Regulatory motif discovery using PWMs and the architecture of eukaryotic core promoters

Holger Hartmann

SIGILL. UNIVERSITAT. LUDOVIC. MAXIM.

München 2012

Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig-Maximilians-Universität München

Regulatory motif discovery using PWMs and the architecture of eukaryotic core promoters

Holger Hartmann

aus

Bad Friedrichshall, Deutschland

2012

Erklärung:

Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Herrn Prof. Dr. Patrick Cramer betreut.

Eidesstattliche Versicherung:

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, am 29. März 2012

Holger Hartmann

Dissertation eingereicht am 29. März 2012

1. Gutachter: Prof. Dr. Patrick Cramer
2. Gutachter: Dr. Johannes Söding

Mündliche Prüfung am 8. Mai 2012

Acknowledgments

During the past years of work on this thesis, I was influenced, assisted and motivated by many people. Having the opportunity to thank all of them is a pleasure not to be missed:

First, I would like to thank Dr. Johannes Söding for being an amazing supervisor, offering me scientific freedom and constant support at the same time. I would like to thank Prof. Dr. Patrick Cramer for being my doctoral supervisor and the possibility to work in many fruitful collaborations. I am also very grateful to Prof. Dr. Ulrike Gaul, Dr. Dietmar Martin, Prof. Dr. Klaus Förstemann, and Prof. Dr. Roland Beckmann for offering their time as members of my dissertation committee.

Furthermore, I would like to thank all students I had the pleasure to supervise (Claudia Gugenmus, Alexej Grjasnow, Lukas Utz, Anja Kiesel, Sebastian Luehr, and Mark Heron) for all input and help during my projects and the knowledge I gained from their analyses. My deepest thanks to all my fellow PhD students and scientists at the Gene Center Munich (Matthias Siebert, Eckhart Guthöhrlein, Michael Remmert, Andreas Biegert, Armin Meier, Maria Hauser, Jörn Marialke, Phillipp Torkler, Björn Schwalb, Sebastian Dümcke, Theresa Niederberger, Andreas Hauser – to name only some) for general and specific support, for discussions, all the fun we had together, the delicious cakes every week, and especially for breathtaking blobby games.

Finally, I thank my parents and Mr & Mrs Petre for their unwavering support throughout all these years, and foremost Cristina, for her love and patience when time was rare, for believing in me when I had doubts, for making things possible that others think impossible.

Summary

To analyze gene regulatory networks, the sequence-dependent DNA/RNA binding affinities of proteins and non-coding RNAs are crucial. Often, these are deduced from sets of sequences enriched in factor binding sites. Two classes of computational approaches exist. The first describe binding motifs by sequence patterns and search the patterns with highest statistical significance for enrichment. The second class use position weight matrices (PWMs) that describe the binding site by a more powerful probabilistic model. They cannot maximize the statistical significance of enrichment but maximize a likelihood instead. In this thesis, I present XXmotif (eXhaustive evaluation of matriX motifs), the first PWM-based motif discovery method that directly optimizes the statistical significance of enrichment. It computes 100,000s of single-site P-values for thousands of candidate PWMs during the refinement. For this purpose, we developed an efficient branch-and-bound algorithm that calculates exact single-site P-values using an eighth-order background model. This approach allows us to naturally combine P-values for motif enrichment, conservation, and localization. When tested on ChIP-chip/seq, miRNA knock-down, and co-expression data sets from yeast and metazoans, XXmotif outperforms state-of-the-art tools, both in numbers of correctly identified motifs and in the qualities of PWMs. In segmentation modules of *D. melanogaster*, XXmotif detects the known key regulators and several new motifs (Hartmann et al., 2012a).

The core promoter is defined as the region where the basal transcription machinery assembles to initiate transcription. Evidence is accumulating for the involvement of core promoters in transcriptional regulation. Several core promoter elements are known in eukaryotes (e. g., TATA-box, Initiator, BRE, DPE). However, in many genes only one or no motif has been identified, prompting the question how the transcription machinery finds the core promoter. Yet unknown motifs or the incorporation of physical properties of the DNA might be an explanation. Our *de novo* motif analysis using XXmotif reveals new, highly significant motifs that are conserved and enriched in promoter regions. In human core promoters, XXmotif reports in a single run most previously described and four novel motifs sharply peaked around the transcription start site (TSS), among them a novel Initiator motif similar to the fly version. Applied on core promoter regions in *D. melanogaster*, XXmotif reveals 12 known and 7 novel elements, all highly significant, conserved, and non-randomly localized with respect to the TSS. In addition, we identified four different classes of core promoters. Each class consists of a defined set of core promoter elements that build a strong indicator

for the width of the TSS cluster, the degree of gene regulation, the stallability, as well as the minimum and maximum transcription rate (Hartmann et al., 2012b).

An interactive web server for XXmotif is available at `http://xxmotif.genzentrum.lmu.de`. It provides (a) free binaries and sources for the command line version of XXmotif, (b) a list of all significantly overrepresented motif PWMs with web logos, number of occurrences, and E-values, (c) a graph with color-coded boxes indicating the positions of user selected motifs in the input sequences, (d) a histogram of the overall positional distribution for user selected motifs, and (e) a page for each motif with all significant motif occurrences, their P-values for enrichment, conservation, and localization, their sequence contexts, as well as their coordinates within the input sequences (Luehr et al., 2012).

Contents

Acknowledgments vii

Summary ix

I. XXmotif - *P*-value Based Motif Identification 1

1. Introduction to Motif Finding 3
- 1.1. Motif Finding Problem 4
- 1.2. Motif Models 5
  - 1.2.1. Consensus string 5
  - 1.2.2. Position weight matrix (PWM) 6
- 1.3. Approaches to Solve the Motif Finding Problem 7
  - 1.3.1. Enumerative approach 7
  - 1.3.2. Iterative approach 7

2. Materials and Methods 9
- 2.1. Availability 9
- 2.2. Statistical Framework 9
  - 2.2.1. Calculating the significance of an l-mer – Background model 9
  - 2.2.2. Calculating the significance of a binding site given a PWM 11
  - 2.2.3. Calculate the significance of a set of binding sites 14
  - 2.2.4. Correcting P-values for multiple testing 15
  - 2.2.5. Calculating conservation P-values 16
  - 2.2.6. Weighted combination of P-values 17
  - 2.2.7. Calculating positional P-values 17
- 2.3. Workflow of XXmotif 24
  - 2.3.1. XXmasker 24
  - 2.3.2. Seeds phase 25
  - 2.3.3. Extension phase 25
  - 2.3.4. Merging phase 28
  - 2.3.5. Refinement phase 29

2.4. Overview of the Used Published Motif Finders 30
  2.4.1. PRIORITY 30
  2.4.2. MEME 31
  2.4.3. Weeder 33
  2.4.4. ERMIT 33
  2.4.5. AMADEUS 34

3. XXmotif Web Server **35**
3.1. Overview of Web Servers 35
3.2. Input 36
3.3. Output 36
3.4. Documentation 39
3.5. Implementation 40
3.6. Conclusion 40

II. Analysis of Transcription Factors and Core Promoters **41**

4. Transcription Factors **43**
4.1. Overview 43
4.2. Families of DNA-Binding Domains 45
  4.2.1. Helix-turn-helix motif 45
  4.2.2. Homeodomain 46
  4.2.3. C_2H_2 zinc finger 46
  4.2.4. C_6-zinc cluster family 47
4.3. Experimental Techniques for in Vitro Analysis of TFBSs 48
  4.3.1. Protein-binding microarrays 49
  4.3.2. HT-SELEX 49
  4.3.3. HiTS-FLIP 50
4.4. Experimental Techniques for in Vivo Analysis of TFBSs 50
  4.4.1. DNase I hypersensitive sites 51
  4.4.2. Chromatin immunoprecipitation (ChIP) 51

5. Performance of XXmotif on TFBSs **53**
5.1. Benchmarks 53
  5.1.1. Drabløs Benchmark 53
  5.1.2. Motif sensitivity benchmark 55
  5.1.3. PWM quality benchmark 58
  5.1.4. Metazoan benchmark 62
5.2. Regulatory Motifs for Early Embryo Segmentation in Flies 64
5.3. Human Core Promoter Motifs 66

5.4. Discussion 69

6. Analysis of the Motif Architecture within Fly Core Promoters 73
- 6.1. Introduction to Core Promoters in *D. melanogaster* 73
- 6.2. Drosophila Core Promoters Fall Into Four Classes 75
  - 6.2.1. TSS cluster width 75
  - 6.2.2. Gene sets 76
  - 6.2.3. Identification of core promoter elements 78
  - 6.2.4. Core promoter elements allow for the prediction of gene properties . 82
  - 6.2.5. Core promoter elements belong to specific architectures 84
  - 6.2.6. Each class of core promoters has defined physical properties 85
- 6.3. Summary 87
- 6.4. Materials and Methods 88
  - 6.4.1. Tag clustering 88
  - 6.4.2. Optimization of minimal score thresholds 88
  - 6.4.3. Conservation scores 88

III. Appendix 89

A. Detailed Performance of XXmotif 91
- A.1. Drabløs Benchmark 91
- A.2. Motif Sensitivity Benchmark 92
- A.3. PWM Quality Benchmark 93
- A.4. Metazoan Benchmark 94
- A.5. Regulatory Motifs For Early Embryo Segmentation in Flies 95

B. Drosophila Core Promoters 97
- B.1. TSS Cluster Width 97
- B.2. Motif Conservation 98
- B.3. Core Promoter Elements vs. Developmental Stages 102
- B.4. Class 4 Core Promoter Elements 103

Bibliography 105

Curriculum Vitae 116

List of Figures

1.1. Example sequence logo 6

2.1. Splitting and recombination of the branch-and-bound P-value calculation . 13
2.2. Weighted combination of P-values 18
2.3. Overview of XXmotif with its main stages 25
2.4. Main data structures in XXmotif elongation phase 27

3.1. XXmotif web server pipeline for job submission 37
3.2. Multi distribution plot 38
3.3. Motif distribution plot 40

4.1. Base recognition in major and minor groove of DNA 43
4.2. Helix-Turn-Helix motif 45
4.3. C_2H_2 zinc finger motif 47
4.4. C_6-zinc cluster motif 48

5.1. Drabløs Benchmark 55
5.2. Sensitivity Benchmark 57
5.3. Dependence on Model Order in Sensitivity Benchmark 58
5.4. ROC curve of a PWM 59
5.5. PWM quality benchmark 61
5.6. Metazoan benchmark Top 4 63
5.7. Runtime of tested tools on the metazoan benchmark 64

6.1. Comparison of MAD score and SI score 76
6.2. Gene Sets 77
6.3. Motif conservation 81
6.4. Correlations between CPEs and gene sets 84
6.5. Correlations between CPEs and gene sets 84
6.6. Each class of core promoters possesses distinct dinucleotide patterns 86
6.7. Four classes of core promoters separated by cluster width and expression strength 87

A.1. Detailed results of the Drabløs benchmark 91

A.2. PWM quality benchmark without using conservation 93
A.3. Metazoan benchmark Top 1 94

B.1. SI score distribution 97
B.2. Conservation of all motifs 101
B.3. Correlations between CPEs and developmental stages 102

List of Tables

1.1. Example motif finding problem 4
1.2. The IUPAC code 5

5.1. Motifs in fly segmentation modules discovered by XXmotif 65
5.2. Motifs in human core promoters discovered by XXmotif 68

6.1. Motifs in fly core promoters discovered by XXmotif 79

A.1. Detailed Results of Sensitivity Benchmark 92
A.2. Used *cis*-regulatory modules in fly segmentation 95

B.1. GO analysis of Class 4 genes 103
B.2. Genes containing RDPE 104

Abbreviations

BLAST	basic local alignment search tool
BP	broad peak
BRE	TFIIB recognition element
CAGE	cap analysis of gene expression
ChIP	chromatin immunoprecipitation
CPE	core promoter element
DBD	DNA-binding domain
DIP	DNA-immunoprecipitation
DNA	deoxyribonucleic acid
DPE	downstream promoter element
DRE	DNA replication-related element
DREF	DRE binding factor
EM	expectation maximization
EMSA	electrophoretic mobility shift assay
ENCODE	encyclopedia of DNA elements
GO	gene ontology
HiTS-FLIP	high-throughput sequencing - fluorescent ligand interaction profiling
HMM	hidden Markov model
HT-SELEX	high-throughput SELEX
HTH	helix-turn-helix
INR	Initiator
IUPAC	international union of pure and applied chemistry

MAD	mean absolute deviation from the median
MEME	multiple EM for motif elicitation
mops	multiple occurrence per sequence
MTE	motif ten element
NP	narrow peak
oops	one occurrence per sequence
pAUC	partial area under ROC curve
PBM	protein binding microarray
PCR	polymerase chain reaction
PDB	protein data bank
PIC	pre-initiation complex
PWM	positional weight matrix
RNA	ribonucleic acid
ROC	receiver operator curve
SELEX	systematic evolution of ligands by exponential enrichment
SI	shape index
TAF	TBP-associated factor
TBP	TATA-binding protein
TF	transcription factor
TFBS	transcription factor binding site
TRF	TBP-related factor
TSS	transcription start site
UML	unified modeling language
XXmotif	eXhaustive evaluation of matriX motifs
zoops	zero or one occurrence per sequence

Part I.

XXmotif - *P*-value Based Motif Identification

1. Introduction to Motif Finding

Understanding the regulatory control mechanisms involved in transcriptional networks is one of the main objectives in computational biology (Wasserman and Sandelin, 2004). Although the decryption of the entire network is still a distant hope, major progress has been made in understanding the key regulators of the network, the transcription factors (TFs). These proteins bind to specific sequence elements in the DNA depending on their intrinsic preferences (see Chapter 4), leading to an activation or repression of the target gene. Revealing these sequence preferences, in the following referred to as sequence motifs, is necessary to predict the binding sites and binding strengths of every TF within an organism. Subsequently, these binding strengths can be used in combination with additional and partly yet unknown sequence properties, to predict the expression strength of every gene.

Motif finding tools have the objective to reveal these sequence motifs. Analyzing data sets enriched in a TF of interest, they aim for detecting sequence patterns more frequent than expected by chance. Experimental proceedings in determining binding sites of TFs *in vitro* or *in vivo* (see Sections 4.3 and 4.4) have led to a tremendous growth of data sets enriched in binding sites, making motif finding tools more valuable than ever. Next to over representation, analyzing divergences between related species might increase the accuracy of a motif search. Since functional regions within the DNA mutate at lower rates than nonfunctional regions (Wang and Stormo, 2003), conserved nucleotides indicate biological importance. In addition, some recently developed algorithms are able to incorporate nucleosome occupancies or experimentally determined binding profiles (Narlikar et al. (2007), Georgiev et al. (2010)) into their *de novo* motif analysis.

This chapter provides a general definition of the motif finding problem and presents frequently used motif models and approaches to solve it. Furthermore, the frequently used sequence logo to graphically visualize the binding affinities of a transcription factor is described.

1.1. Motif Finding Problem

Formally, the motif finding problem can be defined as follows:

> Given a set of X strings of symbols from some alphabet $\mathcal{A}$ that are essentially random except that in some of them is a motif site of length W sampled from some unknown set S, find a model that best describes S.

An example of such a problem is given in Table 1.1. It consists of 5 input sequences of each length 30 over the alphabet $\mathcal{A}$ of all 26 English lower case letters. Predicted motif sites illustrating a sample of S are shown in upper case. But, given the definition, the detection of all motif sites is not sufficient to solve the motif finding problem, it is also necessary to define a model of S able to generalize to unseen instances (see Section 1.2).

Example Problem	Possible Result
`kafemotafsitemakilggahlleghejv`	`kafeMOTAFSITEmakilggahlleghejv`
`qigjbvkajlmotifsiaeajkqlebabbe`	`qigjbvkajlMOTIFSIAEajkqlebabbe`
`katellhallbergebkalhheerokamel`	`katellhallbergebkalhheerokamel`
`allhmenicatlexmooramofifsitefe`	`allhmenicatlexmooraMOFIFSITEfe`
`hmotifsiteekglckhagekdgotojall`	`hMOTIFSITEekglckhagekdgotojall`

predicted sites={`motafsite`, `motifsiae`, `mofifsite`, `motifsite`}

Table 1.1.: Example motif finding problem. Predicted motif sites are shown in upper case.

In the definition, the motif finding problem is described as a zero or one occurrence per sequence (zoops) problem. However, if for an input set prior knowledge is available that all sequences contain exactly one motif occurrence, the motif finding problem can be reformulated to a one occurrence per sequence (oops) problem. Similarly, if more than one motif site might be present within the input sequences, the motif finding problem can be reformulated to the most general case, the multiple occurrence per sequence (mops) problem.

Since nearly no data set exists without false positives, the development of an algorithm solving only the oops problem is almost useless for biological problems. The zoops problem, however, is important for the detection of core promoter elements, as these elements that contact the core transcriptional machinery exist at most once per promoter (see Chapter 6). The mops problem is the typical problem in case of TFs. Many TFs bind in clustered regions consisting of many binding sites of the same element. Since TFs bind to DNA, the considered alphabet consists only of the four nucleotides, adenine (A), cytosine (C), guanine (G), and thymine (T), i. e., $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$.

1.2. Motif Models

Multiple approaches have been described to model the intrinsic sequence preferences of TFs (e. g., MacIsaac and Fraenkel (2006), GuhaThakurta (2006)). All of them make assumptions to simplify the complex binding profile, with the independence assumption of complete independence between all positions being the most widely used one. In general, this assumption is reasonable since most studied cases indeed show no dependencies between positions (Roulet et al. (2002), Benos et al. (2002)). However, for some factors clear dependencies could be observed (Badis et al., 2009), demonstrating the relevance of more sophisticated models than the classical ones: consensus string and positional weight matrix.

1.2.1. Consensus string

The simplest way to represent a motif is a consensus string that only consists of the most common symbol at every position within the binding site. E. g., given the example result of Table 1.1, the consensus string would be `motifsite`. In order to build a consensus string for the sequence specificity of a TF, it is possible to use solely the four nucleotides (A, C, G, T), or to allow degeneracies using IUPAC characters as an extended alphabet (Cornish-Bowden, 1985). A summary of all IUPAC characters and the respective origin of designation is depicted in Table 1.2.

Symbol	**Meaning**	**Origin of designation**
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
W	A or T	Weak interaction (2 H bonds)
S	C or G	Strong interaction (3 H bonds)
R	A or G	puRin
Y	C or T	pYrimidine
K	G or T	Keto
M	A or C	aMino
B	C, G, or T	not A, B follows A
D	A, G, or T	not C, D follows C
H	A, C, or T	not G, H follows G
V	A, C, or G	not T, (not U), V follows U
N	A, C, G, or T	aNy nucleotide

Table 1.2.: The IUPAC code representing an extended alphabet including symbols for degenerate nucleotides.

All sites matching a consensus string are considered binding sites, however, additional binding sites can be selected by tolerating a certain number of mismatches. Tolerating mismatches is especially important if the very specific non-degenerate alphabet is used (e. g., Pavesi and Pesole (2006)).

1.2.2. Position weight matrix (PWM)

Another widely used motif model is the positional weight matrix (PWM). It consists of a two-dimensional matrix, of which every row corresponds to one symbol of the used alphabet, e. g., the four nucleotides, and every column to a position within the motif. The values within the matrix give the likelihood of each symbol at the respective position within a list of motif sites (Stormo et al., 1982). Hence, PWMs allow for a probabilistic description of the binding site without being restricted to degenerate nucleotides of a predefined alphabet. However, as all possible sites match to every given PWM with different probabilities, a threshold is necessary to separate matching binding sites from non-binding sites.

Generating a sequence logo is an elegant way to visualize a PWM, combining several different types of information in one figure (Schneider and Stephens, 1990): (a) the order of frequency of symbols at every position, with the highest frequent symbol at the top (b) the relative frequencies of every symbol at every position (c) the amount of information present at every position in the motif, measured in bits. The sequence logo for the example of Table 1.1 is given in Figure 1.1. As the frequencies of every symbol are ordered, the consensus sequence `motifsite` of the motif can be obtained by combining all symbols at the top of every position. To define the information R at each position within the motif, the difference between the maximum possible entropy and the entropy of the observed symbol distribution is calculated:

$$R = S_{\text{max}} - S_{\text{obs}} = \log_2 N - \left(- \sum_{n=1}^{N} p_n \log_2 p_n \right) \tag{1.1}$$

As the used alphabet consists of $N = 26$ symbols, the maximum possible entropy is 4.7 bits per position. Positions 3, 4, and 8 within the motif have all only two symbols with frequencies $p_0 = 0.25$ and $p_1 = 0.75$. Hence, the information at these positions is 3.89 bits.

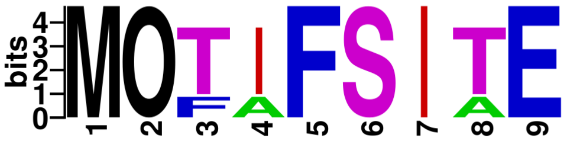


Figure 1.1.: Sequence logo for a possible result of the motif finding problem shown in Table 1.1.

1.3. Approaches to Solve the Motif Finding Problem

Several approaches have been developed to solve the motif finding problem, each tailored to the underlying motif model. In case of consensus strings, the motif space is finite given a certain motif length allowing for an enumeration of all possible consensus strings. In contrast, PWMs have an infinite motif space that necessitates an iterative refinement of the motif model to the optimal solution. Interestingly, despite of the differences between consensus strings and PWMs, both motif models are able to outperform each other depending on the benchmark (Tompa et al. (2005), Sandve et al. (2007)). A possible combination that utilizes the advantages of consensus strings and PWMs is to start with an enumerative phase as a candidate filter based on consensus strings followed by an iterative refinement phase based on PWMs (Linhart et al., 2008).

1.3.1. Enumerative approach

Enumerative approaches are based on an exhaustive enumeration of consensus strings. Some tools use a degenerate alphabet (e. g., Georgiev et al. (2010)), others allow for mismatches to a non-degenerate alphabet (e. g., Pavesi and Pesole (2006)). A main advantage of enumerative approaches is the possibility to cover large regions of the motif space efficiently. Starting with all seed k-mers of a given length k, for instance $k = 5$, seeds are extended as long as a quality score increases. As the number of binding sites can be counted given a k-mer, it is possible to use P-values as the quality measure.

1.3.2. Iterative approach

Iterative approaches are based on an iterative refinement of candidate PWMs. This refinement consists of a site refinement that updates the motif sites being bound by the current PWM. Some algorithms additionally perform a length refinement that optimizes the number of columns in the current PWM. As this iterative refinement is very time consuming, it is not possible to refine all seeds of a given length till convergence. Therefore, the complete refinement is typically done only on a very small subset or even only on the most promising seed. This seed can for instance be selected by the seed that has the best score after one iteration (Bailey and Elkan (1994)).

The two most widely used approaches to iteratively refine a candidate PWM are called expectation maximization EM and Gibbs sampling. However, both approaches are based on the optimization of a score function. How to optimize a solid P-value for a PWM is for the first time described within this work.

EM algorithm

Expectation maximization is a two-step process: In the expectation step, the current model is scanned over all sequences to calculate an estimate for each position to contain the motif. Then, in the maximization step, the model is adjusted to provide a better match to the sequences. Therefore, every position within the motif PWM is updated by the frequency in each motif site weighted by the calculated probability in the expectation step. This cycle of sequence scanning and model updating is repeated until convergence.

The implicit one occurrence per sequence model can be extended to a zero or one occurrence per sequence model by introducing an estimate that the motif PWM matches to background frequencies rather than to the sequence positions. If the likelihood of this estimation is close to one the potential binding sites within the sequence have only small contributions to the next maximization step.

The most widely used tool applying the EM-algorithm is MEME (Bailey and Elkan, 1994). Due to several extensions (Bailey et al. (2010), Bailey (2011)), MEME still remains one of the most powerful motif finding tools available.

Gibbs sampling

Gibbs sampling is in principle very similar to the EM algorithm described above. It also uses a two-step process of first scanning over all sequences to derive probabilities for each position to contain the motif, followed by a model updating step based on the probabilities obtained in the first step. However, in contrast to the EM algorithm, the new model is not a weighted mixture of all possible positions, it is an unweighted mixture of one binding site per sequence. This binding site is selected by sampling over all possible positions within the sequence, whereas the calculated probabilities reflect the likelihood of a position to be selected. In case of a zero or one occurrence per sequence model, the introduced site consisting of only background frequencies might be selected, leading to no contribution of this sequence.

The first motif finding tool that uses this approach is Gibbs Sampler (Lawrence et al., 1993). However, also more recent algorithms are based on Gibbs sampling, for instance PRIORITY (Narlikar et al., 2006).

2. Materials and Methods

This chapter provides details of the theoretical foundation and the efficient realization of significance calculations within XXmotif. Furthermore, it provides all the parameters used for the tested motif finding tools within the benchmarks. This should support repeatability of the results and point out the main ideas that allow for efficient calculations of P-values from PWMs.

The remainder of this methods chapter is organized as follows: Section one describes wherefrom XXmotif can be obtained. Section two provides the statistical framework used for calculating P-values from PWMs followed by the third section that gives a more detailed description of the XXmotif workflow. Finally, section four summarizes the tools used in the benchmarks and lists the chosen parameters.

2.1. Availability

The command line version of XXmotif can be obtained as source code or binaries (64 Bit and 32 Bit versions for UNIX systems) from `ftp://toolkit.lmb.uni-muenchen.de/xxmotif`.

The XXmotif web server is available at `http://xxmotif.genzentrum.lmu.de`.

2.2. Statistical Framework

In order to optimize the P-value of a motif PWM, it has to be possible to calculate the significance of a specific site given the PWM as well as the significance of a set of binding sites. This section provides the theoretical basis to efficiently calculate P-values from PWMs.

2.2.1. Calculating the significance of an l-mer – Background model

To calculate the probability to find a given l-mer x by chance, a background model has to be used. This model should be calibrated on a set having the same DNA properties as the input set, but no motifs (negative set). The simplest background model assumes no correlations between the positions of the l-mer (0^{th}-order background model), and hence, utilizes only monomer probabilities of the nucleotides $f(x_i)$ on the negative set. According to this background model, the probability to find an l-mer x is:

$$P_{\text{bg}}(x) = \prod_{i=1}^{l} f(x_i) \tag{2.1}$$

However, the independence assumption underlying this model is very inaccurate and leads to an overestimation of the significance of poly A/T stretches or dinucleotide repeats, which are very frequent in non-coding DNA.

Therefore, many motif finding tools use higher-order background models to capture these dependencies. E. g., for a k^{th}-order background model all $(k+1)$-mers within the negative set are counted and probabilities $f(x_1 \dots x_{k+1})$ and conditional probabilities $f(x_{k+1}|x_1 \dots x_k)$ are calculated. With these, the probability to find an l-mer x can be calculated as follows:

$$P_{\text{bg}}(x) = f(x_1 \dots x_{k+1}) \prod_{i=k+2}^{l} f(x_i|x_{i-k} \dots x_{i-1}) \tag{2.2}$$

The main drawback of this method is the huge amount of possible $(k+1)$-mers for large k's necessary to estimate from usually limited data. As poly A/T stretches and dinucleotide repeats are usually longer than 6 residues, a k of at least 8 is still useful. However, this leads to very few counts that are indistinguishable from noise for many of the 262144 different 9-mers even for large negative sets.

To overcome this problem, XXmotif uses interpolated Markov models (Salzberg et al., 1998) which automatically use lower-order probabilities if the negative set does not provide enough counts for higher-order k-mers. Given a pseudocount factor α and the number of occurrences of a k-mer $n(x_1 \dots x_k)$, the conditional probability $f(x_{k+1}|x_1 \dots x_k)$ can be calculated as follows:

$$f(x_{k+1}|x_1 \dots x_k) = \frac{n(x_1 \dots x_{k+1}) + 4\alpha f(x_{k+1}|x_2 \dots x_k)}{n(x_1 \dots x_k) + 4\alpha} \tag{2.3}$$

In case of few k-mer counts, i. e., $n(x_1 \dots x_{k+1}) \approx 0$, the formula simplifies to the result of order $k-1$. However, if the k-mer counts are high, i. e., $n(x_1 \dots x_{k+1}) \gg 4\alpha$, the formula corresponds to the one of order k.

We used $\alpha = 10$ as default value for XXmotif as it seems to be a good trade-off between noise reduction and utilization of the counts of higher orders.

2.2.2. Calculating the significance of a binding site given a PWM

To calculate the P-value of a specific site x of interest with length l given a PWM, the P-values of all l-mers z that have a better or equal log-odds score $S(z)$ than the site of interest $S(x)$ have to be summed up:

$$P\text{-value}(x) = \sum_{z \in \{z : S(z) \geq S(x)\}} P_{bg}(z) \tag{2.4}$$

where the log-odds score $S(x)$ is calculated by summing up the logarithm of the probability to have nucleotide x_i at position i within the PWM divided by the background probability of this nucleotide $f(x_i)$:

$$S(x) = \sum_{i=1}^{l} \log\left(\frac{PWM(i, x_i)}{f(x_i)}\right) \tag{2.5}$$

Since it is very time consuming to generate all 4^l l-mers, P-value(x) cannot be efficiently obtained by exhaustive enumeration of all l-mers with score $S(z) \geq S(x)$. However, by using a branch-and-bound technique, it is possible to generate exactly these high-scoring l-mers in linear time with respect to the output size, i. e., the number of l-mers generated.

Branch-and-bound algorithm

Every PWM column contributes independently to the log-odds score (see Equation 2.5). Therefore, given the prefix of length m of an l-mer, the maximum log-odds score of the remaining suffix $S_{\max,m+1}$ is easily calculated by summing up the maximum log-odds value of the corresponding columns:

$$S_{\max,m+1} = \sum_{i=m+1}^{l} \max_{j \in \{A,C,G,T\}} \left\{\log\left(\frac{PWM(i,j)}{f(j)}\right)\right\} \tag{2.6}$$

If the maximum score of the suffix is not high enough to reach the threshold $S(x)$, it is not necessary to enumerate the suffixes and the current path can be abandoned. All paths reaching the l-th column correspond to l-mers that are ‘similar enough’ to the PWM. Pseudocode for the procedure is given in Algorithm 1.

The runtime is approximately linear in the number of branches followed, that is in the number of l-mers generated plus the number of dead-end paths. Two optimizations are used to reduce the number of futile furcations by trying highest scoring nucleotides first: (a) sort each column’s entries in order of descending score, (b) reorder columns according to their highest scoring entry in descending order.

Algorithm 1: CreateSimilarKmers(i, S_{i-1}, Z_{i-1})
Recursive generation of l-mers similar to a PWM with branch-and-bound. The initial call is CreateSimilarKmers($1, 0, \varepsilon$), i. e., the algorithm starts in the first column with the neutral element of addition for the score so far and the empty word for the preceding l-mer.

```
Data: PWM score matrix
      $S_{\max,j}$ maximum possible score for columns $j, \dots, l$
      $S(x)$  similarity threshold: score for site $x$
Input: $i$       current column
       $Z_{i-1}$ generated $(i-1)$-mer
       $S_{i-1}$ score of $Z_{i-1}$
foreach $j \in \{A, C, G, T\}$ do
    $S_i \leftarrow S_{i-1} + PWM_{i,j}$
    if $S_i + S_{\max,i+1} < S(x)$ then continue
    $Z_i \leftarrow Z_{i-1} \cdot j$
    if $i < l$ then
        CreateSimilarKmers($i+1, S_i, Z_i$)
    else
        add $Z_i$ to list of similar $l$-mers
```

Of course, the maximal number of similar l-mers is still 4^l. However, only high scoring matches are of interest during the search for significant motifs. This means that only a small fraction of l-mers has to be considered for stringent thresholds.

Splitting and recombination of branch-and-bound *P*-values

For long l-mers, enumeration of high scoring matches can still be very time consuming, especially if the PWM has many degenerate columns. Therefore, we accelerate the calibration for $l > 8$ by splitting the motif into two parts, calculate P-values for both parts individually and combine them to yield the final P-value (see Figure 2.1). The left part of the l-mer (X_1) is set to length 8, the right part (X_2) to the remaining nucleotides, allowing to calculate P-values for l-mers with up to 17 nucleotides. For longer l-mers, even the calibration of the shorter parts is too time consuming if the PWM is very degenerate.

Calculating independent P-values for both parts as illustrated in Figure 2.1 would neglect higher-order dependencies between both parts of the l-mer. To calculate the exact P-value for the combined l-mer, the P-values of the right part has to be calculated depending on the left part of the l-mer. Therefore, the products for all relevant combinations of left part l-mers Z_1 and right part l-mers Z_2 reaching a total score $S_1 + S_2$ bigger than the score for

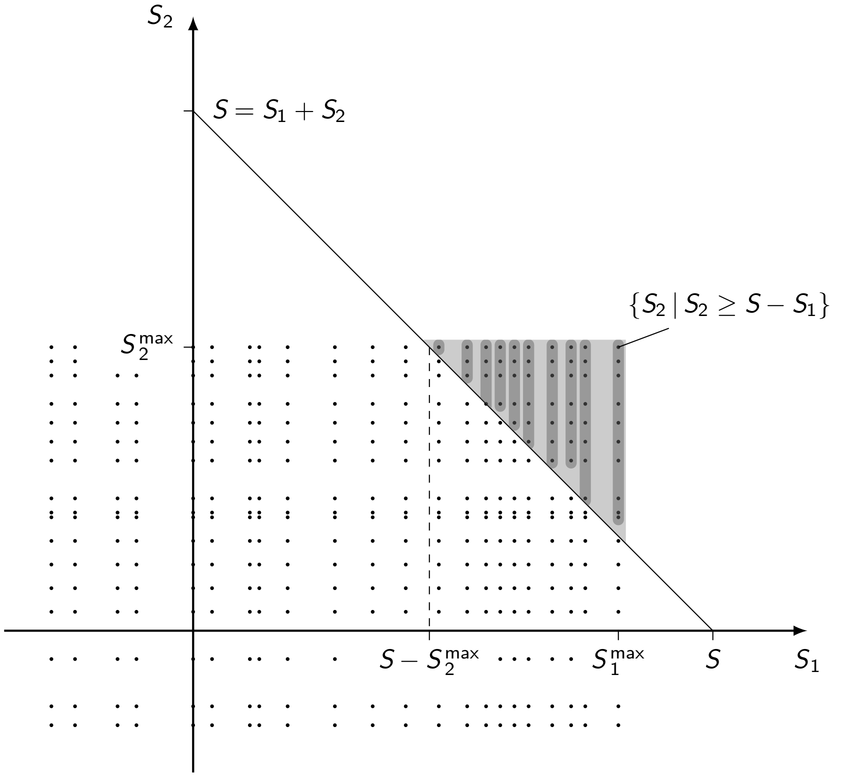


Figure 2.1.: Splitting and recombination of the branch-and-bound P-value calculation for long l-mers (shown in log-space). The l-mer $x = x_1 \dots x_l$ is divided into two parts $X_1 = x_1 \dots x_j$ and $X_2 = x_{j+1} \dots x_l$. For each part, the probabilities of relevant scores S_1 and S_2 are obtained using the branch-and-bound algorithm CreateSimilarKmers. In order to obtain the P-value $P(x)$ for reaching at least a total score of $S(x)$, the probabilities of all pairs (S_1, S_2) with $S_1 + S_2 \geq S(x)$ (light gray area) have to be summed up. Neglecting higher-order dependencies between both parts, this can be calculated as the sum over relevant S_1 columns (shown in dark gray), which in turn can be constructed incrementally from lower to higher S_1:

$$P(x) = \Pr[S_1 + S_2 \geq S(x)] = \sum_{S_1 \geq S(x) - S_2^{\max}}^{S_1^{\max}} \Pr[s_1 = S_1] \; \Pr[S_2 \geq S(x) - S_1].$$

the site of interest $S(x)$ have to be summed up:

$$P(x) = \sum_{(Z_1, Z_2) \in \{S_1 + S_2 \geq S(x)\}} P(Z_1) \; P(Z_2 | Z_1) \tag{2.7}$$

To speed up this calculation, we approximate the dependency of the right part to the average PWM of the left part, allowing for the right part P-values to be summed up independently, as shown in Figure 2.1. For a left part of length 8, the background order k, and the set Ω of all possible k-mers, this gives:

$$P(Z_2 | Z_1) \approx P(Z_2 | \overline{PWM}) = \sum_{z \in \Omega} \left(P(Z_2 | z) \prod_{i=1}^{k} PWM(8{-}k{+}i, z_i) \right) \tag{2.8}$$

2.2.3. Calculate the significance of a set of binding sites

After a P-value is calculated for every site given the PWM, it is necessary to find the optimal P-value threshold to consider a site as significant, i. e., to assign a site as a binding site of the motif. This is determined using so-called order statistics. The possible sites are sorted by their P-value in increasing order (i. e., in order of decreasing significance). For a binomial distribution, the probability to find exactly K sites in a set of N possible sites with P-values at least as small as p is $\binom{N}{K}p^K(1-p)^{N-K}$. Accordingly, the probability to find at least K sites with P-values at least as good as the K-th best, P_K, is given by:

$$P_{\text{overrep}}^{(K)} = \sum_{k=K}^{N} \binom{N}{k} (P_K)^k (1 - P_K)^{N-k} \tag{2.9}$$

The optimal K^* is the one with minimal $P_{\text{overrep}}^{(K)}$:

$$K^* = \underset{K \in \{1,\ldots,N\}}{\operatorname{argmin}} \{P_{\text{overrep}}(K)\} \tag{2.10}$$

The P-value of the most significant set of binding sites is thus $P_{\text{overrep}}^{(K^*)}$, and the K^* best sites are considered to be functional.

Multiple occurrence per sequence model

Using the multiple occurrence per sequence model (mops model) as the motif model of XXmotif permits binding sites simultaneously at every position within the input set. Hence, N, the number of different binding sites, equals the number of nucleotides within the input set M, subtracted by the nucleotides at the sequence ends covered by the motif. Having L sequences and a PWM length W,

$$N = M - L\,(W - 1) \tag{2.11}$$

However, the possibility to have overlapping binding sites is problematic as repetitive motifs fit to more binding sites than expected by the background model. Hence, we preclude overlapping sites, which leads to a maximum number of possible sites K that is smaller than N. If XXmotif is set to search motifs on both strands of DNA, N is increased by a factor of two. To preclude that palindromic motifs fit to the same binding site on both strands leading to too significant P-values , we also preclude overlapping binding sites between both strands.

Zero or one occurrence per sequence model

Using the zero or one occurrence per sequence model (zoops model) as the motif model of XXmotif permits at most one binding site per input sequence. Here, N is the number of sequences and K ranges between 0 and N as overlaps are not possible in this setting. The P-value P_K of a site used in Equation 2.9 now refers to a per sequence P-value that can be calculated from the site P-value p by calculating the probability to find a binding site with length W at least as significant as p within sequence S:

$$P_K = 1 - (1-p)^{|S|-W+1}\,, \tag{2.12}$$

i. e., the probability of the complementary event of not finding it at any of the $|S|-W+1$ possible starting positions of the PWM in sequence S.

Instead of individual sequence lengths, the geometric mean length is used for all sequences. This avoids problems resulting from equally scoring matches becoming disproportionately significant (or insignificant) if found in very short (or very long) sequences. The geometric mean is adequate since lengths are scaling variables that are best compared in terms of factors, not absolute differences.

One occurrence per sequence model

XXmotif also provides a one occurrence per sequence model. It is implemented by using the same framework as for the zero or one occurrence per sequence model, however, the number of motifs K is not optimized using order statistics, but manually set to N. Subsequently, the final P-value is the likelihood to find N times an instance with the N'th best P-value. As only the N'th best P-value contributes to the final result, this option should only be used if it is known that all sequences contain the motif, otherwise the zero or one occurrence per sequence option is recommended.

2.2.4. Correcting *P*-values for multiple testing

Finally, multiple testing has to be taken into account: Any PWM in the whole motif space has the same chance to achieve a certain significance by coincidence. Hence, for calculating the E-value which corresponds to the expected number of motifs with a given P-value, a Bonferroni correction is applied. This is a conservative method that multiplies the calculated P-value by the number of different motif models that could be tested. Since in principle infinite slightly different PWMs exist in the motif space, it is necessary to define a parameter N_{eff} which defines the effective number of possible PWM columns. Hence, for a PWM with length W, the respective E-value is calculated as follows:

$$E\text{-value} = P\text{-value}\, N_{\text{eff}}^{W} \tag{2.13}$$

In the enumeration phase of XXmotif, the motif model consists of one out of ten different IUPAC characters per position (A, C, G, T, M, R, W, S, Y, K). However, only the four nucleotides A, C, G, and T are independent, the remaining characters are partly similar to each other. Hence, N_{eff} should be set to a value between four and ten, with $N_{\text{eff}} = 6$ is used as default. In the refinement phase of XXmotif, we use $N_{\text{eff}} = 10$ to account for the strong similarities between different PWM columns, but still capture the higher number of different PWMs than IUPAC strings of the same length. In case of the enumeration phase, for which gaps are allowed in the IUPAC string, additionally to the factor N_{eff}^{W}, a factor of two per gap position is used to capture the higher amount of motifs to test if a certain number of gaps is present.

2.2.5. Calculating conservation *P*-values

Conservation P-values are like overrepresentation P-values calibrated on the negative set, if available. Otherwise, the input set is used. They are calculated as the probability to find at most m mutations from the first sequence to n other sequences within the alignment, given the frequency of every nucleotide within the site. As this nucleotide composition $c = f_{\text{site}}(A,C,G,T)$ is taken into account, different mutation rates within regions of different A/T content are included.

$$P_{\text{cons}}(m, n, c) = \sum_{i=0}^{m} \frac{f(i,n,c)}{\sum_{j=0}^{\infty} f(j,n,c)} \tag{2.14}$$

For each site it is tested how many sequences have no gaps in the alignment and the maximal n is used. To preclude that related informative sequences are lost if a closely related sequence was not alignable and therefore has only gaps at the site position, a preprocessing step is used to fill gaps in closely related species with the nucleotides of the a more distantly related species. This procedure can be considered as an upper bound estimation for the mutation within the site.

To calculate a combined conservation P-value for the K best sites according to the PWM, we use the formula for the distribution of the product of independent pairwise P-values given by Bailey and Gribskov (1998):

$$P_{\text{cons}}^{(K)} \approx p \frac{\sum_{i=0}^{K-1} (-\log p)^{i}}{i!} \tag{2.15}$$

where p is the product of conservation P-values of the K considered sites:

$$p = \prod_{i=1}^{K} P_{\mathrm{cons},i} \tag{2.16}$$

2.2.6. Weighted combination of P-values

Given two independent P-values p_1 and p_2, these can be combined to a single P-value using the formula:

$$P_{\mathrm{comb}} = \Pr[P_1 P_2 \leq p_1 p_2] = p_1 p_2 \left(1 - \log\left(p_1 p_2\right)\right) \tag{2.17}$$

However, this formula for independent P-value combination implicitly assumes that both P-values are similarly important. If the first source of information p_1 is much more important than the second source of information p_2, meaning that most non-random events (TPs) have much smaller p_1's than p_2's, combined P-values can be even worse in distinguishing non-null-model events than the single P-value of the more important source of information.

E. g., if $p_1 = 10^{-5}$ and $p_2 = 0.1$, $\quad P_{\mathrm{comb}} = 10^{-6}(1 - \log(10^{-6}) = 1.5 \times 10^{-5} > p_1$

This scenario is very common for XXmotif. Here, overrepresentation P-values p_1 are combined with conservation P-values p_2 which often are only slightly significant. This low information stems from TF turnover events, which cancel out any conservation information even for functional binding sites, or, if the species are too closely related, even completely conserved binding sites are not significant.

Therefore, it is desirable to assign a weight $w \in]0, 1[$ to the source of information which is less important and calculate a P-value for the weighted score $\varrho = p_1 p_2^w$ (see Figure 2.2). This can be calculated analytically:

$$P_{\mathrm{comb}} = \Pr[P_1 P_2^w \leq \overbrace{p_1 p_2^w}^{\varrho}] = \varrho^{\frac{1}{w}} + \int_{\varrho^{\frac{1}{w}}}^{1} \frac{\varrho}{P_2^w} \,\mathrm{d}P_2 = \frac{\varrho - w\varrho^{\frac{1}{w}}}{1-w} = \frac{p_1 p_2^w - p_1^{\frac{1}{w}} p_2 w}{1-w}, \tag{2.18}$$

For the weight $w = 1/3$, which is the default value used in XXmotif, the example calculation from above gives $P_{\mathrm{comb}} = 6.96 \times 10^{-6}$, which is 1.44 times more significant than the single P-value.

2.2.7. Calculating positional P-values

If motif instances cluster together at a fixed distance relative to a specified anchor point, e. g., TSSs or nucleosomes, motif identification is facilitated by introducing a P-value that

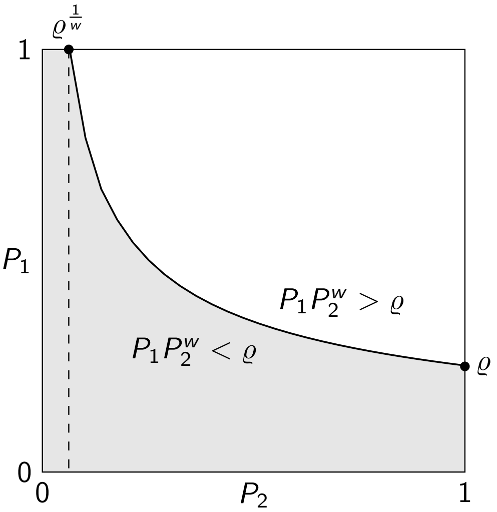


Figure 2.2.: Weighted combination of P-values. The probability that the product of P_1 and P_2^w is less than ϱ is equal to the shaded area. This in turn is composed of a rectangle with area $\varrho^{1/w}$ and the area under $P_1 P_2^w$ over $[\varrho^{1/w}, 1]$.

captures the differences of this clustering to a random distribution. To decide whether for a motif of size l the instances are significantly clustered within a region Δ, a localization P-value P_{loc} is calculated using the binomial distribution:

$$P_{\text{loc}} = \sum_{k=K}^{N} \binom{N}{k} (P_{\text{reg}})^k (1 - P_{\text{reg}})^{N-k} \quad \text{with } P_{\text{reg}} = \frac{|\Delta|}{L - l + 1}$$

where K is the number of motifs within the tested region, N is the number of all motif instances, $|\Delta|$ is the size of the region where the motifs are clustered and L is the length of the sequences. To find the region Δ of highest enrichment, the region with the best P_{loc} is selected if it is significant enough ($P_{\text{loc}} < 10^{-3}$).

Positional quasi *P*-value

Since P_{reg} cannot be smaller than $1/(L - l + 1)$ it is only possible to calculate a quasi P-value, which cannot directly be combined with the overrepresentation P-value by simply using the formula for combining P-values shown in Equation 2.18.

We define our positional quasi P-value for position z_k and a given cluster region Δ_0 ranging from positions z_s to z_e by

$$p(z_k) = \frac{|\{z : 1 \leq z \leq L - l + 1 \wedge |z - \mu| \leq d\}|}{L - l + 1}$$

where

$$d = \max\{|z_k - \mu|, D\}$$

and

$$\mu = \frac{z_e + z_s}{2}$$

$$D = \left\lceil \frac{\Delta_0}{2} \right\rceil = \left\lceil \frac{z_e - z_s + 1}{2} \right\rceil$$

To simplify the expression for $p(z_k)$ and see how it depends on μ, D and $L - l + 1$ one can first note that

$$p(z_k) = \frac{\lfloor \min\{L - l + 1, \mu + d\} \rfloor - \lceil \max\{1, \mu - d\} \rceil + 1}{L - l + 1}$$

and write this formula as

$$p(z_k) = \begin{cases} \dfrac{\Delta_0}{L - l + 1} & \text{for } |z_k - \mu| \leq D \\ \dfrac{\Delta(z_k)}{L - l + 1} & \text{for } |z_k - \mu| \leq D \end{cases} \tag{2.19}$$

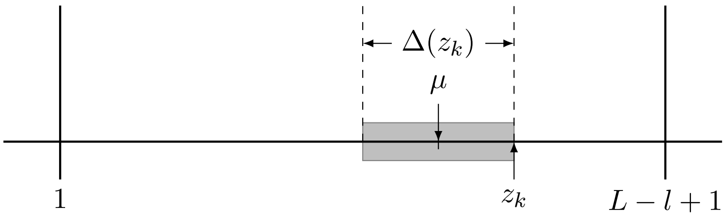
where

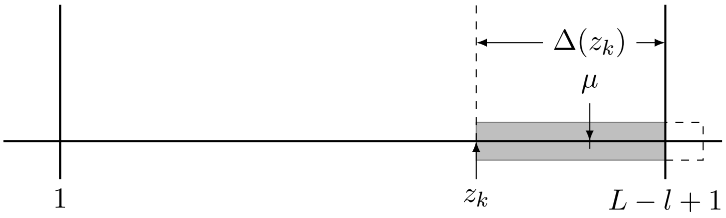
$$\Delta_0 = \lfloor \min\{L - l + 1, \mu + D\} \rfloor - \lceil \max\{1, \mu - D\} \rceil + 1$$

and

$$\Delta(z_k) = \lfloor \min\{L - l + 1, \mu + |z_k - \mu|\} \rfloor - \lceil \max\{1, \mu - |z_k - \mu|\} \rceil + 1$$

as can be seen for two different values for μ and z_k in the following figure:





Then for $z_k \geq \mu$:

$$\begin{aligned}\Delta(z_k) &= \underbrace{\lfloor \min\{L-l+1, z_k\}\rfloor}_{z_k} - \lceil \max\{1, 2\mu - z_k\}\rceil + 1 \\ &= \begin{cases} \lfloor 2(z_k-\mu)+1\rfloor & \text{for } z_k < 2\mu - 1 \\ z_k & \text{for } z_k \geq 2\mu - 1 \end{cases}\end{aligned} \tag{2.20}$$

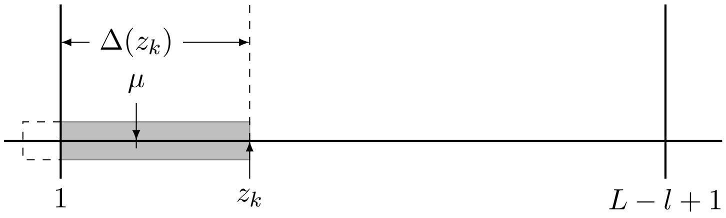
For $z_k \leq \mu$ we obtain:

$$\begin{aligned}\Delta(z_k) &= \lfloor \min\{L-l+1, 2\mu - z_k\}\rfloor - \underbrace{\lceil \max\{1, z_k\}\rceil + 1}_{z_k} \\ &= \begin{cases} \lfloor 2(\mu - z_k)+1\rfloor & \text{for } z_k > 2\mu - L + l - 1 \\ L-l+1-z_k+1 & \text{for } z_k \leq 2\mu - L + l - 1 \end{cases}\end{aligned} \tag{2.21}$$

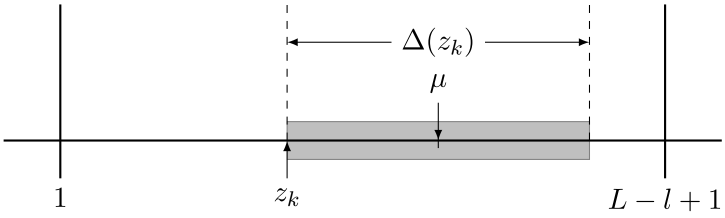
We can summarize Equations 2.20 and 2.21:

$$\Delta(z_k) = \begin{cases} z_k & \text{for } z_k \geq 2\mu - 1 & \text{(a)} \\ \lfloor 2|z_k-\mu|+1\rfloor & \text{for } 2\mu - L + l - 1 < z_k < 2\mu - 1 & \text{(b)} \\ L-l+1-z_k+1 & \text{for } z_k \leq 2\mu - L + l - 1 & \text{(c)} \end{cases} \tag{2.22}$$

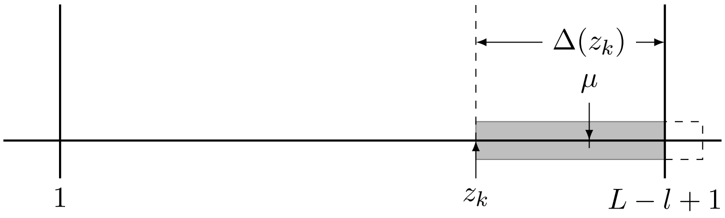
(a)



(b)



(c)



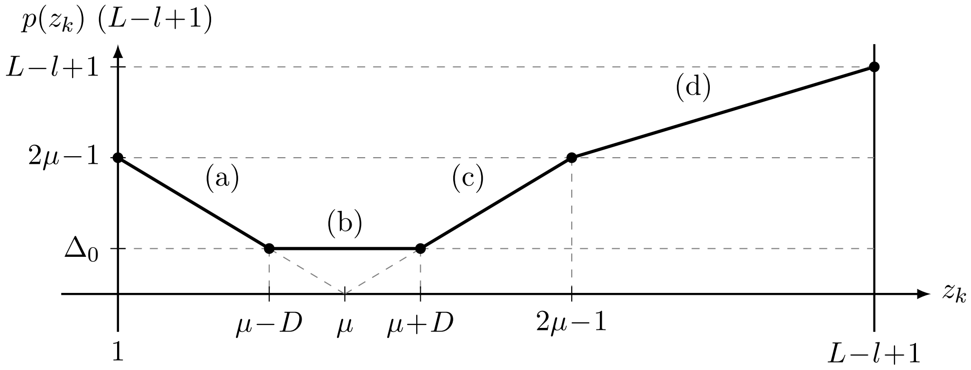
Case (a) can only occur when $L-l+1 \geq 2\mu - 1$, i. e., $\mu \leq \frac{L-l+2}{2}$

Case (c) can only occur when $2\mu - L + l - 1 \geq 1$, i. e., $\mu \geq \frac{L-l+2}{2}$

To substitute Equation 2.22 into Equation 2.19, we can now distinguish two cases:

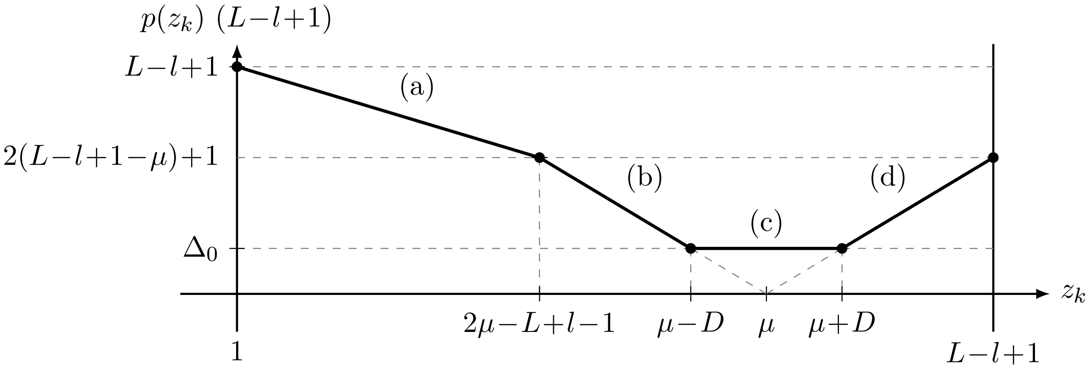
1. $\mu \leq \dfrac{L-l+2}{2}$:

$$p(z_k)\ (L{-}l{+}1) = \begin{cases} \Delta_0 & \text{for } \mu - D \leq z_k \leq \mu + D \quad \text{(b)} \\ z_k & \text{for } 2\mu - 1 \leq z_k \quad \text{(d)} \\ \lfloor 2|z_k - \mu| + 1 \rfloor & \text{for } 1 \leq z_k \leq \mu - D \lor \quad \text{(a)} \\ & \qquad \mu + D \leq z_k \leq 2\mu - 1 \quad \text{(c)} \end{cases}$$



2. $\mu \geq \dfrac{L-l+2}{2}$:

$$p(z_k)\ (L{-}l{+}1) = \begin{cases} \Delta_0 & \text{for } \mu - D \leq z_k \leq \mu + D \quad \text{(c)} \\ L{-}l{+}1{-}z_k{+}1 & \text{for } z_k \leq 2\mu{-}L{+}l{-}1 \quad \text{(a)} \\ \lfloor 2|z_k - \mu| + 1 \rfloor & \text{for } 2\mu{-}L{+}l{-}1 \leq z_k \leq \mu{-}D \lor \quad \text{(b)} \\ & \qquad \mu + D \leq z_k \leq L{-}l{+}1 \quad \text{(d)} \end{cases}$$

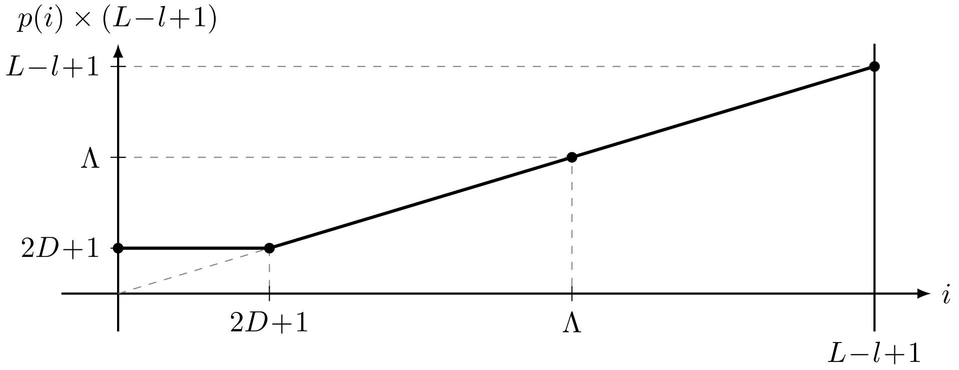


When we sort the values $p(z_k)$ in ascending order (indexed by i), we get

$$p(i) = \left.\begin{cases} 2D+1 & \text{for } 1 \le i \le 2D+1 \\ i & \text{for } 2D+1 \le i \le \Lambda \\ i & \text{for } \Lambda \le L{-}l{+}1 \end{cases}\right\} \times (L{-}l{+}1)^{-1}$$

where

$$\Lambda = \begin{cases} 2\mu - 1 & \text{for } \mu \le \dfrac{L{-}l{+}2}{2} \\ 2(L{-}l{+}1{-}\mu) + 1 & \text{for } \mu > \dfrac{L{-}l{+}2}{2} \end{cases}$$



P-value combination of the positional quasi *P*-value

Suppose our positional P-value $p_2(i)$, which may not be uniformly distributed and therefore is no true P-value, is calculated according to

$$\Delta_0 = \lfloor \min\{L - l + 1, \mu + D\} \rfloor - \lceil \max\{1, \mu - D\} \rceil + 1$$

$$p_2(i) = \begin{cases} \dfrac{\Delta_0}{L - l + 1} & \text{for } |i - \mu| \le D \\ \dfrac{i}{L - l + 1} & \text{for } |i - \mu| > D \end{cases}$$

Suppose the match of the PWM at position $i \in \{1, \ldots, L - l + 1\}$ is quantified by a true P-value $p_1(i)$. We would now like to calculate the true combined P-value for $p = p_1(i)\, p_2(i)$ at a specified position i, which is the probability that a better combined P-value could be achieved at any start position in a sequence of length $L-l+1$:

$$P\text{-value}(p) = P\Big(\min_i\{p_1(i)\ p_2(i)\} < p\Big)$$

We start by calculating $1-P$-value(p):

$$\begin{aligned}
&P\left(\min_i\{p_1(i)\ p_2(i)\} \geq p\right) = \prod_{i=1}^{L-l+1} P\left(p_1(i)\ p_2(i) \geq p\right)\\
&= \prod_{i=1}^{L-l+1} P\left(p_1(i)\ \ \max\left\{\frac{\Delta_0}{L-l+1}, \frac{i}{L-l+1}\right\} \geq p\right)\\
&= \prod_{i=1}^{\Delta_0} P\left(p_1(i) \geq p\ \frac{L-l+1}{\Delta_0}\right) \prod_{i=\Delta_0+1}^{L-l+1} P\left(p_1(i) \geq p\ \frac{L-l+1}{i}\right)\\
&= \left(1-p\ \frac{L-l+1}{\Delta_0}\right)^{\Delta_0} \prod_{i=\Delta_0+1}^{L-l+1} \left(1-p\ \frac{L-l+1}{i}\right)
\end{aligned}$$

$\Rightarrow$ P-value(p) can be calculated as follows:

$$P\text{-value}(p) = 1 - \left(1-p\ \frac{L-l+1}{\Delta_0}\right)^{\Delta_0} \exp\left(\sum_{i=\Delta_0+1}^{L-l+1} \log\left(1-p\ \frac{L-l+1}{i}\right)\right)$$

if $p\ \dfrac{L-l+1}{\Delta_0} \leq 0.1$, we can expand the logarithm into a Taylor series:

$$\log(1+x) = x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \ldots + \frac{x^k}{k!} + \ldots \quad \text{for } x = -p\ \frac{L-l+1}{i}$$

$\Rightarrow$ by using a first-order approximation one finally gets:

$$\begin{aligned}
P\text{-value}(p) &= 1 - \left(1-p\ \frac{L-l+1}{\Delta_0}\right)^{\Delta_0}\\
&\qquad \exp\Bigg(-p\ (L-l+1)\ \underbrace{\sum_{i=\Delta_0+1}^{L-l+1} \frac{1}{i}}_{I_1(\Delta_0+1)}\ + \frac{1}{2}p^2(L-l+1)^2\ \underbrace{\sum_{i=\Delta_0+1}^{L-l+1} \frac{1}{i^2}}_{I_2(\Delta_0+1)}\ + \ldots\Bigg)\\
&= 1 - \left(1-p\ \frac{L-l+1}{\Delta_0}\right)^{\Delta_0} \exp\left(\sum_{k=1}^{\infty} \frac{1}{k!}\left(-p\ (L-l+1)\right)^k I_k(\Delta_0+1)\right)\\
&\approx 1 - \left(1-p\ \frac{L-l+1}{\Delta_0}\right)^{\Delta_0} \exp\left(-p\ (L-l+1) \sum_{i=\Delta_0+1}^{L-l+1} \frac{1}{i}\right)
\end{aligned}$$

2.3. Workflow of XXmotif

This section describes the workflow of XXmotif in more detail. In a nutshell, starting with an optional run of XXmasker, IUPAC strings are extended, merged together to obtain a PWM model and refined by optimizing the P-value (see Figure 2.3).

2.3.1. XXmasker

Since XXmotif ranks all motifs by P-value, it is highly sensitive to even rare motifs if the information content of these motifs is high enough. However, this high sensitivity becomes problematical if the input sequences contain homologous parts, repeats or low complexity regions. These stretches of DNA are typically longer than 20 nucleotides and occur multiple times, typically as perfect repeats. In order to avoid an assignment of high P-values to these features, we have developed XXmasker, an optional tool that masks these sequence regions prior to the main algorithm of XXmotif.

Nucleotides are masked by XXmasker if at least one out of the following three conditions is satisfied:

1. **The nucleotide is within a homologous region:**
   To detect homologous regions, BLAST is used with E-value cutoff 10^{-10} and the soft masking option (`"-F m S"`). For this, a database is incrementally built with all input sequences, where the first input sequences is kept completely and parts of the remaining sequences are masked in all regions in which BLAST detects sufficient homology to any of the already considered sequences. The very stringent E-value cutoff assures that no informative regions are masked, the BLAST masking option avoids that low-complexity segments cause homologous parts to fall below the E-value threshold.

2. **The nucleotide is within a low complexity region:**
   We define a low complexity region to be a DNA stretch of at least 50 nucleotides consisting of at most two different nucleotides.

3. **The nucleotide is within a repeat:**
   We define a repeat region as a DNA stretch of at least 50 nucleotides consisting of perfect repeats with a repeat length in between 3 and 10.

Generally, we chose very strict parameters for all of these conditions. However, in some cases a relaxation of the E-value cutoff, or the masking of imperfect low complexity and repeat regions might be a possibility to further improve the performance.

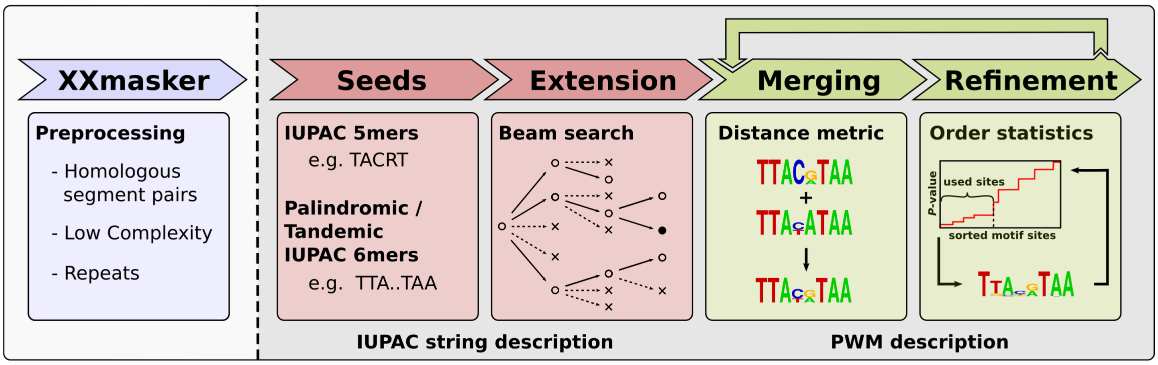


Figure 2.3.: Overview of XXmotif with its main stages. After an optional step to mask confounding sequence regions (blue), P-values of all 5-mers and gapped palindromic 6-mer seed patterns are evaluated, and the best seeds are recursively extended by an optional gap and motif position (red). Patterns are converted into PWMs and fed to the PWM stage (green). Here, similar PWMs are merged and then iteratively refined by optimizing the motif enrichment P-value. Finally, merging and refinement stages are iterated till convergence.

2.3.2. Seeds phase

XXmotif starts by enumerating all 5-mers with at most two IUPAC characters (M, R, W, S, Y, K) as well as all palindromic and tandemic 6-mers with gaps of size $0 - 11$ between the first and last three positions (seeds phase; Figure 2.3). For each of these seed patterns, a P-value is calculated using the binomial distribution which is corrected for multiple testing to obtain the corresponding E-value.

To calculate the P-value for an IUPAC string U, the probabilities of all l-mers x matching to U have to be summed up. Therefore, in case of gaps in the IUPAC string, probabilities of l-mers with any nucleotide at these positions have to be considered:

$$P(U) = \sum_{x \in U} P(x)$$

where $P(x)$ is calculated as shown in Section 2.2.1. The probability to find K out of N possible binding sites matching U is calculated using the binomial distribution.

$$P_{\text{enrichment}} = \sum_{k=K}^{N} \binom{N}{k} (P(U))^k (1 - P(U))^{N-k} \tag{2.23}$$

To account for multiple testing, E-values are calculated as described in Section 2.2.4

2.3.3. Extension phase

Seeds are extended using a beam search approach, i. e., not only the one most promising path is followed, but the B most promising paths are examined. This allows for a very efficient extension, while avoiding local minima which may arise more likely by using only the best path.

As all IUPAC degenerations of a non-degenerate seed are highly overlapping and would therefore extend to similar IUPAC strings, it is possible to reduce runtime by extending only a small subset of these. Therefore, we extend only the five most promising degenerate seeds per non-degenerate seed, giving a total of 5120 5-mer (5×4^5), 3840 gapped palindromic and 3840 gapped tandemic 6-mer degenerate seeds ($5 \times 12 \times 4^3$).

All of these seeds are extended individually as long as the E-value improves. Possible extensions are IUPAC characters (A, C, G, T, M, R, W, S, Y, K) at the beginning and the end of the current IUPAC string, allowing gaps of size zero to three. Larger gap sizes are not necessary as it is very unlikely that the extended IUPAC string is more significant than the unextended version (see Section 2.2.4, multiple testing). Extensions having a lower E-value than the unextended version are sorted and the three most significant ones (circles in Figure 2.3) are iteratively extended. Afterwards, extended IUPAC strings are converted into PWMs by calculating the frequencies of every nucleotide within the matching sites.

Since identical IUPAC strings can be reached from different seeds, all extensions are stored in a hash allowing for a fast extraction of already calculated results.

Implementation details

In order to efficiently elongate thousands of different seeds, it is necessary to store each k-mer within a data structure that needs only a minimum amount of memory and can be modified to an elongated version within a few clock cycles. Furthermore, the possibility to reach one elongated k-mer from different IUPAC seeds makes it desirable to have a unique id for each k-mer. This allows to build a hash of already considered k-mers and to stop the elongation of already traversed paths. The main data structures fulfilling these requirements are shown as UML class diagrams in Figure 2.4.

Kmer is the central IUPAC k-mer representation class. It contains the offset list of Matches as sequence/position pairs and a representation of the current IUPAC k-mer. Since the matches are stored in a doubly linked list, each seed takes 28 bytes of memory on a 64 bit system. During extension millions of matches have to be stored and deleted from the seeds list. In order to avoid time consuming memory allocation and release operations we have developed a pool allocater (Pool_Alloc), which allocates chunks of 8 kilo bytes at once and releases matches by simply removing the pointer to a position within the chunk.

The k-mer representation interface AbstractKmer was designed with two main requirements: the characters and gaps at each position should be accessible and modifiable by index, and

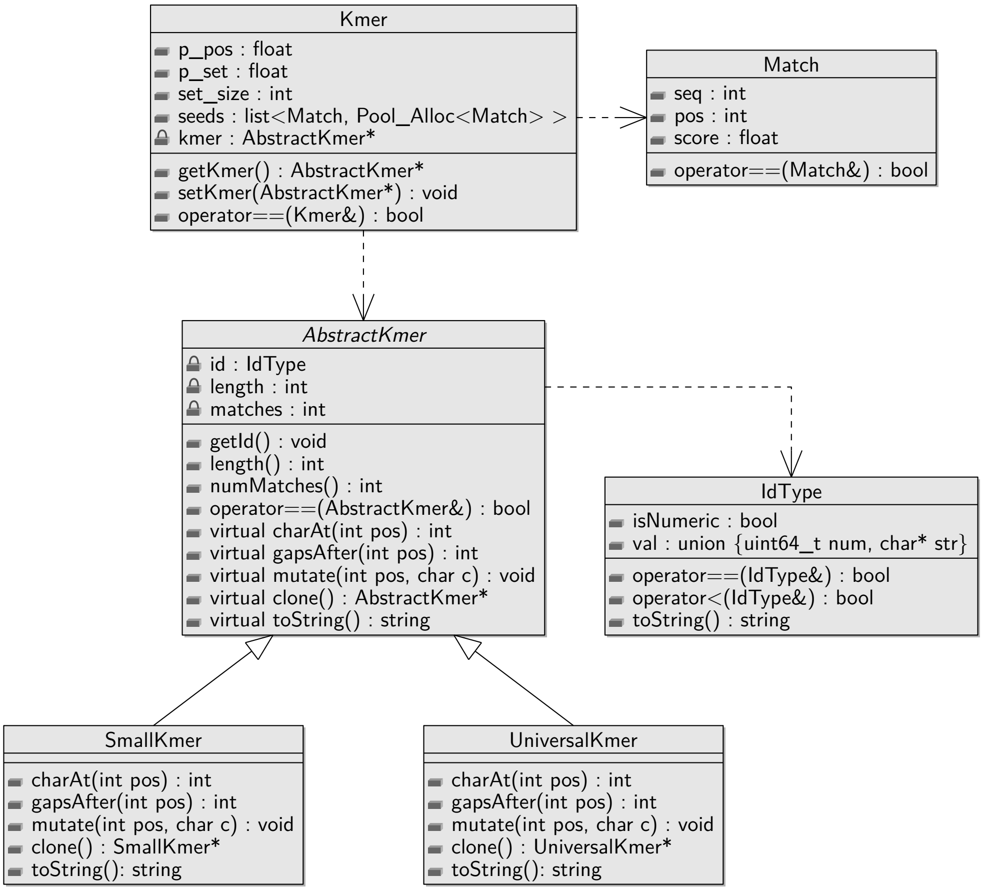

Kmer
p_pos : float
p_set : float
set_size : int
seeds : list<Match, Pool_Alloc<Match> >
kmer : AbstractKmer*
getKmer() : AbstractKmer*
setKmer(AbstractKmer*) : void
operator==(Kmer&) : bool
Match
seq : int
pos : int
score : float
operator==(Match&) : bool
AbstractKmer
id : IdType
length : int
matches : int
getId() : void
length() : int
numMatches() : int
operator==(AbstractKmer&) : bool
virtual charAt(int pos) : int
virtual gapsAfter(int pos) : int
virtual mutate(int pos, char c) : void
virtual clone() : AbstractKmer*
virtual toString() : string
IdType
isNumeric : bool
val : union {uint64_t num, char* str}
operator==(IdType&) : bool
operator<(IdType&) : bool
toString() : string
SmallKmer
charAt(int pos) : int
gapsAfter(int pos) : int
mutate(int pos, char c) : void
clone() : SmallKmer*
toString(): string
UniversalKmer
charAt(int pos) : int
gapsAfter(int pos) : int
mutate(int pos, char c) : void
clone() : UniversalKmer*
toString(): string

Figure 2.4.: Main data structures used for IUPAC k-mers within the XXmotif elongation phase.

each k-mer should have a unique id for hashing.

The SmallKmer implementation is optimized for speed and suitable only for k-mers up to a certain maximum length. It directly uses its id, a 64 bit integer, for representing a k-mer as a bit field: c bits represent a character $s \in \Gamma$ followed by g bits designating the number of gaps after s. The maximum number n of match positions is determined by $nc + (n-1)g \leq 64$. With the default values of $c = 4$ required for 10 IUPAC characters and $g = 2$, all motifs with at most eleven match positions separated by no more than three gaps between each position can be represented. By far the bigger part of motifs are compassed by these limits. The advantage of the SmallKmer implementation is that all access and mutation operations can be performed very efficiently by simple logical and bit shift operations.

For exceptionally occurring motifs with more than eleven match positions, XXmotif switches to the UniversalKmer implementation which uses a list of an arbitrary number of character/gap pairs. The generality is payed for by loss of efficiency due to list traversal upon access. Since typically only very few motifs reach the required length during the XXmotif extension phase the UniversalKmer representation, this does not carry weight as a whole.

UniversalKmer uses the id field for storing a pointer to a string representation of itself. For building hashes of k-mers, our hash function decides from the isNumeric flag whether to compute a value based on the numerical value of the id field or based on the referenced string.

As the number of match positions within the k-mer is stored directly in the class AbstractKmer, it is possible to use static casts to decide whether an AbstractKmer has to be considered as a SmallKmer or a UniversalKmer, making time consuming virtual function calls unnecessary.

2.3.4. Merging phase

Similar PWMs are merged in order to create a list of non-redundant motifs. First, all motifs are ranked by E-value, and, beginning with the motif having highest significance, similarity tests are performed. Therefore, all less significant motifs are compared to it, and, if similar enough, merged. Afterwards, this procedure is repeated for the second most significant motif, and so on. The criteria for these similarity tests are the following:

1. The divergence between the two PWMs calculated with a normalized Euclidian distance is smaller than 0.25 in an overlapping region of at least length six. This criterion is frequently used to assign a motif to be correctly found in diverse benchmarks (e. g., Gordân et al. (2010), Georgiev et al. (2010)):

   $$D(a,b) = \frac{1}{\sqrt{2}w} \sum_{i=1}^{w} \sqrt{\sum_{L \in \{A,C,G,T\}} (a_{i,L} - b_{i,L})^2} \tag{2.24}$$

   where a and b are the regions of both PWMs that are overlapping and w is the size of the overlap.

2. The overlapping region has an average entropy over the six positions with highest information content of at least 0.5 for both PWMs. This assures that the overlap is within important parts of both PWMs. The restriction to only six positions guarantees that uninformative positions within a binding site do not negatively influence the score:

   $$E(a) = \frac{1}{6} \sum_{j=1}^{6} \left(\sum_{L \in \{A,C,G,T\}} a_{j,L} \log_2 \frac{a_{j,L}}{0.25} \right) \tag{2.25}$$

   where a consists of the six PWM columns of the overlap with highest information content. The use of entropy as a criteria for measuring motif similarity was first described by Gordân et al. (2010).

3. The binding sites of the motif with less sites overlap at least 20% with the binding

sites of the other motif. This criterion assures that no motifs are merged that are only similar in a small overlapping region but different in the surrounding.

Criteria 1 and 2 were also used in our motif sensitivity and metazoan benchmark in order to determine successful motif discovery.

The merged PWM is built from all binding sites of both PWMs and 10% pseudocounts (Durbin et al., 2006). If the length of both motifs is not the same, the length of the motif with the better E-value is chosen. Afterwards, an E-value is calculated for the merged motif. If this E-value is better than the E-values of both unmerged motifs, only the merged motif is kept. Otherwise, only the better of the original motifs is kept.

2.3.5. Refinement phase

The set of non-redundant motifs is now iteratively refined by selecting the most significant motif instances and motif lengths. To decide which sites are functional, putative binding sites are sorted by P-value. For each K, we calculate the probability of observing by chance at least K binding sites with a P-value equal to or better than the K-th best. The K that optimizes the P-value is used to select the sites contributing to the refined PWM (see Section 2.2.3, order statistics). Afterwards, the PWM is updated using 10% pseudocounts (Durbin et al., 2006) and the refinement step is repeated.

To decide whether a different motif length is more significant, all PWMs including up to two more or fewer positions at both ends are tested. For every tested length, order statistics is used to select the most significant motif set. However, the refined PWM with this new length might influence the sorting of the putative binding sites and thus the P-value. Therefore, two iterations of motif set optimization and PWM creation are performed. Afterwards, the motif length having the best P-value is chosen for a new iteration of the refinement phase. To improve runtime, only sites of the unoptimized PWM with a log-odds score greater zero are tested for the most significant motif set.

The observed P-values are corrected for multiple testing and the refinement step is repeated as long as the E-value improves. Finally, the merging and refinement phases are iterated until convergence.

2.4. Overview of the Used Published Motif Finders

In this section we give a short overview of the tools used and describe the parameters chosen for all benchmarks. Generally, we used the default parameters of each tool and added useful optional parameters if provided, e. g., the possibility to search on both strands or to use a multiple occurrence per sequence model. The only exception is MEME for which we used additional arguments suggested within Bailey et al. (2010).

2.4.1. PRIORITY

PRIORITY (Narlikar et al., 2006) is a PWM-based method that refines the motif model using Gibbs-sampling, a Markov chain Monte Carlo (MCMC) method that approximates sampling from a joint posterior distribution by sampling iteratively from individual conditional distributions (Gelfand and Smith, 1990). PRIORITY uses informative priors based on common structural classes of transcription factors to improve the sampling and provides the opportunity to add more priors to the procedure if more information is available, e. g., a nucleosomal prior (PRIORITY-$\mathcal{N}$, Narlikar et al. (2007)), a discriminative prior (PRIORITY-$\mathcal{D}$, Gordân et al. (2010)), an alignment-free conservation prior (PRIORITY-$\mathcal{C}$, Gordân et al. (2010)), or a combination of these.

PRIORITY can only be run with a zero-or-one occurrence per sequence model and cannot optimize the motif length. As sampling is not deterministic, PRIORITY is run many times and the resulting motif is set to the best of these trials.

We used PRIORITY version 2.1.0 for all benchmarks.

Parameters PRIORITY

PRIORITY was run using default parameters, consisting of the supplied third order background model, the default motif length 8 and 50 trials. From the command line it is started using

```
java -jar priority.jar -nogui
```

Parameters PRIORITY-$\mathcal{D}$

To start PRIORITY using the discriminative prior (Gordân et al., 2010), the corresponding prior has to be created. This can be done using a Perl script supplied by the PRIORITY package:

```
./discr_from_pos_and_neg.pl 8 input.fasta negset.fasta input.prior
```

Now, we added the directory containing the $\mathcal{D}$-prior to the PRIORITY `params` file and started the tool as before.

Parameters PRIORITY-$\mathcal{DC}$

The first step to start PRIORITY using the discriminative conservation prior (Gordân et al., 2010) is to create a directory (`homologs/`) consisting of all homologous regions. Afterwards, the $\mathcal{DC}$-prior is created using two Perl scripts supplied by the PRIORITY package:

```
./generate_fastalike_cons_simple.pl input.fasta homologs/ 8 input.info
./generate_fastalike_cons_simple.pl negset.fasta homologs/ 8 negset.info
./discr_INFO_from_pos_and_neg.pl 8 input.fasta negset.fasta input.info
     negset.info input.prior
```

PRIORITY is now started as before with the directory containing the $\mathcal{DC}$-prior added to the `params` file.

2.4.2. MEME

MEME (Multiple Em for Motif Elicitation, Bailey and Elkan (1994)) is a PWM-based motif finding tool that iteratively refines candidate PWMs by an expectation maximization (EM) algorithm. It allows to find the optimal motif length and provides the possibility to add higher-order background models and the priors used by PRIORITY to improve the motif search.

MEME was used in version 4.3.0 in our analysis.

Parameters MEME

As suggested by Bailey et al. (2010), depending on the complexity of the organism, two different parameter settings were used:

```
./meme input.fasta -dna -revcomp -mod zoops -minsites 20 -nmotifs 4
```

- Yeast data sets: `-minw 7 -maxw 12`
- Metazoan data sets: `-minw 8 -maxw 20`

Parameters MEME-$\mathcal{M}$

To test MEME with a higher-order background model, we trained a fifth-order background model using the script `fasta-get-markov` supplied with the MEME package:

```
./fasta-get-markov -m 5 > background.b
```

To incorporate the background model to the motif search, MEME was run using an additional argument:

```
   -bfile background.b
```

Parameters MEME-$\mathcal{D}$

Bailey et al. (2010) demonstrated that the performance of MEME can be further improved by using the discriminative prior from the Hartemink lab (Gordân et al., 2010) as additional information. Therefore, we created the prior file `input.prior` using a Perl script from the Hartemink lab as shown in Section 2.4.1.

Afterwards, we used the script `hartemink2psp` supplied with the MEME package to translate this prior to the `psp` format that can be used as input for MEME:

```
cat input.prior | hartemink2psp -mod zoops -revcomp -width 8 > input.psp
```

Now, MEME can be run with the discriminative prior using an additional argument:

```
   -psp input.psp
```

Parameters MEME-$\mathcal{DC}$

Furthermore, MEME can be run using conservation information by incorporating the discriminative conservation prior from the Hartemink lab (Gordân et al., 2010) as additional information. To create this prior, again Perl scripts from the Hartemink lab have to be used as shown in Section 2.4.1.

Now, as for the discriminative prior, the `hartemink2psp` script supplied by the MEME package is used to translate the prior and MEME is started with the `-psp input.psp` argument.

2.4.3. Weeder

Weeder (Pavesi and Pesole, 2006) is a pattern-based motif finding tool that exhaustively enumerates the motif space. It tolerates mismatches within the patterns and does not need the exact pattern length as input. Internally, the sequences are represented as a suffix tree, which also allows to efficiently enumerate longer patterns. However, it is not possible to use conservation information.

Weeder was used in version 1.4.2 for our analysis.

Parameters

Weeder was started using the optional arguments `S`, to process both strands of DNA, and `M`, to use a multiple occurrence per sequence model:

```
./weederlauncher.out input.fasta speciescode medium S M
```

where `speciescode` was replaced by the respective two letter code.

2.4.4. ERMIT

ERMIT (Evidence Ranked Motif Identification Tool, Georgiev et al. (2010)) is a pattern-based motif finding tool that incorporates quantitative experimental evidence to find a motif pattern that is enriched in sequences with high evidence values. It starts with IUPAC 5-mers and elongates them as long as their enrichment score improves. To incorporate conservation information, binding sites are filtered to the ones that fit to the pattern in all species.

ERMIT was used in version 1.01 for our analysis.

Parameters ERMIT

To run ERMIT, input files have to be parsed from FASTA format into a special format, and a summary file has to be created (`sequence_file`) that contains the location of the input file. Furthermore, an evidence file has to be created with the probabilities assigned to every sequence identifier, and a summary file is needed (`evidence_file`) containing the location of this evidence file. Now, ERMIT can be started using the command line statement:

```
./cERMIT evidence_file sequence_file output chip_chip
```

where `chip_chip` specifies the data type of the input sequences. As the required evidence values per sequence were only available for the yeast ChIP-chip experiments from the Harbison data set (Harbison et al., 2004), this data type was always set to `chip_chip`.

Parameters cERMIT

To run ERMIT with conservation information (cERMIT), the homologous regions of an alignment have to be parsed to the file format required by ERMIT and stored separately for each species. Afterwards, the locations of each of these files have to be added to the summary file `sequence_file`. Now, cERMIT can be started as before:

```
./cERMIT evidence_file sequence_file output chip_chip
```

2.4.5. AMADEUS

AMADEUS (Linhart et al., 2008) is both a pattern- and a PWM-based motif finding tool that starts by enumerating all k-mers of a given length. In the following, these k-mers are merged depending on their similarity and refined by an EM-algorithm. However, it is neither possible to optimize the motif length, nor to incorporate conservation information.

AMADEUS was used in version 1.0 for our analysis.

Parameters

To run AMADEUS from the command line, a parameter file (`params.txt`) has to be created. Therefore, paths for files with all sequences, input set identifiers and negative set identifiers have to be supplied. We used the default parameters, motif length `8`, and running mode `normal` for our analysis:

```
java -Xmx3000m -jar AmadeusPBM_v1.0.jar file params.txt
```

3. XXmotif Web Server

3.1. Overview of Web Servers

The most popular web server for motif discovery is the MEME Suite server (Bailey et al. (2006), Bailey et al. (2009)), within which the PWM-based MEME and GLAM2 motif discovery programs can be run (Bailey and Elkan (1994), Frith et al. (2008)), alongside several related tools to compare the discovered motifs to libraries of literature motifs, and to search for matches to the discovered motifs in sequence databases. With a higher-order background model to describe sequences that should not carry the sought motifs, MEME has shown state-of-the-art performance (Bailey et al., 2010). To use higher-order models, users have to upload their own model file generated using a MEME command line tool, which will limit most users to the zero order model with lower sensitivity. The SCOPE web server combines three pattern-based motif discovery tools, which are specialized to find non-degenerate, degenerate, and gapped motifs, into a single prediction using a "winner takes all" learning rule (Carlson et al., 2007). The RegAnalyst server runs a motif discovery method that searches for the most enriched patterns using fixed thresholds for the maximum number of allowed mismatches. It has been developed for mycobacterial and yeast sequences, on which it was reported to have higher sensitivity than SCOPE (Sharma et al., 2009). The WebMOTIFS server takes gene names from human, mouse or *S. cerevisiae* as input, extracts promoter sequences, launches four motif discovery programs, and displays the results in a uniform format (Romer et al., 2007). RSAT is a web toolbox for regulatory sequence analysis that also offers several simple tools and Gibbs sampling for motif discovery (Thomas-Chollier et al. (2008), Thomas-Chollier et al. (2011)).

Although various published tools can score conservation in multiple sequence alignments of related species and a few can exploit the positional clustering of motifs, to our knowledge none of the web services offer this useful functionality. In contrast, the XXmotif web server available at `xxmotif.genzentrum.lmu.de` can combine enrichment P-values for PWMs with P-values for sequence conservation and for positional clustering of motif occurrences.

3.2. Input

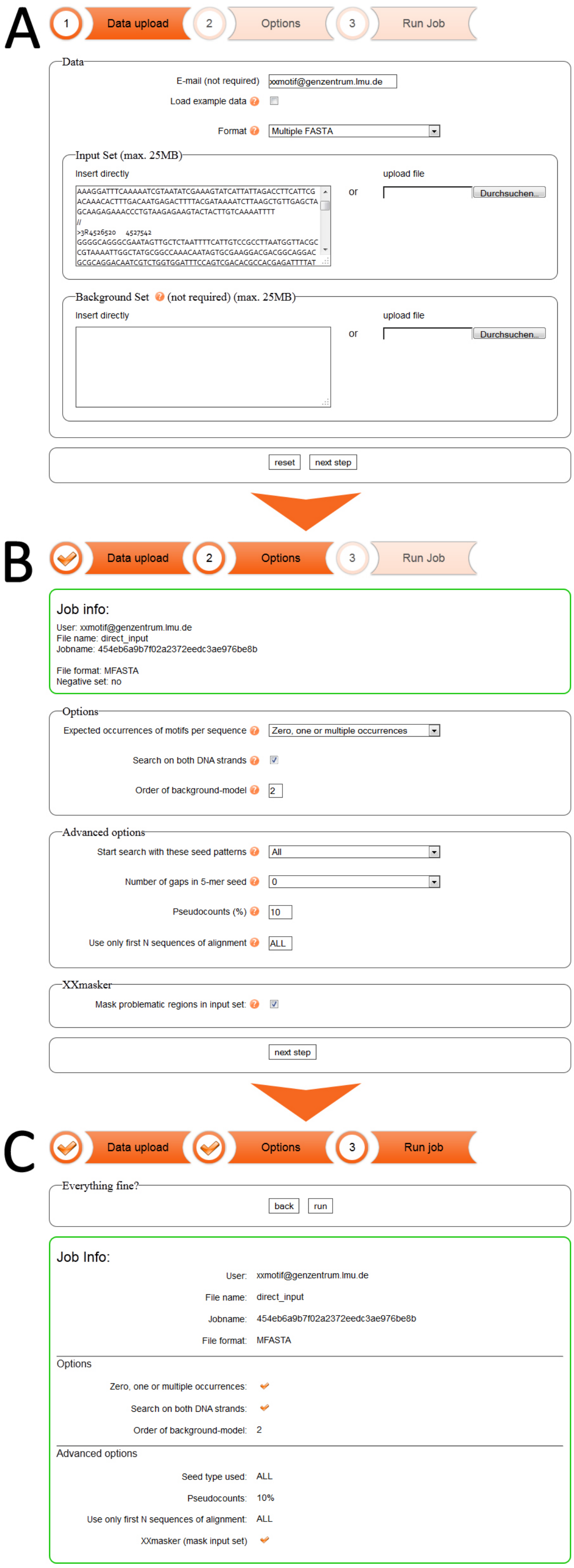
On the “Data upload page” (Figure 3.1A), users can enter the input sequence set and an optional background sequence set. The background sequences are used to learn the statistical background model, which describes how “normal” sequences look like. XXmotif will then try to find motifs that are enriched in the input set in comparison to the expectation derived from the background model. When no background sequences are supplied, a second-order background model is trained from the input sequences. To increase the sensitivity of the motif search, XXmotif can calculate motif conservation P-values during the search, which are combined with the enrichment P-values. In this case, the user can upload a set of input and background multiple sequence alignments, using the “Multiple FASTA” format.

On the “Options” page, the suggested default options can be modified (Figure 3.1B). First, the user can specify how many motif occurrences per input sequence are expected. For most transcription factor and microRNA binding sites we would expect multiple occurrences, for example. For core promoter motifs or splice sites we would expect zero or one occurrence per sequence. When selecting the latter option, only the best occurrence per sequence is scored, whereas with the former option, all occurrences above a certain single-site significance P-value are scored. Searching on both strands is recommended for all motifs that should occur with similar probabilities on both strands (i. e., as reverse complements of each other.) This is true for most transcription factor and microRNA binding sites, for example, but not for core promoter or splice site motifs. The order of the background model specifies how long the patterns are that XXmotif learns from the background sequence set. An eighth-order model learns the frequencies of 9-mer nucleotides to model the correlations between nearby nucleotides. This is the default option selected when a background set is supplied by the user. When the background model has to be learned from the positive set, the default order is set to 2. If an order 8 model were trained in this case from the positive set, no motif shorter than 10 nucleotides could become significant.

Upon pressing “next step”, a summary of all selected options is presented (Figure 3.1C), and corrections can be made using the “back” button. After job submission, the user is directed to a status page, which can be bookmarked and automatically redirects to the results page when the job is finished. If the user has provided an email address, a notification with the result page URL is sent.

3.3. Output

The results page lists the web logos, E-values, and number of sites of matched motifs found up to an E-value of 100, (Figure 3.2A). When both strands were searched, the reverse



E-mail (not required) xxmotif@genzentrum.lmu.de
Load example data ☐
Format Multiple FASTA

```
AAAGGATTTCAAAAATCGTAATATCGAAAGTATCATTATTAGACCTTCATTCG
ACAAACACTTTGACAATGAGACTTTTACGATAAAATCTTAAGCTGTTGAGCTA
GCAAGAGAAACCCTGTAAGAGAAGTACTACTTGTCAAAATTTT
//
>3R4526520    4527542
GGGGCAGGGCGAATAGTTGCTCTAATTTTCATTGTCCGCCTTAATGGTTACGC
CGTAAAATTGGCTATGCGGCCAAACAATAGTGCGAAGGACGACGGCAGGAC
GCGCAGGACAATCGTCTGGTGGATTTCCAGTCGACACGCCACGAGATTTTAT
```

Job info:
User: xxmotif@genzentrum.lmu.de
File name: direct_input
Jobname: 454eb6a9b7f02a2372eedc3ae976be8b

File format: MFASTA
Negative set: no

Expected occurrences of motifs per sequence Zero, one or multiple occurrences
Search on both DNA strands ☑
Order of background-model 2

Start search with these seed patterns All
Number of gaps in 5-mer seed 0
Pseudocounts (%) 10
Use only first N sequences of alignment ALL

Mask problematic regions in input set: ☑

Job Info:
User: xxmotif@genzentrum.lmu.de
File name: direct_input
Jobname: 454eb6a9b7f02a2372eedc3ae976be8b
File format: MFASTA

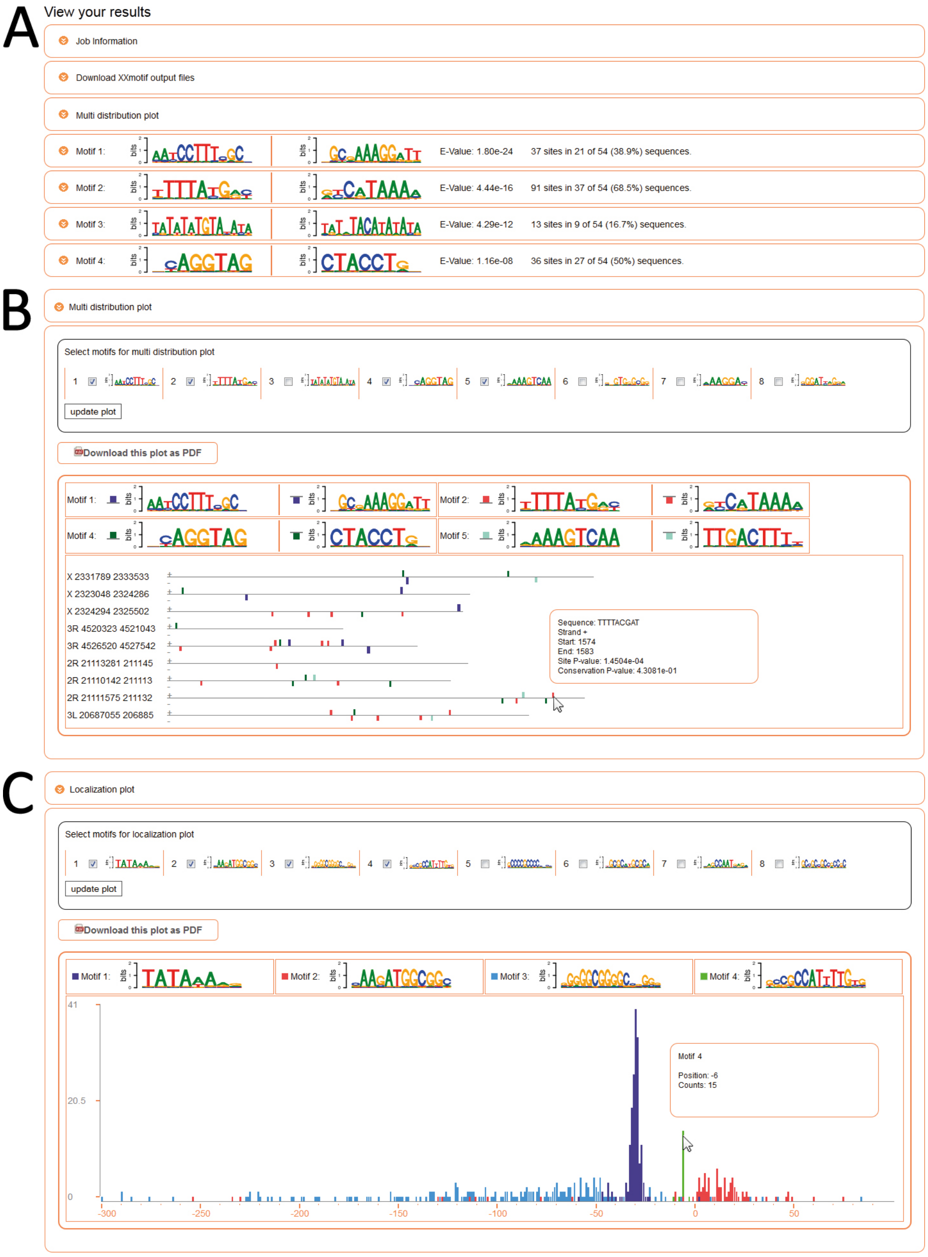
Options
Zero, one or multiple occurrences: ✓
Search on both DNA strands: ✓
Order of background-model: 2

Advanced options
Seed type used: ALL
Pseudocounts: 10%
Use only first N sequences of alignment: ALL
XXmasker (mask input set) ✓

Figure 3.1.: Pages for submitting a job to the XXmotif web server: (A) Upload input and background sequence sets, (B) set options for the motif search, (C) verify and submit.

complement versions of the motifs are also plotted. More detailed results are hidden behind expandable boxes.

The “multi distribution plot” (Figure 3.2B) depicts with colored boxes the position and



Motif 1: E-Value: 1.80e-24 37 sites in 21 of 54 (38.9%) sequences.

Motif 2: E-Value: 4.44e-16 91 sites in 37 of 54 (68.5%) sequences.

Motif 3: E-Value: 4.29e-12 13 sites in 9 of 54 (16.7%) sequences.

Motif 4: E-Value: 1.16e-08 36 sites in 27 of 54 (50%) sequences.

Sequence: TTTTACGAT
Strand +
Start: 1574
End: 1583
Site P-value: 1.4504e-04
Conservation P-value: 4.3081e-01

Motif 4
Position: -6
Counts: 15

Figure 3.2.: Sample results with boxes that can be expanded with the orange buttons on the left. (A) Summary list of discovered motifs sorted by significance (E-value). (B) The “multi distribution plot” depicts positions and strand of motif occurrences on the input sequences. Motifs can be selected in the upper part. The single-site P-values are represented by the height of the box, their length corresponds to the motif length. (C) The “Localization plot” is a histogram view of the positional distribution of selected motifs relative to an anchor point. All plots can be downloaded in PDF format.

strand of significant motif occurrences within the input sequences. The motifs to display in this plot can be selected by the user in the upper part of the plot. This allows to plot clustered binding sites marking, for example, *cis*-regulatory elements, co-occurring pairs of motifs, and other positional biases. Setting the mouse over a particular motif site will show the site's sequence, strand, start and end position, the single-site P-value measuring the match quality with the PWM, and a conservation P-value (if multiple sequence alignments had been supplied). Only sequences with at least one motif site are shown. Most significant motifs are drawn last and may hide less significant ones.

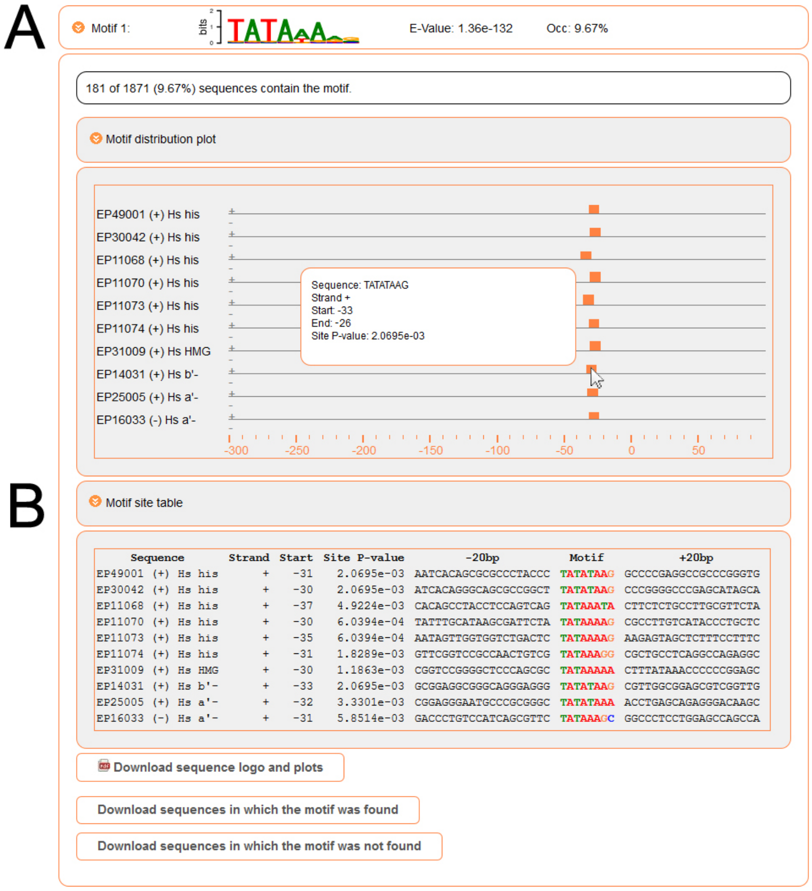
When the input sequences are all of the same length, a "localization plot" can be displayed (Figure 3.2C). This graph is useful to analyze positional preferences with respect to the fixed-length sequence window of the input sequences. It shows in a histogram view the positional distributions of all user-selected motif occurrences with each motif in a different color. For instance, motif 1 (TATA-box) in Figure 3.2B is exactly positioned between -33 to -27 bp with respect to the transcription start site (TSS) at position 0, whereas motif 2 (YY1) is located mainly downstream of the TSS. Mouse-over in the histogram provides the position with respect to the anchor point and the number of counts of the motif. For instance, Motif 4 in Figure 3.2C has 15 counts sharply peaked at position -6 with respect to the TSS, and a "CA" dinucleotide at position -1, indicating an initiator like function.

Detailed information about each motif can be obtained by clicking the expand buttons in the motif summary list. Two single motif graphs can then be viewed (Figure 3.3). The "motif distribution plot" is similar to the "multi distribution plot" and indicates the positions of significant matches of the selected motif on the input sequences. The "motif site table" lists all significant matches with their sequence identifiers, strands, positions, the single-site P-values, and the sequence contexts of the motif.

All plots can be downloaded with the buttons below them. All data files generated by the XXmotif program, such as lists of motifs with their occurrence positions, P-values, and site sequences, PWM weight coefficients, and images of motif logos can be downloaded by expanding the box "Download XXmotif output files".

3.4. Documentation

Two sample input sets and pre-computed results allow the user to get a quick overview of the server's usage and results. Help buttons and mouse-over explanations are available for all input options. More general help is listed on the FAQ page.



181 of 1871 (9.67%) sequences contain the motif.

Sequence: TATATAAG
Strand +
Start: -33
End: -26
Site P-value: 2.0695e-03

Sequence	Strand	Start	Site P-value	-20bp	Motif	+20bp
EP49001 (+) Hs his	+	-31	2.0695e-03	AATCACAGCGCGCCCTACCC	TATATAAG	GCCCCGAGGCCGCCCGGGTG
EP30042 (+) Hs his	+	-30	2.0695e-03	ATCACAGGGCAGCGCCGGCT	TATATAAG	CCCGGGGCCCGAGCATAGCA
EP11068 (+) Hs his	+	-37	4.9224e-03	CACAGCCTACCTCCAGTCAG	TATAAATA	CTTCTCTGCCTTGCGTTCTA
EP11070 (+) Hs his	+	-30	6.0394e-04	TATTTGCATAAGCGATTCTA	TATAAAAG	CGCCTTGTCATACCCTGCTC
EP11073 (+) Hs his	+	-35	6.0394e-04	AATAGTTGGTGGTCTGACTC	TATAAAAG	AAGAGTAGCTCTTTCCTTTC
EP11074 (+) Hs his	+	-31	1.8289e-03	GTTCGGTCCGCCAACTGTCG	TATAAAGG	CGCTGCCTCAGGCCAGAGGC
EP31009 (+) Hs HMG	+	-30	1.1863e-03	CGGTCCGGGGCTCCCAGCGC	TATAAAAA	CTTTATAAACCCCCCGGAGC
EP14031 (+) Hs b'-	+	-33	2.0695e-03	GCGGAGGCGGGCAGGGAGGG	TATATAAG	CGTTGGCGGAGCGTCGGTTG
EP25005 (+) Hs a'-	+	-32	3.3301e-03	CGGAGGGAATGCCCGCGGGC	TATATAAA	ACCTGAGCAGAGGGACAAGC
EP16033 (-) Hs a'-	+	-31	5.8514e-03	GACCCTGTCCATCAGCGTTC	TATAAAGC	GGCCCTCCTGGAGCCAGCCA

Figure 3.3.: Detailed motif view. The first box ("Motif distribution plot") plots the position of significant motif matches within the input sequences. The second box ("motif site table") gives detailed information on all significant motif matches.

3.5. Implementation

The XXmotif web server was mainly implemented by Sebastian Luehr within a practical course. It runs on an Apache server using PHP, PERL, and R scripts. The user interface is dynamically generated HTML content with JavaScripts from the jQuery library. Submitted jobs are processed on a Scientific Linux computer cluster.

3.6. Conclusion

With the XXmotif web server, we aim to make a very sensitive and reliable motif discovery method easily accessible to non-expert users. The server has clearly structured input and results pages and offers various useful interactive analyses. It is unique in being able to include evidence from motif conservation and positional clustering in the motif search.

Part II.

Analysis of Transcription Factors and Core Promoters

4. Transcription Factors

4.1. Overview

Transcription factors (TFs) are key regulators in every biological network. Their specific binding to DNA activates or represses the expression of genes which in turn influences the number of transcribed primary RNAs and proteins (Latchman, 1997). Since regulation occurs in most cases at the level of transcription, understanding the behavior of TFs is the most critical step in understanding the complex mechanisms within biological networks.

The specific binding to DNA is achieved by DNA-binding domains, which combine the recognition of nucleotides with the readout of the DNA shape. The only places on the DNA where TFs are able to contact the DNA specifically enough to recognize distinct nucleotides are the major and minor grooves (see Figure 4.1).

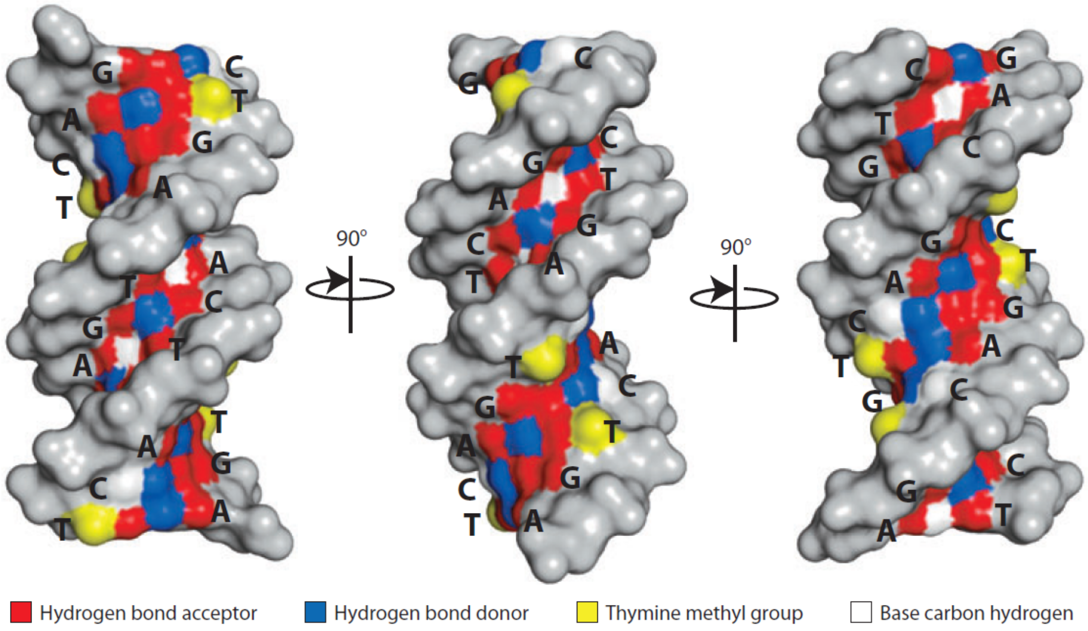


Figure 4.1.: Base recognition in the major and minor groove of DNA (figure taken from Rohs et al. (2010)). The DNA stretch contains the dodecamer d(GACT)$_3$. The three panels show successive rotations of 90° around the helix axis.

In the major groove, hydrogen bonds and hydrophobic contacts allow the TF to specifically recognize which nucleotides are present. All four possible base pairs, i. e., A:T, T:A, C:G, and G:C have a unique pattern of hydrogen bond acceptors (red), hydrogen bond donors (blue), thymine methyl groups (yellow), and base carbon hydrogens (white). In contrast, in the minor groove A:T versus T:A and C:G versus G:C are indistinguishable. Therefore, most DNA-binding domains form hydrogen bonds with bases in the major groove. Examples are HTH domains (e. g., homeodomains, λ repressor), zinc finger domains (e. g., TFIIIA), immunoglobin fold domains (e. g., p53, NFκB, STAT, and NFAT), and the N-terminal end of basic leucine zipper (bZip) domains (Garvie and Wolberger, 2001).

Although the contacts in the minor groove cannot distinguish A:T from T:A and C:G from G:C, some families of DNA-binding domains use the minor groove to improve specificity. For instance, homeodomain proteins have N-terminal arms that dock in the minor groove in addition to their contacts in the major groove (Gehring et al., 1994b). Moreover, a special class of TFs, called "architectural proteins", binds purely in the minor groove. These proteins play crucial roles in the assembly of large protein-DNA complexes and provide the possibility of a controlled assembly of the macromolecular complex by a strong bending of the DNA (e. g., TATA-box, IHF, HMG-box).

Additional to the described "base readout", the specificity of a TF can be further increased by a mechanism called "shape readout". In case of "local shape readout", the TF improves the specificity by exploiting local deviations from ideal B-DNA. As an example, AT-rich sequences lead to a narrowing of the minor groove causing a negative electrostatic potential recognizable by arginines (Rohs et al., 2009). An interesting example for this "local shape readout" are the members of the Hox family of TFs. All of them form the same major group interactions but gain specificity by the Extradenticle (Exd) cofactor that forms contacts to the minor group if presented to the correct DNA structure (Joshi et al., 2007).

In case of "global shape readout" the entire binding site is not in a classical B-form helix. An example for the importance of this readout is the papillomavirus E2 protein that binds as dimer two half sites separated by four nucleotides. However, in order to have a strong interaction with the DNA, the consensus binding site ACCG N_4CGGT is not sufficient; the DNA has also to be bent (Hegde et al., 1992). Intrinsically bent DNA is correlated with sequences containing A-tracts, i. e., stretches of A:T base pairs that include ApA (TpT) and ApT, but not TpA steps (Nelson et al., 1987).

Hence, the binding energy of TFs can be influenced by parts of the DNA that are not directly contributing to the binding site but alter the global DNA structure. To model the binding energy of a TF accurately, it is therefore also necessary to model the regions between and around interacting regions of the TF and the DNA. Moreover, as the DNA shape is mainly based on the succession of two base pairs that do not have to be positioned relative to the interacting regions, modeling PWMs as higher-order PWMs might improve the model accuracy. E. g., in order to generate a bent stretch of DNA it is not important at what position an ApT step is located, however, it is important that the base pairs before this step are exclusively adenines, and the base pairs following this step are exclusively thymines.

4.2. Families of DNA-Binding Domains

This section gives an overview about techniques of TFs to contact the major and minor grooves of DNA, reveals important features of binding sites, and gives information about how to incorporate such features within a *de novo* motif search. As examples, the four most abundant DNA-binding domains (DBDs) are described. All of them are relatively small domains consisting of less than 100 amino acids.

4.2.1. Helix-turn-helix motif

The helix-turn-helix (HTH) structural motif is the most widely used DNA-binding domain in prokaryotes. *Escherichia coli*, for instance, contains 270 TFs that belong to 11 different families, with 10 of them utilize the HTH motif (Babu and Teichmann, 2003). This structural motif is formed by two α-helices connected by a turn and was first described by Anderson et al. in 1981 (see Figure 4.2).

Amino acid residues in the first α-helix, the recognition helix, contact the major groove of DNA and thus mediate sequence specificity. Many TFs bind as dimers with two similar HTH-motifs, for example CRO, CAP, and λ repressor (Brennan and Matthews, 1989). Such a complex binds to two adjacent major grooves, which leads to palindromic patterns for the

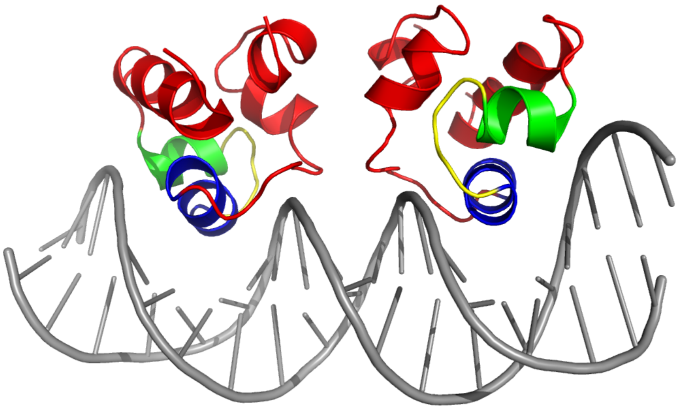


Figure 4.2.: Overall view of the complex between 434 CRO and the OR1 binding site (PDB accession number: 3CRO). The recognition helix is depicted in blue, the turn in yellow and the second helix of the helix-turn-helix motif in green. The TF binds as a dimer, leading to a palindromic binding site.

bound sequence stretches. However, in between the bound regions, the sequence motif has not to be palindromic.

Therefore, in order to find occurrences of HTH motifs using a motif finding tool, it is important to use palindromic seeds allowing gaps between both parts of the two palindromic regions. Furthermore, it is important to allow an extension of these palindromic motifs that does not have to be palindromic anymore.

4.2.2. Homeodomain

Eukaryotes have found many more solutions than prokaryotes to build a scaffold that allows for a sequence specific DNA binding. Nevertheless, the prokaryotic HTH motif is exploited in eukaryotes as well, with homeodomain proteins being the most frequent class. The primary function of this class of proteins is to regulate the expression of other genes involved in development and differentiation. Homeobox (Hox) genes code for homeodomains and are often organized into chromosomal clusters, termed HOX clusters. An intriguing feature of these clusters is the correlation between the physical location of the gene within the cluster and the expression pattern along the anterior-posterior body axis. Moreover, all Hox genes are transcribed in the same direction (Carroll (1995), Kmita and Duboule (2003)).

The homeodomain motif consists of 60 amino acids that are structured as three alpha helices connected by short linkers. The second and third alpha helix form the prokaryotic HTH motif. The main sequence specificity is provided by helix three that binds to the major groove of DNA. The core motif of this interaction is TAAT. However, as this motif matches many sites by chance, amino acid residues at the N-terminus of the homeodomain improve specificity by reaching into the minor groove. In addition, some TFs combine several DBDs or form homodimeric or heterodimeric complexes (Gehring et al., 1994a). For instance, the POU domain (originally defined based on the four genes *Pit-1*, *Oct-1*, *Oct-2*, *unc-86*) combines a classical homeodomain, called POU homeodomain, with a POU-specific domain. This POU-specific domain that is required for a cooperative, high affinity binding, is approximately 80 amino acids long, and consists of four alpha-helices connected to the POU homeodomain by a variable linker (Herr et al., 1988).

4.2.3. C_2H_2 zinc finger

C_2H_2 zinc fingers, first identified in studies of the *Xenopus laevis* TF TFIIIA (Miller et al., 1985) are one of the most common DNA-binding motifs in eukaryotic TFs (Wolfe et al., 2000). Typically, several zinc fingers are arranged in tandem along the DNA with each finger having a conserved $\beta\beta\alpha$ structure. The most abundant subtype, called the “Krüppel-type”, comprises of 28 amino acids that coordinate a single zinc atom by paired cysteine and histidine residues (see Figure 4.3).

Positions −1, 2, 3, and 6 (relative to the N-terminus of the alpha helix) are critical for DNA specificity (Choo and Klug, 1994). Each finger defines binding specificity to around three adjacent nucleotides. However, although several research groups have designed algorithms to predict from the amino acid sequence which nucleotides a given zinc finger protein preferentially binds (e. g., Kaplan et al. (2005), Persikov et al. (2009)), it is still unclear whether these algorithms can accurately predict the binding preferences in normal cellular contexts.

Consequently, the tandemic orientation of C_2H_2 zinc fingers leads to binding sites having multiple important positions next to each other. Hence, in order to find these kind of sites it is best to use gapless seeds as starting points of a motif analysis.

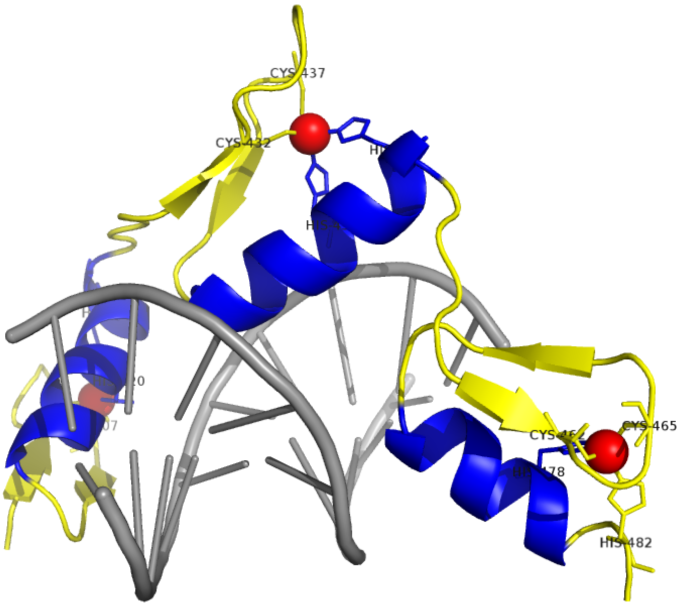


Figure 4.3.: Overall view of the complex between Krüppel-like factor 4 (Klf4) and DNA (PDB accession number: 2WBU). Klf4 consists of three C_2H_2 zinc finger domains arranged in tandem orientation. Each zinc finger consists of two beta sheets (yellow) and an alpha helix (blue). All of them contain two cysteine and two histidine residues that coordinate a single zinc atom (red). Side chains and labels of these residues are shown.

4.2.4. C_6-zinc cluster family

A special class of zinc-containing motifs exists in several TFs from fungi. Here, the DNA binding domain contains a cluster of two zinc atoms liganded to six cysteine residues (see Figure 4.4). The first structure of a protein of this class, GAL4, was determined by Marmorstein et al. in 1992. It consists of two dimerizing subunits, which can be divided into three structurally and functionally distinct regions: a dimerization domain at the C-terminus, a zinc cluster domain at the N-terminus, and a nine-residue long linker region connecting them.

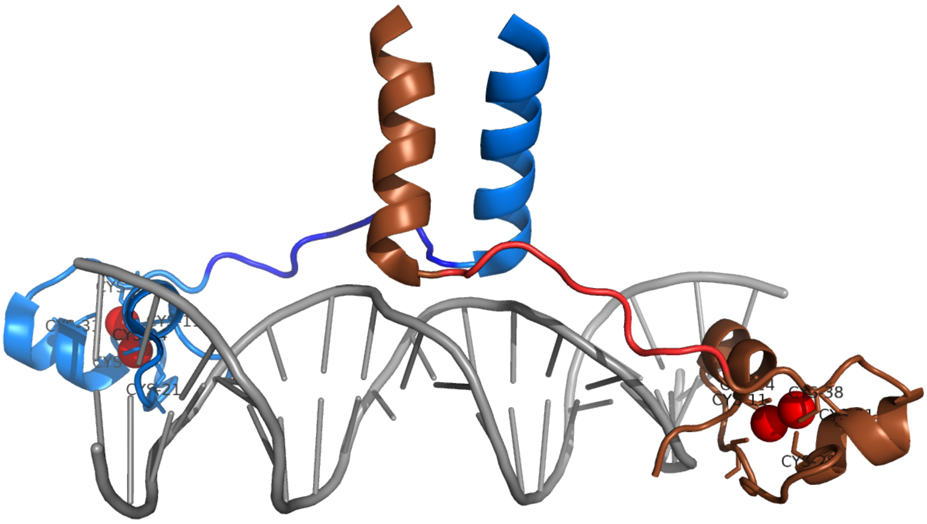


Figure 4.4.: Overall view of the complex between GAL4 and DNA (PDB accession number: 1D66). GAL4 consists of two dimerizing monomers, each divided into three distinct regions: a dimerization domain, a linker region, and a Zn_2Cys_6 zinc cluster region. The two dimerization alpha helices are packed into a coiled coil structure. The length of the linker region determines the distance between both DNA binding sites.

The function of the zinc cluster is to ensure the proper folding of the DBD that binds to a highly conserved CCG sequence. However, also the linker region contributes to DNA specificity. Several nonspecific contacts between this region and phosphate groups of the DNA backbone following the minor groove from the dimerization domain to the zinc cluster allow the TF to recognize the shape of the DNA strand. Moreover, different lengths of this linker region allow both DNA binding sites to be separated by different distances.

In order to find this kind of motifs, a motif finding tool needs seeds allowing for large gaps in between two informative regions. In case of GAL4, the distance between the first CGG site and the second palindromic CCG site is 11. However, in principle all distances are possible.

4.3. Experimental Techniques for in Vitro Analysis of TFBSs

Different experimental approaches have been developed to analyze TFs, their binding affinities and binding sites. However, classical techniques like DNA footprinting (Galas and Schmitz, 1978) or electrophoretic mobility shift assays (EMSA, Hellman and Fried (2007)) are laborious and expensive. A moderate to high resolution binding profile is therefore only available for around 20% of all TFs. For instance, the largest databases for TF binding specificities, Jaspar (Bryne et al., 2008) and UniPROBE (Newburger and Bulyk, 2009), list currently only around 500 profiles for human and mouse. However, mammals have approximately 1300 to 2000 TFs (Vaquerizas et al. (2009), Fulton et al. (2009)).

New assays have been developed that allow for a high-throughput analysis of TF binding specificities *in vitro*. Important techniques are protein-binding microarrays (Bulyk et al., 2001), HT-SELEX (Jolma et al., 2010), and HiTS-FLIP (Nutiu et al., 2011).

4.3.1. Protein-binding microarrays

Protein-binding microarrays (PBMs) are microarrays containing double-stranded DNA sequences at the surface. The TF of interest is incubated to the array and subsequently washed away to remove weakly bound proteins. Now, fluorescent-labeled antibodies are used to measure the amount of proteins at each spot, which can be interpreted computationally as an estimate for the binding specificity of the TF. The most recent designs of PBMs contain around 44000 different spots of 60 bp long double-stranded DNA. The sequences are designed to contain each possible 10 bp sequence, of which the vast majority are in different spots (Berger and Bulyk, 2009).

The main disadvantage of PBMs is the limitation to short binding sites that cannot capture dimeric binding of, for instance, dimeric RFX-proteins, or the characterization of orientation and spacing of heterodimeric TF pairs (Jolma et al., 2010).

4.3.2. HT-SELEX

In HT-SELEX (high-throughput systematic evolution of ligands by exponential enrichment), the binding specificity of a TF is obtained by an *in vitro* selection of target sites from a pool of randomized DNA strands. After complex formation between the TF and the DNA, bound DNA ligands are separated from free DNA using affinity capture, EMSA or materials that bind to proteins but not to free double-stranded DNA. Afterwards, the bound DNA sequences are amplified by polymerase chain reaction (PCR), sequenced and used as the DNA pool in the next round of selection. Hence, the resulting DNA library is enriched with bound DNA sequences. After typically three to seven rounds, the binding profile of the TF of interest can be derived from the enrichment of DNA sequences in each round.

Disadvantages of this approach are possible biases from the resin- or filter-based selection step and the dependencies on the number of performed cycles. In later cycles only the high affinity sites are selected, whereas in the first cycles the number of bound sequences is very low. As all DNA fragments can have a different sequence, the covered sequence space is very high (10^{15} sequences, $\sim$ 25bp). In order to analyze even longer and very specific binding sites, a protocol called DNA immunoprecipitation (DIP) can be applied. Here, fragmented genomic DNA is used instead of synthetic random sequences, which assures that every binding site is present at least once within the DNA pool (Liu et al., 2005).

4.3.3. HiTS-FLIP

High-throughput sequencing - fluorescent ligand interaction profiling (HiTS-FLIP) is an *in vitro* technique for both high-throughput as well as quantitative measurement of binding affinity. In a first step, $\sim$100 million clusters of genomic or random synthetic DNA sequences are anchored on a microfluid flow cell. After denaturing the double-stranded DNA and washing away the second strand, double-stranded DNA is rebuild using sequencing by synthesis (Bentley et al., 2008). Now, hundreds of distinct clusters can be assigned on the flow cell using the Illumina Genome Analyzer, each consisting of hundreds of identical DNA molecules. To analyze DNA binding affinities, fluorescently tagged proteins are added to the flow cell at different concentrations. After an optional washing step, the binding of each protein to a cluster is quantified by visualizing the fluorescent using the same camera as before used in the sequencing step. Binding k-mers are detected by an enrichment analysis within the bound clusters.

The main advantage of HiTS-FLIP is the measurement of tens to hundreds of millions of binding events. This allows for the detection of complex interdependencies between motif positions and the analysis of more complex binding events.

4.4. Experimental Techniques for in Vivo Analysis of TFBSs

In vitro analysis of binding specificities of a TF is not enough to reliably predict the binding patterns of the factor *in vivo*. In the genome, the occupancy of a TF on the DNA is not only determined by its specificity, it is also strongly affected by the occupancy of nucleosomes, higher-order chromatin structure that affects accessibility, protein-protein interactions, and co-operative interactions mediated by DNA bending and/or unwinding. In addition, the genome is not a random sequence, and the accessible regions (e. g., promoters) are not similar in sequence to the whole genome. Hence, it is necessary to devise a background model that corrects for these biases, which is nontrivial (Hughes, 2011).

Several approaches haven been developed to examine the genomic locations bound by a specific TF. One important approach is based on endonucleases, another on cross-linking and antibodies.

4.4.1. DNase I hypersensitive sites

If a TF binds to DNA, it simultaneously protects the bound stretches from digestion when exposed to an endonuclease, in particular DNase I (Galas and Schmitz, 1978). *In vivo*, this general approach of DNase I digestion can be used to get a global view of open versus closed chromatin structures (Shibata and Crawford, 2009).

If the interest is not only on revealing all regulatory regions within the genome but to analyze the binding sites of a specific TF, footprints of all bound factors are not sufficient. In some cases, bioinformatics analyses are able to predict which TFs bind to a particular regulatory region (Quitschke et al., 2000). However, techniques like ChIP-chip or ChIP-seq offer more direct results to solve this kind of problems.

4.4.2. Chromatin immunoprecipitation (ChIP)

Since the 1960s it has been known that formaldehyde can be used to cross-link proteins to nucleic acids (Perry and Kelley, 1966). Hence, using antibodies against a TF in a sheared cell treated with formaldehyde allows for a specific isolation of DNA segments bound by the factor, commonly known as chromatin immunoprecipitation (ChIP) (Solomon and Varshavsky, 1985).

In order to analyze the bound DNA regions, isolated DNA sequences have to be mapped to the genome. The oldest mapping technique is based on microarray chips, called ChIP-chip (Ren et al., 2000). These microarrays cover the whole genome, or at least all non-coding regions, allowing for every DNA sequence to hybridize with its complementary strand. As every spot on the chip is linked to a genomic location, the measurement of fluorescent intensities on the chip can be directly transformed to a binding profile on the genome. However, since the main part of the measured intensities is due to indirect effects, data normalization is very important (Siebert et al., 2010). The second mapping technique uses massive parallel sequencing, called ChIP-seq (Mardis, 2008). Here, over 10 million oligonucleotides between 20 and 100 bases are sequenced per experiment. These stretches are read from the end of the ChIP-isolated DNA and are subsequently mapped to the genome. The binding profile of the TF is resolved by counting the number of mapped DNA stretches per genomic position. The resolution of ChIP-seq experiments is shown to be around 40 nucleotides (Venters and Pugh, 2009). However, this resolution can be dramatically improved using the recently developed ChIP-exo protocol (Rhee and Pugh, 2011). Here, an exonuclease trims the immunoprecipitated DNA to a precise distance to the site of crosslinking, allowing to achieve nearly single base resolution.

A general disadvantage of ChIP approaches is the dependence on an available antibody that has to be very specific for the protein of interest. In addition, the resulting binding profile is not limited to direct binding events, it consists of all binding events of all proteins the protein of interest is interacting with.

5. Performance of XXmotif on TFBSs

5.1. Benchmarks

Several benchmarks have been proposed to assess the performance of motif finding tools. Benchmarks based on artificially created test sequences containing randomly placed occurrences of known motifs (Tompa et al. (2005), Sandve et al. (2007)) have the advantage of being easily evaluable since the true sites are known, but it is questionable how transferable these results are to biological sequences.

In our first benchmark, we utilize one of these artificial benchmark settings by testing XXmotif on the benchmark set described by Sandve et al. (2007). It is a pure motif finding task done on TFBS implanted in artificial or real data sets. In the following, this benchmark will be called the “Drabløs benchmark”, named by the last author.

For our second and third benchmark we use the large biological data set of Harbison et al. (2004). It is the most widely employed test set for motif discovery tools (e. g., Georgiev et al. (2010), Gordân et al. (2010)) and consists of lists of *S. cerevisiae* intergenic regions that were significantly enriched in 352 ChIP-chip experiments using 203 tagged transcription factors, 82 of which assayed under several conditions (Harbison et al., 2004). In our second benchmark, referred to as “Motif sensitivity benchmark”, we evaluate how many times a motif finding tool is able to detect the correct motif within the ChIP-enriched regions. In our third benchmark, referred to as “PWM quality benchmark”, we evaluate how well the ChIP-chip data can be described by the found PWM.

As a fourth benchmark we analyze TFs from higher eukaryotes as well as miRNAs. The benchmark was first described by Linhart et al. (2008) and is also used to measure the runtime of XXmotif and the other used motif finding tools on large data sets.

5.1.1. Drabløs Benchmark

In 2007, Sandve et al. published a benchmark suite consisting of three benchmark sets, called “Algorithm Markov”, “Algorithm Real”, and “Model Real”, which are all artificially generated but consist of real binding site fragments extracted from the TRANSFAC database

(Wingender et al., 1996). Each benchmark set consists of data sets containing between 5 and 78 sequences, in which every sequence contains exactly one implanted motif site.

The performance of the benchmark is measured by the Matthews correlation coefficient working on the nucleotide level (nCC). This allows not only to evaluate whether there is an overlap between the prediction and the correct motif, but also whether the correct motif boundaries are predicted. The Matthews correlation coefficient is a balanced measure that can be used even if the classes are of different size, which is the case as there are only few positions bound by the factor (positives), opposed to the unbound positions (negatives):

$$nCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (FP + TN) \cdot (TN + FN) \cdot (FN + TP)}} \tag{5.1}$$

Sequences were either created as real genomic sequences ("Real" data sets) or as third order Markov sequences ("Markov" data sets). In case of Markov sequences, the motif sites are randomly implanted. Additionally, data sets were separated by a discrimination score measuring the similarity of the binding sites from the surrounding sequences. Data sets with high discrimination scores were used for the "Algorithm" benchmark suite, whereas data sets with low discrimination scores were used for the "Model" benchmark suite specifically built to test more sophisticated motif descriptions like higher-order PWMs. However, as only few motif instances are available within these sets, an improvement by using higher-order descriptions seems unrealistic. The "Algorithm Markov" set as well as the "Algorithm Real" set consist of 50 data sets, each containing between 5 and 18 sequences. As every sequence contains exactly one motif site, motif finding algorithms are challenged to work only on very small sample sizes. The "Model Real" set consists of only 25 data sets, however, each containing between 18 and 78 sequences.

The summarized results for all three benchmark sets are shown in Figure 5.1. For all tested tools, the average nCC value is given for each benchmark set. Detailed results for each data set are shown in Figure A.1. As it is necessary to know the exact binding sites of the predicted motif instances to calculate the Matthews correlation coefficient on the nucleotide level, only tools could be tested that output this information. Therefore, only Weeder, MEME, and XXmotif were used for comparison. On the "Algorithm Markov" and "Algorithm Real" benchmark sets, XXmotif significantly outperforms both of the other tested tools. On the "Model Real" benchmark set, however, XXmotif and Weeder show nearly similar performance, whereas MEME predicts nearly no correct motif instance at all. A possible explanation for the better performance of Weeder for the "Real" data sets in comparison to the "Markov" data set might be the used background model. Weeder uses as background model with precalculated 6-mer and 8-mer counts given an organism. However, as no organism is given for any of the data sets, we tested all possible organisms and chose *M. musculus* which performed best. Using such a background model is an advantage in comparison to MEME and XXmotif which use only a second-order background model trained on the input set. Contrary, on the "Markov" data set which consists of sequences created from a third order Markov model, the information within the higher orders of the background model of Weeder is useless.

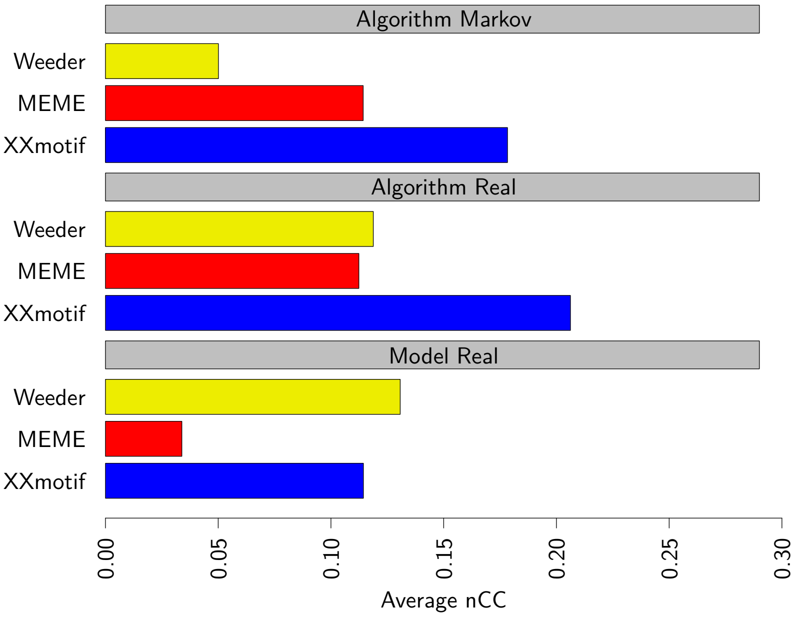


Figure 5.1.: Results of the Drabløs benchmark for three different benchmark sets (Algorithm Markov, Algorithm Real, and Model Real). The quality of motif predictions is measured by the Matthews coefficient on the nucleotide level (nCC).

5.1.2. Motif sensitivity benchmark

To test the sensitivity of XXmotif, we applied the most widely used benchmark (Georgiev et al. (2010), Gordân et al. (2010) on genome-wide yeast ChIP-chip data obtained from Harbison et al. (2004). It consists of lists of *S. cerevisiae* intergenic regions that were significantly enriched in 352 ChIP-chip experiments using 203 tagged transcription factors, 82 of which assayed under several conditions. For a subset of 80 transcription factors and 156 experiments, Harbison et al. (2004) found a published motif as a gold-standard reference. We gave the general purpose motif discovery tools the positive and negative sets of intergenic sequences as described in Harbison et al. (2004), while ERMIT (Georgiev et al., 2010) was supplied with all intergenic sequences and with the set of published ChIP-chip enrichment P-value for each sequence and each experiment. As described by Harbison et al. (2004), only experiments having at least ten sequences with a ChIP-chip P-value < 0.001 were considered.

In addition to the gold standard set of literature motifs described by Harbison et al. (2004) ("Harbison set"), we used two more recent data sets of literature motifs obtained by protein binding microarray (PBM) experiments ("Bulyk set": 56 motifs matching to 101 experiments (Zhu et al., 2009), and "Hughes set": 72 motifs matching to 126 experiments (Badis et al., 2008)). We initially defined a correctly detected motif as having a normalized Euclidean distance smaller than 0.25 in an overlapping region of length ≤ 6, as in Georgiev et al. (2010) and Gordân et al. (2010):

$$D(a,b) = \frac{1}{\sqrt{2}w} \sum_{i=1}^{w} \sqrt{\sum_{L \in \{A,C,G,T\}} (a_{i,L} - b_{i,L})^2} \tag{5.2}$$

where a and b are the regions of both PWMs that are overlapping and w is the size of the overlap. But, when working with the "Bulyk set" of reference motifs, we realized that the definition needs to be extended by the additional requirement of a minimum entropy in the overlapping part of both matrices, as was done by Gordân et al. (2010). This precludes counting motifs as correct that have an overlap only in non-informative regions, which occurred frequently with the "Bulyk set" due to low information content in the outer positions of these PWMs. Therefore, we additionally require the average relative entropy per position over the 6 positions with highest information content in the overlapping region to be at least 0.5 for both PWMs:

$$E(a) = \frac{1}{6} \sum_{j=1}^{6} \left(\sum_{L \in \{A,C,G,T\}} a_{j,L} \log_2 \frac{a_{j,L}}{0.25} \right) \tag{5.3}$$

where a consists of the six PWM columns of the overlap with highest information content.

We measured the sensitivity of the motif discovery tools in the same way as was done previously (Harbison et al. (2004), Linhart et al. (2008), Georgiev et al. (2010), Gordân et al. (2010)). For each tool, we counted the number of successfully identified motifs within the top one and top four predictions (Figure 5.2, Table A.1). Tools that can include conservation information where tested in both versions. When including conservation, the four yeast species in the *sensu strictu Saccharomyces clade* were used for comparison. Alignments were extracted from the UCSC 7-way yeast alignment (sacCer2) (Blanchette et al., 2004).

XXmotif without conservation information found 217 correct motifs cumulated over all three data sets, 39% more than PRIORITY-$\mathcal{D}$ (Gordân et al., 2010) with 156, the next best general-purpose tool, and 22% more than ERMIT (Georgiev et al., 2010), which is specialized to motif discovery on ChIP-chip and ChIP-seq data. With conservation, XXmotif-$\mathcal{C}$ detected 220 correct motifs, 41% more than PRIORITY-$\mathcal{DC}$ (Gordân et al., 2010) with 156 correct motif predictions. Interestingly, the background model is critical to avoid ranking false motifs as top candidates. The standard version of MEME (Bailey and Elkan, 1994) uses a

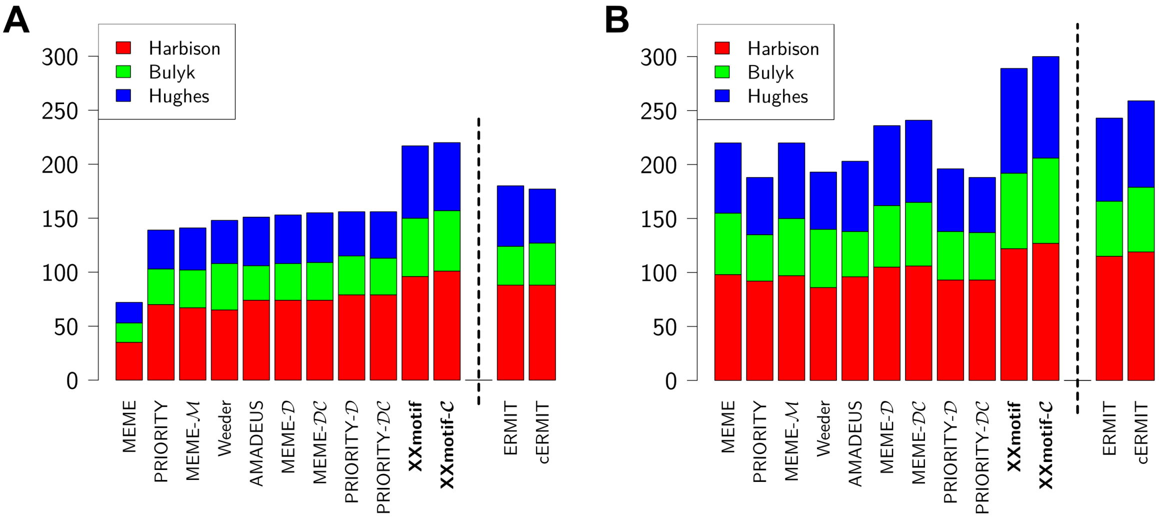


Figure 5.2.: Sensitivity of motif discovery tools on yeast ChIP-chip data. Number of correctly predicted transcription factor binding motifs within the top 1 (A) or top 4 predictions (B). Predictions are based on ChIP-enriched intergenic regions from 352 ChIP-chip experiments (Harbison et al., 2004). Three experimental reference sets are used to judge the correctness of motifs (red, green, blue). The dashed line separates the general-purpose motif discovery tools from ERMIT, which needs ChIP enrichment P-values as additional information. In the tool names, $\mathcal{M}$ indicates a fifth-order Markov model, $\mathcal{C}$ the use of conservation, and $\mathcal{D}$ the discriminative prior from the Hartemink lab (Gordân et al., 2010).

zeroth-order background model trained on the input set and scores only 72 correct motifs among its top predictions. Replacing its zeroth-order background model with a fifth-order Markov model learned from the negative set (MEME-$\mathcal{M}$) raises this number to 141. This can be further increased to 153 by using the discriminative prior from the Hartemink lab (MEME-$\mathcal{D}$, Bailey et al. (2010)). We analyzed the influence of the background model by running XXmotif with interpolated Markov models of order 0 to 9 (Figure 5.3). The improvements were quite dramatic up to order 2, but became much smaller above.

When considering the top four predictions (Figure 5.2B), MEME with a zeroth-order model achieved results nearly as good as the tools using higher-order background models. Hence, the higher-order background model and discriminative prior mainly help to rank down false motifs, which are often repetitive or have a biased nucleotide composition. The sensitivity of Weeder (Pavesi and Pesole, 2006), AMADEUS (Linhart et al., 2008), and PRIORITY (Narlikar et al., 2006) on the top four motifs is lower than that of MEME, as these tools often report different variants of the same motif. Surprisingly, none of the tested tools – including our own – could gain much on this data set by using conservation information. MEME improved from 153 (MEME-$\mathcal{D}$) to 155 (MEME-$\mathcal{DC}$), PRIORITY stayed constant at 156 (PRIORITY-$\mathcal{D}$ and PRIORITY-$\mathcal{DC}$), and ERMIT even decreased from 180 (ERMIT) to 179 (cERMIT). This failure might be due to only weak cross-species conservation of functional binding sites (Borneman et al. (2007), Odom et al. (2007)), but it may also

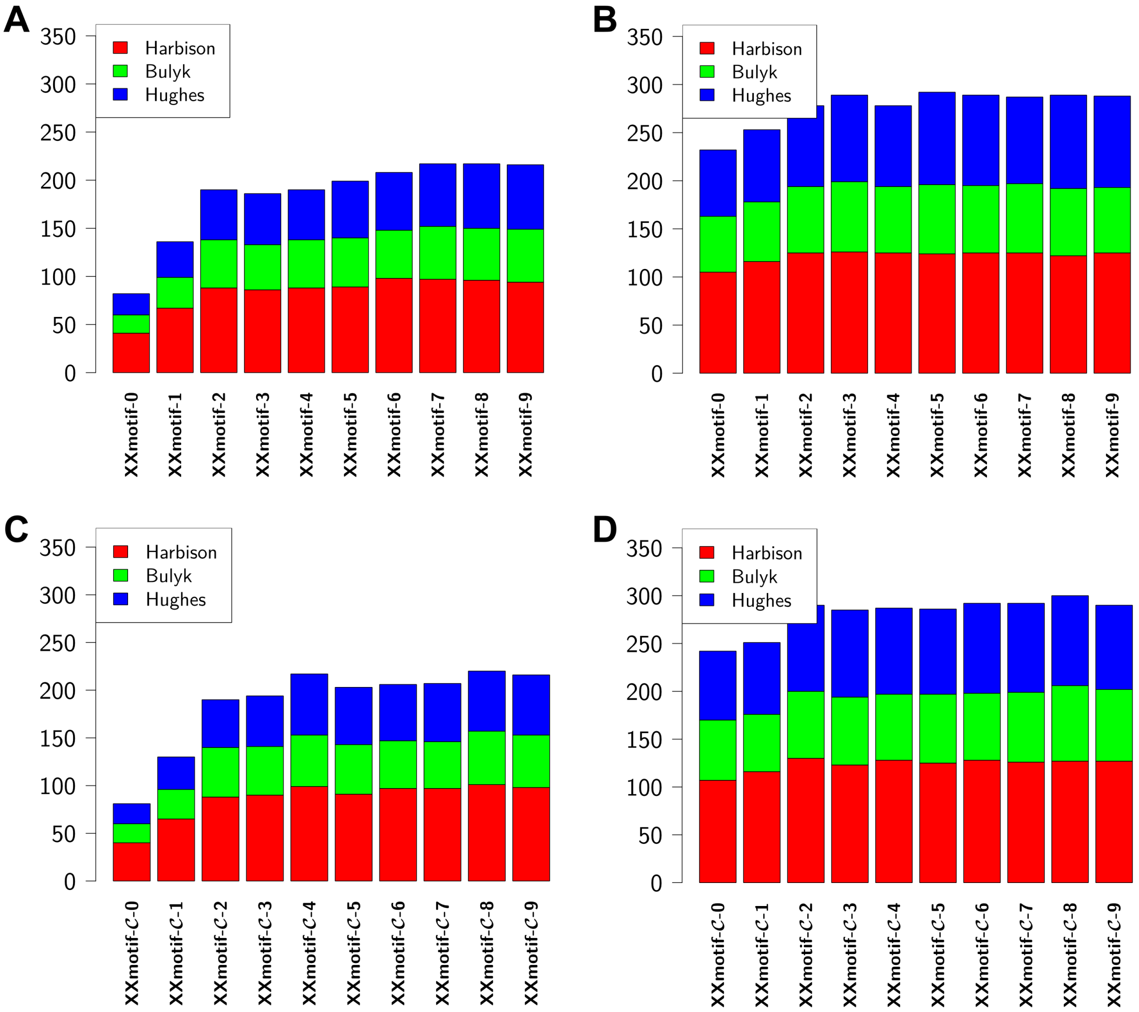


Figure 5.3.: Number of correctly identified motifs of XXmotif on the ChIP-chip data set of Harbison et al. (2004), depending on the order of the background model ranging from zero to nine. Three experimental reference sets are used to judge the correctness of motifs (red, green, blue). (A) Top 1 prediction without conservation, (B) Top 4 prediction without conservation, (C) Top 1 prediction with conservation, (D) Top 4 prediction with conservation

point to limitations of how conservation is evaluated and integrated into the motif search (Mustonen and Lässig (2005), Kim et al. (2009), Shultzaberger et al. (2010)).

To find out how much XXmotif gained by its masking stage, we tested the performance of the other tools on the masked sequence data but observed only minor improvements (see Table A.1).

5.1.3. PWM quality benchmark

To assess the quality of the predicted motifs quantitatively, we could simply evaluate the similarity of the predicted motif PWMs to the reference motifs. However, since some of the reference motifs themselves may be quite inaccurate, we sought a quality measure that is independent of the reference motifs. We were inspired by the reference-free quality assessment presented by Zhu et al. (2009), with which the quality of the PWMs obtained with protein binding arrays was measured.

We analyzed how well the ChIP-enriched regions of Harbison et al. (2004) are predicted using the motif PWMs reported by the tools. We selected the 247 data sets that had at least ten significantly ChIP-enriched regions (P-value < 0.001). The positive and negative sequence sets were generated as in the previous section. We selected the best from each breed of tools, ran these six tools on the 247 sequence sets, and analyzed the PWMs they reported. For this purpose, we ranked all intergenic regions by the best match to the reported PWM. Regions that were significantly ChIP-enriched (P-value < 0.001) were counted as correct predictions, all others as false predictions. A receiver operating characteristic (ROC) curve plots the number of correct predictions over the number of false predictions (Figure 5.4). Usually, only a small fraction of all intergenic regions contain a binding site for a transcription factor. We therefore calculated the partial area under the ROC curve (pAUC) within the best-ranked 5% false predictions. Here, pAUC $= 1$ corresponds to a perfect PWM that scores all significantly ChIP-enriched regions above all other regions. A PWM whose correct predictions are distributed uniformly among the 5% top-scoring regions would achieve a pAUC ≈ 0.5. To avoid rewarding methods that tended to report overly specific motifs, we employed five-fold cross-validation. This technique ensures that PWMs are assessed on a part of the data that was not used to predict these motifs.

Figure 5.5 shows the cumulated distribution of pAUC values, one for each of 247 PWMs

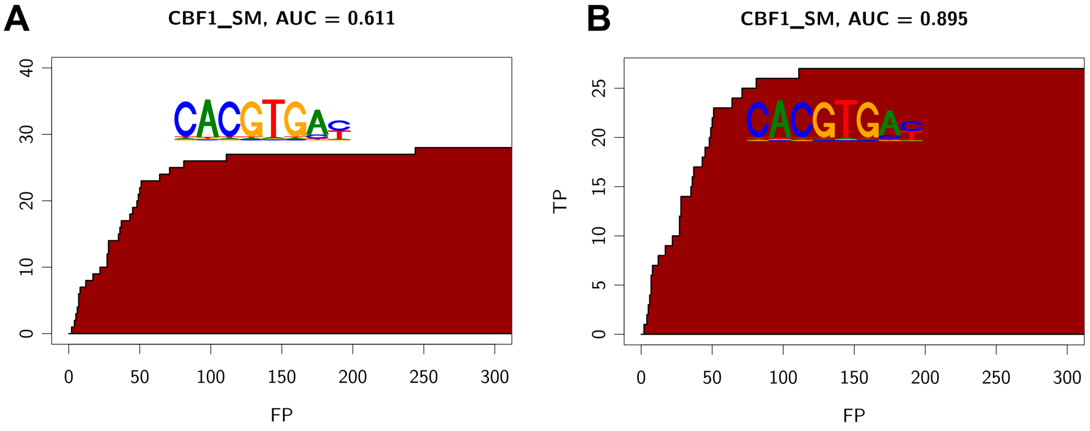


Figure 5.4.: ROC curve of the PWM found by XXmotif in the CBF1_SM ChIP-chip data set and the corresponding partial area under curve (AUC) value calculated from it. (A) All intergenic regions having a ChIP-chip P-value < 0.001 are listed as true positives (TPs) (B) Only those TPs are listed that have a binding site in the region that matches to at least one of the CBF1 PWMs from the "Bulyk", "Hughes", or "Harbison set".

(A, B), and for each of 151 PWMs on a restricted data set (C, D). In A and C, the pAUC values of the top-ranked PWMs are plotted. The average pAUC values are listed in the legend. XXmotif attains an average pAUC value 20% higher than that of MEME-$\mathcal{DC}$, the second best tool, and 40% higher than PRIORITY-$\mathcal{DC}$, the next best. Similar results are obtained on the best out of the four top-ranked PWMs (Figure 5.5B, D).

The biggest differences between top 1 and top 4 predictions are observed for Weeder, scoring 0.071 and 0.172, respectively, though top 1 and top 4 predictions are comparable in the sensitivity benchmark (Figure 5.2). Weeder has the tendency to report short motifs as the top 1 prediction. These PWMs are too unspecific to achieve good pAUC values although they are counted as correct in the sensitivity benchmark. The improvement for the top 4 predictions mainly originates from longer versions of the same motif at lower ranks. In contrast, PRIORITY and AMADEUS have a predefined motif length (eight by default). Since many regulatory elements have more than eight informative positions, their motifs are often less specific than those of tools that optimize the motif length. cERMIT incorporates conservation information into the algorithm by filtering out all non-conserved binding sites. This strategy leads to very specific PWMs that cannot generalize well to weak but functional sites. The result is an average pAUC comparable to MEME-$\mathcal{DC}$, Weeder and PRIORITY-$\mathcal{DC}$, although the algorithm is more sensitive (Figure 5.2). Hence, ERMIT, which does not incorporate conservation information obtains significantly better average pAUC values than are obtained for cERMIT (Figure A.2). XXmotif incorporates conservation information by combining P-values for conservation and motif enrichment (Section 2.2.5). Therefore, conserved and non-conserved sites can contribute to the resulting motif, leading to good motif qualities for both the top 1 and top 4 predictions.

No tool achieved a pAUC value of larger than 0.7 on any of the data sets, although ∼50% of the PWMs are expected to be correct according to Figure 5.2. The low correlation of binding sites predicted using PWMs and *in-vivo* binding sites as measured by ChIP-chip/seq and related techniques is well known, and various causes have been implicated: (1) The predicted PWM might not be specific enough to separate functional, bound sites from nonfunctional, unbound sites. (2) Transcription factors only bind effectively in regions with open, accessible chromatin as measured, for example, by DNase hypersensitivity (Li et al., 2011). (3) Transcription factors compete with nucleosomes for the DNA reducing binding efficiencies around regions of high nucleosome occupancy (Segal and Widom, 2009). (4) Transcription factors with similar sequence specificities can compete for binding sites (Zhou and O'Shea, 2011). (5) The immunoprecipitated transcription factor might bind indirectly to the DNA via other factors.

In particular, we observed that quite often, long CA- and TG-repeats were predicted irrespective of the immunoprecipitated transcription factor. These unspecific motifs are overrepresented in the ChIP-enriched regions of the Harbison data set and therefore obtained

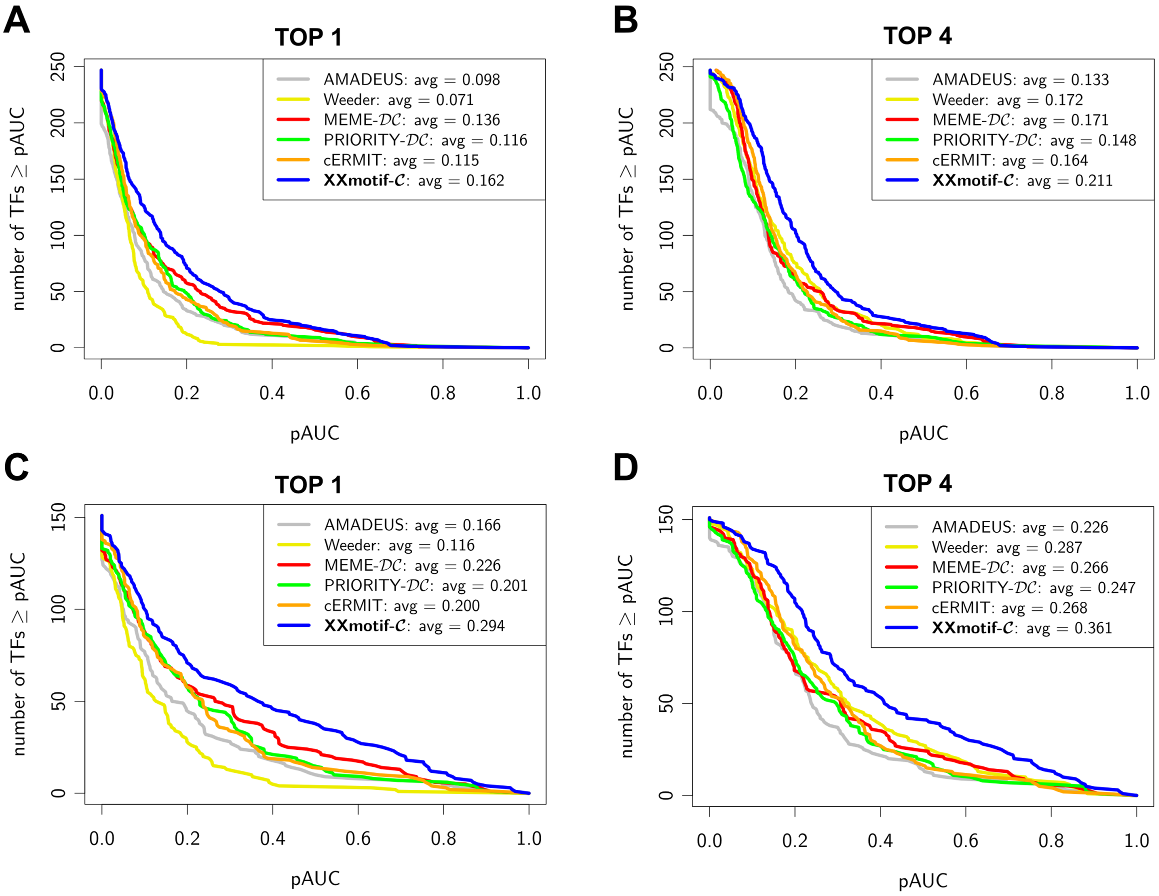


Figure 5.5.: PWM quality assessment on yeast ChIP-chip data from Harbison et al. (2004). The curves quantify how well the scores of the reported PWMs can predict the ChIP enrichment of the sequences. Each PWM is used to rank the intergenic regions by their maximum PWM score. For each predicted PWM, a receiver operator characteristic (ROC) curve with the number of correct predictions over the number of false predictions is computed, and the partial area under the ROC curve (pAUC) deduced from it. The pAUC is the fractional area under the ROC curve within the 5% best-ranked false predictions. For an ideal predictor, pAUC = 1. The average pAUC scores are listed in the figure legends. (A, B) cumulative distribution of the pAUC over all 247 ChIP-chip data sets that had at least ten significantly enriched regions (P-value < 0.001). Regions with ChIP enrichment P-value < 0.001 are defined as correct predictions, all other regions as false predictions. (C, D) As in A, B but using only data sets that have at least five significantly ChIP-enriched regions with matches to the literature motif, and considering only sequences that contain a match to the literature motif.

high pAUC scores (Eden et al., 2007). To reduce these and other potential sources of discrepancies between ChIP-enrichment and binding sites, we restricted the analysis to those 151 data sets which have at least five significantly ChIP-enriched sequences (P-value < 0.001) with matches to one of the reference motifs in the “Harbison set”, the “Bulyk set”, or the “Hughes set”. Here, we defined a match to a reference motif by a log-odds score of at least 70% of the maximum attainable log-odds score for the PWM. We also ignored ChIP-enriched regions without a match to one of the transcription factor’s reference motifs. Figure 5.5 (C, D) shows the resulting pAUC distributions. Around 50% of all top-ranked PWMs reported by XXmotif achieved pAUC values of at least 0.2, compared to 30% in

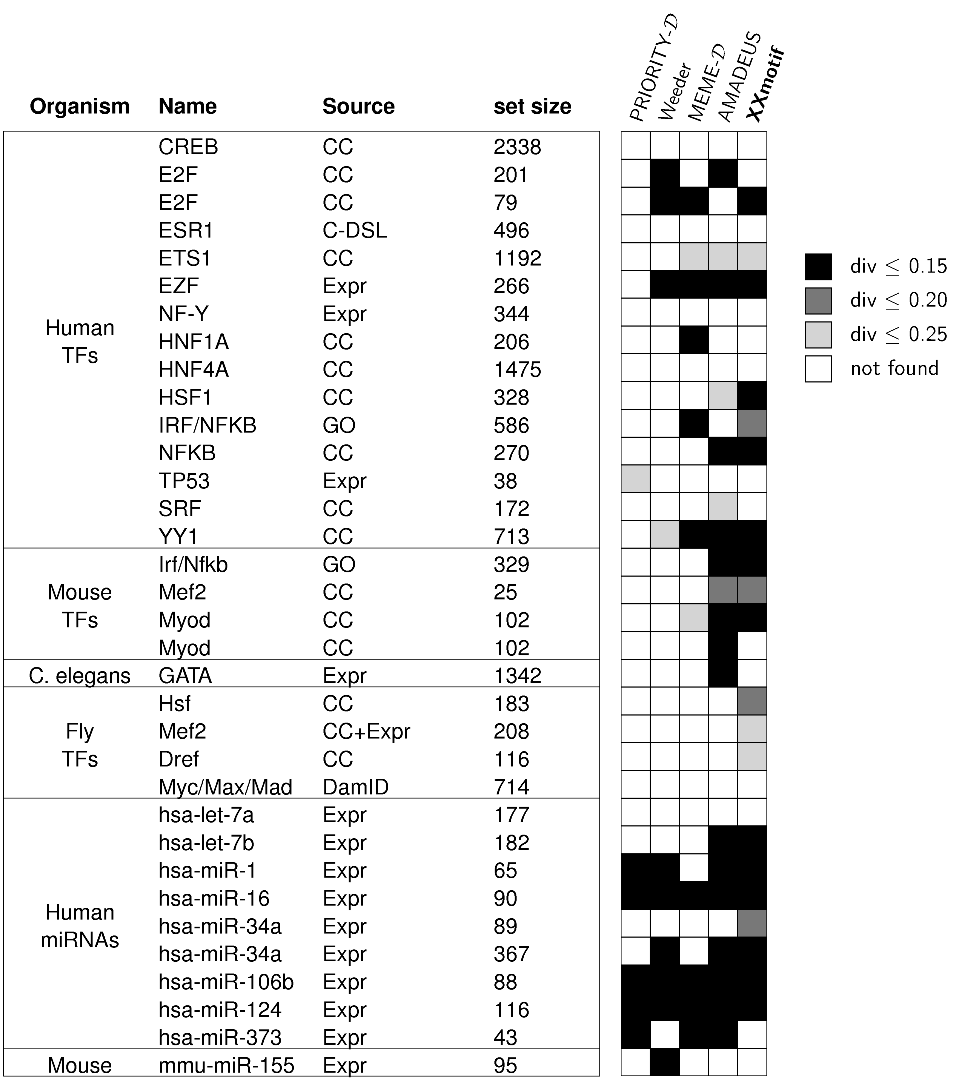
Figure 5.5A. XXmotif improved most in this stricter benchmark setting. In the previous setting (Figure 5.5A, B), the other tools tended to report more low-complexity motifs that achieved good pAUC scores than XXmotif, whose low-complexity filter masks out most dinucleotide repeats.

5.1.4. Metazoan benchmark

The great majority of motif discovery tools has been tested on artificial data sets or on the ChIP-chip data sets of Harbison et al. (2004). Ron Shamir and coworkers therefore assembled a benchmark set ("metazoan target set compendium") with sequences mainly from human and mouse (Linhart et al., 2008): 32 target sets contain enriched transcription factor binding sites from human, mouse, fly (*Drosophila melanogaster*), and worm (*Caenorhabditis elegans*), which are based on ChIP-chip experiments, co-expressed genes, and other data sources. Ten target sets from human and mouse contain genes that are co-regulated under microRNA (miRNA) knock-downs.

The 8-mer miRNA seeds were imported from miRBase 16.0 (Griffiths-Jones et al., 2006). While Linhart et al. (2008) used experimentally validated transcription factor PWMs from release 8.0 of the TRANSFAC database (Wingender et al., 1996), we could only access the latest public release (7.0) and therefore had to remove eight transcription factors from the analysis. We used the benchmark set up as described in Linhart et al. (2008) to evaluate the sensitivity of XXmotif and the best versions of the previously tested tools that do not need multiple sequence alignments to score sequence conservation. ERMIT could not be evaluated on this benchmark since for many target sets no P-values existed. We used the same metric as before to calculate the distance of a predicted motif from a literature motif. If multiple validated motifs were listed in TRANSFAC or miRBase, we took the motif that had the lowest distance to the predicted motif. The motif analysis was performed as described by Linhart et al. (2008).

Figure 5.6 displays the results of the top four predictions in the same way as in Linhart et al. (2008). On the transcription factor target sets, PRIORITY-$\mathcal{D}$ finds only two correct motifs, whereas Weeder, MEME-$\mathcal{D}$, AMADEUS, and XXmotif find 6, 14, 17, and 22, respectively (divergence ≤ 0.25). When counting only highly similar motif predictions (divergence ≤ 0.15), PRIORITY-$\mathcal{D}$ achieves 0, Weeder 3, MEME-$\mathcal{D}$ 8, AMADEUS 10, and XXmotif 15 correct predictions. On the miRNA target sets, PRIORITY-$\mathcal{D}$ and AMADEUS, whose fixed motif length of eight coincides with the length of the miRNA seeds, are able to detect 6 and 9 miRNA seeds, respectively. Weeder and MEME-$\mathcal{D}$ find 6 and 5, respectively, whereas XXmotif finds 8 correct miRNA seeds. The results for the top one predictions show the same trend (Figure A.3).



Organism	Name	Source	set size	PRIORITY-$\mathcal{D}$	Weeder	MEME-$\mathcal{D}$	AMADEUS	**XXmotif**
Human TFs	CREB	CC	2338	□	□	□	□	□
	E2F	CC	201	□	■	□	■	□
	E2F	CC	79	□	■	■	□	■
	ESR1	C-DSL	496	□	□	□	□	□
	ETS1	CC	1192	□	□	░	░	░
	EZF	Expr	266	□	■	■	■	■
	NF-Y	Expr	344	□	□	□	□	□
	HNF1A	CC	206	□	□	■	□	□
	HNF4A	CC	1475	□	□	□	□	□
	HSF1	CC	328	□	□	□	░	■
	IRF/NFKB	GO	586	□	□	■	□	▩
	NFKB	CC	270	□	□	□	■	■
	TP53	Expr	38	░	□	□	□	□
	SRF	CC	172	□	□	□	░	□
	YY1	CC	713	□	░	■	■	■
Mouse TFs	Irf/Nfkb	GO	329	□	□	□	■	■
	Mef2	CC	25	□	□	□	▩	▩
	Myod	CC	102	□	□	░	■	■
	Myod	CC	102	□	□	□	■	□
C. elegans	GATA	Expr	1342	□	□	□	■	□
Fly TFs	Hsf	CC	183	□	□	□	□	▩
	Mef2	CC+Expr	208	□	□	□	□	░
	Dref	CC	116	□	□	□	□	░
	Myc/Max/Mad	DamID	714	□	□	□	□	□
Human miRNAs	hsa-let-7a	Expr	177	□	□	□	□	□
	hsa-let-7b	Expr	182	□	□	□	■	■
	hsa-miR-1	Expr	65	■	■	□	■	■
	hsa-miR-16	Expr	90	■	■	■	■	■
	hsa-miR-34a	Expr	89	□	□	□	□	▩
	hsa-miR-34a	Expr	367	□	■	□	■	■
	hsa-miR-106b	Expr	88	■	■	■	■	■
	hsa-miR-124	Expr	116	■	■	■	■	■
	hsa-miR-373	Expr	43	■	□	■	■	□
Mouse	mmu-miR-155	Expr	95	□	■	□	□	□

■ div ≤ 0.15
▩ div ≤ 0.20
░ div ≤ 0.25
□ not found

Figure 5.6.: Top 4 benchmark results on 24 target sets for transcription factors from human, mouse, worm and fly, as well as 10 target sets for microRNAs from human, and mouse from the metazoan target set compendium (Linhart et al. 2008). The plot is adapted from Linhart et al. (2008): The "Source" column indicates the experimental procedure or database from which the target set was derived: Gene expression microarrays (Expr), ChIP-chip (CC), ChIP-DSL (C-DSL), DamID (van Steensel et al., 2001), or Gene Ontology (GO) database (Ashburner et al., 2000). Black and gray boxes indicate the similarity of the predicted PWM to the reference motif in TRANSFAC or miRBase. Darker shades indicate closer similarity. "Set Size": number of sequences within the input set.

We compared the run times of the five tools on the metazoan target set compendium for a single core Xeon 2.9 GHz CPU (Figure 5.7). AMADEUS is the fastest tool with an average run time per target set of 1m57s. XXmotif comes in second with an average run time of 4m27s, whereas PRIORITY-$\mathcal{D}$ needs on average 13m23s. Neither AMADEUS nor PRIORITY-$\mathcal{D}$ optimizes the motif length, which is the most time consuming step within XXmotif. Weeder and MEME do optimize the motif length and are on average 19 and 700 times slower than XXmotif, respectively.

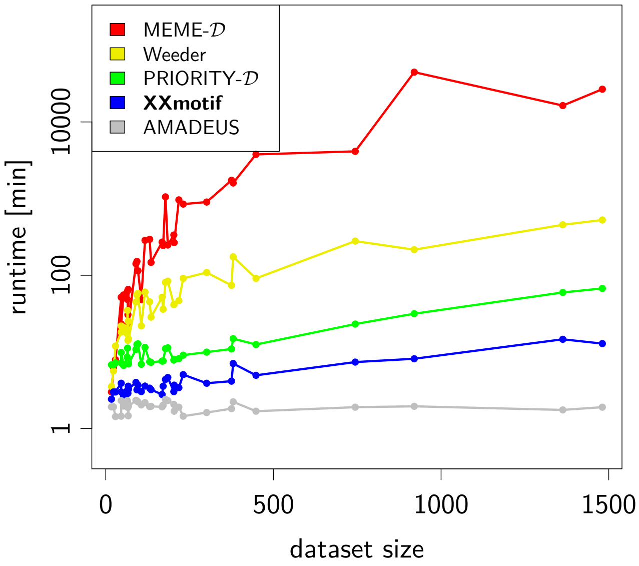


Figure 5.7.: Rutime of tested tools on the metazoan benchmark. All jobs were run on a single core Xeon 2.9 GHz CPU.

5.2. Regulatory Motifs for Early Embryo Segmentation in Flies

One of the most studied model systems of transcription regulatory networks is the network that lays down the segmentation pattern along the anterior-posterior axis in the early *Drosophila* embryo (Jaeger et al. (2004), Zinzen et al. (2009), Li et al. (2011), Perry et al. (2011)). Various transcription factors are known to participate in this network, but also other, as yet unidentified factors are believed to be involved (Segal et al. (2008), He et al. (2010)). The identification of these “missing nodes” in the network would set the stage for more accurate, quantitative models for this network paradigm.

We obtained 54 hand-curated *cis*-regulatory modules comprising sequences that are primarily targeted by maternal and gap genes, and exclude some of the pair rule elements that are primarily targeted by pair rule genes (Schroeder et al. (2004), Schroeder et al. (2011)). The list of all 54 segmentation modules including the reference of each individual element was provided by Mark Schroeder (Table A.5). Alignments for these sequences were generated using the UCSC 14-way multiple sequence alignments (dm3). Since we expect functional binding sites to be more conserved than background, we used the 13 most related species of the UCSC 15-way multiple sequence alignments (Blanchette et al., 2004) consisting of the *Drosophila* group and *Anopheles gambiae* as outgroup. For simplicity, we did not supply a negative sequence set. In this case, XXmotif automatically constructs a second-order background model from the sequences in the positive set.

Table 5.1 lists all motifs reported in a single XXmotif run up to an E-value of 1. First, we note that XXmotif's E-values are quite conservative due to the Bonferroni correction and that most motifs with an E-value below 1 correspond to real binding motifs. Of the 19 predicted motifs, 11 were clearly similar to motifs in the Fly Factor Survey (Noyes et al., 2008) that all correspond to factors known to organize segmentation in the early embryo. Impressively, the list of motifs includes representatives of most classes of the transcription

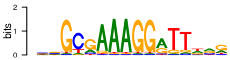
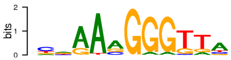
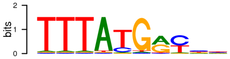
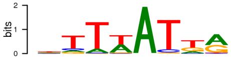
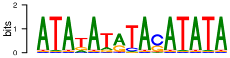
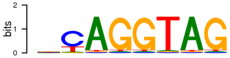
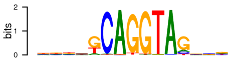
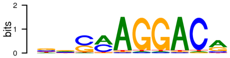
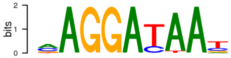
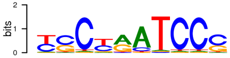
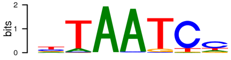
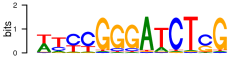
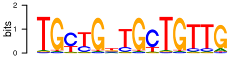
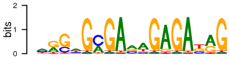
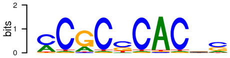
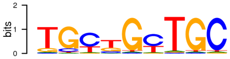
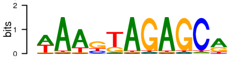
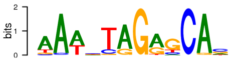
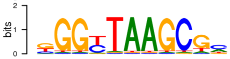
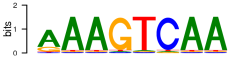
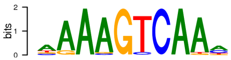
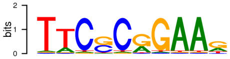
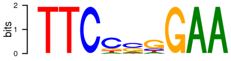
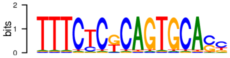
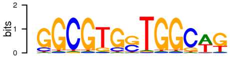
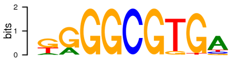
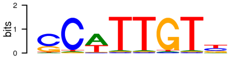
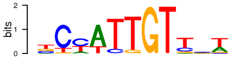
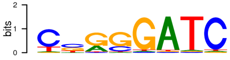
Motif	E-value	Motif PWM	Literature Motif	Name
1	9.16×10^{-25}			Kr
2	1.99×10^{-16}			Cad
3	1.34×10^{-9}		–	–
4	2.69×10^{-9}			Vfl
5	1.81×10^{-7}			Ttk
6	6.56×10^{-7}			Bcd
7	1.15×10^{-6}		–	–
8	3.57×10^{-6}		–	–
9	5.37×10^{-6}		–	–
10	1.88×10^{-5}		–	–
11	2.51×10^{-3}		–	–
12	1.12×10^{-2}			Kni
13	1.50×10^{-2}		–	–
14	3.81×10^{-2}			Tll
15	5.80×10^{-2}			D-Stat
16	1.89×10^{-1}		–	–
17	4.59×10^{-1}			Hkb
18	6.33×10^{-1}			D
19	6.37×10^{-1}		–	–

Table 5.1.: Motifs discovered in *cis*-regulatory modules for fly segmentation. The table lists all motifs that XXmotif reports up to an E-value of 1 in a single run on 54 segmentation modules responsible for patterning the anterior-posterior (AP) axis during early embryogenesis. To score conservation, multiple sequence alignments of *D. melanogaster*, 11 other *Drosophila* species, and *Anopheles gambiae* were supplied as input. For 11 of the 19 predicted motifs, similar literature motifs of transcription factors known to be involved in AP axis segmentation have been found. The literature motifs and names of the transcription factors are shown in the rightmost columns. Nine motifs are predicted that may describe binding affinities of missing nodes in the transcriptional network.

factors that are known to be involved in the segmentation. Factors missing are Forkhead that is underrepresented in the considered sequences and Hunchback, for which an unusual motif with consensus “TTTTTT” was reported in the literature (Gallo et al., 2011). As Hunchback has many binding sites in the segmentation modules, we surmise that in our second-order background model a “TT” is followed with high probability by another T, increasing the *P*-value for matches to a “TTTTTT” motif beyond significance.

Nine of the predicted motifs cannot be matched to known factors. Their *E*-values are of comparable significance as the known motifs. We therefore speculate that many of these novel motifs belong to transcription factors that represent “missing nodes” in the segmentation network. It will be exciting to determine experimentally what factors bind to these motifs, for example using one-hybrid screens (Deplancke et al. (2006), Hens et al. (2011)) or mass spectrometry techniques (Mittler et al., 2009).

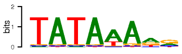
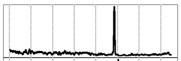
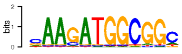
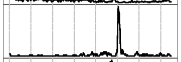
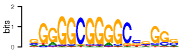
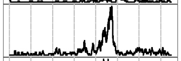
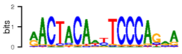
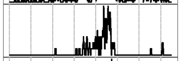
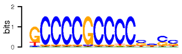
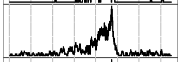
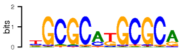
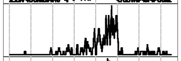
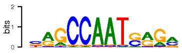
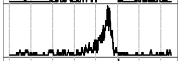
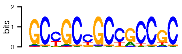
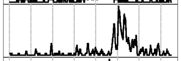
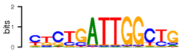
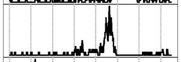
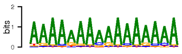
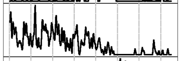
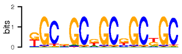
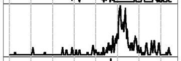
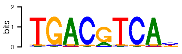
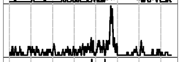
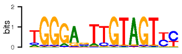
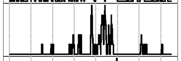
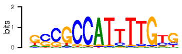
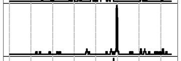
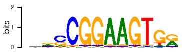
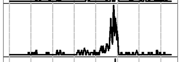
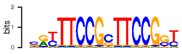
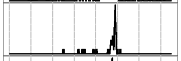
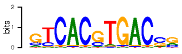
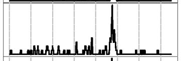
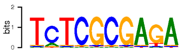
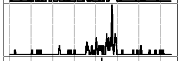
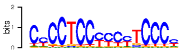
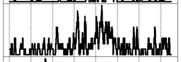
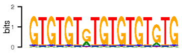
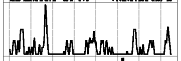
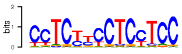
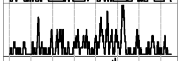
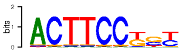
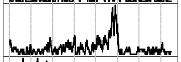
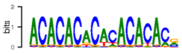
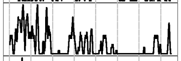
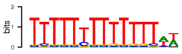
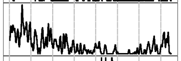
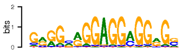
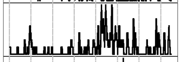
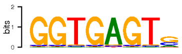
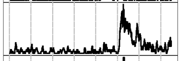
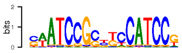
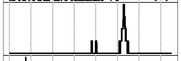
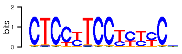
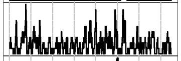
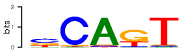
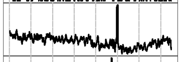
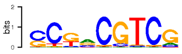
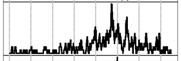
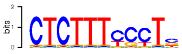
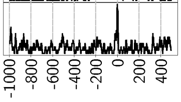
5.3. Human Core Promoter Motifs

Core promoters are the regions around transcription start sites (TSS) to which the general transcription machinery consisting of RNA polymerase and general transcription factors bind. In recent years it has become clear that the motif architecture of core promoters can influence the regulatory behavior of the promoter (Juven-Gershon and Kadonaga, 2010). Around 15 motifs have been discovered in fly and human core promoters that are enriched around human TSSs (Gershenzon and Ioshikhes (2005), FitzGerald et al. (2006), Gershenzon et al. (2006), Xi et al. (2007)), the most frequently occurring ones being the TATA-box (10% occurrence) and the SP1 motif (11%). Most of the elements are rare and not generally conserved within *Animalia*. For example the human Initiator motif reported by Xi et al. (2007) occurs in only about one percent of all core promoters and bears little resemblance to the Initiator found in *D. melanogaster* (consensus TCAGT).

We extracted 1871 core promoter regions from -300 bp to +100 bp around human transcription start sites from the eukaryotic genome database (Schmid et al., 2006) and ran XXmotif using the “zero or one occurrence per sequence” option. As we expect core promoter elements to have a defined distance to the TSS, we used the “localization” option of XXmotif (Section 2.2.7), in which *P*-values for positioning of motif occurrences are combined with the enrichment *P*-values. No negative sequence set was given, therefore XXmotif learned a second-order background model from the positive sequence set.

Table 5.2A shows all enriched PWMs with an *E*-value up to one. 18 of the 31 motifs

A

Motif	Logo	Distribution	occ [%]	E-value	Name
1			9.67	1×10^{-132}	TATA
2			4.92	1×10^{-124}	YY1
3			11.60	3×10^{-123}	Sp1 (rev)
4			2.83	4×10^{-112}	SREBP-1
5			9.51	1×10^{-111}	Sp1
6			6.04	3×10^{-83}	NRF1
7			6.73	1×10^{-62}	CAAT
8			3.69	2×10^{-61}	motif8 (rev)
9			4.76	5×10^{-61}	CAAT (rev)
10			2.14	1×10^{-58}	–
11			4.01	5×10^{-57}	motif8
12			4.22	2×10^{-48}	CREB
13			1.12	1×10^{-32}	SREBP-1 (rev)
14			1.18	1×10^{-30}	YY1 (rev)
15			4.81	2×10^{-28}	NRF2
16			1.12	1×10^{-27}	XX1
17			1.44	4×10^{-22}	USF
18			1.82	1×10^{-19}	CLUS1
19			3.05	9×10^{-19}	–
20			0.43	1×10^{-17}	–
21			2.24	2×10^{-17}	–
22			7.16	2×10^{-17}	NRF2 (rev)
23			0.53	3×10^{-17}	–
24			0.86	4×10^{-17}	–
25			2.57	2×10^{-15}	–
26			6.63	2×10^{-14}	XX2
27			0.64	5×10^{-13}	XX3
28			2.19	6×10^{-9}	–
29			6.52	1×10^{-7}	XX4 (Inr)
30			6.79	6×10^{-3}	–
31			0.53	4×10^{-1}	TCT

B

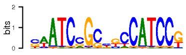
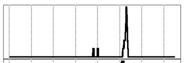
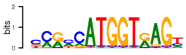
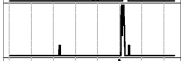
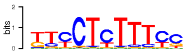
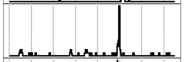
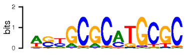
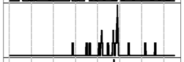
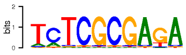
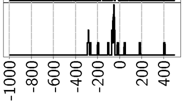
Motif	Logo	Distribution	occ [%]	E-value	Name
1			20.00	2×10^{-27}	XX3
2			20.00	4×10^{-20}	XX5
3			26.15	9×10^{-20}	TCT
4			21.54	2×10^{-6}	NRF1
5			15.38	5×10^{-6}	CLUS1

Table 5.2.: Human core promoter motifs discovered by XXmotif. (A) Motifs reported up to E-value of 1 by a single run of XXmotif on 1871 human core promoter regions (-300 bp to +100 bp around TSS) from the eukaryotic promoter database (EPD, Schmid et al. (2006)). For 18 of the 31 predicted motifs we found similar motifs in the literature. Their names are given in the last column. The motif at position 14, which was originally named Initiator (Xi et al., 2007), is actually the reverse complement of YY1 and is therefore referred to as YY1 (rev) here. Four novel, highly significant motifs, designated XX1 to XX4, show positional distribution peaks near the TSS. XX4 is the canonical Initiator motif similar to elements found in *D. melanogaster* and *S. cerevisiae*. Nine motifs have a broad positional distribution and are not named. The positional distributions of the PWMs were obtained by scanning the PWMs over a larger region (-1000 bp to +500 bp) around the TSS. (B) Top five motifs obtained with the core promoter sequences of the 65 genes annotated as coding for ribosomal proteins in EPD. These motifs occur relatively frequently in ribosomal genes and are likely to be characteristic for constitutively and highly expressed human genes.

are similar to previously described motifs, whose names are given in the last column. These motifs are indeed enriched within the core promoter region, as shown by their positional distribution in a region from -1000 bp to +500 bp around the TSS.

Nine motifs are mostly repetitive, of low compositional complexity, and rather uniformly distributed. We believe that they do not represent functional promoter motifs. Possibly these low complexity regions serve to modulate the physical properties of the DNA double helix near the core promoter, for example in order to attract or repel nucleosomes.

XXmotif further detected five sharply peaked motifs with E-values comparable to those of previously described motifs (XX1 to XX5). XX1 is similar to a tandemic version of NRF2 (rev) missing the first adenine. However, as XX1 has a sharper positional distribution and a higher E-value than NRF2 (rev), we consider it to be a distinct core promoter element. YY1 (rev) was called Initiator element in Xi et al. (2007) due to its precise localization at the TSS. But in contrast to the very specific YY1 (rev) motif, which occurs in only 1.2% of EPD core promoter sequences, motif XX4 occurs in 6.5%, is equally well positioned at the TSS, and is similar to known Initiator elements in flies (Ohler, 2006) and yeast (Zhang and Dietrich, 2005). We therefore suggest that XX4 is the as yet undiscovered canonical human Initiator motif.

XXmotif also finds a second Initiator element (called the “TCT element”) that was discovered in fly promoters of ribosomal protein genes and was shown to also be enriched around human TSSs (Parry et al., 2010). To further analyze the ribosomal system of transcription initiation, we searched for motifs in the subset of 65 promoter regions of genes annotated to code for ribosomal proteins (Table 5.2B). We found four motifs that we had already seen on the large set of core promoters, and their PWMs look almost identical. All four are strongly enriched at ribosomal protein core promoters. As expected, the TCT motif is among them, occurring at 26% of ribosomal protein genes, compared to 0.5% of all genes (50-fold enrichment). Motif XX3 is present in 20% of these promoters, in comparison to 0.6% over all EPD promoters. NRF1 and CLUS1 are enriched 3.5 fold and eight-fold, respectively. XX5 is a novel motif that bears some similarity to XX2.

In summary, in addition to finding almost all motifs known to be enriched in human core promoters, we discovered five new motifs that are strongly peaked around TSSs. It will be exciting to understand the function of these motifs, find the corresponding specific or general transcription factors and to investigate the association of these motifs with regulatory properties of core promoters, such as stress inducibility, degree of tissue- and time-dependent regulation, maximum and basal transcription rates.

5.4. Discussion

We aimed to demonstrate XXmotif's usefulness in several complementary, biological benchmarks and applications. On a large data set of ChIP-chip measurements, we compared the sensitivity and the quality of predicted motifs of various state-of-the-art motif discovery tools. On a smaller data set, the metazoan target set compendium, we could show that XXmotif's sensitivity for detecting the correct motifs was transferable to metazoan and mammalian sequences, and to diverse scenarios for measuring and selecting motif-enriched sequences. We then applied XXmotif to 54 segmentation modules in flies and discovered most of the binding motifs of known segmentation factors.

Finally, we analyzed human core promoter sequences with XXmotif. We found almost all previously described motifs as well as five novel motifs that have sharply peaked positional distributions around the TSS. One of the novel motifs is localized to within ±10 base pairs of the TSS and is similar to the Initiator motif in fly, and yeast, which identifies it as the canonical human Initiator motif. We did not find the BRE, DPE, and MTE elements. However, these were never found by a *de-novo* search on human core promoter sequences. The BRE element was deduced from crystal structures of TFIIB and TBP bound to the DNA (Nikolov et al., 1995) and later shown to be weakly positioned, but enriched around TSSs of several species (Gershenzon et al. (2006), Sandelin et al. (2007)). The MTE and DPE elements were discovered in *D. melanogaster* (Ohler et al., 2002) and, by scanning their PWMs over human core promoter sequences, the DPE element was then found to be slightly enriched around human TSSs (FitzGerald et al., 2006). However, their positioning and signal over background is much weaker than about we observe for the motifs reported by XXmotif. Five motifs, two of them discovered in this study, are found to be strongly enriched in human core promoters of ribosomal protein genes. It is an intriguing possibility to try to combine these motifs into a “super core promoter” that would support extremely high levels of transcription for applications in basic research and biotechnology (Juven-Gershon et al., 2006).

Several design aspects contribute to XXmotif’s performance. First, its pattern-based stage (Figure 2.3, red) is very sensitive and efficient in finding good patterns to be improved in the PWM stage (green). XXmotif employs palindromic and tandemic seeds with gaps of up to 11 positions. Its parallel strategy to extend the best five patterns instead of only the best one and the possibility to extend patterns across gaps allows it to find patterns that do not contain even a single, significant 5-mer seed.

Second, an eighth-order background model gives clear improvements over lower-order models (Figure 5.3). The use of an interpolated Markov model makes it possible to train high orders with limited data (Salzberg et al., 1998). To understand why higher orders help, note that some sequences with relatively low complexity, such as (imperfect) trinucleotide repeats, can be strongly overrepresented in the entire genome and will look enriched in comparison to a first-order background model in any subset of genomic sequences. Some tools, such as AMADEUS, do not train a statistical background model but instead use the negative set directly to determine the P-values of patterns. Therefore, no patterns of any length that are enriched uniformly in the entire genome can become significant. However, this approach has the disadvantage of limiting the significance of P-values that can be calculated. If a pattern does not have a single match in the negative set, it is not possible to decide if it can be improved by extending it. The pattern length is therefore limited to around 8 positions in practice.

Third and most importantly, XXmotif can optimize motif PWMs by statistical significance of enrichment as measured by P-values. It thus combines the solid statistical estimates of pattern-based algorithms with the more powerful representation of motifs by PWMs. In contrast to patterns, PWMs can describe weak and strong binding. In a thermodynamic treatment of factor binding, PWMs emerge naturally, representing the independent energetic contributions of the binding site nucleotides to the binding energy. A challenge in calculating P-values for PWMs is that, unlike for patterns, we cannot simply count matches. Instead, we have to compute P-values for each and every position in the input sequences and somehow combine these P-values for motif occurrences to yield a motif enrichment P-value. We have solved this by applying order statistics and we effectively optimize the score threshold above which potential sites are counted as matches. This procedure can be interpreted from a thermodynamic viewpoint. It is equivalent to the zero-temperature approximation of factor binding, in which sites are either not bound or fully occupied (Homsi et al., 2009). The optimization of K (and hence of the P-value threshold) corresponds to finding the factor concentration at which the total occupancy on the positive sequences differs most significantly from that on the background sequences. But why should it be better to optimize P-values instead of likelihoods in the first place? In lieu of an answer, we note that all popular PWM-based tools in the end rank their motifs by P-value, not by the likelihood.

Fourth, because XXmotif quantifies information from all sources (enrichment, conservation, positioning) by P-values, it is straightforward to combine these without having to resort to loss-prone heuristics. As an additional advantage, other sources of independent information such as chromatin accessibility scores could be easily added in this framework.

In conclusion, XXmotif is a general-purpose method for the discovery of enriched motifs in nucleotide sequences that is based on optimizing the P-values of motif PWMs. In several benchmarks on yeast and metazoan sequences, XXmotif compares favorably with some of the best state-of-the-art motif discovery tools. We hope that in this era of functional genomics and high-throughput, data-driven biology, XXmotif will contribute towards understanding the regulation of our genomes by sequence-specific binding of protein and ncRNA factors.

6. Analysis of the Motif Architecture within Fly Core Promoters

6.1. Introduction to Core Promoters in *D. melanogaster*

Although there has been substantial progress in the structural understanding of eukaryotic transcription (e. g., Kostrewa et al. (2009)), the regulation of this complex process of RNA synthesis from a DNA template still remains poorly understood (Juven-Gershon and Kadonaga (2010), Ohler and Wassarman (2010)). A key step in modulating transcriptional activity is the initiation, thus the assembly of the pre-initiation complex (PIC) consisting of the basal transcription factors TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH at the core promoter (Thomas and Chiang, 2006). Here, within a region of around fifty base pairs upstream and downstream of the transcription start site (TSS), the DNA is enriched with defined sequence elements serving as assembly platforms for the basal transcription factors. Mutations within this so called core promoter elements (CPEs) are able to affect the efficiency of PIC assembly, displace the native TSS, or alter the expression strength (Grosschedl and Birnstiel, 1980).

Several CPEs have been identified using biochemical or computational approaches, e. g., the TATA-box, BRE, Initiator, DPE, MTE, or DRE (Ohler et al. (2002), Smale and Kadonaga (2003)). However, many genes contain only one or no known CPE, prompting the question how the transcription machinery finds these core promoters. Yet unknown motifs or the incorporation of physical properties of the DNA within the core promoter region (Abeel et al., 2008) might be an explanation.

High quality TSS data provide a detailed genome-wide map of single transcription initiation events. In *D. melanogaster*, strong correlations between several CPEs and different initiation patterns could be identified (Hoskins et al., 2011), demonstrating the possibility to enrich CPEs by selecting gene sets with special properties. One approach to perform such a grouping based on TSS data incorporates the width of the region in which initiation occurs and the proportion of TSS counts within a window of 2 nucleotides around the mode of the distribution. Following the classification scheme based on this information, one of three different classes is selected: narrow peak (NP), broad peak (BP), or weak peak (WP) promoter (Ni et al., 2010). A second approach is based on a continuous measure defining the shape of the distribution, the shape index (SI). Promoters with an SI score > -1 are defined as peaked, whereas promoters with an SI score ≤ -1 as broad (Hoskins et al., 2011). However, as the SI score is strongly influenced by the number of counts, highly expressed genes tend to be classified as broad promoters.

Per nucleotide resolution of TSSs is also important for *de novo* identification of CPEs. Already known elements like TATA-box, DPE, or MTE are located almost exclusively at defined positions with respect to the TSS. A previous analysis utilized these distance restrains by specifically searching for 8-mer/6-mer patterns that are non-randomly enriched within bins of 20 bp (FitzGerald et al., 2006). However, a rigid bin size and pattern-based method is prone to be too specific to detect degenerate core promoter elements. The more flexible PWM model was used by Ohler et al. (2002) to analyze the core promoter of *D. melanogaster*. However, as the used motif discovery tool MEME (Bailey and Elkan, 1994) cannot incorporate localization information, they had to focus the analysis on a very small region surrounding the TSS.

In vitro and *in vivo* mutation experiments are frequently used to validate newly identified motifs (e. g., Lim et al. (2004), Parry et al. (2010), Seizl et al. (2011)). In a small subset of genes, these experiments prove the importance of a binding site for transcription initiation. However, mutation experiments are time consuming and do not allow to compare the quality of nearly similar PWMs. In contrast, *in silico* approaches allow for a model comparison on a genome-wide scale. Properties like strand specificity, localization with respect to the TSS, overrepresentation within the core promoter region, as well as conservation to related species are useful features to analyze whether a newly found motif might be functional.

Since general transcription factors such as TBP, or TRF2 act in defined complexes with multiple DNA binding domains (Veenstra and Wolffe (2001), Hochheimer and Tjian (2003)), subsets of CPEs are more frequently found in the same promoter than in different ones. Peaked promoters, for instance, have higher frequencies of location-specific CPEs (e. g., TATA-box, Initiator, DPE, MTE), whereas broad promoters contain the complementary set of known elements (Rach et al., 2009). Stalled genes also contain specific elements (GAGA, Initiator, Pause Button) (Hendrix et al., 2008) indicating that special motif combinations trigger defined gene properties.

In this chapter we provide an improved picture of the motif architecture of eukaryotic core promoters. Based on experimentally derived features, i. e., expression strength, difference of expression within developmental stages, stalling, as well as peakedness of the transcription initiation cluster, we defined 19 gene sets that allow us to analyze correlations to CPEs. To assure high quality sets, we derived an expression independent score for the peakedness of transcription initiation patterns (MAD score), and separated expression classes by analyzing their distribution. In order to reveal novel motifs within these sets, we used XXmotif that combines a P-value that evaluates whether the motif sites are located non-randomly with respect to the TSS with motif overrepresentation and conservation P-values (see Chapter 2). Hence, our *de novo* motif analysis can be performed in a single run on large regions of the core promoter without loosing the descriptive power of a PWM. All analyses in this chapter were done in close collaboration with two Master students I was supervising, Anja Kiesel and Mark Heron.

6.2. Drosophila Core Promoters Fall Into Four Classes

To obtain gene sets likely to consist of different combinations of CPEs, we examined expression data from different developmental stages (Graveley et al., 2011), stalling data (Zeitlinger et al., 2007), and CAGE data measuring single transcription initiation events (Ni et al. (2010), Hoskins et al. (2011)).

6.2.1. TSS cluster width

The CAGE data was used to separate promoters with a defined position of transcription start sites (TSSs) from promoters utilizing several TSSs distributed over a broad genomic region. In a first step, we defined clusters of TSSs belonging to one promoter. Therefore, we smoothed all CAGE tags with a rectangular kernel, and chose continuous regions above the genomic background as clusters if they were close enough to an annotated gene (Section 6.4.1). To quantify the peakedness of a cluster, we utilized a score calculating the mean absolute deviation from the median (MAD score):

$$\mathrm{MAD} = \frac{1}{n}\sum_{i=1}^{n} |x_i - m(X)| \tag{6.1}$$

where n is the number of tags within the cluster, x_i represents the position of the ith tag and $m(X)$ is the median tag position within the cluster. In contrast to the SI score (Hoskins et al., 2011) that has a clear bias towards lower scores if the TSS cluster has many tags, the MAD score is independent of the cluster size (Figure 6.1).

The distribution of cluster widths over all genes is depicted in Figure 6.2A. The local minima at cluster width five was used as threshold to classify core promoters as either a narrow peak (NP) promoter or a broad peak (BP) promoter. As the distribution can be easily fitted by two Gaussians, we do not find any evidence for the existence of a third class of promoters as defined by Ni et al. (2010). For the distribution of SI scores, no clear minimum is observable (Figure B.1), which indicates that the MAD score is a better measure to define the width of TSS clusters. For each cluster the TSS was defined as the mode of the tag distribution.

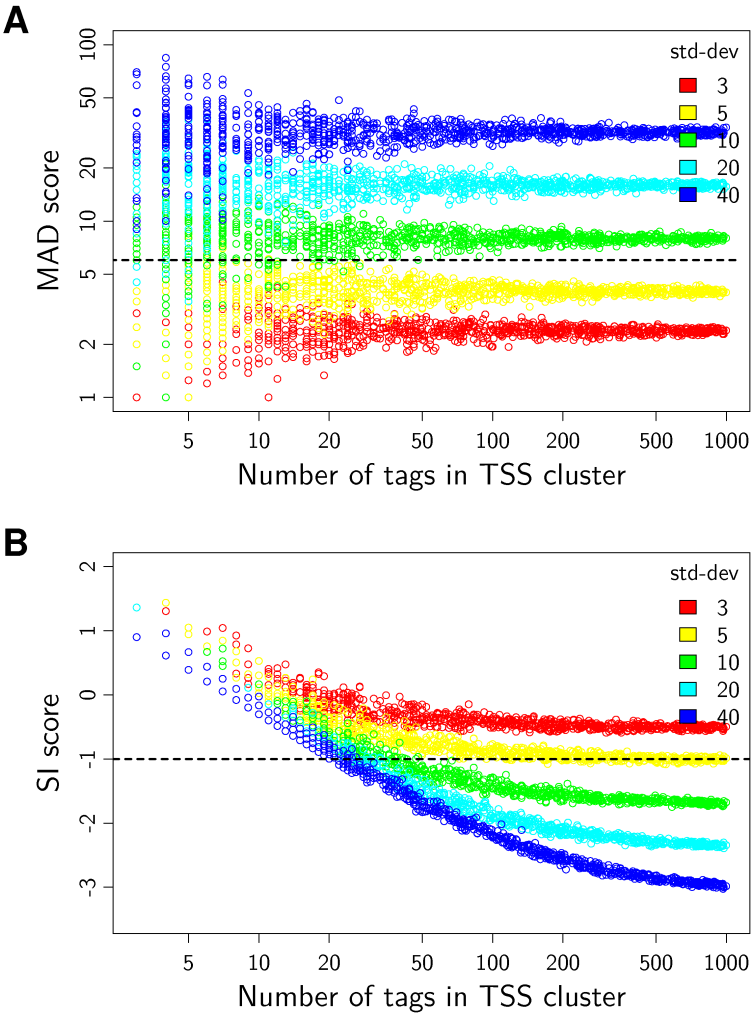


Figure 6.1.: Comparison of the MAD score (A) and SI score (B) metric to estimate the peakedness of a TSS cluster. Each score is calculated on sampled tags from a normal distribution given the standard deviation (std-dev).

6.2.2. Gene sets

Figure 6.2 summarizes the distribution of all promoters depending on the TSS cluster width (A), as well as on different gene expression properties in 30 different developmental stages (B–F). Gene sets were selected depending on the inducibility of gene expression (MAD expression, B), minimum gene expression (C), maximum gene expression (D), gene expression in embryo, larva, or female (E), and gene expression in adult (F). The dashed lines indicate the used thresholds to define each set. If possible, thresholds were chosen at the minima of the distribution (A, C). Otherwise, we used the highest and lowest 10% quantiles to derive sets of genes with special behaviors. As an exception, we divided the tail of the MAD expression distribution (B) into two overlapping classes: the "high" class consists of the 10% genes with highest MAD expression, whereas the "medhigh" class consists of the top 40% of genes. In addition to these 18 sets, we built a set of all genes classified as stalled by Hendrix et al. (2008).

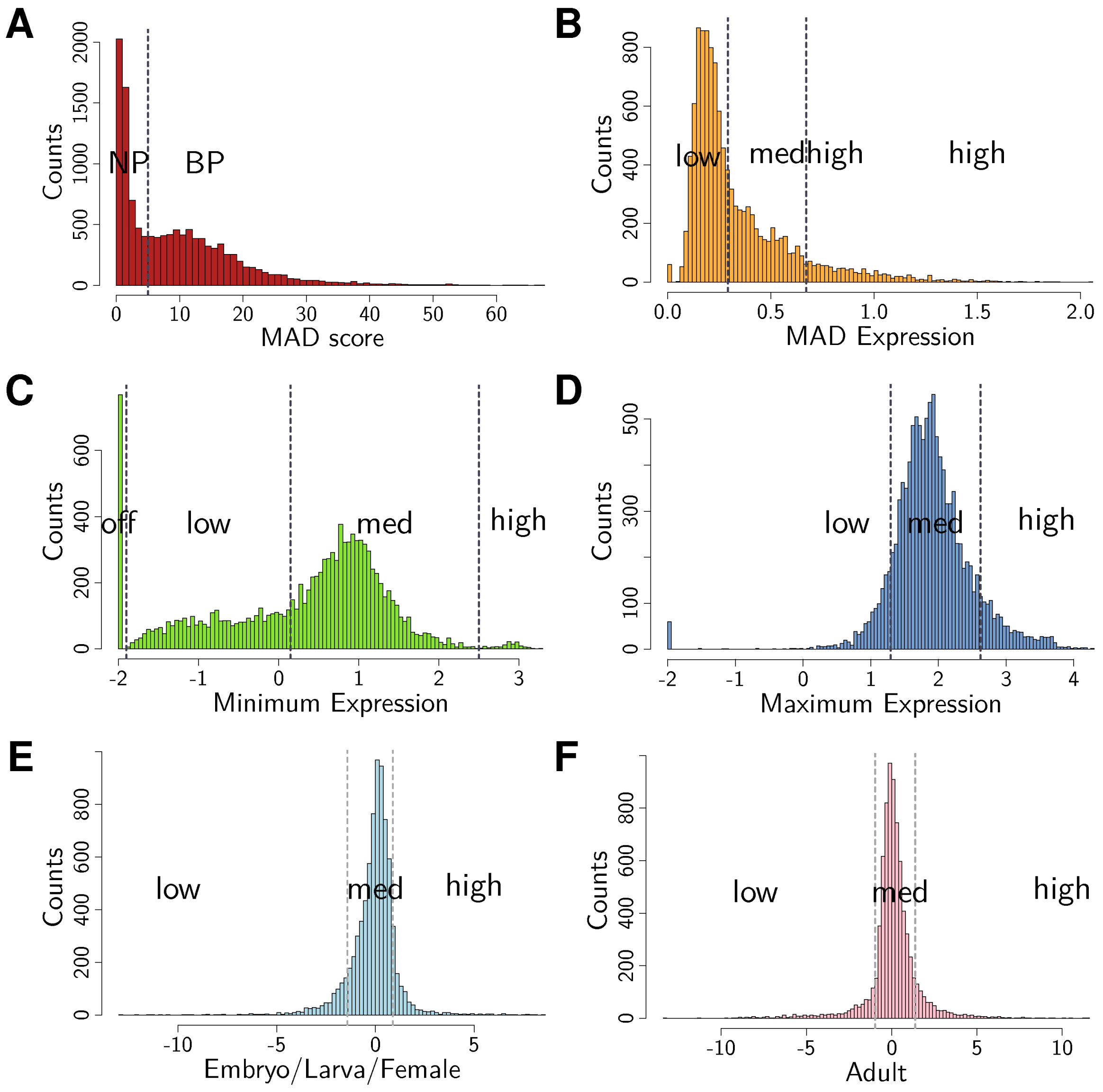


Figure 6.2.: Distribution of all genes depending on TSS cluster width (A), strength of gene regulation (B), minimum and maximum expression rate within 30 different developmental stages (C, D), and gene expression within embryo / larva / female (E) or adult (F). Dotted lines represent the chosen thresholds to separate all genes into different gene sets.

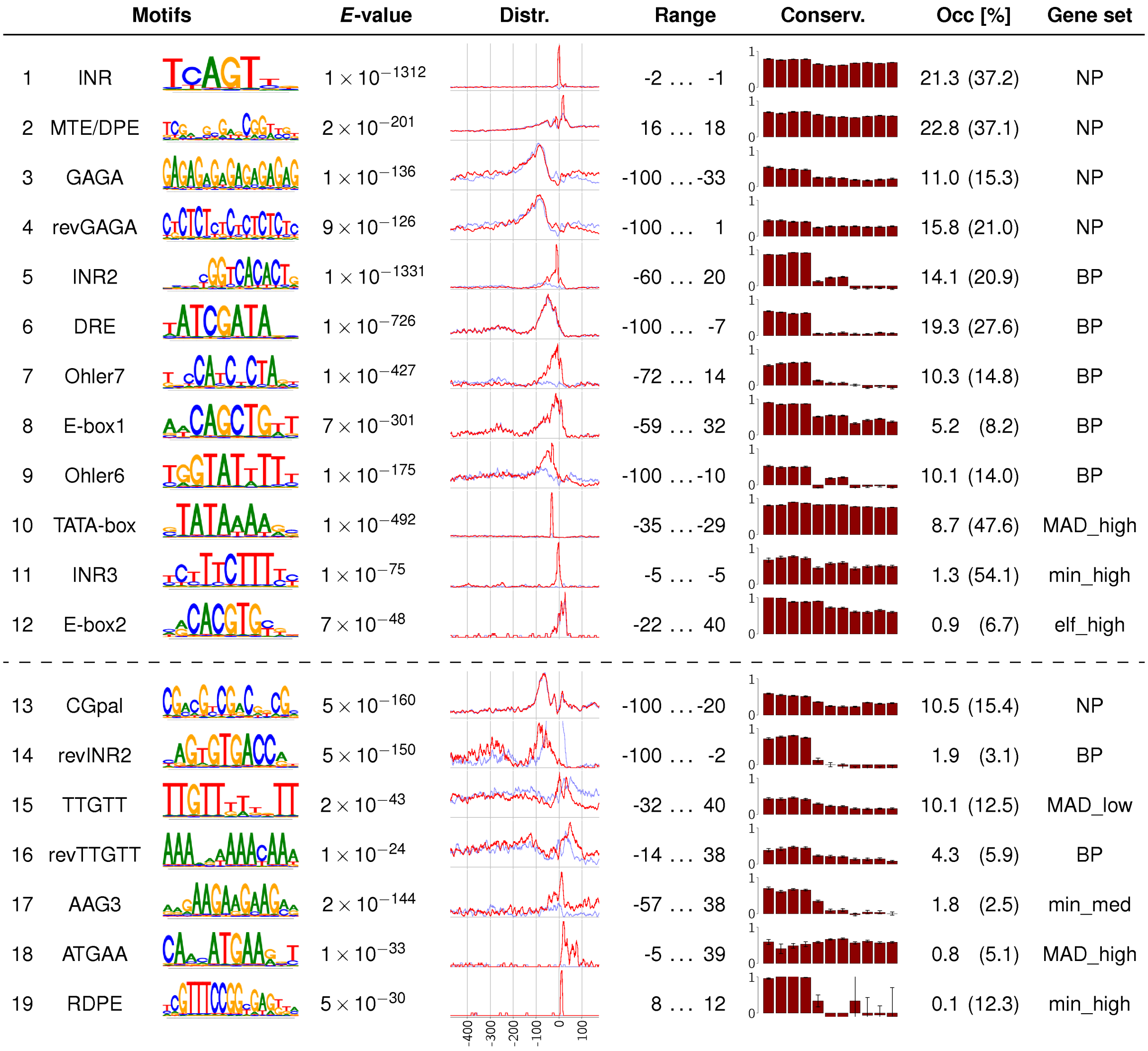
6.2.3. Identification of core promoter elements

To examine whether specific core promoter elements are enriched within the chosen gene sets (Figure 6.2), we performed a *de novo* motif search for each set separately. Therefore, we created alignments to the four most related *Drosophila* species, extracted core promoter sequences from -100 bp to $+50$ bp around the identified TSSs, and used XXmotif with the zero or one occurrence per sequence option (Section 2.2.3). If the same core promoter element was found in more than one set, we selected the motif with the lowest reported E-value as the representative for further analysis.

Table 6.1 summarizes the results of the motif search and subsequently performed analysis to validate the found motifs. In total, we were able to identify 12 previously described CPEs (Ohler et al. (2002), FitzGerald et al. (2006), Parry et al. (2010)) as well as 7 new elements. All identified CPEs are highly significant with E-values ranging from 7×10^{-48} to 1×10^{-1331} for already known motifs, and 1×10^{-24} to 5×10^{-160} for the newly identified motifs. Three CPEs are precisely located at the TSS, which we call based on their localization INR, INR2, and INR3. In contrast to a previous analysis (Ohler et al., 2002), we did not identify two distinct motifs for DPE and MTE, but only one motif that overlaps both elements, hence called MTE/DPE. Moreover, we identified two different E-box variants containing the known E-box consensus CANNTG that is bound by basic helix-loop-helix leucine zipper (bHLH-zip) transcription factors: E-box1 and E-box2. E-box1 consists of the CAGCTG consensus and was computationally identified by FitzGerald et al. (2006). E-box2 consists of the CACGTG consensus, is positioned with respect to the TSS (Hulf et al., 2005), and bound by Myc-Max complexes that activate the transcription of nearby genes (Amati et al., 2001).

To assign a motif match in a promoter region as a binding site we used two criteria: (1) If the motif has a significantly non-random localization, the binding site has to be within the region of enrichment. This region was taken from the output of XXmotif. (2) The match score of the PWM to the binding site has to be above a motif specific minimal score threshold. To determine this minimal score threshold we optimized the mutual information between the motif and all gene sets separately, which corresponds to an optimization of the TF concentration (Section 6.4.2). The gene set with the highest mutual information (and positive correlation) to the motif is given in column "Gene set". As already shown by Hoskins et al. (2011), INR, MTE/DPE, GAGA, and revGAGA are enriched within narrow peak promoters (NP), and INR2, DRE, Ohler7, E-box1, and Ohler6 within broad peak promoters (BP). However, TATA-boxes, INR3, and E-box2 seem to show special properties, as they are specifically enriched in the gene sets MAD high (strongly regulated genes), min high (strongly expressed in every developmental stage), and elf high (upregulated in embryo, larvae, and female), respectively.

Column "Distr" of Table 6.1 depicts the distribution of all assigned binding sites within


-400
-300
-200
-100
0
100
1
0

	Motifs	*E*-value	Distr.	Range	Conserv.	Occ [%]	Gene set
1	INR	1×10^{-1312}		-2 … -1		21.3 (37.2)	NP
2	MTE/DPE	2×10^{-201}		16 … 18		22.8 (37.1)	NP
3	GAGA	1×10^{-136}		-100 … -33		11.0 (15.3)	NP
4	revGAGA	9×10^{-126}		-100 … 1		15.8 (21.0)	NP
5	INR2	1×10^{-1331}		-60 … 20		14.1 (20.9)	BP
6	DRE	1×10^{-726}		-100 … -7		19.3 (27.6)	BP
7	Ohler7	1×10^{-427}		-72 … 14		10.3 (14.8)	BP
8	E-box1	7×10^{-301}		-59 … 32		5.2 (8.2)	BP
9	Ohler6	1×10^{-175}		-100 … -10		10.1 (14.0)	BP
10	TATA-box	1×10^{-492}		-35 … -29		8.7 (47.6)	MAD_high
11	INR3	1×10^{-75}		-5 … -5		1.3 (54.1)	min_high
12	E-box2	7×10^{-48}		-22 … 40		0.9 (6.7)	elf_high
13	CGpal	5×10^{-160}		-100 … -20		10.5 (15.4)	NP
14	revINR2	5×10^{-150}		-100 … -2		1.9 (3.1)	BP
15	TTGTT	2×10^{-43}		-32 … 40		10.1 (12.5)	MAD_low
16	revTTGTT	1×10^{-24}		-14 … 38		4.3 (5.9)	BP
17	AAG3	2×10^{-144}		-57 … 38		1.8 (2.5)	min_med
18	ATGAA	1×10^{-33}		-5 … 39		0.8 (5.1)	MAD_high
19	RDPE	5×10^{-30}		8 … 12		0.1 (12.3)	min_high

Table 6.1.: Core promoter motifs detected by XXmotif on 19 different gene sets in *D. melanogaster*. If the same motif was found in different gene sets, the motif with the best *E*-value was chosen. Motifs above the dashed line are previously described in literature, motifs below are newly detected. The gene set with the highest mutual information (and positive correlation) to the motif is given in column "Gene set". Column "Distr" depicts the distribution of all assigned binding sites within the gene set having the highest mutual information ("Gene set") smoothed over five nucleotides. The region with the highest enrichment of binding sites is taken from the XXmotif out ("Range"). Column "Conserv" indicates the average conservation of binding sites within related species. whereas 1 is perfect conservation, and 0 no conservation. The 11 bars correspond to related *Drosophila* species, ordered by ascending evolutionary distance. Column "Occ [%]" gives the frequency of motif sites within the whole sequence set (the gene set of highest mutual information).

the gene set having the highest mutual information ("Gene set") from −500 bp to +200 bp with respect to the TSS smoothed over ten nucleotides. Considering only the forward strand (red curve), two groups of CPEs can be detected: CPEs that have a per nucleotide defined position with respect to the TSS (i. e., INR, MTE/DPE, INR2, TATA-box, INR3), and CPEs that show an enrichment within 50 – 100 nucleotides (i. e., GAGA, DRE, E-box1, Ohler6, E-box2). Our newly identified motifs mainly fit into the second group (i. e., CGpal, revINR2, TTGTT, AAG3, ATGAA), however, we could also identify a new CPE that is located within a range of only five nucleotides with respect to the TSS (RDPE). The range of the most significant motif enrichment as given by XXmotif is depicted within column "Range". Considering both forward and reverse strand allows for the analysis of strand specificity. Again, two groups can be detected: Strand specific CPEs (e. g., INR, MTE/DPE, Ohler7), and CPEs that are located on both strands (e. g., GAGA, DRE, E-box1). Our newly identified motifs fall into both classes. AAG3, ATGAA, and RDPE show enrichment only on the forward strand, whereas the palindromic motif CGpal and TTGTT are found on both strands with similar frequency. Interestingly, revINR2 has a very low frequency on the forward strand compared to its known reverse complement INR2, and shows an enrichment of binding sites not only within the core promoter region from −100 bp to −2 bp with respect to the TSS, but also between −500 bp and −200 bp.

To measure conservation information, we analyzed the difference of PWM scores in *D. melanogaster* to 11 related species (Figure 6.3). In a first step, we correlated the PWM scores from binding sites of *D. melanogaster* to aligned binding sites from each related species. Column "PWM Scores" shows the scatter plots of these correlations for the known motif MTE/DPE, the newly identified motifs CGpal and ATGAA, as well as a negative control (AACCTTGG). The remaining motifs and more negative controls are shown in Figure B.2. The average PWM score distance between *D. melanogaster* and the related species ordered by evolutionary distance for all binding sites passing the minimal score threshold and enriched region filter is shown in Column "Score Distance" as a red circle. The boxes correspond to the expected score distance calculated on aligned binding sites from sampled sequences (Section 6.4.3). The final conservation plot (column "Conservation") depicts a scaled measure of the sampled and biological score distances, with a one denoting perfect conservation of the PWM score and a zero no conservation (Section 6.4.3). The barplots in column "Conserv." of Table 6.1 summarize these conservation plots. In general, three types of conservation can be identified: First, CPEs that are equally conserved in the whole *Drosophila* group (i. e., INR, MTE/DPE, TATA-box, INR3, E-box2, ATGAA), second, CPEs that are only conserved in the *melanogaster* subgroup consisting of the four closest related species (i. e., INR2, DRE, Ohler7, Ohler6, revINR2, AAG3, RDPE), and third, CPEs that are well conserved within the *melanogaster* subgroup, but only moderately conserved within the whole *Drosophila* group (i. e., GAGA, revGAGA, E-box1, CGpal, TTGTT, revTTGTT).

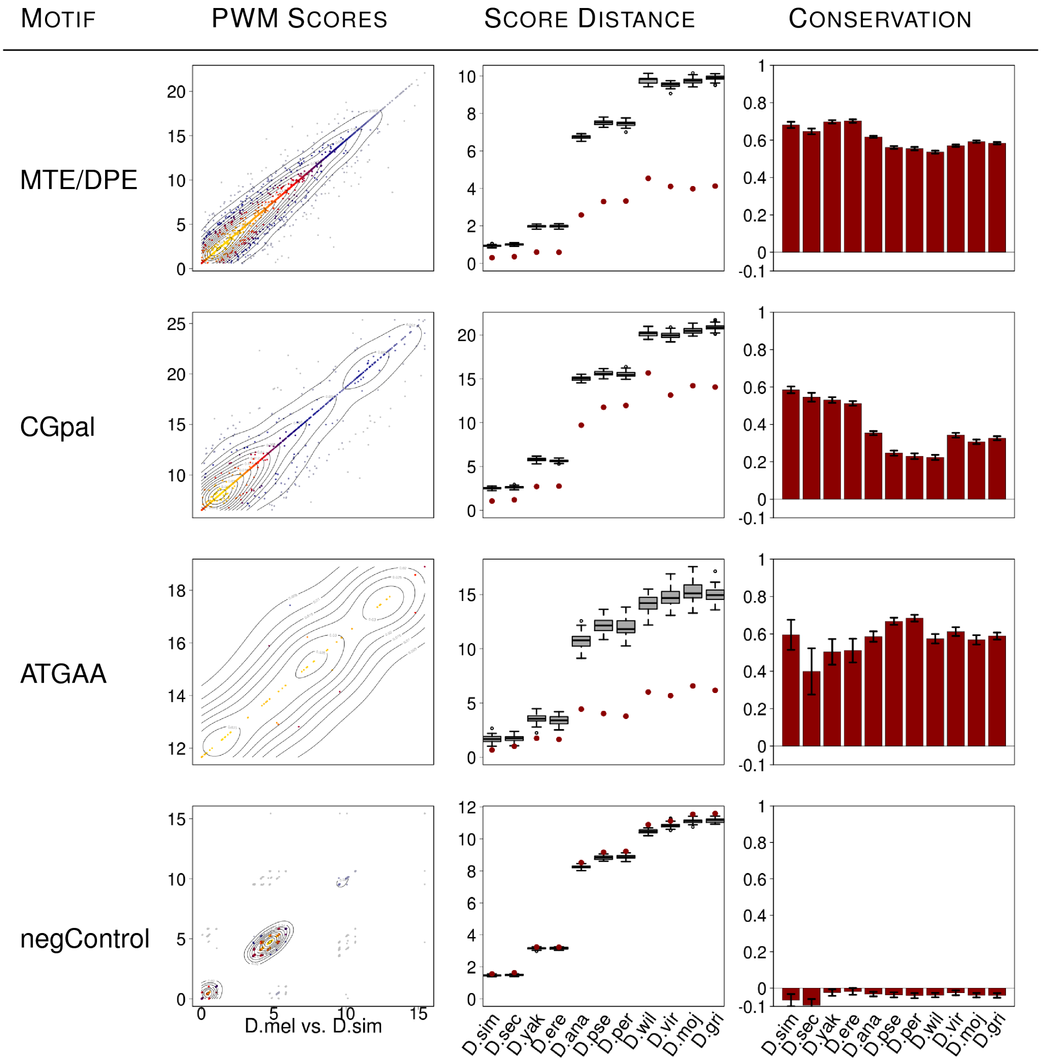


Figure 6.3.: Conservation of newly predicted motifs compared to MTE/DPE and AACCTTGG as a negative control. Column "PWM Scores" depicts the scores of the motif PWM for each site in *D. melanogaster* and the aligned site in *D. simulans*. . The average PWM score distance between *D. melanogaster* and related species is shown in Column "Score Distance" as a red circle. To calculate the average PWM score we use only binding sites with a PWM score above the minimal score threshold and within the enriched region. Related species are ordered by evolutionary distance. The boxes correspond to the expected score distance calculated on aligned binding sites from sampled sequences. Column "Conservation" depicts a scaled measure of the sampled and biological score distances, giving a one for perfect conservation of the PWM score and a zero for conservation as expected from background. Error bars indicate the standard deviation over all sampled conservation scores (Section 6.4.3).

6.2.4. Core promoter elements allow for the prediction of gene properties

To analyze the influence of TSS cluster width, expression in developmental stages, and stalling index on the enrichment of CPEs, we ordered all genes depending on each property (Figure 6.4, Figure B.3) and calculated Z-scores for the enrichment of every CPE within bins of 50 genes. Strikingly, sets of CPEs show transitions from correlated to anticorrelated and vice versa at defined scores. An example is the cluster width of five that is also the value estimated from the distribution in Figure 6.2A to separate NP and BP promoters. More specific sets of CPEs are enriched in stalled genes (INR, MTE/DPE, GAGA) (B), genes with a high expression in every developmental stage (Ohler6, Ohler7, INR3, RDPE) (C), genes with a high expression in at least one developmental stage (TATA-box, ATGAA, INR3, RDPE) (D), and within the most regulated genes (TATA-box, ATGAA) (E). Correlating all CPEs to all gene sets from Figure 6.2 provides a global view on the architecture of core promoters in *D. melanogaster* (F). In total, four sets of CPEs with similar correlations to all considered gene sets are observable, and E-box2 as a specific element for the elf high gene set consisting of up-regulated genes within embryo, larvae, and female. This special role of E-box2 fits to the importance of the Myc protein that binds the E-box2 consensus, and is crucial in controlling cellular proliferation and growth during development (Oster et al., 2002).

Class 1 (INR, MTE/DPE, GAGA, CGpal) and Class 2 (TATA-box, ATGAA) CPEs are both present in genes with narrow peak promoters (NP). However, whereas Class 1 enriched genes show no extreme expression values (max high, min low), are only slightly regulated (MAD medhigh), and show strong correlations to stalled genes (stalledPol), Class 2 enriched genes belong to the 10% most highly expressed genes in at least one developmental stage (max high), and to the most strongly regulated genes (MAD high). Class 3 (INR2, Ohler6, DRE, Ohler7, E-box1, revINR2, TTGTT) is the only class of CPEs present in broad peak promoters (BP). Enriched genes are similarly expressed in all developmental stages (MAD low), a main feature of genes having housekeeping function. Class 4 (INR3, RDPE) is like Class 2 enriched within the 10% most strongly expressed genes (max high) and mainly present in narrow peak promoters (NP). However, Class 4 genes are in contrast to all other classes strongly expressed in all developmental stages (min high). A gene ontology (GO) analysis of this set of genes clearly describes them as being ribosomal (Table B.1). The only identified CPE that shows no significant enrichment to any of the 19 gene sets is AAG3. It might either be of general importance, or specific for a gene set not included in our analysis.

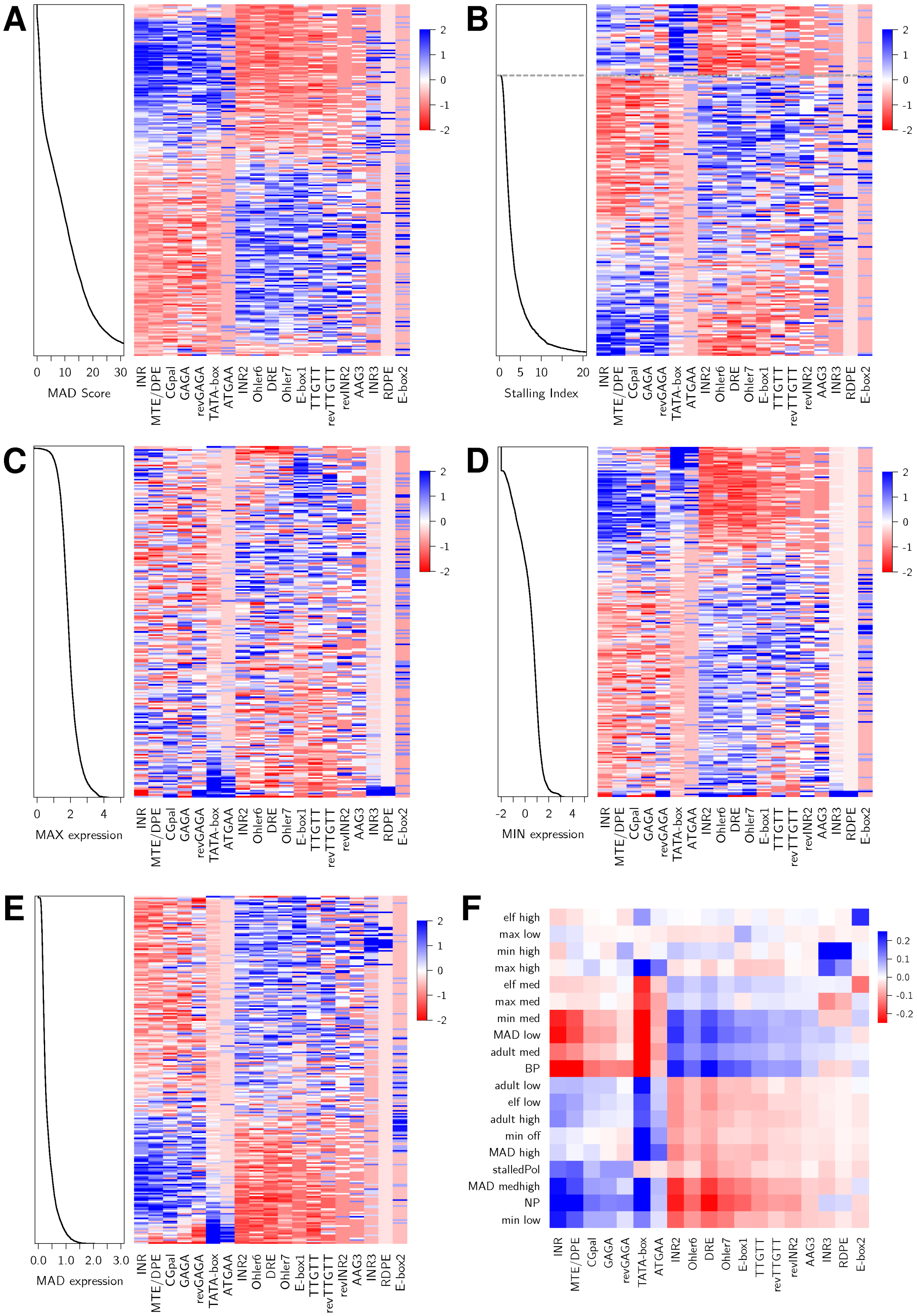

A
0 10 20 30
MAD Score
B
0 5 10 20
Stalling Index
C
0 2 4
MAX expression
D
-2 0 2 4
MIN expression
E
0.0 1.0 2.0 3.0
MAD expression
F
elf high
max low
min high
max high
elf med
max med
min med
MAD low
adult med
BP
adult low
elf low
adult high
min off
MAD high
stalledPol
MAD medhigh
NP
min low
0.2
0.1
0.0
-0.1
-0.2
2
1
0
-1
-2
INR
MTE/DPE
CGpal
GAGA
revGAGA
TATA-box
ATGAA
INR2
Ohler6
DRE
Ohler7
E-box1
TTGTT
revTTGTT
revINR2
AAG3
INR3
RDPE
E-box2

Figure 6.4.: Fly core promoter motifs are correlated to distinct gene properties. In (A–E), genes are sorted by five scores and the frequency of the core promoter motifs on the x-axis within bins of 50 genes is indicated from red (depleted) to blue (enriched) by Z-scores. (A) TSS cluster width derived from CAGE data (Ni et al. (2010), Hoskins et al. (2011)). (B) Stalling index (Zeitlinger et al., 2007), (C) minimum expression, (D) maximum expression, (E) strength of regulation over 30 developmental time points (Graveley et al., 2011). (F) Correlation of core promoter elements to 19 different gene sets reveals four classes of elements with similar MCC values to all sets.

6.2.5. Core promoter elements belong to specific architectures

To analyze whether different CPEs occur within the same core promoter, we calculated the correlations between all CPEs (Figure 6.5). In agreement with the four identified classes, most CPEs are positively correlated to all elements within their class and negatively correlated to CPEs belonging to other classes. Only the Class 4 elements are positively correlated to some motifs of especially Class 3.

The co-occurrence of CPEs within one class indicates the usage of a specific transcription initiation complex utilizing several binding sites for each class. One such example is the TFIID complex that assembles at the DNA due to interactions to the Class 1 element INR bound by the subunits TAF1 and TAF2 (Smale and Baltimore, 1989), and the DPE element bound by the subunits TAF6 and TAF9 (Burke and Kadonaga, 1996). Hence, it is likely that the remaining Class 1 elements also contribute to the binding of TFIID. Within Class 2,

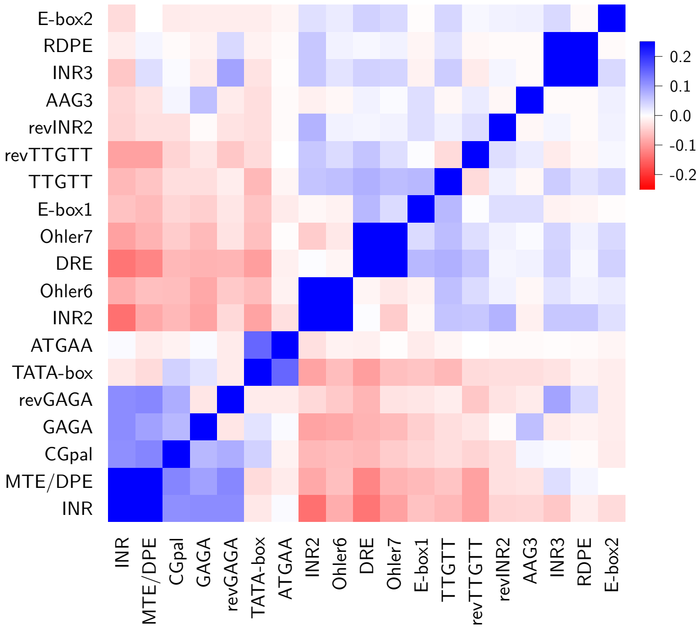


Figure 6.5.: Fly core promoter elements occur in defined architectures. Correlation of all CPEs to each other reveals elements that occur preferentially within the same promoter (positive MCC values, blue) or avoid each other (red).

TATA-boxes are known to be bound by TBP (Goldberg, 1979) that also belongs to the TFIID complex. However, TATA-boxes are anti-correlated to the MTE/DPE element, suggesting that the new element ATGAA replaces the MTE/DPE element in Class 2 promoters. Similar hypothesis can be stated for Class 4 promoters consisting of INR3 and RDPE. As shown by Parry et al. (2010), genes containing the INR3 are not regulated by TFIID, but by a special RNA polymerase II system for ribosomal protein genes. As genes containing the RDPE element are clearly ribosomal (Table B.2), we named the downstream of the TSS located element RDPE (ribosomal downstream promoter element) by its potential function to substitute DPE in the ribosomal system of transcription initiation.

Negative correlations between elements within the same class are only found for elements located on both strands (e. g., GAGA vs. revGAGA, TTGTT vs. revTTGTT) and for two groups of elements within Class 3. The first group of elements (Class 3A) consists of INR2 and Ohler6. Both elements are strongly correlated, which indicates their binding to the same complex. In contrast, the second group of elements, DRE, Ohler7, and E-box1 (Class 3B), are all anticorrelated to the Class 3A elements, indicating a different mechanism of transcriptional initiation. DRE promoters are known to be bound by TBP-related factor 2 (TRF2) associating with the DRE-binding factor DREF (Hochheimer et al., 2002). The remaining elements of Class 3 (TTGTT, revTTGTT) show weak positive correlations to all elements of their class suggesting that both transcription initiating complexes present in this class share DNA binding subunits.

6.2.6. Each class of core promoters has defined physical properties

To analyze whether the identified classes of core promoters differ not only in their motif composition, but also in the physical properties of their surrounding promoter region, we examined the average dinucleotide frequency over all promoters within each of the four classes of core promoters (Figure 6.6). All classes show a strong composition bias for A and T containing dinucleotides, preferentially for 'AA' and 'TT' adjacent to the core promoter region located between −100 to +50 bps with respect to the TSS. However, the classes vary strongly in the shape of A/T enrichment and the most frequently occurring dinucleotides. Class 1 promoters (A) show a strong 'AA' vs 'TT' bias within 500 bps downstream of the TSS that is reduced to around 150 bps for Class 2 promoters. Class 3 promoters possess two peaks of A/T enrichment, one ∼150 bp upstream and the other 75 bp downstream of the TSS. The downstream peak consists preferentially of 'AA' and 'TT' dinucleotides, whereas the 'TT' peak is located around 20 bps upstream of the 'AA' peak. In addition to the two peaks of Class 3, Class 4 promoters show a second 'AA' peak located at around 150 bps downstream of the TSS.

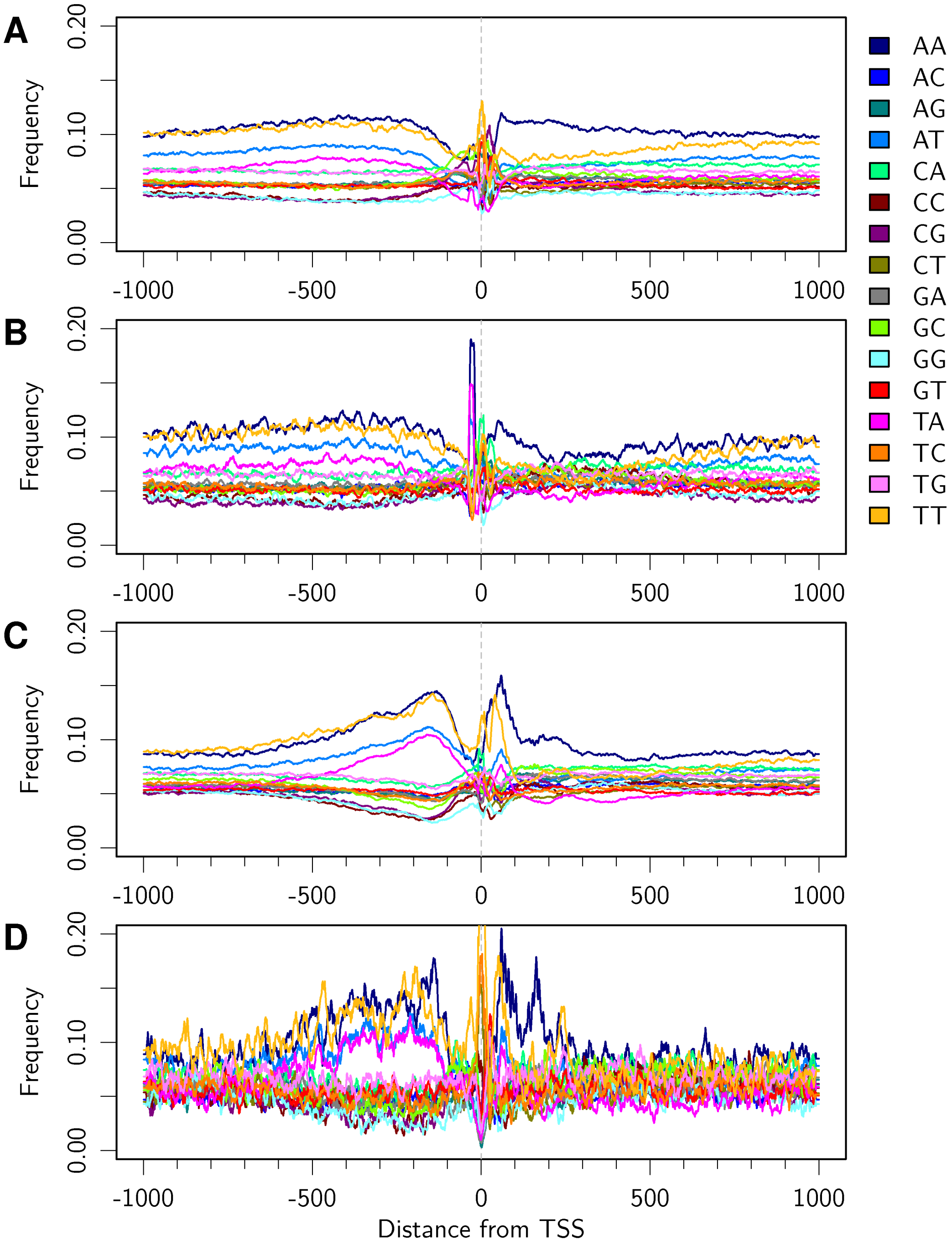


Figure 6.6.: Each class of core promoter elements belongs to genes with distinct dinucleotide patterns. Each line corresponds to the average frequency of a dinucleotide over all promoters within the respective class, smoothed over 15 nucleotides. (A) Class 1 promoters (regulated genes), (B) Class 2 promoters (highly regulated genes), (C) Class 3 promoters (housekeeping genes), (D) Class 4 promoters (ribosomal genes).

6.3. Summary

Our analysis revealed 12 known and 7 novel core promoter motifs in *D. melanogaster* that are all conserved and show a clear localization to the TSS. Strikingly, the major core promoter element composition is sufficient to define fairly reliably to which of four classes a gene belongs (Figure 6.7). Genes with narrow, focused TSS clusters are strongly enriched in the green Class 1 and orange Class 2 motif groups, whereas genes with broad TSS clusters are enriched in the blue Class 3 motif group. Genes constitutively expressed at high levels (mostly ribosomal protein genes) are strongly enriched in the red Class 4 motif group, whereas genes that are switched off completely in some conditions are strongly enriched for motifs from the orange and green groups. Genes that can be regulated by stalling after initiation are highly enriched for green motifs. These four classes are characterized by their minimum expression, maximum strength, variability of expression, stallability, and TSS cluster width. Our results thus demonstrate the importance of core promoter sequences in shaping transcriptional behavior.

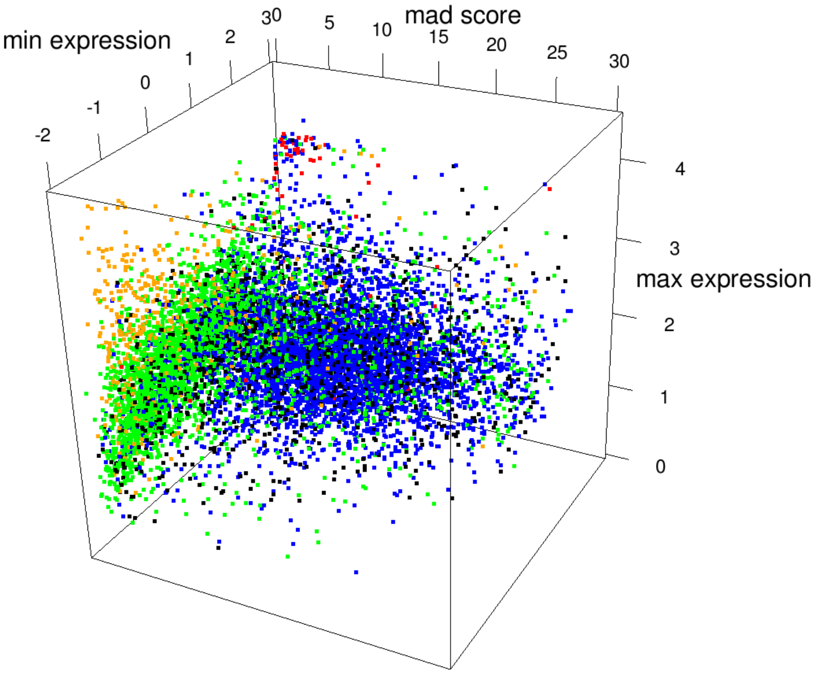


Figure 6.7.: Four classes of core promoters are separated by cluster width (x-axis), minimum expression strength (y-axis), and maximum expression strength (z-axis). Each point corresponds to a promoter colored by the core promoter class with the highest scoring motif. Green: Class 1 (regulated genes), yellow: Class 2 (highly regulated genes), blue: Class 3 (housekeeping genes), red: Class 4 (ribosomal genes), black: no binding site of any motif.

6.4. Materials and Methods

6.4.1. Tag clustering

All tags from two CAGE datasets (Ni et al. (2010), Hoskins et al. (2011)) were pooled and smoothed using a square kernel function of width 41. Clusters were defined as continuous regions with a tag distribution higher than the genomic average. For each cluster, the TSS was declared at the position with the most assigned tags. For further analysis, we only used clusters with at least five annotated tags and no other TSS within a range of 150 bps. Furthermore, we only considered clusters with either an annotated gene start within 250 bps downstream of the TSS, have the TSS within an annotated 5'UTR, or contain an annotated FlyBase (McQuilton et al., 2012) TSS within the cluster. The clustering resulted in 12061 different TSS clusters for 8502 different genes.

6.4.2. Optimization of minimal score thresholds

To determine the minimal log-odd score of a PWM indicating the presence of a motif, we calculated the mutual information between the motif and all gene sets to which the PWM has a positive correlation given all minimal score thresholds ranging from -15 to 30 with step size 0.1. For each motif we chose the minimal score threshold leading to the highest mutual information in a gene set. Low complexity motifs (TTGTT, revTTGTT) that have the highest mutual information by accounting only for their nucleotide composition (threshold around zero) were manually set to 50% of their maximal log-odds score.

6.4.3. Conservation scores

For each motif, conservation scores were calculated on the best scoring site above the minimal score threshold within the defined range of motif enrichment on all 12061 *D. melanogaster* core promoter sequences. Alignments were generated using the UCSC 14-way multiple sequence alignments (dm3). The conservation score $S_{\text{cons}}(X)$ for each of the 11 Drosophila species X was calculated as follows:

$$S_{\text{cons}}(X) = \frac{1}{N} \sum_{i=1}^{N} \frac{B_i(X) - S(X)}{B_i(X)} \tag{6.2}$$

whereas $S(X)$ is the average log-odds score difference between *D. melanogaster* and species X from the alignment, and $B_i(X)$ is the expected average log-odds score difference from a null distribution based on the ith of N sets of sampled binding sites. Each binding site is sampled from a position specific substitution matrix learned on the alignment to species X at the respective position +/- 10 bps. We used $N = 50$ for the analysis.

Part III.

Appendix

A. Detailed Performance of XXmotif

A.1. Drabløs Benchmark

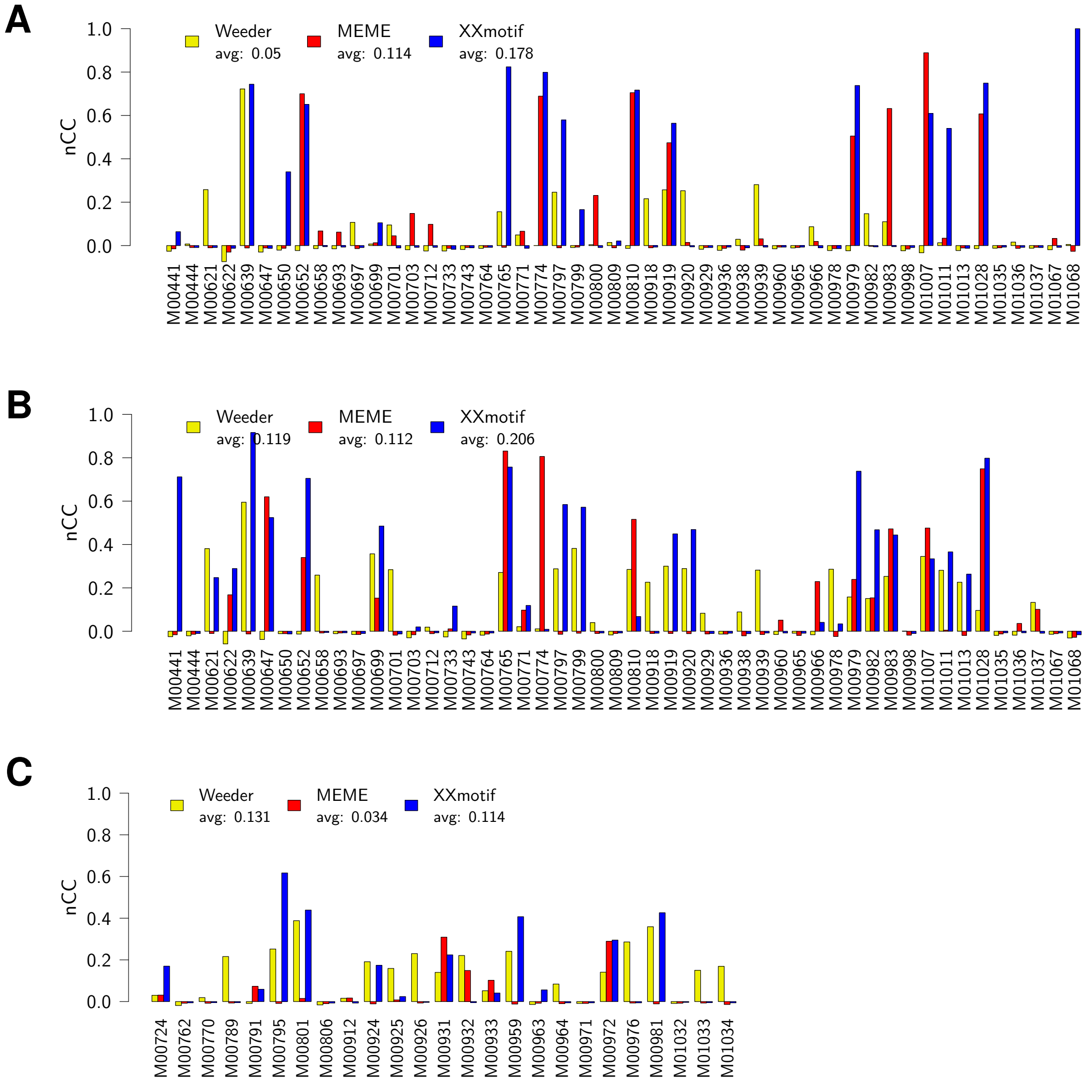


Figure A.1.: Detailed results of the Drabløs benchmark. The x-axis gives the names of the used data sets, the y-axis the Matthews correlation coefficient on the nucleotide level (A) Algorithm Markov, (B) Algorithm Real, (C) Model Real

A.2. Motif Sensitivity Benchmark

A

	Harbison		Bulyk		Hughes	
	TOP1	TOP4	TOP1	TOP4	TOP1	TOP4
MEME	35	98	18	57	19	65
PRIORITY	70	92	33	43	36	53
MEME-$\mathcal{M}$	67	97	35	53	39	70
Weeder	65	86	43	54	40	53
AMADEUS	74	96	32	42	45	65
MEME-$\mathcal{D}$	74	105	34	57	45	74
MEME-$\mathcal{DC}$	74	106	35	59	46	76
PRIORITY-$\mathcal{D}$	79	93	36	45	41	58
PRIORITY-$\mathcal{DC}$	79	93	34	44	43	51
XXmotif	96	122	54	70	**67**	**97**
XXmotif-$\mathcal{C}$	**101**	**127**	**56**	**79**	63	94
ERMIT	88	115	36	51	56	77
cERMIT	88	119	39	60	50	80

B

	Harbison		Bulyk		Hughes	
	TOP1	TOP4	TOP1	TOP4	TOP1	TOP4
MEME	33	99	18	54	20	65
PRIORITY	68	90	32	44	37	51
MEME-$\mathcal{D}$	68	98	31	55	44	67
PRIORITY-$\mathcal{D}$	73	87	32	38	40	48
MEME-$\mathcal{M}$	69	103	38	58	41	72
AMADEUS	73	97	31	46	46	67
Weeder	70	87	40	55	41	52
MEME-$\mathcal{DC}$	75	105	38	64	49	80
PRIORITY-$\mathcal{DC}$	82	95	37	44	47	55
XXmotif	96	122	54	70	**67**	**97**
XXmotif-$\mathcal{C}$	**101**	**127**	**56**	**79**	63	94
ERMIT	91	117	37	63	48	81
cERMIT	94	117	43	63	53	83

Table A.1.: Detailed results of motif sensitivity benchmark. The tools are sorted by the sum of the top 1 predictions. Highest number per benchmark set is given in bold face. Methods above the separator take only intergenic regions with a ChIP-chip P-value $< 10^{-3}$ as input, methods below the separator take all intergenic regions and require the associated P-valueas additional information. (A) XXmasker is applied only to the input of XXmotif. (B) XXmasker is applied to the input of all tools.

A.3. PWM Quality Benchmark

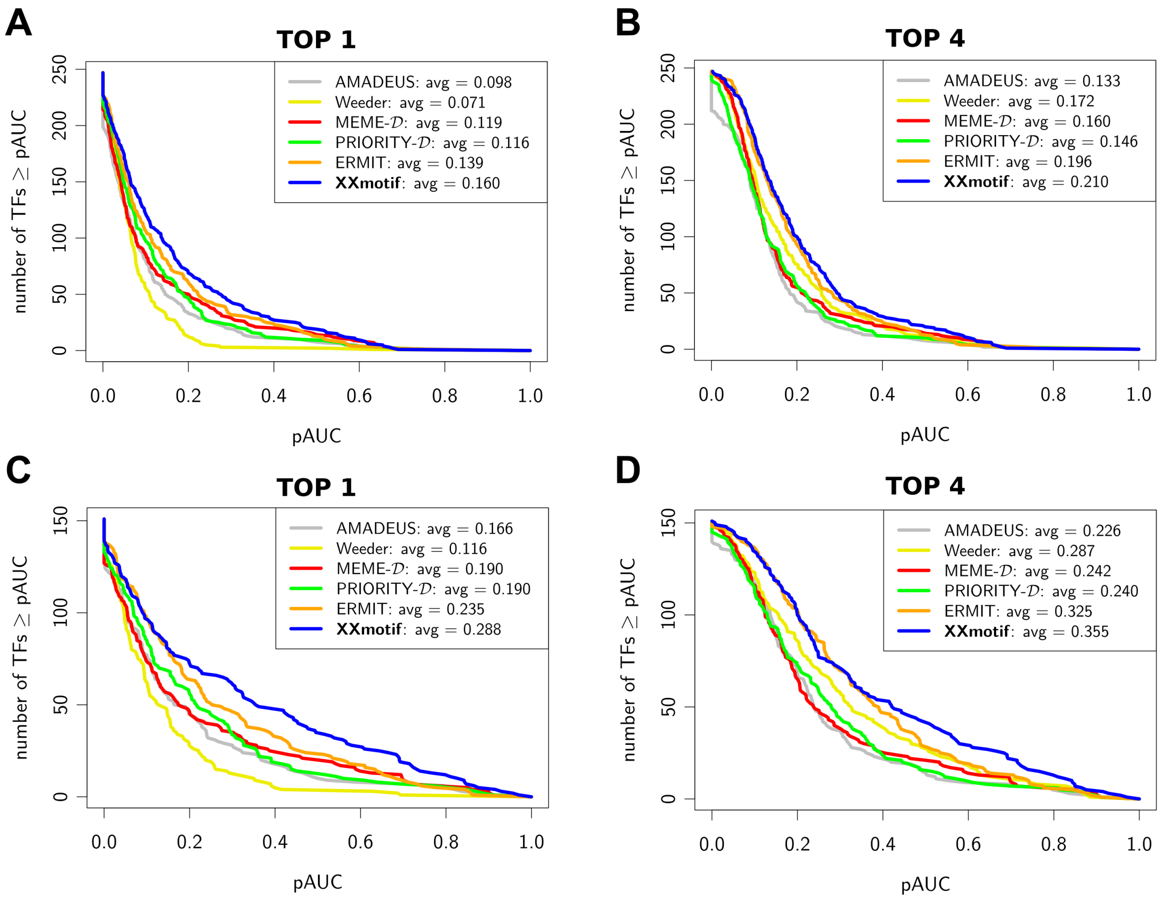
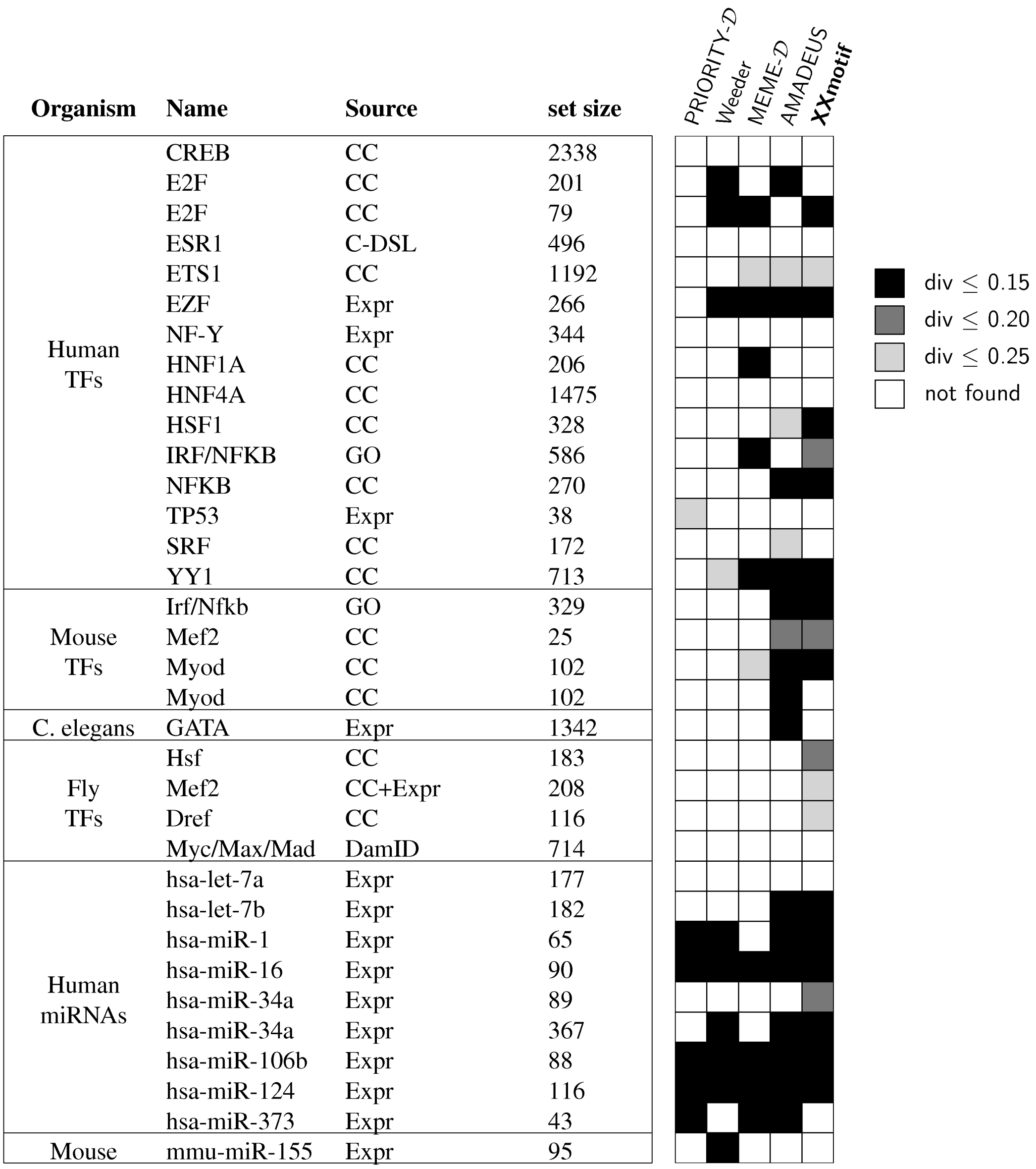


Figure A.2.: PWM quality assessment on yeast ChIP-chip data from Harbison et al. (2004). The curves quantify how well the scores of the reported PWMs can predict the ChIP enrichment of the sequences. Each PWM is used to rank the intergenic regions by their maximum PWM score. For each predicted PWM, a receiver operator characteristic (ROC) curve with the number of correct predictions over the number of false predictions is computed, and the partial area under the ROC curve (pAUC) deduced from it. The pAUC is the fractional area under the ROC curve within the 5% best-ranked false predictions. For an ideal predictor, pAUC=1. The average pAUC scores are listed in the figure legends. (A, B) cumulative distribution of the pAUC over all 247 ChIP-chip datasets that had at least ten significantly enriched regions (P-value < 0.001). Regions with ChIP enrichment P-value < 0.001 are defined as correct predictions, all other regions as false predictions. (C, D) As in A, B but using only datasets that have at least five significantly ChIP-enriched regions with matches to the literature motif, and considering only sequences that contain a match to the literature motif.

A.4. Metazoan Benchmark



Organism	Name	Source	set size	*PRIORITY-𝒟*	*Weeder*	*MEME-𝒟*	*AMADEUS*	***XXmotif***
Human TFs	CREB	CC	2338	□	□	□	□	□
	E2F	CC	201	□	■	□	■	□
	E2F	CC	79	□	■	■	□	■
	ESR1	C-DSL	496	□	□	□	□	□
	ETS1	CC	1192	□	□	▨	▨	▨
	EZF	Expr	266	□	■	■	■	■
	NF-Y	Expr	344	□	□	□	□	□
	HNF1A	CC	206	□	□	■	□	□
	HNF4A	CC	1475	□	□	□	□	□
	HSF1	CC	328	□	□	□	▨	■
	IRF/NFKB	GO	586	□	□	■	□	▩
	NFKB	CC	270	□	□	□	■	■
	TP53	Expr	38	▨	□	□	□	□
	SRF	CC	172	□	□	□	▨	□
	YY1	CC	713	□	▨	■	■	■
Mouse TFs	Irf/Nfkb	GO	329	□	□	□	■	■
	Mef2	CC	25	□	□	□	▩	▩
	Myod	CC	102	□	□	▨	■	■
	Myod	CC	102	□	□	□	■	□
C. elegans	GATA	Expr	1342	□	□	□	■	□
Fly TFs	Hsf	CC	183	□	□	□	□	▩
	Mef2	CC+Expr	208	□	□	□	□	▨
	Dref	CC	116	□	□	□	□	▨
	Myc/Max/Mad	DamID	714	□	□	□	□	□
Human miRNAs	hsa-let-7a	Expr	177	□	□	□	□	□
	hsa-let-7b	Expr	182	□	□	□	■	■
	hsa-miR-1	Expr	65	■	■	□	■	■
	hsa-miR-16	Expr	90	■	■	■	■	■
	hsa-miR-34a	Expr	89	□	□	□	□	▩
	hsa-miR-34a	Expr	367	□	■	□	■	■
	hsa-miR-106b	Expr	88	■	■	■	■	■
	hsa-miR-124	Expr	116	■	■	■	■	■
	hsa-miR-373	Expr	43	■	□	■	■	□
Mouse	mmu-miR-155	Expr	95	□	■	□	□	□

■ div ≤ 0.15
▩ div ≤ 0.20
▨ div ≤ 0.25
□ not found

Figure A.3.: Top 1 benchmark results on 24 target sets for transcription factors from human, mouse, worm and fly, as well as 10 target sets for microRNAs from human, and mouse from the metazoan target set compendium (Linhart et al. 2008). The plot is adapted from Linhart et al. (2008): The "Source" column indicates the experimental procedure or database from which the target set was derived: Gene expression microarrays (Expr), ChIP-chip (CC), ChIP-DSL (C-DSL), DamID (van Steensel et al., 2001), or Gene Ontology (GO) database (Ashburner et al., 2000). Black and gray boxes indicate the similarity of the predicted PWM to the reference motif in TRANSFAC or miRBase. Darker shades indicate closer similarity. "Set Size": number of sequences within the input set.

A.5. Regulatory Motifs For Early Embryo Segmentation in Flies

Number	Chr	Start	End	Reference
1	X	2331789	2333533	Schroeder et al. (2004)
2	X	2323048	2324286	Schroeder et al. (2004)
3	X	2324294	2325502	Schroeder et al. (2004)
4	X	2327322	2329503	Schroeder et al. (2004)
5	3R	4520323	4521043	Hox_pro
6	3R	4526520	4527542	Hox_pro
7	2R	21113281	21114511	Hoch et al. (1991)
8	2R	21110142	21111300	Hoch et al. (1991)
9	2R	21111575	21113281	Hoch et al. (1991)
10	3L	20687055	20688533	Schroeder et al. (2004)
11	3L	20692603	20694005	Schroeder et al. (2004)
12	3L	20689640	20690516	Pankratz et al. (1992)
13	3R	26676777	26677272	Liaw and Lengyel (1993)
14	3R	26677663	26678030	Liaw and Lengyel (1993)
15	3R	26675265	26675744	Rudolph et al. (1997)
16	3R	173891	174480	Häder et al. (2000)
17	3L	14166136	14167860	Schroeder et al. (2004)
18	2L	20784670	20786306	Schroeder et al. (2004)
19	3R	19020990	19022410	Schroeder et al. (2004)
20	3R	24411719	24413426	Schroeder et al. (2004)
21	3L	20604991	20606288	Schroeder et al. (2004)
22	2L	12615792	12617776	Schroeder et al. (2004)
23	X	8537082	8538914	Schroeder et al. (2004)
24	2L	12678898	12680520	Schroeder et al. (2004)
25	2L	3832698	3835337	Schroeder et al. (2004)
26	X	584106	9585905	Wimmer et al. (1995)
27	3R	9720485	9720788	Hartmann et al. (2001)
28	3R	12636231	12637975	Shimell et al. (2000)
29	X	8547931	8548807	Gao and Finkelstein (1998)
30	2L	11455640	11455917	Kühnlein et al. (1997)
31	2L	11455917	11456155	Kühnlein et al. (1997)
32	2L	21851517	21853665	Coré et al. (1997)
33	2L	3608812	3610461	Schroeder et al. (2004)
34	2L	3610420	3611803	Schroeder et al. (2004)
35	3R	2692616	2694360	Schroeder et al. (2011)
36	3R	2683373	2684612	Schroeder et al. (2011)
37	3R	2681761	2683378	Schroeder et al. (2011)
38	X	20523501	20524783	Schroeder et al. (2011)
39	X	20533075	20535598	Schroeder et al. (2011)
40	X	20548261	20549257	Schroeder et al. (2011)
41	X	20555735	20556596	Schroeder et al. (2011)
42	X	20594595	20597303	Schroeder et al. (2011)
43	X	20551039	20552655	Klingler et al. (1996)
44	X	20552655	20553990	Klingler et al. (1996)
45	2R	5863006	5863516	Small et al. (1996)
46	2R	5865217	5865879	Stanojevic et al. (1991)
47	2R	5871404	5872005	Fujioka et al. (1999)
48	2R	5873440	5874240	Fujioka et al. (1999)
49	2R	5874147	5874946	Fujioka et al. (1999)
50	3L	8657463	8658374	Howard and Struhl (1990)
51	3L	8657938	8659411	Howard and Struhl (1990)
52	3L	8659411	8660491	Howard and Struhl (1990)
53	3L	8662058	8665028	Howard and Struhl (1990)
54	2L	12080376	12081687	Schroeder et al. (2011)

Table A.2.: Coordinates of the 54 hand-curated *cis*-regulatory modules in fly segmentation

B. Drosophila Core Promoters

B.1. TSS Cluster Width

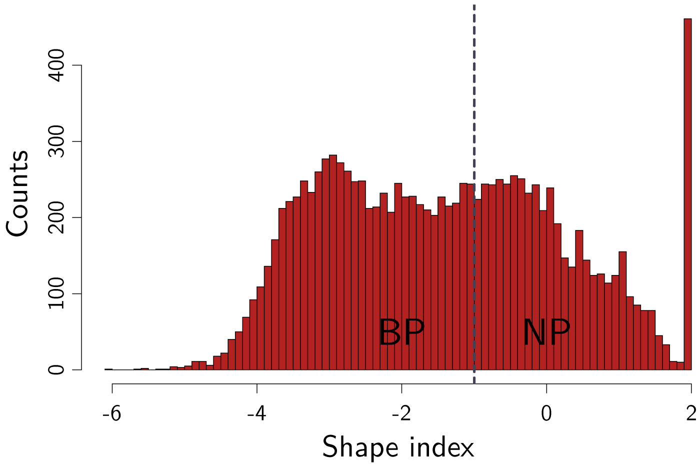
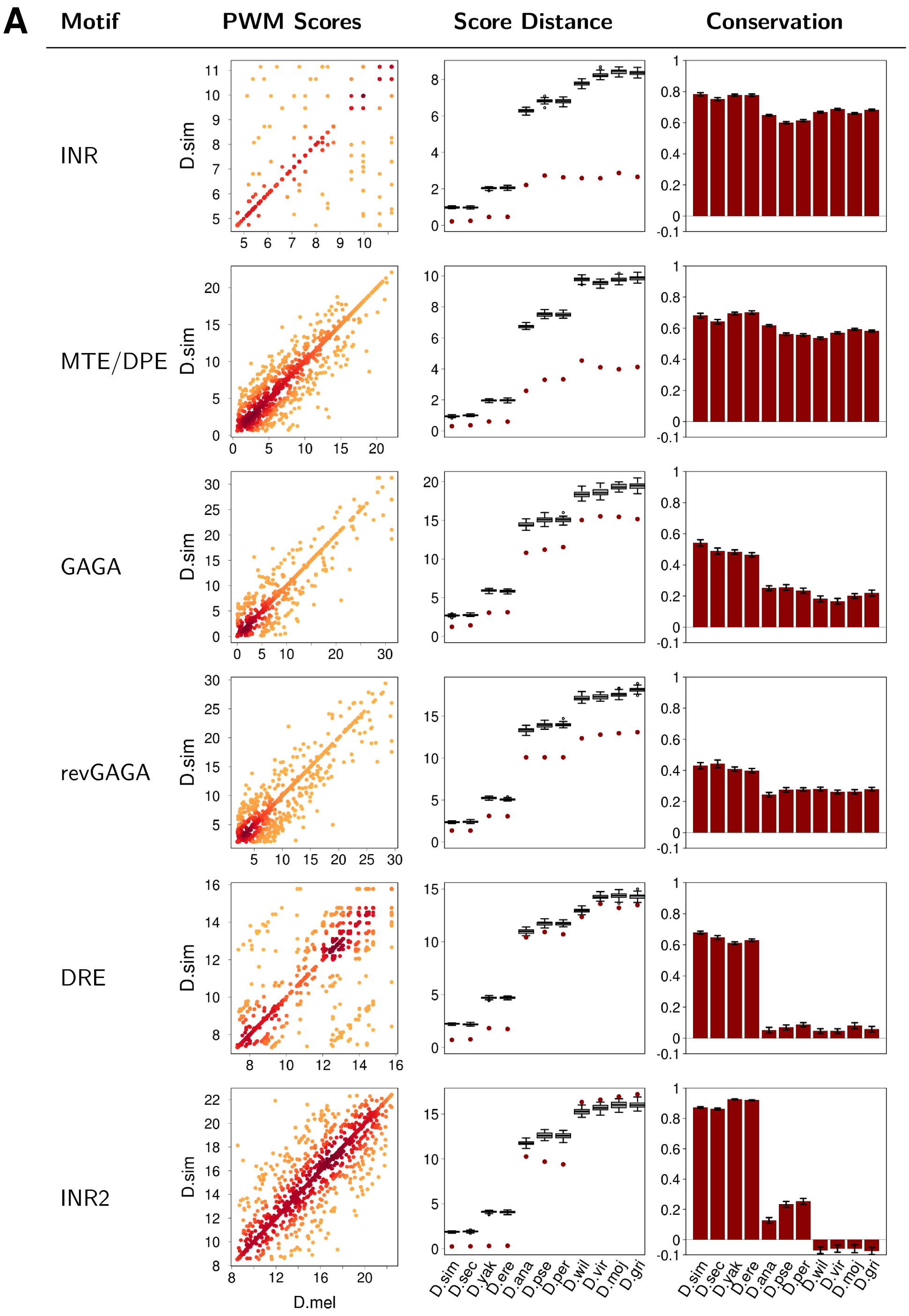
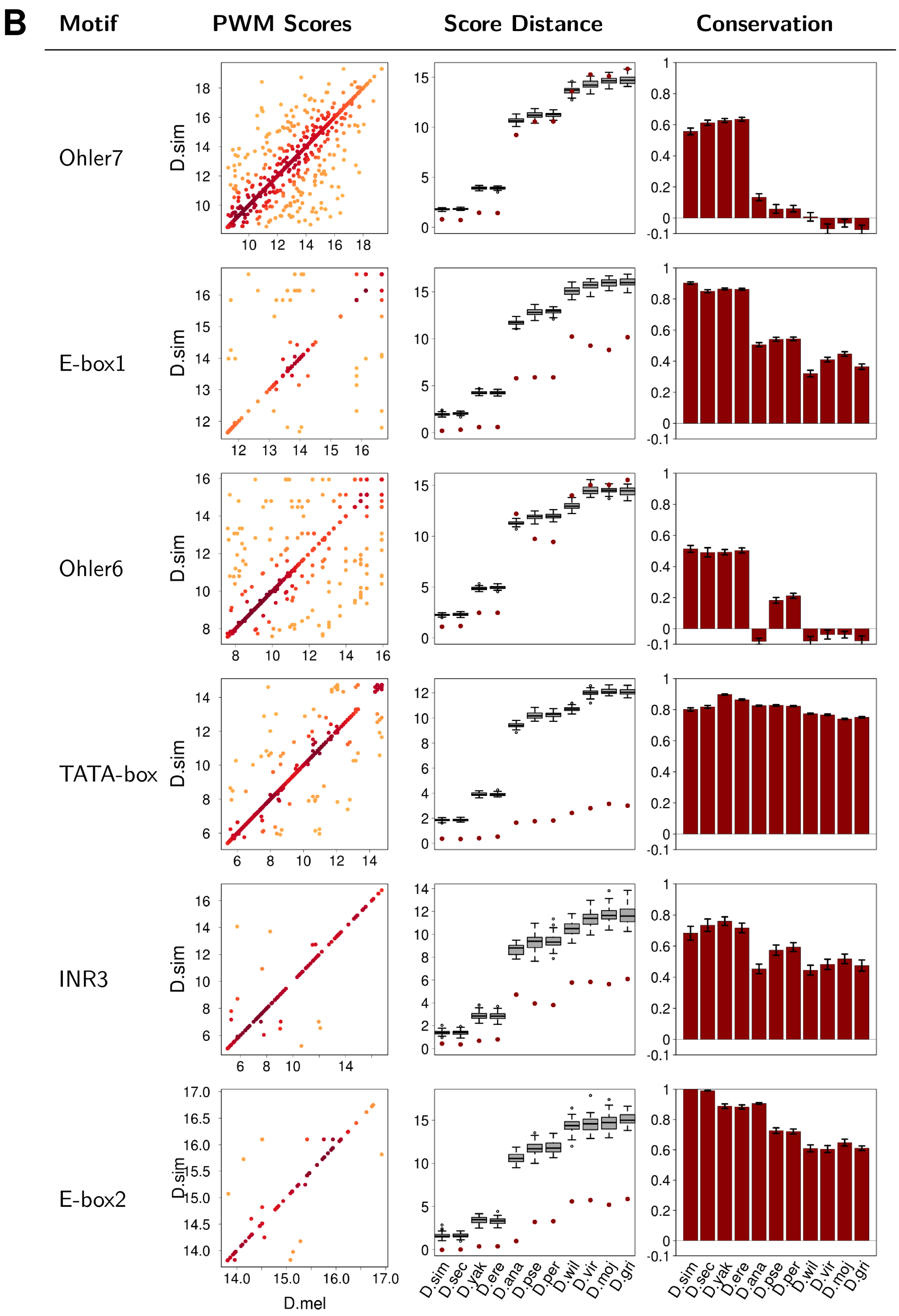
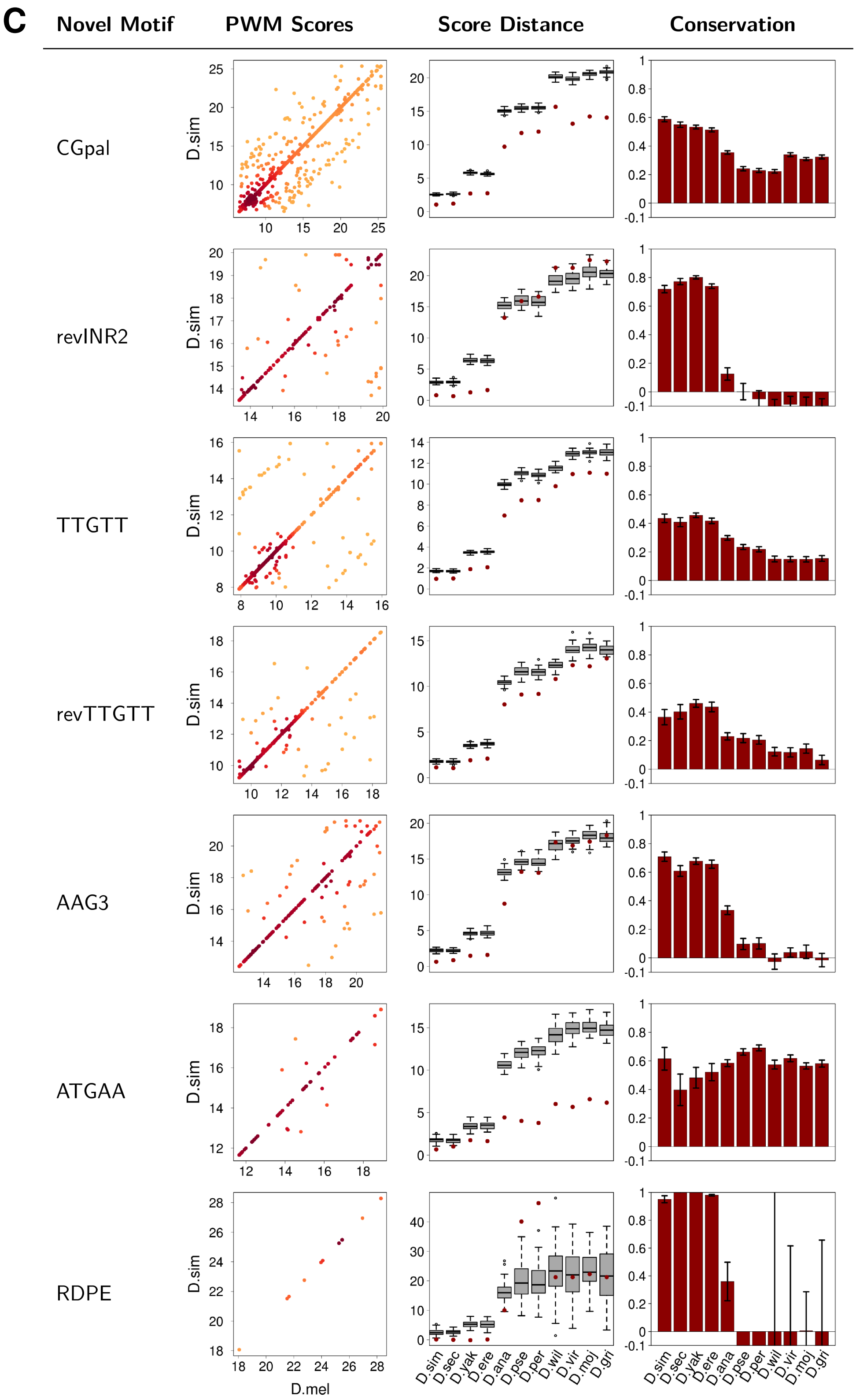


Figure B.1.: Distribution of shape index (SI) scores (Hoskins et al., 2011) over all genes. The vertical line corresponds to the threshold chosen to separate broad peak (BP) promoters from narrow peak (NP) promoters.

B.2. Motif Conservation




B
Motif
PWM Scores
Score Distance
Conservation
Ohler7
E-box1
Ohler6
TATA-box
INR3
E-box2
D.sim
D.mel
D.sim
D.sec
D.yak
D.ere
D.ana
D.pse
D.per
D.wil
D.vir
D.moj
D.gri


C
Novel Motif
PWM Scores
Score Distance
Conservation
CGpal
revINR2
TTGTT
revTTGTT
AAG3
ATGAA
RDPE
D.sim
D.mel
D.sim
D.sec
D.yak
D.ere
D.ana
D.pse
D.per
D.wil
D.vir
D.moj
D.gri
1
0.8
0.6
0.4
0.2
0
-0.1

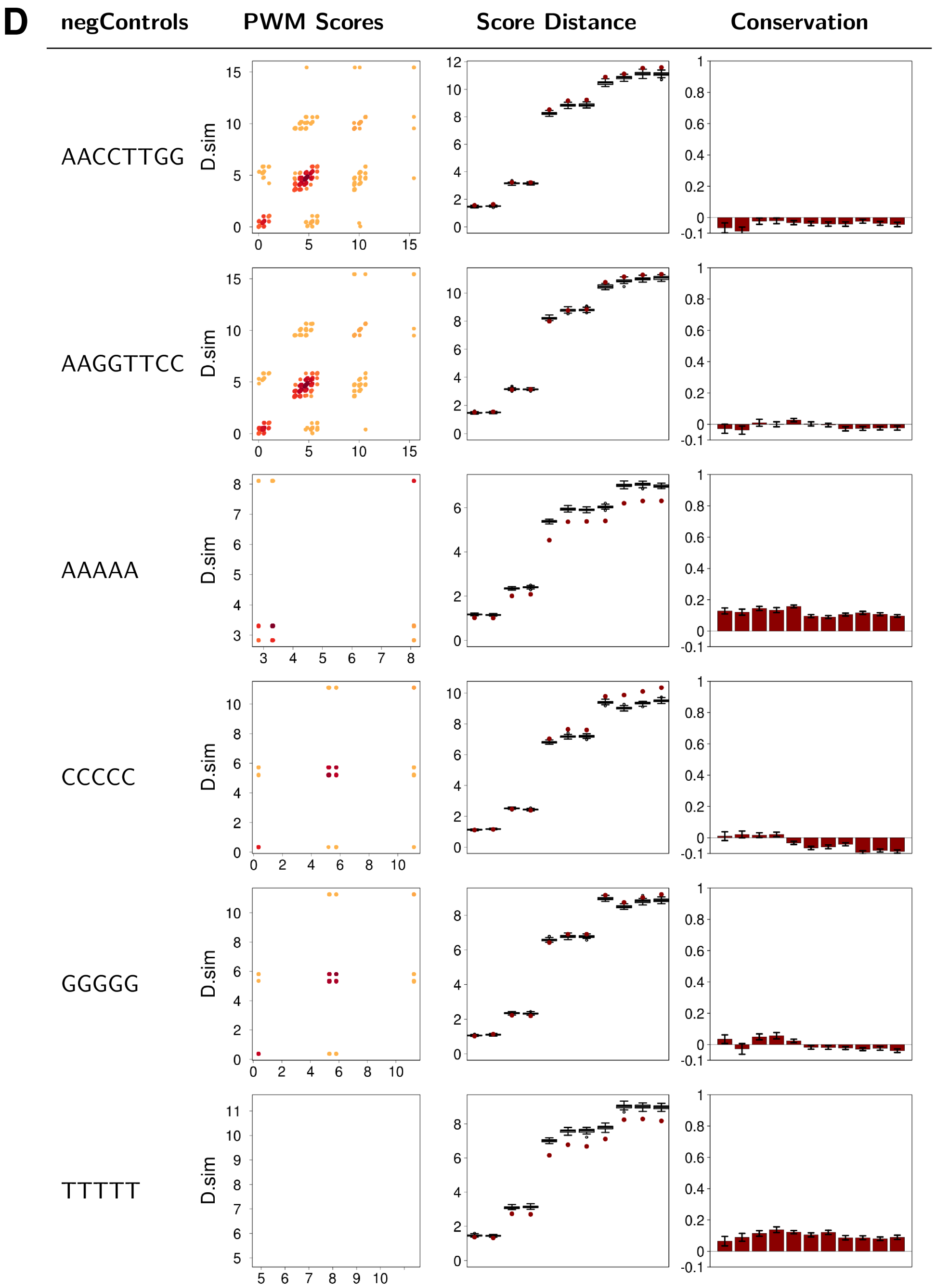


Figure B.2.: Conservation of predicted core promoter motifs and six negative controls. Column "PWM Scores" depicts the scores of the motif PWM for each site in *D. melanogaster* and the aligned site in *D. simulans*. The average PWM score distance between *D. melanogaster* and related species is shown in Column "Score Distance" as a red circle. To calculate the average PWM score we use only binding sites with a PWM score above the minimal score threshold and within the enriched region. Related species are ordered by evolutionary distance. The boxes correspond to the expected score distance calculated on aligned binding sites from sampled sequences. Column "Conservation" depicts a scaled measure of the sampled and biological score distances, giving a one for perfect conservation of the PWM score and a zero for conservation as expected from background. Error bars indicate the standard deviation over all sampled conservation scores (section 6.4.3). (A, B) known motifs, (C) newly identified motifs, (D) negative controls.

B.3. Core Promoter Elements vs. Developmental Stages

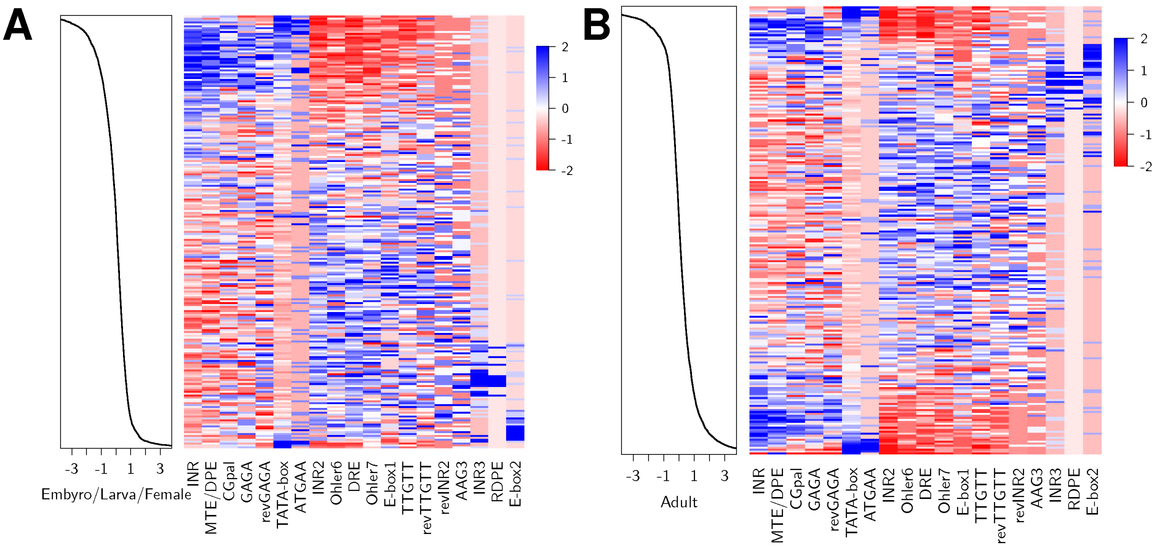


Figure B.3.: Fly core promoter motifs are correlated to distinct gene properties. Genes are sorted by the enrichment within a developmental stage (Graveley et al., 2011) and the frequency of the core promoter motifs on the x-axis within bins of 50 genes is indicated from red (depleted) to blue (enriched) by Z-scores. (A) Enrichment within embryo / larva / female, (B) Enrichment within adult.

B.4. Class 4 Core Promoter Elements

GO ID	Term	Counts	Category	P-value
GO:0022626	cytosolic ribosome	65/100	CC	1×10^{-96}
GO:0044445	cytosolic part	65/100	CC	3×10^{-83}
GO:0003735	structural constituent of ribosome	65/122	MF	6×10^{-77}
GO:0033279	ribosomal subunit	65/100	CC	5×10^{-74}
GO:0005840	ribosome	65/100	CC	1×10^{-69}
GO:0022625	cytosolic large ribosomal subunit	41/100	CC	4×10^{-59}
GO:0006412	translation	75/120	BP	1×10^{-58}
GO:0030529	ribonucleoprotein complex	68/100	CC	5×10^{-52}
GO:0005829	cytosol	71/100	CC	6×10^{-52}
GO:0000022	mitotic spindle elongation	39/120	BP	1×10^{-50}
GO:0051231	spindle elongation	39/120	BP	2×10^{-50}
GO:0005198	structural molecule activity	65/122	MF	4×10^{-48}
GO:0015934	large ribosomal subunit	41/100	CC	8×10^{-44}
GO:0043232	intracellular non-membrane-bounded organelle	78/100	CC	2×10^{-36}
GO:0043228	non-membrane-bounded organelle	78/100	CC	2×10^{-36}
GO:0007052	mitotic spindle organization	39/120	BP	4×10^{-33}
GO:0022627	cytosolic small ribosomal subunit	24/100	CC	1×10^{-30}
GO:0007051	spindle organization	39/120	BP	1×10^{-30}
GO:0000226	microtubule cytoskeleton organization	39/120	BP	7×10^{-26}
GO:0000278	mitotic cell cycle	40/120	BP	5×10^{-24}

Table B.1.: Top 20 GO terms from GO analysis of Class 4 genes. The analysis was done with DAVID using the GO FAT category (Huang et al., 2009). Column "Counts" gives the number of genes that fit to the respective GO term vs. the number of genes that can be mapped to the GO category in column "Category" (CC: cellular component, MF: molecular function, BP: biological process).

FlyBase ID	**Gene Description**
FBgn0064225	ribosomal protein L5
FBgn0014026	ribosomal protein L7A
FBgn0261602	ribosomal protein L8
FBgn0015756	ribosomal protein L9
FBgn0013325	ribosomal protein L11
FBgn0029897	ribosomal protein L17
FBgn0035753	ribosomal protein L18
FBgn0010078	ribosomal protein L23
FBgn0002626	ribosomal protein L32
FBgn0037328	ribosomal protein L35A
FBgn0261608	ribosomal protein L37A
FBgn0261593	ribosomal protein S10b
FBgn0005533	ribosomal protein S17
FBgn0010411	ribosomal protein S18
FBgn0029176	elongation factor 1-gamma

Table B.2.: Genes containing the newly identified RDPE (ribosomal downstream promoter element. 14 of 15 genes are described as ribosomal, all of them contain additionally the INR3 element within the same promoter.

Bibliography

T. Abeel, Y. Saeys, E. Bonnet, P. Rouzé, and Y. V. de Peer. Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Res* **2008**;*18*(2):310–323.

B. Amati, S. R. Frank, D. Donjerkovic, and S. Taubert. Function of the c-Myc oncoprotein in chromatin remodeling and transcription. *Biochim Biophys Acta* **2001**;*1471*(3):M135–M145.

W. F. Anderson, D. H. Ohlendorf, Y. Takeda, and B. W. Matthews. Structure of the cro repressor from bacteriophage lambda and its interaction with DNA. *Nature* **1981**;*290*(5809):754–758.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **2000**;*25*(1):25–29.

M. M. Babu and S. A. Teichmann. Evolution of transcription factors and the gene regulatory network in Escherichia coli. *Nucleic Acids Res* **2003**;*31*(4):1234–1244.

G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C.-F. Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes, and M. L. Bulyk. Diversity and complexity in DNA recognition by transcription factors. *Science* **2009**;*324*(5935):1720–1723.

G. Badis, E. T. Chan, H. van Bakel, L. Pena-Castillo, D. Tillo, K. Tsui, C. D. Carlson, A. J. Gossett, M. J. Hasinoff, C. L. Warren, M. Gebbia, S. Talukder, A. Yang, S. Mnaimneh, D. Terterov, D. Coburn, A. L. Yeo, Z. X. Yeo, N. D. Clarke, J. D. Lieb, A. Z. Ansari, C. Nislow, and T. R. Hughes. A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol Cell* **2008**;*32*(6):878–887.

T. L. Bailey. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* **2011**; *27*(12):1653–1659.

T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **2009**; *37*(Web Server issue):W202–W208.

T. L. Bailey, M. Bodén, T. Whitington, and P. Machanick. The value of position-specific priors in motif discovery using MEME. *BMC Bioinformatics* **2010**;*11*:179.

T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **1994**;*2*:28–36.

T. L. Bailey and M. Gribskov. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **1998**;*14*(1):48–54.

T. L. Bailey, N. Williams, C. Misleh, and W. W. Li. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res* **2006**;*34*(Web Server issue):W369–W373.

P. V. Benos, M. L. Bulyk, and G. D. Stormo. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res* **2002**;*30*(20):4442–4451.

D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, J. M. Boutell, J. Bryant, R. J. Carter, R. K. Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, N. A. Gormley, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. J. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. D. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, P. A. Baybayan, V. A. Benoit, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, J. A. Bridgham, R. C. Brown, A. A. Brown, D. H. Buermann, A. A. Bundu, J. C. Burrows, N. P. Carter, N. Castillo, M. C. E. Catenazzi, S. Chang, R. N. Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. F. Fajardo, W. S. Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, P. A. Granieri, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, T. A. H. Jones, G.-D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, M. A. Laurent, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, J. A. Loch, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. L. Ng, S. M. Novo, M. J. O'Neill, M. A. Osborne, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. Pike, A. C. Pike, D. C. Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Raczy, V. H. Rae, S. R. Rawlings, A. C. Rodriguez, P. M. Roe, J. Rogers, M. C. R. Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, M. A. Smith, J. E. S. Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, and A. J. Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **2008**; *456*(7218):53–59.

M. F. Berger and M. L. Bulyk. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat Protoc* **2009**; *4*(3):393–411.

M. Blanchette, W. J. Kent, C. Riemer, L. Elnitski, A. F. A. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. D. Green, D. Haussler, and W. Miller. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **2004**;*14*(4):708–715.

A. R. Borneman, T. A. Gianoulis, Z. D. Zhang, H. Yu, J. Rozowsky, M. R. Seringhaus, L. Y. Wang, M. Gerstein, and M. Snyder. Divergence of transcription factor binding sites across related yeast species. *Science* **2007**;*317*(5839):815–819.

R. G. Brennan and B. W. Matthews. The helix-turn-helix DNA binding motif. *J Biol Chem* **1989**; *264*(4):1903–1906.

J. C. Bryne, E. Valen, M.-H. E. Tang, T. Marstrand, O. Winther, I. da Piedade, A. Krogh, B. Lenhard, and A. Sandelin. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res* **2008**;*36*(Database issue):D102–D106.

M. L. Bulyk, X. Huang, Y. Choo, and G. M. Church. Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc Natl Acad Sci U S A* **2001**;*98*(13):7158–7163.

T. W. Burke and J. T. Kadonaga. Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev* **1996**;*10*(6):711–724.

J. M. Carlson, A. Chakravarty, C. E. DeZiel, and R. H. Gross. SCOPE: a web server for practical de novo motif discovery. *Nucleic Acids Res* **2007**;*35*(Web Server issue):W259–W264.

S. B. Carroll. Homeotic genes and the evolution of arthropods and chordates. *Nature* **1995**; *376*(6540):479–485.

Y. Choo and A. Klug. Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. *Proc Natl Acad Sci U S A* **1994**;*91*(23):11163–11167.

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

N. Coré, B. Charroux, A. McCormick, C. Vola, L. Fasano, M. P. Scott, and S. Kerridge. Transcriptional regulation of the Drosophila homeotic gene teashirt by the homeodomain protein Fushi tarazu. *Mech Dev* **1997**;*68*(1-2):157–172.

B. Deplancke, A. Mukhopadhyay, W. Ao, A. M. Elewa, C. A. Grove, N. J. Martinez, R. Sequerra, L. Doucette-Stamm, J. S. Reece-Hoyes, I. A. Hope, H. A. Tissenbaum, S. E. Mango, and A. J. M. Walhout. A gene-centered C. elegans protein-DNA interaction network. *Cell* **2006**;*125*(6):1193–1205.

R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. Biological sequence analysis. eleventh edition, **2006**.

E. Eden, D. Lipson, S. Yogev, and Z. Yakhini. Discovering motifs in ranked lists of DNA sequences. *PLoS Comput Biol* **2007**;*3*(3):e39.

P. C. FitzGerald, D. Sturgill, A. Shyakhtenko, B. Oliver, and C. Vinson. Comparative genomics of Drosophila and human core promoters. *Genome Biol* **2006**;*7*(7):R53.

M. C. Frith, N. F. W. Saunders, B. Kobe, and T. L. Bailey. Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Comput Biol* **2008**;*4*(4):e1000071.

M. Fujioka, Y. Emi-Sarker, G. L. Yusibova, T. Goto, and J. B. Jaynes. Analysis of an even-skipped rescue transgene reveals both composite and discrete neuronal and early blastoderm enhancers, and multi-stripe positioning by gap gene repressor gradients. *Development* **1999**;*126*(11):2527–2538.

D. L. Fulton, S. Sundararajan, G. Badis, T. R. Hughes, W. W. Wasserman, J. C. Roach, and R. Sladek. TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol* **2009**;*10*(3):R29.

D. J. Galas and A. Schmitz. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res* **1978**;*5*(9):3157–3170.

S. M. Gallo, D. T. Gerrard, D. Miner, M. Simich, B. D. Soye, C. M. Bergman, and M. S. Halfon. REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. *Nucleic Acids Res* **2011**;*39*(Database issue):D118–D123.

Q. Gao and R. Finkelstein. Targeting gene expression to the head: the Drosophila orthodenticle gene is a direct target of the Bicoid morphogen. *Development* **1998**;*125*(21):4185–4193.

C. W. Garvie and C. Wolberger. Recognition of specific DNA sequences. *Mol Cell* **2001**;*8*(5):937–946.

W. J. Gehring, M. Affolter, and T. Bürglin. Homeodomain proteins. *Annu Rev Biochem* **1994**a; *63*:487–526.

W. J. Gehring, Y. Q. Qian, M. Billeter, K. Furukubo-Tokunaga, A. F. Schier, D. Resendez-Perez,

M. Affolter, G. Otting, and K. Wüthrich. Homeodomain-DNA recognition. *Cell* **1994**b;*78*(2):211–223.

A. E. Gelfand and A. F. M. Smith. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association* **1990**;*85*(410):pp. 398–409.

S. Georgiev, A. P. Boyle, K. Jayasurya, X. Ding, S. Mukherjee, and U. Ohler. Evidence-ranked motif identification. *Genome Biol* **2010**;*11*(2):R19.

N. I. Gershenzon and I. P. Ioshikhes. Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. *Bioinformatics* **2005**;*21*(8):1295–1300.

N. I. Gershenzon, E. N. Trifonov, and I. P. Ioshikhes. The features of Drosophila core promoters revealed by statistical analysis. *BMC Genomics* **2006**;*7*:161.

M. L. Goldberg. Ph.D. thesis, Stanford University, **1979**.

R. Gordân, L. Narlikar, and A. J. Hartemink. Finding regulatory DNA motifs using alignment-free evolutionary conservation information. *Nucleic Acids Res* **2010**;*38*(6):e90.

B. R. Graveley, A. N. Brooks, J. W. Carlson, M. O. Duff, J. M. Landolin, L. Yang, C. G. Artieri, M. J. van Baren, N. Boley, B. W. Booth, J. B. Brown, L. Cherbas, C. A. Davis, A. Dobin, R. Li, W. Lin, J. H. Malone, N. R. Mattiuzzo, D. Miller, D. Sturgill, B. B. Tuch, C. Zaleski, D. Zhang, M. Blanchette, S. Dudoit, B. Eads, R. E. Green, A. Hammonds, L. Jiang, P. Kapranov, L. Langton, N. Perrimon, J. E. Sandler, K. H. Wan, A. Willingham, Y. Zhang, Y. Zou, J. Andrews, P. J. Bickel, S. E. Brenner, M. R. Brent, P. Cherbas, T. R. Gingeras, R. A. Hoskins, T. C. Kaufman, B. Oliver, and S. E. Celniker. The developmental transcriptome of Drosophila melanogaster. *Nature* **2011**; *471*(7339):473–479.

S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **2006**;*34*(Database issue):D140–D144.

R. Grosschedl and M. L. Birnstiel. Identification of regulatory sequences in the prelude sequences of an H2A histone gene by the study of specific deletion mutants in vivo. *Proc Natl Acad Sci U S A* **1980**;*77*(3):1432–1436.

D. GuhaThakurta. Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic Acids Res* **2006**;*34*(12):3585–3598.

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

B. Hartmann, H. Reichert, and U. Walldorf. Interaction of gap genes in the Drosophila head: tailless regulates expression of empty spiracles in early embryonic patterning and brain development. *Mech Dev* **2001**;*109*(2):161–172.

H. Hartmann, E. Guthöhrlein, M. Siebert, S. Luehr, and J. Söding. P-value based regulatory motif dicovery using positional weight matrices, **2012**a. Genome Research, under Review.

H. Hartmann, M. Heron, A. Kiesel, L. Utz, C. Gugenmus, and J. Söding. Drosophila core promoters fall into four classes distinguished by their motif architecture, degree of regulation, and maximum transcription rate, **2012**b. Genome Biology, in Preparation.

X. He, M. A. H. Samee, C. Blatti, and S. Sinha. Thermodynamics-based models of transcriptional regulation by enhancers: the roles of synergistic activation, cooperative binding and short-range repression. *PLoS Comput Biol* **2010**;*6*(9).

R. S. Hegde, S. R. Grossman, L. A. Laimins, and P. B. Sigler. Crystal structure at 1.7 A of the bovine papillomavirus-1 E2 DNA-binding domain bound to its DNA target. *Nature* **1992**; *359*(6395):505–512.

L. M. Hellman and M. G. Fried. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat Protoc* **2007**;*2*(8):1849–1861.

D. A. Hendrix, J.-W. Hong, J. Zeitlinger, D. S. Rokhsar, and M. S. Levine. Promoter elements associated with RNA Pol II stalling in the Drosophila embryo. *Proc Natl Acad Sci U S A* **2008**; *105*(22):7762–7767.

K. Hens, J.-D. Feuz, A. Isakova, A. Iagovitina, A. Massouras, J. Bryois, P. Callaerts, S. E. Celniker, and B. Deplancke. Automated protein-DNA interaction screening of Drosophila regulatory elements. *Nat Methods* **2011**;*8*(12):1065–1070.

W. Herr, R. A. Sturm, R. G. Clerc, L. M. Corcoran, D. Baltimore, P. A. Sharp, H. A. Ingraham, M. G. Rosenfeld, M. Finney, and G. Ruvkun. The POU domain: a large conserved region in the mammalian pit-1, oct-1, oct-2, and Caenorhabditis elegans unc-86 gene products. *Genes Dev* **1988**;*2*(12A):1513–1516.

M. Hoch, E. Seifert, and H. Jäckle. Gene expression mediated by cis-acting sequences of the Krüppel gene in response to the Drosophila morphogens bicoid and hunchback. *EMBO J* **1991**; *10*(8):2267–2278.

A. Hochheimer and R. Tjian. Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression. *Genes Dev* **2003**;*17*(11):1309–1320.

A. Hochheimer, S. Zhou, S. Zheng, M. C. Holmes, and R. Tjian. TRF2 associates with DREF and directs promoter-selective gene expression in Drosophila. *Nature* **2002**;*420*(6914):439–445.

D. S. F. Homsi, V. Gupta, and G. D. Stormo. Modeling the quantitative specificity of DNA-binding proteins from example binding sites. *PLoS One* **2009**;*4*(8):e6736.

R. A. Hoskins, J. M. Landolin, J. B. Brown, J. E. Sandler, H. Takahashi, T. Lassmann, C. Yu, B. W. Booth, D. Zhang, K. H. Wan, L. Yang, N. Boley, J. Andrews, T. C. Kaufman, B. R. Graveley, P. J. Bickel, P. Carninci, J. W. Carlson, and S. E. Celniker. Genome-wide analysis of promoter architecture in Drosophila melanogaster. *Genome Res* **2011**;*21*(2):182–192.

K. R. Howard and G. Struhl. Decoding positional information: regulation of the pair-rule gene hairy. *Development* **1990**;*110*(4):1223–1231.

D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **2009**;*4*(1):44–57.

T. Hughes. A Handbook of Transcription Factors. Subcellular Biochemistry. Springer, **2011**.

T. Hulf, P. Bellosta, M. Furrer, D. Steiger, D. Svensson, A. Barbour, and P. Gallant. Whole-genome analysis reveals a strong positional bias of conserved dMyc-dependent E-boxes. *Mol Cell Biol* **2005**;*25*(9):3401–3410.

T. Häder, D. Wainwright, T. Shandala, R. Saint, H. Taubert, G. Brönner, and H. Jäckle. Receptor tyrosine kinase signaling regulates different modes of Groucho-dependent control of Dorsal. *Curr Biol* **2000**;*10*(1):51–54.

J. Jaeger, S. Surkova, M. Blagov, H. Janssens, D. Kosman, K. N. Kozlov, Manu, E. Myasnikova, C. E. Vanario-Alonso, M. Samsonova, D. H. Sharp, and J. Reinitz. Dynamic control of positional information in the early Drosophila embryo. *Nature* **2004**;*430*(6997):368–371.

A. Jolma, T. Kivioja, J. Toivonen, L. Cheng, G. Wei, M. Enge, M. Taipale, J. M. Vaquerizas, J. Yan, M. J. Sillanpää, M. Bonke, K. Palin, S. Talukder, T. R. Hughes, N. M. Luscombe, E. Ukkonen, and J. Taipale. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res* **2010**;*20*(6):861–873.

R. Joshi, J. M. Passner, R. Rohs, R. Jain, A. Sosinsky, M. A. Crickmore, V. Jacob, A. K. Aggarwal, B. Honig, and R. S. Mann. Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell* **2007**;*131*(3):530–543.

T. Juven-Gershon, S. Cheng, and J. T. Kadonaga. Rational design of a super core promoter that enhances gene expression. *Nat Methods* **2006**;*3*(11):917–922.

T. Juven-Gershon and J. T. Kadonaga. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev Biol* **2010**;*339*(2):225–229.

T. Kaplan, N. Friedman, and H. Margalit. Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput Biol* **2005**;*1*(1):e1.

J. Kim, X. He, and S. Sinha. Evolution of regulatory sequences in 12 Drosophila species. *PLoS Genet* **2009**;*5*(1):e1000330.

M. Klingler, J. Soong, B. Butler, and J. P. Gergen. Disperse versus compact elements for the regulation of runt stripes in Drosophila. *Dev Biol* **1996**;*177*(1):73–84.

M. Kmita and D. Duboule. Organizing axes in time and space; 25 years of colinear tinkering. *Science* **2003**;*301*(5631):331–333.

D. Kostrewa, M. E. Zeller, K.-J. Armache, M. Seizl, K. Leike, M. Thomm, and P. Cramer. RNA polymerase II-TFIIB structure and mechanism of transcription initiation. *Nature* **2009**;*462*(7271):323–330.

R. P. Kühnlein, G. Brönner, H. Taubert, and R. Schuh. Regulation of Drosophila spalt gene expression. *Mech Dev* **1997**;*66*(1-2):107–118.

D. S. Latchman. Transcription factors: an overview. *Int J Biochem Cell Biol* **1997**;*29*(12):1305–1312.

C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **1993**;*262*(5131):208–214.

X.-Y. Li, S. Thomas, P. J. Sabo, M. B. Eisen, J. A. Stamatoyannopoulos, and M. D. BigginNoyes. The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. *Genome Biol* **2011**;*12*(4):R34.

G. J. Liaw and J. A. Lengyel. Control of tailless expression by bicoid, dorsal and synergistically interacting terminal system regulatory elements. *Mech Dev* **1993**;*40*(1-2):47–61.

C. Y. Lim, B. Santoso, T. Boulay, E. Dong, U. Ohler, and J. T. Kadonaga. The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes Dev* **2004**;*18*(13):1606–1617.

C. Linhart, Y. Halperin, and R. Shamir. Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. *Genome Res* **2008**;*18*(7):1180–1189.

X. Liu, D. M. Noll, J. D. Lieb, and N. D. Clarke. DIP-chip: rapid and accurate determination of DNA-binding specificity. *Genome Res* **2005**;*15*(3):421–427.

S. Luehr, H. Hartmann, and J. Söding. The XXmotif web server for eXhaustive, weight matriX-based motif discovery in nucleotide sequences, **2012**. NAR, under Review.

K. D. MacIsaac and E. Fraenkel. Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Comput Biol* **2006**;*2*(4):e36.

E. R. Mardis. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* **2008**; *9*:387–402.

R. Marmorstein, M. Carey, M. Ptashne, and S. C. Harrison. DNA recognition by GAL4: structure of a protein-DNA complex. *Nature* **1992**;*356*(6368):408–414.

P. McQuilton, S. E. S. Pierre, J. Thurmond, and F. Consortium. FlyBase 101–the basics of navigating FlyBase. *Nucleic Acids Res* **2012**;*40*(Database issue):D706–D714.

J. Miller, A. D. McLachlan, and A. Klug. Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *EMBO J* **1985**;*4*(6):1609–1614.

G. Mittler, F. Butter, and M. Mann. A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. *Genome Res* **2009**;*19*(2):284–293.

V. Mustonen and M. Lässig. Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. *Proc Natl Acad Sci U S A* **2005**;*102*(44):15936–15941.

L. Narlikar, R. Gordân, and A. J. Hartemink. A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Comput Biol* **2007**;*3*(11):e215.

L. Narlikar, R. Gordân, U. Ohler, and A. J. Hartemink. Informative priors based on transcription factor structural class improve de novo motif discovery. *Bioinformatics* **2006**;*22*(14):e384–e392.

H. C. Nelson, J. T. Finch, B. F. Luisi, and A. Klug. The structure of an oligo(dA).oligo(dT) tract and its biological implications. *Nature* **1987**;*330*(6145):221–226.

D. E. Newburger and M. L. Bulyk. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res* **2009**;*37*(Database issue):D77–D82.

T. Ni, D. L. Corcoran, E. A. Rach, S. Song, E. P. Spana, Y. Gao, U. Ohler, and J. Zhu. A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nat Methods* **2010**; *7*(7):521–527.

D. B. Nikolov, H. Chen, E. D. Halay, A. A. Usheva, K. Hisatake, D. K. Lee, R. G. Roeder, and S. K. Burley. Crystal structure of a TFIIB-TBP-TATA-element ternary complex. *Nature* **1995**; *377*(6545):119–128.

M. B. Noyes, X. Meng, A. Wakabayashi, S. Sinha, M. H. Brodsky, and S. A. Wolfe. A systematic characterization of factors that regulate Drosophila segmentation via a bacterial one-hybrid system. *Nucleic Acids Res* **2008**;*36*(8):2547–2560.

R. Nutiu, R. C. Friedman, S. Luo, I. Khrebtukova, D. Silva, R. Li, L. Zhang, G. P. Schroth, and C. B. Burge. Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. *Nat Biotechnol* **2011**;*29*(7):659–664.

D. T. Odom, R. D. Dowell, E. S. Jacobsen, W. Gordon, T. W. Danford, K. D. MacIsaac, P. A. Rolfe, C. M. Conboy, D. K. Gifford, and E. Fraenkel. Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* **2007**;*39*(6):730–732.

U. Ohler. Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction. *Nucleic Acids Res* **2006**;*34*(20):5943–5950.

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

U. Ohler and D. A. Wassarman. Promoting developmental transcription. *Development* **2010**; *137*(1):15–26.

S. K. Oster, C. S. W. Ho, E. L. Soucie, and L. Z. Penn. The myc oncogene: MarvelouslY Complex. *Adv Cancer Res* **2002**;*84*:81–154.

M. J. Pankratz, M. Busch, M. Hoch, E. Seifert, and H. Jäckle. Spatial control of the gap gene knirps in the Drosophila embryo by posterior morphogen system. *Science* **1992**;*255*(5047):986–989.

T. J. Parry, J. W. M. Theisen, J.-Y. Hsu, Y.-L. Wang, D. L. Corcoran, M. Eustice, U. Ohler, and J. T. Kadonaga. The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes Dev* **2010**;*24*(18):2013–2018.

G. Pavesi and G. Pesole. Using Weeder for the discovery of conserved transcription factor binding sites. *Curr Protoc Bioinformatics* **2006**;*Chapter 2*:Unit 2.11.

M. W. Perry, A. N. Boettiger, and M. Levine. Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. *Proc Natl Acad Sci U S A* **2011**;*108*(33):13570–13575.

R. P. Perry and D. E. Kelley. Evidence for specific association of protein with newly formed ribosomal subunits. *Biochem Biophys Res Commun* **1966**;*24*(3):459–465.

A. V. Persikov, R. Osada, and M. Singh. Predicting DNA recognition by Cys2His2 zinc finger proteins. *Bioinformatics* **2009**;*25*(1):22–29.

W. W. Quitschke, M. J. Taheny, L. J. Fochtmann, and A. A. Vostrov. Differential effect of zinc finger deletions on the binding of CTCF to the promoter of the amyloid precursor protein gene. *Nucleic Acids Res* **2000**;*28*(17):3370–3378.

E. A. Rach, H.-Y. Yuan, W. H. Majoros, P. Tomancak, and U. Ohler. Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the Drosophila genome. *Genome Biol* **2009**;*10*(7):R73.

B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of DNA binding proteins. *Science* **2000**;*290*(5500):2306–2309.

H. S. Rhee and B. F. Pugh. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. *Cell* **2011**;*147*(6):1408–1419.

R. Rohs, X. Jin, S. M. West, R. Joshi, B. Honig, and R. S. Mann. Origins of specificity in protein-DNA recognition. *Annu Rev Biochem* **2010**;*79*:233–269.

R. Rohs, S. M. West, A. Sosinsky, P. Liu, R. S. Mann, and B. Honig. The role of DNA shape in protein-DNA recognition. *Nature* **2009**;*461*(7268):1248–1253.

K. A. Romer, G.-R. Kayombya, and E. Fraenkel. WebMOTIFS: automated discovery, filtering and scoring of DNA sequence motifs using multiple programs and Bayesian approaches. *Nucleic Acids*

Res **2007**;*35*(Web Server issue):W217–W220.

E. Roulet, S. Busso, A. A. Camargo, A. J. G. Simpson, N. Mermod, and P. Bucher. High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat Biotechnol* **2002**;*20*(8):831–835.

K. M. Rudolph, G. J. Liaw, A. Daniel, P. Green, A. J. Courey, V. Hartenstein, and J. A. Lengyel. Complex regulatory region mediating tailless expression in early embryonic patterning and brain development. *Development* **1997**;*124*(21):4297–4308.

S. L. Salzberg, A. L. Delcher, S. Kasif, and O. White. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res* **1998**;*26*(2):544–548.

A. Sandelin, P. Carninci, B. Lenhard, J. Ponjavic, Y. Hayashizaki, and D. A. Hume. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet* **2007**; *8*(6):424–436.

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

C. D. Schmid, R. Perier, V. Praz, and P. Bucher. EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucleic Acids Res* **2006**;*34*(Database issue):D82–D85.

T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* **1990**;*18*(20):6097–6100.

M. D. Schroeder, C. Greer, and U. Gaul. How to make stripes: deciphering the transition from non-periodic to periodic patterns in Drosophila segmentation. *Development* **2011**;*138*(14):3067–3078.

M. D. Schroeder, M. Pearce, J. Fak, H. Fan, U. Unnerstall, E. Emberly, N. Rajewsky, E. D. Siggia, and U. Gaul. Transcriptional control in the segmentation gene network of Drosophila. *PLoS Biol* **2004**;*2*(9):E271.

E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature* **2008**;*451*(7178):535–540.

E. Segal and J. Widom. From DNA sequence to transcriptional behaviour: a quantitative approach. *Nat Rev Genet* **2009**;*10*(7):443–456.

M. Seizl, H. Hartmann, F. Hoeg, F. Kurth, D. E. Martin, J. Söding, and P. Cramer. A Conserved GA Element in TATA-Less RNA Polymerase II Promoters. *PLoS One* **2011**;*6*(11):e27595.

D. Sharma, D. Mohanty, and A. Surolia. RegAnalyst: a web interface for the analysis of regulatory motifs, networks and pathways. *Nucleic Acids Res* **2009**;*37*(Web Server issue):W193–W201.

Y. Shibata and G. E. Crawford. Mapping regulatory elements by DNaseI hypersensitivity chip (DNase-Chip). *Methods Mol Biol* **2009**;*556*:177–190.

M. J. Shimell, A. J. Peterson, J. Burr, J. A. Simon, and M. B. O'Connor. Functional analysis of repressor binding sites in the iab-2 regulatory region of the abdominal-A homeotic gene. *Dev Biol* **2000**;*218*(1):38–52.

R. K. Shultzaberger, D. S. Malashock, J. F. Kirsch, and M. B. Eisen. The fitness landscapes of cis-acting binding sites in different promoter and environmental contexts. *PLoS Genet* **2010**; *6*(7):e1001042.

M. Siebert, M. Lidschreiber, H. Hartmann, and J. Söding. A Guideline for ChIP-Chip Data Quality Control and Normalization. *Epigenome Network of Excellence* **2010**;*protocol 47*.

S. T. Smale and D. Baltimore. The "initiator" as a transcription control element. *Cell* **1989**; *57*(1):103–113.

S. T. Smale and J. T. Kadonaga. The RNA polymerase II core promoter. *Annu Rev Biochem* **2003**; *72*:449–479.

S. Small, A. Blair, and M. Levine. Regulation of two pair-rule stripes by a single enhancer in the Drosophila embryo. *Dev Biol* **1996**;*175*(2):314–324.

M. J. Solomon and A. Varshavsky. Formaldehyde-mediated DNA-protein crosslinking: a probe for in vivo chromatin structures. *Proc Natl Acad Sci U S A* **1985**;*82*(19):6470–6474.

D. Stanojevic, S. Small, and M. Levine. Regulation of a segmentation stripe by overlapping activators and repressors in the Drosophila embryo. *Science* **1991**;*254*(5036):1385–1387.

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

M. C. Thomas and C.-M. Chiang. The general transcription machinery and general cofactors. *Crit Rev Biochem Mol Biol* **2006**;*41*(3):105–178.

M. Thomas-Chollier, M. Defrance, A. Medina-Rivera, O. Sand, C. Herrmann, D. Thieffry, and J. van Helden. RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Res* **2011**;*39*(Web Server issue):W86–W91.

M. Thomas-Chollier, O. Sand, J.-V. Turatsinze, R. Janky, M. Defrance, E. Vervisch, S. Brohée, and J. van Helden. RSAT: regulatory sequence analysis tools. *Nucleic Acids Res* **2008**;*36*(Web Server issue):W119–W127.

M. Tompa, N. Li, T. L. Bailey, G. M. Church, B. D. Moor, E. Eskin, A. V. Favorov, M. C. Frith, Y. Fu, W. J. Kent, V. J. Makeev, A. A. Mironov, W. S. Noble, G. Pavesi, G. Pesole, M. Régnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* **2005**;*23*(1):137–144.

B. van Steensel, J. Delrow, and S. Henikoff. Chromatin profiling using targeted DNA adenine methyltransferase. *Nat Genet* **2001**;*27*(3):304–308.

J. M. Vaquerizas, S. K. Kummerfeld, S. A. Teichmann, and N. M. Luscombe. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet* **2009**;*10*(4):252–263.

G. J. Veenstra and A. P. Wolffe. Gene-selective developmental roles of general transcription factors. *Trends Biochem Sci* **2001**;*26*(11):665–671.

B. J. Venters and B. F. Pugh. A canonical promoter organization of the transcription machinery and its regulators in the Saccharomyces genome. *Genome Res* **2009**;*19*(3):360–371.

T. Wang and G. D. Stormo. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* **2003**;*19*(18):2369–2380.

W. W. Wasserman and A. Sandelin. Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet* **2004**;*5*(4):276–287.

E. A. Wimmer, M. Simpson-Brose, S. M. Cohen, C. Desplan, and H. Jäckle. Trans- and cis-acting requirements for blastodermal expression of the head gap gene buttonhead. *Mech Dev* **1995**; *53*(2):235–245.

E. Wingender, P. Dietze, H. Karas, and R. Knüppel. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* **1996**;*24*(1):238–241.

S. A. Wolfe, L. Nekludova, and C. O. Pabo. DNA recognition by Cys2His2 zinc finger proteins. *Annu Rev Biophys Biomol Struct* **2000**;*29*:183–212.

H. Xi, Y. Yu, Y. Fu, J. Foley, A. Halees, and Z. Weng. Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. *Genome Res* **2007**;*17*(6):798–806.

J. Zeitlinger, A. Stark, M. Kellis, J.-W. Hong, S. Nechaev, K. Adelman, M. Levine, and R. A. Young. RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. *Nat Genet* **2007**;*39*(12):1512–1516.

Z. Zhang and F. S. Dietrich. Mapping of transcription start sites in Saccharomyces cerevisiae using 5' SAGE. *Nucleic Acids Res* **2005**;*33*(9):2838–2851.

X. Zhou and E. K. O'Shea. Integrated approaches reveal determinants of genome-wide binding and function of the transcription factor Pho4. *Mol Cell* **2011**;*42*(6):826–836.

C. Zhu, K. J. R. P. Byers, R. P. McCord, Z. Shi, M. F. Berger, D. E. Newburger, K. Saulrieta, Z. Smith, M. V. Shah, M. Radhakrishnan, A. A. Philippakis, Y. Hu, F. D. Masi, M. Pacek, A. Rolfs, T. Murthy, J. Labaer, and M. L. Bulyk. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res* **2009**;*19*(4):556–566.

R. P. Zinzen, C. Girardot, J. Gagneur, M. Braun, and E. E. M. Furlong. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* **2009**;*462*(7269):65–70.

Holger Hartmann

MSc in Bioinformatics
BSc in Applied Informatics

Personal Details

DATE OF BIRTH	September 27, 1983
PLACE OF BIRTH	74177 Bad Friedrichshall, Germany
NATIONALITY	German

Education

JUN 2008 – NOW	**PhD in Bioinformatics (Dr. rer. nat.)**, *Gene Center LMU Munich*, Prof. Dr. Patrick Cramer and Dr. Johannes Söding. Topic: „Regulatory motif discovery using PWMs and the architecture of core promoters“ Developed a novel bioinformatics method (XXmotif) for de novo motif discovery. First tool that refines motif PWMs by iteratively optimizing the *P*-value.
OCT 2006 – MAY 2008	**MSc in Bioinformatics**, *LMU/TU Munich*. Thesis: „Design and Analysis of Meta-Models for the Classification of Membrane Proteins“. Master grade: 1.1 (A)
OCT 2003 – APR 2006	**BSc in Applied Informatics**, *University of Bayreuth*. Thesis: „Progress measurement of agile software development projects using open-source tools“. Bachelor grade: 1.2 (A)
AUG 1994 – JUN 2003	**Grammar school**, *Friedrich-von-Alberti-Gymnasium, Bad Friedrichshall*. Grade of university-entrance diploma: 1.4 (A)

Professional Experience

FEB 2008 – MAY 2008	**Research assistant**, *Computational Biology, LMU Munich*, Support of research projects with ChIP-chip data analysis.
APR 2007 – MAY 2007	**Research assistant**, *Genome Oriented Bioinformatics, GSF, TU Munich*, Development of a web browser to interactively browse isochore maps (ISOBASE).
MAY 2006 – AUG 2006	**Bioinformatics software developer**, *Proteome Systems, Sydney, Australia*, Responsible for the design, implementation and integration of a biostatistical package with a scientific data management system. Dr. Keith Junius
JAN 2006 – FEB 2006	**Student assistant**, *Databases and Information systems, University of Bayreuth*, Preparation of databases for practical courses.
SEP 2004 – DEC 2004	**Student assistant**, *Parallel and Distributed computing, University of Bayreuth*, Supervision of practical courses and exercises.

Additional Qualifications

LANGUAGES	German (native), English (fluent), Romanian (advanced), French (basic)
PROGRAMMING	C/C++, Perl, Java, Shell, R, JavaScript, HTML, XML, XSLT, SQL, LaTeX (all at least 5 years experience and in constant use)

Publications

2012 **Drosophila core promoters fall into four classes distinguished by their motif architecture, degree of regulation, and maximum transcription rate**,
H. Hartmann*, M. Heron*, A. Kiesel*, L. Utz, C. Gugenmus, J. Söding, in Preparation.

2012 **The XXmotif web server for eXhaustive, weight-matriX-based motif discovery in nucleotide sequences**,
S. Luehr, **H. Hartmann**, J. Söding, under Review.

2012 ***P*-value based regulatory motif discovery using positional weight matrices**,
H. Hartmann, E. W. Guthöhrlein, M. Siebert, S. Luehr, J. Söding, under Review.

2012 **DBIRD protein complex integrates alternative mRNA splicing with RNA polymerase II transcript elongation**,
P. Close, P. East, A. B. Dirac-Svejstrup, **H. Hartmann**, M. Heron, S. Maslen, A. Chariot, J. Söding, M. Skehel, J. Q. Svejstrup, *Nature*, doi:10.1038/nature10925.

2012 **CAMPS 2.0: exploring the sequence and structure space of prokaryotic, eukaryotic and viral membrane proteins**,
S. Neumann, **H. Hartmann**, A. J. Martin-Galiano, A. Fuchs, D. Frishman, *Proteins 80*(3):839–857.

2012 **The MOF-containing NSL complex associates globally with housekeeping genes, but activates only a defined subset**,
C. Feller, M. Prestel, **H. Hartmann**, T. Straub, J. Söding, P. B. Becker, *Nucleic Acids Res 40*(4):1509–1522.

2011 **A conserved GA element in TATA-less RNA polymerase II promoters**,
M. Seizl*, **H. Hartmann***, F. Hoeg*, F. Kurth, D. E. Martin, J. Söding, P. Cramer, *PLoS One 6*(11):e27595.

2010 **A Guideline for ChIP-Chip Data Quality Control and Normalization**,
M. Siebert, M. Lidschreiber, **H. Hartmann**, J. Söding, *Epigenome Network of Excellence*, Protocol 47.

2008 **Genome-associated RNA polymerase II includes the dissociable Rpb4/7 subcomplex**,
A. J. Jasiak, **H. Hartmann**, E. Karakasili, M. Kalocsay, A. Flatley, E. Kremmer, K. Strässer, D, E. Martin, J. Söding, P. Cramer, *J Biol Chem 31*(1):309–18.

Invited Talks & Posters

OCT 2011 **3rd International Gene Center/SFB646 Symposium**, *Munich, Germany*,
Talk: „Regulatory motif discovery using PWMs and the architecture of core promoters“.

JUL 2011 **ISMB & ECCB 2011**, *Vienna, Austria*,
Poster: „*P*-value based motif identification using positional weight matrices“.

MAY 2010 **Gene Center Retreat**, *Wildbad Kreuth, Germany*,
Poster: „Systematic Discovery of Linear Motifs for Protein Interactions“.

SEP 2009 **1st SFB646 PhD/Postdoc Retreat**, *Bad Wiessee, Germany*,
Talk: „Motif discovery in Eukaryotes“.

JUN 2009 **ISMB & ECCB 2009**, *Stockholm, Sweden*,
Poster: „Genome-wide computational analysis of eukaryotic core promoters“.

SEP 2008 **GCB 2008**, *Dresden, Germany*,
Poster: „Organization of the membrane protein fold jungle“.