in the corresponding set, and predicted binding sites in the sequences are considered to be instances of the same motif even if the binding sites are predicted by different PWMs from the equivalence set.

As an alternative to these TRANSFAC sets, we also constructed custom PWMs for the eleven motifs involved in our module classes. For each motif we extracted the corresponding annotated binding sites plus four flanking bases on each side from our sequences and used MEME [31] to align them and infer a PWM model for the motif. Constructing matrices from the same binding sites they will later on be used to detect introduces a circularity which will probably make these sites easier to find than if the PWMs had been constructed from independent sequences. This was intentional, however. Since the purpose of our study was to assess the methods' abilities to find significant *combinations* of binding sites rather than individual sites, we wanted the individual sites to be easily detectable. To verify that the annotated single binding sites in the TRANSCompel dataset were indeed detectable by our matrices, we used an algorithm from the "TFBS" package [32] to match the PWMs against the sequences. Of the 81 single binding sites in the dataset, all but ten could be detected with an 85% relative cut-off threshold. When we lowered the cut-off to 75%, all sites could be detected. Single binding sites were considered to be detected if a predicted match to the corresponding PWM overlapped with the annotated binding site. For the TRANSFAC matrices, we regarded it as sufficient if any one of the matrices in the equivalence set made a prediction that overlapped with the annotated site.

Liver and muscle datasets

The liver dataset was based on a set of regulatory regions used as a positive training set to develop a model of liver specific regulation in the paper by Krivan and Wasserman [23]. Sequence data as well as PWM models of four TFs implied in liver specific regulation (C/EBP, HNF-1, HNF-3 and HNF-4) was downloaded from their supplementary web site [33]. After inspection of the sequence annotations, we discarded from further consideration those regulatory regions that only contained a single TFBS and also smaller annotated regions that were completely overlapped by larger regions. Furthermore, we ignored a small set of TFs that only had one binding site each in the whole dataset. This left us with regulatory regions consisting of two or more binding sites for the four TFs previously mentioned. The start position of the first TFBS and the end position of the last TFBS in each region were used as module boundaries, and the modules thus obtained varied in length from 26 to 176 bp with and average of 112 bp. Long sequences were cropped to a maximum of 1000 bp. The resulting dataset after curation consisted of 14 modules in 12 sequences with 51 binding sites for 4 different TFs. Eight of the sequences were human, two were from rat and the last two from mouse and chicken.

For the muscle dataset we selected a subset of the regulatory regions from the paper by Wasserman and Fickett [7] obtained from their web site [34]. Five motifs (Mef-2, Myf, Sp1, SRF and Tef) were reported as important in muscle regulation, and PWMs for these motifs were downloaded from the same site. We chose regions that had at least two annotated binding sites for motifs in this set and used the first and last binding site in the regions to delimit the modules. Since most of the sequences at the website were excerpts and rather short, we tried to extend them where possible by obtaining the original sequences from EMBL, though limiting the sequences to a maximum of 1000 bp as usual. The final muscle dataset used contained 24 sequences with one module in each and a total of 84 TFBS for 5 motifs. The smallest module spanned 14 bp and the longest 294 bp (average 120 bp). 10 sequences were from the mouse genome, 6 from human, 5 from rat, 2 from chicken and 1 from cow.

Further statistics on the datasets and PWMs used are summarized in Table 1 and Additional File 1.

Running the programs

Most of the methods tested could be run directly from the input sequences and a set of PWMs. Both CMA and ModuleSearcher, however, rely on separate programs to match the PWMs against the sequences in a preprocessing step. For ModuleSearcher we used the program MotifScanner since both of these methods are part of the Toucan tools suite for regulatory sequence analysis [35]. MotifScanner was run with a third order background model based on vertebrate promoter sequences, which was also available with Toucan. CMA comes bundled with Match [36] for PWM scanning. Match utilizes two different threshold values which should be individually fitted for each specific PWM. Preconstructed cut-off profiles for TRANSFAC matrices are available for different conditions, for instance to minimize either the false positive or false negative discovery rate or to minimize the sum of these two rates. As suggested in the CMA publication, we used cut-off profiles designed to minimize the false negative discovery rate. Similar cut-off profiles for the liver, muscle and custom matrices were estimated according to the procedure described for Match [36]. For each PWM we generated 50000 random oligos by sampling from the PWM distribution. The PWM was then scored against these oligos with Match, and a cut-off threshold was chosen so that 90% of the oligos obtained a match score above this threshold. Since CMA is based on a discriminative model, it also requires a set of negative sequences along with the positive dataset. As negative data we selected 1000 bp promoter segments from 50 random housekeeping genes that were part of the default negative gene set included with the method's web service [37].

Availability and requirements

The web service for assessing composite motif discovery tools, as well as all the results from our benchmark test, is available at http://tare.medisin.ntnu.no/composite.

Abbreviations

ASP, average site performance (defined as (Sn + PPV)/2); bp, base pair; FN, false negative; FP, false positive; HMM, hidden Markov model; mCC, motif-level correlation coefficient; nCC, nucleotide-level correlation coefficient; PC, performance coefficient (defined as TP/(TP + FN + FP)); PPV, positive predictive value (defined as TP/(TP + FP)); PWM, position weight matrix; Sn, sensitivity (defined as TP/(TP + FN)); Sp, specificity. (defined as TN/(TN + FP)); TF, transcription factor; TFBS, transcription factor binding site; TN, true negative; TP, true positive.

Authors' contributions

GKS and OA conceived of the study. All authors participated in the design of the study. KK and GKS assembled the datasets. JJ implemented the web service and ran all the tests together with KK. KK drafted the manuscript. FD was the project supervisor. All authors helped revise and approved the final manuscript.

Additional material

Additional File 1

Dataset statistics. *This supplementary table includes information about the datasets and modules therein, the matrices used to represent the true motifs and the number of matrices in the PWM sets at various noise levels on the TRANSCompel dataset.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-9-123-S1.xls]

Acknowledgements

Kjetil Klepper, Jostein Johansen and Finn Drabløs were all supported by The National Programme for Research in Functional Genomics in Norway (FUGE) in The Research Council of Norway. Finn Drabløs was additionally supported by The Svanhild and Arne Must Fund for Medical Research. Osman Abul has been fully supported by an ERCIM fellowship.

References

1. Werner T: **Models for prediction and recognition of eukaryotic promoters.** *Mammalian Genome* 1999, **10(2):**168-175.
2. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA: **The Evolution of Transcriptional Regulation in Eukaryotes.** *Mol Biol Evol* 2003, **20(9):**1377-1419.
3. Sandve GK, Drabløs F: **A survey of motif discovery methods in an integrated framework.** *Biol Direct* 2006, **1:**11.
4. GuhaThakurta D, Stormo GD: **Identifying target sites for cooperatively binding factors.** *Bioinformatics* 2001, **17(7):**608-621.
5. Xing EP, Wu W, Jordan MI, Karp RM: **Logos: a modular bayesian model for de novo motif detection.** *J Bioinform Comput Biol* 2004, **2(1):**127-154.
6. Zhou Q, Wong WH: **CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling.** *Proc Natl Acad Sci USA* 2004, **101(33):**12114-12119.
7. Wasserman WW, Fickett JW: **Identification of regulatory regions which confer muscle-specific gene expression.** *J Mol Biol* 1998, **278(1):**167-181.
8. Chan BY, Kibler D: **Using hexamers to predict cis-regulatory motifs in Drosophila.** *BMC Bioinformatics* 2005, **6:**262.
9. Frech K, Danescu-Mayer J, Werner T: **A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter.** *J Mol Biol* 1997, **270(5):**674-687.
10. Aerts S, Van Loo P, Thijs G, Moreau Y, De Moor B: **Computational detection of cis-regulatory modules.** *Bioinformatics* 2003, **19(suppl 2):**ii5-14.
11. Stormo GD: **DNA binding sites: representation and discovery.** *Bioinformatics* 2000, **16(1):**16-23.
12. Johansson , Alkema WBL, Wasserman WW, Lagergren J: **Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm.** *Bioinformatics* 2003, **19(Suppl. 1):**i169-i176.
13. Kel AE, Konovalova T, Waleev T, Cheremushkin E, Kel-Margoulis OV, Wingender E: **Composite Module Analyst: a fitness-based tool for identification of transcription factor binding site combinations.** *Bioinformatics* 2006, **22(10):**1190-1197.
14. Sharan R, Ovcharenko I, Ben-Hur A, Karp RM: **CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments.** *Bioinformatics* 2003, **19(suppl 1):**i283-291.
15. Sze SH, Gelfand MS, Pevzner PA: **Finding weak motifs in DNA sequences.** In: *Proceedings of the Pacific Symposium on Biocomputing* 2002:235-246 [http://helix-web.stanford.edu/psb02/sze.pdf]. Lihue, Hawaii

16. Hu J, Li B, Kihara D: **Limitations and potentials of current motif discovery algorithms.** *Nucl Acids Res* 2005, **33(15):**4899-4913.
17. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Régnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z: **Assessing computational tools for the discovery of transcription factor binding sites.** *Nat Biotechnol* 2005, **23(1):**137-144.
18. Bailey TL, Noble WS: **Searching for statistically significant regulatory modules.** *Bioinformatics* 2003, **19(Suppl. 2):**ii16-ii25.
19. Perco P, Kainz A, Mayer G, Lukas A, Oberbauer R, Mayer B: **Detection of coregulation in differential gene expression profiles.** *Biosystems* 2005, **82(3):**235-247.
20. Sandve GK, Drabløs F: **Generalized composite motif discovery.** In In: *Proceedings of the 9th International Conference on Knowledge-Based Intelligent Information and Engineering Systems* Melbourne, Australia; 2005:763-769.
21. Burset M, Guigó R: **Evaluation of gene structure prediction programs.** *Genomics* 1996, **34(3):**353-367.
22. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E: **TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes.** *Nucl Acids Res* 2006, **34:**D108-D110.
23. Krivan W, Wasserman WW: **A predictive model for regulatory sequences directing liver-specific transcription.** *Genome Res* 2001, **11(9):**1559-1566.
24. Sandelin A, Alkema WBL, Engström P, Wasserman WW, Lenhard B: **JASPAR: an open-access database for eukaryotic transcription factor binding profiles.** *Nucl Acids Res* 2004, **32:**D91-D94.
25. Frith MC, Hansen U, Weng Z: **Detection of cis-element clusters in higher eukaryotic DNA.** *Bioinformatics* 2001, **17(10):**878-889.
26. Frith MC, Li MC, Weng Z: **Cluster-buster: finding dense clusters of motifs in DNA sequences.** *Nucl Acids Res* 2003, **31(13):**3666-3668.
27. Sinha S, van Nimwegen E, Siggia ED: **A probabilistic method to detect regulatory modules.** *Bioinformatics* 2003, **19(Suppl.1):**i292-i301.
28. Kielbasa SM, Gonze D, Herzel H: **Measuring similarities between transcription factor binding sites.** *BMC Bioinformatics* 2005, **6:**237.
29. Duret L, Bucher P: **Searching for regulatory elements in human noncoding sequences.** *Curr Opin Struct Biol* 1997, **7(3):**399-406.
30. Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, Bates K, Browne P, van den Broek A, Castro M, Cochrane G, Duggan K, Eberhardt R, Faruque N, Gamble J, Diez FG, Harte N, Kulikova T, Lin Q, Lombard V, Lopez R, Mancuso R, McHale M, Nardone F, Silventoinen V, Sobhany S, Stoehr P, Tuli MA, Tzouvara K, Vaughan R, Wu D, Zhu W, Apweiler R: **The EMBL nucleotide sequence database.** *Nucl Acids Res* 2005, **33:**D29-D33.
31. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** In: *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* 1994:28-36 [http://research1t.imb.uq.edu.au/bailey/t.bailey/papers/ismb94.pdf]. Stanford, California
32. Lenhard B, Wasserman WW: **TFBS: computational framework for transcription factor binding site analysis.** *Bioinformatics* 2002, **18(8):**1135-1136.
33. Krivan W, Wasserman WW: **Liver model, supplementary material.** [http://www.cisreg.ca/tjkwon].
34. Wasserman WW, Fickett JW: **Catalogue of Regulatory Elements.** [http://www.cbil.upenn.edu/MTIR/HomePage.html].
35. Aerts S, Van Loo P, Thijs G, Mayer H, de Martin R, Moreau Y, De Moor B: **TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis.** *Nucl Acids Res* 2005, **33:**W393-W396.
36. Kel AE, Gößling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E: **MATCH: a tool for searching transcription factor binding sites in DNA sequences.** *Nucl Acids Res* 2003, **31(13):**3576-3579.
37. Waleev T, Shtokalo D, Konovalova T, Voss N, Cheremushkin E, Stegmaier P, Kel-Margoulis O, Wingender E, Kel A: **Composite Module Analyst: identification of transcription factor binding site combinations using genetic algorithm.** *Nucleic Acids Res* 2006, **34(Suppl 2):**W541-545.
38. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A: **Information content of binding sites on nucleotide sequences.** *J Mol Biol* 1986, **188(3):**415-431.

Paper 6 O. Abul, F. Drabløs and G. K. Sandve. A methodology for motif discovery employing iterated cluster re-assignment. *Series on Advances in Bioinformatics and Computational Biology*. 2006;4:257–268.

Paper 7 O. Abul, G. K. Sandve and F. Drabløs. False discovery rates in identifying functional DNA motifs. In *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering (BIBE)*. 2007;387–394.

Are not included due to copyright

Accelerating Motif Discovery: Motif Matching on Parallel Hardware

Geir Kjetil Sandve[1], Magnar Nedland[2], Øyvind Bø Syrstad[1], Lars Andreas Eidsheim[1], Osman Abul[3], and Finn Drabløs[3]

[1] Department of Computer and Information Science,
Norwegian University of Science and Technology, Trondheim, Norway
sandve@idi.ntnu.no, {syrstad, eidsheim}@stud.ntnu.no
[2] Interagon A.S.,
Trondheim, Norway
magnar.nedland@interagon.com
[3] Department of Cancer Research and Molecular Medicine,
Norwegian University of Science and Technology, Trondheim, Norway
{osman.abul, finn.drablos}@ntnu.no

Abstract. Discovery of motifs in biological sequences is an important problem, and several computational methods have been developed to date. One of the main limitations of the established motif discovery methods is that the running time is prohibitive for very large data sets, such as upstream regions of large sets of cell-cycle regulated genes. Parallel versions have been developed for some of these methods, but this requires supercomputers or large computer clusters. Here, we propose and define an abstract module PAMM (Parallel Acceleration of Motif Matching) with motif matching on parallel hardware in mind. As a proof-of-concept, we provide a concrete implementation of our approach called MAMA. The implementation is based on the MEME algorithm, and uses an implementation of PAMM based on specialized hardware to accelerate motif matching. Running MAMA on a standard PC with specialized hardware on a single PCI-card compares favorably to running parallel MEME on a cluster of 12 computers.

1 Introduction

Computational discovery of motifs in biological sequences has many important applications, the best known being discovery of transcription factor binding sites (TFBS) in DNA and active sites in proteins. More than a hundred methods have been developed for this problem, all with different strengths and characteristics. Methods that use probabilistic motifs (typically PWMs) are often favored because of their high expressibility. One of the best known and most widely used methods is MEME [1]. MEME is a flexible tool that uses Expectation Maximization (EM) to discover motifs as position weight matrices (PWMs) in both proteins and DNA.

One of the main limitations of current PWM-based motif discovery methods is that the running time is prohibitive for large datasets such as upstream regions

of large sets of cell-cycle regulated genes. Parallel versions have been developed for some methods, for instance the paraMEME [2] version of MEME, but this typically requires supercomputers or computer clusters. Specialized hardware, such as Field Programmable Gate Arrays (FPGAs), may be a very viable alternative to this. FPGAs have previously been used in bioinformatics for instance to accelerate homology search [3], multiple sequence alignment [4] and phylogeny inference [5].

In this paper, we propose and define an abstract module PAMM (Parallel Acceleration of Motif Matching). Proposing the PAMM module serves two purposes. Firstly, it introduces acceleration of motif matching by parallel hardware to the motif discovery field. Secondly, PAMM serves as an interface between the development of modules for parallel matching of motifs and the development of algorithms that can make use of parallel motif matching.

As a first implementation of our methodology, we propose a method MAMA (Massively parallel Acceleration of the Meme Algorithm) that accelerates MEME by the use of an existing pattern matching hardware called the Pattern Matching Chip (PMC) [6]. The PMC can match a subset of regular expressions with massive parallelization[1]. Since this chip was not intended for weighted pattern matching, some transformations are needed when representing and matching motifs. Nonetheless, with these transformations in place we achieve very efficient matching of PWMs against sequences. Running MAMA on a standard PC with specialized hardware on a single PCI-card compares favorably to running paraMEME on a cluster of 12 computers.

2 Parallel acceleration of motif matching

An ever increasing number of computing platforms offer capabilities for parallel execution of programs. Specialized hardware exists to relieve the main CPU of specific tasks, and FPGAs allow the creation of modules for application specific hardware acceleration. To allow the field of motif discovery to realize the full potential of modern computing hardware, the algorithms need to take advantage of this. Here we propose and define an abstract module PAMM that can be used for accelerating motif discovery by matching motifs against sequences in parallel. The purpose of PAMM is to serve as an interface between development of modules for parallel matching of motifs and the development of algorithms that can make use of parallel motif matching. An overview of the PAMM module is presented in Figure 1. The input to PAMM is a set of motifs M and a set of sequences S, while the output depends on the requirements of the algorithm in question. Each motif is represented as a matrix.

As the figure shows, there are two main parts in the PAMM module; a motif matcher and a post processing unit. The motif matcher calculates the match scores for each motif, while the post processing unit refines the results.

[1] More information at http://www.interagon.com

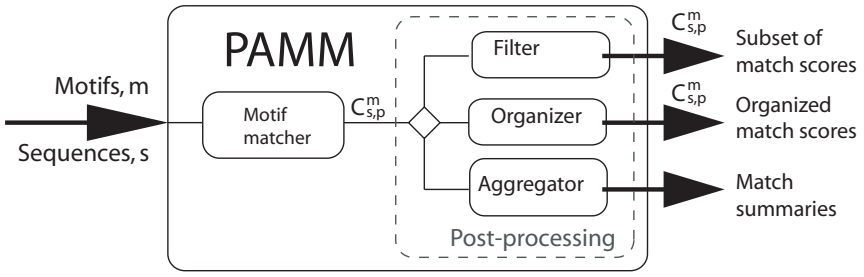


Fig. 1. The structure of the PAMM module.

2.1 Motif matching

The core of a PAMM implementation is a motif matcher that determines match scores $c^m_{s,p}$ for each motif m when aligned at each position p in each sequence s. As the number of motifs and sequences that can be processed in parallel will be limited in any practical implementation of the module, the algorithm must partition the inputs accordingly.

As a standard set-up we propose that a limited number of motifs are first loaded into the PAMM, and that sequence data are then streamed through. The motif matcher will continually calculate match scores for each motif against the sequences. When all motifs have been matched against the complete sequence data, a new set of motifs can be loaded into the module and matched against the sequences. As this means that the same sequences will typically be streamed through the PAMM many times, practical implementations could have an option to store a limited amount of sequence data in local memory to further accelerate matching and reduce bandwidth usage. This set-up is illustrated in Figure 2(a).

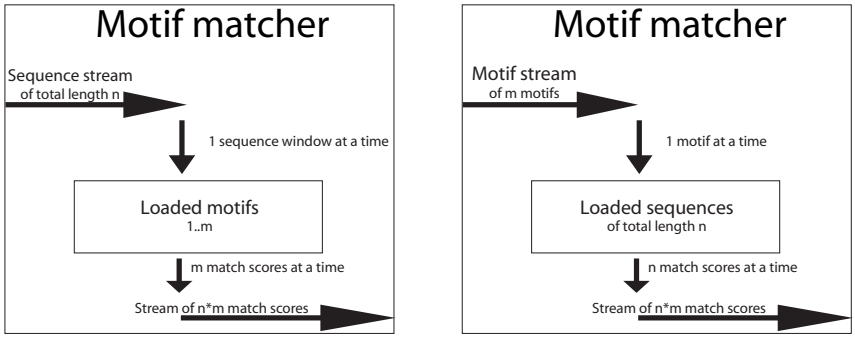
An alternative set-up could be to first load a limited amount of sequence data into the PAMM, and then stream motifs through the module. This could be an effective solution for cases with relatively short sequence data and large number of motifs. This setup is illustrated in Figure 2(b).

2.2 Post-processing of match scores

The number of results from the motif matcher is $|M| * |S|$, where M is the set of motifs and S is the set of all sequence data. This potentially large amount of results must somehow be processed by the system. By incorporating post processing, the number of results returned from a PAMM implementation can be reduced substantially. This reduces result processing in the algorithm module, as well as bandwidth requirements in the case where the PAMM and algorithm modules reside on different (sub)systems.

We envision three main branches of post processing for PAMM implementations; organizing, filtering, or aggregating (or a combination of these).

An organizing post processor organizes the results in a way that facilitates efficient further processing of results outside the PAMM module. It could for



(a) Motif matcher with loaded motifs and sequences streamed through

(b) Motif matcher with loaded sequences and motifs streamed through

Fig. 2. Two possible set-ups of the motif matcher

instance return the match scores sorted by value. Although this does not decrease bandwidth usage, it may allow the CPU to process the results more efficiently.

A filtering post processor filters out uninteresting match scores to save processing time outside the PAMM module. It could for instance make the PAMM return only match scores above a threshold given for each motif. Although this discards some information, our own experiments (not presented here) show that the normalized match scores typically follow a distribution where most sequence offsets have a negligible likelihood of being motif locations. In combination with an organizing post processor, the k highest match scores could be returned, or all scores at most l lower than the highest match score.

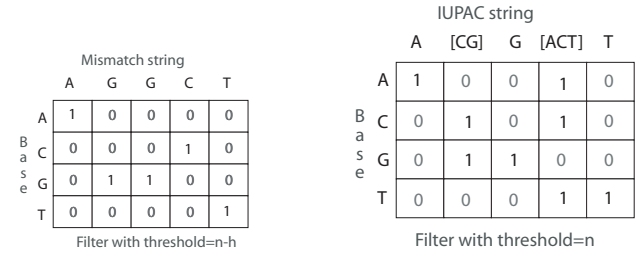
An aggregating post processor is tailored to a specific motif discovery algorithm and may be particularly (computationally) effective. If the PAMM is to be used in connection with stochastic optimization methods like Gibbs sampling, it can be set to return one sequence offset per sequence, with offsets chosen randomly based on the normalized probabilities of motif occurrences. Alternatively, if the PAMM is used in connection with EM methods, a new motif may be constructed from the match scores directly in hardware (maximization step of EM). This new motif would represent a weighted average of every window in the sequences, with windows weighted by the match score of a previous motif.

2.3 Motif representations

The representation of a motif in PAMM is as a motif matrix $m \in M$ with element values $m_{i,x}$, where i is motif position and x is a symbol from the alphabet, *i.e.* $x \in \{A, C, G, T\}$. The element values represent individual scores for each symbol x from the alphabet at each position i in the motif. The motif is aligned against sequences as a sliding window. For a given alignment at position p in sequence s, the score of motif position i is $m_{i,x}$, where x is the symbol at position $p+i$ in sequence s. The match score $c^m_{s,p}$ of the motif is the sum of scores at each motif

position. This motif representation maps directly to PWMs (log-likelihood or log-odds) that are often used for motif discovery.

In addition to PWMs, strings allowing mismatches [7, 8] (a consensus string allowing a certain hamming distance to an occurrence) and IUPAC strings [9, 10] (strings of characters and character classes) are commonly used models in motif discovery. Both of these can be represented by a motif matrix. For a motif matrix representing a mismatch string, elements $m_{i,x}$ corresponding to the consensus symbol at a position have value 1, and all other matrix elements are 0. Matrix scores $c >= n - h$ corresponds to a hit for the mismatch expression, where n is motif length and h is allowed number of mismatches. This is shown in Figure 3(a). For a motif matrix representing an IUPAC string, elements $m_{i,x}$ corresponding to symbols in the character class at a position are valued 1, and all other matrix elements are 0. Matrix scores $c = n$ corresponds to a hit for the IUPAC expression. This is shown in Figure 3(b)



(a) mismatch strings (b) IUPAC strings

Fig. 3. Matrix representation of discrete motif models

Other and more complex motif models could also be represented with such a matrix (variants of Markov models and bayesian trees have for instance been used in motif discovery). This will typically require a larger motif matrix and some preprocessing of the sequence data. Such preprocessing could be done by additional hardware modules within the PAMM. The generality of the matrix representation makes it suitable as a standard motif representation for the PAMM module.

3 Practical implementation

This section describes a motif discovery algorithm that uses a PAMM implementation to accelerate motif matching. To explore the potential of PAMM in motif discovery, we have used available hardware (PMC) to implement a PAMM module.

We have analyzed the running time of the MEME algorithm and developed a motif discovery algorithm MAMA based on MEME that uses the PAMM

implementation for motif matching in the performance-critial parts. As this is a first implementation and a proof-of-concept, we have only made adjustments to the MEME algorithm that make it run faster while not altering which motifs are discovered.

3.1 Motif discovery using the PAMM module

MEME is a a motif discovery algorithm based on Expectation Maximization (EM) that match motifs against sequences in the expectation step. Profiling of the MEME implementation showed that matching initial motifs (starting points) against sequences consumed most of the total running time. We have therefore made the necessary adjustments to allow parallel acceleration of this first iteration of MEME.

MEME running time EM was first used for motif discovery by Lawrence *et al.* [11]. As EM is easily trapped in local minima, they used several random starting points (initial PWMs) for EM. This was improved in the MEME algorithm of Bailey and Elkan [1], which use every substring of a given length in the data set as starting point. More specifically, for every substring a PWM is constructed with a fixed weight to the elements in the matrix corresponding to symbols in the substring, and another, lower fixed weight to the other elements. As this typically amounts to very many starting points, they run EM for one iteration from each starting point, and then only continue with those PWMs that seem most promising.

Inspection of the MEME implementation[2] shows that specialized code is used for this first iteration, using dynamic programming to exploit overlap between starting points. PWMs generated from each substring in the data set are first matched against the sequences (expectation step). For each PWM, the sequence offsets are then sorted by match score and the k highest scoring offsets used to generate a PWM candidate for the next iteration (maximization step). Finally, the significance values for all candidate PWMs are computed, and the most significant ones kept and refined (iterated until convergence).

MEME tries a very large number of starting points in the first iteration, and only continues with a few most promising motifs. Our profiling showed that the first iteration amounted to around 97% of total running time in our tests, using data sets supplied with MEME, the TCM model, and otherwise default parameters. Although this number might vary for different test cases and parameter settings, it shows that the first iteration is the bottleneck when it comes to running time of the algorithm. Furthermore, matching motifs against sequences and sorting offset scores dominate the running time.

Exploration of starting points As the first iteration dominates the running time of MEME, we have focused on accelerating this part. More specifically, we

[2] Version 3.5.0, downloaded from http://meme.nbcr.net/downloads/

have used the PAMM module to match PWMs and sort offset scores in the first iteration, and left the remaining parts of MEME unaltered.

Exploration of starting points differs a bit from all other iterations in MEME. First, all matrix elements of starting point PWMs has one of two values: a fixed high value for elements corresponding to the symbol of the substring it is based on, and a fixed low value for every other element. Thus, all sequence windows at a given hamming distance from the substring a PWM is based on will get the same PWM score. Ranking of sequence offsets based on PWM score will therefore in the first iteration be equal to ranking of sequences windows based on hamming distance. Secondly, in a general EM iteration each sequence window is used in the maximization step (weighted by the expectation values). When maximizing the PWMs in the first iteration, however, only the sequence windows corresponding to the top k expectation values are used.

These properties are exploited in MAMA by using a PAMM implementation that represents motifs efficiently and returns sequence offsets sorted by match score. The motif discovery algorithm thus only needs to consider the first k sequence offsets returned by the PAMM implementation.

3.2 Implementation of the PAMM module

We have implemented PAMM using available hardware for parallel pattern matching. This hardware, The Pattern Matching Chip (PMC) [6], is a multiple instruction single data (MISD) parallel hardware on a PCI card. One PCI-card can match up to one thousand simple patterns against 100 MB of sequences per second, and it is quite straightforward to set up searches. Because of its efficiency and ease of use, we have used the PMC for this first implementation of the PAMM module. The PMC implementation covers both motif matching and organization of match scores.

Motif matching As the PMC only supports binary matching of patterns, and integer summation, the PWM match scores need to be discretized. The discretization is based on the fact that the log-likelihood for any base pair in any location is in the interval $\left[log(\frac{\beta}{n+4\beta}), log(\frac{n+\beta}{n+4\beta})\right]$, where β is the pseudo-count and n is the number of motif sites, given as parameters to MEME. Instead of using a fixed granulation of the interval, we define a granulation parameterized with ϵ. Then, each value $m_{i,x}$ in the PWM m is represented by a number $c_{i,x} = \lfloor \frac{log(m_{i,x}) - log(\frac{\beta}{n+4\beta})}{\epsilon} \rfloor$ of processing elements (PEs) in the specialized hardware. The number of PEs matching a symbol of the alphabet at a given position is thus proportional to the log-likelihood value of that symbol at that position. When the PWM is aligned with a sequence window, the sum of PE match scores at a motif position then corresponds to the *score* at that position. Note that since only one of the four nucleotides can match at a position, the other three do not contribute to the *score*. Furthermore, as PWM log-likelihood is the the sum of log-likelihoods for each position, the total PWM score is given by the sum of *scores* of all positions.

Two optimizations are worth mentioning. First, if the minimum score $c_i = min_x(c_{i,x})$ at a given position i is higher than zero, we may subtract c_i from each score value at that position, and then add c_i to the score after the search. Secondly, if $c = max_{i,x}(c_{i,x})$ is the maximum score value of the motif, and more score values are close to c than are close to zero, we then use transformed score values $c'_{i,x} = c - c_{i,x}$ and compute total PWM score as: $c \cdot I - \sum_i \sum_x c'_{i,x}$, where i runs over all I positions of m. Both optimizations give equivalent results to the basic method while using less PEs on the PMC, thus allowing more matrices to be matched simultaneously.

The discretization method considered above can be used generally for matching arbitrary PWMs against sequences. The approximation accuracy clearly depends on the granulation parameter ϵ. As discussed in section 3.1, the PWMs are regular in the first iteration of MEME. Motif matching can then be done with degenerate use of discretization, thereby avoiding approximation problems. To ensure that MAMA gives the same results as MEME, we have therefore only used hardware-acceleration in the first iteration, and used a standard software solution for motif matching in the remaining iterations. Since the running time of MEME is strongly dominated by the first iteration, we still achieve significant speed-ups.

Organizing match scores As the PMC provides massive parallelity, we are able to calculate expectation values for many PWMs in parallel. We also use this parallelity to scan each PWM against the sequences several times with different hit thresholds. By searching with several thresholds in parallel, we can make the PMC return sequence offsets sorted by decreasing match score. This corresponds to a PAMM organizing module for post-processing of match scores, and avoids CPU-intensive sorting of offsets after the expectation step.

4 Results

We have compared the performance of our hardware accelerated version MAMA with the CPU based version of MEME on data sets of different sizes. On all test referred to here we have used the TCM model of MEME, which is the most general model and presented as the main model in the original MEME article [1]. We ran our tests with the following hardware configuration:

- MAMA: 2.8 Ghz Pentium4 PC with 1 GB memory and the specialized hardware on a single PCI card.
- MEME: 2.8 Ghz Pentium4 PC with 1 GB memory.
- ParaMEME: a cluster of 12 computers, each 3.4 Ghz Pentium4 PC with 1 GB memory.

We evaluated the performance of MAMA on the largest data set (minidrosoph) supplied with MEME and on 5 data sets of human promoter regions, consisting of from 100 to 1600 sequences of 5000 base pair length from cell cycle

regulated genes (J.P.Diaz, in preparation). Data sets, sizes and running times are given in Table 1 for both MEME, paraMEME and MAMA. We see that MAMA gives a significant speed-up compared to MEME on all datasets, and that the speed-up increases with data set size. On the 1 Mbp (Million base pairs) data set, MAMA is more than twenty times as fast as MEME, and on the 8 Mbp data set it is even four times as fast as paraMEME on the 12-computer cluster. For all data sets, standard MEME and the hardware-accelerated version MAMA discovers the same motifs.

Table 1. Results for MEME, paraMEME and MAMA on 6 data sets.

		Running time (hours)		
Data set	Size (Mbp)	MEME	paraMEME	MAMA
mini-drosoph	0.5	2.6	0.19	0.27
hs_100	0.5	2.7	0.20	0.23
hs_200	1	11	0.87	0.50
hs_400	2	104	3.6	1.7
hs_800	4	X[3]	15	6.4
hs_1600	8	X[3]	64	13

5 Discussion and conclusion

We have in this paper proposed an abstract module PAMM for parallel hardware-acceleration of motif discovery. This module could be used for acceleration of many different motif discovery methods. The acceleration could be especially large if post-processing of match scores is tailored to a specific algorithm.

As an exemplification and proof-of-concept we have developed a version of the MEME algorithm called MAMA that uses available hardware to implement a PAMM module. As shown in section 4, MAMA achieves a speed-up of more than a factor of 10 as compared to MEME on a single CPU. Our working implementation thus shows that the PAMM module indeed has a potential.

Furthermore, our work shows examples of both problematic issues and potential rewards in connection with hardware acceleration of algorithms within bioinformatics. Since we have implemented weighted motif matching on hardware that was not specifically built for that purpose, we had to do some transformations of the problem. The issues and solutions with regards to discretization and parallelization are relevant for many algorithmic solutions involving specialized hardware.

A natural continuation of the work presented in this paper is to develop a FPGA-based implementation of PAMM. Such a solution would be more readily available for practical use and further refinement by the scientific community.

[3] Not tested due to excessive running times

It could potentially also give even higher speed-ups. On the other hand, such a solution presumes a solution of representing PWMs on FPGA that is both efficient and flexible. We have ongoing work in this direction that shows promising results.

References

1. Bailey, T.L., Elkan, C.: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol **2**(1553-0833) (1994) 28–36
2. Grundy, W.N., Bailey, T.L., Elkan, C.P.: ParaMEME: a parallel implementation and a web interface for a DNA and protein motif discovery tool. Comput Appl Biosci **12**(0266-7061 (Print)) (1996) 303–10
3. Yamaguchi, Y., Miyajima, Y., Maruyama, T., Konagaya, A.: High speed homology search using run-time reconfiguration. In: Lecture Notes in Computer Science. Volume 2438. (2002) 281–291
4. Oliver, T., Schmidt, B., Nathan, D., Clemens, R., Maskell, D.: Using reconfigurable hardware to accelerate multiple sequence alignment with ClustalW. Bioinformatics **21**(16) (2005) 3431–2
5. Mak, T.S.T., Lam, K.P.: Embedded computation of maximum-likelihood phylogeny inference using platform FPGA. In: Computational Systems Bioinformatics Conference (CSB), 2004 IEEE. (2004) 512–514
6. Halaas, A., Svingen, B., Nedland, M., Sætrom, P., Snøve Jr., O., Birkeland, O.R.: A recursive MISD architecture for pattern matching. IEEE Trans. Very Large Scale Integr. Syst. **12**(7) (2004) 727–734
7. Marsan, L., Sagot, M.F.: Extracting structured motifs using a suffix tree-algorithms and application to promoter consensus identification. In: RECOMB '00: Proceedings of the fourth annual international conference on Computational molecular biology, New York, NY, USA, ACM Press (2000) 210–219
8. Blanchette, M., Tompa, M.: Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Res **12**(5) (2002) 739–48
9. Sinha, S., Tompa, M.: YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. Nucleic Acids Res **31**(13) (2003) 3586–8
10. Bortoluzzi, S., Coppe, A., Bisognin, A., Pizzi, C., Danieli, G.: A multistep bioinformatic approach detects putative regulatory elements in gene promoters. BMC Bioinformatics **6**(1) (2005) 121
11. Lawrence, C.E., Reilly, A.A.: An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins **7**(1) (1990) 41–51