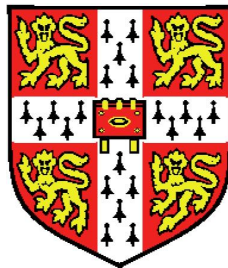


Inference and classification of eukaryotic *cis*-regulatory motifs



Matias Piipari

Wellcome Trust Sanger Institute

University of Cambridge

A thesis submitted for the degree of

Doctor of Philosophy

6th September, 2010

With ❤️ to Kaisa.

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

Acknowledgements

I would like to thank my supervisors Dr Tim Hubbard and Dr Thomas Down for the valuable advice and support I have been given during my project. I am also grateful to members of my PhD committee: Dr Derek Stemple, Dr Alex Bateman and Dr Jurg Bahler. I am deeply honoured to have studied my four years in Cambridge with the generous Wellcome Trust studentship.

My sincere thanks go to the Hubbard research group members Markus Brosch, Mutlu Dogruel, Jenny Mattison, and especially Daniel James who read the preliminary versions of this dissertation. My time at Sanger has been made both happy and productive thanks to many of the other Sanger Institute future doctors. I would especially like to thank my dear friends Alexandra Nica, Leopold Parts, Sergei Manakov, Steve Pettitt and Marija Buljan for the many conversations, lessons and laughs they have let me experience.

I will always be grateful for the love and endless encouragement for intellectual pursuit that I have received from my parents, brother and sister. Finally, I am indebted to the love and friendship of my dear wife Kaisa.

Matias Piipari, September 2010, Cambridge, UK.

Abstract

Regulation of gene expression by networks of sequence specific transcription factors is one of the most important control mechanisms that defines the expression pattern of a genome. Describing transcriptional regulatory networks requires a near complete knowledge of the transcription factors present in the cell, as well the DNA binding sites to which each of the TFs is able to bind. Recent years have witnessed advances in both directions. High coverage transcription factor annotations have become available for many sequenced eukaryotic genomes. Improvements have also been made in profiling DNA specificity motifs for eukaryotic transcription factors, *in vitro* and *in vivo*.

The theme of my work has been the application and development of computational methods for inferring regulatory motifs from promoter sequence, and finding clues to the function of computationally inferred DNA motifs. Functional annotation of inferred motifs led me to conduct a comparative study of the familial relationships between regulatory motifs, the conclusion of which was a probabilistic motif family model I call the ‘metamotif’. The metamotif, I will show, allows improved prediction of the DNA binding domain family for *de novo* inferred motifs, and is an effective way of encoding prior information about known DNA binding domain families to a motif inference algorithm. The use of familial prior information improves the sensitivity to detect regulatory motifs contained in the large promoter sequences that are common to higher eukaryotic genomes. The metamotif guides motif inference towards types of sequence signal that are expected *a priori* to be present in the sequence set of interest, thereby improving and supplementing traditional regulatory motif inference algorithms.

I have also assessed several published *de novo* DNA motif inference algorithms by challenging them to infer a complete set of regulatory motifs from a large series of *Saccharomyces cerevisiae* promoters. This work provides a novel way to assess performance of regulatory motif inference methods, and is made possible by the availability of an experimentally determined regulatory motif dictionary for the *S. cerevisiae* genome. In addition to benchmarking motif inference methods compared to a reference motif set, I make use of many of the rich genomics resources available for study of the budding yeast. These include curated lists of TF target genes based on ChIP-chip and gene expression studies of wild type and knockout yeasts, a close-to-complete list of TF motif from the JASPAR database, and a 7-way sequence conservation score across the genome, as well as sequence variation data from the *Saccharomyces* Genome Resequencing Project.

Development of sensitive regulatory motif inference algorithms continues to be important in gaining understanding of eukaryotic gene regulation by sequence specific transcription factors. In particular I believe that methods that integrate different sources of biological evidence, such as metamotifs, gene expression and ChIP-seq, to sequence motif inference will be highly important to the field.

Contents

Contents	VI
List of Figures	X
1 Introduction	1
1.1 Gene regulation by control of transcription	2
1.1.1 Sequence specific transcription factors	6
1.1.2 Binding specificity of transcription factors	11
1.2 Computational inference of transcription factor binding site motifs	13
1.2.1 The position weight matrix	16
1.3 Computational methodology	18
1.3.1 Hidden Markov Models in motif inference	18
1.3.2 Nested sampling	23
1.3.3 The NestedMICA algorithm	26
1.3.4 Random forest classification	28
1.4 Biological datasets and resources	31
1.4.1 Ensembl	31
1.4.2 Regulatory motif databases	32
1.4.2.1 TRANSFAC	33
1.4.2.2 JASPAR	34
1.4.2.3 UniPROBE	35
1.5 Contributions of this thesis	37
2 Metamotifs - a generative model for building families of nu- cleotide position weight matrices	38

2.1	Background	38
2.1.1	Previous work on motif family models	39
2.2	The metamotif	42
2.2.1	Formulation of the model	45
2.2.2	Visual representation of the model	46
2.2.3	Aligning motifs and estimating metamotifs from a motif multiple alignments	46
2.2.4	Metamotif inference by nested sampling	50
2.2.5	The likelihood function	53
2.2.6	Monte Carlo sampling moves	55
2.2.7	Accounting for incomplete metamotif hits	56
2.3	Evaluating the metamotif nested sampler algorithm	56
2.3.1	A single metamotif	60
2.3.2	Multiple metamotifs	62
2.3.3	Inferring metamotifs from TRANSFAC	63
2.4	Summary	65
3	Metamotifs in motif inference	66
3.1	Previous work on biologically informative motif prior functions . .	67
3.2	Materials & Method	69
3.2.1	The metamotif prior function	69
3.2.2	Measuring motif inference sensitivity with synthetic sequence	71
3.3	Results & Discussion	74
3.3.1	Performance effect of a correct motif family prior function	74
3.3.2	Performance effect of an incorrect motif family prior function	74
3.3.3	Making the metamotif prior available	77
3.3.4	Using the metamotif prior with the NestedMICA algorithm	77
3.3.5	Using the metamotif prior with iMotifs	78
4	Metamotifs in motif classification	81
4.1	Previous work on motif family classification	81
4.2	Materials & Method	83
4.2.1	Training data	84

4.2.2	The classifier feature set	84
4.3	Results & Discussion	85
4.3.1	Performance comparison with previous methods	85
4.3.1.1	MotifPrototyper	86
4.3.1.2	Sparse Multinomial Logistic Regression	88
4.3.2	Performance measurement of two large homeodomain datasets	91
4.3.2.1	Classifying homeodomain motifs by their specificity group	92
4.3.2.2	Clustering of motifs prior to metamotif training .	95
4.3.3	Comparing a metamotif density based classification to a Cartesian distance based classifier	95
4.3.4	Making metamatti available	96
4.3.4.1	The metamatti R package	96
4.3.5	The metamatti web server	97
5	Genome scale motif inference in <i>Saccharomyces cerevisiae</i>	99
5.1	Background	99
5.1.1	Genome scale motif inference	100
5.1.2	Performance inference method assessments	102
5.1.3	The Tompa <i>et al.</i> (2005) assessment	103
5.2	Materials & Method	108
5.2.1	Sequence and annotation retrieval	108
5.2.2	Motif inference	111
5.2.2.1	Unsuccessfully run algorithms	112
5.2.3	Motif comparison	114
5.2.3.1	Motif clustering with the SSD metric	116
5.2.4	Motif scanning	116
5.2.5	Predicted binding site overlap	117
5.2.6	Association of motif hits to transcription factor target genes	118
5.2.6.1	YEASTRACT	119
5.2.6.2	Reimand <i>et al.</i> (2010) TF knockout and expression data based target set	120
5.2.6.3	Harbison <i>et al.</i> (2004) ChIP-chip dataset	120

5.2.6.4	Relationship between discovered motifs and inter-species sequence conservation	121
5.2.7	Relationship between discovered motifs and sequence variation in <i>cerevisiae</i> strains	121
5.2.7.1	Positional bias of motifs	122
5.2.8	Classification of motifs with metamatti	123
5.3	Results & Discussion	124
5.3.1	Properties of inferred motifs	124
5.3.2	Finding matches to known regulatory motifs amongst <i>de novo</i> motif discoveries	128
5.3.3	TF target gene associations of the discovered motifs	141
5.3.4	Clustering of motifs and their binding sites	145
5.3.5	Comparing motifs by the overlap of their genomic matches	149
5.3.6	Looking for evidence of function for the inferred motifs	155
5.3.6.1	Inter-species conservation of the inferred motifs	158
5.3.6.2	SNP rates of the inferred motifs	161
5.3.6.3	Positional bias of motif matches close to the TSS	161
5.3.6.4	Combining the conservation, SNP rate and positional bias to highlight potentially functional motifs	164
5.3.6.5	Classification of the inferred motifs with metamatti	168
5.4	Summary	171
6	Conclusions	173
6.1	Future work	176
Appendix A - iMotifs		182
Appendix B - The motif inference tutorial		188
Appendix C - Motif inference algorithm assessment parameters		196
References		202

List of Figures

1.1	Key regulatory interactions which modulate transcription initiation.	4
1.2	TF counts versus gene counts.	7
1.3	The TF domain coverage of genomes.	10
1.4	Strongly constrained PWM motif, and one with degenerate positions in the middle.	15
1.5	The zero-or-one occurrences per sequence–motif model (ZOOOPS).	21
1.6	The multiple-uncounted sequence-motif mixture model (MUSMM).	22
1.7	The likelihood contour.	24
1.8	The NestedMICA model components: the motif set and the mixing matrix. An ensemble of three states is shown (states labelled 1,2,3).	27
2.1	A forkhead-like metamotif (inferred from an alignment of motifs) is shown alongside selection of motif samples drawn from it.	44
2.2	Visual representations of metamotifs.	47
2.3	Schematic explaining the MLE metamotif inference algorithm.	49
2.4	Example metamotifs for forkhead (A) and HSF (C) motif families from the TRANSFAC database (Matys et al., 2006).	51
2.5	The multiple-uncounted motif–metamotif mixture HMM (MUMM).	54
2.6	Incomplete hits are handled by padding the input motifs with additional columns that fit the background model optimally.	57
2.7	Motifs which were aligned and the multiple alignment summarised as an MLE metamotif with the program <i>nmalign</i> .	59
2.8	Metamotifs estimated with the metamotif nested sampler algorithm with varying relative frequency of metamotif samples.	61

LIST OF FIGURES

2.9	The metamotifs predicted at relative frequency of 0.2 are shown alongside the source metamotifs.	62
2.10	Examples of metamotifs inferred with the nested sampler algorithm from clustered motifs deposited in the TRANSFAC database (Matys et al., 2006).	64
3.1	Metamotif densities with all offsets of the metamotif (shown above the PWM) are summed over the length of the motif.	70
3.2	Synthetic metamotifs contributing to the motif prior functions used in the assessment.	73
3.3	Informative weight matrix prior improves NMICA’s sensitivity to resolve motifs present in human intronic sequence in low frequency (0.2 frequency).	75
3.4	The closest motif match to the invalid motif pattern (ZAP1) shown alongside the ZAP1 motif.	76
3.5	A NestedMICA motif inference run can be configured and run directly in iMotifs.	80
4.1	Accuracy comparison between TF domain superfamily level classification with metamatti and MotifPrototyper (10-fold cross-validation).	87
4.2	Accuracy comparison between the TF domain family classification with metamatti , and SMLR (k-fold cross-validation).	89
4.3	Confusion matrix of the 6-way TRANSFAC motif classification with the metamatti classifier.	90
4.4	Misclassified homeodomain motifs in the A) Noyes et al. (2008a) and the B) Berger et al. (2008) datasets.	92
4.5	Confusion matrix of the homeodomain specificity group classifier. Columns represent the real class, and rows represent the predicted class.	94
4.6	The metamatti motif classification web server.	98
5.1	The sequence retrieval tools included in iMotifs.	110

LIST OF FIGURES

5.2	The number of motifs from different experimental sources in the JASPAR 2010 non-redundant fungal motif dataset.	115
5.3	The ten closest matches between inferred motif sets, and JASPAR motifs. The JASPAR motifs are shown on green background. . . .	126
5.4	Summary of the average lengths and information contents of the different inferred motifs.	128
5.5	The number of statistically significant matches of the predicted motifs with A) JASPAR, and B) Zhu et al. (2009) PBM motifs. . .	129
5.6	The number of reciprocal matches between the predicted motifs and A) JASPAR, and B) Zhu et al. (2009) PBM motifs.	131
5.7	Overlap of significant matches to the JASPAR database between the three top performing motif prediction methods: NestedMICA, MEME and SOMBRERO.	133
5.8	Distribution of SSD distances of predicted motifs to significant matches in the A) JASPAR and B) Zhu et al. (2009) PBM motif sets.	134
5.9	A heatmap showing the JASPAR motifs found or missed by each of the prediction methods ($p < 0.05$).	135
5.10	Different algorithms find matches to partially overlapping subsets of the JASPAR motif set.	137
5.11	JASPAR motifs and computationally predicted motif, grouped according to their A) domain family and B) the motif set.	138
5.12	Differences in length, information content, and column-wise information content between the predicted and the JASPAR reference motifs.	140
5.13	Some <i>de novo</i> inferred motifs are able to distinguish putative TF target genes from non-target genes by the maximum bit scores achieved by the gene promoter sequences (500bp upstream promoter sequences considered).	142
5.14	Motif158 is closely similar to both the CBF1 and PHO4 motifs. . .	143
5.15	TF–target associations of the inferred motifs, when compared to JASPAR motifs.	146

LIST OF FIGURES

5.16	Dendrogram of a complete linkage clustering of all predicted motif sets with the JASPAR motifs, with the SSD metric from Down et al. (2007)	148
5.17	Clustering of JASPAR motifs with results of A) AlignACE, B) Weeder, C) MotifSampler, D) MEME, E) NestedMICA, F) YMF, G)Oligoanalysis H) SOMBRERO.	150
5.18	Numbers of clusters that contain at least one or more inferred, and one or more JASPAR motifs. Four different distance cutoffs are shown.	151
5.19	Motif redundancy as judged by the motif-to-motif SSD distance. A) Fraction of motifs which have at least one pair B) Average motif clique size.	152
5.20	The fraction of motifs with at least one matching pair, at three different significance cutoffs. The consensus string based YMF and Oligoanalysis are omitted from this analysis, because the empirical significance score used here does not behave reliably for PWMs derived from IUPAC consensus strings.	153
5.21	Motif binding site overlap of A) SOMBRERO and B) NestedMICA motifs. The rows represent inferred motifs, and the columns are JASPAR motifs. They are ordered based on an euclidian distance between the overlap patterns, with complete linkage clustering (Johnson, 1967).	154
5.22	Predicted motif similarity to JASPAR motif set on the level of binding site overlap. The bars represent the numbers of motifs which show overlap above 0.10, 0.30, 0.70, 0.90 to JASPAR motifs with the metric described in Section 5.2.5.	155
5.23	The overlap of genomic matches within motif sets. A) SOMBRERO and B) Weeder motifs are shown as examples of the predicted motif sets, and binding site overlap of JASPAR motifs are in panel C. SOMBRERO and Weeder differ in the degree of redundancy amongst the motif set. 500bp upstream sequences were analysed.	156

LIST OF FIGURES

5.24	Predicted motif redundancy on the level of binding site overlap. The bars represent the numbers of motifs which show binding site overlap with the metric described in Section 5.2.5.	157
5.25	Conservation of motifs predicted by NestedMICA.	159
5.26	The number of motifs from each of the predicted motif sets that are found more conserved than intergenic sequence of the same length.	160
5.27	The number of motifs predicted by each of the methods with lower SNP rates than randomly selected intergenic sequence of the matching length. See Section for a description of the bootstrapping based significance scores.	161
5.28	A heat map depiction of the positional bias trends of the motifs inferred with the A) SOMBRERO and B) Weeder algorithms. . .	163
5.29	The fraction of motifs output by each of the eight methods, which show a preference for positions -500 to 0. See Section 5.2.7.1 for details regarding the method.	164
5.30	Overlap of motifs predicted by A) NestedMICA and B) SOMBRERO, that have lower SNP rate than intergenic sequence, higher conservation than intergenic sequence, and are preferential placed within -500 to 0 of TSS.	165
5.31	Motifs predicted by different methods which have lower SNP rate than intergenic sequence, higher conservation than intergenic sequence, and preferential placement close to the TSS.	166
5.32	The ABF1 motif in the JASPAR database. Data originates from the CSI, PBM and Dip-CHIP based study by Badis et al. (2008)	167
5.33	Performance measures of metamatti classification of JASPAR motifs.	169
5.34	Variable importances of a JASPAR family classifier.	170
5.35	Metamatti classification of the predicted motifs at the 0.6 classification probability cutoff.	171
6.1	Three Markov chains aiming to draw a sample from $\mathbb{P}(\mathbf{M} \mathbf{G}, p)$. .	178
6.2	Mixing matrices and their correlations.	179

LIST OF FIGURES

6.3	The sampling algorithm produces mixing matrices that are closely related in correlation pattern to the target (gene expression) correlation matrix.	180
A1	iMotifs can present motif sets and alignments.	184
A2	Output of the nmevaluatebg command plotted in R.	191
A3	The predicted motif alongside known STAT motifs from the TRANS-FAC database.	195
A4	Evaluation of sequence background model class counts at Markov chain order 1.	197
A5	Parameter choices used with Oligo-analysis.	201

Chapter 1

Introduction

The genetic information stored in our DNA is transcribed into RNA by large molecular holoenzymes called RNA polymerases. In eukaryotic organisms there are three types of RNA polymerases, out of which RNA polymerase II (Pol II) is the one responsible for transcribing protein-coding genes and many noncoding RNAs such as micro-RNAs (Megraw et al., 2009; Saltzman and Weinmann, 1989). Pol II activity is highly regulated at the level of the individual transcript, and this regulation is essential for both cellular homeostasis and development of multicellular organisms (Fuda et al., 2009). The most central and best understood mechanisms of gene regulation is mediated by the interaction of sequence specific transcription factors (TFs) with DNA target sequences, each other and with other members of the Pol II complex (Mitchell and Tjian, 1989). Transcription factors orchestrate the transcription cycle because their activities are in turn controlled by cellular signals, for instance on the level of post-transcriptional modifications and protein-protein interactions. Each factor has a preference towards a specific set of DNA words which dictates the positions at which it is recruited to the genome. As this mechanism of DNA site recognition acts in part to choose the target genes of the transcription factors, the DNA patterns are commonly known as ‘regulatory motifs’.

In this introduction I firstly outline the known regulatory mechanisms acting on the level of transcription to highlight the importance of and challenges in the study of transcriptional regulatory mechanisms (Section 1.1). I then briefly review the previous literature on computational regulatory motif inference (Section 1.2),

before introducing the specific computational methodology used in the project (Section 1.3). I then discuss the biological resources which were applied in this work (Section 1.4), and finally introduce the specific contributions in this work to the inference and classification of regulatory motifs (Section 1.5).

1.1 Gene regulation by control of transcription

Transcription factors act by promoting or inhibiting the recruitment of Pol II to the gene’s promoter, to initiate RNA transcription at the transcription start site (TSS) of the gene, eventually leading to the generation of a full-length RNA transcript. This classical understanding of eukaryotic transcriptional regulation – involving only proximally located transcription factor binding sites (TFBS) – has had to give way to a more complex view of regulatory interactions. Firstly, factors which interact with Pol II not only act to recruit it to the complex, but can also affect its post-initiation clearance from the promoter, elongation of the transcript, and its termination, all of which are found to be rate-limiting and therefore highly likely regulated steps in the case of some genes (Venters and Pugh, 2008). Secondly, regulatory regions are found not only proximal to the TSS, but also kilobases further upstream, or even downstream, of their target genes in an orientation independent manner (Banerji et al., 1981).

The more distal regulatory regions are known as “enhancer” regions when they have an activatory role, and “silencer” regions when they inhibit recruitment of the transcriptional machinery (Visel et al., 2009). Several large studies have been conducted and are currently underway to systematically discover and catalogue tissue specific enhancers acting in mammalian and fish genomes (Ellingsen et al., 2005; Pennacchio et al., 2006; Visel et al., 2008). Enhancer- and silencer-like regions, as well as insulators which set the ‘borders’ of the chromatin domains regulated by enhancers and silencers, have also been described in yeasts (Bi and Broach, 2001; Buchman et al., 1988). The chromatin packaging of the genome sets limits to the regions that are available for transcription factor binding, and regulatory interactions that control this process can both activate and repress expression (Li et al., 2007; Steinfeld et al., 2007; Venters and Pugh, 2008). Figure 1.1A depicts these various factors and interactions involved in transcriptional

regulation.

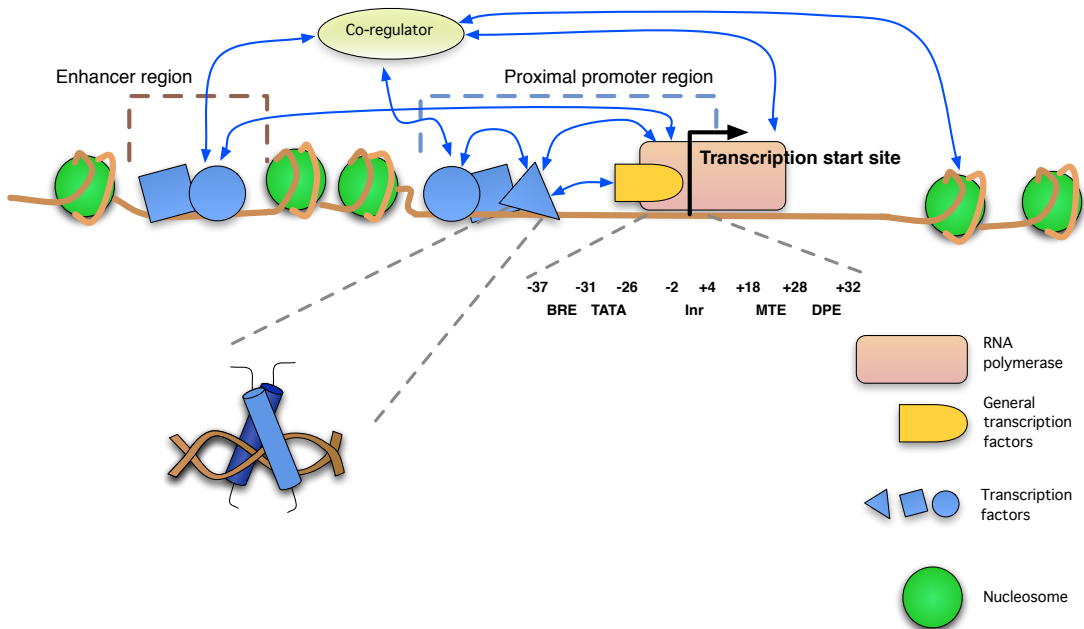
General transcription factors (GTFs) bind to specific target sequences close to the transcription start site (TSS) at defined locations (Venters and Pugh, 2008), as shown in Figure 1.1A. Names and approximate positions are shown for the GTF target sequences. Regulatory transcription factors bind either to activate or repress the transcription of the target gene by binding to their target DNA sequences either near the core promoter or more distally (enhancers). Interactions between the TFs, GTFs and the Pol II are also important for regulation. Co-regulators which do not themselves bind DNA in a sequence-specific manner also interact with GTFs, TFs and nucleosomes (via modified histone tails). Both activation and repression can occur via each of these interactions.

Trans-acting enhancer regions are thought to contribute to eukaryotic gene regulation by looping DNA to promote the recruitment of the transcription machinery at a TSS (Figure 1.1B). Many genes are known to achieve their observed expression patterns through the combination of weak promoters and enhancer regions, which supplement them. In this example the expression pattern of a gene is modulated by both a promoter, as well as brain and limb specific enhancer elements. Silencer elements, which were not depicted here, can also act from a large distance to the TSS.

Enhancers and silencers rely on the organisation of genes into chromosomal domains that can in part be co-regulated. However, it has also been suggested that TF target genes are organised non-randomly for the majority of TFs, even in *S. cerevisiae* with its compact non-coding genome (Janga et al., 2008), short promoter sequences and relatively few examples of long-distance enhancer or silencers. The organisation of targets of a TF along chromosomes, possibly through their association in shared three-dimensional ‘chromosomal territories’ (Cremer and Cremer, 2001; Gasser, 2002; Lieberman-Aiden et al., 2009), could pose yet another largely uncharacterised level of regulatory information. The effect of neighbouring genes sharing similar promoter motifs has also been shown in *D. melanogaster* (Zhu and Halfon, 2009).

Another mechanism of transcriptional regulation not depicted above is the tissue or time specific use of alternative TSSs. The majority of human and mouse Pol II promoters have clusters of close TSSs instead of a single one (Frith et al.,

A)



B)

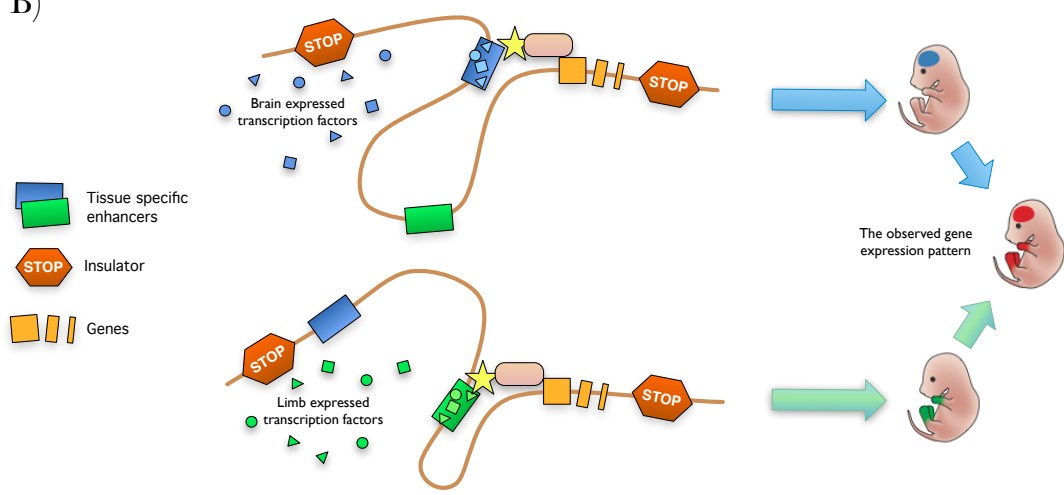


Figure 1.1: Key regulatory interactions which modulate transcription initiation. A) A promoter centric view on transcriptional regulation. Transcription factors interact with DNA and other regulatory factors to modulate the action of the RNA polymerase. B) An enhancer centric view on transcriptional regulation. Figure adapted from [Visel et al. \(2009\)](#) and [Fuda et al. \(2009\)](#).

2008). Larger scale TSS usage variation also occurs. Alternative promoter usage can in fact act as a mechanism for creating variant isoforms of gene products (Carninci et al., 2006), and changes in alternative TSS use are found associated to tissue and developmental stage specific dynamics of transcription (Consortium et al., 2009; Valen et al., 2008).

The identification and study of gene regulatory sequence is more difficult than protein-coding sequence because of several factors. Perhaps most importantly, the conservation pattern of regulatory sequence does not resemble that of protein coding sequence. Purifying selective constraint in regulatory sequences is often seen between closely related species (Hardison, 2000; Loots et al., 2000; Ludwig, 2002), but genomic TF binding studies suggest that turnover of regulatory elements occurs at remarkably high rate even when expression pattern (i.e. the connectivity of the TF network) shows little change (Schmidt et al., 2010). Indeed, changes in regulatory interactions have been hypothesised to be a cause of species divergence both in fungi (Borneman et al., 2007) and in animals (Carroll et al., 2000; Galant and Carroll, 2002). Furthermore, regulatory elements are often not constrained in the ordering, orientation or number of functional sites (Ludwig, 2002; Markstein and Levine, 2002). Consequently, alignment based comparative methods, which have been largely developed for the study of protein coding DNA, suffer from misalignments. For instance only 59% agreement is found between methods in the case of the 12 whole-genome *Drosophila* genomes aligned in the study by Stark et al. (2007). Detecting selective constraint acting on short blocks – often less than 20bp long (Bergman and Kreitman, 2001) – is not easy. Indeed, alignment based comparative analyses can only identify a small fraction of functional elements (Siggia, 2005). Alignment free *cis*-regulatory motif discovery methods which can consider recurring signals between related species to be conserved regardless of alignment or orientation are only beginning to appear (Gordan et al., 2010; Kim et al., 2010; Xie et al., 2009).

TF binding sites frequently occur in clusters – homotypic or heterotypic (Gotea et al., 2010). Site proximity of different TFs can modulate both cooperative and repressive interactions between different TFs (Kulkarni and Arnosti, 2005; Lebrecht et al., 2005), and competition of TFs for overlapping TFBSs is known to contribute for instance to *Drosophila* embryo segmentation (Walter

et al., 1994). Repetitive (homotypic) clustering of sites for the same TF is also well documented and can act to ensure stable binding (Cunningham and Cooper, 1993) or modulate a graded transcriptional response (Donahue et al., 1983). Interestingly, it has been suggested that even proximal or overlapping spacing of sites might be produced by selection mechanisms acting to maintain the overall composition of TFBSs in *cis*-regulatory elements instead of a constraint acting to maintain binding site position or orientation (Lusk and Eisen, 2010).

1.1.1 Sequence specific transcription factors

Understanding properties of *cis*-regulatory sequences is an ongoing challenge faced by the field of regulatory genomics. Another challenge which similarly continues to require extensive experimental and computational work is the annotation of transcription factors in genomes. High coverage annotations of TF genes are available for some well studied organisms in manually curated databases, ranging from RegulonDB for *Escherichia coli* (Huerta et al., 1998; Salgado et al., 2006), DBTBS for *Bacillus subtilis* (Ishii et al., 2001; Sierro et al., 2008), FlyBase (Wilson et al., 2008b) and FlyTF (Adryan and Teichmann, 2006; Pfreundt et al., 2010) for *Drosophila*, TFdb (Kanamori et al., 2004) and TFCat (Fulton et al., 2009) for human and mouse.

Advanced comparative sequence analysis techniques based on the use of protein domain profile Hidden Markov models have been helpful in systematically predicting large numbers of transcription factors for many sequenced genomes, both eukaryotic and prokaryotic (Kummerfeld and Teichmann, 2006; Wilson et al., 2008a). To illustrate the insight that TF annotation gives about transcription regulation, a comparison is shown below between the number of predicted sequence specific transcription factor genes out of the total number of protein coding genes for four eukaryotic species, as well as the *E. coli* (K12). The data presented is from the DBD database (Wilson et al., 2008a) (Release 2.0, downloaded 12/6/2010) which predicts TFs based on statistically significant matches to protein domain models from either the PFAM (Finn et al., 2010; Sonnhammer et al., 1997) or the SUPERFAMILY (Gough et al., 2001; Wilson et al., 2009) databases.

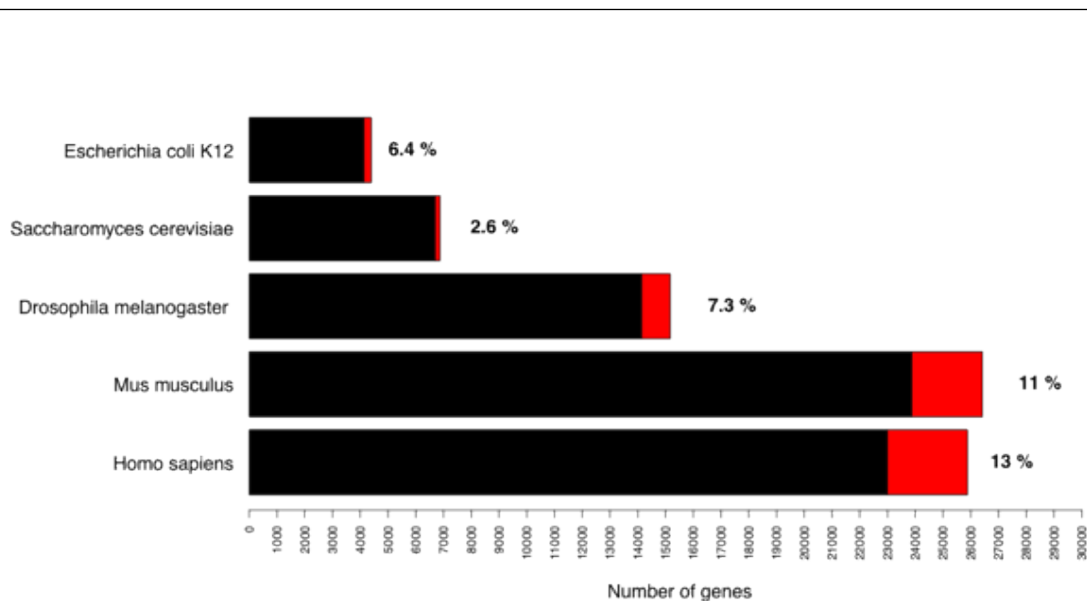


Figure 1.2: TF counts versus gene counts. The data presented is from the DBD database ([Wilson et al., 2008a](#)).

The TF number comparison shown in Figure 1.2 highlights several properties of transcriptional regulation. Firstly, the number, and more interestingly the proportion of TFs, increases for large genomes. For example in the case of the human genome, 13% of its approximately 23,000 genes are predicted to be TFs, whereas only 2.6% out of the 6,700 *Saccharomyces cerevisiae* genes are annotated as TFs. The increase in fraction of regulatory factors from the total number of genes amongst eukaryotes is thought to be a manifestation of the increased need to specifically regulate genes in larger, more complex organisms.

Single-cellular eukaryotic genomes contain a smaller fraction of TFs from total gene number when compared to bacteria (*S. cerevisiae* at 2.6%, *E. coli* at 6.4%). This is a well documented observation and thought to be a result of tissue and condition specific combinatorial regulation of genes in eukaryotes ([van Nimwegen, 2003](#)), epigenetic regulation ([Choi and Kim, 2008](#)), as well as the additional post-transcriptional control mechanisms such as microRNAs that are abundant in higher eukaryotes but absent in some fungi such as *S. cerevisiae* ([Grimson et al., 2008](#)). A power-law relationship has been described between the genome size and the number of TFs present in a genome, both in eukaryotes and prokaryotic organisms, with a lower exponent in eukaryotes ([van Nimwegen, 2003](#)).

Known binding site motifs of eukaryotic TFs tend to be less constrained than bacterial motifs (Wunderlich and Mirny, 2009). This together with the much larger genome sizes of eukaryotes also points at the requirement for additional levels of regulation. To put it simply, the DNA motif of a eukaryotic TF does not contain enough information to help it distinguish its cognate sites from non-functional sites that could occur as often as every $10^3 - 10^4$ nucleotides (assuming a simple genomic background model parameterised by average GC content). This view is supported by *in vivo* ChIP-seq binding studies of genomic binding sites of several eukaryotic TFs: assumably non-functional binding far from genes is found to be abundant in several studies (Robertson et al., 2007; yong Li et al., 2008). Abundant non-functional binding of TFs was in fact observed already in a much more laborious UV-crosslinking and Southern blot study by Walter et al. (1994).

Clustering of TFBSs can provide additional regulatory information by allowing combinatorial binding of TFs (Georges et al., 2010; Makeev et al., 2003; Papatsenko, 2009). More recently, a large scale analysis of human and mouse TF protein-protein interactions and expression measurements of the factors strongly suggests the combined action of sequence specific TF complexes, most importantly homeobox factors, in cell fate specific regulation of target genes (Ravasi et al., 2010). Homeobox factors are interesting in this context because they are especially common in mammals (Wilson et al., 2008a), they have short five or six nucleotide long motifs (Affolter et al., 2008) and they often bind with an additional, specific co-factor in a manner specific to cell-type (Ravasi et al., 2010). In conclusion, in higher eukaryotes it is important to consider gene regulation as a combination of multiple mechanisms including for instance increased combinatorial interactions of TFs, multiple classes of noncoding RNAs (Jacquier, 2009), epigenetic mechanisms (Jaenisch and Bird, 2003) and alternative transcripts (Carninci et al., 2006).

When the TFs of each organism are grouped by the content of their DNA binding families (Figure 1.3), it becomes apparent that TFs of all the organisms shown here fall into a much smaller number of DNA binding domains (e.g. 155 domains in 2886 human TFs, or 46 domains in 177 *S. cerevisiae* TFs). The low overlap between TF domain content of different genomes highlights that many of

the TF families have expanded within specific lineages ([Babu et al., 2004](#)). For example, the overlap between domains annotated in *H. sapiens* and *E. coli* is only four domains (*HTH₃*, *HTH₁₁*, *CSD* and *PAS* domains) whereas the mammals *H. sapiens* and *M. musculus* share 151 domains. The reader is referred to [Wilson et al. \(2008a\)](#) for a more thorough discussion of the kingdom specific expansion of DNA binding domains.

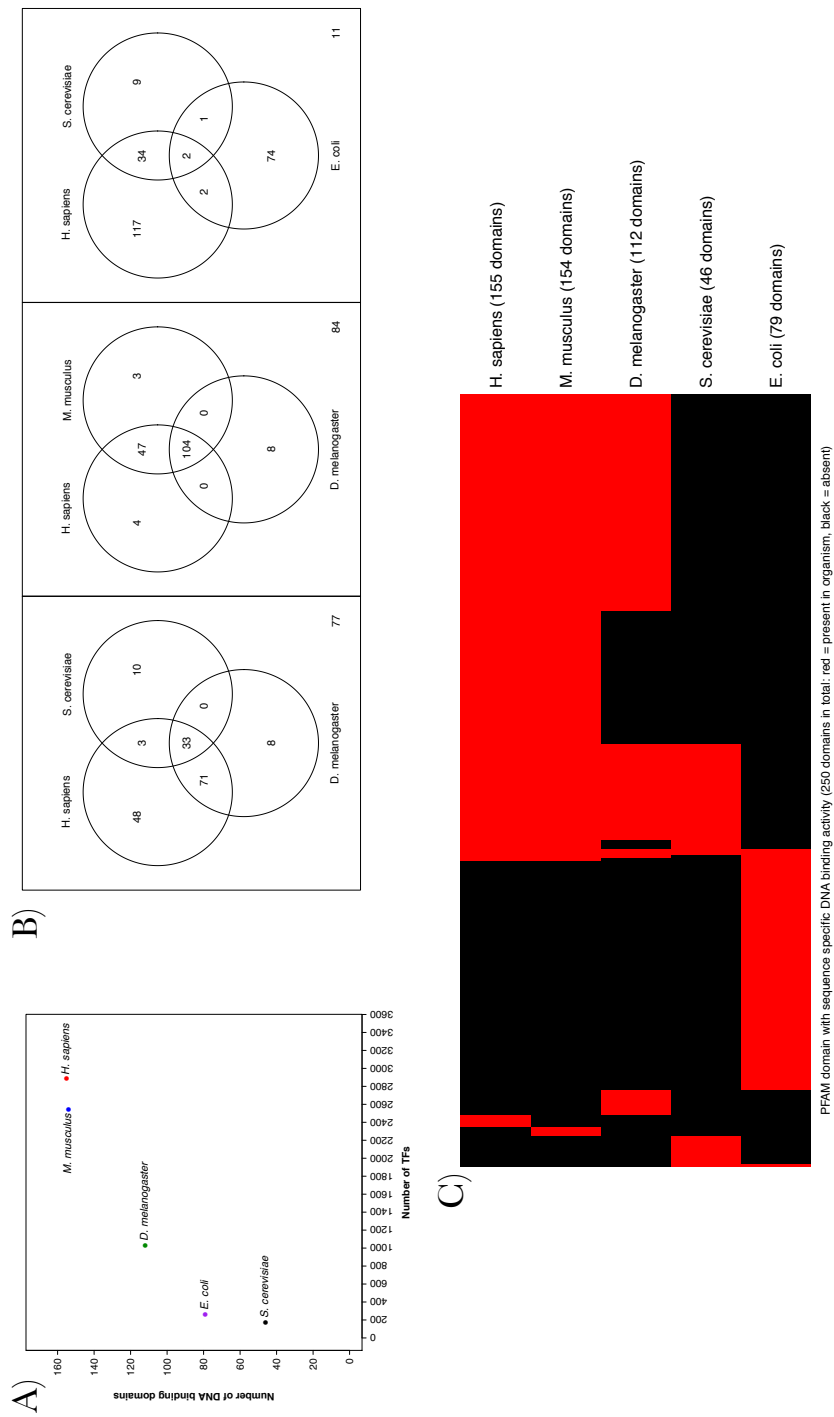


Figure 1.3: The TF domain coverage of genomes. A) Number of TF domain families vs TF gene counts in a genome. B) Overlap of TF domains between different organisms, represented in a Venn diagram. C) Overlap of TF domains between different organisms, represented as a heatmap (red = domain present, black = domain absent in organism). Data presented here originates from the DBD database ([Wilson et al., 2008a](#)).

1.1.2 Binding specificity of transcription factors

Determining the genomic binding sites and modelling sequence specificity patterns of TFs has proven a formidable task. Currently the only eukaryotic organism for which binding specificity of the large majority of its transcription factors has been determined based on DNA–protein interaction assays is *Saccharomyces cerevisiae*, which has a small genome by eukaryotic standards (12 Mbases with 6,532 protein coding genes according to the Ensembl release 58.1j (Hubbard et al., 2009)). I will take special interest here in discussing *S. cerevisiae* because what is already known of its transcriptional regulation is the closest that we currently have to a ‘regulatory code’ of any eukaryotic genome, and because computational genome scale regulatory motif inference in *S. cerevisiae* is the focus of the work described in Chapter 5.

The first large scale effort towards the *in vivo* profiling of TF binding on a genome scale was the study by Harbison et al. (2004), where ChIP-chip assays were conducted with 203 sequence specific TFs, each factor’s binding profile being measured in one or more of 12 different growth conditions. The original analysis of the paper detailed a high confidence motif for 63 of the 203 TFs studied. MacIsaac et al. (2006) then provided a re-analysis of the large dataset with two phylogenetic footprinting based inference algorithms. PhyloCon (Wang and Stormo, 2003) and Converge (MacIsaac et al., 2006) yield motifs for an additional 36 TFs. The resolution of the ChIP-chip assay however does not reach beyond 500nt due to the limitations set by the use of randomly sheared genomic DNA fragments and tiling arrays (Sikder and Kodadek, 2005). ChIP-chip in other words is not ideal for determining accurate binding site profiles for TFs. ChIP followed by sequencing (ChIP-seq) offers a partial solution to the resolution problem, and allows more accurate and quantifiable *in vivo* study of protein-DNA binding. ChIP-seq assays with TFs have been to date conducted with TFs of larger, higher eukaryote genomes¹, with the exception of Lefrançois et al. (2009) who assayed a series of budding yeast TFs as a proof of concept of a multiplexed ChIP-seq experiments (a single sequencing experiment contains samples for multiple TFs). *In vitro*

¹Large scale efforts to profile sequence specific TF specificity in human and several model organisms *in vivo* with ChIP-seq have begun as part of the ENCODE and modENCODE projects. See <http://www.genome.gov/10005107> for more information.

measurements of TF DNA specificity however already provide a close-to-complete, high resolution dataset for the *S. cerevisiae*: a protein binding microarray (PBM) [Mukherjee et al. \(2004\)](#) based study by [Zhu et al. \(2009\)](#), and a study by [Badis et al. \(2008\)](#) using a combination of PBMs, cognate site identifier microarrays ([Warren et al., 2006](#)), and DIP-chip ([Liu et al., 2005](#)).

Our knowledge of sequence specific protein–DNA interactions is far less complete in the case of larger eukaryotic genomes than it is in the budding yeast. The JASPAR database ([Portales-Casamar et al., 2010](#)), which contains a high quality non-redundant resource of TFBS motifs for different kingdoms of life, contains only 75 TFBS motifs for the 2886 TFs in human. For mouse there are only 40 TFs present in JASPAR (out of 2548 TFs). Furthermore, most high throughput studies to date have concentrated on a small number of highly expanded TF domain families, such as homeodomains ([Noyes et al., 2008a](#)) and basic helix-loop-helix factors ([Grove et al., 2009](#); [Maerkl and Quake, 2009](#)), with the exception of [Badis et al. \(2009\)](#) whose 104 TFBS motifs cover 22 different families of TFs. New high-throughput methods for studying DNA–protein interactions are becoming available in addition to universal PBMs which currently provide majority of the publicly available high-throughput TF–DNA specificity data. These new promising methodologies include ChIP-seq ([Robertson et al., 2007](#)), bacterial one-hybrids ([Meng and Wolfe, 2006](#); [Noyes et al., 2008a,b](#)), multiplexed massively parallel SELEX ([Jolma et al., 2010](#)) and a microfluidic molecular interaction assay platform by [Maerkl and Quake \(2007a\)](#).

Although new protein–DNA interaction probing technologies have the potential to transform our knowledge of eukaryotic transcriptional regulation, it is also clear that efficient computational methods for motif inference and classification continue to be of key importance. My aim in Chapter 2 is to present a new class of motif family models that can be learned using experimentally determined PWM motifs, such as those derived from new HT technologies. In Chapters 3 and 4 I present applications of motif family models for sensitively inferring motifs from genomic sequence, and for classifying computationally inferred motifs by their DNA binding domain type, respectively. In both of these lines of work use experimentally determined motif data to provide a comparison for evaluating computational predictions. Experimentally determined regulatory motifs are also

central to the *S. cerevisiae* motif inference performance benchmark in Chapter 5, where *de novo* predictions are compared to experimental motifs.

1.2 Computational inference of transcription factor binding site motifs

Computational inference of TFBSs by applying short motif inference algorithms to pieces of genomic DNA sequence is a long-standing research problem. It has motivated computational biologists to propose literally hundreds of algorithms over the course of more than 30 years. Many of these algorithms are introduced in previous reviews (Das and Dai, 2007; MacIsaac and Fraenkel, 2006; Nguyen and Androulakis, 2009; Sandve and Drabløs, 2006), and therefore only essentials of different approaches are covered here.

The first motif inference algorithm was published in the landmark paper by Korn et al. (1977) where pairwise comparisons of aligned sequence immediately close to prokaryotic transcription start sites (TSS) and terminator sequences were used to infer recurring motifs. The Korn et al. (1977) approach, which simply lists recurring sequence words found by pairwise comparisons of noncoding DNA sequence, is the earliest precursor to oligonucleotide word enumeration based motif inference algorithms. Such algorithms aim to exhaustively list possible k -mers that satisfy an objective function such as a conservation or significance score, commonly allowing a certain maximum number of mismatches. This approach is still taken in several recently published algorithms, ranging from reporting ranked k -mers of a specified length (Helden et al., 1998; van Dongen et al., 2008) to IUPAC consensus strings that allow for describing degeneracy in positions (Marschall and Rahmann, 2009; Xie et al., 2005, 2007). In fact the Tompa et al. (2005) *ab initio* motif inference method benchmark showed the word-enumeration based Weeder (Pavesi et al., 2001) as one of the best performing inference method of the 13 methods that were tested. The Tompa et al. (2005) benchmark is discussed in more detail in Section 5.1.2. Enumeration based methods can be made computationally very fast through the use of modern computers with access to a large volume of runtime memory together with highly optimised look-up data

structures, such as suffix trees which were originally introduced in computational biology detection of repeat elements (Sagot, 1998).

Word enumeration methods however have certain inherent limitations. Firstly, the reliance on lookup based data structures make them incapable of modelling very long TFBS patterns – 8-mers or 10-mers are typically studied – which are known to be present amongst eukaryotic TFBS motifs of many TF families. Cys₂His₂ zinc finger motifs for instance can be as long as 15 or 20 nucleotides due to the common architecture of their protein–DNA interaction which involves several zinc finger domains binding in tandem (LeClerc et al., 1991; Wolfe et al., 2000). Motifs with a large number of weakly constrained positions are also problematic for word enumeration methods which generally require sequence word clustering based on edit distance to group individual related sequence words to motif models to describe degeneracy. The great majority of TFs do not bind to a unique DNA ‘word’, but instead they show a distribution of binding affinity across a number of possible sites (known as ‘degeneracy’). Degenerate positions are well known to occur in TFBS motifs (examples with degenerate motifs are shown in Figure 1.4), and the information content of a position has been shown to correlate with its conservation (Moses et al., 2003) and the number of contacts the base makes with amino acid residues (Gelfand and Mirny, 2002). Genome scale *in vivo* profiling of transcriptional control is rapidly forming an image of transcriptional control where not only is a large spectrum of possible binding sequences observed (Badis et al., 2009), but also that even weak binding sites can exert a regulatory response (Gertz et al., 2009) and therefore are biologically meaningful. Therefore, models of sequence motifs should ideally represent the sequence specificity distribution as completely as possible, whilst being able to weight strongly binding sequences above weakly binding sequences, neither of which is possible with *k*-mer enumeration based models.

The above-mentioned limitations of word enumeration methods in describing transcription regulatory motifs resulted in development of probabilistic motif inference methods, which most commonly use the position weight matrix (PWM) as the motif model. The PWM is described in more detail in Section 1.2.1, and examples of PWMs are shown in Figure 1.4 as sequence logos (Schneider and Stephens, 1990).

improve sensitivity to detect regulatory motif and *cis*-regulatory modules have also been developed (Siddharthan, 2008; Sinha et al., 2004; Wang and Stormo, 2003).

In conclusion, a multitude of different approaches have been applied to regulatory motif inference. Finding a suitable algorithm for a biological problem at hand can be a daunting task for a researcher, and indeed one might expect that standard benchmarking methods would have surfaced in the literature of motif inference algorithms. However, the great majority of the above mentioned publications describing motif inference algorithms are either:

1. applied to a specific biological problem without an explicit performance assessment with other algorithms.
2. compared with a publication specific biological dataset with one, two or a handful of different common tools such as MEME (Bailey and Elkan, 1995).
3. compared with a synthetic sequence set with one, two or a handful of different common tools.

Performance comparison of motif inference tools is itself a non-trivial problem. Very few comprehensive attempts have been made to date to systematically assess different tools (Li and Tompa, 2006; Pevzner and Sze, 2000; Sinha and Tompa, 2003a; Tompa et al., 2005). The assessment by Tompa et al. (2005) is perhaps the most comprehensive to date, covering 13 different algorithms. In Chapter 5 I discuss the challenges of measuring motif inference performance with synthetic and real promoter sequence (Section 5.1.3), and describe a new, large scale motif inference benchmark challenge (Section 5.3.2).

1.2.1 The position weight matrix

The PWM, also known as a position specific scoring matrix (PSSM) or a gapless profile, is a commonly used probabilistic model used in motif inference algorithms. It has been found to preserve more of the information of individual motif positions (columns) than consensus string motifs, and to systematically perform better in

describing regulatory binding site patterns (Osada et al., 2004). It is also the motif model of choice in my work.

PWMs are probabilistic sequence motif models that can be scanned along sequence to assign a score for a sequence window to contain a motif match. Commonly a threshold is determined for the sequence window scores, such that windows where the threshold is exceeded are called motif matches (potential binding sites). A large part of my work has revolved around analysing properties of inferred PWM motifs and their connection to previously known motifs (Chapters 2, 3, 4) with the use of motif family models. In addition, in Chapter 5 I present an assessment of the prediction performance of several *de novo* motif discovery algorithms. A formal definition of the PWM is therefore in place, and provided below (adapted from Rahmann et al. (2003)).

Let \mathbb{A} be a finite alphabet with cardinality $|\mathbb{A}|$ ($|\mathbb{A}| = 4$ for DNA and RNA). If \mathbb{A}^k represents the space of all string of k symbols from \mathbb{A} , a PWM \mathbf{M} is a probability distribution over all of the sequence positions i of \mathbb{A}^k . More specifically, \mathbf{M} is an $|\mathbb{A}| \times k$ matrix where each column vector \mathbf{M}_i represents the weights $m_{i,j}$ (nucleotide j at sequence position i) for a multinomial distribution, i.e. $\mathbf{M}_{i,j}$ are nonnegative such that $(\sum_{i \in \mathbb{A}} \mathbb{A}_i = 1)$.

\mathbf{M} is thought of as a generative model for sequences from \mathbb{A}^k such that symbol s at each position i is generated independently according to the multinomial distribution parametrised by \mathbf{M}_i . The probability $\mathbb{P}_{\mathbf{M}}(S)$ of a sequence S from \mathbb{A}^k being generated by \mathbf{M} is $\mathbb{P}_{\mathbf{M}}(S) = \prod_{i=1}^k M_{i,S_i}$. \mathbf{M} is in other words a product multinomial distribution over \mathbb{A}^k . The probability $\mathbb{P}_{\mathbf{M}}(S)$ score is often used as the match score. The NestedMICA suite motif scanning algorithm which I have used, provided in the program `nmScan` (Down and Hubbard, 2005), transforms the scores to bit scores and transforms them such that maximum score reported is 0 (Function 1.1).

$$W(S, p) = \prod_{i=1}^{|W|} W_i(S_{p+i-1}) \quad (1.1)$$

In brief, the PWM is a model for gapless position-specific probability distributions of nucleotides which assumes independence of nucleotide positions (Rahmann et al., 2003). Departures of the position independence assumption have

been reported in the form of variable length linkers, interdependencies between nucleotides at different binding site positions (Badis et al., 2009; Benos et al., 2002a; Bulyk et al., 2002), and compensatory mutations that maintain the binding energy and function of binding sites (Mustonen et al., 2008). More complex probabilistic motif models based on for instance Bayesian (Barash et al., 2003; Ben-Gal et al., 2005) and Markov networks (Sharon et al., 2008) have been developed to fit these observations. With the exception of the newest DNA–protein interaction assays which provide direct binding energy measurements of a protein with a large spectrum of different DNA binding sites (Berger et al., 2006; Maerkl and Quake, 2007b), parameter estimation of motif models more complex than the PWM is hard with often scarce biological data. The PWM therefore remains the model of choice for most large scale motif inference tools; it is intuitive to interpret as a sequence logo (Schneider and Stephens, 1990) and retains more of the information contained in binding site patterns than sequence word based models (Osada et al., 2004).

1.3 Computational methodology

Several lines of the work I describe in the later chapters builds on previously described computational frameworks, the most important of which I will summarise below. Firstly, Hidden Markov models are used for modelling sequential data (described in Section 1.3.1, applied in Chapter 2 for inferring motif family models). Secondly, the nested sampling Monte Carlo method used for drawing samples from complex probability distributions that are not analytically tractable (described in Section 1.3.2, applied in Chapters 2 and 3). Thirdly, random forest classification is applied in Chapters 4 and 5 for the supervised machine learning task of predicting TF domain labels for regulatory motifs (Section 1.3.4).

1.3.1 Hidden Markov Models in motif inference

A Hidden Markov Model (HMM) is a model for sequential signals. It is a stochastic finite automaton consisting of finite number of states. Each state has an associated probability distribution, and the distribution is typically multidimensional

(Dogruel, 2008). The HMM was originally developed and described in a series of papers by Baum *et al.* (Baum, 1972; Baum and Petrie, 1966; Baum *et al.*, 1970; Baum and Eagon, 1967; Baum and Sell, 1968), and it quickly developed into a popular model in speech recognition (Baker, 1975). Applications to biological pattern recognition problems from data such as protein and DNA sequence arrived much later, sparked by several widely circulated papers from Haussler and others (Brown *et al.*, 1993; Krogh *et al.*, 1994). In these papers HMMs were described as a superset of the profile multiple alignment methods which were already commonly used in modelling protein sequence. Indeed, HMM profile based protein domain families computed with tools such as HMMER (Eddy, 1998) and stored in databases such as Pfam (Finn *et al.*, 2010; Sonnhammer *et al.*, 1997) and SUPERFAMILY (Wilson *et al.*, 2009) are perhaps the most ubiquitous biological application of HMMs in computational biology, in addition to other common uses such as gene finding (Stanke and Waack, 2003). The HMM is also a commonly used formalism in regulatory motif inference problems. Firstly however let us arrive at a formal definition of an HMM and some of the common terminology used in connection to them.

For an observable sequence $O = O_1O_2 \dots O_T$ emitted by HMM λ , each of its observables (symbols) is said to be emitted by a sequence of T hidden states from a finite set of N hidden states $S = S_1, S_2, \dots, S_N$. As described by Rabiner (1989), the model is parameterised by three types of parameters:

- 1) The transition probability distribution A_{ij} (Equation 1.2)

$$a_{ij} = P[q_{t+1} = S_j | q_t = S_i], 1 \leq i, j \leq N \quad (1.2)$$

HMMs are often depicted as a diagram with directed, weighted edges showing transitions a_{ij} between nodes representing states. The missing edges between states correspond to transitions with probability 0 (see Figures 1.5 and 1.6).

- 2) The observable emission probability distribution $B = b_j(k)$ (Equation 1.3)

$$b_j(k) = P[v_k \text{ at } t | q_t = S_j] 1 \leq j \leq N \cap 1 \leq k \leq M. \quad (1.3)$$

3) The initial state distribution $\pi = \pi_i$ (Equation 1.4)

$$\pi_i = P[q_1 = S_i], 1 \leq j \leq N \text{ and } 1 \leq k \leq M. \quad (1.4)$$

A HMM can be used to solve several types of problems in relation to the observable sequence and the hidden state path, the three most common of which are:

1. Given a sequence of observations $O = O_1O_2 \dots O_T$ and a HMM $\lambda = (A, B, \pi)$, compute the probability of the observation sequence, given the model λ , that is, $P(O|\lambda)$. Computing $P(O|\lambda)$ involves integrating the possible state paths through the model with their likelihood (also known as the forward algorithm).
2. Given a sequence of observations $O = O_1O_2 \dots O_T$ and λ , how do we find the most likely hidden state path $Q = q_1q_2 \dots q_T$ (the ‘Viterbi path’) that generates (‘explains’) a sequence of observables. The algorithm that solves this problem is known as Viterbi decoding.
3. Adjusting λ parameters (A, B, π) such as to maximise $P(O|\lambda)$.

My work with the motif family model estimation problem has involved working on the first of the three above problems: defining a likelihood function over the sequence of nucleotide sequence motif columns and expressing it as an HMM forward algorithm. This work is described in more detail in Chapter 2, and its applications into motif inference and motif classification are described in Chapters 3 and 4.

Motif inference algorithms are also often expressed with an HMM model. The most common such sequence model, used for example in MEME (Bailey and Elkan, 1994), is the zero-or-one occurrences per sequence model, or ZOOPS (Figure 1.5). The common feature of the sequence models used in probabilistic motif inference algorithms is that they express biological sequence (e.g. DNA) as a string of symbols emitted by a series of emissions from a background model and a sequence motif. The background state generates the ‘un-interesting’ symbols

in the analysed sequence (the non-motif containing positions, which in most promoter analysis problems constitute the bulk of the sequence). The ‘interesting’ states are the overrepresented motifs, which are parameterised most commonly as a position weight matrix (PWMs described in Section 1.2.1).

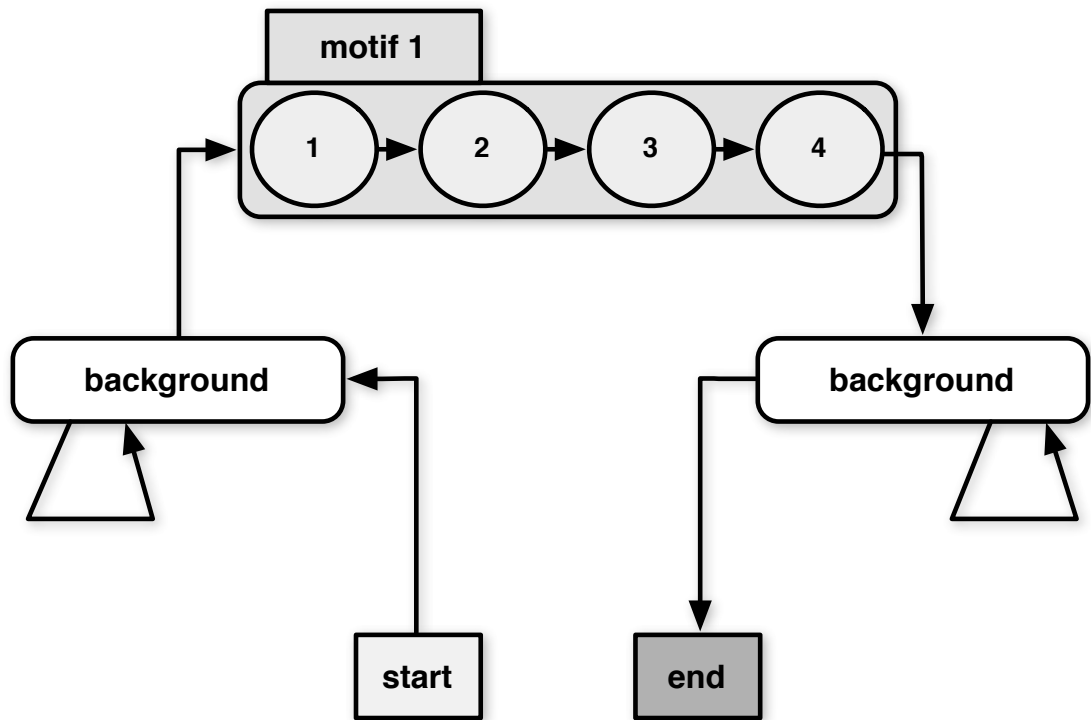


Figure 1.5: The zero-or-one occurrences per sequence-motif model (ZOOPS).

The sequence HMM used in the NestedMICA motif inference algorithm (Down and Hubbard, 2005) which I have also expanded as part of my project is slightly more complex, allowing multiple motifs to be modelled simultaneously. An example of these ‘multiple-uncounted sequence-motif mixture models’ (MUSMM) are shown in Figure 1.6.

The important improvement of the MUSMM model over the ZOOPS model is that it allows simultaneous motif learning from sequence data. In other words parameter estimation of each of the motifs is not done in iterations of learning a motif, masking its putative hit positions from the sequence, before repeating

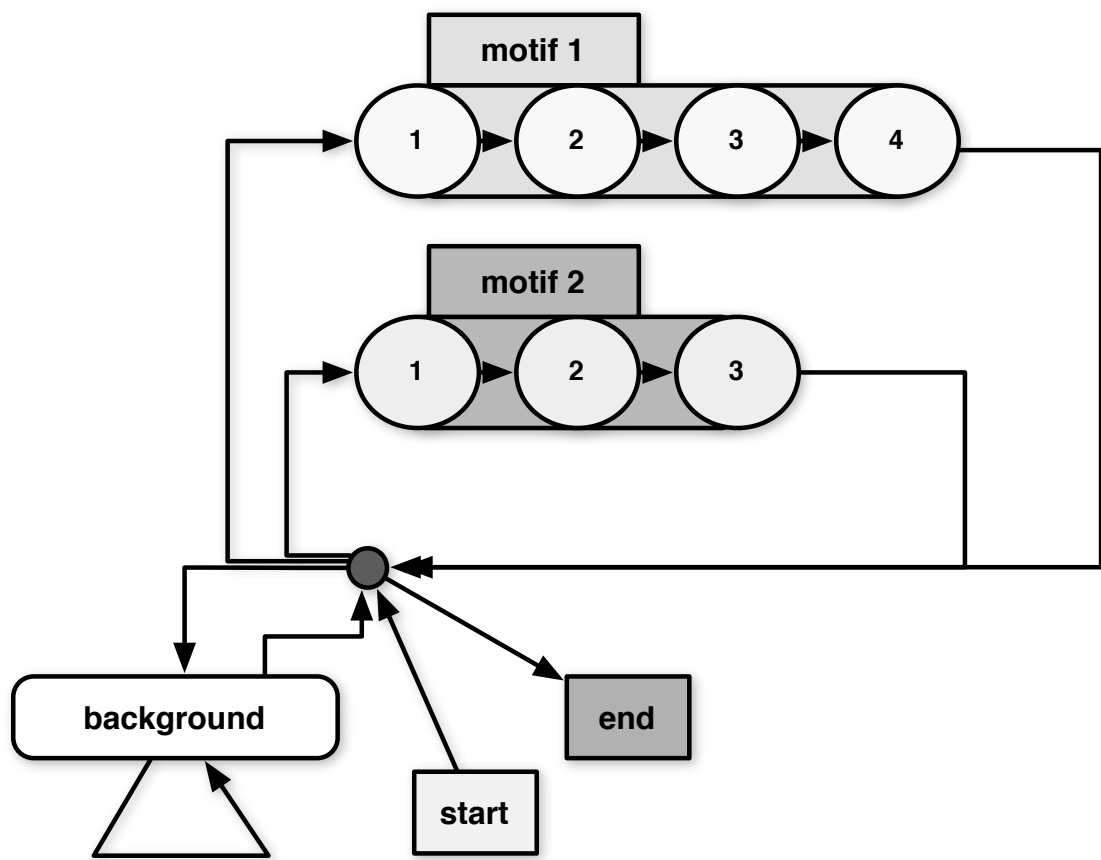


Figure 1.6: The multiple-uncounted sequence-motif mixture model (MUSMM).

the parameter estimation for the next, less strong motif. A greedy motif learning that requires repeated masking of sequence will incur an unpredictable sensitivity drop when multiple motifs are inferred: less and less sequence which is masked based on previously predicted is available for subsequent iterations. As I will show in Chapter 2, the metamotif inference framework I have developed also uses an analogous design to the NestedMICA algorithm to allow multiple metamotifs to be inferred simultaneously, with what I call the multiple-uncounted motif-motif mixture model, or MUMMM.

1.3.2 Nested sampling

Inference of parameters for Bayesian probabilistic models is often difficult, particularly for high dimensional models that are common in biology. Analytical solutions are almost always intractable. Most commonly approximate solutions are estimated using different Monte Carlo (MC) sampling techniques. I will below describe a state-of-the-art MC method, called nested sampling. Nested sampling is an MC technique originally introduced by [Skilling \(2004\)](#), and it is used in the metamotif inference algorithm I discuss in Chapter 2, as well as the NestedMICA motif inference algorithm which I expand in Chapter 3, and use for a large motif inference problem in Chapter 5.

As described by [Dogruel et al. \(2008\)](#), nested sampling is a MC method applied to an ensemble of e solutions (e typically ranges in hundreds to thousands). A nested sampler is firstly initialised with samples drawn from the prior distribution of states. After sampling, states are sorted by their likelihood and the member with lowest likelihood is removed from ensemble and replaced with a new sample, with the constraint that the new state has a higher likelihood than the removed state (Figure 1.7).

Samples are drawn from the prior distribution subject to the constraint that the likelihood of the new state must exceed that of the discarded state. This is done initially with rejection sampling ([von Neumann, 1951](#)), but after a certain number of iterations (the number of which is decided dynamically by measuring the rejection rate of the proposals), new samples begin to be generated with MCMC moves from other members of the ensemble because simple rejection sam-

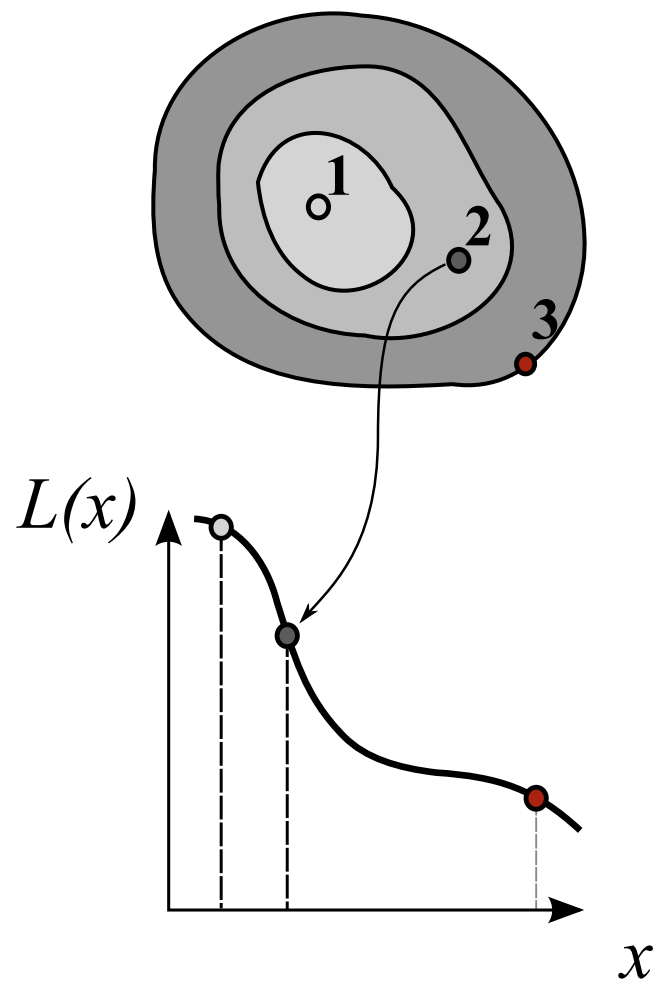


Figure 1.7: The likelihood contour. Lowest likelihood state is removed and a new state sampled on every iteration.

pling from prior with this increasing constraint becomes progressively ‘harder’ as the minimum likelihood threshold increases. As the sampling progresses through repeated iterations (typically in the range in tens to hundreds of thousands of iterations), more and more prior mass is excluded and the sampler reaches higher likelihood regions of the space. This is in a way analogous to simulated annealing, except progress occurs automatically without applying a temperature gradient to ‘heat’ or ‘cool’ the process (assuming that there are no complete plateaus in the space). Notably, nested sampling has demonstrated good performance in avoiding strictly local optima (Mukherjee and Parkinson, 2006; Shaw et al., 2007; Vegetti and Koopmans, 2009), unlike for instance Gibbs sampling which is a common MC strategy in motif inference. The fraction of prior mass removed from consideration at step t tends towards W_t (Equation 1.5).

$$W_t = \frac{1}{e} \left(\frac{e}{e+1} \right)^t \quad (1.5)$$

A particular strength of the nested sampling technique is that it allows direct estimation of the Bayesian evidence of the model, something which Monte Carlo methods do not traditionally do. Assuming that the likelihood of states removed at step t is approximately equal at L_t , the Bayesian evidence Z of the model can be estimated as described in Equation 1.6.

$$Z = \sum_{t=1}^{\infty} W_t L_t \quad (1.6)$$

The estimate of Z becomes progressively more accurate as sampling progresses, and indeed Z can be used for comparing models (motif set models derived with different input parameters for instance can be assessed by their Bayesian evidence). Furthermore, change in the evidence estimate Z_t (evidence at step t) is the criterion used for terminating the sampling (Equation 1.7). This same criterion is used with the DNA, protein and metamotif samplers in the NestedMICA suite (Dogruel et al., 2008; Down and Hubbard, 2005; Piipari et al., 2010a).

$$\frac{1}{Z_t} L_t \left(\frac{e}{e+1} \right)^t < 0.01 \quad (1.7)$$

1.3.3 The NestedMICA algorithm

NestedMICA applies nested sampling to motif inference, using an independent component analysis (Comon, 1994) like formulation of the motif inference problem: input sequences are modelled as a mixture of a number of independent motif signals and random noise (the background model). As described by Down and Hubbard (2005), in linear ICA, a matrix of observations X is approximated as a linear mixture A of some sources s and a noise matrix ν :

$$x = As + \nu \tag{1.8}$$

The noise matrix ν represents errors in the linear approximation. A commonly described example application of ICA is the “cocktail party problem”: a set of M microphones record different mixtures of the voices of N speakers. Given samples from these microphones at t time points, ICA methods attempt to factorize the $M \times t$ observation matrix into an $N \times t$ source matrix and an $M \times N$ mixing matrix. One can map the motif inference problem to an independent component analysis like formulation where the observations are a series of nucleotide strings, the sources are short sequence motifs, and a sequence background model represents the random noise. The mixing operation in motif ICA however is not simply a matrix multiplication.

The simplest mixing operation, and the one used by default, is simply a binary weighting: a motif has either a zero or ‘full’ weight in contributing to the likelihood of a sequence. That means that the mixing matrix (depicted in Figure 1.8) informs for each motif and sequence pair if a motif is expected to be a match in the sequence, according to a MUSMM-like sequence mixture model (Figure 1.6, where there are two motifs in the sequence with a nonzero weight). More complex mixing matrices, such as logistic function based weighting, are also included in the NestedMICA suite.

The model parameters – the motifs and the mixing matrix which describes pairing of motifs to sequences – are estimated with the nested sampling strategy (Section 1.3.2). Nested sampling allows inference to be made without heuristics to provide local starting points for motif search. Similarly, repeated runs of the algorithm are unnecessary, unlike with the commonly used Gibbs sampling Smith

(1987) based motif inference algorithms pioneered by [Lawrence et al. \(1993\)](#), or greedy expectation maximization ([Dempster et al., 1977](#)) based algorithms such as MEME ([Bailey and Elkan, 1995](#); [Bailey et al., 2006](#)). A schematic of the motif ICA and nested sampling, is provided in [Figure 1.8](#).

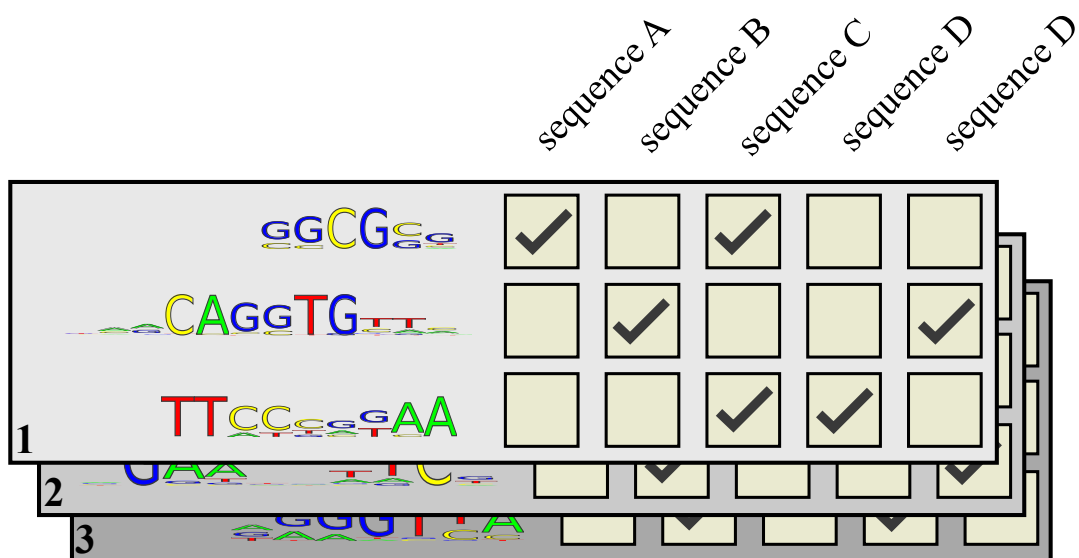


Figure 1.8: The NestedMICA model components: the motif set and the mixing matrix. An ensemble of three states is shown (states labelled 1,2,3).

A realistic model of the genomic sequence is also a key consideration with motif inference algorithms. The sequence background model in these algorithms is commonly modelled with a stationary Markov chain, and therefore depending on the order of the Markov chain it is parameterised simply by the nucleotide, dinucleotide, ... frequencies of the sequence. Real promoter sequence however is not uniform, and instead contains, for instance, discrete regions of GC-richness and AT-richness. NestedMICA uses a sequence background model that allows for compositionally distinct regions, for example the variation in GC content that is known to occur on multiple scales ([FitzGerald et al., 2006](#); [Thompson and Rouchka, 2003](#)). In addition to simply varying GC content, dinucleotide content can also be used to subdivide promoters according to their CpG content to two groups: those with exceptionally high frequency of CpG dinucleotide content,

and those with average genomic CpG content (Saxonov et al., 2006). Other regional biases in di-, tri- and tetranucleotides have also been described (Burge et al., 1992). The NestedMICA background model is referred to as mosaic to highlight its capability to describe sequence as a mixture of multiple generative processes (Markov chains). Use of multiple Markov chains, or ‘classes’, that are weighted per sequence position, improves the capacity of the background to describe compositional biases and is a considerably less complex model than higher order Markov chain backgrounds which are commonly used in motif inference algorithms.

A recently published motif inference algorithm BayesMD, which similarly as NestedMICA applies a Monte Carlo sampling method that is resilient to local maxima (parallel tempering, Gregory (2005)), and a sequence background model related to NestedMICA but trained from a larger selection of noncoding sequences, improves sensitivity over MEME, Align-ACE, MDScan, and also against NestedMICA in most benchmarks (Tang et al., 2008).

NestedMICA has been implemented in the Java programming language in a modular fashion where the definition of the model and the nested sampling framework are separate. As I will show in Chapter 2, this has made it possible to replace the NestedMICA motif model (the PWM) with a different space of models, and to therefore allow applying the nested sampling algorithm in the space of motif family models I have termed ‘metamotifs’. Furthermore, using existing nested sampling framework has also had the benefit of high runtime performance and scalability: the original NestedMICA algorithm and my variants of it make use of multiple CPUs when available, and the computational load can be distributed over multiple computers, scaling to up to 40 CPU cores (unpublished data).

1.3.4 Random forest classification

Supervised machine learning techniques aim to build a function based on input training data to predict the state of a response variable (the output). The response can be either continuous, at which case the procedure is called regression, or discrete, at which case the procedure is called classification. The function

should fit the training closely, but it should also generalise to other unseen data (Bhaskar et al., 2006). A compromise therefore needs to be made between a function which memorises the feature value combinations from training data but is incapable of generalising it to new input (an effect often referred to as ‘overfitting’), and one which generalises but is not necessarily able to fit all the training examples (training error). In Chapter 4 I use a supervised machine learning technique called random forest classification to learn the mapping from a motif (PWM) to the likely DNA binding domain which binds it.

A random forest is an ensemble machine learning technique, meaning that the classification function itself is a function of a number of independent classifier functions. The technique can be applied to either regression or classification, but we will concentrate on random forest classification, as regression techniques were not used in this work. According to Breiman (2001b), random forests follow in the line of three types of ensemble classification techniques noted below, all acting on ensembles of classification trees. Any of the three methods noted below are also sometimes confusingly referred to as a type of random forest.

1. “Random subspace” methods, where randomness is applied to subsets of features to use to grow trees (Ho, 1998).
2. Bagging methods, where randomness is applied to the choice of training data examples used to grow classification trees (Breiman, 1996).
3. A method where the splits made at tree nodes are made randomly according to voting (Dietterich, 1998).

The common factor between all of the above methods is that for the k^{th} classification tree, a random vector θ_k is generated independent of past vectors $\theta_1, \dots, \theta_{k-1}$ but with the same distribution (i.i.d.); A tree is grown using the training set (or its subset) and θ_k , resulting in a classifier $h(\mathbf{x}, \theta_k)$ where \mathbf{x} is an input vector. The nature of θ varies between the different tree construction methods. For instance, in bagging it can thought to be generated as the counts in N boxes resulting from N darts thrown at random at the boxes, where N is number of examples in the training set.

In Breiman’s random forest, each θ_k is trained from random selection of features from a subset x_k of bootstrapped examples in x (Equation 1.9) (Breiman, 2001b). Each x_k are taken from roughly two thirds of the examples, and the rest are used for the so-called out-of-bag error estimates (see below).

$$\{h(\mathbf{x}, \theta_k), k = 1, \dots\} \tag{1.9}$$

The set of i.i.d. random vectors noted above are noted as θ_k . In a classification problem, a random forest is a collection of decision tree predictors, and the response value is simply chosen by popular vote for the most popular label from the ensemble of k trees (the ensemble is referred to as a ‘forest’). The relative frequency at which the winning vote was made in the ensemble gives a confidence estimate for the decision. In regression the response value is the average of the response values in the forest.

A random forest classification has a number of attractive properties as a generic supervised machine learning framework:

1. An unbiased generalisation error estimate is made without the need for separate cross validation. This is achieved by leaving approximately one third of the training data x out from the bootstrapped examples x_k and they are labelled with the k^{th} classification tree. The error rate of this classification is the out-of-bag (oob) prediction error rate.
2. Its generalisation error tends to perform comparably to SVMs (Meyer et al., 2003) and favourably to related ensemble methods such as Adaboost (Freund and Schapire, 1996) or bagging.
3. It is naturally suited for multiclass problems (such as the motif domain labelling problem in Chapter 4), and provides a confidence estimate for the classification decisions regardless of the dimensionality of the class variable.
4. It is simple to understand, and provides insight into the importance of different classifier features (several kinds of proximity measures of training examples can also be computed). This is in contrast with kernel methods whereby variable importances are not straightforward to derive or interpret when one needs to resort to nonlinear kernel functions (usually for

improved classification performance), or multiclass classification. The variable importance measure in Breiman's `randomForest` package (Liaw and Wiener, 2002) which I use in my project is based on permutation testing: for each tree, all values of the m^{th} predictor variable are permuted, classification is made, and internal error rate computed as normally. The difference between correct (unpermuted) and incorrect (permuted) classifications is then computed and averaged over all trees, and normalised by the standard error. The margin is defined as the proportion of votes for true class minus maximum proportion of each of the other classes.

5. Although several adjustable input parameters are made available, only one is generally adjusted (`mtry`, which denotes the number of variables randomly sampled as candidates at each split), values of which the classification is also often robust to (Breiman, 2001a; Liaw and Wiener, 2002). This is in contrast with kernel method based classification, where a grid search of kernel function parameters is always necessary.

1.4 Biological datasets and resources

The most important biological datasets and resources which I have made use of during the course of my project are introduced below. Additional resources used in individual analyses are described in later chapters.

1.4.1 Ensembl

Ensembl is an open access database which provides access to eukaryotic genome sequence and annotation (Birney et al., 2004; Hubbard et al., 2009). Originally developed for analysis of the human genome, the current release 58 now contains 49 annotated eukaryotic genomes. The genome annotations provided by Ensembl are a key resource for large scale regulatory motif inference studies. For instance, all promoter sequences used for predicting motifs in my project have been chosen dependant on the transcription start site predictions provided by Ensembl. The resulting promoter regions are therefore a result of a combination of computational predictions and manual curation. Similarly, masking protein coding

sequences and sequence repeats is made possible by annotations retrieved from Ensembl.

In addition to the web site at <http://www.ensembl.org>, Ensembl offers programmatic access with a publicly supported Perl API (Stabenau *et al.*, 2004). Several other language specific APIs unsupported by the Ensembl project have also surfaced, including Ensembl Core for Ruby ¹ and `biojava-ensembl` ². Both of the above unsupported libraries proved useful in my work, and in the course of my project I in fact developed simple Ensembl database backed tools on top of `biojava-ensembl` for regulatory motif inference oriented tasks, which in turn were used in all of the peer-reviewed, published work which I have taken part in (Lewis *et al.*, 2009; Piipari *et al.*, 2011, 2010a,b), and Murray *et al.* (in press). See Section 5.2.1 and Appendices A, B for more detail.

1.4.2 Regulatory motif databases

Experimentally determined regulatory motifs have been another key resource in my work, both with motif family classification (Chapter 4) and validation of *de novo* inferred motifs (Chapter 5). The different TFBS motif databases I have resorted to in my work, and the rationale for choosing the individual datasets for analyses, are summarised below.

The regulatory genomics community suffers at the moment from the absence of a single authoritative database, data format, or minimal publishable requirements for distributing experimentally validated regulatory motifs or associated metadata (e.g. species information, experimental method). This is in notable contrast to for instance sequence, protein structure, or gene expression microarray data, each data type of which is generally required to be made available in a public database upon publication in a peer reviewed journal. TFBS motif data is scattered between individual publications, several databases in different partially overlapping subsets, and the standard of data and curation quality varies.

¹<http://www.github.com/jandot/ensembl>

²<http://www.derkholm.net/svn/repos/biojava-ensembl>

1.4.2.1 TRANSFAC

Currently the largest single dataset of eukaryotic TFBS motifs is contained in the TRANSFAC database, which is a commercial, curated database of eukaryotic gene regulation maintained by BIOBASE Ltd (Matys et al., 2006; Wingender et al., 2001). TRANSFAC contains a curated set of TFs, known TF–target gene regulatory relationships, and TFBS motifs as position frequency matrices (PFM). Most of the TFBS data stored in TRANSFAC originates from individual small-scale studies, including electrophoretic mobility shift assays (Fried and Crothers, 1981b; Garner and Revzin, 1981), DNase I foot-printing (Brenowitz et al., 1986), immunoprecipitation (Hecht and Grunstein, 1999) and some from higher throughput approaches such as *in vitro* selection (SELEX) (Oliphant et al., 1989). The more recently released TRANSFAC versions have begun expanding the database with ChIP-seq and various other HT methodologies discussed in Section 1.1.2.

TRANSFAC also defines its own structural taxonomy for classifying TF motifs by the structural class and family of binding TF. The structural taxonomy is largely similar on the level of TF domain families to the taxonomy used in the JASPAR database (Section 1.4.2.2), but the coarser level of the hierarchy (‘superfamilies’ in the TRANSFAC terminology, ‘structural classes’ in the JASPAR terminology) differs both in the divisions of TF domains and the terminology used.

The species covered by TRANSFAC are primarily vertebrates. Other animal TFs, as well as some plant and fungal TFs are included but in smaller scale. For instance, the database release 12.2, which my analysis in Chapter 4 is based on, contains a mere 38 motifs annotated with the species *S. cerevisiae*, and the same number of motifs for *Arabidopsis thaliana*, 68 for *D. melanogaster*, but 409 for mouse and 455 annotated with *H. sapiens*.

Due to the license fee associated with TRANSFAC, and its closed nature, an open access alternative to TRANSFAC could be beneficial for the research community. Attempts have been made to create alternatives, the most interesting being perhaps ORegAnno (Griffith et al., 2008), which is a community curation based database of transcriptional regulation. The ORegAnno dataset however has unfortunately not progressed to a form that is usable for most researchers.

The JASPAR database (Section 1.4.2.2), which similarly to TRANSFAC relies on a dedicated team of curators, has perhaps the best potential in providing an alternative to TRANSFAC’s collection of TFBS motifs.

I made use of TRANSFAC motifs for the motif domain family classification analysis conducted in Chapter 4 primarily because it allowed a direct comparison to previous motif classification methods MotifPrototyper (Xing and Karp, 2004) and SMLR (Narlikar and Hartemink, 2006), and because at the time it contained a considerably larger training and cross-validation dataset than the open-access alternative JASPAR: TRANSFAC 12.2 contained 848 structurally classified motifs (Wingender, 2008) versus a total of 138 in JASPAR 2008 (Bryne et al., 2008). In Chapter 5 I however describe more recent work where I built a motif family classifier based on the most recent JASPAR release, which has been expanded to include for instance many of the high-throughput datasets noted in Section 1.1.2.

1.4.2.2 JASPAR

JASPAR is another commonly used database of TFBS motifs (Bryne et al., 2008; Portales-Casamar et al., 2010; Sandelin et al., 2004). JASPAR distinguishes itself from TRANSFAC in several important aspects:

1. The structural terminology of TF domains, which covers most of its motifs, differs from that of TRANSFAC. JASPAR uses a two-level DNA binding structural mode taxonomy introduced by Luscombe et al. (2000). This classification terminology extends an earlier taxonomy created by Harrison (1991) on a smaller number of crystal structures. The Luscombe et al. (2000) taxonomy describes ‘classes’ and ‘families’ for TFs. Classes are defined by a manual, visual comparison of structures and families by a computational clustering of the domain structures with the SSAP secondary structure alignment algorithm (Orengo and Taylor, 1996). The taxonomy in TRANSFAC extends to more detailed levels, but past the class and family-like levels appears to be defined on a rather *ad hoc* basis by the TRANSFAC curators based on the terminology introduced in literature.
2. The data is open access, and its curation is of high quality. Key annotations such as species, experimental method and primary publications which

describe the data in the database are included almost with no exceptions, unlike TRANSFAC, where for instance only 490 of the 848 records contain a reference to a peer reviewed publication.

3. JASPAR, unlike TRANSFAC, is a non-redundant database, and aims to cover different kingdoms of life with separate non-redundant datasets (currently for mammals, insects, fungi and plants are covered). This is an important effort because of the lineage specific expansion of TF domains (Wilson et al., 2008a): TF domains utilised preferentially by different kingdoms of life differ substantially (discussed in 1.1.2).
4. JASPAR 2010 contains a near to complete non-redundant motif dataset of 177 *S. cerevisiae* motifs, compared to only 38 *S. cerevisiae* motifs in TRANSFAC 12.2 which emphasises vertebrate genomes.

I used the JASPAR database in Chapter 5 to train a motif family classifier to assess computationally inferred *S. cerevisiae* motifs most importantly because of the last two points above; for an accurate organism specific classifier it is important to have a good coverage of the TF domains that are specific to the lineage being studied. For instance, there are 47 known TFs with the fungal specific zinc cluster domain (Macpherson et al., 2006) in the *S. cerevisiae* genome out of the total 99 *S. cerevisiae* zinc finger motifs. TRANSFAC 12.2 includes motifs for only 9 of them, whereas JASPAR 2010 contains 38.

1.4.2.3 UniPROBE

UniPROBE (Universal PBM Resource for Oligonucleotide Binding Evaluation) is, as the name suggests, a database containing protein binding microarray derived motifs. At the time of writing, the database included motifs for 391 proteins from eight different studies, originating from affinity tagged TFs from human (Berger et al., 2006; Scharer et al., 2009), mouse (Badis et al., 2009; Berger et al., 2008), *C. elegans* (Grove et al., 2009), budding yeast (Zhu et al., 2009), the parasites *Malaria falciparum* and *Cryptosporidium parvum* (Silva et al., 2008), as well as the Gram-negative bacterium *Vibrio harveyi* (Pompeani et al., 2008). Its focus is simply to provide a repository for downloading and searching raw PBM data,

and PWM models derived from the data with the Seed-and-wobble algorithm (Berger et al., 2006). It does not attempt to provide a rich annotated reference database of TFBS motifs, like JASPAR or TRANSFAC. I have used two motif datasets from the UniPROBE database:

1. The 168 mouse homeodomain TF motifs by Berger et al. (2008). This dataset is one of the two high-throughput studies published in 2008 of the developmentally important homeodomain TFs, in addition to the bacterial one-hybrid dataset of *D. melanogaster* homeodomain TFs (Noyes et al., 2008a). The Berger et al. (2008) dataset covers 65% of the 260 known homeodomain proteins in the mouse genome. I apply both of the above mentioned homeodomain datasets in Chapter 4 for evaluating the capacity of the **metamatti** classifier in distinguishing homeodomain motifs from members of five other common TF domain families.
2. The 89 *S. cerevisiae* TF motifs (Zhu et al., 2009). This study provides the largest protein–DNA interaction dataset recovered with a single methodology, and it is therefore a convenient comparison dataset for comparing *ab initio* predicted regulatory motifs with. The slightly larger study by Berger et al. (2008) covers 112 yeast TFs, with a combination of different high-throughput methods.

1.5 Contributions of this thesis

My goal in this dissertation is to, firstly, present a new probabilistic model for familial relationships between regulatory motifs (Chapter 2). I then apply this familial motif model to sensitively infer motifs from novel sequence (Chapter 3), and to predict the DNA binding domain responsible for binding different regulatory motifs (Chapter 4).

Finally, I conduct a *de novo* motif inference study of the budding yeast genome to infer a large regulatory motif set from its promoters with a number of commonly used motif inference tools (Chapter 5). This is done primarily to assess the ability of the different motif inference tools to discover motifs that are consistent with previously known motifs from this particularly well studied eukaryotic regulatory genome.

Chapter 2

Metamotifs - a generative model for building families of nucleotide position weight matrices

2.1 Background

¹ A fundamental difficulty in studying DNA specificity of TFs is the absence of a simple, universal recognition code from the protein sequence or tertiary structure of the TF to its DNA recognition motif (Smith, 1998). Comparative studies of TF domains and their crystal structures with bound DNA have shown certain recurring rules for protein-DNA interactions (Jones et al., 1999; Kono and Sarai, 1999; Nadassy et al., 1999), for instance commonly occurring hydrogen bond mediated interactions between the base guanine, and arginine, lysine, histidine or serine residues (Luscombe and Thornton, 2002). However, the stronger patterns predictive of DNA specificity of proteins are highly TF domain family specific (Kono and Sarai, 1999; Luscombe and Thornton, 2002). That these interactions are domain specific, and sometimes non-additive (Badis et al., 2009; Benos et al.,

¹This chapter, and the two following two, were partly published in BMC Bioinformatics (Piipari et al., 2010a), by the author of this PhD thesis (MP), Dr. Thomas Down (TD), and my thesis supervisor Dr. Tim Hubbard (TH). Authors' contributions are as follows: TH, TD and MP conceived the work, MP developed the software, performed the tests and wrote the manuscript.

2002a), should not come as a surprise; protein and DNA interactions form a dynamic three dimensional network of contacts between the protein residues, the DNA sugar–phosphate backbone, bases and water residues in the binding interface (Luscombe and Thornton, 2002). Substantial conformational changes of both the protein and the DNA also often occur upon binding (Kim, 1995; Percipalle et al., 1995).

Even though a universal recognition code of protein DNA binding is unlikely to surface, familial patterns of DNA binding specificity can still be made use of to provide biological insight about newly presented data. The interaction rules of the DNA-binding residues are understood well in the case of some extensively studied domains like Cys₂His₂ zinc fingers (Wolfe et al., 2000). The DNA specificity of a Cys₂His₂ domain can be predicted based on sequence (Benos et al., 2002b; Kaplan et al., 2005; Mandel-Gutfreund et al., 2001; Persikov et al., 2008), and altogether new transcription factors can be engineered by mutating the DNA binding residues (Pabo et al., 2001). More interestingly from the point of view of my work, however, familial patterns of DNA specificity can be taken to infer TFBS motifs from genomic sequence with greater sensitivity. Several algorithms have been designed that take into account previous knowledge of TF domain DNA specificity to find motifs which fit familial patterns, or to label newly discovered motifs to TF families with classification methods (Narlikar et al., 2006; Sandelin and Wasserman, 2004; Xing and Karp, 2004).

2.1.1 Previous work on motif family models

The most widely applicable model for short regulatory motifs is the position weight matrix, or PWM (see Section 1.2.1), originally introduced by Stormo et al. (1982). Methods have been developed for comparing and clustering PWMs. The earliest such methods were made for protein domain model comparison (Pietrokovski, 1996). In the case of DNA motifs, clustering can be used to infer information about possible function of *de novo* predicted motifs, such as to find clusters of closely related motifs to known data. Although DNA binding domains vary widely, familial tendencies exist in DNA sequence motifs that are predictive of the family of transcription factors which bind them (Narlikar et al.,

2006; Narlikar and Hartemink, 2006). This makes clustering useful for inferring potential binding partners for discovered motifs of interest.

Familial binding profiles (FBP) offer perhaps the earliest solution for summarising familial patterns in nucleotide PWMs (Sandelin and Wasserman, 2004). FBPs are weighted averages of aligned sets of motifs. All motif pairs in the set are aligned with a variant of the Needleman & Wunsch global alignment algorithm (Needleman and Wunsch, 1970), using the score defined in Equation 2.1 to minimise the sum of squared deviations between the aligned motif columns amongst a familial alignment of PWMs, allowing for a single gap (with a stringent but arbitrarily chosen gap opening penalty). The significance of scores is measured with an empirical distribution of motif pair scores derived from shuffled motifs of the same length (Sandelin and Wasserman, 2004). Motifs are then added to a multiple alignment in the order of decreasing significance, and finally the motif columns are averaged, with contribution of each motif V weighted according to $w_V = 1 - p_v$, where p_v is the average of p -values of motif V with all the other motifs.

$$S = 2 - \sum_{b \in \{A,C,G,T\}} (M_b - N_b)^2 \quad (2.1)$$

FBPs for 11 metazoan transcription factor families are made available through the JASPAR motif database (Portales-Casamar et al., 2010). However, the FBP-based approach suffers from certain inherent limitations; Firstly it is not a probabilistic method but uses an arbitrary distance metric between motif columns, necessitating an empirical significance score computation and an arbitrary weighting of motif contributions to the FBP. Secondly, a global alignment is assumed between all motif columns, which means that only patterns common to all members of the family can be reliably modeled in this fashion. Sandelin and Wasserman (2004) only present FBPs for a small number of metazoan specific groups of DNA binding domains (11 FBPs, built from a total of 63 closely related motifs). Incidentally, many of the DNA binding domains in these 11 (e.g. ETS, Rel, MADS) have been classified as ‘highly specific’ to their DNA binding sites already by Luscombe et al. (2001), meaning that TFs in these families have a closely similar distribution of binding site specificities (motifs) with little variation.

More generally, motif comparison methods also suffer from the absence of a natural distance metric between motifs, although many different metrics have been proposed for this problem. For instance, a χ^2 -based distance metric was found an effective measure by [Kielbasa et al. \(2005\)](#). A metric based on Pearson correlations of motif columns was also described in the same publication. Various other distance metrics were suggested and systematically evaluated in a study by [Mahony et al. \(2007\)](#), where a sum of squared deviations based metric was found to be the best single metric. The asymptotic covariance between hits of two motifs in an infinitely long sequence parameterised by its nucleotide content has also been applied as a distance measure ([Pape et al., 2008](#)). The most recent motif distance metric and clustering methods are probabilistic and draw special attention to the uncertainty in motif comparison and the importance of high-information columns in measuring distances of sequence motifs: a Bayesian probability distance metric between motif columns ([Habib et al., 2008](#)) and a fuzzy integral based metric ([Garcia et al., 2009](#)). In this work I also explore a probabilistic solution for comparing motifs. Unlike any of the above motif-to-motif distance work, I however do not apply the developed method to a motif clustering problem. Instead, I attempt to solve the supervised learning problem of classifying motifs to their TF domain families probabilistically (Chapter 4). Classification based learning can be arguably more informative when predicting the likely function of motifs. This is because assigning a motif to a motif family has an associated uncertainty. Therefore finding closely similar known motifs by clustering does not always allow precise conclusions to be made regarding the binding partner of a discovered motif.

Supervised learning strategies have been applied to classify motifs and infer motifs similar to previously known motifs from novel sequences. Self-organising maps ([Kohonen and Somervuo, 2002](#)) have been applied for classification of binding sites for the purposes of semi-supervised motif inference in the SOMBRERO algorithm ([Mahony et al., 2005a](#)). Other notable methods include a Sparse Multinomial Regression (SMLR) based binding site sequence classification described in [Narlikar and Hartemink \(2006\)](#), and an application of this method to motif inference; The motif inference program PRIORITY assigns an SMLR-derived prior probability for each sequence position for its potential to fit a motif of a given

transcription factor family (Narlikar et al., 2006).

I present here a probabilistic model for describing motif families and measuring relatedness of sequence motifs – the metamotif. Metamotifs can be used to summarise gapless alignments of motifs of a given length, similar to an FBP. In contrast to the FBP framework introduced by Sandelin and Wasserman (2004), I do not model the recurring patterns found amongst a related set of motifs necessarily as a single motif alignment. Furthermore, the metamotif includes a vector of column wise mean nucleotide weights, as well as a variance parameter for each column. Variance is not modelled for example by the FBP or other non-probabilistic methods. Inclusion of motif column variances as part of the model makes it unnecessary to derive empirical significance estimates of motif similarity. In this respect a metamotif is similar to the hierarchical profile hidden Markov–Dirichlet multinomial model used by MotifPrototyper (Xing and Karp, 2004): both describe familiar prototypes of PWMs that are estimated probabilistically with a sequence of position specific probability distributions and can yield a Bayesian prior on the weight matrix columns (a ‘structural prior’ for the weight matrices in the terminology used by Xing and Karp (2004)). In contrast to MotifPrototyper, however, the metamotif inference algorithm I developed (Section 2.2.4) can account for intra-motif structure such as repeating or palindromic segments by treating motifs as a series of potentially several metamotif instances (i.e. learning several prototype patterns rather than only one), and positions emitted by a background model. In other words, in our framework, not all positions are generated from a single metamotif, and I additionally model some motif positions as noise emitted by a background model.

2.2 The metamotif

A metamotif is a generative model for PWM motif columns that can be used to represent a gapless alignment of position weight matrices. For each PWM position i (multinomial column) there exists a Dirichlet distribution in the metamotif column at position i . A metamotif is therefore a parameter configuration for a product Dirichlet distribution where position i of the motif alignment model corresponds to parameters α_i . More intuitively, consider that in a metamotif,

nucleotides at all positions have an associated average weight (depicted in Figure 2.1A as the symbol heights) and a variance (the error bars). It is in other words a probability distribution over PWM motifs of a given length. A metamotif of length k therefore allows drawing motifs of length k from it (Figure 2.1B), and querying for the probability of the metamotif being the source distribution for any motif of the same length. This is analogous to computing a probability score for a sequence k -mer to measure the probability of the k -mer having been generated by a PWM.

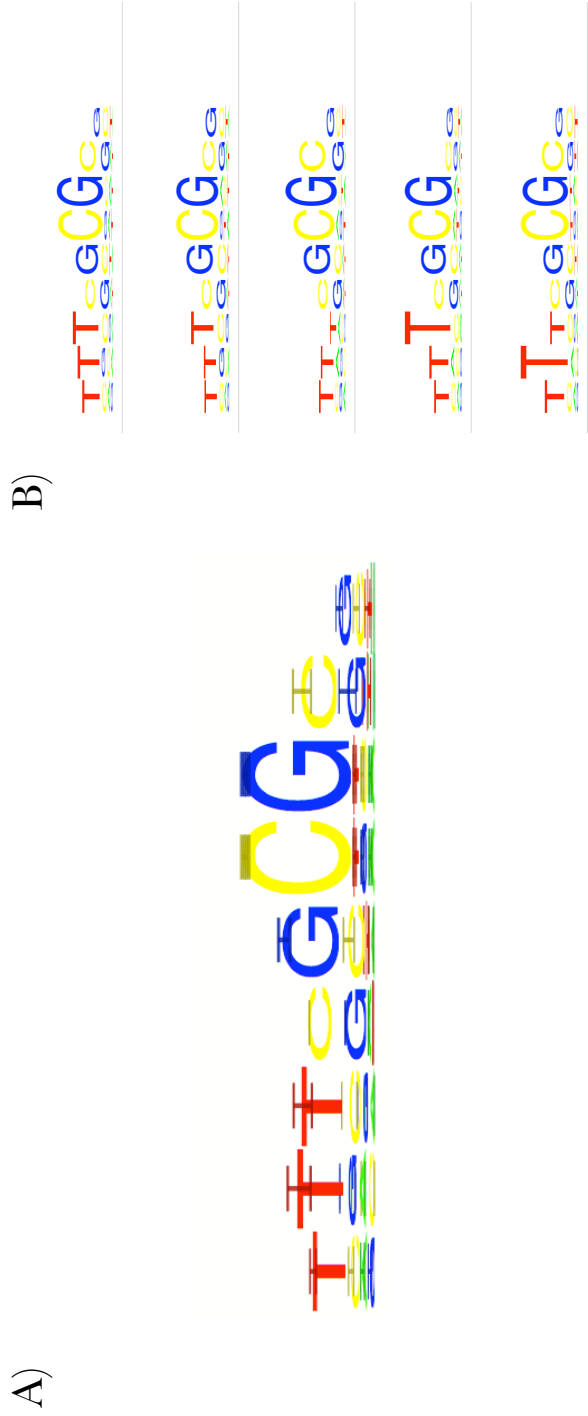


Figure 2.1: A forkhead-like metamotif (inferred from an alignment of motifs) is shown alongside selection of motif samples drawn from it. The wider error bars (representing 95% confidence intervals of nucleotide weights) of the thymine-rich 5' end of the metamotif is found consistent with the variation in the motif column heights. A) metamotif is a column-wise model with average nucleotide weights and variance associated per nucleotide column. B) A metamotif is a probability distribution from which motifs of the same length can be drawn.

Below I first formally define the metamotif (Section 2.2.1) and present a simple maximum likelihood method for estimating metamotifs from aligned motif data. In Section 2.2.2 I present a form of visualisation for the metamotif akin the sequence logo (Schneider and Stephens, 1990), and then expand the use of the model beyond simply constructing metamotifs from aligned motifs (Section 2.2.4). This expansion is made possible by a Monte Carlo metamotif inference algorithm that simultaneously estimates multiple weakly represented metamotifs from a potentially large set of motifs.

2.2.1 Formulation of the model

A metamotif α is a matrix of L columns, each defining a Dirichlet distribution over \mathbb{R}^K where K is the size of the alphabet (Equation 2.2).

$$\alpha = \begin{pmatrix} \alpha_{11} & \dots & \alpha_{1L} \\ \vdots & & \vdots \\ \alpha_{K1} & \dots & \alpha_{KL} \end{pmatrix} \quad (2.2)$$

A motif $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ is a set of column vectors over the same alphabet. The probability of observing the column \mathbf{x}_i from the metamotif α is given by the density of the Dirichlet distribution with parameters α_i at weights \mathbf{x}_i (Equation 2.3). The normalising constant $B(\alpha)$ is the multinomial beta function, expressed in Equation 2.4 via the Gamma function.

$$\mathbb{P}(\mathbf{x}_i | \alpha_i) = \text{Dir}(\mathbf{x}_i; \alpha_i) = \frac{1}{B(\alpha)} \prod_{j=1}^K x_{ij}^{\alpha_{ij}-1} \quad (2.3)$$

$$B(\alpha) = \frac{\prod_{j=1}^K \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^K \alpha_j)} \quad (2.4)$$

The log probability of observing a motif of length L is then given by Equation 2.5.

$$\log \mathbb{P}(\mathbf{X} | \alpha) = \sum_{i=1}^L \log(\text{Dir}(\mathbf{x}_i; \alpha_i)) \quad (2.5)$$

To motivate the use of the metamotif we note that the metamotif column α_i can be understood as a combination of the mean nucleotide weights $\mathbb{E}[x_{mk}]$ and precision $\alpha_{0m} = \sum_{j=1}^K \alpha_j$ (Equation 2.6) where $m \in [1, M]$ and $k \in [1, K]$.

$$\mathbb{E}[x_{ij}] = \alpha_{ij}/\alpha_{0j} \tag{2.6}$$

2.2.2 Visual representation of the model

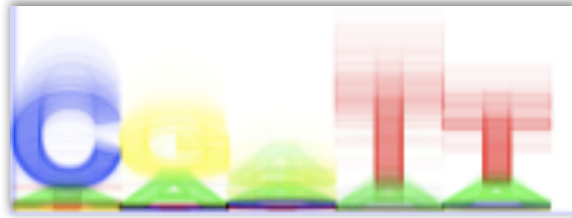
The visual representations I developed for the metamotif model are both based on the sequence logo (Schneider and Stephens, 1990). The metamotif visualisation was implemented as part of the iMotifs sequence motif visualisation environment (Piipari et al., 2010b) with Apple’s C-based Quartz API and the Objective-C based Cocoa drawing APIs. Metamotif model visualisation was in fact originally implemented in a Java based cross-platform motif visualisation tool mXplor, which I created as a precursor to iMotifs (available openly at <http://www.github.com/mz2/mxplor>). The representation evolved from a ‘fuzzy sequence logo’, where a number of sequence logos are overlaid on top of each other (Figure 2.2A), to a sequence logo with confidence intervals being drawn on the motifs (Figure 2.2B). Notably iMotifs supports both the error bar and fuzzy motif representations.

Both visual forms shown in Figure 2.2 communicate the mean weights $\mathbb{E}[\mathbf{X}|\alpha]$ and precision α_0 aspects of the metamotif. A sequence logo is drawn for PWM with nucleotide weights $\mathbb{E}[\mathbf{X}|\alpha]$. In the error bar enabled sequence logo in Figure 2.2B the error bars are shown to highlight 95% confidence intervals of nucleotide weights of the Dirichlet density at α_i for each symbol (Figure 2.2B).

2.2.3 Aligning motifs and estimating metamotifs from a motif multiple alignments

Given that a metamotif is a probability distribution over motifs of length k , it should be possible to estimate a metamotif from a series of aligned motif columns of matching length (see for example Figure 2.1B). Indeed, during my project I firstly designed a simple maximum likelihood metamotif inference algorithm for

A)



B)



C)

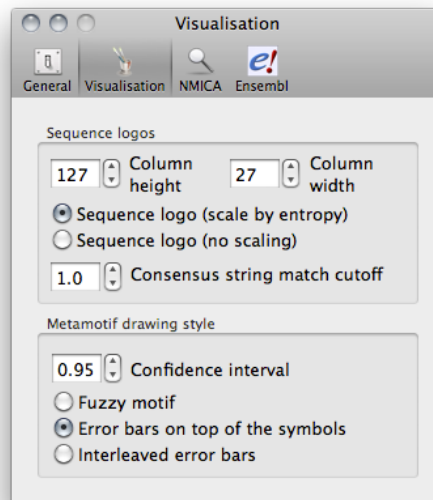


Figure 2.2: Visual representations of metamotifs. A) A ‘fuzzy sequence logo’ representation of a metamotif drawn with mXplor. One hundred samples are drawn per column, and sequence logos of the resulting PWMs are overlaid with low opacity on top of each other. The symbols in the sampled PWMs are ordered according to the decreasing nucleotide weight of the average weights in the distribution. B) Metamotif represented by a sequence logo with error bars (5% – 95% confidence intervals are presented with the error bars). C) The confidence intervals presented for a metamotif, i.e. the ‘height’ of the error bars in (B), can be configured in iMotifs.

the purpose. It is described in brief in Figure 2.3.

Firstly, a distance distribution is computed between the input motifs according to the column-wise sum of squared differences (SSD) motif distance metric from Down et al. (2007), which is noted below in Equation 2.7. P and Q are distributions from the two compared motifs, and ϵ is an adjustable modifier on the exponent. When it has the value 1.0, the distance computed is the Cartesian distance. Similar to Down et al. (2007) I use $\epsilon = 2.5$.

$$D(P||Q) = \left(\sum_{s \in A} (P(s) - Q(s))^2 \right)^{\epsilon/2} \quad (2.7)$$

When comparing the distance, all possible offsets with at least one overlapping column are considered between motif pairs (the unmatched columns are treated as a multinomial distribution with uniform nucleotide weights [0.25, 0.25, 0.25, 0.25]). Then, beginning from the closest motif pair, motifs are progressively added to the alignment, one by one in the order of increasing distance to motifs already present in the alignment. This is analogous to the progressive multiple alignment strategy used in many protein sequence multiple alignment algorithms (Chenna et al., 2003; Notredame et al., 2000). The resulting gapless alignment is simply defined by the offsets and reverse complement operations required to minimise the distance between the closest pairs (reverse-complementing motifs, i.e. allowing matches on either strand, is optional). Computing the metamotif is in fact simply a post-processing step done after aligning motif columns and cutting the motifs to a fixed length (Step 2 in Figure 2.3): a maximum likelihood Dirichlet distribution is computed using the Newton iteration method described in Minka (2003) due to the lack of a closed form solution. The motif set alignment algorithm which I implemented was also made to allow outputting an average PWM (a familial binding profile -like construct, see Section 2.1.1), or the alignment as a series of aligned motifs. All of these output options (a metamotif, an average motif, and an aligned set of motifs) are also available in iMotifs (Piipari et al., 2010b).

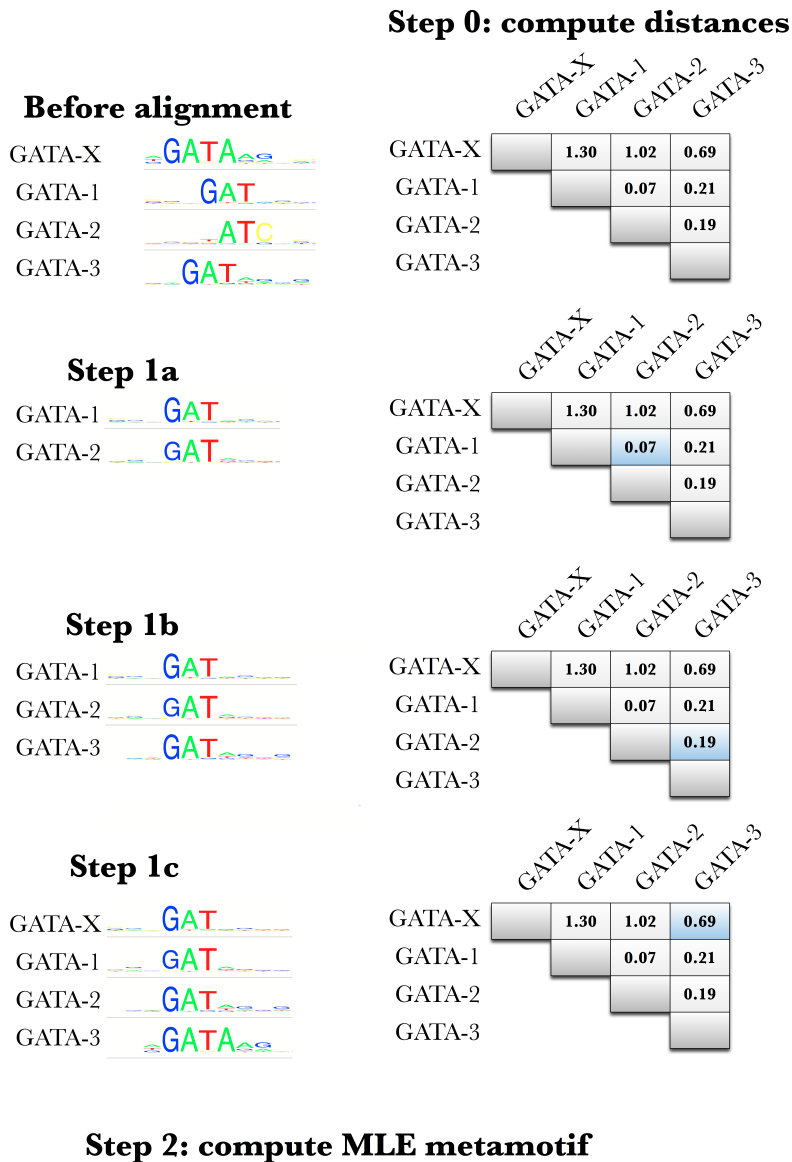


Figure 2.3: Schematic explaining the MLE metamotif inference algorithm. Firstly a distance matrix is computed between the motifs (Step 0). Motifs are added to the alignment in the order of increasing minimal pairwise distance to the motifs already in the alignment (steps 1a,1b,1c). Note that the motif GATA-2 is reverse-complemented upon adding it to the alignment. Motif ends are optionally cut such as to arrive at a motif alignment with no ‘hanging end columns’ (a minimum number of motifs with a supporting column can be defined to choose the threshold). A MLE Dirichlet distribution is then estimated for all motif columns using the method described in [Minka \(2003\)](#).

2.2.4 Metamotif inference by nested sampling

The metamotif can be seen as a way to summarise a gapless alignment of motifs of a certain length, to yield a probability distribution of motifs. However, my goal in designing the metamotif framework was to describe recurring patterns seen in sequence motif data deposited in public motif databases such as TRANSFAC (Matys et al., 2006), JASPAR (Portales-Casamar et al., 2010) or UniPROBE (Newburger and Bulyk, 2009). Many sequence motif families cannot be described accurately by global gapless multiple alignments of motifs at a fixed length. Motifs can for example consist of shorter repetitive signals, such as in the case of the heat-shock factor (HSF) motifs (Figure 2.4D), or the basic Helix-Loop-Helix (bHLH) motif family that are completely or partially palindromic due to their dimeric binding mode (Anthony-Cahill et al., 1992). Inspection of the HSF motif set shows that a global alignment of its columns does not describe the regularly spaced five-base repeat that is observed as part of the motifs in opposing orientations (aGAAn / nTTcT) (Kroeger and Morimoto, 1994). Furthermore, even non-repetitive and non-palindromic motifs present challenges for gapless multiple alignments: the span of informative columns contributing to familial patterns in publicly available PWM data is often unclear because of different signal-to-noise ratios and varying information content criteria used for calling motif ends.

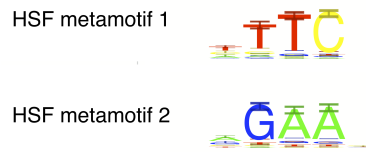
I wanted to develop an inference algorithm that allows simultaneous detection of n short metamotif signals from a set of motif data, allowing for varying length for different metamotifs, and optionally free orientation (signal present on either strand). The metamotif count n is a fixed, user settable parameter to the algorithm. For metamotif inference problems where n is expected to be large, the choice for the parameter should be informed by prior information of the motif set under study, for example clustering of the motifs to estimate a rough number of recurring motif segments. Each metamotif has *a priori* an unknown length between l_{min} and l_{max} columns, and is expected to contain one or more matches in a fraction f of motifs. Motifs in the framework are thought to be generated by recurring metamotif patterns, each of which is potentially shorter than any of the motifs, and background positions that model “uninteresting” sections of the motifs (positions not emitted by any of the metamotifs). The background

model in the framework is the maximum likelihood (MLE) Dirichlet distribution estimated from all the motif columns in the input data. It is computed with the optimisation procedure described in [Minka \(2003\)](#), which is also used in the simpler MLE metamotif inference algorithm described in Section 2.2.3.

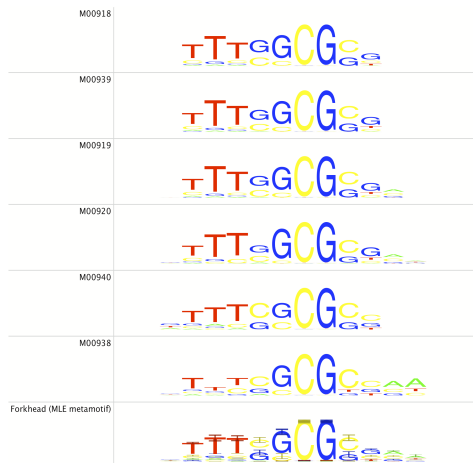
A) Forkhead metamotif (MLE)



C) HSF metamotifs (nested sampling)



B) Forkhead family motifs



D) HSF family motifs

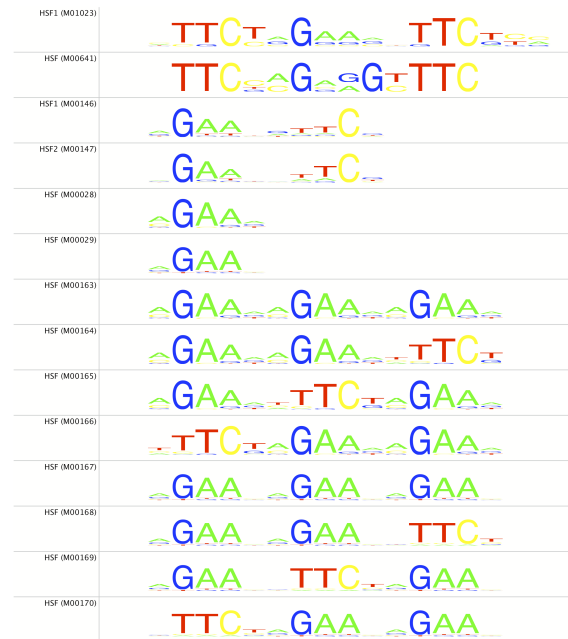


Figure 2.4: Example metamotifs for forkhead (A) and HSF (B) motif families from the TRANSFAC database ([Matys et al., 2006](#)). A) The MLE metamotif estimated for a subset of forkhead motifs (B) in the TRANSFAC 12.2 ([Matys et al., 2006](#)) regulatory motif database. C) Two HSF metamotifs estimated using the metamotif nested sampling algorithm from a subset of HSF motifs (D) in the TRANSFAC regulatory motif database.

The metamotif inference algorithm which I developed is a variant of the NestedMICA nested sampling algorithm described in Section 1.3.3. Nested sampling, originally introduced by Skilling (2004), is a generic Bayesian MCMC sampling strategy that allows drawing samples from a posterior distribution and directly estimating the evidence (marginal likelihood) of the model.

The metamotif nested sampler takes recurring intra-motif structure into account and allows detection of multiple metamotifs from a set of motifs. Motif sets are treated as a combination of short recurring patterns emitted by metamotifs, and background positions. The recurring signal can also optionally be allowed to be present on either strand, further improving the ability to detect repeating features. Recurring metamotif signals of interest are modelled separately from the “uninteresting” sections of the motifs that are taken as having been generated by a background model. The background model is the maximum likelihood (MLE) Dirichlet distribution estimated from all the motif columns in the input data. It is computed with the optimisation procedure described in Minka (2003), which is also used in the simpler MLE metamotif inference algorithm described in Section 2.2.3.

The algorithm allows estimating n metamotifs for a set of p motifs, with a variable metamotif length between l_{min} and l_{max} columns, and an expected fraction f of motifs containing any one of the n metamotifs. This is analogous to the NestedMICA motif inference algorithm that estimates multiple motifs with varying length from an expected fraction of nucleotide or protein sequence data. The posterior distribution being sampled is over the sets of n metamotifs and so-called mixture matrices, given the motif data and a background model for the motifs. The mixture (or occupancy) matrix describes the pairing between metamotifs and motifs. The term mixing matrix is a reference to the algorithm treating pattern recognition as an independent component analysis problem similar to the NestedMICA motif inference algorithm (Section 1.3.3): a likelihood function is written for the observations (the motif set) and the motif set is assumed to be generated as a mixture of independent metamotif contributions and noise represented by the background model. Each element $\mathbf{Q}_{i,j}$ in the $n \times p$ mixing matrix \mathbf{Q} is a binary indicator of the metamotif j being present one or more times in the motif i . If the metamotif is present, $\mathbf{Q}_{i,j} = 1$, otherwise $\mathbf{Q}_{i,j} = 0$. The likelihood of

the motif set given the metamotif set is simply the product of likelihoods of each individual motif given the metamotif set and the mixture matrix.

2.2.5 The likelihood function

The likelihood of a motif given a set of metamotifs is calculated assuming the motif is emitted by the multiple-uncounted motif–metamotif mixture model (a MUMM with two metamotifs is given in Figure 2.5). This formulation allows for each motif to contain multiple metamotifs simultaneously, without the need to iteratively repeat sampling after masking previously inferred stronger signals.

Computing the likelihood of a motif given metamotifs under the MUMM model involves completing one-dimensional dynamic programming from the beginning of the motif to column c , closely in the same form as the protein or nucleotide sequence likelihood function described for the NestedMICA algorithm in [Dogruel et al. \(2008\)](#) (Equation 2.8).

$$L_c = (1 - t)B_{c-1}L_{c-1} + \frac{t}{|M|} \sum_{\alpha \in M} \mathbb{P}(\mathbf{X}_{c-l_\alpha+1}^{c-1})L_{c-l_\alpha} \quad (2.8)$$

L_c represents the likelihood of all metamotif and background column arrangements (paths) in the input motif up to the column c . M is the set of metamotifs that have a mixing coefficient of 1 for the motif under consideration (i.e. metamotifs marked to be present in the motif in the mixing matrix \mathbf{Q}), and $|M|$ is the number of metamotifs that have a mixing coefficient 1. The length of the metamotif α is represented by l_α . B_c is the probability that the motif column at position c was emitted by the background. For the motif \mathbf{X} of length $l_{\mathbf{X}}$ the transition probability t to a metamotif is defined as $t = 1/l_{\mathbf{X}}$, i.e. one metamotif is expected per motif, and any motif position is equally likely to contain a transition. $\mathbb{P}(\mathbf{X}_i^j)$ is the probability that the motif segment from i to j was emitted by a metamotif m , and it is given by the metamotif density function (Equation 2.5). A metamotif can optionally be allowed to be present on either strand to improve the ability to detect repeating (e.g. palindromic) features. Alternating orientation of metamotifs are achieved simply by summing the probability contributions $\mathbb{P}(\mathbf{X}_i^j)$ of the metamotif α and its reverse complement at all possible

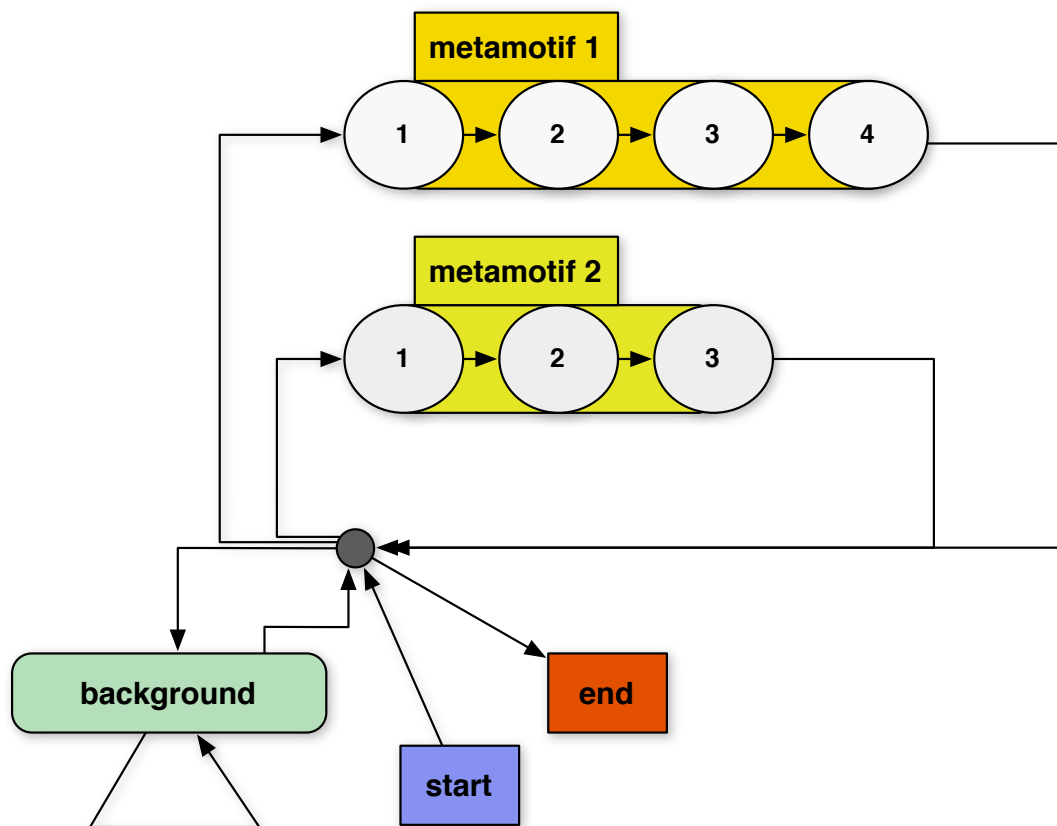


Figure 2.5: The multiple-uncounted motif-metamotif mixture HMM (MUMM). Numbered steps model the columns of the metamotif signals of interest and the background states are responsible for the “uninteresting” positions. Motif columns are emitted from a selection of metamotifs of varying lengths, and background positions. Note the similarity to the sequence-motif mixture model used in the NestedMICA motif discovery algorithm for motifs embedded in sequence (Figure 1.6).

offsets. Incomplete metamotif hits are also accounted for (Section 2.2.7).

2.2.6 Monte Carlo sampling moves

The metamotif nested sampler algorithm evolves metamotif parameters, and the mixture matrix state, with Monte Carlo sampling moves. Most of the proposal types alter the metamotif column parameters. The metamotif proposals are selected randomly from amongst the following set of moves:

- a small perturbation is made to a randomly selected metamotif column nucleotide mean weight: perturbation is made according to a randomly chosen nucleotide α weight α_i , nucleotide mean weights adjusted so they again sum to 1, and α_i of the column adjusted accordingly, maintaining precision unchanged.
- a small perturbation is made to a randomly selected metamotif column precision α_0 : α_0 is perturbed, and α adjusted such as to maintain the mean nucleotide weights unchanged with a new precision.
- a small perturbation is made to a randomly selected metamotif column nucleotide weight α_i , thereby indirectly changing the precision.
- replacing a metamotif column with a new one, sampled from an uninformative simplex prior (nucleotide weights on the range [0.1, 40.0] are allowed).
- removing a column in one end of a metamotif while adding another one to the other end.
- adjusting motif length, by adding or removing a column from either end.

The two update operations that use an alternative parameterisation of α with precision and the mean nucleotide weights, i.e. updating the precision whilst maintaining mean weights unchanged, and altering the mean weights whilst maintaining the precision unchanged, proved beneficial for achieving convergence of the algorithm. When these moves were included, the algorithm converged consistently with smaller number of iterations than when only the more naive method

of updating α_i with random perturbations was included (data not shown). The prior function over the Dirichlet distribution parameters was an uninformative 'clipped' simplex prior: all values for the nucleotide weight parameters α_i of the distribution are allowed on the range $[0.1, 40.0]$ and equally likely. Parameter values above or below this range are clipped such as to avoid numerical instability.

Sampling moves are also done in the space of mixture matrices by flipping states of randomly selected elements in the mixture matrix similarly as done in [Dogruel et al. \(2008\)](#) for the NestedMICA algorithm.

2.2.7 Accounting for incomplete metamotif hits

Accounting for incomplete metamotif matches in a motif is an important consideration. This is because we wish to analyse data from different experimental and computational sources where motif start or end positions have not been chosen consistently, for instance with an information content criterion. Incomplete hits are accounted for by adding additional "un-informative" columns in the input motifs in both the 5' and the 3' motif ends. The un-informative columns are multinomial distributions that match the mean nucleotide weights of the background model Dirichlet distribution. This effectively allows all possible offsets of the metamotif that overlap the motif with at least one column, whilst associating more uncertainty to those columns supported by only a subset of the motif data (Figure 2.6).

2.3 Evaluating the metamotif nested sampler algorithm

Performance of the metamotif inference algorithm was tested using synthetic motif sets where samples from metamotifs were inserted, or "spiked", similarly as done by [Dogruel \(2008\)](#); [Tang et al. \(2008\)](#) with synthetic sequences and samples from motifs. The aim was to measure the relative frequency of metamotifs at which the expected metamotifs could be recovered by the algorithm from synthetic motif data containing metamotif instances. The evaluations were done in two stages. The ability of the algorithm to infer a single metamotif presented to

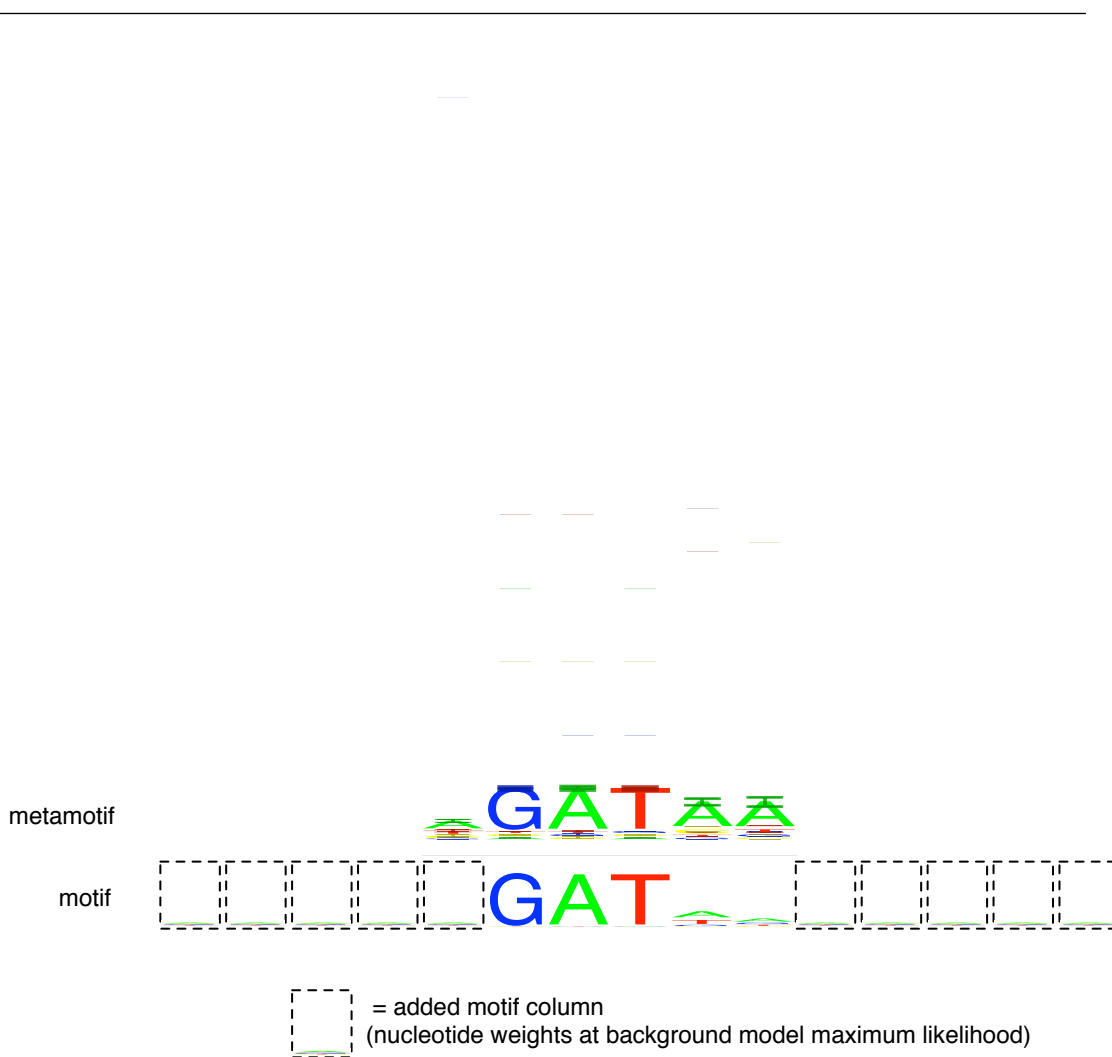


Figure 2.6: Incomplete hits are handled by padding the input motifs with additional columns that fit the background model optimally. All metamotif hits are required to be at minimum two columns long, which means that all input motifs are (optionally) capped with $l_{min} - 1$ additional columns, where l_{min} is the user settable minimum metamotif length parameter (which also has a minimum allowed value of 2).

it was tested first (Section 2.3.1). After that, several metamotifs were presented to the algorithm to assess the ability of the algorithm to infer multiple metamotifs simultaneously (Section 2.3.2). Metamotifs were then also inferred from the TRANSFAC database (Section 2.3.3).

To prepare the synthetic motif sets, metamotifs were first generated of examples of three structurally diverse TRANSFAC 12.2 PWM families: six forkhead motifs (class 3.3 in TRANSFAC classification), six GATA-like Cys₄ zinc finger motifs (class 2.1) and five MADS box motifs (class 4.4) were used (source motifs shown in Figure 2.7). This was done by aligning each of the three input motif sets with a greedy gapless sequence motif multiple alignment method related to the one utilised in STAMP motif toolkit (Mahony and Benos, 2007). A metamotif was then estimated from the motif multiple alignments with the MLE method from Minka (2003): MLE Dirichlet distribution was computed for motif alignment columns (example seen in Figure 2.4A), with each motif column in the alignment mapping to a MLE Dirichlet distribution in the resulting metamotif.

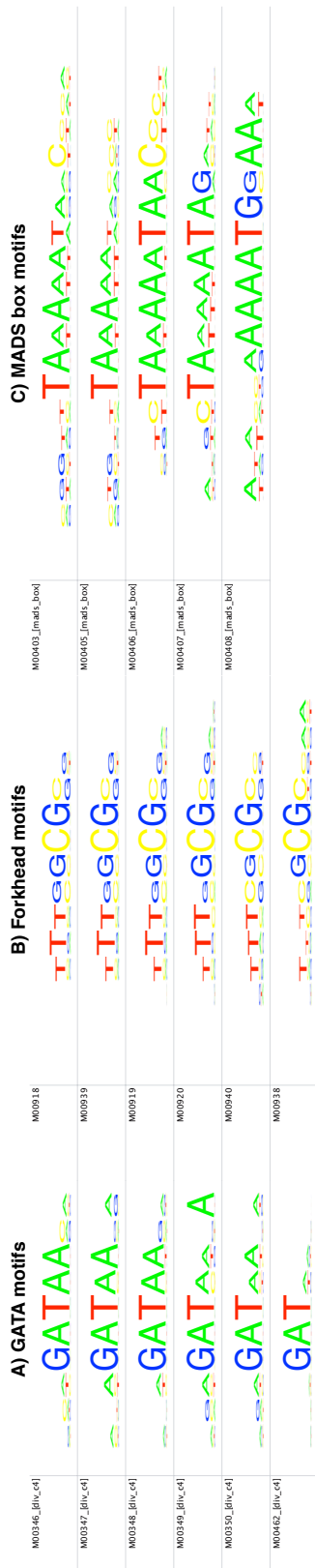


Figure 2.7: These motifs were aligned and the multiple alignment summarised as an MLE motif with the program *mmalign*. See the topmost motifs in Figure 2.8 for the resulting target motifs that were spiked into synthetic motif sets.

Motifs (PWMs) from each of the three familial metamotifs were sampled in relative frequencies of 0%, 10%, 20%, ..., 100%, into synthetic input motif sets (separate input motif set per motif family). Each synthetic motif set contained 60 motifs, each 20 nucleotide columns long, with a maximum of one metamotif instance allowed per input motif. The synthetic motif columns in the input motif sets are samples from a Dirichlet distribution with parameters $\alpha = \{0.5, 0.5, 0.5, 0.5\}$. The metamotif sample PWMs were inserted at random positions within the 20 nucleotide long synthetic motifs. The metamotif inference algorithm was then run on the motif set to infer a single metamotif between length ranges 4 and 14, allowing for the signal to be present in either orientation (`-numMetamotifs 1 -revComp -minLength 4 -maxLength 14`).

Metamotif inference performance was measured qualitatively with visual inspection comparing the inferred metamotifs to the known spiked metamotifs, and quantitatively measuring the Cartesian distance between the metamotif mean nucleotide weights.

2.3.1 A single metamotif

The metamotif nested sampler algorithm was used to infer metamotifs from the synthetic motif sets to evaluate how well the spiked metamotif patterns could be recovered. Performance was measured qualitatively with visual inspection comparing the inferred metamotifs to the known spiked metamotifs, and quantitatively measuring the Cartesian distance between the metamotif mean nucleotide weights. The visual comparison, Cartesian distances and empirical p -values for observed metamotif-metamotif distances are presented in Figure 2.8. The evaluation shows that metamotifs can be inferred from motif sets that contain them with relative frequencies of even 10%. At a relative frequency of 40% and above all three recovered metamotifs are very similar to the respective source metamotif (Figure 2.8).

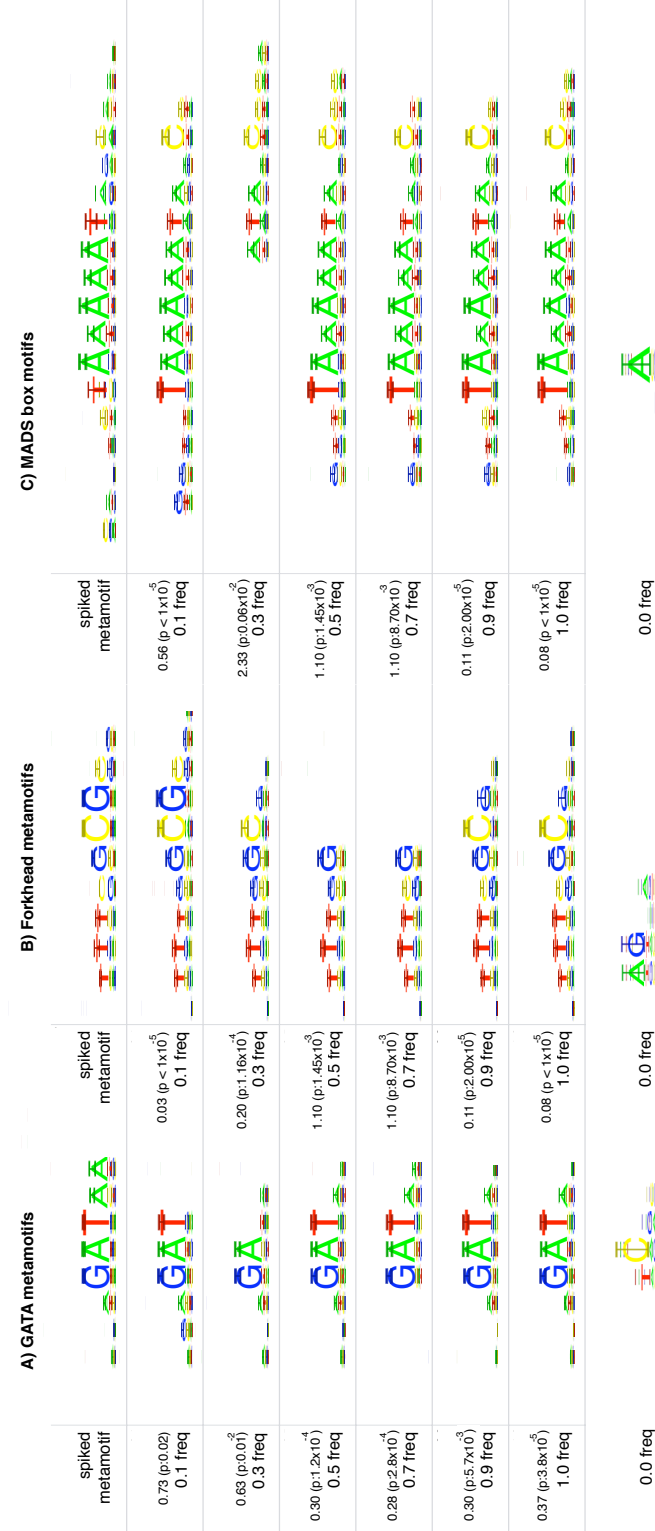


Figure 2.8: Metamotifs estimated with the metamotif nested sampler algorithm with varying relative frequency of metamotif samples. The top row in each metamotif alignment contains the “correct” metamotif that was sampled to the input weight matrix data in six different relative frequencies: 0.0, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0. Frequency 0.0 which is shown in the bottom of the graph refers to a control experiment where all columns of the motif set are samples from the background (a Dirichlet distribution with parameters $\alpha = \{0.5, 0.5, 0.5, 0.5\}$). A Cartesian-like distance between the sampled metamotif column mean nucleotide weights of the shown metamotif and the spiked metamotif mean nucleotide weights is presented above the relative frequency. An empirical p -value as described by (Down et al., 2007) is also shown for the Cartesian distances (100,000 shuffles made for each motif).

2.3.2 Multiple metamotifs

The ability to predict multiple metamotifs was demonstrated in a second evaluation experiment where instances of all the three motif families were inserted into synthetic motif sets and the algorithm was required to infer three metamotifs. It was shown that the algorithm was able to infer multiple metamotif models concurrently with correct lengths at a relative frequency as low as 20% (Figure 2.9).

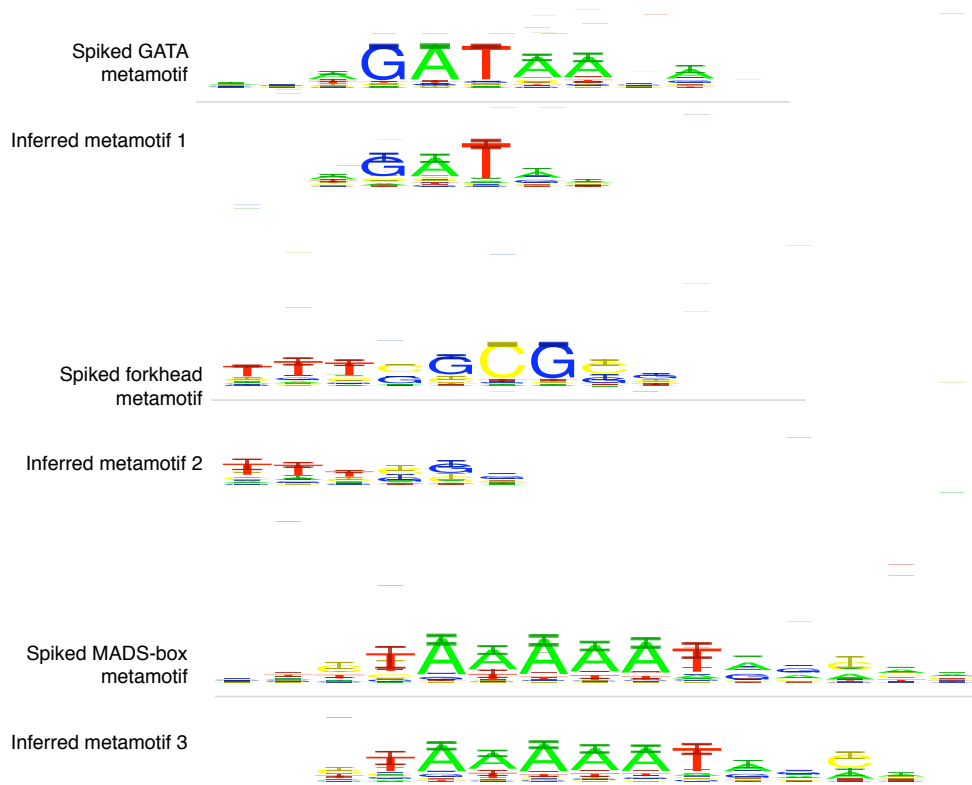


Figure 2.9: The metamotifs predicted at relative frequency of 0.2 are shown alongside the source metamotifs.

2.3.3 Inferring metamotifs from TRANSFAC

I demonstrated use of the metamotif nested sampling algorithm in inferring familial metamotifs from known experimentally determined regulatory motifs from the TRANSFAC database (Matys et al., 2006). Motifs retrieved from TRANSFAC were first divided to clusters with the SSD distance by Down et al. (2007) with cutoff 6.0. Three metamotifs were then inferred from each of the resulting clusters. Examples of metamotifs inferred are shown in Figure 2.10. The metamotif nested sampler algorithm was found capable of detecting several recurring patterns from the motif clusters that are clear upon visual inspection of the motifs, in addition to finding overlappers from the motif sets (Figure ??B).

Evaluation of the nested sampling based metamotif inference algorithm suggests that it is able to correctly infer familial metamotif patterns. It performs both in the case of a single recurring motif family, and in the case of motif sets with examples of multiple motif families. This makes it potentially applicable for instance for finding redundant motif patterns from large scale *de novo* inferred sets of motif predictions from different algorithms, or for inferring a complete set of familial metamotifs from a set of motifs. Metamotif inference is also conducted from clustered motifs from the TRANSFAC database.

2.4 Summary

In this chapter I introduce a generative model for PWM motif columns, called the metamotif. The metamotif is a probability distribution over PWM motifs of a given length. I also present a nested sampling based algorithm for inferring metamotif parameters from a set of motifs.

All of the following chapters make use of the metamotif in one way or another: Chapter 3 introduces a variant of the NestedMICA motif inference algorithm with an informative motif prior based on the metamotif likelihood function (Equation 2.5). Chapter 4 presents a motif family classification method based around metamotifs. In Chapter 5 I then experiment with using the metamotif based classification method with *de novo* discovered motifs.

Chapter 3

Metamotifs in motif inference

A central goal in modelling genome regulation is the identification of TFs and their target DNA binding sites, expressed as short nucleotide sequence motif models. This goal is becoming tractable even for higher eukaryotic genomes due to the availability of reference genomes for numerous organisms, development of high-throughput methods for measuring DNA interactions of transcription factors, and with computational advances in short sequence motif inference algorithms. The lack of sensitivity to detect weakly represented motifs from noncoding sequence however remains a key challenge when applying computational motif inference on a large scale. One way to tackle this problem is through informing the inference process of prior biological information of known motif families – for instance through the use of metamotifs.

This chapter describes the addition of a metamotif based motif prior to the NestedMICA algorithm. This modification to the algorithm diversifies its use from hypothesis-free discovery of motif collections from large scale sequence data to answering specific questions about possible regulators acting in the sequences (“Is there a motif roughly like this present?”). To achieve this, I extended the NestedMICA motif inference algorithm to accept a series of metamotifs as a position specific prior probability function for motifs. The NestedMICA algorithm was chosen for the purpose, because it is known to perform well in large scale motif inference tasks (Down et al., 2007; Down and Hubbard, 2005). It was also straightforward to adapt the existing clipped simplex motif prior probability function to a function based on column-specific biologically informative Dirichlet

distributions. The prior function, which allows multiple types of motif families to contribute to it simultaneously, could also be applied more generally to bias the search space of a larger motif inference problem to ‘biologically plausible’ motifs (instead of for instance repeat-like).

3.1 Previous work on biologically informative motif prior functions

De novo motif inference approaches show promise in finding motifs that determine gene regulatory programs. The NestedMICA algorithm for instance has been used in a number of regulatory genomics studies of both human and other organisms. Examples include analysis of Polycomb and Trithorax binding sites in *Drosophila* (Kwong et al., 2008), zebrafish distal enhancers (Rastegar et al., 2008), targets of the transcription factor Ntl (Morley et al., 2009), indirect targets of the deafness associated micro-RNA miRNA-96 in mouse (Lewis et al., 2009), as well as transcription factors involved in determination of ES cell transcriptional programs in mouse (Chen et al., 2008; Loh et al., 2006). NestedMICA, similar to other *de novo* motif inference algorithms, however commonly suffers from lack of sensitivity when applied to large collections of long eukaryotic promoter sequences where the TFBS motifs are weakly represented. This makes it difficult to describe complete sets of regulatory motifs from sequence alone with it. I therefore wanted to see if prior biological knowledge in the form of familial metamotifs could be used to improve its sensitivity. This was motivated primarily by the work of Xing and Karp (2004) and Narlikar et al. (2006) who both showed that tendencies in the motifs of sequence specific transcription factors can improve the sensitivity of probabilistic motif inference algorithms. Earlier instances of biologically informed motif prior functions and position specific parameter constraints have however also been presented.

The earliest instance of a method which uses column-specific information in a probabilistic motif inference method was the MEME program (Bailey and Elkan, 1995), which has been extended to include an optional palindromic constraint on the motif nucleotide weights (Bailey and Elkan, 1995); The last column is taken

as an complemented version of the first column, the second last is the second, and so on. The same paper also describes a Dirichlet mixture prior used specifically in protein motif inference, inspired by the Dirichlet mixture priors developed originally to help in deriving protein domain HMM models (Brown et al., 1993; Krogh et al., 1994).

More advanced hierarchical Dirichlet mixture based motif models and motif prior functions were later developed by Xing *et al.* in a series of papers (Xing et al., 2003a; Xing and Karp, 2004; Xing et al., 2003b). The hidden Markov-Dirichlet multinomial based framework, coined as ‘MotifPrototyper’ (Xing and Karp, 2004), allows for training a family-specific prior function that is parameterised with column-specific weights over a small number of prototypical Dirichlet distributions trained from a database of PWMs. This is somewhat related to the metamotif based approach which uses column-specific Dirichlet distributions trained from motif data. The Gibbs Recursive Sampler algorithm also reportedly includes a column-specific Dirichlet prior, described by Thompson and Rouchka (2003) as follows: “informed prior models provide clues to the expected patterns in DNA binding motifs that influence but do not control posterior inference of sites and motifs. The Gibbs Recursive Sampler permits incorporation of informed motif priors and gives the user control over the strength of the clue.” The paper describes no further description to the exact approach used, nor offers an assessment of its performance impact. Sandelin and Wasserman (2004) present such an assessment for the Gibbs sampler, as well as the neural network based ANN-Spec (Workman and Stormo, 2000), which also contains an otherwise unreported feature to include target PWMs as initial neural network weights. Both ANN-Spec and the Gibbs sampler show a measurable sensitivity gain. Median 200% and 140% sensitivity improvement for the ANN-Spec and Gibbs sampler algorithms was observed, respectively, in an evaluation which was made roughly with similar principles as that described in Section 3.2.2 for the NestedMICA algorithm.

Some of the previous motif prior enabled methods allow simultaneous inclusion of prior information for more than one motif family during motif inference. One example of such methods is the neural network based SOMBRERO algorithm which uses prior information of PWMs for initialising a self-organising map used for motif discovery (Mahony et al., 2005a). The most recent example is

the Bayesian phylogenetic foot printing method, Phylogibbs-MP, which can use PWMs as a prior (Siddharthan, 2008). The motif prior function in the PRIORITY algorithm (Narlikar et al., 2006), which is based on a series of binary logistic regression functions trained from binding site instances, also allow multiple classes to be specified, although the sequence model itself greedily infers motifs one by one (with a ZOOPS-like sequence model, see Section 1.3.1); Narlikar et al. (2006) also concede that the Gibbs sampling based parameter estimation method would struggle beyond the tested class count of three.

3.2 Materials & Method

Below, I will introduce the metamotif based motif prior function which I incorporated into the NestedMICA algorithm (Section 3.2.1), and then describe the method devised for assessing its effect on the performance of NestedMICA in Section 3.2.2.

3.2.1 The metamotif prior function

The prior probability of motif \mathbf{X} given a metamotif α is taken as the sum of metamotif densities of α with all continuous motif segments contained in \mathbf{X} that have the same length l as the metamotif (log of the density is given by Equation 2.5). A segment of motif \mathbf{X} refers to a motif formed from columns of the motif starting from column i and ending at position $i + l - 1$. The prior probability of a motif given a series of metamotifs is simply the sum of prior density contributions of each of the metamotifs. A schematic showing summation of one metamotif of five columns ($l = 5$) over an eight-column PWM is shown in Figure 3.1.

The prior function described above can be summarised simply as a summation of a number of different, potentially overlapping metamotifs over the length of the motif. There are alternative, more computationally demanding but potentially more meaningful ways to compute a prior function with multiple metamotifs. One possibility would be to apply the “motif probability given a series of independent, non-overlapping metamotifs” function described in Section 2.2.4 as a motif prior function in the NestedMICA algorithm. That is, the motif would be treated

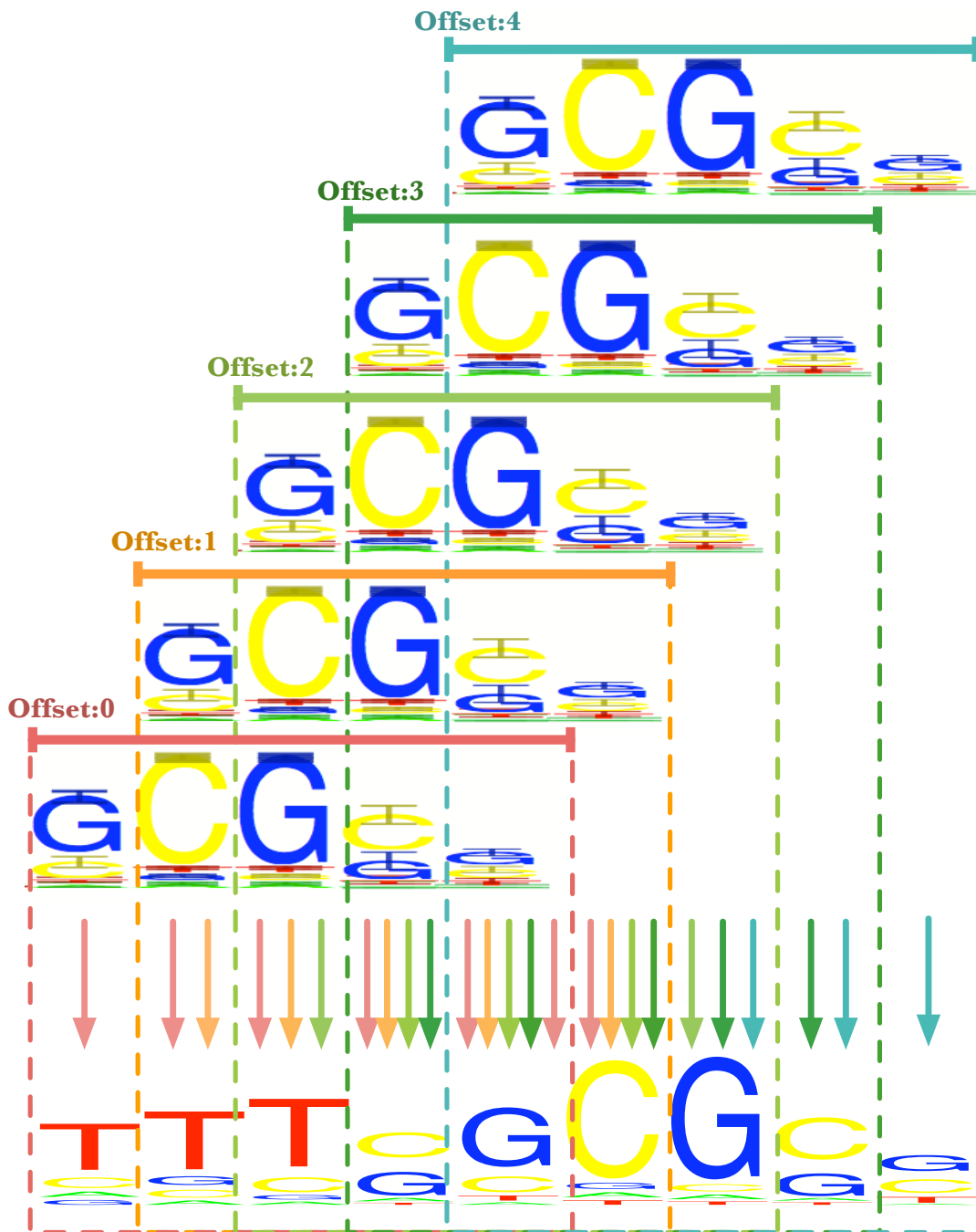


Figure 3.1: Metamotif densities with all offsets of the metamotif (shown above the PWM) are summed over the length of the motif (five different offsets shown, with different colours).

as an HMM of background multinomial positions and independent metamotif segments which can be ordered freely but cannot overlap (the multiple-uncounted motif metamotif mixture model). This formulation could potentially be more appropriate to cases where short metamotif components are applied as a motif prior (e.g. half sites). However, the already considerable run time that the NestedMICA algorithm requires for completing on large sequence and motif sets could be increased further by this prior function. This is because another costly dynamic programming step to compute the metamotif density function would be needed, as the prior function is computed on every iteration of the nested sampling for all motifs in the ensemble of potentially several hundred solutions. I therefore concentrated on the simple motif prior function presented here. This algorithm scales well to large sequence sets, and it is unlikely that the more complex metamotif density HMM prior would be practically useful in genome scale motif inference tasks without substantial optimisation. The optimisation work would likely include at least caching prior contributions of individual motifs.

3.2.2 Measuring motif inference sensitivity with synthetic sequence

To test the performance of the metamotif prior function, I conducted simulation experiments following the same principle as described for the NestedMICA (Down and Hubbard, 2005) and the BayesMD (Tang et al., 2008) algorithms. Human intronic nucleotide sequence fragments randomly chosen from the *Homo sapiens* Ensembl database release 50 (Flicek et al., 2008) were ‘spiked’ with five different types of motifs. The motifs used were those of ZAP1, HIF1, TBX5, TAL1 and NF- κ B transcription factors. These motifs were selected because they showed little similarity with each other when aligned, and because this set contains examples of differing motif length and information content. All sequence sets used contained 200 sequences, and the length of the sequences was varied between 100, 200, ..., 2000 nucleotides. The nucleotide k -mers sampled from each of the five PWMs in the evaluation were inserted at a constant relative frequency of 20% of the sequences, with a maximum of one motif present per sequence. In other words, motif density was varied by inserting the motif instances to back-

ground sequences of different lengths. Motifs of only one kind were present in each synthetic sequence set.

Motif inference with three types of motif prior functions were tested with the sequences:

1. A single familial metamotif contributing to the prior function.
2. A prior function with all of the five unrelated metamotifs contributing to the prior, with instances of only one motif family being actually present represented in the sequences.
3. An uninformative Dirichlet prior similar to the previously published NestedMICA version 0.8.

In each of the motif inference runs, the longest sequence length at which the algorithm infers the correct motif of interest is reported as a measure of sensitivity ($p < 0.05$), with motif comparison p -values computed, as described in [Down et al. \(2007\)](#). In all cases, five motifs were inferred from the sequences. Five motifs, as opposed to for example only one, were inferred, because recurring sequence motifs tend to be found from even intronic sequences, and I therefore cannot assume that the spiked motif would be the only motif signal present. The sequence background model used in all evaluations of the algorithm was a 4-class 1st order trained from the 2000nt long intronic sequences with `nmmakebg`.

The source motifs (ZAP1, HIF1, TBX5, TAL1, NF- κ B) were transformed to metamotifs to be used in the metamotif prior function by applying a pseudocount of 0.1 to the motif column weights, and interpreting the resulting motif nucleotide weights as mean nucleotide weights in Dirichlet distributions with precision set at 4.0 (metamotifs used in the experiment shown in [Figure 3.2](#)). The metamotif priors used in the prior function evaluation were constructed from known PWMs with a set precision and pseudocounts to assess the hypothesis testing use of a motif prior function: user is aware of a set of potentially relevant motifs or consensus strings present in a sequence set and wants to inform the algorithm of them to increase its sensitivity to detect the signal.

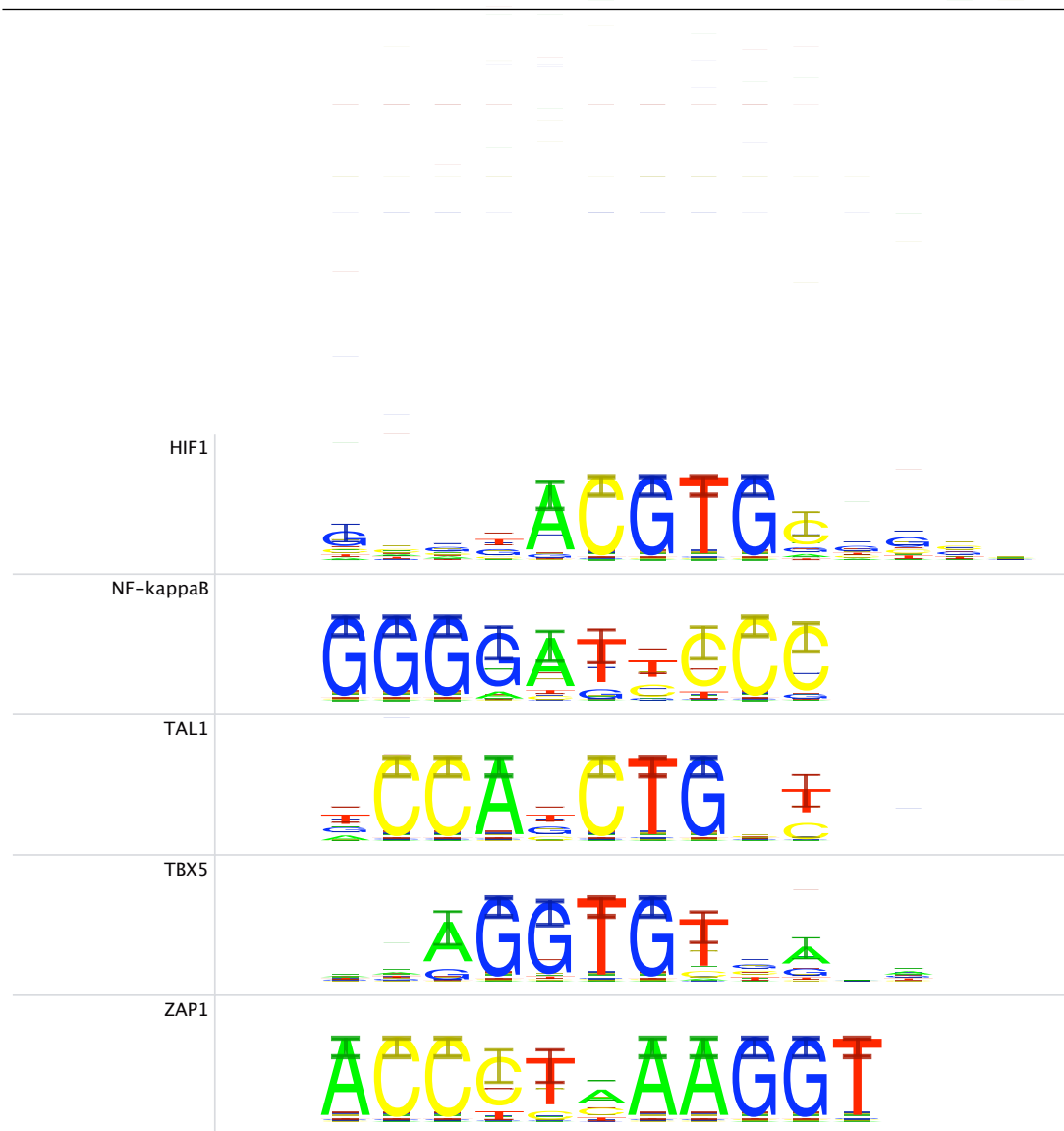


Figure 3.2: Synthetic metamotifs contributing to the motif prior functions used in the assessment. Error bars represent 95% confidence intervals.

3.3 Results & Discussion

Results of applying the metamotif based motif prior function are shown in Sections 3.3.1 and 3.3.2. Several ways to use the motif prior with the NestedMICA suite (Down and Hubbard, 2005) and the graphical iMotifs motif inference environment (Piipari et al., 2010b) are introduced in Section 3.3.3.

3.3.1 Performance effect of a correct motif family prior function

Results of the motif prior comparison are shown in Figure 3.3. It is evident that when the correct motif prior function is used on its own (the rightmost bars), improvement in the motif inference performance is seen across the line, when compared to the uninformative prior (the leftmost bars). When the correct motif is introduced amongst a set of ‘decoy motif’ contributions in the prior function, improved performance over the uninformative prior is seen with all motifs but TBX5, which is unchanged. The effect size, in terms of the difference between maximum sequence lengths at which the motif is detected in the informative and uninformative cases, depends on the motif; Some motifs appear inherently ‘harder’ to discover even when a biologically informed prior function is available. The most likely reason for the variability both in the baseline motif inference sensitivity, and the effect of the informative weight matrix prior, is in the difference in length and information content of the motifs, ranging from as high as fourfold difference in the motif recovery length for TAL1 and NFKappa- β , to only a 1/3 improvement from 400bp to 600bp sequence between the uninformative and the ‘single’ informative metamotif prior for the TBX5 motif. The presence of ‘decoy’ metamotif patterns decreases the effect size in all cases.

3.3.2 Performance effect of an incorrect motif family prior function

I also wanted to ensure that the metamotif prior did not have the propensity to bias motif inference to an incorrect solution, i.e. that it does not encourage the inference of a motif not supported by the sequence data. I tested this by

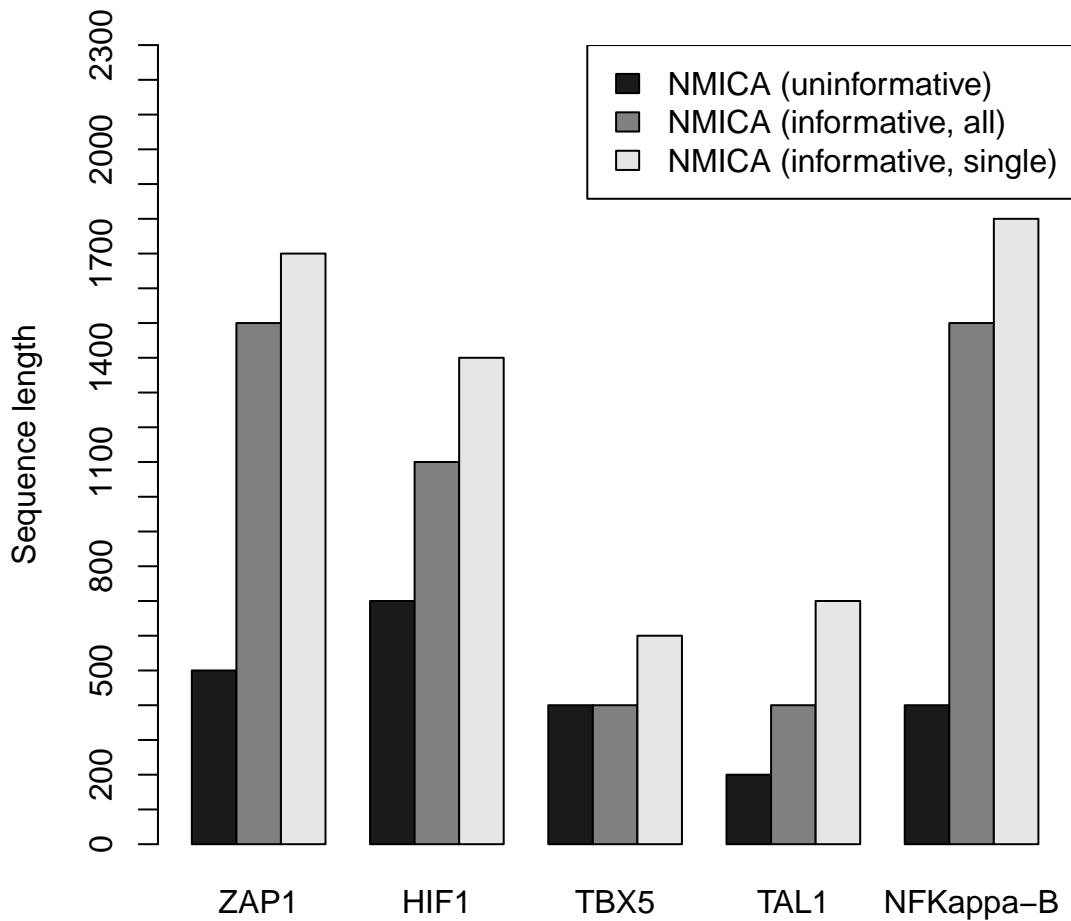


Figure 3.3: Informative weight matrix prior improves NMICA’s sensitivity to resolve motifs present in human intronic sequence in low frequency (0.2 frequency). The bars represent the sequence length at which a motif closely similar to the input motif was successfully recovered ($p < 0.05$, empirical p -value defined in (Down et al., 2007)).

spiking intronic sequence with the NF- κ B motif, and using the ZAP1-like metamotif in the prior function. No motifs similar to ZAP1 (whose instances were not present in the sequences) were recovered from the spiked intronic sequence between lengths 100 and 2000 (comparison with distances and p -values shown in 3.4), indicating that the metamotif prior function does not have an adverse effect on inference specificity. A number of other combinations of spiked motifs and inaccurate informative metamotif prior functions were also tested, with no observed tendency for the algorithm to infer a motif that is not supported by the sequence data (data not shown).








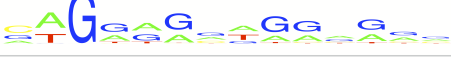

	ZAP1
100nt (distance to ZAP1: 8.78, p: 0.21)	
300nt (distance to ZAP1: 6.78, p: 0.26)	
500nt (distance to ZAP1: 12.81, p: 0.50)	
700nt (distance to ZAP1: 7.50, p: 0.66)	
900nt (distance to ZAP1: 9.01, p: 0.30)	
1100nt (distance to ZAP1: 7.51, p: 0.20)	
1300nt (distance to ZAP1: 6.91, p: 0.20)	
1500nt (distance to ZAP1: 13.02, p: 0.48)	
	

Figure 3.4: The closest motif match to the invalid motif pattern (ZAP1) shown alongside the ZAP1 motif. No pattern like ZAP1 should be seen, and indeed is not seen. Five motifs were inferred at each sequence length (100nt, ...,1500nt).

3.3.3 Making the metamotif prior available

As the ultimate aim of the metamotif prior function work was to provide tools useful for motif inference related hypothesis testing, to answer questions such as “Are there motifs present in this sequence set that are related to what I am expecting?”, I developed several ways in which other researchers can effectively make use of this work that are detailed in the sections below.

It should also be noted that metamotif models inferred from motif sets with the nested sampler framework introduced in Chapter 2 can be incorporated in a reduced PWM representation to other motif inference algorithms which accept PWM based motif prior functions or initialisation values, for instance the ANN-Spec (Workman and Stormo, 2000) and Gibbs Sampler (Qin et al., 2003) variants created by (Sandelin and Wasserman, 2004), the SOMBRERO (Mahony et al., 2005b) variant by Mahony et al. (2005a), or Phylogibbs-MP (Siddharthan, 2008). This is because a metamotif is a product Dirichlet distribution model of motif families, which contains an implicit familial binding profile like average motif (see Section 3.1 for a discussion of FBPs). Using metamotifs in external programs is made especially easy because of the way the metamotif models are stored in the same XML-based XMS format used by NestedMICA (Down and Hubbard, 2005) and iMotifs (Piipari et al., 2010b) to store PWMs; The metamotif’s average column weights (the implicit ‘average motif’) are in fact stored identically to a PWM, and the α_0 precision values are stored as additional key-value based annotations in the file, only applicable for tools which are ‘metamotif aware’.

3.3.4 Using the metamotif prior with the NestedMICA algorithm

Support for the metamotif prior function was integrated into the NestedMICA suite ¹ with a series of command line arguments. The metamotif prior extension to the NestedMICA tool was also designed to function with any number of metamotif models, or input PWMs or IUPAC consensus sequences ‘converted to’ metamotifs. PWMs are treated as metamotif priors by interpreting its columns i

¹The NestedMICA suite is available at <http://www.sanger.ac.uk/resources/software/nestedmica/>

as the $\mathbb{E}[\mathbf{x}_m]$ of a metamotif and applying a constant precision α_0 to all columns of the metamotif. IUPAC consensus sequences are first transformed to PWMs by applying pseudocounts and then transformed similarly as PWMs. Metamotifs inferred with our framework can also be potentially used with other Bayesian motif inference algorithms that model a prior distribution over motif positions. Metamotifs could therefore be of general use in building large and complete regulatory binding site motif libraries for novel genomes. Usage examples are shown below for the three ways in which the NestedMICA motif inference tool `nminfer` can be used with metamotifs.

1. An XMS file containing metamotif models (consult NestedMICA manual for more detail for including per-column precision information in the XMS format):

```
nminfer -priorMetamotifs y.xms -seqs input_sequences.fasta \  
-numMotifs 3 -minLength 6 -maxLength 14
```

2. An XMS file containing motif models, with an added pseudocount and precision parameter set to transform PWMs to metamotif models:

```
nminfer -priorMotifs x.xms -priorPseudocount 0.1 \  
-priorPrecision 4.0 -seqs input_sequences.fasta -numMotifs 3 \  
-minLength 6 -maxLength 14
```

3. An IUPAC consensus string, with an added pseudocount and precision parameter set to transform PWMs to metamotif models:

```
nminfer -consensus gataa -priorPseudocount 0.1 \  
-priorPrecision 4.0 -seqs input_sequences.fasta \  
-numMotifs 3 -minLength 6 -maxLength 14
```

Notably the IUPAC consensus string support allows inputting not only A, C, G, T, N, R (purine), Y (pyrimidine) but also all the other degenerate symbols in the IUPAC DNA code standard (e.g. S which corresponds to C or G).

3.3.5 Using the metamotif prior with iMotifs

The motif set visualisation environment iMotifs, which I developed during this project ([Piipari et al., 2010b](#)), was expanded with support for the metamotif prior

function driven motif inference (Figure 3.5). This was done to make it easy for a user with little prior experience of the NestedMICA suite to deploy and try it with the informative prior extension. More information about iMotifs is available in Appendix A.

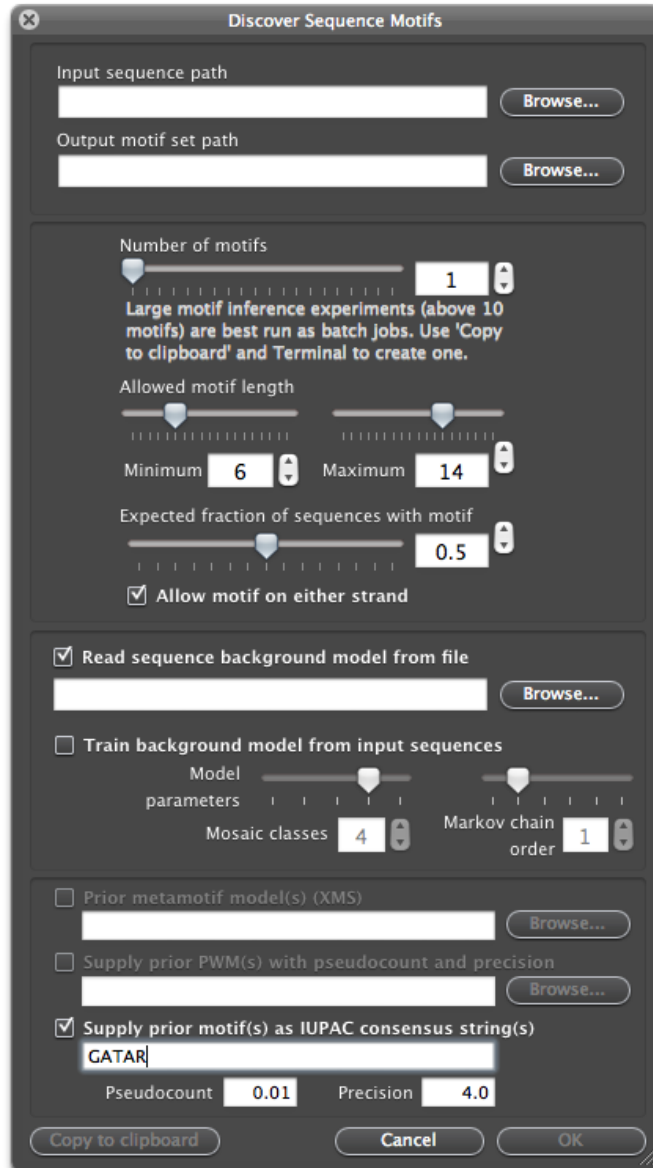


Figure 3.5: A NestedMICA motif inference run can be configured and run directly in iMotifs. Alternatively the NestedMICA run can be configured in iMotifs (Analysis >Discover Motifs from Sequence) and executed in the terminal after using the ‘Copy to clipboard’ function. A metamotif prior with one or more metamotifs can also be specified, either by specifying a file that contains metamotif model(s) as an XMS formatted file, as a series of PWMs in an XMS formatted file, or as IUPAC consensus strings. In the last two cases, pseudocounts and the prior precision (α_0) can also be specified.

Chapter 4

Metamotifs in motif classification

Metamotifs are shown in the previous chapter to significantly improve the sensitivity to infer motifs from sequence, when applied as a Bayesian PWM prior in the NestedMICA algorithm. Here I will show that metamotifs can also be applied to form functional predictions for motifs. Metamotifs are applied to a motif classification problem where features extracted from regulatory motifs (PWMs) are used to predict the family of protein DNA binding domains which is likely to interact with them. I will refer to this problem as ‘motif family classification’. The features I used in my motif family classifier are metamotif densities, and I therefore call the method **metamatti**, for **metamotif** based **automated** transcription factor **type** inference.

4.1 Previous work on motif family classification

Motif family classification is not a new idea. In particular, the following three studies provided an inspiration for the work described here:

- The hidden Markov Dirichlet-multinomial based MotifPrototyper framework (Xing and Karp, 2004), which is also used to provide the PWM column-specific Bayesian prior function discussed in the previous chapter (Section 3.1). The MotifPrototyper based motif classification is presented as a cross-validation based exercise where motifs from the TRANSFAC database are labelled with their superclass (one of basic, zinc coordinated,

helix-turn-helix, or β -scaffold domains, see Section 1.4.2 for a further discussion on structural taxonomies of TFs). The ability to classify motifs on the level of their superclass is discussed by [Xing and Karp \(2004\)](#) mostly as an interesting side-product of co-evolution of transcription factors and their binding sites, and the authors do not make available the motif classifier for other researchers to use.

- The sparse multinomial logistic regression (SMLR) based motif classifier by [Narlikar and Hartemink \(2006\)](#). Similarly as above, the emphasis of this work is not in constructing a publicly available motif family classification tool for the research community, but to present the classification problem as a side-product of the evolutionary pressures acting of TFs and their binding sites. The paper also acts as a biological application to a novel sparse, probabilistic supervised machine learning method developed by the authors (SMLR). The classification is done, as in the case of MotifPrototyper, to motifs in the TRANSFAC database (its six largest classes Cys₂His₂ and Cys₄ zinc fingers, homeodomains, forkhead domains, basic helix-loop-helices and basic zipper domains), but the classifier labels the motifs with their TRANSFAC class (not superclass, as done by MotifPrototyper). Notably, the same authors also published a separate paper ([Narlikar et al., 2006](#)) where they present an informative PWM prior enabled motif inference algorithm which also labels the discovered motifs with their domain family. This paper is discussed in the context of motif priors in Chapter 3.
- [Sandelin and Wasserman \(2004\)](#) are the earliest at suggesting a computational motif family labelling method, in the same familial binding profile paper which was discussed in the previous motif prior chapter. It is however limited to a small number of metazoan TFs (63 in total) which are closely similar in the clustering chosen by the authors (bZIP motifs for instance are subdivided to three subgroups). Due to the limited scope of this classification study, and the biased choice of the motifs in this study, I decided not to assess my method against it (similar choice was also made by [Narlikar and Hartemink \(2006\)](#)).

In contrast to the previous studies, my goal in this work has been to both rigorously test my method in context of the earlier work where applicable, and to also present a tool for motif family classification that other researchers can use in the comparative study of regulatory motifs. Indeed, **metamatti** can be distributed as an R package (Section 4.3.4.1), and as a remotely available motif classification web server (Section 4.3.5).

In this chapter I firstly introduce the **metamatti** classifier, and compare its performance to two of the methods noted above: MotifPrototyper (Xing and Karp, 2004) and SMLR (Narlikar and Hartemink, 2006). I also validate the classification method’s performance with two independent, experimentally validated homeodomain datasets, and give a brief introduction to the usage of the classification tool. In the next chapter I then apply the method to a series of computationally predicted motifs, to showcase **metamatti**’s ability to predict the class of *de novo* predicted motifs from a genome scale motif inference study. In addition to assigning clues of function to large sets of *de novo* motifs, I believe that family classification of motifs could for instance become a useful diagnostic method when working with TFBS motifs predicted from genomic ChIP-chip or ChIP-seq data; with it, one could test how closely motifs predicted from the DNA fragments bound by a TF of interest match the expected familial pattern of the DNA binding domain under study. This can be helpful in identifying the relevant motif from potentially many that are over-represented in DNA fragments bound in a ChIP assay. This idea has been explored by MacIsaac et al. (2006) with a familial binding profile based method.

4.2 Materials & Method

The principle of my motif classifier is to compute the density function (Equation 2.5) of a large dictionary of familial metamotifs along the length of training set motifs, effectively “scanning” weight matrices with metamotifs. The optimal (maximum) and average metamotif densities of each metamotif with the motif are then included as features in a random forest classifier that tries to infer the TRANSFAC superfamily (Figure 4.1) or TRANSFAC family (Figure 4.2) of the motifs. Random forest classification was chosen as the machine learning frame-

work, most importantly because it generalises naturally to multi-class problems and provides reliable error estimates as part of model training (Breiman, 2001b). The framework also controls the sparsity of the feature set used for classification (see Section 1.3.4 for an introduction to random forests).

4.2.1 Training data

All motif families with at least 10 representatives were retrieved from the TRANSFAC 12.2 database (Matys et al., 2006), totalling 623 motifs of 13 domain families (see Section 1.4.2.1 for more information about the TRANSFAC database). For the motif domain superfamily classifier comparison made with MotifPrototyper Xing and Karp (2004) (Figure 4.1), the set of motifs was reduced further to include only motifs annotated in TRANSFAC with the four superfamilies classified in (Xing and Karp, 2004). For the motif TRANSFAC class prediction comparison with SMLR (Figure 4.2), only motifs of the same six major classes classified with SMLR in Narlikar and Hartemink (2006) were included in our training set. The feature set is discussed in Section 4.2.2.

The **metamatti** motif type classifier training and cross-validation were implemented in the Ruby and R (Team, 2007) programming languages. Random forest classification was done using the package `randomForest` (Liaw and Wiener, 2002). Pseudocounts of 0.01 were added to all training set metamotifs, and the *mtry* parameter of the random forest classifier training was optimised by testing $0.1 \times \sqrt{p}, 0.2 \times \sqrt{p} \dots, 2.0 \times \sqrt{p}$ with intervals of 0.1, where p is the number of features in the classifier (the default value for *mtry* is \sqrt{p}). The *ntree* parameter that controls the number of trees to grow was set at 5000.

4.2.2 The classifier feature set

Most features in **metamatti** are metamotif probability density scores (Table 4.1). To compute the metamotif density features for the classifier, we chose to first divide the motifs into sets by complete linkage hierarchical clustering (Johnson, 1967) with the SSD metric described in Down et al. (2007) and cutting the clusters at a lenient clustering cutoff of 6.0. This resulted in 54 motif clusters. Three metamotifs were trained from each motif cluster with `nmmetainfer`, resulting in

195 metamotifs to be used in the motif classifier (examples seen in Figure 2.10). Metamotif length was constrained between 6 and 15 columns, and the expected usage fraction was set at 0.5.

Feature type	Description
Maximum metamotif hit scores with all of the familial metamotifs	Motifs were scanned with all input metamotifs, and the optimal score was chosen.
Per-column average entropy	Average Shannon entropy of columns.
MLE Dirichlet parameters	A maximum likelihood Dirichlet distribution is estimated as described in Minka (2003), and the parameters of this distribution are used as features $(\alpha_A, \alpha_G, \alpha_C, \alpha_T)$.
Symmetric Dirichlet background parameters	A symmetric Dirichlet distribution is estimated.

Table 4.1: Features used in the **metamatti** classifier.

4.3 Results & Discussion

The main results in this chapter are threefold: the comparisons of the developed method with previous methods (Section 4.3.1), an independent validation of the performance with two large homeodomain datasets (Section 4.3.2), and a brief explanation of the publicly available implementation of the classification method (Section). Additionally, I also discuss the reasoning behind choosing an appropriate motif cluster count (Section 4.3.2.2), and compare the classifier to the more naive option of simply scoring motifs with average motifs derived from clustered, aligned motifs (Section 4.3.3).

4.3.1 Performance comparison with previous methods

Classification performance of **metamatti** was compared to two methods with a related goal: MotifPrototyper (Xing and Karp, 2004) which classifies motifs into four TRANSFAC superfamilies (zinc coordinated, helix-turn-helix, β -

scaffold,basic), and SMLR which classifies motifs into six major classes of TF domains (Cys₂His₂ and Cys₄ zinc fingers, homeodomains, forkhead domains, basic helix-loop-helices and basic zipper domains) (Narlikar and Hartemink, 2006).

4.3.1.1 MotifPrototyper

Classification accuracy comparison shows that **metamatti** outperforms MotifPrototyper (Xing and Karp, 2004) (Figure 4.1) across all four TF domain superfamilies. The margin between the two methods is especially clear when one compares **metamatti** with the ‘full’ dataset classification made by Xing and Karp (2004), which contains all members of the four superfamilies in the TRANSFAC class, as opposed to the reduced ‘major class’ set which contains all motifs with at least 10 examples in the dataset. The **metamatti** classification was made with the full dataset.

There are several possible reasons for the substantial difference in performance. Firstly, the MotifPrototyper classification is made simply with a maximum a posteriori scheme: each TRANSFAC superclass corresponds to a MotifPrototyper model, and motifs are assigned to the superclass which has the highest maximal posterior probability to be generated by the corresponding MotifPrototyper. **metamatti** instead uses the metamotif densities as a features in a more sophisticated, discriminative random forest based classifier, which assigns the class labels to a motif. Secondly, the metamotif inference algorithm I developed is not constrained to a fixed motif family column count, unlike the algorithm utilised in MotifPrototyper which estimates model parameters from aligned motifs. The method by which motifs are aligned and trimmed to equal length is not specified by Xing and Karp (2004). Thirdly, training several metamotifs per motif family, **metamatti** also accounts for the fact that not all columns in motif families can be accurately expressed as a single column wise probability distribution. Instead, recurring patterns in a motif set can be generated by multiple potentially shorter familial metamotif components in my model. Furthermore, the metamotif estimation algorithm treats some motif columns as noise with a column background model, improving the capacity to find recurring patterns from sequence motif sets and reducing over-fitting of familial models due to reporting

weak or nonexistent recurring trends.

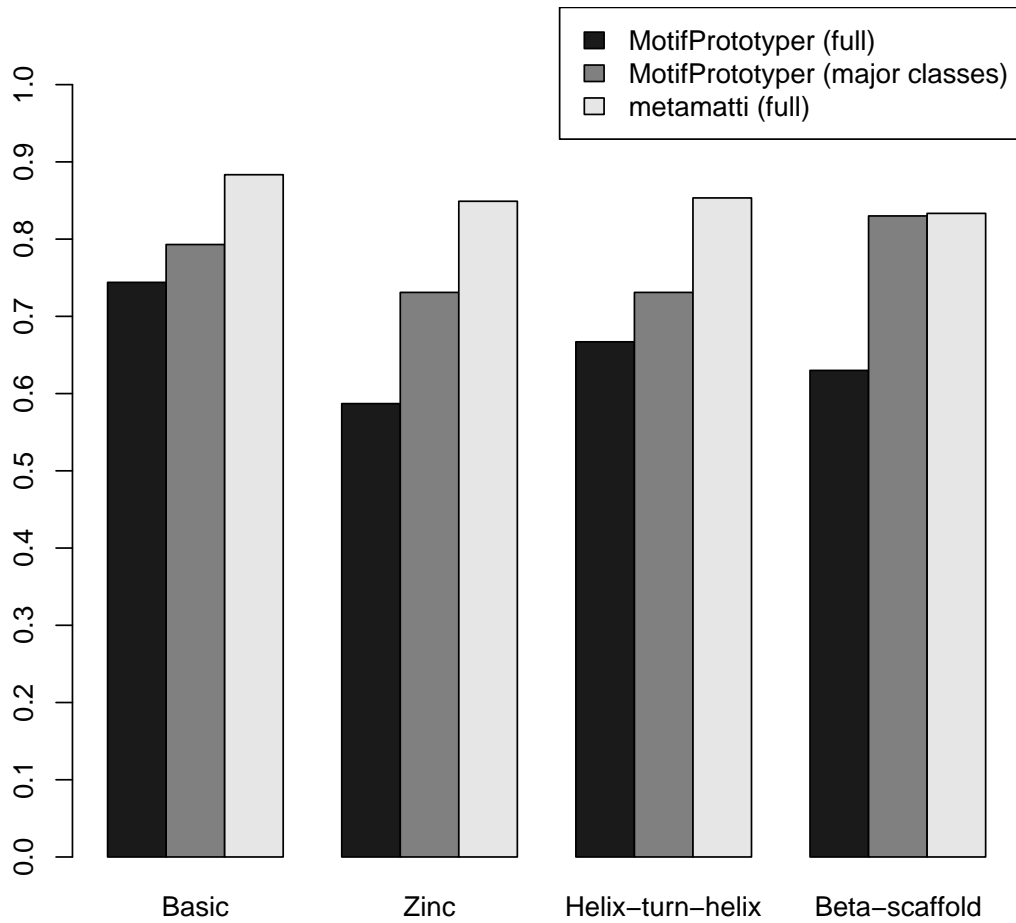


Figure 4.1: Accuracy comparison between TF domain superfamily level classification with **metamatti** and MotifPrototyper (10-fold crossvalidation). The 'major classes' refers to MotifPrototyper's reported performance for all motif families which include at least ten motif instances (Xing and Karp, 2004) in the TRANSFAC database (Matys et al., 2006) from the four superfamilies basic, zinc, helix-turn-helix and β -scaffold. 'Full' refers to a classification of all motifs in the four superfamilies, instead of just the major classes.

4.3.1.2 Sparse Multinomial Logistic Regression

To compare **metamatti** with SMLR (Narlikar and Hartemink, 2006), I conducted the TRANSFAC class level classification with the same subset of TRANSFAC 12.2 PWMs that were classified with SMLR. The overall classification accuracy comparison shows that **metamatti** has a marginally improved performance at 89.5% classification accuracy over the 87% reported for SMLR. The class-by-class accuracy figures (Figure 4.2) and the confusion matrix of the 6-way TRANSFAC motif family classifier (Table 4.3) however make it evident firstly that the ability of sequence motif properties to distinguish motifs by binding domain varies considerably depending on the domain both for **metamatti** and SMLR, and secondly that the higher classification accuracy comes at the cost of a 14% drop in the classification accuracy of the bHLH family (89% accuracy with SMLR, 75% with **metamatti**). The partially palindromic E-box motif CAGGTG appears to be the most common type misclassified in the erroneous bHLH motif cases. Inspection of family assignments of motifs in the TRANSFAC database shows that closely similar motifs with the CAGGTG consensus have been annotated with all of bHLH and C_2H_2 zinc finger families, highlighting a general limitation of a sequence PWM feature based motif family classification methods. Overall, the variability in accuracy across classes is not surprising: Luscombe and Thornton (2002) already describe sequence-specific DNA binding motifs into ‘highly specific’ (e.g. TATA binding protein and the basic zipper domain) and ‘multi-specific’ (e.g. homeodomain, C_2H_2 and Cys_4 type zinc finger domains), i.e. that different domains show different degree of constraint in the binding profiles seen in nature, which can make some domains harder to classify even with sophisticated methods. Random forest classification in fact outputs a classification probability for each of the potential classes. I in fact use this property of random forest classification in Chapter 5 (Section 5.3.6.5) to choose a confidence level for classification decisions, instead of reporting a class for all input motifs regardless of the uncertainty.

Motif family prediction methods ultimately rely on the structural mode of interaction by a protein DNA binding domain being reflected as a DNA sequence specificity pattern, and that pattern being distinct to each motif family as a

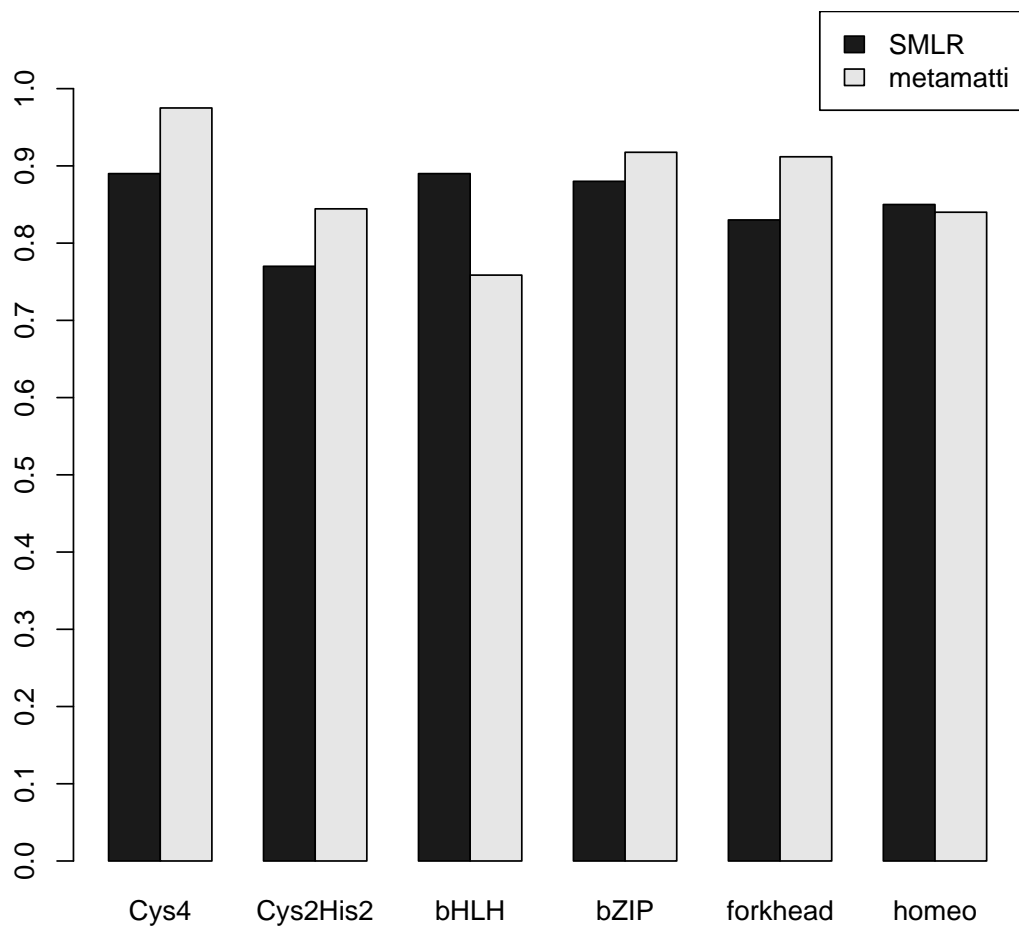


Figure 4.2: Accuracy comparison between the TF domain family classification with metamatti, and SMLR (k-fold cross-validation).

	Cys4	C2H2	bHLH	bZIP	Forkhead	Homeodomain	Class error
Cys4	39	0	0	0	0	1	2.5%
C2H2	0	38	3	0	1	3	15.6%
bHLH	0	2	22	5	0	0	24.0%
bZIP	0	3	0	78	0	4	8.0%
Forkhead	0	0	0	0	31	2	9.0%
Homeodomain	2	1	1	3	0	37	16.0%
Totals	41	43	26	86	32	47	

Figure 4.3: Confusion matrix of the 6-way TRANSFAC motif classification with the **metamatti** classifier. Columns correspond to the real class, and rows to the predicted class.

result of co-evolution of the two protein and its binding sites. As the above example of CANNTG sites shows, this is not always the case in nature: certain bHLH and Snail-like C₂H₂ like factors for example are thought to bind with closely similar specificities to compete for the same binding site positions (Nieto, 2002). The familial tendencies observed for DNA binding sites of transcription factors are thought to be due to both biophysical constraints on the possible DNA binding site patterns of a certain binding domain and evolutionary forces that maintain the familial DNA specificities distinct. Such forces range from functional redundancy of paralogous factors with overlapping binding sites (Kafri et al., 2005) to antagonistic regulation by opposing activators and repressors (Tanaka et al., 1993). To give an example of the inherent differences between TF domains, the C₂H₂ domain noted above has been found to be extremely plastic and a number of individual zinc fingers have even combined to very long (18bp) binding site patterns in a highly modular fashion (Dreier et al., 2001, 2000). In contrast, the bHLH domain has been observed to be much more strongly constrained in its DNA binding tendencies in a thorough mutagenesis study of the DNA contacting residues of the Max transcription factor (Maerkl and Quake, 2009). Further work is clearly needed to cover the full spectrum of binding site patterns explored by

sequence specific DNA binding domains, which also highlights the need for models such as the metamotif that describe recurring patterns in sequence motifs.

4.3.2 Performance measurement of two large homeodomain datasets

The previous motif classification work, which I compare my method with, has relied on cross-validation based estimation of classification accuracy from a single public database (Narlikar and Hartemink, 2006; Sandelin and Wasserman, 2004; Xing and Karp, 2004). Recent advances in protein-DNA interaction assaying have however resulted to the availability of several new experimental regulatory motif data sets that are not deposited in TRANSFAC. I wanted to assess the performance of **metamatti** with two homeodomain motif sets recovered from different species and via different experimental methods. The evaluation also allowed me to compare classification error rates achieved in independent datasets to the error rate predicted by **metamatti** classification for the homeodomain motif family. I applied **metamatti** to the *Mus musculus* PWMs constructed from the Berger et al. (2008) protein binding microarray motif data and reported the relative frequency at which the motifs were classified by **metamatti** with the homeodomain label (out of the six possible classes). Similarly, I classified motifs from the Noyes et al. (2008a) *Drosophila melanogaster* bacterial one-hybrid motif datasets.

The classification accuracy rates for both homeodomain motif sets were shown to be high, and in good agreement with the out-of-bag accuracy estimate of 91.3% reported by the **metamatti** random forest classifier during classifier training: 92.1% and 91.7% of the homeodomain motifs in the Berger et al. (2008) set of 84 motifs, and the Noyes et al. (2008a) set of 177 motifs, were correctly classified, respectively. I studied the misclassified examples from the *Drosophila melanogaster* homeodomain datasets in more detail to see where the misclassified motifs lie in the homeodomain specificity group clustering presented in Noyes et al. (2008a). Interestingly, the misclassifications were shown to be atypical homeodomains which do not contain the canonical TAATTA core and fall amongst the smaller specificity groups. The misclassified motifs included three

TGIF-Exd-like motifs (Vis, Hth, Exd), two Iroquois-like (Ara, Mirr), one Six-like (Optix) and an outlier from the specificity group clustering (Figure 4.4A). A similar trend of non-canonical homeodomains being primarily amongst the misclassified was also noted for the *Mus musculus* homeodomain motifs (4.4B). This is most likely explained by atypical homeodomain motifs not being well covered well by the TRANSFAC 12.2 training set; No closely matching homeodomain motifs were observed in TRANSFAC 12.2 to many of the misclassified motifs.

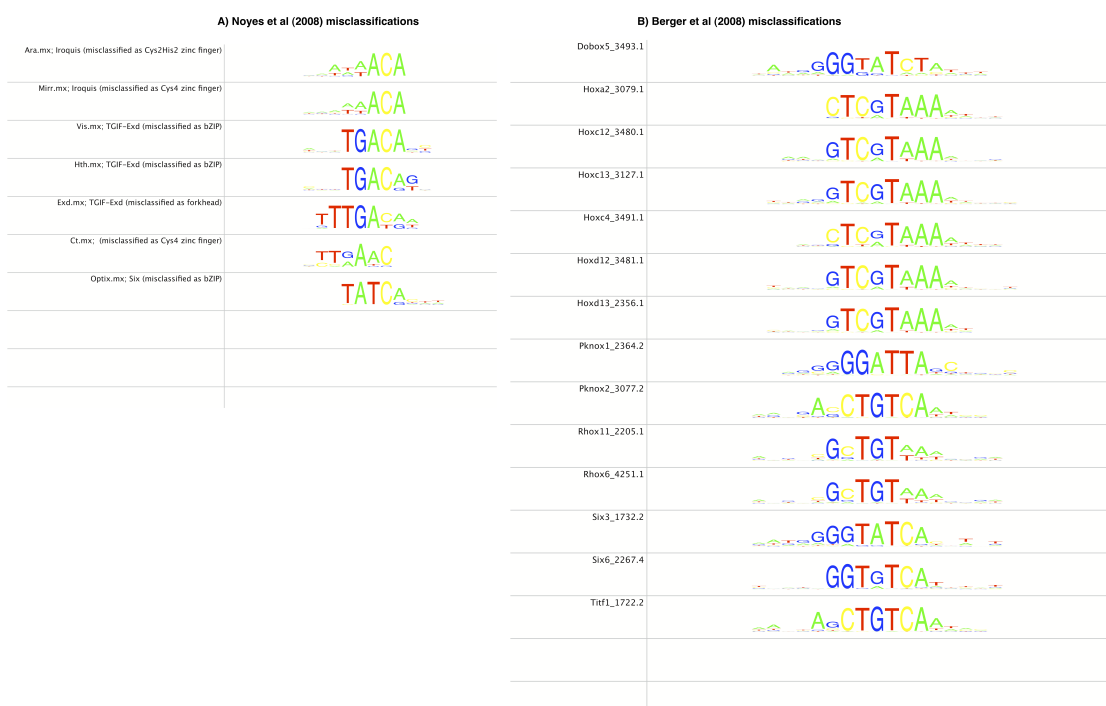


Figure 4.4: Misclassified homeodomain motifs in the A) Noyes et al. (2008a) and the B) Berger et al. (2008) datasets.

4.3.2.1 Classifying homeodomain motifs by their specificity group

I also wanted to test if a **metamatti**-like classifier could be trained to detect more detailed differences between motif groups than motif family or superfamily, a question which the previous methods have not addressed. I therefore labelled the *Drosophila melanogaster* homeodomain motifs with the homeodomain specificity

groups suggested by [Noyes et al. \(2008a\)](#) and estimated a single metamotif with `nmmetainfer` from each of the specificity groups. A single metamotif was used because of the small total number of motifs in the training data. I then trained a **metamatti** classifier with these metamotifs similarly as described above in Section 4.2. A remarkably high accuracy of 84% (confusion matrix shown in Table 4.5), when all [Noyes et al. \(2008a\)](#) homeodomain motifs with 3 or more examples per specificity group were included in the classification (9-way classification). The applicability of supervised machine learning strategies that aim to learn motif type labels more precise than the DNA binding domain family are however currently limited by the amount of available training data. For instance, the 84 motifs in the [Noyes et al. \(2008a\)](#) dataset contain examples of 11 specificity groups which are very biased to the two largest groups (Antennapedia and Engrailed, with 25 and 15 examples, respectively), with several specificity groups containing as few as two to four examples (Ladybird, Iroquis, NK-1, NK-2, TGIF-Exd, Bcd). This makes classifier error estimation imprecise especially for the weakly represented classes and results in the major classes, which have as much as eightfold as many examples present in the training dataset, to have considerable weight in predictions over the smaller classes (such as to maximise overall classification accuracy). Methods like **metamatti** can however become increasingly relevant once more high-throughput TF DNA specificity data becomes available.

	AbdB	Antp	Bar	Bed	Engrailed	Iroquis	NK-1	NK-2	TGIF-Exd	Class
AbdB	5	0	0	0	0	0	0	0	0	0.00
Antp	0	15	0	0	2	0	0	0	0	0.12
Bar	0	0	5	0	1	0	0	0	0	0.17
Bed	0	0	0	4	0	0	0	0	0	0.00
Engrailed	0	1	1	0	23	0	0	0	0	0.08
Iroquis	0	0	0	0	0	3	0	0	0	0.00
NK-1	0	0	0	0	3	0	2	0	0	0.60
NK-2	0	0	0	1	0	0	0	2	0	0.33
TGIF-Exd	0	0	0	0	0	0	0	1	3	0.25
Totals	5	16	6	5	29	3	2	3	3	

Figure 4.5: Confusion matrix of the homeodomain specificity group classifier. Columns represent the real class, and rows represent the predicted class.

4.3.2.2 Clustering of motifs prior to metamotif training

Clustering of the motifs, and training metamotifs from motif clusters, was motivated by the requirement to choose a value for the metamotif count parameter of the metamotif inference algorithm, and to limit the metamotif search space. Inspection of clusters at cutoff 6.0 showed no clusters with more than three strongly distinct recurring patterns. Although for many motif clusters there were clearly less than three distinct recurring metamotif patterns present at the clustering cutoff of 6.0, the metamotif inference algorithm was found to treat these cases by either inferring closely similar duplicate metamotifs (such as metamotifs 1 and 2 in Figure 2.10A) or short metamotifs with mean nucleotide weights with low information content, or occasionally splitting the metamotif segments in several independent parts. This suggested that together with a sparse machine learning strategy such as a random forests, it would be advantageous to choose a high metamotif count that would describe the input motif set in as much detail as possible, with the price of some potentially redundant features in the feature set (densities for duplicate or low information metamotifs). I validated this assumption by retraining the classifier with two metamotifs per cluster (a total of 130 metamotifs). The classifier trained with two metamotifs per family resulted in a mild decrease in the classification accuracy (88.4%, as opposed to 89.5% with three metamotifs per cluster), suggesting that the additional metamotifs were indeed informative.

4.3.3 Comparing a metamotif density based classification to a Cartesian distance based classifier

I assessed the importance of the metamotif density score in the **metamatti** classifier by comparing it to a more naive classifier where we replace the metamotif average and maximum scores with average and maximum SSD distances computed between the training set motifs and ‘average motifs’ of each of the motif families. The average motifs used in the more naive classifier were the mean PWMs of the metamotifs trained with **nmmetainfer**. They were used for classification by scoring the training set motifs with an SSD distance metric with each of the metamotifs. We found that the classifier accuracy achieved with the

SSD metric was lower to the metamotif density based classifier by 1.4% (accuracy of 88.1%), suggesting that both the metamotif mean and the column wise precision values which contribute to the metamotif density scores are partially responsible for **metamatti**'s high performance. Furthermore, I tested training a classifier with cluster average motifs instead of the metamotif segments, resulting in an accuracy figure of 86.5%, suggesting that not only is the metamotif density a suitable score, but that the motif segments identified by the metamotif inference algorithm provide a classifier that generalises better than simply using average motifs inferred by clustering and collapsing clustered motifs to an average representation.

4.3.4 Making metamatti available

Once I had shown the favourable performance of **metamatti** with respect to previous related methods, it became important to make the classification method readily available. Much like with the familial PWM prior work described in the previous chapter, I wanted to make it usable for both experienced and inexperienced users, with as low a barrier to installing and using it as possible. The following sections describe two ways in which **metamatti** can be taken advantage of.

4.3.4.1 The metamatti R package

The metamotif based classifier was initially developed as a series of R and ruby scripts. Distributing the tool as an R package was therefore a natural choice. The R package can be used to predict using classifiers either packaged in the software (included as R datasets loadable with the `data()` function), or ones trained with the package based on training data. The classifier training procedure also optionally plots a precision-recall curve and a variable importance graph, similar to those shown in Chapter 5. Furthermore, the JASPAR based classifier noted in this example is introduced and applied in Chapter 5 (Section 5.3.6.5).

The package source code, installation instructions and documentation is available at <http://www.github.com/mz2/metamatti>. A brief usage example is provided below.

```

#Load the library
library(metamatti)

# Get a list of available metamatti classifiers
# alternatively way to accomplish this is:
# try(data(package="metamatti"))'
# Due to the licensing terms of the TRANSFAC database,
# the TRANSFAC based classifiers are not made publicly available.
# Additional classifiers can however be trained
# as shown below.
getAvailableMetamattiClassifiers()
#"transfac-class-6-way", "transfac-superclass-4-way", "jaspar-5-way"

# Extract features from your motifs of interest
features <-
  extractMetamattiFeatures("your-motifs.xml", "jaspar-5-way")

# trainMetamattiForest(features, classifierName) can be used to
# train a new random forest classifier. Classifier training will
# also output a precision-recall graph
# (in this case jaspar-5-way-prec-recall.pdf,
# and a graph of variable importances
# ("jaspar-5-way-importances.pdf")
# in the working directory.
forest <- trainMetamattiForest(features, "jaspar-5-way")

# Alternatively, you can retrieve a jaspar-5-way classifier which is
# packaged alongside metamatti.
# Because the training sets are exposed as standard R datasets,
# you can also accomplish this with data("jaspar-5-way")'
forest <- getMetamattiForest("jaspar-5-way")

# Predict the class for the motifs
# Note that this is in fact a function from the randomForest package
# (the package is loaded upon loading the metamatti' library)
preds <- predict(features, forest)

```

4.3.5 The metamatti web server

In addition to the **metamatti** R package, I also created a simple web server application for motif family prediction. This was done most importantly because the outside dependencies required for installing the R package can act as a barrier of entry for inexperienced users, and because a web based application makes it possible to expose the TRANSFAC family classification to outside users (re-distributing the training data needed for it in the R package is impossible due to the licensing terms). The **metamatti** server can be used with a web browser (Figure 4.6) with a rather Spartan form based user interface. It also responds to a JavaScript Object Notation (JSON¹) based response format to web service API calls. Documentation for using the web service API is included alongside the freely available (LGPL licensed) source code of the project at

¹<http://www.json.org>

<http://www.github.com/mz2/metamatti>. It was implemented using Ruby on Rails (<http://www.rubyonrails.org>).

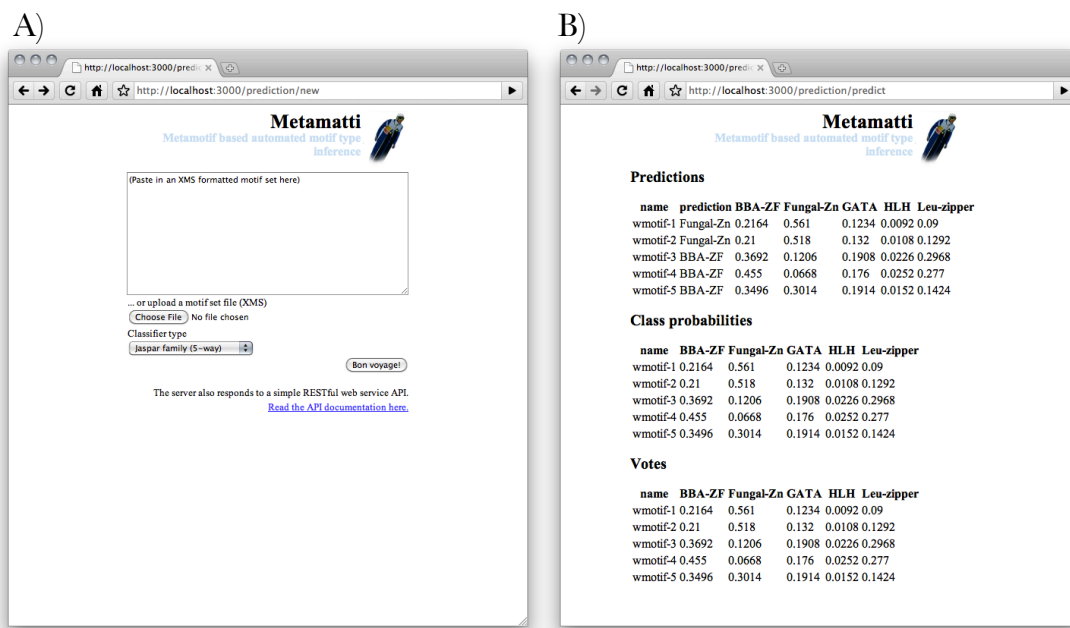


Figure 4.6: The **metamatti** motif classification web server. A) A screenshot of the prediction submission form. A motif set is entered either by pasting it to the form, uploaded as a file, or sent in a web service API call. B) A screenshot of the prediction report view. The tabular reports can be copied and pasted from (for instance to MS Excel), and they are also made available in a machine readable tabular (tab separated value) format through the web service API.

Chapter 5

Genome scale motif inference in *Saccharomyces cerevisiae*

The algorithmic work presented in the previous chapters, particularly the **meta-matti** motif classification framework, was partly motivated by the needs of the sequence analysis projects in which I have been involved. One part of this has been involvement in collaborative projects, where I have analysed human and mouse noncoding sequence with computational regulatory motif inference, scanning and statistical testing tools ¹, some of which I had developed for the purpose. The more substantial part has however been the genome scale *de novo* regulatory motif inference work with the *Saccharomyces cerevisiae* genome that will be discussed in this chapter.

5.1 Background

Budding yeast is an organism of great interest for regulatory genomics, given its small genome, amenability to genetic manipulation, and relatively simple regulatory mechanisms including a small total number of transcription factors (Goffeau et al., 1996). The DNA specificity of many of its TFs has been characterised in a combination of several high throughput *in vitro* studies (Badis et al., 2008; Zhu et al., 2009), providing a high quality reference set of regulatory motifs that

¹Majority of this work is now published in (Lewis et al., 2009) and Murray *et al.* (in press)

are useful for comparison with *de novo* discoveries. Information on the genomic binding positions for many of its TFs are also known from large scale ChIP-chip based studies (Harbison et al., 2004; Lee et al., 2002). Gene expression studies comparing knock-out lines for nearly all of its known sequence specific TFs to the wild-type are available (Hu et al., 2007; Reimand et al., 2010). Furthermore, many of the target genes of these TFs are known, as a result of the above ChIP-chip and expression studies, and literally thousands of other primary publications that have been manually curated (Teixeira et al., 2006). The *in vivo* DNA specificity of many budding yeast TFs is yet to be studied in high resolution, but nevertheless, budding yeast currently offers the best available knowledge base of TFs, TF target genes and binding site specificity, of any eukaryotic genome.

These resources together allow us to assess the ability of *de novo* motif inference algorithms to find large collections of regulatory motifs on a genome scale. Information from this large scale study is valuable most importantly because it indicates which of the algorithms, if any, are sufficiently accurate for complex regulatory problems that are aplenty in large genomes of multicellular eukaryotes.

5.1.1 Genome scale motif inference

Motif inference studies have traditionally been made to infer one or more recurring signals from a sequence set – of dozens to at most a few hundred – of sequences assumed to be co-regulated or involved in the same biological process. The rapid expansion in the number of complete genomes and computational power has however made it possible to use motif inference for a more ambitious goal: genome scale inference of comprehensive motif collections or ‘dictionaries’ from a significant subset of promoter sequences of a genome. I will below review a selection of previous literature on genome scale motif inference – both *ab initio* methods¹, and methods which apply gene expression or sequence conservation as a guide. See Section 1.2 for a more general discussion of motif inference methods.

To my knowledge, the earliest motif discovery study which fits the above criteria of *de novo* genome-scale motif inference is that of Brazma et al. (1998), who

¹*Ab initio* suggests in this context that no other information but the reference genome sequence and the predicted transcription start sites (putative promoter locations) are used as input for inferring the motifs.

predicted a series of regular expression like patterns from the *S. cerevisiae* genome using the SPEXS algorithm (Vilo, 1998), in an experiment where the algorithm was run ‘blindly’ with 6,000 upstream sequences. Assessing the significance of the found patterns, however, proved troublesome: top scoring regular expressions are matched to TRANSFAC binding site entries, but the authors attempted to draw few conclusions based on the found matches, except to note the surprise at being able to discover TFBS-like patterns with sequence information alone. Bussemaker et al. (2000) also presented a word enumeration based study where they found 11 known matching k -mers from a genome-wide study of *S. cerevisiae* promoters.

Several large, gene expression cluster-driven motif inference studies have been published. Among the earliest were Roth et al. (1998), who successfully recapitulated motifs of some of the key regulators of galactose response, heat shock and mating type regulatory systems in the *S. cerevisiae*, using the Gibbs sampling based AlignACE algorithm. Vilo et al. (2000) on the other hand used a word enumeration based method to find 62 clustered consensus strings reported to be match words in the SCPD database (Zhu and Zhang, 1999). Methods that go beyond clustering genes (and applying motif inference algorithms separately per cluster) have also been developed: Bussemaker et al. (2001) introduced a gene expression correlation based method REDUCE, which they apply to *S. cerevisiae* cell cycle regulation (Bussemaker et al., 2001). Elemento and Tavazoie (2005) use mutual information between gene expression patterns and the absence or presence of motifs as a means to infer *cis*-regulatory elements, in both mammalian, the yeast, and the *Plasmodium falciparum* genomes.

Whereas gene expression patterns are useful in inferring regulators which act in a certain state of the cell, use of sequence conservation has been used as a general ‘cell state blind’ informant for large scale motif inference. One of the earliest studies was Kellis et al. (2003) with a study of *S. cerevisiae*: a whole-genome multiple alignment of *S. cerevisiae* with *S. paradoxus*, *S. mikatae* and *S. bayanus*, which identified highly conserved consensus strings by clustering instances of shorter ‘mini-motifs’. Amongst the 78 motifs found, 28 closely match known TFBS consensus strings. Comparative techniques were later used by the same authors and others (Elemento and Tavazoie, 2005; Ettwiller, 2005; Jones

and Pevzner, 2006; Xie et al., 2005, 2007).

In conclusion, different large scale approaches to inferring *cis*-regulatory elements have been proposed, and several of them have been applied to the *S. cerevisiae* genome. In contrast to these previous studies, my perspective to inferring motif dictionaries from the budding yeast is primarily to find out how different previously published algorithms perform at this task, rather than setting out to discover novel functional motifs. This assessment is now made possible due to the availability of regulatory motifs, and sets of target genes for many of the budding yeast TFs. This is important, because performance of *de novo* motif inference methods have not previously been systematically assessed on biologically relevant, realistic problems.

5.1.2 Performance inference method assessments

Publications describing regulatory motif inference algorithms typically contain a comparison of the algorithm introduced with at least some previously published ones. Standard assessment criteria or benchmark datasets have not surfaced, and new methods are often compared only with a small number of common existing methods, so it is not always clear how they compare with the state of the art. An objective assessment of the merits of the hundreds of different available algorithms is therefore difficult. To my knowledge, the most comprehensive *de novo* motif inference algorithm benchmark, involving 13 different methods and discussed in more detail below, has been conducted by Tompa et al. (2005). As more and more motif inference methods are published on top of the hundreds already available, being able to assess the performance of methods relative to each other becomes increasingly important.

Two types of approaches have been used in previous literature for ranking methods:

1. Finding TFBS motifs from motifs from well studied collections of *cis*-regulatory elements (Ao et al., 2004; Liu et al., 2002; Roth et al., 1998; Thijs et al., 2002).
2. Finding TFBS motifs from synthetic sequence created by planting, or ‘spiking’ motifs into background sequence. The background is usually some neu-

tral sequence thought to be devoid of other motifs (e.g. intronic sequence). This approach is taken for instance by [Down and Hubbard \(2005\)](#); [Pevzner and Sze \(2000\)](#); [Workman and Stormo \(2000\)](#).

Measuring the performance of algorithms in either of the above cases is done most often by counting instances of motifs above some significance level, and comparing the overlap of the list of predicted motif instances to a reference binding site collection. The reference is either a set of known sites, if the assessment is made with real sequence, or a known set of planted instances of the target motif in the case of synthetic sequence. Some commonly used metrics derived from comparing binding site matches on nucleotide and binding site level are discussed below in Section 5.1.3. Testing a motif discovery algorithm in its capacity to find motifs from unmodified biological sequence would perhaps seem as the most intuitive approach. However, to date, performance assessment with unmodified biological sequence has been limited to small numbers of individual genomic regions because of our limited knowledge of regulatory regions. Perhaps for this reason, synthetic regulatory sequence is often used, and is also the primary type of sequence used in the [Tompa et al. \(2005\)](#) assessment, detailed below. Regardless of the sequence type, the above assessment criteria also make the assumption that a motif inference algorithm should be able to partition sequences into binding sites and background sequence. The appropriateness of this partitioning assumption is also discussed below.

5.1.3 The Tompa *et al.* (2005) assessment

[Tompa et al. \(2005\)](#) compared 13 different motif inference methods in their ability to predict motif binding sites from mostly synthetic promoter sequence sets. The authors assessed the algorithms with summary statistics derived from motif hit instances predicted in the sequences. A thorough review of the assessment is provided here, because it is the most comprehensive performance assessment of its kind, and has been influential for performance assessments presented in later publications. It also suffers from a number of self-professed flaws, some of which I intend to address in the present work.

The binding site sequences used in their assessment were retrieved from the

TRANSFAC database (Matys et al., 2006), and inserted into a mixture of the types of background sequences: 1) randomly chosen promoter sequences from the same genome, or 2) sequences generated from a 3^{rd} order Markov chain. Unmodified binding site sequences are used in a third type of benchmark dataset. In total, 52 datasets were created for different TFs of fly, human, mouse, rat and yeast (one dataset per TF), and four negative control sequence sets created from the Markov chain background were added to the set. The benefit of testing algorithms with synthetic sequences (types 2 and 3) is the controlled environment they provide: inserted binding site positions are known, and motif frequency or sequence length can be varied at will. This is the reason that a benchmark with synthetic sequences, consisting of sampled TFBS hits in intronic background sequence, is also used in my work in Chapter 3 to allow the known motif frequency (sequence length) to be varied in a predictable way. Making sure that synthetic benchmarking sequence sets are realistic is not possible, especially in a genome scale problem, because of our limited understanding of regulatory sequences. In this case the background sequence is sampled from a 3^{rd} order Markov chain (trained from genomic sequence) in the Tompa et al. (2005) assessment are almost certainly not closely related to real promoter sequence in their properties (nucleotide content in genomic sequence varies in discrete regions, as discussed in Section 1.3.3).

At the nucleotide level, four types of measurements were defined, to measure the overlap of real binding sites with those predicted:

- **nTP**: the number of nucleotide positions in both known sites and predicted sites.
- **nFN**: the number of nucleotide positions in known sites but not in predicted sites.
- **nFP**: the number of nucleotide positions not in known sites but in predicted sites.
- **nTN**: the number of nucleotide positions in neither known sites nor predicted sites.

Similar metrics were also defined for binding site overlap, with an arbitrarily chosen 25% overlap required between the nucleotides of the sites to be considered overlapping.

Tompa et al. (2005) then defined a number of further statistics based on nTP , nFN , nFP , nTN . Firstly, sensitivity nSn , specificity nSp , and positive predictive value $nPPV$:

$$nSn = nTP / (nTP + nFN) \quad (5.1)$$

$$nSp = nTN / (nTN + nFP) \quad (5.2)$$

$$nPPV = nTP / (nTP + nFP) \quad (5.3)$$

A nucleotide level performance coefficient nPC , intended to “in some sense average (some of) [the above] quantities”, is also reported (Equation 5.4), following the work of Pevzner and Sze (2000).

$$nPC = nTP / (nTP + nFN + nFP) \quad (5.4)$$

Following Burset and Guigó (1996), the authors also report a nucleotide level Pearson product-moment correlation coefficient (Equation 5.5), and an average site performance $sASP$ (Equation 5.6).

$$nCC = \frac{nTP \times nTN - nFN \times nFP}{\sqrt{(nTP + nFN)(nTN + nFP)(nTP + nFP)(nTN + nFN)}} \quad (5.5)$$

$$sASP = (sSn + sPPV) / 2 \quad (5.6)$$

The measures nSn , nSp , $nPPV$, nPC , nCC , $sASP$ are then summarised in three different ways per tool across the datasets: either as an average, as a Z-score, or a ‘combined’ weighted average score where all the measures are computed as if the real and predicted sites were part of one large dataset instead of 56 individual ones. Most of the chosen performance measures however present problems with the four negative control datasets with no motifs: nSn , nCC ,

sASP are not defined, and *nPPV*, *nPC* and *sPPV* are uninformative. Most troubling however is that when a tool makes no prediction in datasets containing motif instances, $TP + FP = 0$, causing *nPPV*, *nCC*, *sPPV* to be undefined and *nSn*, *nPC* and *sSN* to be uninformative. The ‘combined’ average score works around this to the extent where these predictions consisting of entirely false negative predictions do not contribute at all. The score does however still penalise methods which make a small number of false positive prediction against those which attempt to make no predictions whatsoever (also pointed out by the authors). The statistics used also leave no intuition for how any of the tools performed on any individual dataset, and no guidance is given by the authors for the interpretation or relative importance of the various different measures.

A further problem with the above performance measures is that if the binding site positions called either positive or negative for a predicted binding event are dramatically affected by the motif significance thresholds used (high significance cutoff increases the false positive rate). Indeed, given that different experts ran the experiments, it is possible that this assessment tested not only the ability to detect recurring motifs with different algorithms, but also the stringency and parameter choices involved in deciding which of the potential binding site matches to report based on the inferred motifs. The problem of inferring a motif, and finding its binding site matches are independent in the formulation used by many motif inference algorithms. Some Bayesian motif inference algorithms do not in fact report individual binding site matches as part of the motif inference process ([Down and Hubbard, 2005](#)). Furthermore, when the above binding site level measures are computed for real promoter sequence with experimentally determined TFBSs, the quality of binding site data affects all of the above-mentioned measures. For example, some of the false positives can in fact be true, unknown binding sites.

The authors cite several gene finding assessments ([Burge and Karlin, 1997](#); [Bures and Guigó, 1996](#); [Reese et al., 2000](#)) as the inspiration for their approach. In those studies protein coding gene models are inserted to large sets of vertebrate sequence. I question the analogy between gene finding and TFBS finding, and advocate the use of comparison of motifs, rather than comparison of individual motif matches, as the primary means to benchmark motif inference performance.

TFBSs are several orders of magnitude shorter and lower in information, transient and turned over during evolutionary time scale, tend to co-occur, and vary in frequency and stringency of matches, depending on the TFBS in ways that are not well understood (see Section 1.1). Furthermore, weak binding sites which can be very ‘distant’ matches to the motif, and therefore both difficult to find experimentally or by scanning computational motifs, can also contribute to regulatory responses (Gertz et al., 2009). A motif match alone does not determine if a genomic position binds a TF or not; other levels of information relevant for regulation is stored in genomes, including for instance tissue specific epigenetic marks and the DNA melting propensity. Making use of such additional sources of evidence substantially improves classification of sites as either binding or non-binding Ernst et al. (2010); Lähdesmäki et al. (2008); Ramsey et al. (2010). For many eukaryotic TFs, even a perfect motif inference algorithm cannot predict its binding sites accurately, in turn raising questions about the use of binding site or nucleotide level based methods for their performance assessment.

The authors required the experts applying prediction methods to report a single high confidence prediction. Especially when inferring motifs from real-world genomic sequence, one cannot be sure of the absence of unexpected ‘real’ sequence motifs, which a good computational motif prediction tools should in fact be able to report. Indeed, the authors also state that “no attempt was made to eliminate sequences that might contain additional transcription factor binding sites, since our ability to identify such sites accurately is limited.” Therefore, methods which were (correctly) able to report additional motifs present in the sequence, but where the genomic matches of the correct motif was not submitted for analysis, can in fact be penalised for it heavily, perhaps explaining in part the reportedly bad prediction performance seen with the real sequences. Inferring motifs, and ranking them, should be considered independently. I would argue also that the algorithm assessment should be made with a collection of inferred motifs per method, instead of a single motif per method. Otherwise the assessment measures, in part, the correctness of post-processing and motif ranking steps which can be made by the experts – and were not detailed by the authors.

In conclusion, the design of the Tompa et al. (2005) study suffers from certain troubling assumptions and sources of potential bias. It is also inconclusive; the

authors do not offer direct advice or a ranking of methods based on the measures, and point out many of the study’s shortcomings also themselves. To my surprise, I have been unable to find later performance assessments which would directly try to address these shortcomings, apart from [Li and Tompa \(2006\)](#); [Sandve et al. \(2007\)](#) who mostly confirm problems apparent in the [Tompa et al. \(2005\)](#) assessment, but do not offer a new thorough assessment. On the contrary, several motif inference method publications after this paper have used the same statistical measures or synthetic datasets provided by [Tompa et al. \(2005\)](#), as supporting evidence for the favourable performance of their computational tools to previous work ([Chan et al., 2009](#); [Fauteux et al., 2008](#); [Gunewardena and Zhang, 2008](#); [Hu et al., 2006](#); [Klepper et al., 2008](#); [Lu et al., 2008](#); [Peng et al., 2006](#); [Reddy et al., 2007](#); [Robinson et al., 2006](#); [Sandve et al., 2008](#); [Wang and Zhang, 2006](#); [Wijaya et al., 2008](#); [Zare-Mirakabad et al., 2009](#)).

5.2 Materials & Method

This project had two phases: running a number of DNA motif inference algorithms on a large series of genomic sequence, and then assessing the discovered motifs. The sections below firstly describe the sequence sets used in the project (Section 5.2.1), before giving an account of the tested motif inference algorithms (Section 5.2.2). The remaining sections then detail the methodology of the various analyses conducted on the predicted motif sets. Notably, the performance assessment of methods is made in a parameter free manner when possible. Motif scanning with a motif hit significance cutoff parameter is done primarily for exploration of the data, for instance to find subsets of potentially interesting motifs which do not match the reference motif sets (Section 5.3.6.4).

5.2.1 Sequence and annotation retrieval

The *S. cerevisiae* promoter sequence used in all motif inference runs consisted of 200 base long upstream sequences from 1,000 randomly chosen protein coding genes with 5-way orthologs between the hemiascomycetous yeast species *S. cerevisiae*, *Candida glabrata*, *Kluyveromyces lactis*, *Debaryomyces hansenii* and

Yarrowia lipolytica. These sequence sets were collated by Dr Thomas Down. Briefly, Ensembl Compara (Birney et al., 2004) formatted database schemas were created of the genomic sequence data retrieved from the hemiascomycete comparative genomics database Genolevures (Sherman et al., 2004). BLASTP (Altschul and Gish, 1996) and reciprocal matching was then used to assign orthology between genes. *S. cerevisiae* sequences for orthologous genes were then retrieved, and a randomly selected subset of 1,000 200 bases long promoters chosen from the subset (other organisms were only used for selecting candidate genes).

I fetched additional sequence sets from the Ensembl database Hubbard et al. (2009) for the purposes of assessing the motifs (e.g. positional bias in Section 5.2.7.1, or the conservation analysis in Section 5.3.6.1). Most sequence fetching tasks were done from the Ensembl database with tools which I created with Dr Thomas Downs help using the BioJava toolkit Holland et al. (2008). Sequences for the assessment originated from version 57 of the Ensembl Core database. An usage example for the **nmensemblseq** retrieval tool is provided below:

```
nmensemblseq \  
-database saccharomyces_cerevisiae_core_57_1j \  
-host ensembldb.ensembl.org \  
-user anonymous \  
-port 5306 -noRepeatMask \  
-noExcludeTranslations \  
-proteinCoding -known \  
-fivePrimeUTR 500 0 -type protein_coding
```

The genomic coordinates for the sequence regions were also retrieved for the sequences, similarly using **nmensemblseq**, by adding the command line flag **-outputType gff**. A more thorough tutorial on using this utility, as well as some of the others included in the *nmica-extra* package I created during my project, are provided in Appendix B. The sequence retrieval tools were also integrated with the iMotifs sequence motif visualization and inference environment which I created during my project (Piipari et al., 2010b) (Figure 5.1A,B).

A)

B)

Figure 5.1: The sequence retrieval tools included in iMotifs. A) Configuration dialog for the 5' / 3' UTR sequence retrieval tool `nmensemblseq`. B) Configuration dialog for the GFF/BED sequence feature and ChIP-seq peak retrieval tools (`nmensemblfeat` and `nmensemblpeakseq`).

5.2.2 Motif inference

I tested predicting motifs with all of the thirteen motif inference algorithms from the Tompa et al. (2005) assessment, as well as SOMBRERO (Mahony et al., 2005b), PRIORITY (Narlikar et al., 2006), MoAn (Valen et al., 2009) and BayesMD (Tang et al., 2008). The Tompa et al. (2005) methods were chosen because it is perhaps the most comprehensive assessment to date, and the additional methods (NestedMICA, SOMBRERO, PRIORITY, BayesMD, MoAn) were tested because of their reported favourable performance in comparison to those tested in Tompa et al. (2005). The input parameters used for all of the successfully run algorithms are described in Appendix C. All inference experiments were made with the random orthologous promoter sequence set detailed in Section 5.2.1. If possible, each algorithm was made to predict 200 motifs. In case this was not possible, the largest motif set output by the tool was used for evaluation.

The PWMs output by each of the programs were converted to the XMS format used by the NestedMICA suite and iMotifs, with scripts that use the libxms Ruby bindings which I wrote Piipari et al. (2010b). Two of the algorithms which successfully returned results use a consensus string representation of their output (YMF and Oligoanalysis). These were converted to a PWM representation, applying a very small pseudo-count of 0.001 to the motifs.

I ran all of the motif inference programs myself after consulting the publications describing the algorithms, and other available documentation regarding each of them. This is in contrast with the Tompa et al. (2005) assessment, which was a large collaborative project where outside experts (the authors of the algorithms) created the motif predictions, which were assessed independently.

Conservation of noncoding sequence has been applied in some earlier studies as a means of selecting candidate sequences for motif inference (Elemento and Tavazoie, 2005; Hardison, 2000; Kellis et al., 2003; Xie et al., 2005). However, I decided not to choose or weight promoter sequences for my study according to conservation. There were several reasons for this decision. Firstly, leaving sequence conservation aside from the motif inference step allows it to be used as an independent way of assessing the motifs. Secondly, the traditional con-

ervation scoring methods, such as the PhastCons (Siepel et al., 2005) used in the present study, assume an alignment between the sequences; given the small alphabet size of DNA, and repetitive nature of genomic sequence, alignment errors are inevitable. Thirdly, biologically active TFBSs are known to be turned over quickly, and some experience near to neutral mutation rates (Kunarso et al., 2010; Schmidt et al., 2010). Although success has been reported in studies using conservation as a criterion of choosing motifs amongst candidates (Xie et al., 2005), it does not always lead to detection of correct ones. For instance, Li et al. (2005) suggest that a simple conservation based significance score would lead to the selection of an incorrect TFBS motif in 28% of cases with yeast ChIP-chip data of Lee et al. (2002).

The rate of binding site turnover has been studied in high resolution with ChIP-seq assaying in the CEBPA and HNF4A transcription factors, which are strongly conserved across placental mammals (Schmidt et al., 2010). Less than 0.3% of binding events were shown to be conserved in all assayed species. A study by Kunarso et al. (2010) finds that in the case of Oct4 and Nanog, 2.0% of sites are conserved in sequence. The binding regions however are functionally conserved at a much higher rate of between 50% and 10% depending on the chosen stringency of statistical significance. The strength of binding was not seen to associate with conservation, suggesting that the wide binding site spectrum of TFs is important (Schmidt et al., 2010), and that weak binding sites can have a biological effect. Several studies of human (Kasowski et al., 2010; McDaniell et al., 2010) and yeast (Zheng et al., 2010) individuals and related yeast species (Borneman et al., 2007) have shown results pointing in the same direction: individual TFBS events undergo rapid divergence, but a weak conservation signal tends to be found from a collection of TFBSs. The excess conservation of motifs is considered here, in combination with other lines of evidence, as a potential sign of function for computationally predicted motifs.

5.2.2.1 Unsuccessfully run algorithms

Several motif inference programs which were assessed in the Tompa et al. (2005) assessment by the authors of each of the algorithms were unsuccessfully attempted

to be used in the assessment, due to various reasons. Firstly, ANN-SPEC (Workman and Stormo, 2000) and Improbizer (Ao et al., 2004) are not distributed in binary or source code form without request from their authors, and the web servers provided are not suitable for discovering motifs on a genome scale. MITRA (Eskin and Pevzner, 2002) was not available at the URL noted by the authors¹, and no suitable online prediction server was found. QuickScore (Egner, 2004) is only available as an online prediction server, and it was found not to handle the large (200,000nt) input sequence size. CONSENSUS (Hertz and Stormo, 1999) failed to compile on either 32 or 64 bit Linux or Mac OS X with the available compiler versions (gcc 4.2 and 4.3), and I was unable to find a binary distribution, or an online CONSENSUS prediction server suitable for the large analysis task at hand.

MoAn (Valen et al., 2009), PRIORITY (Narlikar et al., 2006), and SeSiMCMC (Favorov et al., 2005) were each successfully run with example data sets, but each only allowed for a single motif to be estimated.

BioProspector (Liu et al., 2001) was attempted to be run (`BioProspector -i orthologs-sc-1000.fa -r 200 -f yeast_all.bg -n 100 -h 1`). The currently distributed version of the program² does not parse the FASTA files used in the assessment. The file did appear to conform to the required variant of the file format given in the program's example file, and all of the other attempted tools processed it without problems. Furthermore, the BioProspector web server (<http://robotics.stanford.edu/~xslui/BioProspector/>) only allows reporting a maximum of ten motifs (and its documentation specifically warns against specifying too large an input sequence set), which made it inapplicable for this benchmark (the target is 200 motifs). The same reason also made it impossible to run MDscan from the same authors (Liu et al., 2002)³.

The Bayesian motif inference method BayesMD, which reportedly performs better with long promoter sequence than NestedMICA (Tang et al., 2008), was also tested, but it failed to report any output motifs due to persistently running

¹<http://www.cs.columbia.edu/compbio/mitra>

²'BioProspector.2004.zip', downloaded 1st June, 2010 from <http://motif.stanford.edu/distributions/bioprospector/>

³'MDScan.2004.zip', download made 1st June, 2010 from <http://motif.stanford.edu/distributions/mdscan/>

out of runtime memory, even with cluster nodes with 15.5G of allocatable memory.

5.2.3 Motif comparison

The computationally inferred *S. cerevisiae* motifs were compared to two different, partially overlapping reference sets of regulatory motifs: the JASPAR 2010 database (Portales-Casamar et al., 2010), and the Zhu et al. (2009) PBM motifs (some of which are included in the JASPAR dataset). The discovered motifs were also compared against one another to measure the level of redundancy across the sets.

To study the capacity of each of the motif inference methods to detect motifs that resemble known regulatory motifs, I compared them to motifs in the JASPAR 2010 database (Portales-Casamar et al., 2010). The JASPAR fungal motif dataset was chosen as the primary gold standard comparison set because it covers the great majority of all *S. cerevisiae* transcription factor motifs (177 TFBS non-redundant motifs in the database). It is an open access database, and its curation appears to be of more uniform quality than its competitor TRANSFAC which suffers from infrequent missing annotations such as species or publication references. Furthermore, JASPAR 2010, unlike previous versions of the database, includes a high coverage, non-redundant¹ set of *S. cerevisiae* motifs. The dataset originates mostly from two large scale studies; The single largest set included, and one preferred by Portales-Casamar et al. (2010) in case of conflicts, is the set of motifs from a study by Badis et al. (2008). This study includes data for a total of 112 TFs (107 of which are included in the non-redundant dataset, see Figure 5.2) from a combination of universal protein binding microarray assays (Berger et al., 2006; Mintseris and Eisen, 2006), cognate site identifier (CSI) microarrays (Warren et al., 2006), and DIP-chip (Liu et al., 2005) assays. The second large dataset included in JASPAR 2010 is the PBM based study by Zhu et al. (2009) (89 motifs). The remaining motifs from two datasets containing primarily literature based motifs from the SCPD binding profile database and literature (Zhu and Zhang, 1999), and the ChIP-chip based SwissRegulon database (Pachkov et al.,

¹In this context, non-redundant means that only one motif prediction is included in the set for each TF.

2007) as well as computationally inferred motif dataset from the genome-wide ChIP-chip study of *S. cerevisiae* by MacIsaac et al. (2006). The motif comparisons presented in this chapter rely on these original studies and the manual curation conducted for the JASPAR database.

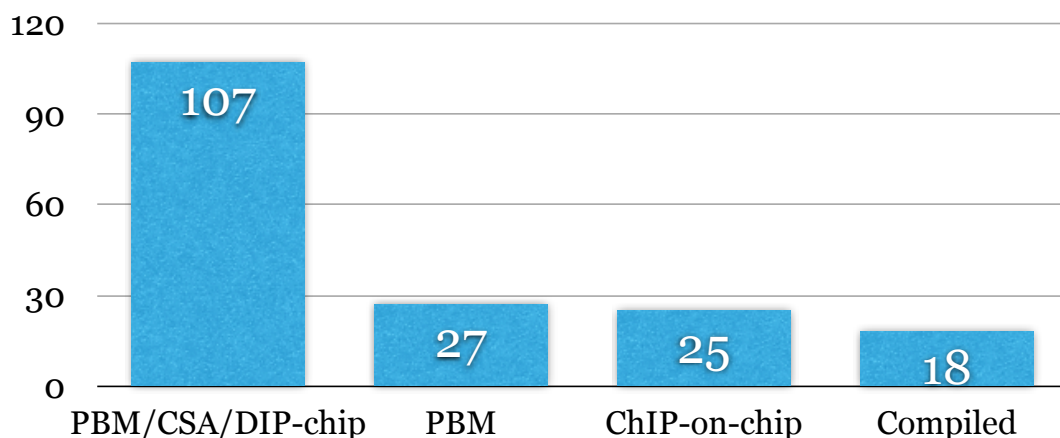


Figure 5.2: The number of motifs from different experimental sources in the JASPAR 2010 non-redundant fungal motif dataset. Note that some datasets contain motifs for TFs covered by other datasets. The PBM/CSA/Dip-chip dataset of Badis et al. (2008) for example contains in total 112 motifs, but only 107 of these are used in the non-redundant dataset by Portales-Casamar et al. (2010).

I also compared the inferred motifs to the Zhu et al. (2009) PBM motifs because they form the highest coverage regulatory motif dataset originating from a single type of experiment in the *S. cerevisiae*; the Badis et al. (2008) dataset with 112 motifs is in fact larger than the 89 motifs estimated by Zhu et al. (2009), but Badis et al. (2008) apply a combination of three different high-throughput methods, rather than one. A reference dataset additional to JASPAR was useful also because some of the JASPAR motifs could in fact originate from one of the tested algorithms (the 25 ChIP-on-chip and 18 ‘other’ motifs in JASPAR are suspect). In contrast, the Zhu et al. (2009) motifs are all estimated from PBM data with the Seed-and-wobble algorithm (Berger et al., 2006), and these data should therefore not suffer from circularity in the comparison of the *de novo*

predictions to a reference.

5.2.3.1 Motif clustering with the SSD metric

The pairwise sum of squared differences (SSD) metric between PWMs, introduced by [Down et al. \(2007\)](#) (Equation 2.7 in Section 2.2.3), was computed systematically between all pairs of motifs. The distance matrix between all inferred motifs and JASPAR reference motifs were computed. Motif-to-motif distances allowed probing the redundancy of motifs within inferred sets with complete linkage clustering ([Johnson, 1967](#)). All motif sets were also clustered together with the JASPAR reference set, to summarise and visualise the trends in motif types found by each of the algorithms.

5.2.4 Motif scanning

After predicting sequence motifs with a selection of motif inference algorithms from the putative *S. cerevisiae* promoters, I scanned all putative promoter sequences of lengths 200bp, 500bp and 2000bp for the inferred motifs using the `nmscan` program included in the NestedMICA suite ([Down and Hubbard, 2005](#)). Sequences on the reverse strand with respect to the reference genome were reverse-complemented. The 200bp and 500bp sequence ends were aligned to the TSS. 2000bp sequences were centered on the TSS (i.e. they contain 1000bp upstream and 1000bp downstream sequence). The motif bit score function evaluated by `nmscan` for all PWMs W at positions p in sequence S is explained in Section 1.2.1 (Equation 1.1).

I also scanned all the sequences again to report the maximum bit score achieved in 200nt and 500nt upstream sequence regions of all *S. cerevisiae* genes (the `-maxPerSeq` mode in `nmscan`). Maximum bit scores were computed because they allow a parameter free comparison of score distributions between groups of promoters (genes). In Section 5.2.6 the maximum bit scores achieved by promoters are used to compare putative target genes of TFs to non-target genes (to see if the maximum bit scores discriminate TF targets from non-targets).

The match positions identified are dependent on the choice of the bit score threshold chosen for each of the motifs. Finding a meaningful statistical measure

of significance for motifs found from genomic DNA sequence itself is an active research problem. Approximate (Thijs et al., 2001) and even exact P -value calculation of PWM matches in DNA sequence (Zhang et al., 2007) is possible for PWMs given a sequence background with independent and identically distributed (i.i.d.) nucleotides, but i.i.d. is not a realistic model of background genomic DNA (Section 1.3.3). I therefore used a method for assigning the significance threshold of motif hits which can account for varying DNA dinucleotide content (Down et al., 2007).

In brief, the significance scores are computed with respect to a 1st order mosaic sequence background model. I compare the score distribution of k -mers drawn from a 1st order Mosaic sequence background model to the motif matches in each bin (both the expected and the observed score distribution are binned on 1 bit intervals). The benefit of this approach is that it allows a comparison to be made to a more representative background model of nucleotide sequence than what is commonly done (with a GC-content based background model). The drawback is that the computation is not exact, and the scores are reliant on the score bin sizes, and the total number of hits. This led to some difficulties, particularly with the motifs output by MEME, which are discussed below.

The total number of motif hits identified at different confidence thresholds varies dramatically. For instance, in the case of the 200 motifs predicted by NestedMICA, the total genomic hit count in 200 base upstream sequences ranges from 47,312 with the 0.01 confidence threshold to 139,312 hits with the 0.05 threshold. All analyses presented here were made with a stringent 0.01 cutoff.

5.2.5 Predicted binding site overlap

I computed the overlap between matches of different motifs within the inferred sets, and with the JASPAR database motifs, with a score similar to the one used by (Down et al., 2007) (Equation 5.7). In brief, the overlap score O , between binding sites B_1 of motif 1 and binding sites B_2 of motif 2, is the fraction of overlapping predicted sites which are hits for motif 1. O is 0 when the sets are disjoint, and 1 when a motif matches all of the other ones sites.

$$O = \frac{|B_1 \cap B_2|}{\min(|B_1|, |B_2|)} \quad (5.7)$$

This allows the detection of similar motifs within the inferred motif sets, and also between the inferred and the experimentally validated JASPAR motifs. The overlap scores were considered for binding sites at the 0.01 significance cutoff (see Section 5.2.4 for discussion of determining motif hit significance). Overlapping motifs were analysed in an orientation independent manner, simply as chromosomal coordinate ranges with no strand information. This was done because all of the motif inference algorithms were run in a mode which allows for matches of a motif to occur in either orientation.

5.2.6 Association of motif hits to transcription factor target genes

A set of target genes is known for the great majority of *S. cerevisiae* regulatory TFs. For many of them, there is also an experimentally verified DNA motif in the JASPAR database. This makes it possible to judge if high-scoring matches of the predicted motifs distinguish target promoters of their likely TFs from non-target promoters. That is, for each computationally predicted motif with a closely related known TFBS motif, I test if the distribution of its maximum scoring occurrences differs between targets of the likely TF genes, and non-target genes.

I considered three different TF target gene datasets in this work. These datasets were:

1. **YEAst Search for Transcriptional Regulators And Consensus Tracking** database (Teixeira et al., 2006). Introduced in Section 5.2.6.1.
2. TF target calls from a reanalysis (Reimand et al., 2010) of a sequence specific TF knockout expression dataset Hu et al. (2007). Introduced in Section 5.2.6.2.
3. The Harbison et al. (2004) dataset of genome-wide location analysis by ChIP-chip (Iyer et al., 2001; Lieb et al., 2001). Introduced in Section 5.2.6.3.

For all of the target gene sets (introduced below), I extracted the curated TF–target dataset for all of the factors which also had a corresponding motif available in the JASPAR database. For each of these JASPAR motifs, I then calculated the closest motif from each predicted motif set (using the SSD distance metric by [Down et al. \(2007\)](#)). Maximum bit scores of the computationally predicted motifs were then compared in 500 base upstream regions of the *S. cerevisiae* genome using a two-sample single-tailed Kolmogorov-Smirnov (KS) test. The target genes of the TF, and the non-target genes, were the two different sets whose maximum bit score distributions were compared for each motif. In the KS test a low p-value indicates skewing of the bit score distribution of TF target promoters to the high bit-score end when compared to non-target genes. In addition to the two-sample KS-test, the rank-based two-sample Mann-Whitney (MW) test was computed for the maximum bit score distributions to see if the ranks of the maximum motif bit scores would be higher amongst the TF target genes. The non-parametric KS and MW tests were used due to the non-normal shape of the maximum bit score distribution.

5.2.6.1 YEASTRACT

YEast Search for **T**ranscriptional **R**egulators **A**nd **C**onsensus **T**racking database is a curated repository of transcriptional regulatory interactions in the *S. cerevisiae* genome ([Teixeira et al., 2006](#)). It currently collates a total of 12,346 TF–target associations for 149 TFs, each derived from one of a number of possible experimental sources, described in as many as 861 primary publications (download date 18/3/2010). The possible lines of evidence accepted as support of a target association in it are either:

1. change in the expression of the gene of interest owing to deletion or mutation of the TF gene (as measured by either gene by gene or genome-wide microarray).
2. binding of the transcription factor to the promoter region of the target gene, as supported by a band-shift assay ([Fried and Crothers, 1981a](#)), DNase footprinting ([Brenowitz et al., 1986](#)), or ChIP assaying ([Harbison et al., 2004](#)).

In other words, the evidence sources in this dataset range from detailed individual genetic or physical interaction studies to high throughput ChIP-chip experiments.

5.2.6.2 Reimand *et al.* (2010) TF knockout and expression data based target set

Reimand *et al.* (2010) present a reanalysis of the sequence specific TF knockout expression dataset by Hu *et al.* (2007) of 269 sequence specific regulatory factors, including both general and specific TFs and factors involved in regulating chromatin state. The re-analysed dataset applied a series of corrections and processing steps to the expression data which were not made by original authors. These include a correction for non-specific background and print-tips (Huber *et al.*, 2002), as well as correction for multiple-testing which was not made by false-discovery rate estimates (Reiner *et al.*, 2003). TF target calls made by Reimand *et al.* (2010) were downloaded from the ArrayExpress database (Parkinson *et al.*, 2009). Genes called as targets for a TF have a highly significant expression difference between the knock-out and the wild-type, with a 0.05 p -value cutoff. The problem of possible indirect targets being included amongst the predicted target genes is however not directly addressed by Reimand *et al.* (2010).

5.2.6.3 Harbison *et al.* (2004) ChIP-chip dataset

The Harbison *et al.* (2004) dataset of genomic occupancy of 203 TFs is a result of genome-wide location analysis by ChIP-chip (Iyer *et al.*, 2001; Lieb *et al.*, 2001). They made measurements in a number of growth conditions (1 to 12 conditions, depending on the TF). I use a re-analysis of the Harbison *et al.* (2004) dataset by MacIsaac *et al.* (2006). This dataset contains lists of ORFs likely to be regulated by the TFs, based on conservation in other related yeasts, and a significance cutoff of the signals identified close to the ORFs in the ChIP-chip measurements. The analysis I present was made with the most stringent dataset provided by MacIsaac *et al.* (2006): ChIP-chip signal significance $p < 0.001$, with the binding site conserved in at least 2 other yeast species.

5.2.6.4 Relationship between discovered motifs and inter-species sequence conservation

The relationship between discovered motifs and sequence conservation were studied with 7-way phastCons conservation scores (Nielsen, 2005; Siepel et al., 2005) derived of an alignment of the *S. cerevisiae* genome with genomes of six other *Saccharomyces* species (*S. paradoxus*, *S. kudriavzeii*, *S. bayanus*, *S. castelli*, and *S. kluyveri*). The phastCons scores were retrieved from the UCSC Genome Browser FTP server (sacCer2 conservation track, available at <ftp://hgdownload.cse.ucsc.edu/goldenPath/sacCer1/phastCons/>, downloaded on 12/02/2010).

The conservation scores of motif match positions at the stringent confidence cutoff of 0.01 were contrasted with phastCons scores of 10,000 randomly sampled intergenic regions of the same lengths (10,000 regions were sampled at all lengths between 6 and 20 nucleotides). The random intergenic regions were sampled and retrieved from Ensembl (Hubbard et al., 2009) with the help of tools I wrote as part of the project. See Appendix B for usage examples for some of the tools included in the nmica-extra toolkit. The difference in conservation score distributions of the motif matches and random intergenic sequences were measured with the single-tailed two-sample Kolmogorov-Smirnov test.

5.2.7 Relationship between discovered motifs and sequence variation in *cerevisiae* strains

The *S. cerevisiae* reference genome was the first eukaryotic genome to be published (Goffeau et al., 1996; Mewes et al., 1997). Because the budding yeast is so amenable for genomic study and manipulation, and because its association to human activity and migration, its genetic variation in and between its different populations has also been studied. Large genetic studies began from typing microsatellites of over 600 *S. cerevisiae* strains (Legras et al., 2007). In this work I however use the more recent whole genome sequencing data from 42 *S. cerevisiae* strains conducted by the *Saccharomyces* genome resequencing project (SGRP) (Liti et al., 2009). This study presents the 1x to 4x coverage whole-genome capillary sequencing of the *S. cerevisiae* strains. Genotypes reported by Liti et al. (2009) for individual positions in the multiply aligned strains were imputed using

ancestral recombination graphs (Minichiello and Durbin, 2006) and the sequencing traces, instead of ‘trusting’ the base calls alone. On top of the low coverage sequence, the PALAS alignment method built for assembling and aligning the low coverage sequences is not a principled, probabilistic method with predictable properties, but instead an ad hoc iterative algorithm. The common occurrence of binding sites with large numbers of mismatches in aligned binding site matches suggested that alignment errors were prevalent (Edmund Duesbury, personal communication), especially between the *S. cerevisiae* and *paradoxus* strains. Because of the limitations of the low coverage data and the SNP calls derived from it, I resorted to a simple comparative study between the SNP rates in binding sites when compared to intergenic sequence, with the aim of detecting motifs with likely function (those which show lower SNP rate than intergenic sequence). Only the *S. cerevisiae* strains were considered (no *S. paradoxus* strains), with two or less SNPs per regions of interest, as well as filtering out SNPs with less than 1×10^{-6} error probability. Putative TFBS matches with more than two SNPs were rejected because they are most likely caused by misalignments.

I applied a simple bootstrapping based statistical test to assess the significance of the difference of SNP rates seen in motif matches and random intergenic regions of the matching length. This was done for each predicted motif by counting the number of SNPs in a randomly chosen sub-selection of binding sites of the same length as the motif, and repeating this 10,000 times. The number of binding sites in each of the 10,000 random intergenic region sets was matched to the number of motif hits above the significance cutoff of 0.01. The significance score was derived as the fraction of the 10,000 sets where the mean SNP rate was higher than that observed for the motif’s binding sites. Higher coverage Solexa based resequencing data, which (at the time of writing) is expected soon, could allow a more detailed analysis, for instance using the mutation spectra of motifs.

5.2.7.1 Positional bias of motifs

Regulatory motifs often match positions close to transcription start sites. Many cases of characteristic positional biases have been described for TFs, especially for elements bound by the general TFs, such as TATA-box (at around -30) or the

B-recognition element (BRE) which is found immediately upstream from TATA (Lagrange et al., 1998). An inverse linear association between the distance of the binding site to the TSS and its effect on gene expression has been suggested based on an *in vivo* study of factors acting in the liver and the immune system (MacIsaac et al., 2010). An earlier *in vitro* study of differently spaced Gal4 activator sites upstream to Gal4 also suggest a simple inverse relation between the distance of binding site to the transcription start site and its gene expression activating effect (Ross et al., 2000). I therefore analysed the positional bias of the computationally discovered motifs as an indicator of potential function.

I counted the motif matches in all matches overlapping 100-base windows between -1000 to 1000 from the TSS of all known protein-coding genes in the *S. cerevisiae* genome, and tested for the enrichment of sites within the region -500–0 with respect to the TSS, compared to sequence regions outside this window. I used the exact one tailed binomial test with the null hypothesis success probability of 0.25 (the interval -500 to 0 covers a quarter of the 2000 base sequence length of interest). The interval was chosen because it is expected to contain the great majority of *S. cerevisiae* TFBSs (Venters and Pugh, 2008).

5.2.8 Classification of motifs with metamatti

Metamotifs were constructed from the JASPAR 2010 motif dataset similarly as described in chapter 4: motifs were labelled with their structural class, and clustered at cutoff 4.0 (complete linkage clustering) using the SSD metric from Down et al. (2007). However, in this classification exercise I did not use the structural classification terminology from the TRANSFAC database, but instead the binding structural mode taxonomy introduced by Luscombe et al. (2000), which is included for majority of motifs in JASPAR 2010. The Luscombe et al. (2000) classification terminology describes ‘classes’ and ‘families’ for TFs. Classes are defined by a manual, visual comparison of protein structures, and families by a computational clustering of the domain structures with the SSAP secondary structure alignment algorithm (Orengo and Taylor, 1996).

The JASPAR database was used for building a *S. cerevisiae* motif classifier because it contains the largest selection of high quality training data for the *S.*

cerevisiae genome; The emphasis in TRANSFAC is on vertebrate genomes, and as of version 12.2 its non-redundant coverage of the *S. cerevisiae* genome is only 43 as opposed to 177 motifs in JASPAR 2010. As described in Section 1.1.1, eukaryotic genomes have experienced lineage specific expansion of TF domains. Therefore for an accurate organism specific TFBS motif classifier it important to have a good coverage of the domains that are present in that genome. For example in the case of *S. cerevisiae* the largest domain class is that of zinc coordinated domains, especially the fungal specific zinc cluster (Macpherson et al., 2006) (47 of 99 *S. cerevisiae* zinc finger motifs belong to this family, and very few are present in TRANSFAC).

Metamotifs were trained from each of the motif clusters with `nmmetainfer` (minimum length 6, maximum length 15) and metamotif density features were then computed per training set motif as described in Section 4. Based on the classification labels and probabilities that the random forest classifier produces, I computed a precision-recall curve using the ROCR R package (Sing et al., 2005), and applied a probability cutoff to the classification decisions such to provide a high confidence labelling of motifs.

5.3 Results & Discussion

I apply eight motif inference tools in this work primarily as a genome scale performance benchmark. To my knowledge, these algorithms have not been judged before on problems involving the prediction of large motif collections from promoter sequence. The rationale in the assessment is simple: a well performing *de novo* motif discovery algorithm should find as many as possible motifs closely matching known TFBS motifs in the *S. cerevisiae* genome (Section 5.3.2).

5.3.1 Properties of inferred motifs

The motifs predicted by different computational methods were found to differ clearly by visual inspection. A selection of the top matches between the inferred motifs and motifs in the JASPAR database are shown in Figure 5.3. The closest matches identified vary considerably between different methods. The familial

patterns of motifs found by different methods is also apparent amongst the closest matches; MEME, in particular, shows clear preference towards discovering GC-rich fungal Zn cluster motifs, whereas SOMBRERO and NestedMICA show more variability amongst the closest matches.

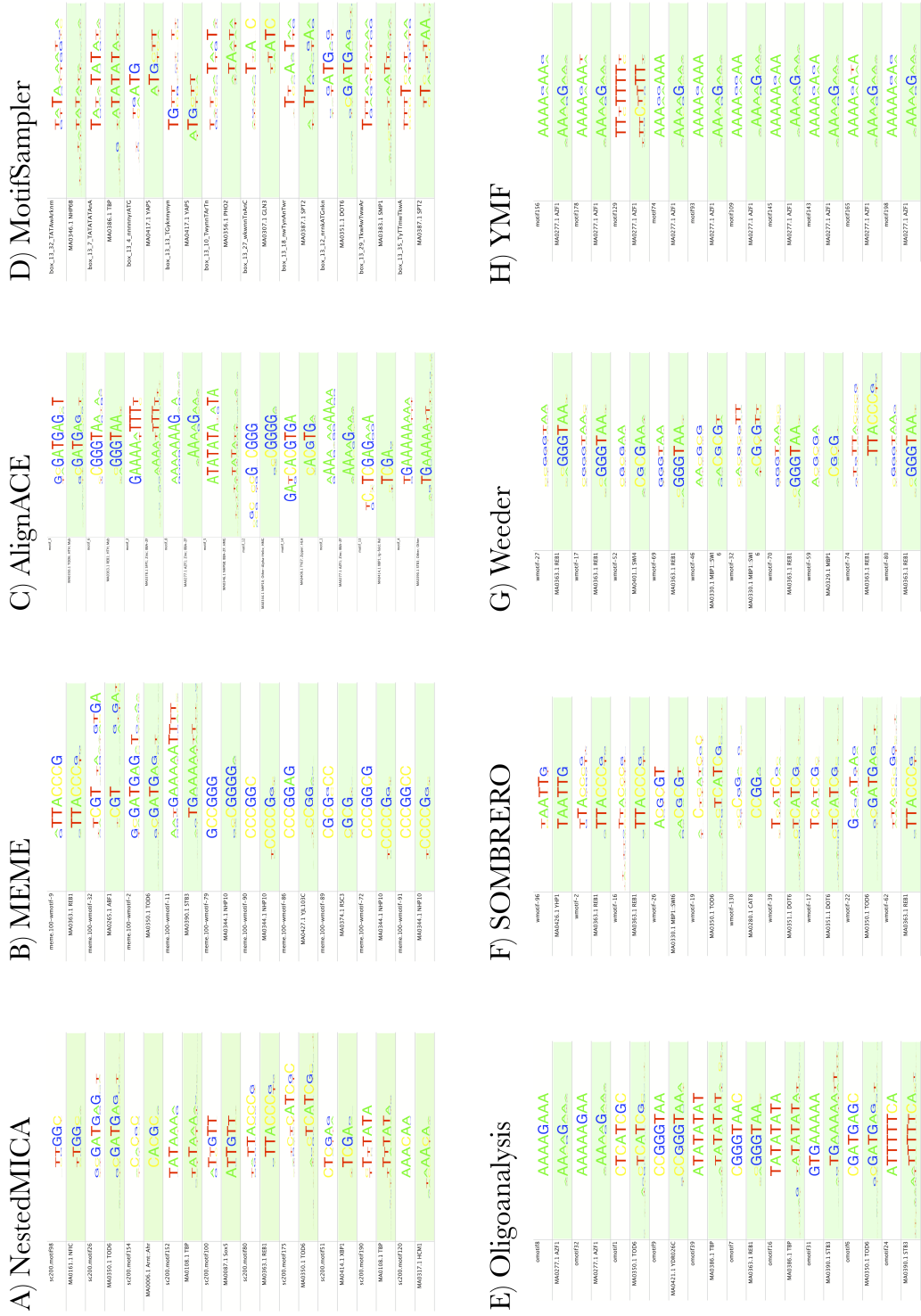


Figure 5.3: The ten closest matches between inferred motif sets, and JASPAR motifs. The JASPAR motifs are shown on green background.

The lengths, information contents and column wise average information contents are summarised for reference motifs and all inferred motif sets in Table 5.4. NestedMICA predicts the shortest motifs (6.6 columns), whereas Weeder has both the smallest information content (7.1 bits) and lowest per-column information content (0.9 bits per column). In contrast, MEME’s motifs are almost twice as long as those of NestedMICA, at 12.6 columns, and they have the highest information content (over three times as high on average as motifs predicted by Weeder, at 21.7 bits). It should be noted that these motif set summary statistics and the relative performance measures reported in the following sections also depend on the chosen input parameters (Appendix C).

In terms of information content, the methods are divided to two groups: SOMBRERO, MotifSampler, NestedMICA and Weeder all predict motifs with smaller information content than their closest JASPAR matches, whereas AlignACE, Oligoanalysis, MEME and YMF have higher information content. The median per column information content is slightly higher with the JASPAR motifs with all but Weeder and MotifSampler. The combination of short motif lengths, with less information in total but with higher per-column information could be explained by the computational motifs lacking ends with low information columns, which are common in the experimentally verified motifs. The systematically low information content seen in the case of Weeder and MotifSampler is apparent already by visual inspection of the sequence logos: the columns tend to be less constrained than those in the reference set, or those output by the other methods.

Oligo-analysis and YMF results are included in this study for the sake of completeness: both are word enumeration based methods, and therefore not strictly comparable to the other methods which output a PWM, but they could be run also on my benchmarking dataset. Oligo-analysis motifs are in fact individual 8-mers (not IUPAC consensus strings, like those predicted by YMF). This inflates its information and per-column average information content measures shown in Table 5.4.

Motif set	Average length	Information content	Average column info content
NestedMICA (200 motifs)	6.6	9.5	1.5
AlignACE (16 motifs)	11.6	17.2	1.5
MEME (100 motifs)	12.6	21.7	1.7
MotifSampler (37 motifs)	10.0	10.0	1.0
Oligoanalysis (50 motifs)	8.0	16.0	2.0
SOMBRERO (200 motifs)	9.4	9.6	1.1
Weeder (200 motifs)	8.3	7.1	0.9
YMF (200 motifs)	8.6	14.2	1.6
JASPAR (177 motifs)	10.3	11.6	1.3
Zhu et al. (2009) PBM motifs (89 motifs)	9.6	11.7	1.3

Figure 5.4: Summary of the average lengths and information contents of the different inferred motifs, and the two reference datasets (JASPAR and [Zhu et al. \(2009\)](#) PBM motifs, shown on a grey background in the bottom).

5.3.2 Finding matches to known regulatory motifs amongst *de novo* motif discoveries

The number of JASPAR motifs with matches in each of the predicted motif sets ($p < 0.05$) are shown in Figure 5.5. Results appear to be rather consistent with two different reference databases (JASPAR in Figure 5.5A, and [Zhu et al. \(2009\)](#) PBM motifs in Figure 5.5B). The top performers, by a clear margin, are NestedMICA (54 matches to JASPAR amongst its 200 motifs, 44 matches with 100 motifs), MEME (39 matches) and SOMBRERO (38 matches). NestedMICA was tested with two different motif set sizes, in part to measure its robustness with differing motif count, and also to allow direct comparison with MEME which was incapable of predicting more than 100 motifs. AlignACE reports a mere 16 motifs, but surprisingly, these map to 31 JASPAR motifs; almost all of the motifs predicted by AlignACE are in fact contributing to the JASPAR matches (14 out of 16 motifs). With the ([Zhu et al., 2009](#)) PBM motifs as a reference, NestedMICA is consistently the top performer, with SOMBRERO outperforming MEME.

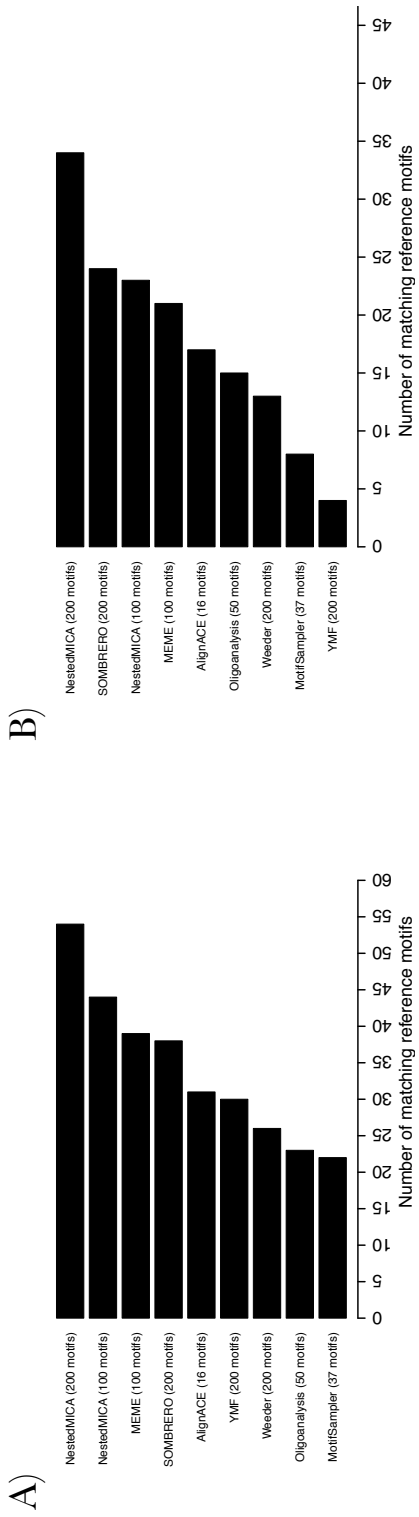


Figure 5.5: The number of statistically significant matches of the predicted motifs with A) JASPAR, and B) Zhu et al. (2009) PBM motifs.

Counting reciprocal matches between the predictions and the reference motifs is a more stringent way to assess motif relatedness (Figure 5.6). This measure penalises motif sets containing several closely related motifs. Some of the motifs amongst the reference motif sets are also highly similar to one another. NestedMICA also tops this ranking. With the JASPAR dataset of 177 motifs, it has 14 reciprocal matches, with SOMBRERO behind it, again with a clear margin (10 reciprocal matches) and MEME and AlignACE third (both with 6 reciprocal matches). Note again that the AlignACE program, which outputs a small motif set and has little redundancy in its predictions (Sections 5.3.4 and 5.3.5), is more likely to perform well by chance in this comparison than MEME with 100 motifs with several closely related motifs. Overall, the most likely reason for low numbers of reciprocal matches seen is due to the partial redundancy and large size of the experimental and inferred motif sets. NestedMICA however outperforms MEME and AlignACE also with a 100 motif count which matches that of MEME (9 reciprocal matches). There is little qualitative difference between the rankings with JASPAR or [Zhu et al. \(2009\)](#) PBM dataset as the reference.

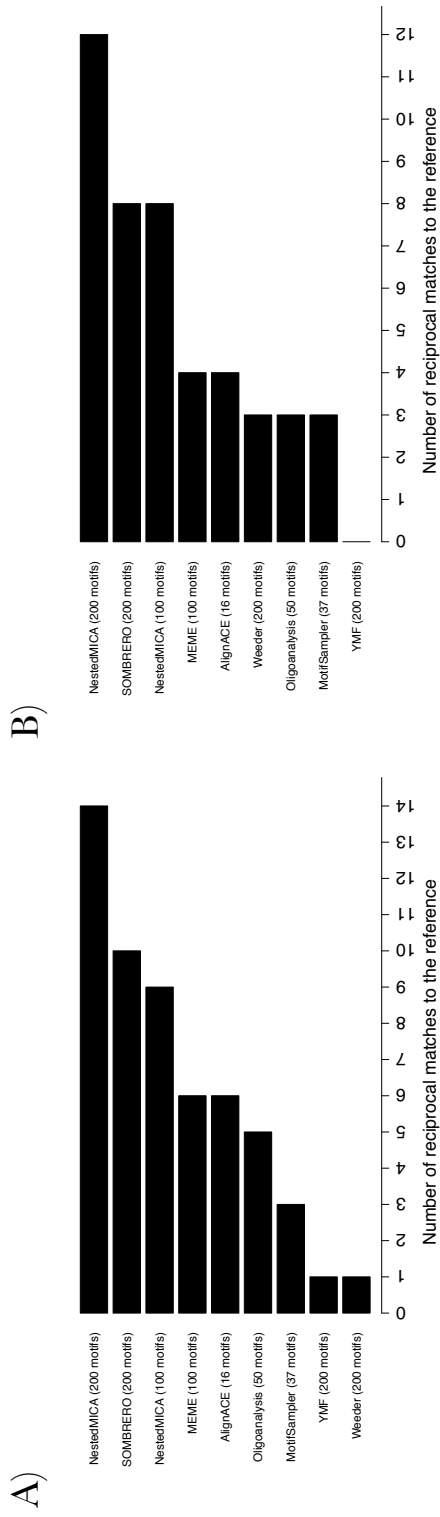


Figure 5.6: The number of reciprocal matches between the predicted motifs and A) JASPAR, and B) Zhu et al. (2009) PBM motifs.

The significant JASPAR and PBM motif matches suggest NestedMICA, SOMBRERO and MEME as the top performing methods. I also studied the overlap between the reference motifs covered by the different methods. I did this by computing the numbers of overlapping motifs between the top performers with the JASPAR motifs (Figure 5.7). NestedMICA has the highest overlap with the two other top performing methods (13 overlapped with SOMBRERO, and 9 with MEME). The number of motifs predicted by it and not covered by the other top performers (22 motifs) is also higher than either of MEME or SOMBRERO (14, and 9 motifs respectively), suggesting it covers more reference motifs than either of the other two top performers. Ten JASPAR motifs are found by all of SOMBRERO, NestedMICA, and MEME.

The number of statistically significant matches is informative of the extent to which the predictions cover the reference motif sets with detectably related motifs. The distribution of SSD distances between the inferred motifs, and their significant reference motif matches however also varies between algorithms (Figure 5.8). These results are consistent with above ranking in that NestedMICA also tends to have the shortest median distance, with SOMBRERO ranking the second. Once again the top performers are also consistent between the two different reference motif sets (JASPAR and the [Zhu et al. \(2009\)](#) PBM motifs).

The substantial disjunction of discoveries between the top-performing NestedMICA, MEME and SOMBRERO suggests that differences exist in the types of motifs that different algorithms are capable of finding. To study this further, I visualised the JASPAR dataset matches as a heatmap of matching or non-matching states, labelling the JASPAR motifs with its associated structural taxonomy of TFs, and clustering the motifs (Figure 5.9).

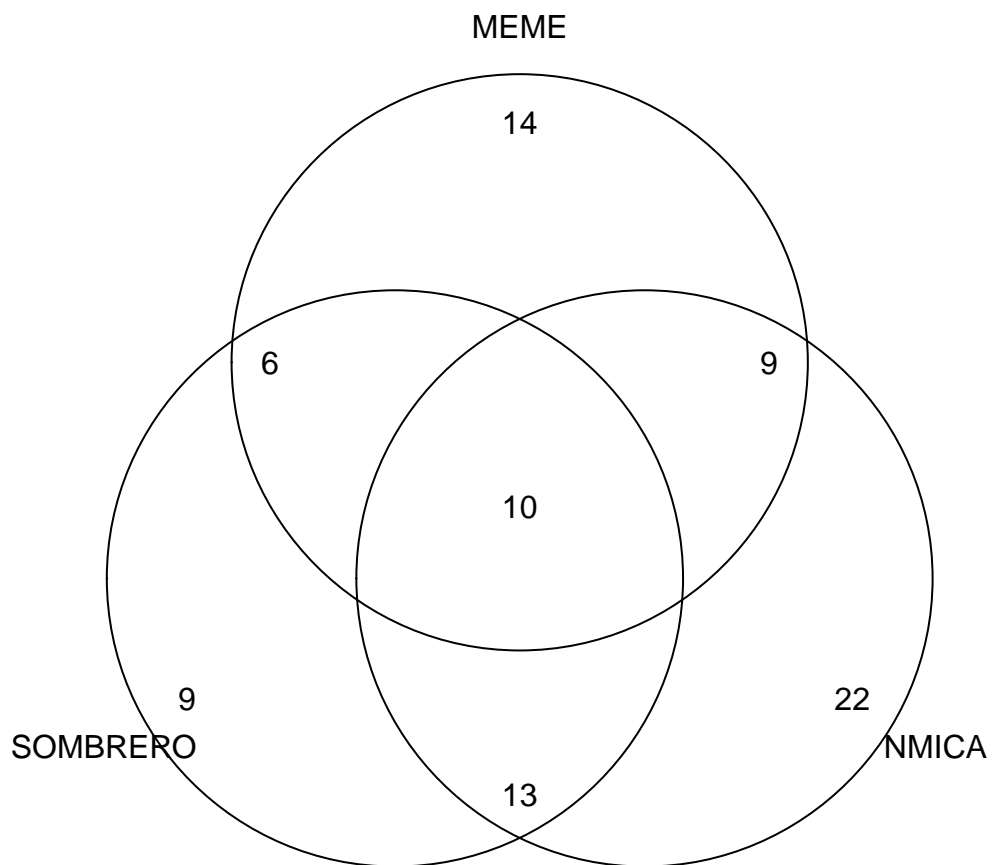


Figure 5.7: Overlap of significant matches to the JASPAR database between the three top performing motif prediction methods: NestedMICA, MEME and SOMBRERO.

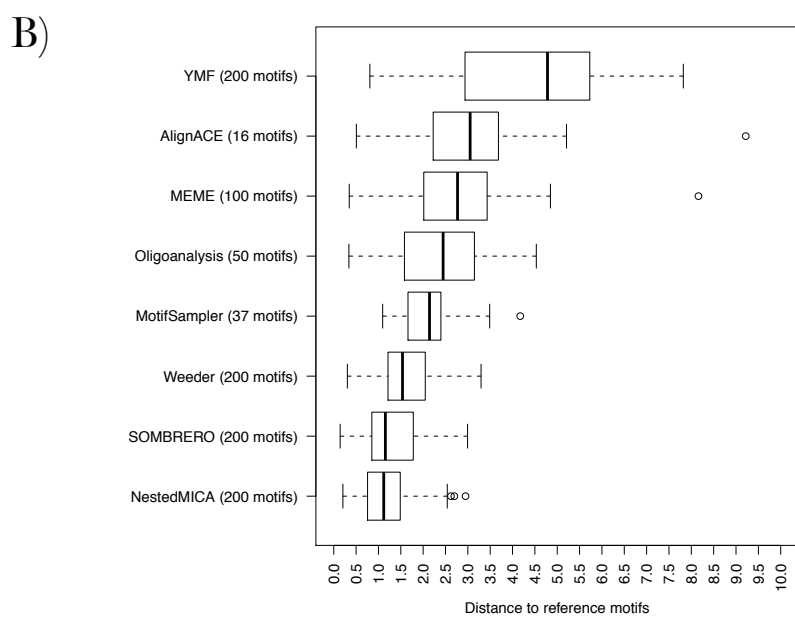
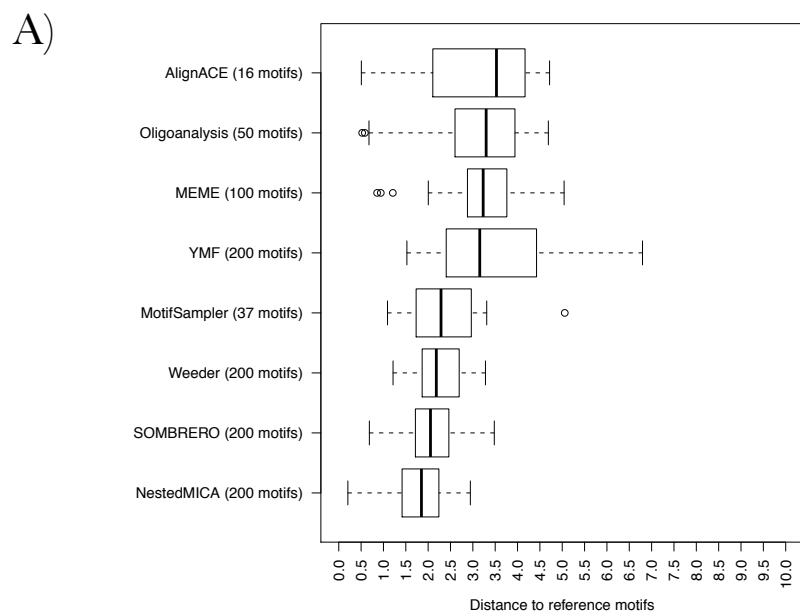


Figure 5.8: Distribution of SSD distances of predicted motifs to significant matches in the A) JASPAR and B) [Zhu et al. \(2009\)](#) PBM motif sets.

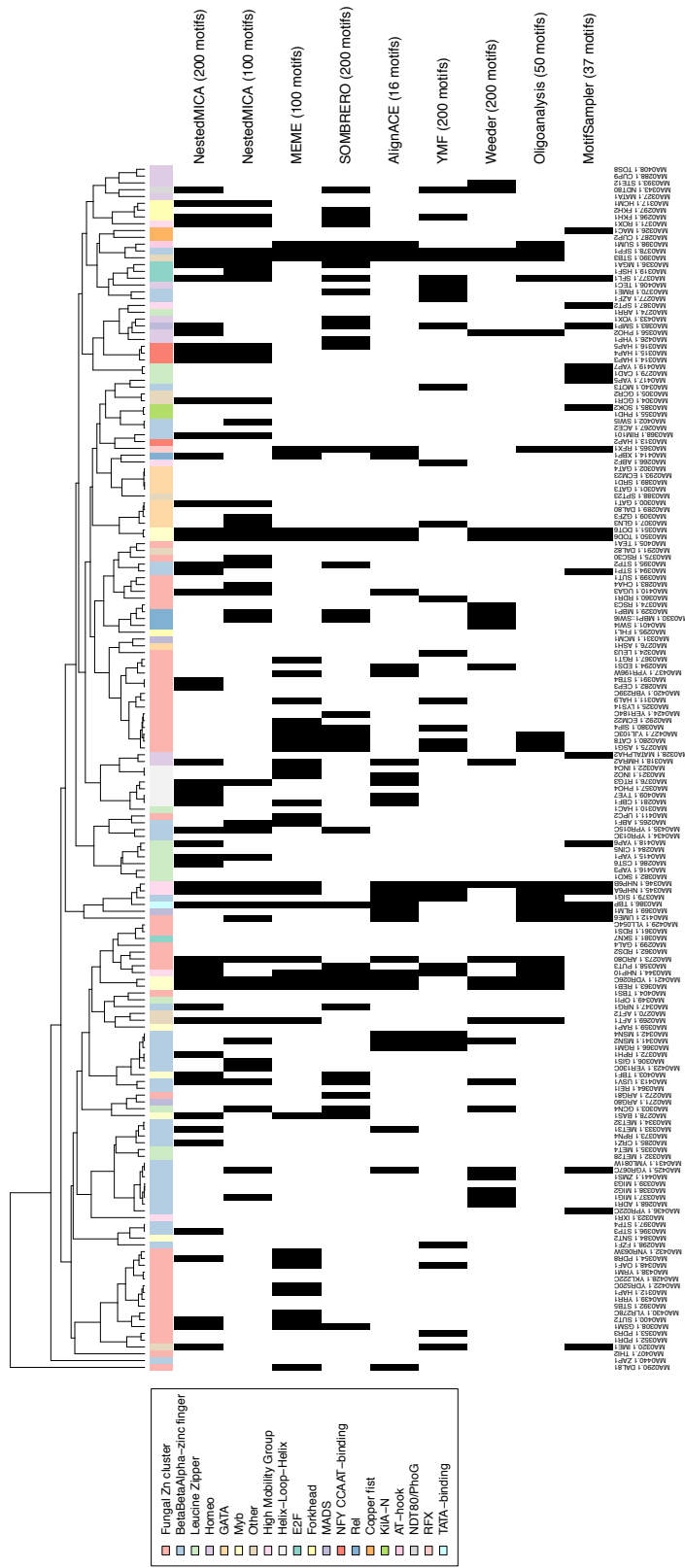


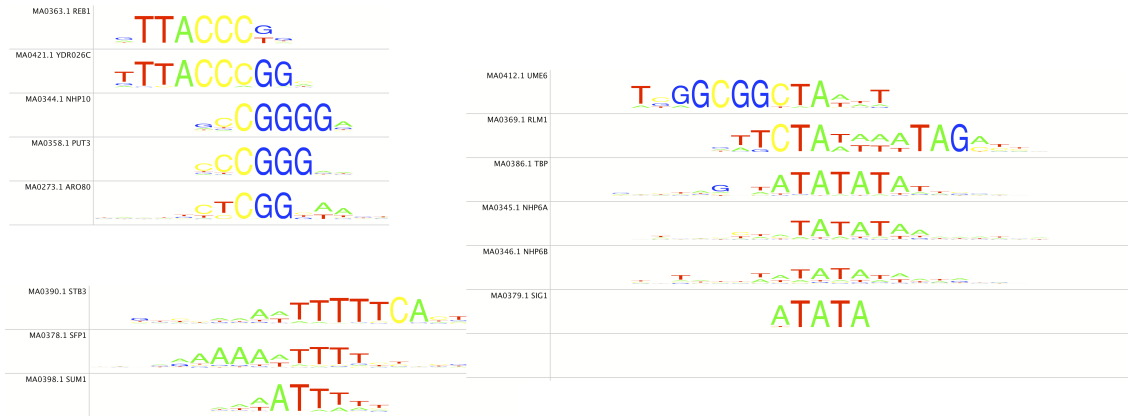
Figure 5.9: A heatmap showing the JASP motifs found or missed by each of the prediction methods ($p < 0.05$). Black cells in the matrix correspond to statistically significant matches ($p < 0.05$) between the JASP reference motifs (columns) and the computationally inferred motifs (rows). The reference motifs are ordered based on hierarchical clustering with the SSD distance. The inferred motif sets are ordered by their number of matches.

Some clustering of shared predictions by different computational methods is evident. Examples of JASPAR motifs predicted by different subsets of the methods are shown in Figure 5.10. Few clusters are covered by the majority of the algorithms, in fact only four such clusters appear. However, most JASPAR motifs in fact match by two or more methods, suggesting that consensus based predictions could perhaps be developed for more successful large scale motif inference, using combinations of different agreeing predictions. For example, SOMBRERO, and especially MEME, succeed with a large homogeneous cluster of 15 Zn cluster motifs (MEME identifies matches to 9, SOMBRERO to 5), to which NestedMICA predicts only two matches (CEP3, STB4). In contrast, NestedMICA shares motifs with SOMBRERO which match the FKH1 and FKH2 forkhead motifs, and the relatively closely related ROX1 motif, matches to which are not discovered by any of the other algorithms. All of the top performing methods also have motifs unique to them. Some examples of these motifs are also shown in Figure 5.10.

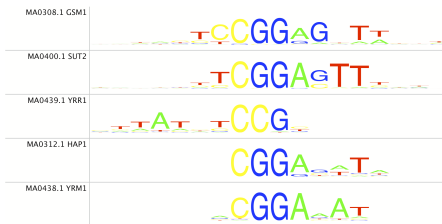
I studied the motif families predicted by the different methods, using the structural taxonomy provided by JASPAR. Some of these families, such as fungal Zn clusters, or $\beta\beta\alpha$ -zinc fingers are present in high numbers in the yeast genome. I separated the JASPAR motifs to groups based on their structural family, and counted the numbers of matches to each of these families (Figure 5.11). Stratification of the matches by motif family provides another natural way of ranking the motif inference methods.

Most methods (MEME especially) appear to find several of the fungal Zn cluster motifs (the single most abundant TF domain family in the yeast (Wilson et al., 2008a)). The $\beta\beta\alpha$ zinc finger, Myb and HMG motifs are also covered with predictions by most methods. Substantial differences between methods do however exist. MEME, for instances, appears to be unable to find any instances of E2F, forkhead, MADS, or NFY CCAAT-binding domains, whereas it discovers motifs similar to the only AT-hook and RFX-like motifs present in the JASPAR motif set. NestedMICA and SOMBRERO find the most varied collection of motifs: 16 different structural families, whereas AlignACE only finds 12, and MEME 11.

Motifs identified by almost all methods



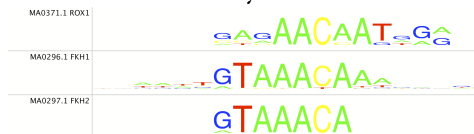
MEME only



SOMBRERO and MEME



SOMBRERO only



NestedMICA and SOMBRERO



NestedMICA only



Figure 5.10: Different algorithms find matches to partially overlapping subsets of the JASPAR motif set. Example motif clusters found by different subsets of the algorithms are presented.

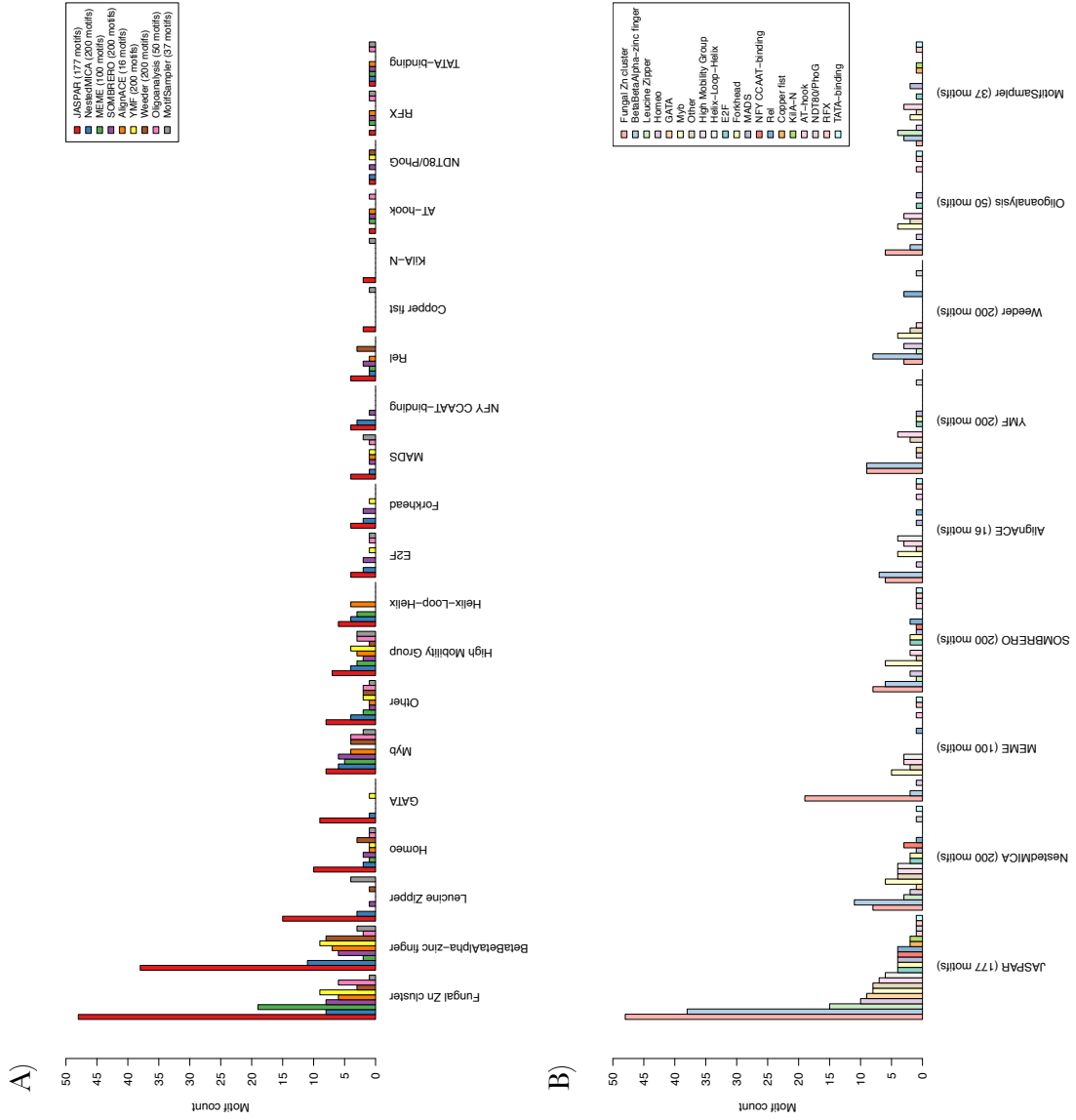


Figure 5.11: JASPAR motifs and computationally predicted motifs, grouped according to their A) domain family and B) the motif set.

Figure 5.12 summarises the differences seen between the motif inferred by the eight different methods, and their closest, statistically significant reference motif matches. The properties shown are the motif lengths, information contents, and per-column information contents, similarly as shown above in Table 5.4. Once again, the analysis conducted with the JASPAR reference motif set is largely consistent with the [Zhu et al. \(2009\)](#) PBM motif set.

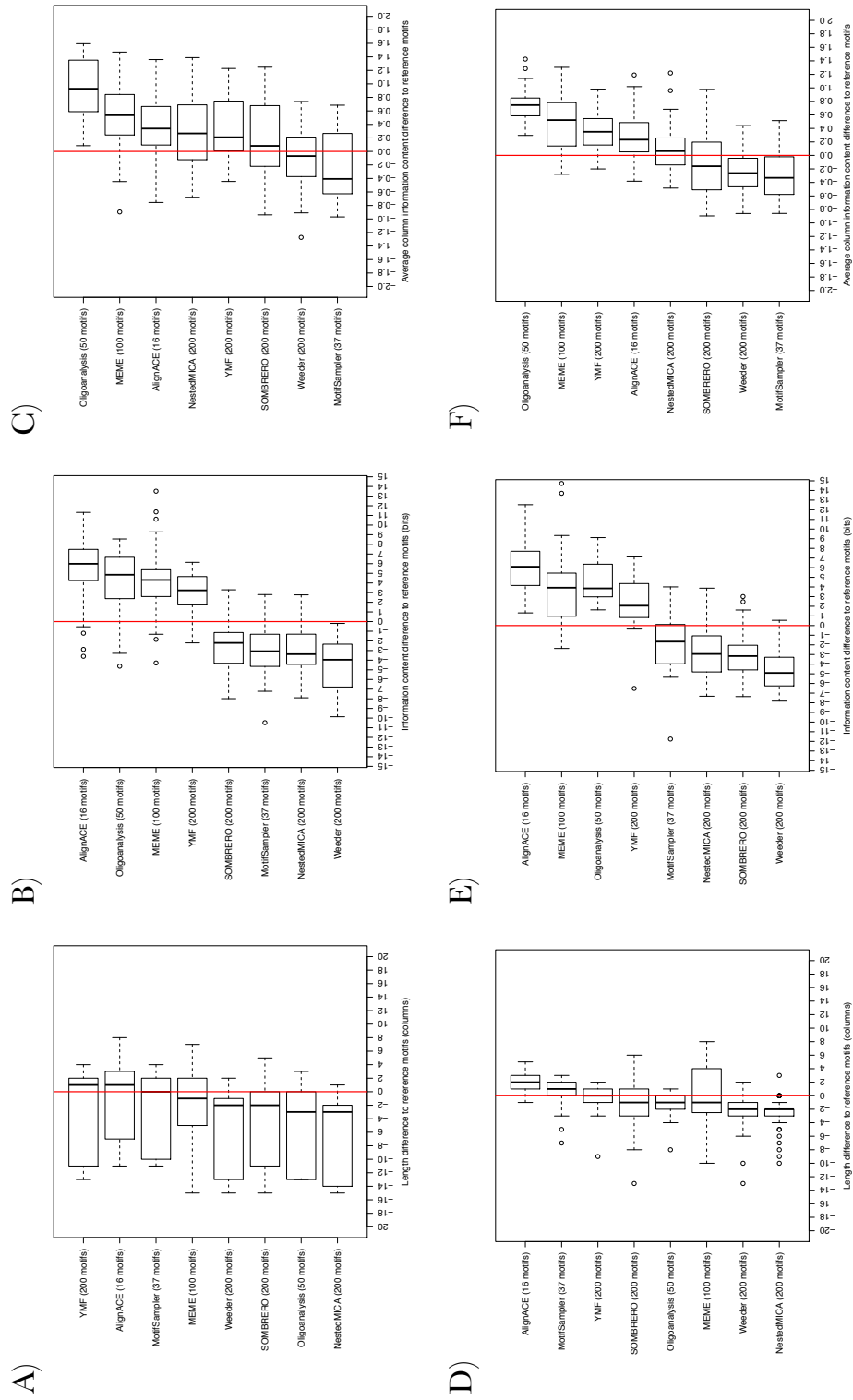


Figure 5.12: Differences in length, information content, and column-wise information content between the predicted and the JASPAR reference motifs. Panels A,B and C show comparisons of the predicted motifs with JASPAR motifs. Panels D,E and F are for comparisons of the predicted motifs with [Zhu et al. \(2009\)](#) PBM motifs. A & D: length difference; B & E: information content difference; C & F: average column-wise information content.

5.3.3 TF target gene associations of the discovered motifs

I tried to associate the genomic matches of inferred motifs with known target genes of TFs in the yeast genome (see Section 5.2.6 for details regarding the method). I did this with a parameter-free approach, assuming no significance threshold for the genomic matches of a motif. Each inferred motif was paired with its closest match in the non-redundant JASPAR database. With one exception (the MBP1:SWI6 complex), the 177 motifs in the JASPAR motif sets correspond to individual TFs, which in turn have associated target gene data available. The distribution of maximum bit scores are then compared with the non-targets to identify differences. Because there is no single authoritative source of TF–target gene pairings for the yeast genome, as discussed in Section 5.2.6, I therefore studied three alternative datasets. It is possible to rank methods based on the number of motifs identified by each, where a statistically significant difference is observed between the maximum bit score distribution of the target versus the non-target genes. Results with the three alternative datasets are shown in Figure 5.15.

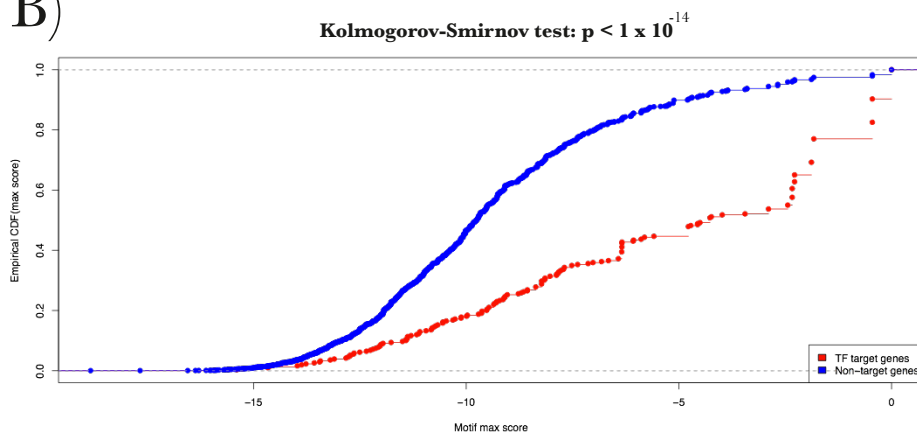
An illustrative example of the maximum bit score distribution difference between target and non-target genes of a TF is shown in Figure 5.13, where motif58 from the NestedMICA 200 motif set is studied with the targets of the REB1 TF (the REB1 motif is the closest match to motif58). There is a highly significant difference between the maximum score distributions.

High scoring TFBS motifs are not expected to cleanly partition promoter sequences of the yeast to disjoint target and non-target gene sets. For instance, motif158 from NestedMICA’s prediction set is found to be a close match to both the CBF1 and the PHO4 helix-loop-helix domain containing TFs (Figure 5.14). A statistically significant pattern is seen for the enrichment of motif158 with both CBF1 and PHO4. The DNA motifs of these two factors have been previously described as being closely similar, but they are known to act under different conditions and have partially different target gene sets; CBF1 acts under sulphur limitation, and PHO4 under phosphorus limitation [Clements et al. \(2007\)](#). High scoring motif matches of motif158 score highly for both of these only partially overlapping gene sets. One can therefore imagine that the motifs alone – espe-

A)



B)



C)

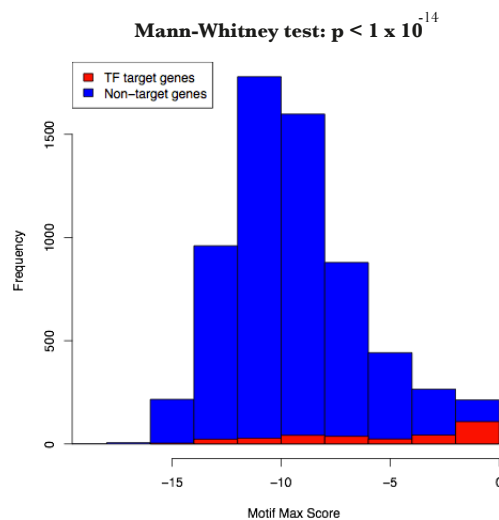


Figure 5.13: Some *de novo* inferred motifs are able to distinguish putative TF target genes from non-target genes by the maximum bit scores achieved by the gene promoter sequences (500bp upstream promoter sequences considered). A) Motif 83 predicted by NestedMICA is one such motif. B) The cumulative distribution of the maximum bit scores of non-targets (blue) and targets (red) as judged by the YEASTRACT database. C) A histogram of the bit score distributions of non-target promoter sequences (blue) and target sequences (red).

cially in the case of highly expanded TF families – do not have the discriminatory power to determine the target gene relationships of a TF (see Section 1.1 for a discussion on the various additional gene regulation mechanisms additional to TF binding).



Figure 5.14: Motif158 is closely similar to both the CBF1 and PHO4 motifs.

Different stringency of calling genes either TF targets or non-targets can affect this analysis: if large number of TF targets are found in the non-target set, or vice versa, the separation between the target and non-target scores diminishes. This can be also caused by limitations in our knowledge of targets of some less studied TFs when compared to others, when dealing with hand-curated datasets. I considered three different TF target gene datasets in this study: a manually curated YEASTRACT dataset (Teixeira et al., 2006), the gene expression study based target set by Reimand et al. (2010), and the ChIP-chip data by Harbison et al. (2004). As a fourth set, I also attempted to retrieve the TF target predictions by Beyer et al. (2006), which are a result of integrating diverse lines of evidence into a probabilistic TF target prediction. Unfortunately however the dataset originally made available by the authors at <http://www.fli-leibniz.de/tsb/tfb> was not found anymore (authors were contacted). Several datasets were considered here because the coverage and confidence of TF–target associations included in each of them is not necessarily uniform across the TFs that each covers. The environmental states (e.g. growth conditions) covered by the datasets for instance are a factor: some factors bind their targets in an environment specific manner. According to Harbison et al. (2004), TFs fall into four groups with regards their target gene sets:

-
- Condition-invariant housekeeper TFs that bind target genes regardless of conditions. For instance Leu3, which regulates amino-acid biosynthesis (Kirkpatrick and Schimmel, 1995))
 - Condition-enabled, for instance MSN2 which only enters nucleus to regulate target genes when the cell is under stress (Beck and Hall, 1999; Chi et al., 2001).
 - Condition-expanded, which bind an expanded set of target genes under specific conditions. These include for instance Gen4, which binds an expanded set of target genes under limited nutrients (Albrecht et al., 1998).
 - Condition-altered, for instance Ste12 whose targets vary depending on condition-specific interaction partners (Zeitlinger et al., 2003).

Given the above categorisation of TFs by their ranges of target genes, one can imagine that there is variation between TFs in the power to detect a difference between promoters of target genes and non-target genes with high-scoring TFBS motif matches.

The largest number of TFs with a significant difference between the maximum bit score distributions of the target and non-target genes is seen consistently for all the algorithms with TF calls from the YEASTRACT dataset. This could be attributable for the manually curated YEASTRACT dataset being the most extensive and accurate resource of TF target calls, as it considers evidence from several sources. The ranking of motif inference algorithms relative to each other varies considerably depending on the source of TF target calls, with NestedMICA performing the best with the YEASTRACT and ChIP-chip based TF target calls, both in the case of the Kolmogov-Smirnov and the Mann-Whitney tests. AlignACE, with its mere 16 predicted motifs, also performs also remarkably well with this metric, outperforming all of MEME, SOMBRERO and the 100 motif NestedMICA prediction with the YEASTRACT dataset (Figure 5.15A). With the Reimand et al. (2010) expression based target calls, AlignACE outperforms NestedMICA with eight TFs ($p < 0.05$), with NestedMICA identifying only six differences at the same significance level. AlignACE and NestedMICA share the top rank with the Mann-Whitney test at this same significance level. Interestingly

though the reference JASPAR motifs identify a significant difference for only two more TFs than AlignACE, with this same dataset and statistical test. One feasible interpretation for this general failure of a motif match based approach to identify differences between the two populations of promoters with the [Reimand et al. \(2010\)](#) TF target calls is that the target list contains indirect downstream targets of the actual TF (possible because the dataset is expression effect based).

As an alternative to studying the closest JASPAR matches, all motifs could have been tested ‘blindly’ against all TF target sets. This however would necessitate a considerably larger number of statistical tests and make correcting for multiple testing more difficult. Furthermore, combinatorial regulation by TFs could potentially lead to statistical associations being called between TFs and motifs that are unrelated in binding specificity, but which tend to co-occur in promoters with the real motif.

5.3.4 Clustering of motifs and their binding sites

Some closely related patterns are expected amongst *de novo* predicted TFBS motifs, due to the shared evolutionary history of TFs. However, when challenged to infer a collection of motifs from a large series of genomic sequence, a motif inference algorithm should ideally find a wide spectrum of motifs, instead of predicting large numbers of redundant copies of a small number of patterns. I therefore measured the relatedness of motifs, not only to the JASPAR reference motifs, but also to other predicted motifs. I did this in two different ways: firstly by computing distance matrices between motifs with the SSD motif distance metric ([Down et al., 2007](#)), and secondly with an genomic match overlap score (Section 5.3.5). To begin with, I studied motif relatedness in a visual, qualitative way by drawing dendrograms of all of the motif sets together with JASPAR motifs (Figure 5.16), and with each of the sets separately with JASPAR motifs (Figure 5.17).

The dendrogram of all predicted motif sets with JASPAR shows – similarly as the analysis presented in Section 5.3.2 – that overall there are few large clusters of experimentally validated motifs with no related predicted motifs from one of the inferred motif sets. Redundant clusters by some of the motif predictions

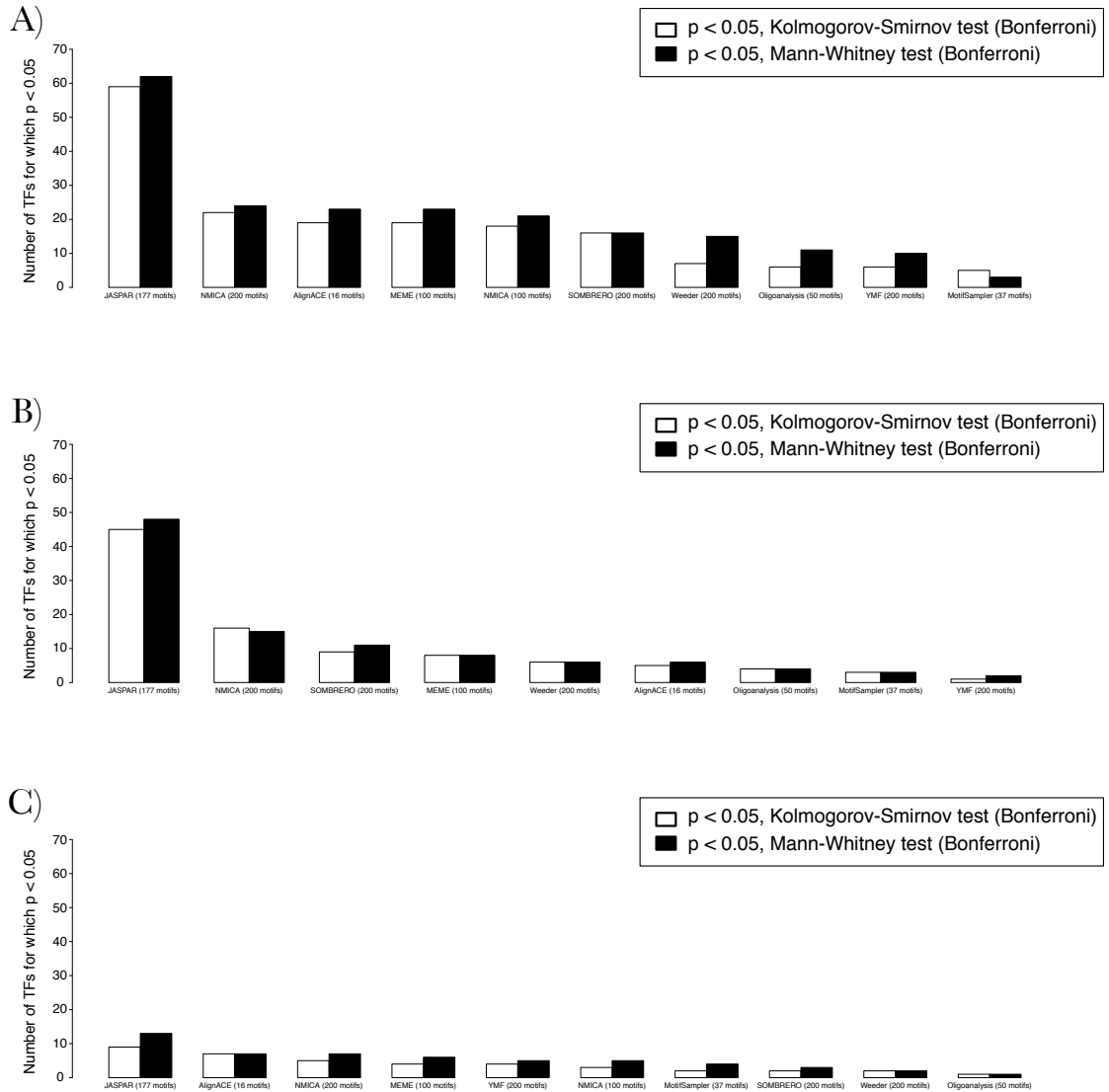


Figure 5.15: TF–target associations of the inferred motifs, when compared to JASPAR motifs (leftmost). The bars represent the number of TFs for which the computationally inferred motif shows a significantly different distribution of maximum bit scores, when target and non-target genes are compared. Motif sets are ordered by decreasing number of TFs with a significant effect. The p-values are Bonferroni corrected (divided by 176, which is the number of TFs tested).

are also apparent, especially in the case of YMF and Weeder. Conversely, the clustering pattern of NestedMICA and SOMBRERO motifs shows the predicted motifs much more ‘intertwined’ with the reference JASPAR motifs. Individual dendrograms are drawn in Figure 5.17 for each of the motif sets to make this pattern clearer to see.

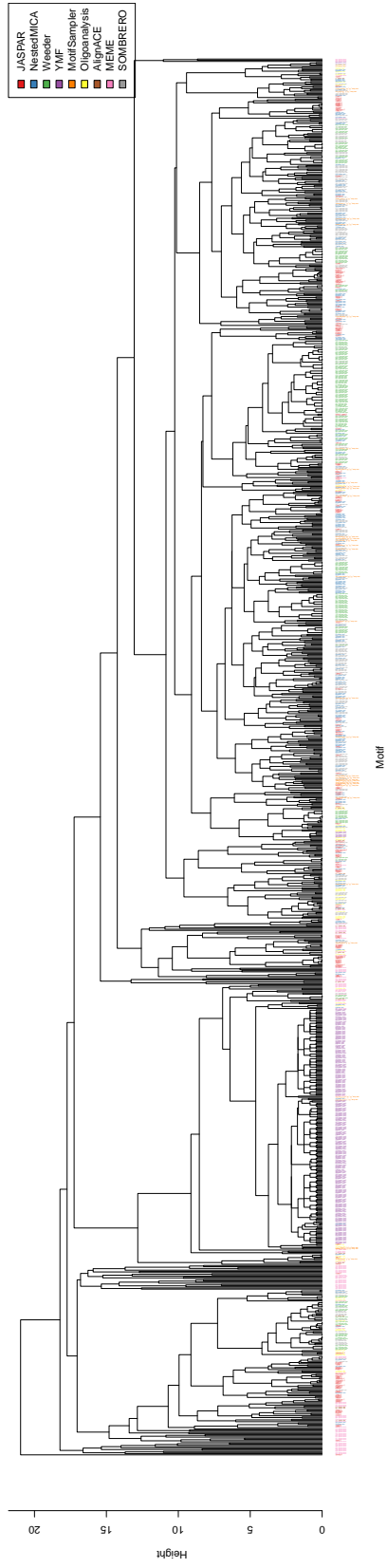


Figure 5.16: Dendrogram of a complete linkage clustering of all predicted motif sets with the JASPAR motifs, with the SSD metric from [Down et al. \(2007\)](#).

The motif clustering tree can be cut at different heights. I counted the numbers of cases where a JASPAR motif is clustered together with any of the other methods at varying heights. By this measure, SOMBRERO and especially NestedMICA perform favourably to the other algorithms (Figure 5.18).

Whereas Figure 5.18 measures inferred motif similarity to JASPAR motifs, the closest pairings of motifs within the predicted sets can also be studied using the distance matrix of the predicted motifs with each others (Figure 5.19). As one would already predict based on the motif dendrograms in Figures 5.16 and 5.17, YMF and Weeder predict considerably larger numbers of overlapping patterns than the other methods. At the 2.0 SSD distance cutoff for example, the average clique size of the motif distance matrix for YMF is above 40, compared to roughly 5 for Weeder, and between 2 and 1 for all of the other methods. Weeder and YMF appear essentially incapable of large scale motif inference as conducted in the present study, either due to my parameter choices for running the tool, or due to intrinsic problems with the algorithms.

The empirical significance values presented in Section 5.3.2 can be estimated for the closest pairs of motifs within each predicted sets, with the same protocol as used for comparing predicted motifs to reference motifs in Section 5.3.2. The ‘uniqueness’ of motifs varies considerably: almost all of Weeder motifs contain a statistically significant match, whereas MotifSampler and MEME have hardly any statistically significant matches regardless of the significance chosen. The JASPAR motif set also contains many motifs with close pairs; depending on the significance scores used, roughly 45% to 75% of JASPAR motifs have at least one match (Figure 5.20). This fraction is in fact higher for JASPAR than any of the other analysed methods but Weeder (the consensus based YMF and Oligoanalysis methods were omitted from the significance score analysis).

5.3.5 Comparing motifs by the overlap of their genomic matches

I measured the fraction of overlapping binding sites shared by motifs as a measure of motif similarity, complementary to the SSD distance matrices and motif clustering shown above in Section 5.3.4. I studied binding site overlap patterns

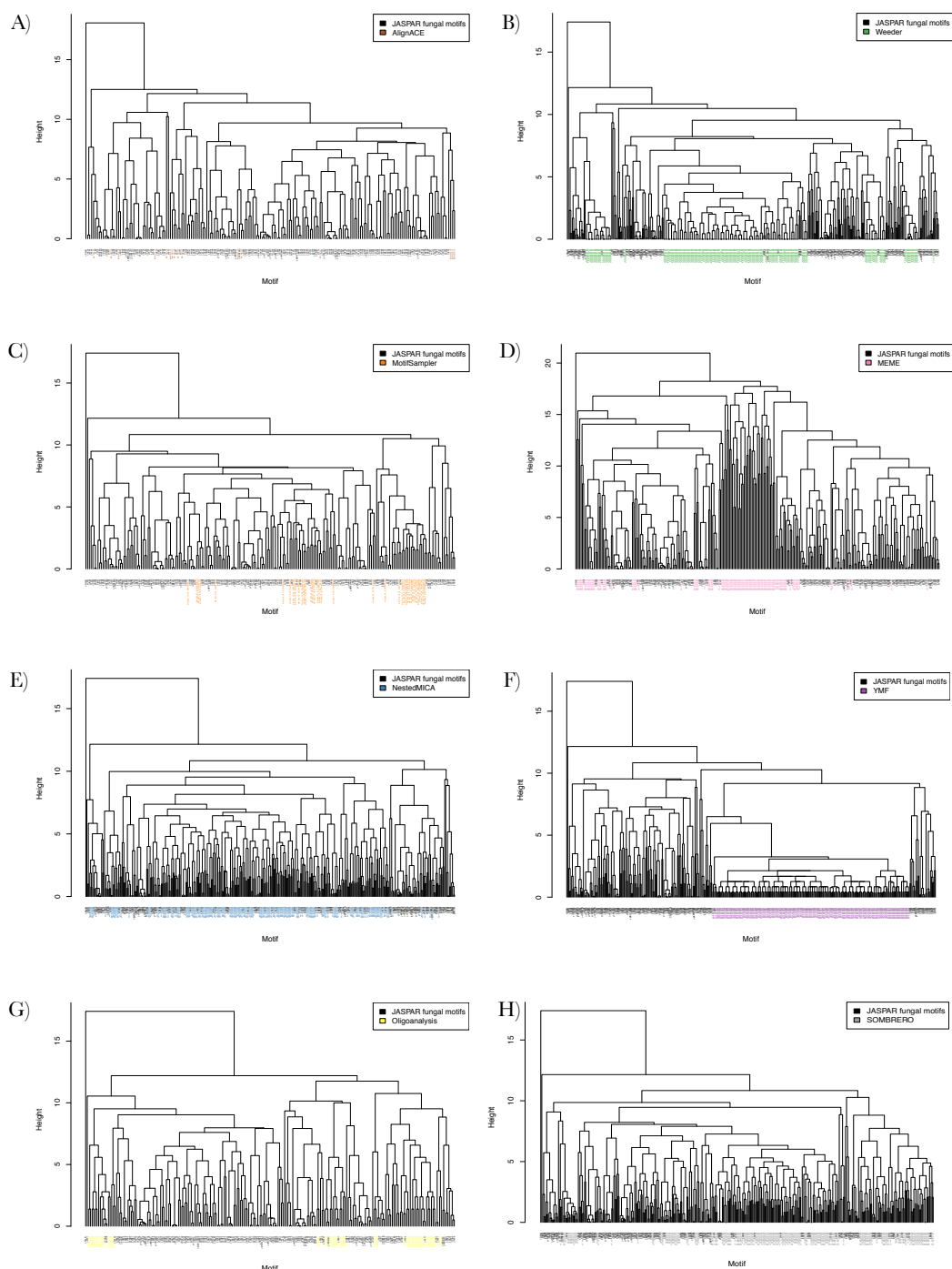


Figure 5.17: Clustering of JASPAR motifs with results of A) AlignACE, B) Weeder, C) MotifSampler, D) MEME, E) NestedMICA, F) YMF, G) Oligoanalysis H) SOMBRERO. The motif names are coloured according to the motif set where they originate from. They are shown as a quick visual summary of the clustering of the inferred motifs, rather than trying to present readable names.

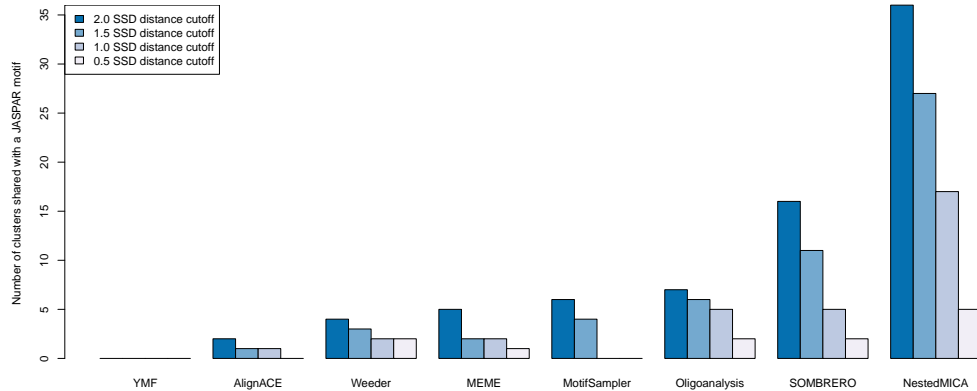


Figure 5.18: Numbers of clusters that contain at least one or more inferred, and one or more JASPAR motifs. Four different distance cutoffs are shown.

firstly visually, using dendrograms and heatmaps. The binding site overlap of two different inferred motif sets with the JASPAR reference motif set are shown in Figure 5.21.

Visual inspection of the heatmaps in Figure 5.21 suggests a higher overlap between NestedMICA and the JASPAR motifs, than between SOMBRERO and the JASPAR motifs. I quantified the binding site overlap by counting the numbers of motifs output by each of the eight methods, which overlap a JASPAR motif above a binding site score overlap. I repeated this analysis with five different overlap score cutoffs (Figure 5.22). The results are largely consistent with the clustering based motif similarity measures, suggesting NestedMICA is the method with the highest fraction of overlapping binding sites by this measure, followed by SOMBRERO, Weeder and MEME. Note that this similarity measure between the inferred and the reference motifs does not account for motif redundancy. This is the reason that Weeder for instance receive relatively high overlap scores with JASPAR motifs, when in fact its motifs map to a relatively small number of known TFBS motifs in the JASPAR set. The motif match significance score cutoff parameter, of both the reference and the inferred motifs, can also affect the results of this analysis.

Overlap of genomic matches between motifs can also be used as another means

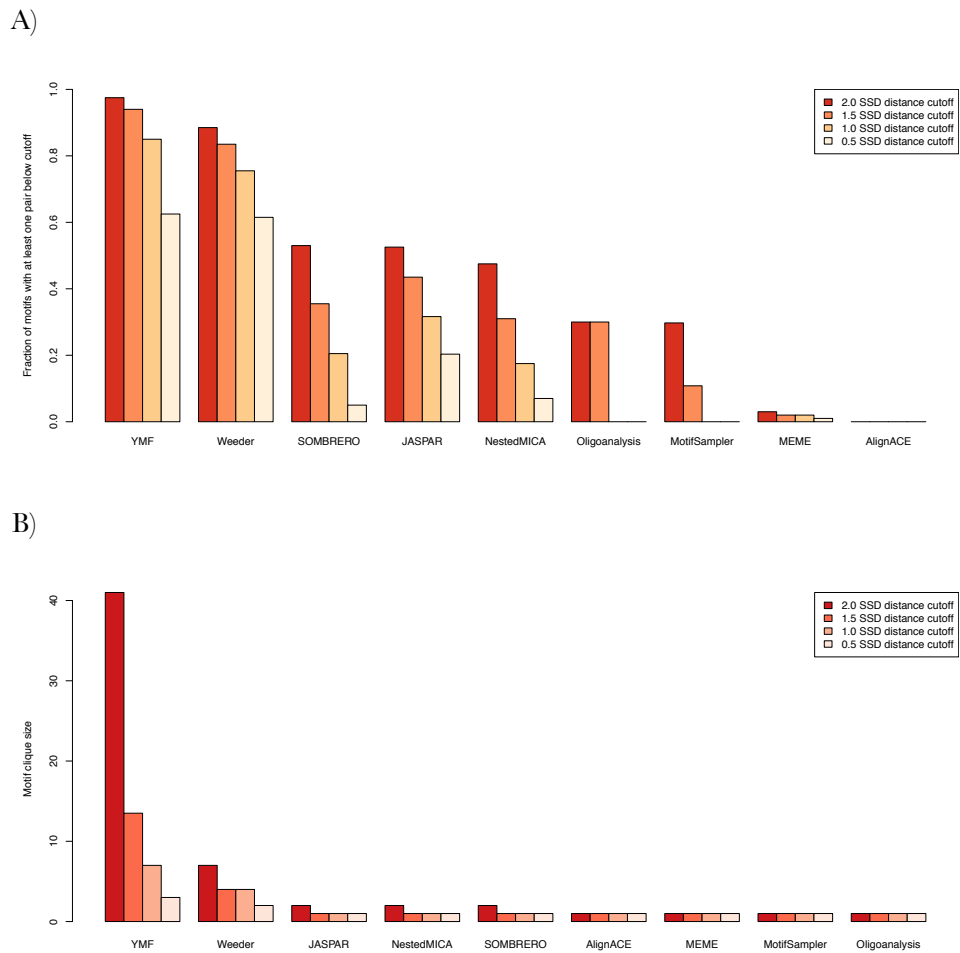


Figure 5.19: Motif redundancy as judged by the motif-to-motif SSD distance. A) Fraction of motifs which have at least one pair B) Average motif clique size.

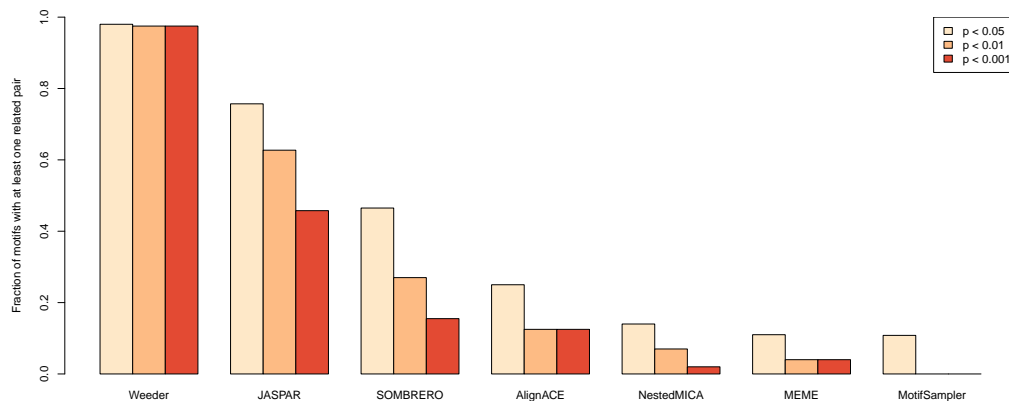


Figure 5.20: The fraction of motifs with at least one matching pair, at three different significance cutoffs. The consensus string based YMF and Oligoanalysis are omitted from this analysis, because the empirical significance score used here does not behave reliably for PWMs derived from IUPAC consensus strings.

of measuring motif similarity within sets. To illustrate this, Figure 5.23 shows the genomic match overlap of the SOMBRERO, Weeder and JASPAR motif sets. As discussed in Section 5.1.3, binding site level comparisons are not necessarily robust to the significance cutoffs used for genomic motif matches, and I do not advocate the use of these measures for ranking inference methods.

By this measure, Weeder receives the highest ‘redundancy scores’: for instance at the 10% overlap score cutoff, nearly all of the 200 weeder motif predictions have at least one motif pair which overlaps (Figure 5.24). The average number of motifs which all share a given fraction of their binding site matches (the motif clique size) however varies dramatically depending on the chosen binding site cutoff.

The present analysis of genomic match overlap between motifs is indeed a cautionary tale of assessing motifs based on their binding site overlaps: performance measures derived from genomic matches are not robust to the bit score significance cutoff chosen for a motif. This is an especially pressing concern for motif inference assessments such as Tompa et al. (2005), where experts applied many of these same algorithms, each with independently chosen motif match significance

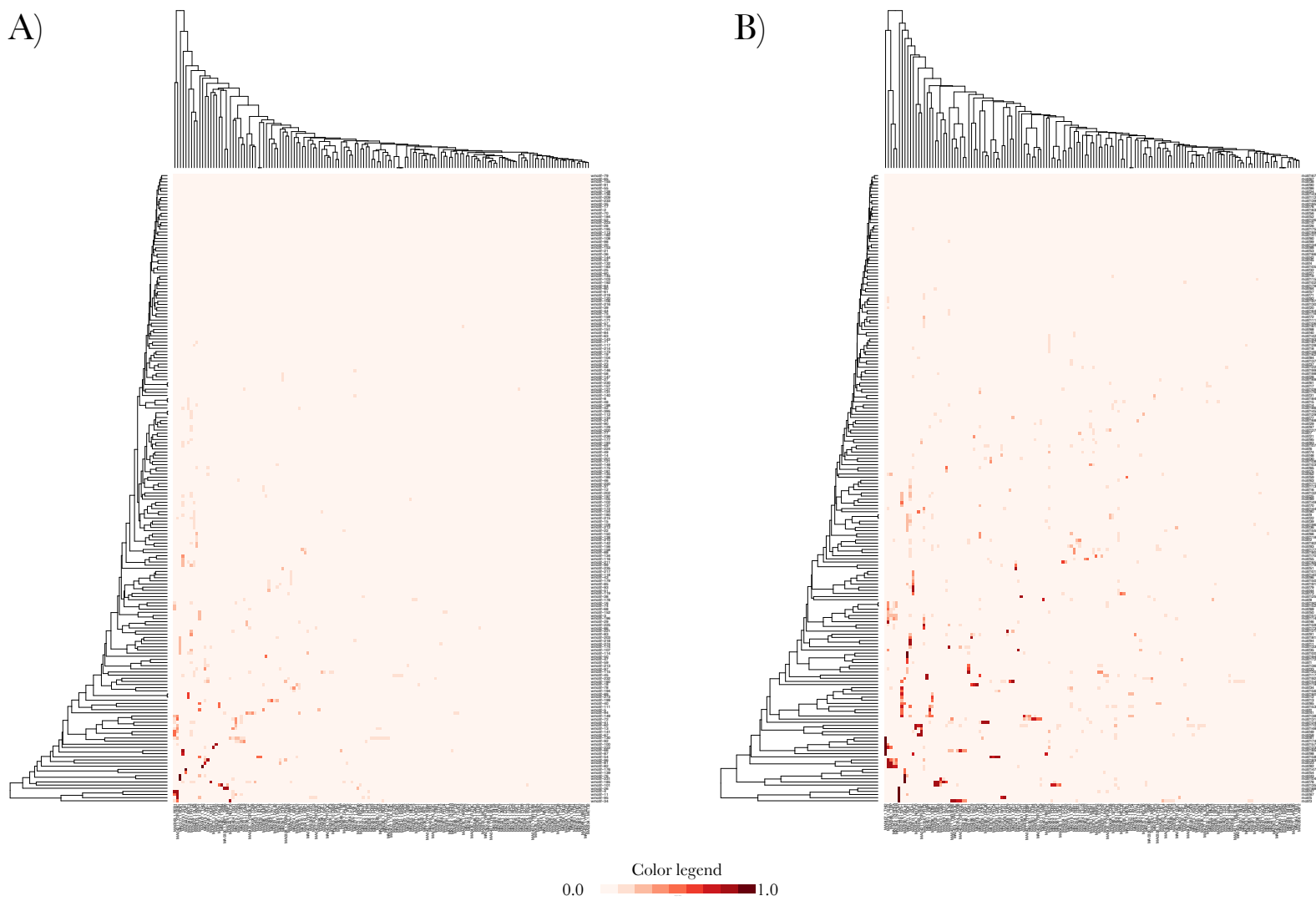


Figure 5.21: Motif binding site overlap of A) SOMBRERO and B) NestedMICA motifs. The rows represent inferred motifs, and the columns are JASPAR motifs. They are ordered based on an euclidian distance between the overlap patterns, with complete linkage clustering ([Johnson, 1967](#)).

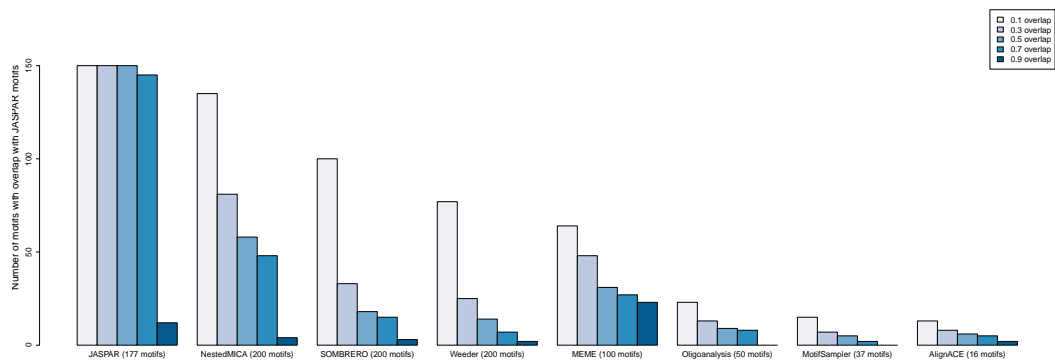


Figure 5.22: Predicted motif similarity to JASPAR motif set on the level of binding site overlap. The bars represent the numbers of motifs which show overlap above 0.10, 0.30, 0.70, 0.90 to JASPAR motifs with the metric described in Section 5.2.5.

parameters.

5.3.6 Looking for evidence of function for the inferred motifs

On top of the 177 TFBS motifs included in JASPAR, the yeast genome contains others. The transcription factor database DBD (Wilson et al., 2008a) for instance contains 177 likely regulatory TFs for the genome, but its DNA binding domain model based predictions are estimated to cover only 2/3 of the genome Wilson et al. (2008a). The Harbison et al. (2004) ChIP-chip study on the other hand includes the binding profile of 203 putative regulatory TFs. It is therefore possible, even likely, that the promoters used in the study contain motifs for TFs which are not included in the 177 motifs of the JASPAR database. Therefore, I do not believe that all the apparent false positives (which do not match reference motifs) are false positives, and I wanted to identify a subset of particularly likely functional motifs from these unknown motifs.

I studied three different aspects of the computationally predicted motifs as signs of potential function: interspecies conservation (Section 5.3.6.1), SNP rate in yeast strains (Section 5.3.6.2), and positional bias of the motifs with respect

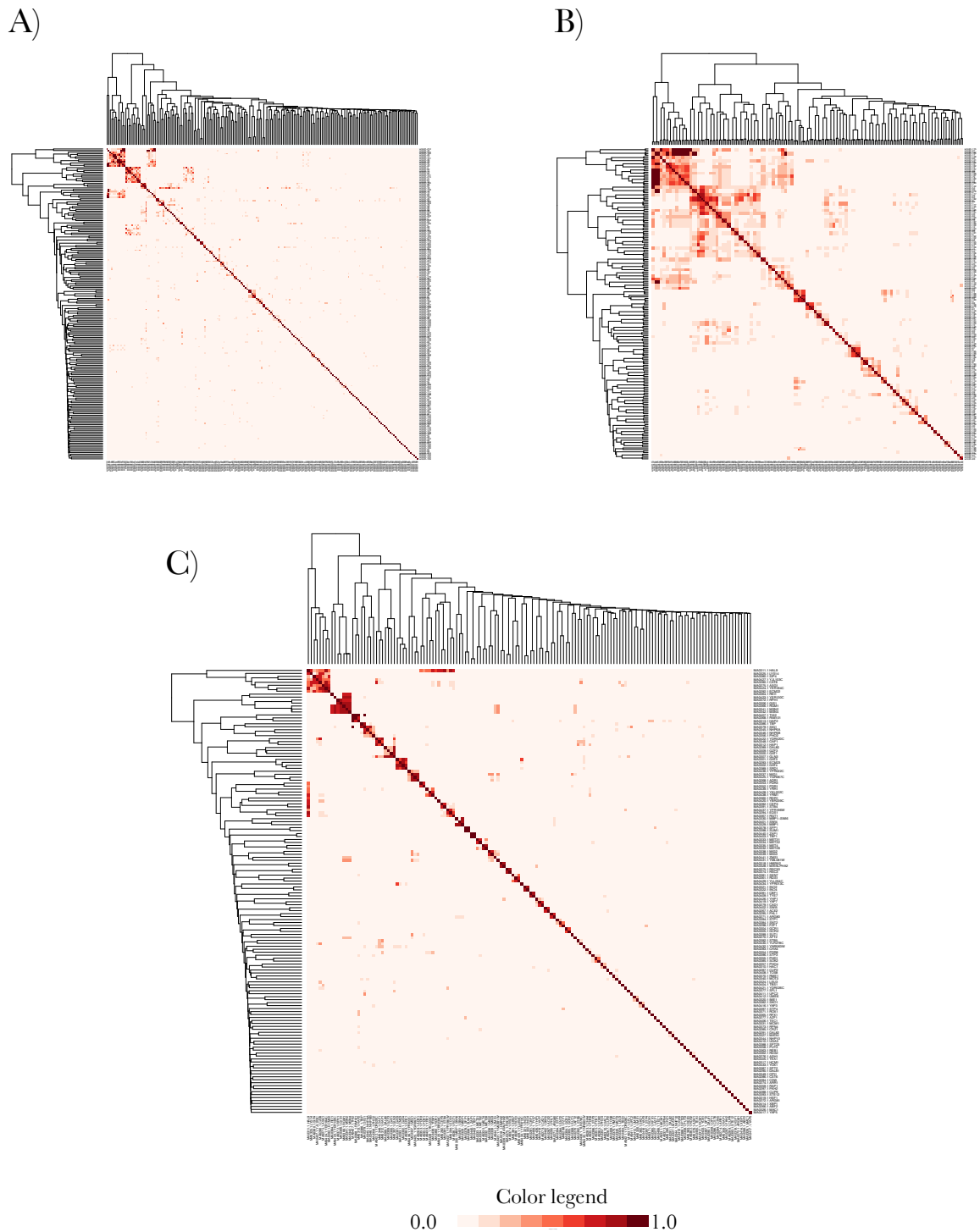


Figure 5.23: The overlap of genomic matches within motif sets. A) SOMBRERO and B) Weeder motifs are shown as examples of the predicted motif sets, and binding site overlap of JASPAR motifs are in panel C. SOMBRERO and Weeder differ in the degree of redundancy amongst the motif set. 500bp upstream sequences were analysed.

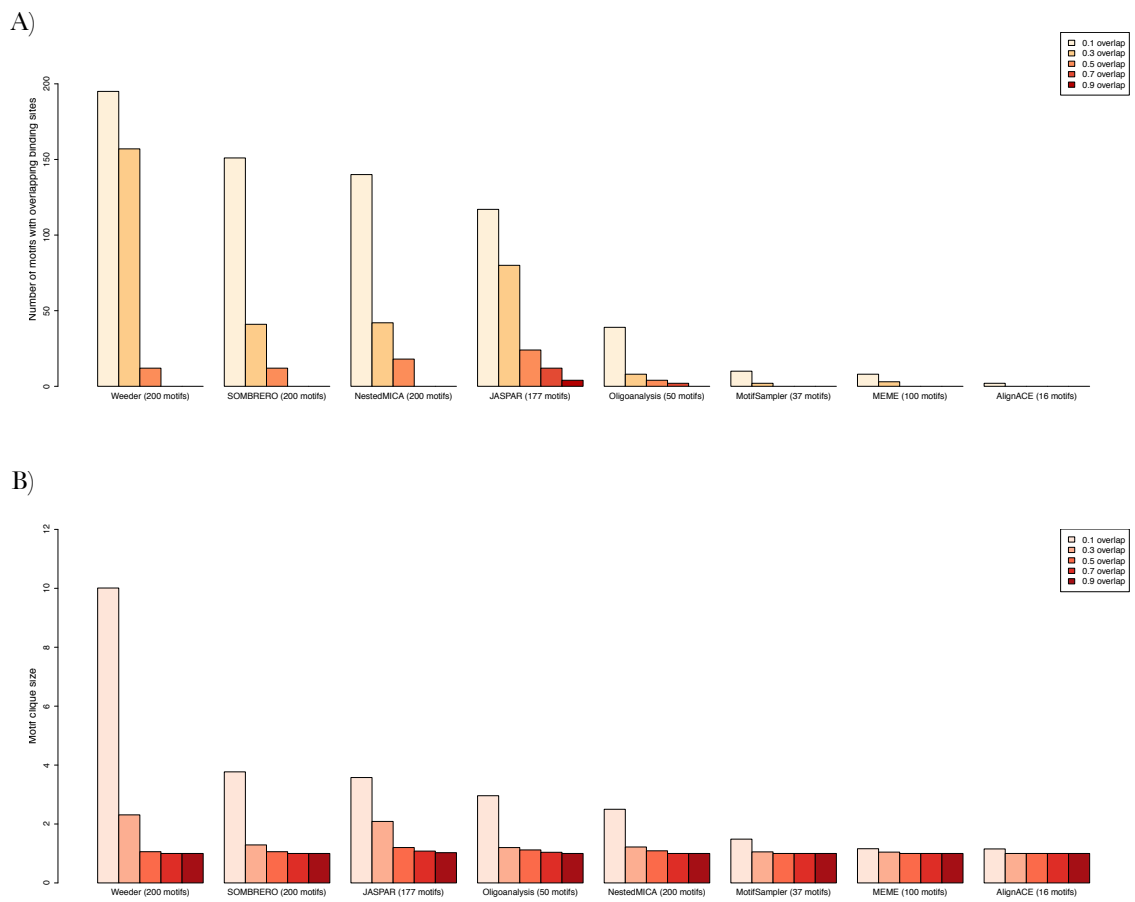


Figure 5.24: Predicted motif redundancy on the level of binding site overlap. The bars represent the numbers of motifs which show binding site overlap with the metric described in Section 5.2.5.

to the closest transcription start sites (Section 5.3.6.3). The motifs which fit all the criteria of high conservation, low SNP rate, and positional bias were then analysed in Section 5.3.6.4. Furthermore, I attempted to use the `metamatti` motif classification framework presented in Chapter 4 to predict the domain family of the motifs as a further sign of function (Section 5.3.6.5).

5.3.6.1 Inter-species conservation of the inferred motifs

The conservation scores for all of the 200 NestedMICA motifs at a 0.05 significance level are shown in Figure 5.25, as an example. A similar analysis was conducted also for all of the other methods (summarised in Figure 5.26). Figure 5.26 shows the fraction of motifs predicted by each method with a significantly higher conservation rate than random intergenic sequences of the same length. Note that for some of the methods, the fraction which matches known TFBS motifs in the JASPAR database (Section 5.3.2) is much smaller than the fraction which shows excess conservation. This could be explained by some of the predicted motifs being weak, undetected matches to real TFBS motifs, or artifacts of the multiple alignment based conservation PhastCons scores. Alternatively it could be that there are other potentially functional motifs within the motif predictions.

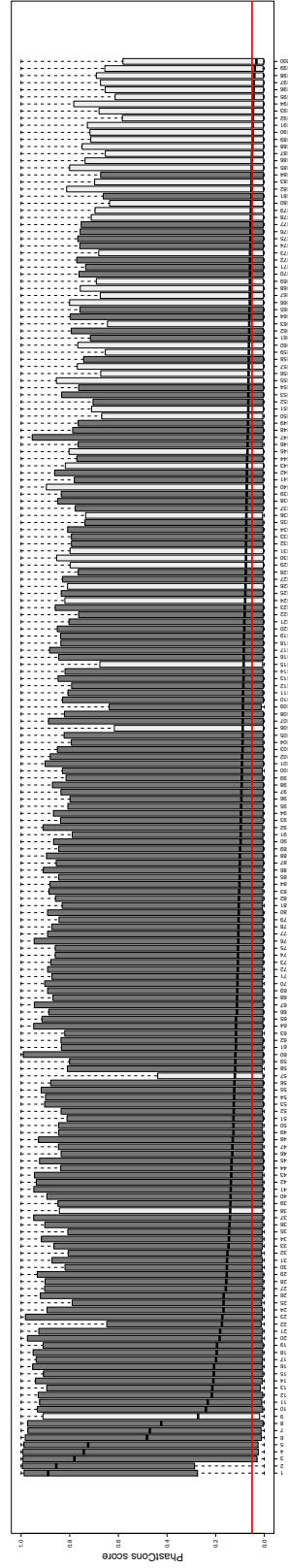


Figure 5.25: Conservation of motifs predicted by NestedMICA. The red horizontal line represents the mean PhastCons score for random intergenic sequence fragments of length 14 (the maximum length of motifs predicted by NestedMICA). The black markers are the median scores. The darker bars are motifs which are significantly more conserved than the intergenic sequence of the same length. The lighter bars are motifs for which this is not the case. Conservation of each motif was tested with a single-tailed two-sample Kolmogorov-Smirnov test between matches of each of the predicted motif sets, and random intergenic sequence positions of the matching length.

NestedMICA and Weeder show a roughly comparable fraction of significantly conserved motifs, between 60% and 80%, depending on the significance threshold which is varied between $p < 0.01$ and $p < 0.0001$. Overall the fraction of conserved motifs fits between 40% to 80% for all but two methods, which are overliers in the opposite ends of the scale; all of the YMF motifs show excess conservation, whereas only 8 of motifs inferred by MEME are significantly conserved. The results seen for YMF are in part explained by its highly redundant motif set, which shows variants of essentially one evidently highly conserved motif. The remarkably low figure of 8 motifs in the case of MEME is most likely due to its long motifs with high information content. This in combination with the stringent bit score cutoff determination method I used (Section 5.2.4) causes only a small number of hits to be reported and compared with the intergenic sequence regions, decreasing the sensitivity to detect differences between the distributions. An inspection of the median motif hit counts indeed shows alarmingly low figures for MEME's motifs at the 0.01 confidence threshold used: median motif hit count with the 200 base long upstream sequences is 2. This means that the significance score determination method used in the present study has largely failed with the motifs output by MEME. This, yet again, is an indication of problems associated with genomic hit based assessment of computationally inferred motifs.

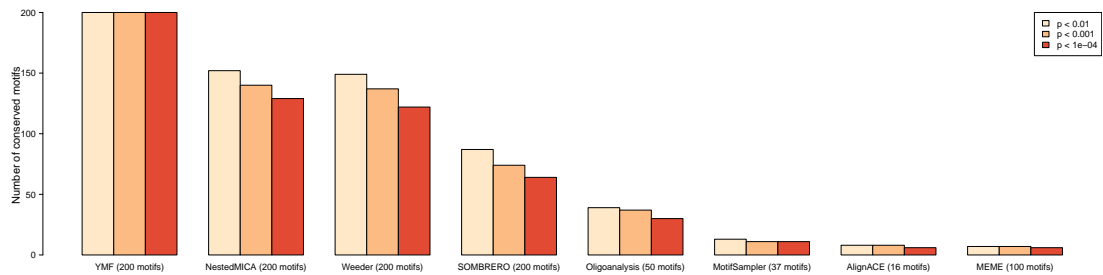


Figure 5.26: The number of motifs from each of the predicted motif sets that are found more conserved than intergenic sequence of the same length. Three different significance thresholds are shown. See Section 5.2.6.4 for details regarding the statistical testing.

5.3.6.2 SNP rates of the inferred motifs

A summary of the SNP rate analysis is shown in Figure 5.27. YMF and MEME are at the opposites of this scale, similarly as in the case of conservation patterns in Section 5.3.6.1. When compared with the inter-species conservation patterns, smaller fraction of motifs inferred by any of the methods show a significant difference to intergenic sequence. NestedMICA, SOMBRERO and Weeder identify the largest numbers of motifs with a significant difference. Similarly as in the case of interspecies conservation, the redundancy of motifs is not taken into account in the numbers reported, and they are not to be interpreted as a measure of the relative performance of the tools.

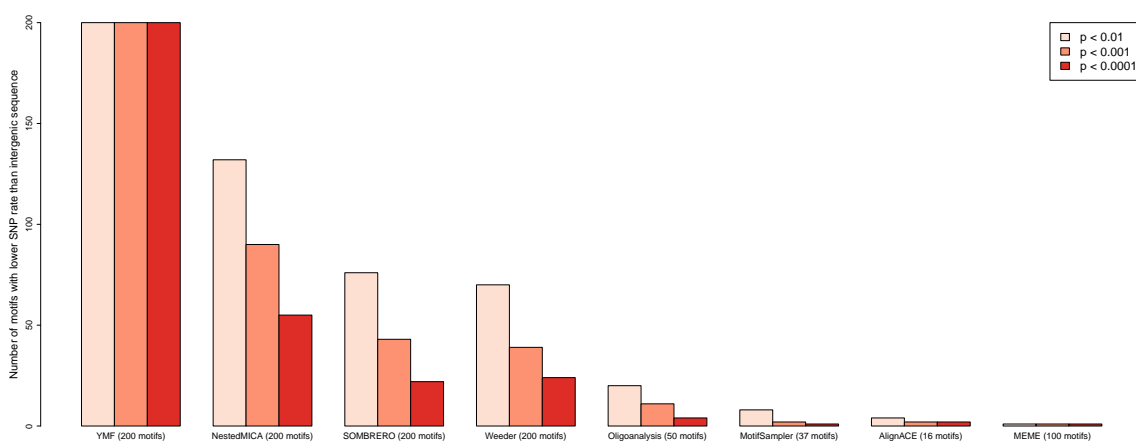


Figure 5.27: The number of motifs predicted by each of the methods with lower SNP rates than randomly selected intergenic sequence of the matching length. See Section for a description of the bootstrapping based significance scores.

5.3.6.3 Positional bias of motif matches close to the TSS

Many of the computationally inferred motifs were found to match preferentially upstream of the TSS. As examples of the typical positional bias trends which were seen, Figure 5.28 show the positional bias patterns in the case of SOMBRERO and Weeder. A summary of the positional bias trends of all of the methods are shown in 5.29, as the fraction of motifs with a statistically significant preference for positions -500 to 0. It is perhaps not surprising that a positional bias is seen

for many of the motifs, given that the motif search was made in the space of promoter sequences that span -200 to 0 from TSSs.

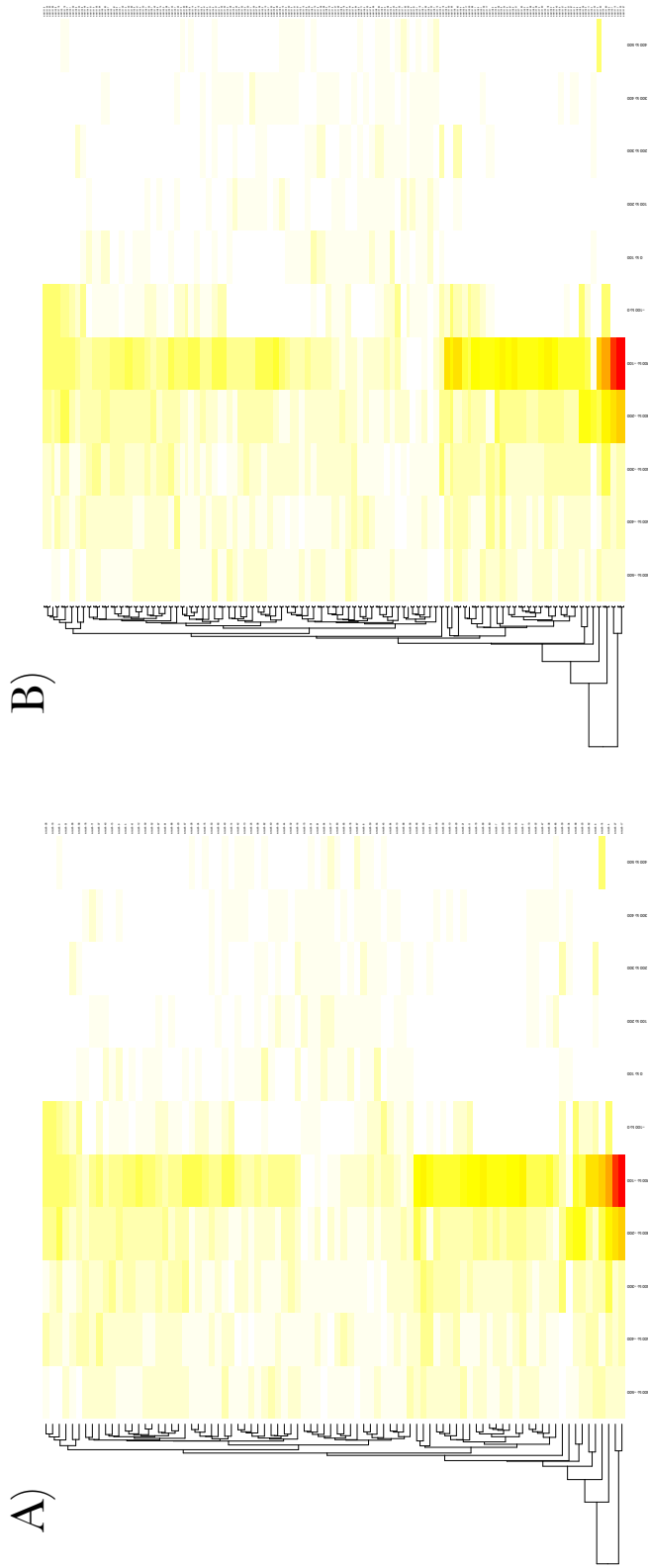


Figure 5.28: A heat map depiction of the positional bias trends of the motifs inferred with the A) SOMBRERO and B) Weeder algorithms. The columns in the heat map are 100 nucleotides long bins from -1000 to 1000, with respect to the TSS. Rows are individual motifs. Rows are ordered by a complete linkage clustering with an Euclidian distance of the relative frequencies at each bin.

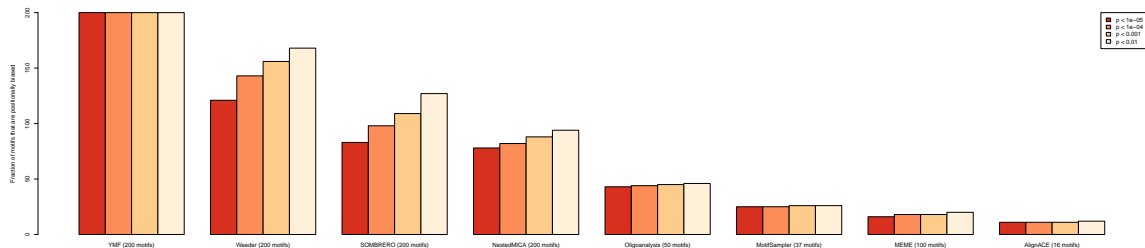


Figure 5.29: The fraction of motifs output by each of the eight methods, which show a preference for positions -500 to 0. See Section 5.2.7.1 for details regarding the method.

5.3.6.4 Combining the conservation, SNP rate and positional bias to highlight potentially functional motifs

I measured three aspects of the computationally predicted motifs as a sign of potential function: interspecies conservation, SNP rate, and positional bias of the motifs with respect to the closest transcription start sites. These properties do not appear to be randomly distributed amongst the motifs, with many motifs showing combinations of these features (Figure 5.30 shows SOMBRERO and NestedMICA motifs as an example). As also found by Down et al. (2007) in the *de novo* inference study of *D. melanogaster* regulatory motifs, a large fraction of motifs exhibit excess inter-species conservation, when compared to other intergenic sequence. The SNP rate and inter-species conservation are also closely associated, as expected.

I selected and counted motifs predicted by each of the methods which are not matches to JASPAR motifs, but show a combination of higher inter-species conservation than intergenic sequence ($p < 0.0001$), lower SNP rate than intergenic sequence in *S. cerevisiae* strains ($p < 0.0001$), and preferentially match close to the TSS ($p < 0.001$). Motifs which fit all of these criteria are shown in Figure 5.31. MEME, most likely because of its low total number of hits above the stringent bit score cutoff, did not find any such motifs.

NestedMICA found the largest number of unknown motifs of potential function (20), apart from YMF with its 182 highly redundant motifs with no sig-

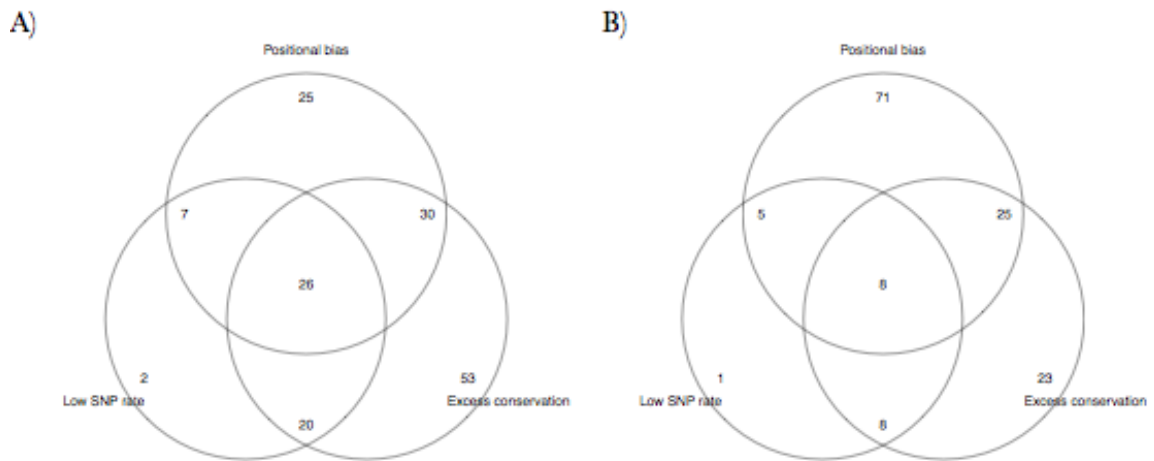


Figure 5.30: Overlap of motifs predicted by A) NestedMICA and B) SOMBRERO, that have lower SNP rate than intergenic sequence ($p < 0.0001$), higher conservation than intergenic sequence ($p < 0.0001$), and are preferential placed within -500 to 0 of TSS ($p < 0.001$).

nificant matches to known TFBS motifs (Figure 5.31G). I conducted literature searches to look for potential supporting information about the function of each of these motifs.

The TGAAAATT motif (motif12 in the NestedMICA set, motif24 in the OligoAnalysis set) is perhaps the most interesting of the patterns. It is found by two previous *S. cerevisiae* motif inference studies (Li et al., 2005; Sudarsanam et al., 2002) to be associated with the TF ABF1. The ABF1 motif in the JASPAR database, derived from the high-throughput study by Badis et al. (2008), is however markedly different (Figure 5.32).

Other potentially functional motifs are also amongst the set. NestedMICA motifs motif152 and motif190 have the consensus TATAAAA and TATAAAG. Both of these sequences have been found to bind the TATA-binding protein (Kim and Burley, 1994; Starr and Hawley, 1991). The motifs both also show a highly significant orientational bias. 60% of the 1864 hits of both motif152 and motif190 in 200bp upstream sequence regions appear as TATAAAA and TATAAAG – as opposed to TTTTATA and CTTTATA – on the same strand as the closest ORF ($p = 2.94 \times 10^{-17}$, binomial test).

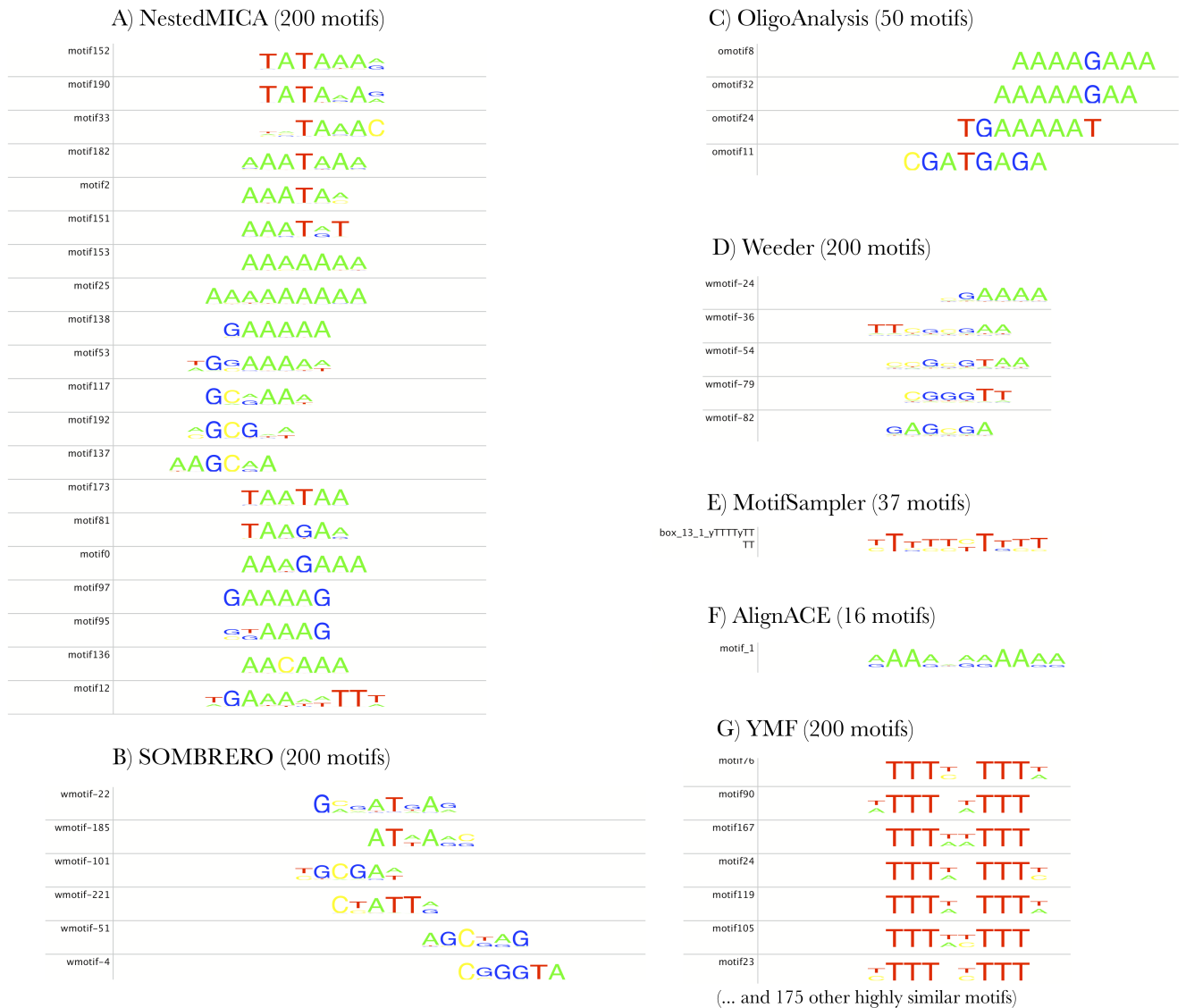


Figure 5.31: Motifs predicted by different methods which have lower SNP rate than intergenic sequence ($p < 0.0001$), higher conservation than intergenic sequence ($p < 0.0001$), and preferential placement close to the TSS ($p < 0.001$). Motifs have been aligned with iMotifs (Piipari et al., 2010b).



Figure 5.32: The ABF1 motif in the JASPAR database. Data originates from the CSI, PBM and Dip-CHIP based study by [Badis et al. \(2008\)](#).

The NestedMICA motif33 (consensus NNTAAAC) matches the motif TAAAC, which has been suggested as the motif for the yeast TF ‘Swi five factor’, or SFF ([Pic et al., 2000](#); [Tamada et al., 2003](#)). The 1252 instances of this motif in 200bp upstream sequence regions of the yeast genome show a highly significant bias in their orientation with respect to the closest ORF (56% of its instances are NNTAAAC, $p = 8.55 \times 10^{-5}$). SOMBRERO also finds a motif with a related, weaker consensus of ATAAAC.

Motif173 from the NestedMICA set has the consensus TAATAA. It has been described as a motif for the BAS2 homeobox TF ([Rolfes et al., 1997](#); [Tice-Baldwin et al., 1989](#)). Interestingly, matches of this motif are also associated with the orientation of the closest gene (54% of its instances are TAATAA, $p = 4.00 \times 10^{-5}$).

The AAAGAAA motif (motif9 in NestedMICA’s set, motifs 8 and 32 in the OligoAnalysis set) has been previously described in a phylogenetic foot printing study of the *S. cerevisiae* genome as a motif associated with genes involved in amino acid transport ([Cliften et al., 2003](#)). The reverse complement of motif motif138 (TTTGTT) corresponds to the consensus string of an HMG like TF domain ([Grosschedl et al., 1994](#)).

Several of the methods also find A- or T-rich motifs, such as AAAAAA, AAAAAAAAAA, AAATAAA or AAATAA. Although I did find publications linking some of these sequence signals, or their reverse complements, to transcriptional control, it could also be that the high conservation and low SNP rate observed for these are artefacts caused by for example the genomic multiple sequence alignment procedures which both of the conservation and SNP rate criteria depend on.

In summary, the motif inference methods studied here find several putatively functional motifs not covered by the JASPAR motif set. NestedMICA – which

is consistently the top performer in the JASPAR based performance measures shown in Sections 5.3.2, 5.3.3 and 5.3.4 – finds a varied selection of 20 motifs with high conservation, low SNP rate and a preference for matching close upstream to the TSS, but with no known regulatory motif matches in the JASPAR database. Several of the other algorithms found different subsets of these 20 motifs identified by NestedMICA. SOMBRERO finds the second largest set of motifs which fit the criteria (6 motifs).

5.3.6.5 Classification of the inferred motifs with **metamatti**

I used the **metamatti** motif classification framework presented in Chapter 4 to predict the domain family of the motifs as another way of assigning function to them (see Section 5.2.8 for a description of the method), and comparing the motifs inferred by different methods to what is known about the yeast regulatory motifs.

The random forest based **metamatti** classifier outputs a probability for each classification decision, based on votes that each of the classes received in its ensemble of classification trees. This allows for the classification to be made at a chosen level of confidence. To aid the choice of the classification probability cutoff, I plotted a number of diagnostic curves, shown in Figure 5.33. Based on the analysis, I chose the lowest classification probability cutoffs for classifying the motifs predicted by each of the eight *de novo* motif prediction methods. I set the lowest probability at 0.60. I did this because the classification accuracy drops dramatically below this probability, and effectively plateaus after it, whereas the recall stays rather stable around this classification probability, but drops rapidly from around 70%. Results were also reported at 80% classification probability.

I profiled the importances of predictor variables in a separate JASPAR motif family classification exercise, to show that several different metamotifs per class contribute strongly to the classification (see Section 1.3.4 for a discussion of the variable importance measure used). The results of this analysis are shown in Figure 5.34. For instance, all of the top ranked six features are from different fungal Zinc cluster derived metamotifs.

The classification results at the 0.6 probability cutoff are shown in Figure 5.35.

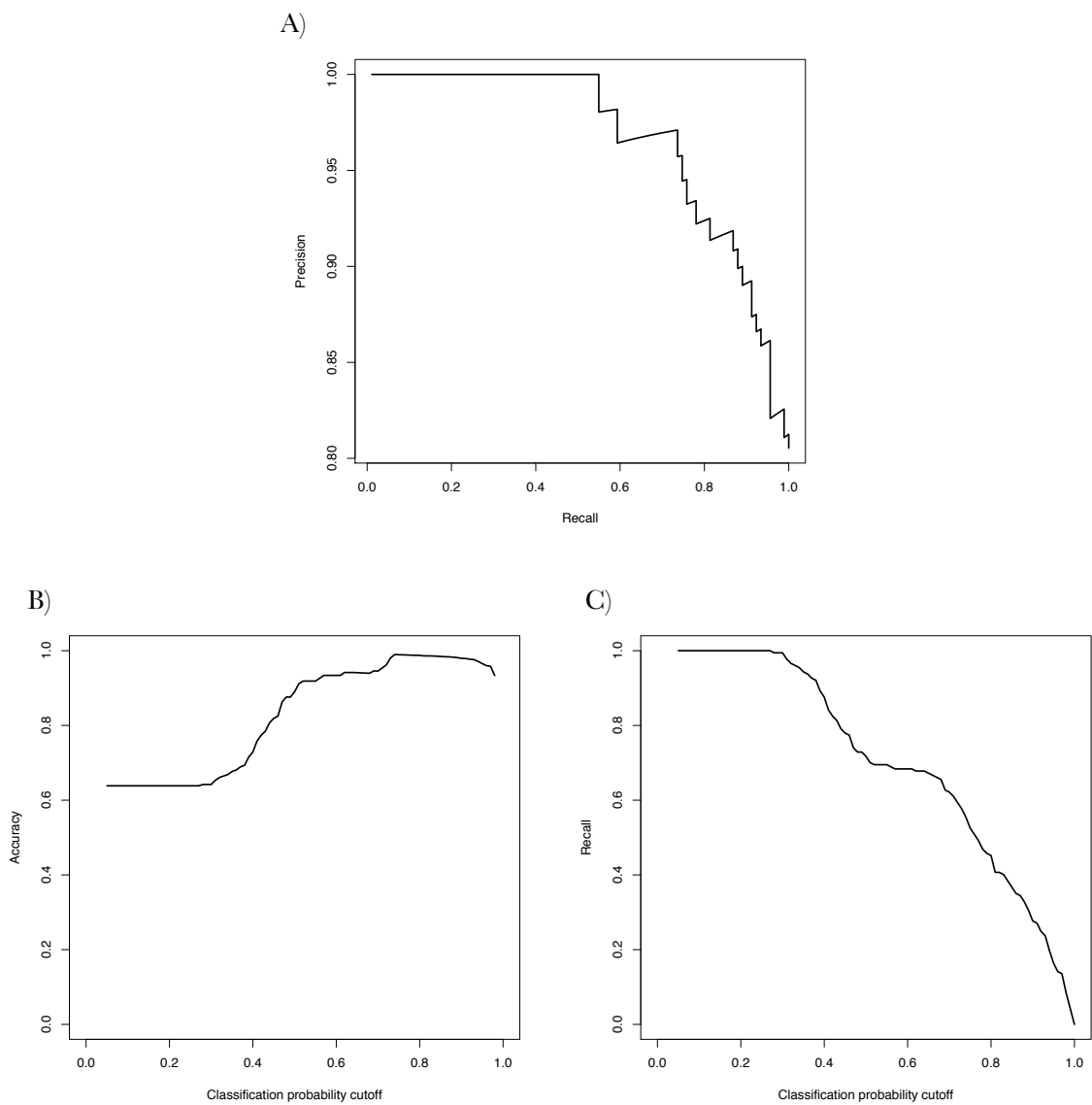


Figure 5.33: Performance measures of **metamatti** classification of JASPAR motifs. A) Precision-recall curve of 5-way JASPAR family classification training with fungal motifs in the JASPAR database. B) Accuracy as a function of the random forest classification probability cutoff. C) Recall rate as a function of the classification probability cutoff.

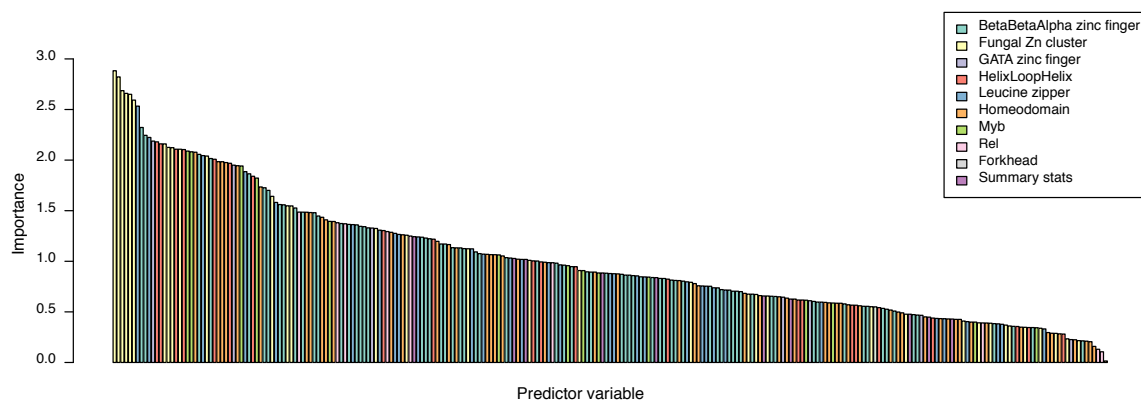


Figure 5.34: Variable importances of a JASPAR family classifier. The importance measure is described in Section 1.3.4. Metamotifs derived from ten major families of motifs in the JASPAR database were included in this exercise. One bar in the classification represents one metamotif density feature.

Instances of only two of the motif families in the 5-way classifier were found to be predicted above the cutoff, by any of the motif inference algorithms (Figure 5.35). It is disappointing that only fungal Zinc cluster motifs and $\beta\beta\alpha$ zinc finger motifs – which dominate the DNA binding domain of JASPAR motifs (Section 5.3.2) – can be detected from the *de novo* predictions at this probability cutoff. These two DNA binding domain families dominate the distribution of DBD families in the JASPAR motif set. It is however reassuring to see that in cases where there is a statistically significant close match to a JASPAR motif, the predictions are largely consistent between the `metamatti` TF family prediction (6 / 8 in the case of NestedMICA, 4 / 6 in the case of SOMBRERO, 3 / 3 in the case of Weeder), and the family of the closest JASPAR motif match. Furthermore, NestedMICA and SOMBRERO, which both show remarkably low distances to their closest JASPAR matches (Section 5.3.1), output the largest numbers of motifs which can be classified by `metamatti` at this confidence cutoff, followed by Weeder (18, 14 and 9, respectively).

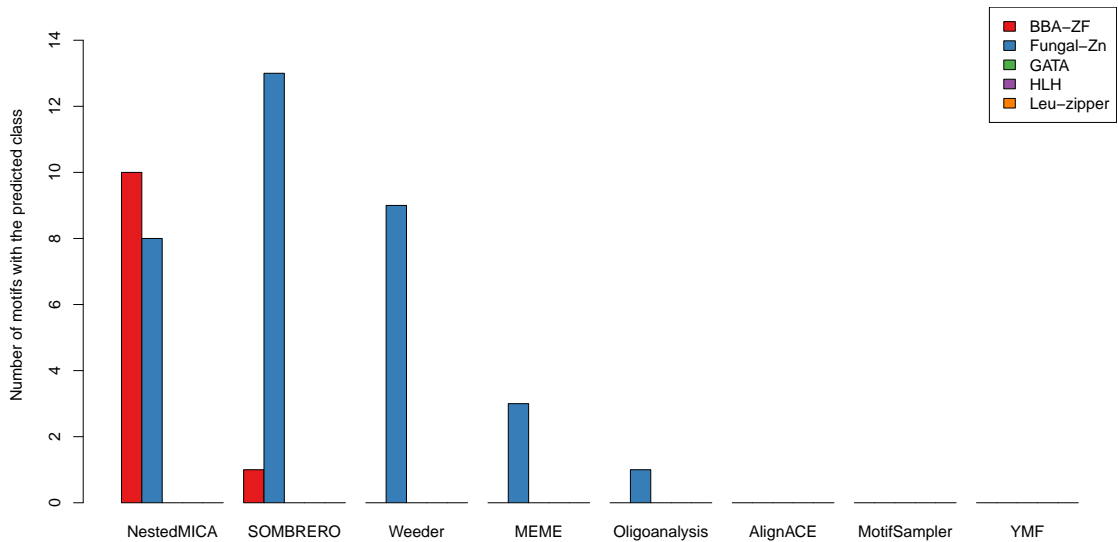


Figure 5.35: Metamatti classification of the predicted motifs at the 0.6 classification probability cutoff.

5.4 Summary

The work described in this chapter deals with large scale prediction of regulatory motifs in the *S. cerevisiae* genome, with the primary focus being a motif level performance assessment of several previously published *de novo* algorithms. The large scale motif comparison based performance assessment shown in Section 5.3.2 is in notable contrast to the binding site or nucleotide level assessments that is commonplace in motif inference literature (see Section 5.1.2). The association of a large collection of *de novo* predicted motifs with putative target genes (Section 5.3.3) has also not been previously tested in a comprehensive manner between a number of algorithms. The results of the performance assessment are rather consistent: especially NestedMICA but also SOMBRERO and MEME appear to perform adequately in finding motifs matching known regulatory motifs. None of the tested algorithms shows strong performance with the yeast genome to suggest wide applicability of *de novo* motif inference algorithms for large scale study of higher eukaryote regulatory genomes. NestedMICA's 54 statistically significant

matches to the 177 TFBS motifs in the JASPAR database is still however a surprisingly positive result for a *de novo* method when it is compared to previous work. For instance the ChIP-chip study by [Harbison et al. \(2004\)](#) reports a confident motif for 31% of 203 TFs, based on the output of six motif inference algorithms in the much easier case of finding motifs from sequence regions with ChIP based evidence of TF binding. It is also interesting to see that the arguable top performer of the ([Tompa et al., 2005](#)) assessment, Weeder, performs rather weakly using the metrics presented here. Indeed, NestedMICA, SOMBRERO and MEME are consistently the top performers in my assessment.

In addition to the performance assessment, I also profiled the conservation, SNP rate and positional bias trends of the motifs, to find motifs unknown to the JASPAR motif database but which are particularly likely to be functional (Section [5.3.6.4](#)). This analysis also showed NestedMICA with the largest collection of conserved motifs with low SNP rates and evidence for preference to genomic positions close to TSSs. This analysis however depends on the criteria used for determining a significance cutoff for genomic matches of motifs, a parameter which the especially motifs predicted by MEME did not show robustness to.

Chapter 6

Conclusions

The work in this thesis has concentrated on modelling regulatory motif families, and inferring motifs on a genome scale. Firstly, in Chapter 2 I present a novel motif family model, the metamotif. In Chapter 3 I then describe a metamotif based informative motif prior, and show its use in the NestedMICA motif discovery algorithm. The prior function substantially improves the sensitivity to detect motifs from genomic sequence.

In Chapter 4 I present another application for the metamotif: a motif classification method based on metamotif density features. I show that the metamotif based motif classifier compares favourably to previously published methods. Its performance with two novel experimental TFBS motif datasets is also found to be high, and consistent with expected error estimates. Motif classification involves learning models from highly imbalanced training datasets, simply because DNA specificity of some highly expanded TF domains has been sampled more than others. In the future, this problem will be partly addressed by increased availability of experimental motif data. In addition to expansion of the available training data, one could also take use of extensions to the random forest classification algorithm designed for learning from imbalanced training data [Chen et al. \(2004\)](#).

I introduced a visual representation for the metamotif akin to the sequence logo, with the addition of confidence intervals for symbol weights. The metamotif inference and visualisation tools have all been made openly available as part of the NestedMICA motif inference suite ([Piipari et al., 2010a](#)), the interactive

motif inference analysis environment iMotifs (Piipari et al., 2010b), as well as a **metamatti** motif classification R package and web server (manuscript in preparation). I envisage that the metamotif will have further machine learning related uses in addition to the Bayesian prior and motif family classification method I have presented. Large scale computational motif inference frameworks especially could benefit from metamotif driven semi-supervised methods to either estimate complete motif sets from novel sequence sets, or on the contrary discriminatively infer motifs not closely matching a previously described sequence motif.

As well as developing methods for motif family modelling, I conducted a large motif inference study of the *Saccharomyces cerevisiae* genome (Chapter 5), using several existing *de novo* motif inference methods. The primary motivation of this work was realistic benchmarking of *de novo* motif inference algorithms, using the *S. cerevisiae* genome as a benchmarking resource. I believe that challenging motif inference methods with large genomic sequence sets provides an objective and readily interpretable test of their abilities. Previous dedicated motif inference performance measurements (Pevzner and Sze, 2000; Tompa et al., 2005) have suffered from a self professed difficulty to define metrics to judge the algorithms with, largely caused by our lack of understanding of the principles of TF binding and properties of regulatory sequence, which hinders also creating synthetic promoter sequences. As the processes which create and constrain regulatory sequences are not well understood, the present study attempts to avoid these problems by not treating individual genomic motif hits as a primary item of interest. Instead, I judge motifs primarily based on the properties of the overall pattern, the PWM (similarly as also done in Chapter 3, and by (Down and Hubbard, 2005; Piipari et al., 2010a; Tang et al., 2008)).

Algorithms are challenged to find a collection from a single, large, real sequence dataset whose ‘motif content’ is not known accurately. Tompa et al. (2005) test the ability of algorithms to find instances of a single motif from a series of small, mostly synthetic sequence sets (each with tens to hundreds of sequences), where at least one instance of the sought after motif is present in all sequences with a motif. Furthermore, the performance measures made here are made primarily on the motif level, rather than the binding site or nucleotide level. This study addresses directly some of the problems associated with the (Tompa

[et al., 2005](#)) assessment, which is the most comprehensive motif inference method assessment to date (see Section 5.1.3).

The most important distinction of this work to previous motif inference benchmarks is that the present study allows clear conclusions to be made regarding applicability of motif inference methods – with my parameter choices – to genome scale motif inference problems. Out of the eight methods successfully tested, especially NestedMICA but also SOMBRERO and MEME appear to perform adequately, with NestedMICA discovering statistically significant matches to 30% of the motifs in the JASPAR database.

The consistently high performance observed with the NestedMICA algorithm, when compared to the other tested algorithms, is most likely attributable to a combination of factors; A state of the art Monte Carlo sampling strategy, that is robust to local maxima, is used. The sequence–motif mixture model which allows concurrent inference of a large number of motifs is also likely to be of benefit in large scale problems. Interestingly SOMBRERO, whose self-organising map based inference strategy is also clearly aimed at concurrent, ‘non-greedy’ motif inference problems, performs well in the problem. The NestedMICA sequence background model which accounts for nucleotide content variation observed in genomic DNA is also a likely contributing factor to high sensitivity from large set of promoters. Importantly, the assessment also suggests certain improvements to how the algorithms should be run; NestedMICA for instance predicts systematically shorter motifs than the matching JASPAR motifs, and therefore for large scale studies it’s minimum motif length parameter should be increased from 6 (which was used in this study).

I also conducted experiments with the inferred motifs involving scanning with a significance cutoff, mostly as a data exploration exercise. This was done in cases where a non-parametric alternative was not apparent (e.g. positional bias). The scanning based analyses highlight the difficulties involved in determining a meaningful significance cutoff for motifs output by a number of algorithms, with different lengths and information content profiles. Problems encountered with genomic motif match based analyses, with the MEME algorithm ([Bailey et al., 2006](#)) in particular, demonstrate the need for parameter free performance assessment of motif inference methods.

6.1 Future work

Much of the work that I did during my project relied on a gene regulatory motif inference strategy whereby regulatory sequence motifs are sought from promoter sequence by looking for overrepresented sequence signals. This strategy has been successfully applied to many problems in regulatory genomics, as has been discussed in the previous chapters, but it clearly has its limitations.

1. Higher eukaryotes that have large genomes and a multitude of gene regulatory mechanisms, including several thousands of TFs. As my work from Chapter 5 suggests, finding complete higher eukaryotic regulatory motif dictionaries with a purely reference genome based strategy is not realistic, given that current algorithms struggle already with the yeast genome of approximately 200 TFs.
2. Overrepresentation of a motif in genomic sequence does not necessarily imply action in gene regulation. Solely sequence based methods do not distinguish motifs acting in transcriptional regulation from other possible recurring signals.
3. Expression patterns of eukaryotic cells are not regulated by independent factors, but by multiple factors that bind in complexes. Complex combinatorial regulatory programs consisting of specific TF complexes are known to be responsible for instance for tissue (Ravasi et al., 2010) or development stage (Levine and Davidson, 2005) specificity of gene regulation. When information is available of potential combinatorial regulation of genes by a group of TFs, it should be possible to input this information for a motif inference algorithm.

Towards the end of my project I became interested of developing methods which address the above limitations by allowing use of gene expression patterns as an evidence source in a probabilistic motif inference algorithm capable of large scale inference. In particular I wanted to test if the NestedMICA algorithm could be modified to include a prior probability function over the motif-to-gene mixing matrices (see Section 1.3.3 for a discussion of the NestedMICA algorithm),

which would encode information derived from a gene expression correlation pattern. More specifically, I consider that mixing matrix states where the correlation of occupancy (presence or absence) of motifs in promoter sequences mimics the correlation of the gene expression states should be more likely states than those where the mixing state correlations differ significantly from the gene expression correlations. I began an effort in developing and optimising a variant of the algorithm for this purpose, and although I did not complete this work, I did solve some sub-problems. I will discuss my proposed method here because its definition could be helpful for others aiming to implement a related stochastic motif inference strategy that acts on regulatory sequence with correlated combinations of motif instances.

The particular prior probability function $\mathbb{P}(\mathbf{M}|\mathbf{G}, p)$ which I developed is noted in Equation 6.1. The probability is over the space of motif-to-gene occupancy matrices \mathbb{M} , given the gene expression matrix \mathbf{G} and an adjustable precision parameter p . The root mean square deviation (*RMSD*) of a gene expression correlation matrix, and the correlation of the occupancy matrix \mathbf{M} follows a Gaussian distribution with precision p (an adjustable parameter). Dimensions of an occupancy matrix \mathbf{M} is $m \times g$, where m is the number of motifs and g the number of genes. The gene expression matrix \mathbf{G} has the dimensionality $g \times n$ (n measurements).

$$\mathbb{P}(\mathbf{M}|\mathbf{G}, p) = \text{Gauss}(\text{RMSD}(\text{corr}(\mathbf{G}), \text{corr}(\mathbf{M})), p) \quad (6.1)$$

I implemented a Metropolis-Hastings algorithm (Hastings, 1970) to draw samples from $\mathbb{P}(\mathbf{M}|\mathbf{G}, p)$. A naive implementation of the occupancy prior sampling by MH proved prohibitively costly in computational time due to the order of n^2 time complexity of the RMSD computation required during each iteration of the long burn-in phase required by the MH algorithm. Therefore I optimised the algorithm to only update contributions of the changed elements in the mixture matrix. Several important steps were also made to decrease the runtime memory use of the algorithm. The end result of my work is an algorithm which performs with sufficiently low CPU and runtime memory requirements to be applied in the NestedMICA algorithm comfortably with several thousands of sequences and

10–100 motifs. Figure 6.1 shows three different Markov chains of the $\mathbb{P}(\mathbf{M}|\mathbf{G}, p)$ sampling algorithm which I developed, with different values of the precision (p) parameter.

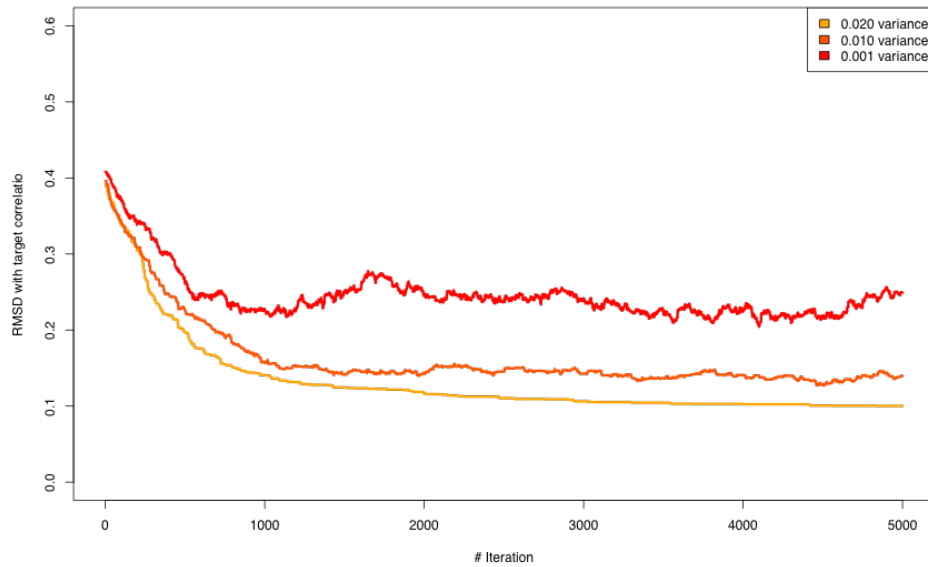


Figure 6.1: Three Markov chains aiming to draw a sample from $\mathbb{P}(\mathbf{M}|\mathbf{G}, p)$, each with a different p parameter.

Figure 6.2 shows an example of the mixture matrix sampling. The end result of sampling is shown in Figure 6.2D, and its correlation matrix is in 6.2C. Figure 6.3 shows an example mixing matrix created by the sampler as being closely related in its correlation pattern to the target correlation pattern given as input to it.

I believe that development of motif inference methods which are capable of integrating several sources of experimental evidence with a well performing probabilistic *de novo* motif inference method have a lot to offer in regulatory motif inference problems, as more and more genome-wide regulatory data becomes available. The metamotif prior function can be considered one such source of experimental evidence. Other sources could be for instance epigenetic marks, or gene expression data as discussed above. Whether a variant of the NestedMICA

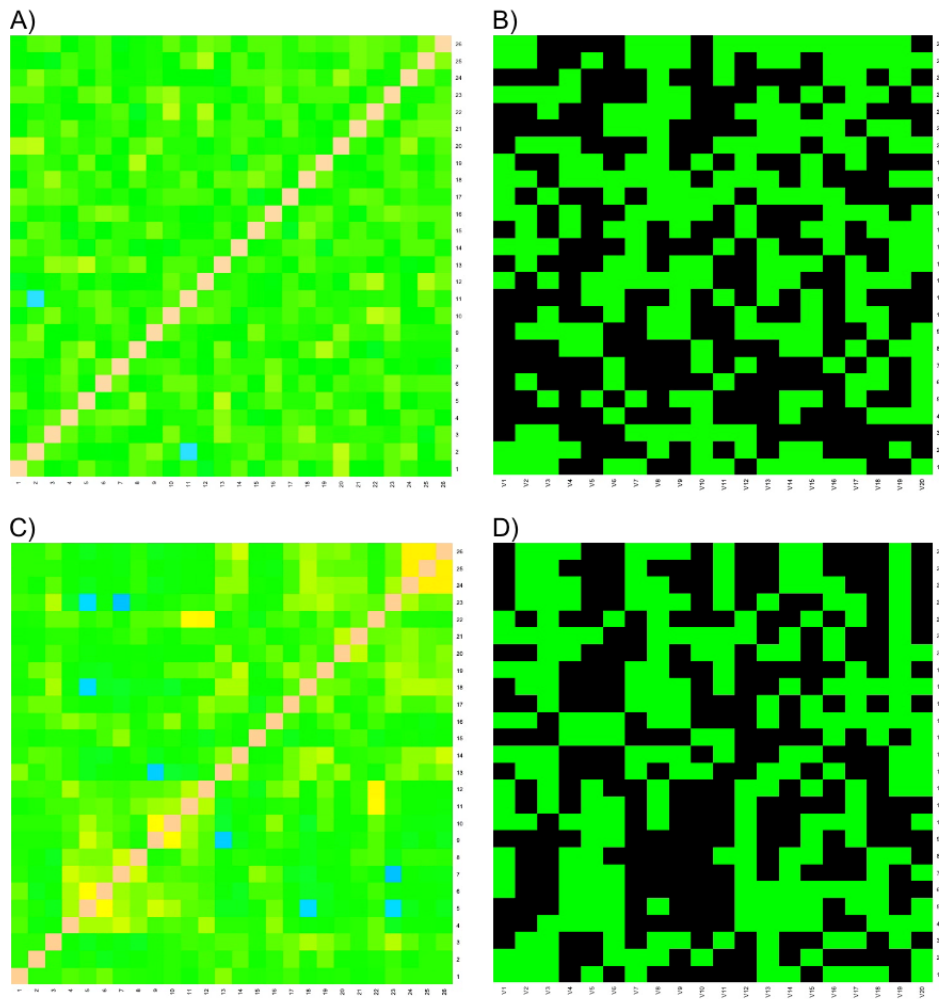


Figure 6.2: Mixing matrices and their correlations. The correlation matrices (panels A and C) of the start and end state of one of the 5000 step long MC chains from Figure 6.1. Panels B and D show the mixing matrices at the start (A) and end (D) of the sampling. Black states in panels B and D are mixing matrix elements with value 0 and green states those with value 1.

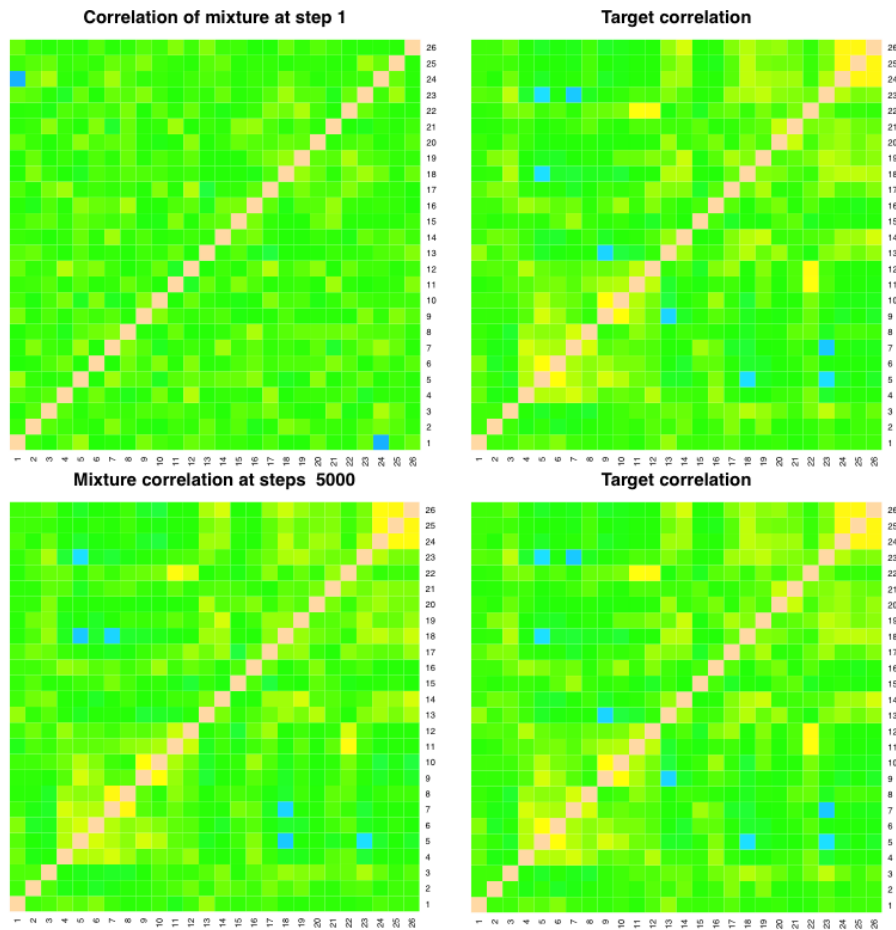


Figure 6.3: The sampling algorithm produces mixing matrices that are closely related in correlation pattern to the target (gene expression) correlation matrix. Gene expression correlations are shown on the right, and the mixture matrix correlations in the left. Whereas there are hardly any correlated states in the mixture matrix at step 1 of the algorithm, after the burn-in (at step 5000) the correlation pattern of the mixture state closely corresponds to the target correlation.

algorithm with a ‘target correlation aware’ mixing matrix prior function turns out to perform well with real genomic sequence remains to be seen. Other potentially more natural formulations could also be used to ‘inject’ gene expression information into a Bayesian motif inference method such as NestedMICA. For instance the mutual information between gene expression patterns and motif occurrences could be used, as done with a greedy motif estimation algorithm in [Elemento et al. \(2007\)](#). Alternatively, the independent component analysis like formulation in NestedMICA could be extended to learn, simultaneously, patterns of gene expression and motifs associated with these patterns. Further work in the direction of data integration in computational motif inference has great potential in improving our understanding of the regulation of genomes.

References

- Adryan, B. and Teichmann, S.A. FlyTF: a systematic review of site-specific transcription factors in the fruit fly *Drosophila melanogaster*. *Bioinformatics*, 2006. URL <http://www.google.com/search?client=safari&rls=en-us&q=FlyTF:+a+systematic+review+of+site-specific+transcription+factors+in+the+fruit+fly+Drosophila+melanogaster&ie=UTF-8&oe=UTF-8>. 6
- Aerts, S., Thijs, G., Coessens, B., Staes, M., et al. Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic Acids Res*, 31(6):1753–1764, 2003. 183
- Affolter, M., Slattery, M., and Mann, R.S. A lexicon for homeodomain-DNA recognition. *Cell*, 133(7):1133–5, 2008. doi:10.1016/j.cell.2008.06.008. URL [http://linkinghub.elsevier.com/retrieve/pii/S0092-8674\(08\)00760-5](http://linkinghub.elsevier.com/retrieve/pii/S0092-8674(08)00760-5). 8
- Albrecht, G., Mösch, H., Hoffmann, B., and Reusser, U. Monitoring the Gcn4 Protein-mediated Response in the Yeast *Saccharomyces cerevisiae*. *Journal of Biological ...*, 1998. URL <http://www.jbc.org/content/273/21/12696.full>. 144
- Altschul, S.F. and Gish, W. Local alignment statistics. *Methods in Enzymology*, 266:460–80, 1996. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=8743700&dopt=abstractplus. 109
- Anthony-Cahill, S.J., Benfield, P.A., Fairman, R., Wasserman, Z.R., et al. Molec-

REFERENCES

- ular characterization of helix-loop-helix peptides. *Science*, 255(5047):979–83, 1992. URL <http://www.sciencemag.org/cgi/reprint/255/5047/979>. 50
- Ao, W., Gaudet, J., Kent, W.J., Muttumu, S., et al. Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. *Science*, 305(5691):1743–6, 2004. doi:10.1126/science.1102216. URL <http://www.sciencemag.org/cgi/content/full/305/5691/1743>. 102, 113
- Babu, M., Luscombe, N., Aravind, L., Gerstein, M., et al. Structure and evolution of transcriptional regulatory networks. *Current Opinion in Structural Biology*, 14(3):283–291, 2004. 9
- Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., et al. Diversity and complexity in DNA recognition by transcription factors. *Science*, 324(5935):1720–3, 2009. doi:10.1126/science.1162327. URL <http://www.sciencemag.org/cgi/content/full/324/5935/1720>. 12, 14, 18, 35, 38
- Badis, G., Chan, E.T., van Bakel, H., Pena-Castillo, L., et al. A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters. *Mol Cell*, 32(6):878–87, 2008. doi:10.1016/j.molcel.2008.11.020. XIV, 12, 99, 114, 115, 165, 167
- Bailey, T.L. and Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36, 1994. 20
- Bailey, T.L. and Elkan, C. The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol*, 3:21–29, 1995. 16, 27, 67
- Bailey, T.L., Williams, N., Misleh, C., and Li, W.W. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, 34(Web Server issue):W369–73, 2006. doi:10.1093/nar/gkl198. URL http://nar.oxfordjournals.org/cgi/content/full/34/suppl_2/W369. 27, 175, 198

REFERENCES

- Baker, J. Stochastic modeling as a means of automatic speech recognition. *oai.dtic.mil*, 1975. URL <http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA013808>. 19
- Banerji, J., Rusconi, S., and Schaffner, W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, 27(2 Pt 1):299–308, 1981. URL <http://www.cell.com/retrieve/pii/009286748190413X>. 2
- Barash, Y., Elidan, G., Friedman, N., and Kaplan, T. Modeling dependencies in protein-DNA binding sites. *Proceedings of the seventh annual international conference . . .*, 2003. URL <http://portal.acm.org/citation.cfm?id=640079>. 18
- Baum, L. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, 1972. URL <http://research.microsoft.com/apps/pubs/default.aspx?id=64753>. 19
- Baum, L. and Petrie, T. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 1966. URL <http://www.jstor.org/stable/2238772>. 19
- Baum, L., Petrie, T., Soules, G., and Weiss, N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical . . .*, 1970. URL <http://www.jstor.org/stable/2239727>. 19
- Baum, L. and Eagon, J. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bull. Amer. Math. Soc*, 73(3):360–363, 1967. 19
- Baum, L. and Sell, G. Growth transformations for functions on manifolds. *Pac. J. Math*, 27:211–227, 1968. 19
- Beck, T. and Hall, M. The TOR signalling pathway controls nuclear localization of nutrient-regulated transcription factors. *Nature*, 1999. URL <http://www.nature.com/nature/journal/v402/n6762/abs/402689a0.html>. 144

REFERENCES

- Ben-Gal, I., Shani, A., Gohr, A., Grau, J., et al. Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics*, 21(11):2657–66, 2005. doi:10.1093/bioinformatics/bti410. URL <http://bioinformatics.oxfordjournals.org/cgi/content/full/21/11/2657>. 18
- Benos, P.V., Bulyk, M.L., and Stormo, G.D. Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res*, 30(20):4442–4451, 2002a. 18, 38
- Benos, P.V., Lapedes, A.S., and Stormo, G.D. Probabilistic code for DNA recognition by proteins of the EGR family. *J Mol Biol*, 323(4):701–27, 2002b. 39
- Berger, M., Badis, G., Gehrke, A., Talukder, S., et al. Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences. *Cell*, 133(7):1266–1276, 2008. doi:10.1016/j.cell.2008.05.024. XI, 35, 36, 91, 92
- Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., et al. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotechnol*, 24(11):1429–35, 2006. doi:10.1038/nbt1246. URL <http://www.nature.com/nbt/journal/v24/n11/abs/nbt1246.html>. 18, 35, 36, 114, 115, 183
- Bergman, C. and Kreitman, M. Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Res*, 2001. URL <http://genome.cshlp.org/content/11/8/1335.full>. 5
- Beyer, A., Workman, C., Hollunder, J., Radke, D., et al. Integrated Assessment and Prediction of Transcription Factor Binding. *PLoS Computational Biology*, 2(6):e70, 2006. doi:10.1371/journal.pcbi.0020070.st005. 143
- Bhaskar, H., Hoyle, D.C., and Singh, S. Machine learning in bioinformatics: a brief survey and recommendations for practitioners. *Comput Biol Med*, 36(10):1104–1125, 2006. doi:10.1016/j.compbiomed.2005.09.002. URL <http://dx.doi.org/10.1016/j.compbiomed.2005.09.002>. 29

REFERENCES

- Bi, X. and Broach, J.R. Chromosomal boundaries in *S. cerevisiae*. *Curr Opin Genet Dev*, 11(2):199–204, 2001. 2
- Birney, E., Andrews, T.D., Bevan, P., Caccamo, M., et al. An overview of Ensembl. *Genome Res*, 14(5):925–8, 2004. doi:10.1101/gr.1860604. URL <http://genome.cshlp.org/cgi/content/full/14/5/925>. 31, 109
- Blanchette, M. and Sinha, S. Separating real motifs from their artifacts. *Bioinformatics*, 17 Suppl 1:S30–8, 2001. URL http://bioinformatics.oxfordjournals.org/cgi/reprint/17/suppl_1/S30?view=long&pmid=11472990. 199
- Borneman, A.R., Gianoulis, T.A., Zhang, Z.D., Yu, H., et al. Divergence of transcription factor binding sites across related yeast species. *Science*, 317(5839):815–9, 2007. doi:10.1126/science.1140748. URL <http://www.sciencemag.org/cgi/content/full/317/5839/815>. 5, 112
- Brazma, A., Jonassen, I., Vilo, J., and Ukkonen, E. Predicting gene regulatory elements in silico on a genomic scale. *Genome Res*, 8(11):1202–15, 1998. 100
- Breiman, L. Bagging predictors. *Machine Learning*, 1996. URL <http://www.springerlink.com/index/L4780124W2874025.pdf>. 29
- Breiman, L. Random Forests. *Machine Learning*, 2001a. URL <http://www.springerlink.com/index/U0P06167N6173512.pdf>. 31
- Breiman, L. Random Forests. *Machine Learning*, 45(1):5–32, 2001b. URL citeseer.ist.psu.edu/breiman01random.html. 29, 30, 84
- Brenowitz, M., Senear, D.F., Shea, M.A., and Ackers, G.K. Quantitative DNase footprint titration: a method for studying protein-DNA interactions. *Meth Enzymol*, 130:132–81, 1986. 33, 119
- Brown, M., Hughey, R., Krogh, A., Mian, I.S., et al. Using Dirichlet mixture priors to derive hidden Markov models for protein families. *Proc Int Conf Intell Syst Mol Biol*, 1:47–55, 1993. 19, 68

REFERENCES

- Bryne, J.C., Valen, E., Tang, M.H.E., Marstrand, T., et al. JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res*, 36(Database issue):D102–6, 2008. doi: 10.1093/nar/gkm955. URL http://nar.oxfordjournals.org/cgi/content/full/36/suppl_1/D102. 34, 183
- Buchman, A.R., Kimmerly, W.J., Rine, J., and Kornberg, R.D. Two DNA-binding factors recognize specific sequences at silencers, upstream activating sequences, autonomously replicating sequences, and telomeres in *Saccharomyces cerevisiae*. *Mol Cell Biol*, 8(1):210–25, 1988. 2
- Bulyk, M.L., Johnson, P.L.F., and Church, G.M. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res*, 30(5):1255–1261, 2002. 18
- Burge, C., Campbell, A.M., and Karlin, S. Over- and under-representation of short oligonucleotides in DNA sequences. *Proc Natl Acad Sci USA*, 89(4):1358–62, 1992. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=1741388&dopt=abstractplus. 28
- Burge, C. and Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*, 268(1):78–94, 1997. doi:10.1006/jmbi.1997.0951. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WK7-45VGF7T-9&_user=10&_coverDate=04rch&_sort=d&_docanchor=&view=c&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=9a6fd068b3be7c44c65d6d208d7d58dc. 106
- Burset, M. and Guigó, R. Evaluation of gene structure prediction programs. *Genomics*, 34(3):353–67, 1996. doi:10.1006/geno.1996.0298. 105, 106
- Bussemaker, H.J., Li, H., and Siggia, E.D. Regulatory element detection using correlation with expression. *Nat Genet*, 27(2):167–71, 2001. doi:10.1038/84792. URL http://www.nature.com/ng/journal/v27/n2/abs/ng0201_167.html;jsessionid=CC8DF0DE6E3E8EB39B739F0A6F599051. 15, 101

REFERENCES

- Bussemaker, H., Li, H., and Siggia, E. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proceedings of the National Academy of Sciences*, 97(18):10096, 2000. 101
- Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*, 38(6):626–635, 2006. doi:10.1038/ng1789. URL <http://dx.doi.org/10.1038/ng1789>. 5, 8
- Carroll, S., Grenier, J., and Weatherbee, S. From DNA to diversity: The primacy of regulatory evolution. 2000. URL <http://www.google.com/search?client=safari&rls=en-us&q=From+DNA+to+diversity:+The+primacy+of+regulatory+evolution&ie=UTF-8&oe=UTF-8>. 5
- Chan, T., Li, G., Leung, K., and Lee, K. Discovering multiple realistic TFBS motifs based on a generalized model. *BMC bioinformatics*, 10(1):321, 2009. 108
- Chen, C., Liaw, A., and Breiman, L. Using Random Forest to Learn Imbalanced Data. *Unpublished manuscript*, p. 12, 2004. 173
- Chen, X., Xu, H., Yuan, P., Fang, F., et al. Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, 133(6):1106–1117, 2008. 67
- Chenna, R., Sugawara, H., Koike, T., Lopez, R., et al. Multiple sequence alignment with the Clustal series of programs. *Nucleic acids research*, 31(13):3497–500, 2003. 48
- Chi, Y., Huddleston, M.J., Zhang, X., Young, R.A., et al. Negative regulation of Gcn4 and Msn2 transcription factors by Srb10 cyclin-dependent kinase. *Genes & Development*, 15(9):1078–92, 2001. doi:10.1101/gad.867501. URL <http://genesdev.cshlp.org/content/15/9/1078.long>. 144
- Choi, J.K. and Kim, Y.J. Epigenetic regulation and the variability of gene expression. *Nat Genet*, 40(2):141–7, 2008. doi:10.1038/ng.2007.58.

REFERENCES

- URL <http://www.nature.com/ng/journal/v40/n2/abs/ng.2007.58.html;jsessionid=1E050424C5A8D78C23141658C4087ECD>. 7
- Clements, M., van Someren, E.P., Knijnenburg, T.A., and Reinders, M.J.T. Integration of known transcription factor binding site information and gene expression data to advance from co-expression to co-regulation. *Genomics Proteomics Bioinformatics*, 5(2):86–101, 2007. doi:10.1016/S1672-0229(07)60019-9. CBF1 and PHO4 have closely similar motifs. 141
- Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., et al. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. *Science*, 301(5629):71–6, 2003. doi:10.1126/science.1084337. URL <http://www.sciencemag.org/cgi/content/full/301/5629/71>. 167
- Comon, P. Independent component analysis, a new concept? *Signal processing*, 1994. URL <http://linkinghub.elsevier.com/retrieve/pii/0165168494900299>. 26
- Conlon, E.M., Liu, X.S., Lieb, J.D., and Liu, J.S. Integrating regulatory motif discovery and genome-wide expression analysis. *Proc Natl Acad Sci USA*, 100(6):3339–44, 2003. doi:10.1073/pnas.0630591100. URL <http://www.pnas.org/content/100/6/3339>. 15
- Consortium, F., Suzuki, H., Forrest, A.R.R., van Nimwegen, E., et al. The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat Genet*, 41(5):553–62, 2009. doi:10.1038/ng.375. 5
- Cremer, T. and Cremer, C. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet*, 2(4):292–301, 2001. doi:10.1038/35066075. URL http://www.nature.com/nrg/journal/v2/n4/full/nrg0401_292a.html. 3
- Cunningham, T.S. and Cooper, T.G. The *Saccharomyces cerevisiae* DAL80 repressor protein binds to multiple copies of GATAA-containing sequences (URSGATA). *Journal of Bacteriology*, 175(18):5851–61, 1993. URL <http://jb.asm.org/cgi/reprint/175/18/5851?view=long&pmid=8376332>. 6

REFERENCES

- Das, M.K. and Dai, H.K. A survey of DNA motif finding algorithms. *BMC Bioinformatics*, 8 Suppl 7:S21, 2007. doi:10.1186/1471-2105-8-S7-S21. URL <http://www.biomedcentral.com/1471-2105/8/S7/S21>. 13, 186
- Dempster, A., Laird, N., and Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. 27
- Dietterich, T. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 1998. URL <http://www.mitpressjournals.org/doi/abs/10.1162/089976698300017197>. 29
- Dogrue, M. Motif based computational identification of protein subcellular localisation. pp. 1–255, 2008. 19, 56
- Dogrue, M., Down, T., and Hubbard, T. NestedMICA as an ab initio protein motif discovery tool. *BMC Bioinformatics*, 9(1):19, 2008. doi:10.1186/1471-2105-9-19. URL <http://dx.doi.org/10.1186/1471-2105-9-19>. 23, 25, 53, 56
- Donahue, T.F., Daves, R.S., Lucchini, G., and Fink, G.R. A short nucleotide sequence required for regulation of HIS4 by the general control system of yeast. *Cell*, 32(1):89–98, 1983. 6
- Down, T.A., Bergman, C.M., Su, J., and Hubbard, T.J.P. Large-Scale Discovery of Promoter Motifs in *Drosophila melanogaster*. *PLoS Comput Biol*, 3(1):e7, 2007. doi:10.1371/journal.pcbi.0030007. URL <http://dx.doi.org/10.1371/journal.pcbi.0030007>. XIII, 48, 61, 63, 66, 72, 75, 84, 116, 117, 119, 123, 145, 148, 164, 185, 186
- Down, T.A. and Hubbard, T.J.P. NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Res*, 33(5):1445–1453, 2005. doi:10.1093/nar/gki282. URL <http://dx.doi.org/10.1093/nar/gki282>. 17, 21, 25, 26, 66, 71, 74, 77, 103, 106, 116, 174, 185, 190
- Dreier, B., Beerli, R.R., Segal, D.J., Flippin, J.D., et al. Development of zinc finger domains for recognition of the 5'-ANN-3' family of DNA sequences and

REFERENCES

- their use in the construction of artificial transcription factors. *J Biol Chem*, 276(31):29466–78, 2001. doi:10.1074/jbc.M102604200. 90
- Dreier, B., Segal, D.J., and Barbas, C.F. Insights into the molecular recognition of the 5'-GNN-3' family of DNA sequences by zinc finger domains. *J Mol Biol*, 303(4):489–502, 2000. doi:10.1006/jmbi.2000.4133. URL <http://dx.doi.org/10.1006/jmbi.2000.4133>. 90
- Eddy, S.R. Profile hidden Markov models. *Bioinformatics*, 14(9):755–63, 1998. URL <http://bioinformatics.oxfordjournals.org/cgi/reprint/14/9/755?view=long&pmid=9918945>. 19
- Egnier, M.R. Rare Events and Conditional Events on Random Strings. 2004. URL <http://citeseer.ist.psu.edu/637853>. 113
- Elemento, O., Slonim, N., and Tavazoie, S. A Universal Framework for Regulatory Element Discovery across All Genomes and Data Types. *Molecular Cell*, 28(2):337–350, 2007. doi:10.1016/j.molcel.2007.09.027. 181
- Elemento, O. and Tavazoie, S. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol*, 6(2):R18, 2005. doi:10.1186/gb-2005-6-2-r18. 101, 111
- Ellingsen, S., Laplante, M.A., König, M., Kikuta, H., et al. Large-scale enhancer detection in the zebrafish genome. *Development*, 132(17):3799–811, 2005. doi:10.1242/dev.01951. 2
- Ernst, J., Plasterer, H.L., Simon, I., and Bar-Joseph, Z. Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res*, 20(4):526–36, 2010. doi:10.1101/gr.096305.109. URL <http://genome.cshlp.org/content/20/4/526.long>. 107
- Eskin, E. and Pevzner, P.A. Finding composite regulatory patterns in DNA sequences. *Bioinformatics*, 18 Suppl 1:S354–63, 2002. 113
- Ettwiller. *Genome Res*, 2005. 101

REFERENCES

- Fauteux, F., Blanchette, M., and Strömviik, M.V. Seeder: discriminative seeding DNA motif discovery. *Bioinformatics*, 24(20):2303–7, 2008. doi:10.1093/bioinformatics/btn444. URL <http://bioinformatics.oxfordjournals.org/cgi/content/full/24/20/2303>. 108
- Favorov, A.V., Gelfand, M.S., Gerasimova, A.V., Ravcheev, D.A., et al. A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. *Bioinformatics*, 21(10):2240–5, 2005. doi:10.1093/bioinformatics/bti336. 113
- Fejes, A.P., Robertson, G., Bilenky, M., Varhol, R., et al. FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, 24(15):1729–1730, 2008. doi:10.1093/bioinformatics/btn305. 185, 189
- Finn, R.D., Mistry, J., Tate, J., Coghill, P., et al. The Pfam protein families database. *Nucleic Acids Res*, 38(Database):D211–D222, 2010. doi:10.1093/nar/gkp985. 6, 19
- FitzGerald, P.C., Sturgill, D., Shyakhtenko, A., Oliver, B., et al. Comparative genomics of *Drosophila* and human core promoters. *Genome Biol*, 7(7):R53, 2006. doi:10.1186/gb-2006-7-7-r53. URL <http://dx.doi.org/10.1186/gb-2006-7-7-r53>. 27
- Flicek, P., Aken, B.L., Beal, K., Ballester, B., et al. Ensembl 2008. *Nucleic Acids Res*, 36(Database issue):D707–14, 2008. doi:10.1093/nar/gkm988. URL http://nar.oxfordjournals.org/cgi/content/full/36/suppl_1/D707. 71
- Foat, B.C., Houshmandi, S.S., Olivas, W.M., and Bussemaker, H.J. Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. *Proc Natl Acad Sci USA*, 102(49):17675–80, 2005. doi:10.1073/pnas.0503803102. URL <http://www.pnas.org/content/102/49/17675.long>. 15
- Freund, Y. and Schapire, R. Experiments with a new boosting algorithm. *MACHINE LEARNING-INTERNATIONAL ...*, 1996. URL <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.90.4143&rep=rep1&type=pdf>. 30

REFERENCES

- Fried, M. and Crothers, D. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Res*, 1981a. URL <http://nar.oxfordjournals.org/cgi/content/abstract/9/23/6505>. 119
- Fried, M. and Crothers, D.M. Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. *Nucleic Acids Res*, 9(23):6505–25, 1981b. 33
- Frith, M.C., Valen, E., Krogh, A., Hayashizaki, Y., et al. A code for transcription initiation in mammalian genomes. *Genome Res*, 18(1):1–12, 2008. doi:10.1101/gr.6831208. URL <http://dx.doi.org/10.1101/gr.6831208>. 3
- Fuda, N.J., Ardehali, M.B., and Lis, J.T. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature*, 461(7261):186–192, 2009. doi:10.1038/nature08449. 1, 4
- Fulton, D.L., Sundararajan, S., Badis, G., Hughes, T.R., et al. TFCat: the curated catalog of mouse and human transcription factors. *Genome Biol*, 10(3):R29, 2009. doi:10.1186/gb-2009-10-3-r29. 6
- Galant, R. and Carroll, S. Evolution of transcriptional repression domain in an insect Hox protein. *Nature*, 2002. URL <http://www.google.com/search?client=safari&rls=en-us&q=Evolution+of+transcriptional+repression+domain+in+an+insect+Hox+protein.&ie=UTF-8&oe=UTF-8>. 5
- Garcia, F., Lopez, F.J., Cano, C., and Blanco, A. FISim: a new similarity measure between transcription factor binding sites based on the fuzzy integral. *BMC bioinformatics*, 10:224, 2009. doi:10.1186/1471-2105-10-224. URL <http://www.biomedcentral.com/1471-2105/10/224>. 41
- Garner, M.M. and Revzin, A. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Res*, 9(13):3047–60, 1981. 33

REFERENCES

- Gasser, S.M. Visualizing chromatin dynamics in interphase nuclei. *Science*, 296(5572):1412–6, 2002. doi:10.1126/science.1067703. URL <http://www.sciencemag.org/cgi/content/full/296/5572/1412>. 3
- Gelfand, M. and Mirny, L. Structural analysis of conserved base pairs in protein-DNA complexes. *Nucleic Acids Res*, 2002. URL <http://www.ingentaconnect.com/content/oup/nar/2002/00000030/00000007/art01704>. 14
- Georges, A.B., Benayoun, B.A., Caburet, S., and Veitia, R.A. Generic binding sites, generic DNA-binding domains: where does specific promoter recognition come from? *The FASEB Journal*, 24(2):346–356, 2010. doi:10.1096/fj.09-142117. 8
- Gertz, J., Siggia, E.D., and Cohen, B.A. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. *Nature*, 457(7226):215–218, 2009. doi:10.1038/nature07521. 14, 107
- Goffeau, A., Barrell, B., Bussey, H., Davis, R., et al. Life with 6000 genes. *Science*, 274(5287):546, 1996. 99, 121, 198
- Gordan, R., Narlikar, L., and Hartemink, A.J. Finding regulatory DNA motifs using alignment-free evolutionary conservation information. *Nucleic Acids Research*, 38(6):e90–e90, 2010. doi:10.1093/nar/gkp1166. 5, 15
- Gordân, R. and Hartemink, A.J. Using DNA duplex stability information for transcription factor binding site discovery. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, pp. 453–64, 2008. URL <http://www.ncbi.nlm.nih.gov/pubmed/18229707?dopt=abstract>. 15
- Gotea, V., Visel, A., Westlund, J.M., Nobrega, M.A., et al. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome Research*, 20(5):565–577, 2010. doi:10.1101/gr.104471.109. 5
- Gough, J., Karplus, K., Hughey, R., and Chothia, C. Assignment of homology to genome sequences using a library of hidden Markov models that represent

REFERENCES

- all proteins of known structure1. *Journal of molecular biology*, 313(4):903–919, 2001. 6
- Gregory, P. Bayesian logical data analysis for the physical sciences: a comparative approach with Mathematica support. *books.google.com*, 2005. URL http://books.google.com/books?hl=en&lr=&id=yJ_5VFo0zGMC&oi=fnd&pg=PR13&dq=+Sciences.&ots=V6MKQvT1Ds&sig=g5r6eIqk_m4J6fiHZRF07AgaNqk. 28
- Griffith, O., Montgomery, S., and Bernier, B. ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic acids ...*, 2008. URL http://nar.oxfordjournals.org/cgi/content/abstract/36/suppl_1/D107. 33
- Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B.J., et al. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*, 455(7217):1193–1197, 2008. doi:10.1038/nature07415. 7
- Grosschedl, R., Giese, K., and Pagel, J. HMG domain proteins: architectural elements in the assembly of nucleoprotein structures. *Trends Genet*, 10(3):94–100, 1994. 167
- Grove, C.A., Masi, F.D., Barrasa, M.I., Newburger, D.E., et al. A multiparameter network reveals extensive divergence between *C. elegans* bHLH transcription factors. *Cell*, 138(2):314–27, 2009. doi:10.1016/j.cell.2009.04.058. URL <http://www.cell.com/retrieve/pii/S0092867409005194>. 12, 35
- Gunewardena, S. and Zhang, Z. A hybrid model for robust detection of transcription factor binding sites. *Bioinformatics*, 24(4):484–91, 2008. doi:10.1093/bioinformatics/btm629. URL <http://bioinformatics.oxfordjournals.org/cgi/content/full/24/4/484>. 108
- Habib, N., Kaplan, T., Margalit, H., Friedman, N., et al. A Novel Bayesian DNA Motif Comparison Method for Clustering and Retrieval. *PLoS Computational Biology*, 4(2):e1000010, 2008. doi:10.1371/journal.pcbi.1000010. 41

REFERENCES

- Harbison, C.T., Gordon, D.B., Lee, T.I., Rinaldi, N.J., et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 431(7004):99–104, 2004. doi: 10.1038/nature02800. URL <http://dx.doi.org/10.1038/nature02800>. 11, 100, 118, 119, 120, 143, 155, 172
- Hardison, R.C. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet*, 16(9):369–72, 2000. 5, 111
- Harrison, S.C. A structural taxonomy of DNA-binding domains. *Nature*, 353(6346):715–9, 1991. doi:10.1038/353715a0. 34
- Hastings, W. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970. 177
- Hecht, A. and Grunstein, M. Mapping DNA interaction sites of chromosomal proteins using immunoprecipitation and polymerase chain reaction. *Methods in Enzymology*, 304:399–414, 1999. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B7CV2-4B5PGMJ-48&_user=10&_coverDate=12arch&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=b77cb9d67f8cc5fbc57550afaac012d8&searchtype=a. 33
- Helden, J.V., André, B., and Collado-Vides, J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol*, 281(5):827–42, 1998. doi:10.1006/jmbi.1998.1947. 13
- Hertz, G.Z. and Stormo, G.D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15(7-8):563–77, 1999. URL <http://bioinformatics.oxfordjournals.org/cgi/reprint/15/7/563?view=long&pmid=10487864>. 113
- Ho, T. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis . . .*, 1998. URL <http://machine-learning.martinsewell.com/ensembles/rsm/Ho1998.pdf>. 29

REFERENCES

- Holland, R.C.G., Down, T.A., Pocock, M., Prlić, A., et al. BioJava: an open-source framework for bioinformatics. *Bioinformatics*, 24(18):2096–7, 2008. doi:10.1093/bioinformatics/btn397. 109
- Hu, J., Yang, Y., and Kihara, D. EMD: an ensemble algorithm for discovering regulatory motifs in DNA sequences. *BMC bioinformatics*, 7(1):342, 2006. 108
- Hu, Z., Killion, P.J., and Iyer, V.R. Genetic reconstruction of a functional transcriptional regulatory network. *Nat Genet*, 39(5):683–7, 2007. doi:10.1038/ng2012. URL <http://www.nature.com/ng/journal/v39/n5/abs/ng2012.html>. 100, 118, 120
- Hubbard, T.J.P., Aken, B.L., Ayling, S., Ballester, B., et al. Ensembl 2009. *Nucleic Acids Research*, 37(Database):D690–D697, 2009. doi:10.1093/nar/gkn828. 11, 31, 109, 121, 185
- Huber, W., von Heydebreck, A., Sülthmann, H., Poustka, A., et al. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1:S96–104, 2002. URL http://bioinformatics.oxfordjournals.org/cgi/reprint/18/suppl_1/S96?view=long&pmid=12169536. 120
- Huerta, A.M., Salgado, H., Thieffry, D., and Collado-Vides, J. RegulonDB: a database on transcriptional regulation in Escherichia coli. *Nucleic Acids Res*, 26(1):55–59, 1998. 6
- Ishii, T., Yoshida, K., Terai, G., Fujita, Y., et al. DBTBS: a database of Bacillus subtilis promoters and transcription factors. *Nucleic Acids Research*, 29(1):278–80, 2001. URL <http://nar.oxfordjournals.org/cgi/content/full/29/1/278?view=long&pmid=11125112>. 6
- Iyer, V.R., Horak, C.E., Scafe, C.S., Botstein, D., et al. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, 409(6819):533–8, 2001. doi:10.1038/35054095. URL <http://www.nature.com/nature/journal/v409/n6819/full/409533a0.html>. 118, 120

REFERENCES

- Jacquier, A. Applications of next-generation sequencing: The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nature Reviews Genetics*, 10(12):833–844, 2009. doi:10.1038/nrg2683. URL <http://dx.doi.org/10.1038/nrg2683>. 8
- Jaenisch, R. and Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet*, 33 Suppl:245–54, 2003. doi:10.1038/ng1089. 8
- Janga, S., Collado-Vides, J., and Babu, M. Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes. *Proceedings of the National Academy of Sciences*, 105(41):15761, 2008. 3
- Johnson, S. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967. XIII, 84, 116, 154
- Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., et al. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res*, pp. 1–14, 2010. doi:10.1101/gr.100552.109. 12
- Jones and Pevzner. *Bioinformatics*, 2006. 101
- Jones, S., van Heyningen, P., Berman, H.M., and Thornton, J.M. Protein-DNA interactions: A structural analysis. *J Mol Biol*, 287(5):877–96, 1999. doi:10.1006/jmbi.1999.2659. 38
- Kafri, R., Bar-Even, A., and Pilpel, Y. Transcription control reprogramming in genetic backup circuits. *Nat Genet*, 37(3):295–9, 2005. doi:10.1038/ng1523. 90
- Kanamori, M., Konno, H., Osato, N., Kawai, J., et al. A genome-wide and nonredundant mouse transcription factor database. *Biochem Biophys Res Commun*, 322(3):787–93, 2004. doi:10.1016/j.bbrc.2004.07.179. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WBK-4D4PPRP-J&_user=10&_coverDate=09rch&_sort=d&_docanchor=&view=c&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=247cd1528a22107a65007eaa408a4b93. 6

REFERENCES

- Kaplan, T., Friedman, N., and Margalit, H. Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput Biol*, 1(1):e1, 2005. doi:10.1371/journal.pcbi.0010001. URL <http://dx.doi.org/10.1371/journal.pcbi.0010001>. 39
- Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., et al. Variation in Transcription Factor Binding Among Humans. *Science*, 328(5975):232–235, 2010. doi:10.1126/science.1183621. 112
- Kechris, K. and Li, H. c-REDUCE: incorporating sequence conservation to detect motifs that correlate with expression. *BMC bioinformatics*, 9:506, 2008. doi:10.1186/1471-2105-9-506. URL <http://www.biomedcentral.com/1471-2105/9/506>. 15
- Keleş, S., van der Laan, M., and Eisen, M.B. Identification of regulatory elements using a feature selection method. *Bioinformatics*, 18(9):1167–75, 2002. URL <http://bioinformatics.oxfordjournals.org/cgi/reprint/18/9/1167>. 15
- Kellis, M., Patterson, N., Endrizzi, M., Birren, B., et al. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937):241–54, 2003. doi:10.1038/nature01644. 101, 111
- Kielbasa, S.M., Gonze, D., and Herzel, H. Measuring similarities between transcription factor binding sites. *BMC Bioinformatics*, 6:237, 2005. doi:10.1186/1471-2105-6-237. URL <http://dx.doi.org/10.1186/1471-2105-6-237>. 41
- Kim. Crystal structure of a yeast TBP/TATA-box complex. *Nature*, 1995. 39
- Kim, J.L. and Burley, S.K. 1.9 Å resolution refined structure of TBP recognizing the minor groove of TATAAAAG. *Nat Struct Biol*, 1(9):638–53, 1994. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=7634103&dopt=abstractplus. 165
- Kim, J., Cunningham, R., James, B., Wyder, S., et al. Functional characterization of transcription factor motifs using cross-species comparison across

REFERENCES

- large evolutionary distances. *PLoS Comput Biol*, 6(1):e1000652, 2010. doi:10.1371/journal.pcbi.1000652. 5
- Kirkpatrick, C.R. and Schimmel, P. Detection of leucine-independent DNA site occupancy of the yeast Leu3p transcriptional activator in vivo. *Mol Cell Biol*, 15(8):4021–30, 1995. URL <http://mcb.asm.org/cgi/reprint/15/8/4021?view=long&pmid=7623798>. 144
- Klepper, K., Sandve, G., Abul, O., Johansen, J., et al. Assessment of composite motif discovery methods. *BMC Bioinformatics*, 9(1):123, 2008. doi:10.1186/1471-2105-9-123. URL <http://www.biomedcentral.com/1471-2105/9/123>. 108
- Kohonen, T. and Somervuo, P. How to make large self-organizing maps for nonvectorial data. *Neural Netw*, 15(8-9):945–52, 2002. 41
- Kono, H. and Sarai, A. Structure-based prediction of DNA target sites by regulatory proteins. *Proteins*, 35(1):114–31, 1999. 38
- Korn, L.J., Queen, C.L., and Wegman, M.N. Computer analysis of nucleic acid regulatory sequences. *Proc Natl Acad Sci USA*, 74(10):4401–5, 1977. 13
- Kroeger, P.E. and Morimoto, R.I. Selection of new HSF1 and HSF2 DNA-binding sites reveals difference in trimer cooperativity. *Mol Cell Biol*, 14(11):7592–603, 1994. URL <http://mcb.asm.org/cgi/reprint/14/11/7592?view=long&pmid=7935474>. 50
- Krogh, A., Brown, M., Mian, I.S., Sjölander, K., et al. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol*, 235(5):1501–31, 1994. doi:10.1006/jmbi.1994.1104. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WK7-45NSKPC-N4&_user=10&_coverDate=02arch&_sort=d&_docanchor=&view=c&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=dfe80e989968625b45adca07c6681ed4. 19, 68
- Kulkarni, M.M. and Arnosti, D.N. cis-regulatory logic of short-range transcriptional repression in *Drosophila melanogaster*. *Molecular and Cellular Biology*,

REFERENCES

- 25(9):3411–20, 2005. doi:10.1128/MCB.25.9.3411-3420.2005. URL <http://mcb.asm.org/cgi/content/full/25/9/3411?view=long&pmid=15831448>. 5
- Kummerfeld, S.K. and Teichmann, S.A. DBD: a transcription factor prediction database. *Nucleic Acids Res*, 34(Database issue):D74–D81, 2006. doi:10.1093/nar/gkj131. URL <http://dx.doi.org/10.1093/nar/gkj131>. 6
- Kunarso, G., Chia, N.Y., Jeyakani, J., Hwang, C., et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet*, 42(7):631–4, 2010. doi:10.1038/ng.600. URL <http://www.nature.com/ng/journal/v42/n7/full/ng.600.html>. 112
- Kwong, C., Adryan, B., Bell, I., Meadows, L., et al. Stability and dynamics of polycomb target sites in *Drosophila* development. *PLoS Genetics*, 4(9), 2008. 67
- Lagrange, T., Kapanidis, A.N., Tang, H., Reinberg, D., et al. New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev*, 12(1):34–44, 1998. 123
- Lähdesmäki, H., Rust, A.G., and Shmulevich, I. Probabilistic inference of transcription factor binding from multiple data sources. *PLoS ONE*, 3(3):e1820, 2008. doi:10.1371/journal.pone.0001820. URL <http://www.plosone.org/article/infojournal.pone.0001820>. 107
- Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., et al. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262(5131):208–14, 1993. URL <http://www.sciencemag.org/cgi/reprint/262/5131/208>. 27
- Lebrecht, D., Foehr, M., Smith, E., Lopes, F.J.P., et al. Bicoid cooperative DNA binding is critical for embryonic patterning in *Drosophila*. *Proc Natl Acad Sci USA*, 102(37):13176–81, 2005. doi:10.1073/pnas.0506462102. URL <http://www.pnas.org/content/102/37/13176.long>. 5
- LeClerc, S., Palaniswami, R., Xie, B.X., and Govindan, M.V. Molecular cloning and characterization of a factor that binds the human glucocorticoid

REFERENCES

- receptor gene and represses its expression. *J Biol Chem*, 266(26):17333–17340, 1991. URL <http://www.google.com/search?client=safari&rls=en-us&q=Molecular+cloning+and+characterization+of+a+factor+that+binds+the+human+glucocorticoid+receptor+gene+and+represses+its+expression.&ie=UTF-8&oe=UTF-8>. 14
- Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., et al. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298(5594):799–804, 2002. doi:10.1126/science.1075090. URL <http://dx.doi.org/10.1126/science.1075090>. 100, 112
- Lefrançois, P., Euskirchen, G.M., Auerbach, R.K., Rozowsky, J., et al. Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing. *BMC Genomics*, 10:37, 2009. doi:10.1186/1471-2164-10-37. URL <http://www.biomedcentral.com/1471-2164/10/37>. 11
- Legras, J., Merdinoglu, D., and Cornuet, J. Bread, beer and wine: *Saccharomyces cerevisiae* diversity reflects human history. *Molecular ...*, 2007. URL <http://www3.interscience.wiley.com/journal/117989623/abstract>. 121
- Levine, M. and Davidson, E.H. Gene regulatory networks for development. *Proc Natl Acad Sci U S A*, 102(14):4936–4942, 2005. doi:10.1073/pnas.0408031102. URL <http://dx.doi.org/10.1073/pnas.0408031102>. 176
- Lewis, M.A., Quint, E., Glazier, A.M., Fuchs, H., et al. An ENU-induced mutation of miR-96 associated with progressive hearing loss in mice. *Nat Genet*, 41(5):614–8, 2009. doi:10.1038/ng.369. URL <http://www.nature.com/ng/journal/v41/n5/abs/ng.369.html>. 32, 67, 99
- Li, B., Carey, M., and Workman, J. The role of chromatin during transcription. *Cell*, 128(4):707–719, 2007. 2
- Li, N. and Tompa, M. Analysis of computational approaches for motif discovery. *Algorithms Mol Biol*, 1:8, 2006. doi:10.1186/1748-7188-1-8. URL <http://dx.doi.org/10.1186/1748-7188-1-8>. 16, 108

REFERENCES

- Li, X., Zhong, S., and Wong, W.H. Reliable prediction of transcription factor binding sites by phylogenetic verification. *Proc Natl Acad Sci USA*, 102(47):16945–50, 2005. doi:10.1073/pnas.0504201102. 112, 165
- Liaw, A. and Wiener, M. Classification and Regression by randomForest. *R News*, 2(3):18–22, 2002. URL <http://CRAN.R-project.org/doc/Rnews/>. 31, 84
- Lieb, J.D., Liu, X., Botstein, D., and Brown, P.O. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat Genet*, 28(4):327–34, 2001. doi:10.1038/ng569. URL <http://www.nature.com/doifinder/10.1038/ng569>. 118, 120
- Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–93, 2009. doi:10.1126/science.1181369. URL <http://www.sciencemag.org/cgi/pmidlookup?view=short&pmid=19815776>. 3
- Liti, G., Carter, D.M., Moses, A.M., Warringer, J., et al. Population genomics of domestic and wild yeasts. *Nature*, 458(7236):337–41, 2009. doi:10.1038/nature07743. 121
- Liu, X., Brutlag, D.L., and Liu, J.S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, pp. 127–38, 2001. URL <http://www.ncbi.nlm.nih.gov/pubmed/11262934?dopt=abstract>. 113
- Liu, X., Noll, D., Lieb, J., and Clarke, N. DIP-chip: rapid and accurate determination of DNA-binding specificity. *Genome Res*, 15(3):421, 2005. 12, 114
- Liu, X., Brutlag, D., and Liu, J. An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature biotechnology*, 20(8):835–839, 2002. 15, 102, 113

REFERENCES

- Loh, Y., Wu, Q., Chew, J., Vega, V., et al. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet*, 38(4):431–440, 2006. [67](#)
- Loots, G., Locksley, R., Blankespoor, C.M., Wang, Z., et al. Identification of a coordinate regulator of interleukins 4, 13 and 5 by cross-species sequence comparisons. *Science*, 2000. URL <http://www.google.com/search?client=safari&rls=en-us&q=Identification+of+a+coordinate+regulator+of+interleukins+4,+13+and+5+by+cross-species+sequence+comparisons.&ie=UTF-8&oe=UTF-8>. [5](#)
- Lu, C.C., Yuan, W.H., and Chen, T.M. Extracting transcription factor binding sites from unaligned gene sequences with statistical models. *BMC Bioinformatics*, 9(Suppl 12):S7, 2008. doi:10.1186/1471-2105-9-S12-S7. [108](#)
- Ludwig, M. Functional evolution of noncoding DNA. *Current opinion in genetics & development*, 12(6):634–639, 2002. [5](#)
- Luscombe, N.M., Austin, S.E., Berman, H.M., and Thornton, J.M. An overview of the structures of protein-DNA complexes. *Genome Biol*, 1(1):REVIEWS001, 2000. [34](#), [123](#)
- Luscombe, N.M., Laskowski, R.A., and Thornton, J.M. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res*, 29(13):2860–2874, 2001. [40](#)
- Luscombe, N.M. and Thornton, J.M. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *Journal of molecular biology*, 320(5):991–1009, 2002. [38](#), [39](#), [88](#)
- Lusk, R.W. and Eisen, M.B. Evolutionary mirages: selection on binding site composition creates the illusion of conserved grammars in Drosophila enhancers. *PLoS Genet*, 6(1):e1000829, 2010. doi:10.1371/journal.pgen.1000829. [6](#)
- MacIsaac, K., Lo, K., Gordon, W., Motola, S., et al. A Quantitative Model of Transcriptional Regulation Reveals the Influence of Binding Location on Expression. 2010. [123](#)

REFERENCES

- MacIsaac, K.D. and Fraenkel, E. Practical strategies for discovering regulatory DNA sequence motifs. *PLoS Computational Biology*, 2(4):e36, 2006. doi:10.1371/journal.pcbi.0020036. 13
- MacIsaac, K.D., Wang, T., Gordon, D.B., Gifford, D.K., et al. An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics*, 7:113, 2006. doi:10.1186/1471-2105-7-113. URL <http://dx.doi.org/10.1186/1471-2105-7-113>. 11, 83, 115, 120
- Macpherson, S., Larochele, M., and Turcotte, B. A Fungal Family of Transcriptional Regulators: the Zinc Cluster Proteins. *Microbiology and Molecular Biology Reviews*, 70(3):583–604, 2006. doi:10.1128/MMBR.00015-06. 35, 124
- Maerkl, S.J. and Quake, S.R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315(5809):233–237, 2007a. doi:10.1126/science.1131007. URL <http://dx.doi.org/10.1126/science.1131007>. 12
- Maerkl, S.J. and Quake, S.R. A systems approach to measuring the binding energy landscapes of transcription factors. *Science*, 315(5809):233–7, 2007b. doi:10.1126/science.1131007. 18
- Maerkl, S.J. and Quake, S.R. Experimental determination of the evolvability of a transcription factor. *Proc Natl Acad Sci USA*, 106(44):18650–5, 2009. doi:10.1073/pnas.0907688106. 12, 90
- Mahony, S., Auron, P.E., and Benos, P.V. DNA Familial Binding Profiles Made Easy: Comparison of Various Motif Alignment and Clustering Strategies. *PLoS Comput Biol*, 3(3):e61, 2007. doi:10.1371/journal.pcbi.0030061. URL <http://dx.doi.org/10.1371/journal.pcbi.0030061>. 41
- Mahony, S. and Benos, P.V. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res*, 35(Web Server issue):W253–W258, 2007. doi:10.1093/nar/gkm272. URL <http://dx.doi.org/10.1093/nar/gkm272>. 58, 183

REFERENCES

- Mahony, S., Golden, A., Smith, T.J., and Benos, P.V. Improved detection of DNA motifs using a self-organized clustering of familial binding profiles. *Bioinformatics*, 21 Suppl 1:i283–i291, 2005a. doi:10.1093/bioinformatics/bti1025. URL <http://dx.doi.org/10.1093/bioinformatics/bti1025>. 41, 68, 77
- Mahony, S., Hendrix, D., Golden, A., Smith, T.J., et al. Transcription factor binding site identification using the self-organizing map. *Bioinformatics*, 21(9):1807–14, 2005b. doi:10.1093/bioinformatics/bti256. 77, 111, 200
- Makeev, V.J., Lifanov, A.P., Nazina, A.G., and Papatsenko, D.A. Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. *Nucleic Acids Res*, 31(20):6016–26, 2003. URL <http://nar.oxfordjournals.org/cgi/content/full/31/20/6016?view=long&pmid=14530449>. 8
- Mandel-Gutfreund, Y., Baron, A., and Margalit, H. A structure-based approach for prediction of protein binding sites in gene upstream regions. *Pac Symp Biocomput*, pp. 139–50, 2001. 39
- Markstein, M. and Levine, M. Decoding cis-regulatory DNAs in the Drosophila genome. *Current opinion in genetics & development*, 12(5):601–606, 2002. 5
- Marschall, T. and Rahmann, S. Efficient exact motif discovery. *Bioinformatics*, 25(12):i356–64, 2009. doi:10.1093/bioinformatics/btp188. URL <http://bioinformatics.oxfordjournals.org/cgi/content/full/25/12/i356>. 13
- Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34(Database issue):D108–D110, 2006. doi:10.1093/nar/gkj143. URL <http://dx.doi.org/10.1093/nar/gkj143>. X, XI, 33, 50, 51, 63, 64, 84, 87, 104, 183, 194
- McDaniell, R., Lee, B.K., Song, L., Liu, Z., et al. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science*, 328(5975):235–9, 2010. doi:10.1126/science.1184655. URL <http://www.sciencemag.org/cgi/content/abstract/328/5975/235>. 112

REFERENCES

- Megraw, M., Pereira, F., Jensen, S.T., Ohler, U., et al. A transcription factor affinity-based code for mammalian transcription initiation. *Genome Res*, 19(4):644–56, 2009. doi:10.1101/gr.085449.108. URL <http://genome.cshlp.org/content/19/4/644.long>. 1
- Meng, X., Brodsky, M.H., and Wolfe, S.A. A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nat Biotechnol*, 23(8):988–994, 2005. doi:10.1038/nbt1120. 183
- Meng, X. and Wolfe, S.A. Identifying DNA sequences recognized by a transcription factor using a bacterial one-hybrid system. *Nat Methods*, 1(1):30–45, 2006. doi:10.1038/nprot.2006.6. 12
- Mewes, H., Albermann, K., Bähr, M., Frishman, D., et al. Overview of the yeast genome. *Nature*, 387(6632):7–8, 1997. 121
- Meyer, D., Leisch, F., and Hornik, K. Benchmarking. 2003. URL <http://citeseer.ist.psu.edu/619009>. 30
- Minichiello, M.J. and Durbin, R. Mapping trait loci by use of inferred ancestral recombination graphs. *Am J Hum Genet*, 79(5):910–22, 2006. doi:10.1086/508901. URL <http://www.ncbi.nlm.nih.gov/pubmed/17033967?dopt=abstract>. 122
- Minka, T.P. Estimating a Dirichlet distribution, 2003. URL <http://research.microsoft.com/users/Cambridge/minka/papers/dirichlet/minka-dirichlet.ps>. 48, 49, 51, 52, 58, 85
- Mintseris, J. and Eisen, M.B. Design of a combinatorial DNA microarray for protein-DNA interaction studies. *BMC bioinformatics*, 7:429, 2006. doi:10.1186/1471-2105-7-429. URL <http://www.biomedcentral.com/1471-2105/7/429>. 114
- Mitchell, P.J. and Tjian, R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*, 245(4916):371–8, 1989. 1

REFERENCES

- Montgomery, S.B., Astakhova, T., Bilenky, M., Birney, E., et al. Sockeye: a 3D environment for comparative genomics. *Genome Res*, 14(5):956–62, 2004. doi:10.1101/gr.1890304. URL <http://genome.cshlp.org/content/14/5/956.long>. 183
- Morley, R.H., Lachani, K., Keefe, D., Gilchrist, M.J., et al. A gene regulatory network directed by zebrafish No tail accounts for its roles in mesoderm formation. *Proc Natl Acad Sci USA*, 106(10):3829–34, 2009. doi:10.1073/pnas.0808382106. 67
- Moses, A., Chiang, D., and Kellis, M. Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evolutionary ...*, 2003. URL <http://www.biomedcentral.com/1471-2148/3/19>. 14
- Mukherjee, P. and Parkinson, D. A nested sampling algorithm for cosmological model selection. *The Astrophysical Journal ...*, 2006. URL <http://iopscience.iop.org/1538-4357/638/2/L51>. 25
- Mukherjee, S., Berger, M., Jona, G., and Wang, X. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat Genet*, 2004. URL <http://www.nature.com/ng/journal/vaop/ncurrent/full/ng1473.html>. 12
- Mustonen, V., Kinney, J., Callan, C.G., and Lässig, M. Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. *Proc Natl Acad Sci USA*, 105(34):12376–81, 2008. doi:10.1073/pnas.0805909105. 18
- Nadassy, K., Wodak, S.J., and Janin, J. Structural features of protein-nucleic acid recognition sites. *Biochemistry*, 38(7):1999–2017, 1999. doi:10.1021/bi982362d. 38
- Narlikar, L., Gordân, R., and Hartemink, A.J. A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Comput Biol*, 3(11):e215, 2007. doi:10.1371/journal.pcbi.0030215. URL <http://dx.doi.org/10.1371/journal.pcbi.0030215>. 15

REFERENCES

- Narlikar, L., Gordân, R., Ohler, U., and Hartemink, A.J. Informative priors based on transcription factor structural class improve de novo motif discovery. *Bioinformatics*, 22(14):e384–e392, 2006. doi:10.1093/bioinformatics/btl251. URL <http://dx.doi.org/10.1093/bioinformatics/btl251>. 15, 39, 42, 67, 69, 82, 111, 113
- Narlikar, L. and Hartemink, A.J. Sequence features of DNA binding sites reveal structural class of associated transcription factor. *Bioinformatics*, 22(2):157–163, 2006. doi:10.1093/bioinformatics/bti731. URL <http://dx.doi.org/10.1093/bioinformatics/bti731>. 34, 40, 41, 82, 83, 84, 86, 88, 91
- Needleman, S. and Wunsch, C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970. 40
- Newburger, D.E. and Bulyk, M.L. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, 37(Database):D77–D82, 2009. doi:10.1093/nar/gkn660. 50, 183
- Nguyen, T.T. and Androulakis, I.P. Recent Advances in the Computational Discovery of Transcription Factor Binding Sites. *Algorithms*, 2(1):582–605, 2009. doi:10.3390/a2010582. URL <http://www.mdpi.com/1999-4893/2/1/582>. 13
- Nielsen, R. Statistical methods in molecular evolution. p. 504, 2005. URL <http://books.google.com/books?id=nJipT3toWFAC&printsec=frontcover>. 121
- Nieto, M.A. The snail superfamily of zinc-finger transcription factors. *Nat Rev Mol Cell Biol*, 3(3):155–66, 2002. doi:10.1038/nrm757. 90
- Notredame, C., Higgins, D.G., and Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology*, 302(1):205–17, 2000. doi:10.1006/jmbi.2000.4042. 48
- Noyes, M.B., Christensen, R.G., Wakabayashi, A., Stormo, G.D., et al. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell*, 133(7):1277–89, 2008a. doi:10.1016/j.

REFERENCES

- cell.2008.05.023. URL [http://linkinghub.elsevier.com/retrieve/pii/S0092-8674\(08\)00682-X](http://linkinghub.elsevier.com/retrieve/pii/S0092-8674(08)00682-X). XI, 12, 36, 91, 92, 93
- Noyes, M.B., Meng, X., Wakabayashi, A., Sinha, S., et al. A systematic characterization of factors that regulate *Drosophila* segmentation via a bacterial one-hybrid system. *Nucleic Acids Research*, 36(8):2547–60, 2008b. doi:10.1093/nar/gkn048. URL <http://nar.oxfordjournals.org/cgi/content/full/36/8/2547>. 12
- Oliphant, A.R., Brandl, C.J., and Struhl, K. Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. *Molecular and Cellular Biology*, 9(7):2944–9, 1989. URL <http://mcb.asm.org/cgi/reprint/9/7/2944?view=long&pmid=2674675>. 33
- Orengo, C.A. and Taylor, W.R. SSAP: sequential structure alignment program for protein structure comparison. *Meth Enzymol*, 266:617–35, 1996. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=8743709&dopt=abstractplus. 34, 123
- Osada, R., Zaslavsky, E., and Singh, M. Comparative analysis of methods for representing and searching for transcription factor binding sites. *Bioinformatics*, 20(18):3516–25, 2004. doi:10.1093/bioinformatics/bth438. URL <http://bioinformatics.oxfordjournals.org/cgi/reprint/20/18/3516>. 17, 18
- Pabo, C.O., Peisach, E., and Grant, R.A. Design and selection of novel Cys2His2 zinc finger proteins. *Annu Rev Biochem*, 70:313–40, 2001. doi:10.1146/annurev.biochem.70.1.313. 39
- Pachkov, M., Erb, I., Molina, N., and van Nimwegen, E. SwissRegulon: a database of genome-wide annotations of regulatory sites. *Nucleic Acids Res*, 35(Database issue):D127–D131, 2007. doi:7. URL <http://dx.doi.org/7>. 114
- Papatsenko. Organization of developmental enhancers in the *Drosophila* embryo. *Nucleic Acids Res*, 2009. URL <http://www.google.com/search?client=safari&rls=en-us&q=Organization+of+developmental+enhancers+in+the+Drosophila+embryo.&ie=UTF-8&oe=UTF-8>. 8

REFERENCES

- Pape, U.J., Rahmann, S., and Vingron, M. Natural Similarity Measures between Position Frequency Matrices with an Application to Clustering. *Bioinformatics*, 2008. doi:10.1093/bioinformatics/btm610. URL <http://dx.doi.org/10.1093/bioinformatics/btm610>. 41
- Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G., et al. ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic acids research*, 37(Database issue):D868–72, 2009. doi:10.1093/nar/gkn889. URL http://nar.oxfordjournals.org/cgi/content/full/37/suppl_1/D868?view=long&pmid=19015125. 120
- Pavesi, G., Mauri, G., and Pesole, G. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics*, 17 Suppl 1:S207–14, 2001. URL http://bioinformatics.oxfordjournals.org/cgi/reprint/17/suppl_1/S207?view=long&pmid=11473011. 13, 197
- Peng, C., Hsu, J., Chung, Y., Lin, Y., et al. Identification of degenerate motifs using position restricted selection and hybrid ranking combination. *Nucleic acids research*, 2006. 108
- Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., et al. In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, 444(7118):499–502, 2006. doi:10.1038/nature05295. URL <http://www.nature.com/nature/journal/v444/n7118/abs/nature05295.html>. 2
- Percipalle, P., Simoncsits, A., Zakhariyev, S., Guarnaccia, C., et al. Rationally designed helix-turn-helix proteins and their conformational changes upon DNA binding. *EMBO J*, 14(13):3200–5, 1995. 39
- Persikov, A.V., Osada, R., and Singh, M. Predicting DNA recognition by Cys2His2 zinc finger proteins. *Bioinformatics*, 25(1):22–29, 2008. doi:10.1093/bioinformatics/btn580. 39
- Pevzner, P.A. and Sze, S.H. Combinatorial approaches to finding subtle signals in DNA sequences. *Proc Int Conf Intell Syst Mol Biol*, 8:269–78, 2000. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=10977088&dopt=abstractplus. 16, 103, 105, 174

REFERENCES

- Pfreundt, U., James, D.P., Tweedie, S., Wilson, D., et al. FlyTF: improved annotation and enhanced functionality of the Drosophila transcription factor database. *Nucleic Acids Res*, 38(Database issue):D443–7, 2010. doi:10.1093/nar/gkp910. URL http://nar.oxfordjournals.org/cgi/content/full/38/suppl_1/D443?view=long&pmid=19884132. 6
- Pic, A., Lim, F.L., Ross, S.J., Veal, E.A., et al. The forkhead protein Fkh2 is a component of the yeast cell cycle transcription factor SFF. *The EMBO Journal*, 19(14):3750–61, 2000. doi:10.1093/emboj/19.14.3750. URL <http://www.nature.com/emboj/journal/v19/n14/abs/7593192a.html>. 167
- Pietrokovski, S. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Research*, 24(19):3836, 1996. 39
- Piipari, M., Down, T., and Hubbard, T. Large-scale gene regulatory motif discovery with NestedMICA. *Advances in Genomic Sequence Analysis and pattern Discovery*, 2011. 32, 188
- Piipari, M., Down, T.A., and Hubbard, T.J. Metamotifs - a generative model for building families of nucleotide position weight matrices. *BMC Bioinformatics*, 11(348):1–24, 2010a. doi:10.1186/1471-2105-11-348. 25, 32, 38, 173, 174
- Piipari, M., Down, T.A., Saini, H., Enright, A., et al. iMotifs: an integrated sequence motif visualization and analysis environment. *Bioinformatics*, 26(6):843–4, 2010b. doi:10.1093/bioinformatics/btq026. URL <http://bioinformatics.oxfordjournals.org/cgi/content/full/26/6/843?view=long&pmid=20106815>. 32, 46, 48, 74, 77, 78, 109, 111, 166, 174, 182
- Pompeani, A.J., Irgon, J.J., Berger, M.F., Bulyk, M.L., et al. The *Vibrio harveyi* master quorum-sensing regulator, LuxR, a TetR-type protein is both an activator and a repressor: DNA recognition and binding specificity at target promoters. *Mol Microbiol*, 70(1):76–88, 2008. doi:10.1111/j.1365-2958.2008.06389.x. URL <http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2958.2008.06389.x/abstract>. 35

REFERENCES

- Portales-Casamar, E., Thongjuea, S., Kwon, A.T., Arenillas, D., et al. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 38(Database):D105–D110, 2010. doi:10.1093/nar/gkp950. [12](#), [34](#), [40](#), [50](#), [114](#), [115](#)
- Qin, Z., McCue, L., Thompson, W., Mayerhofer, L., et al. Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nat Biotechnol*, 21(4):435–439, 2003. [77](#)
- Rabiner, L. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE DOI - 10.1109/5.18626*, 77(2):257–286, 1989. URL [10.1109/5.18626](#). [19](#)
- Rahmann, S., Müller, T., and Vingron, M. On the power of profiles for transcription factor binding site detection. *Stat Appl Genet Mol Biol*, 2:Article7, 2003. doi:10.2202/1544-6115.1032. URL [http://dx.doi.org/10.2202/1544-6115.1032](#). [17](#)
- Ramsey, S.A., Knijnenburg, T.A., Kennedy, K.A., Zak, D.E., et al. Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. *Bioinformatics*, 26(17):2071–2075, 2010. doi:10.1093/bioinformatics/btq405. [107](#)
- Rastegar, S., Hess, I., Dickmeis, T., Nicod, J.C., et al. The words of the regulatory code are arranged in a variable manner in highly conserved enhancers. *Dev Biol*, 318(2):366–77, 2008. doi:10.1016/j.ydbio.2008.03.034. [67](#)
- Ravasi, T., Suzuki, H., Cannistraci, C.V., Katayama, S., et al. An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man. *Cell*, 140(5):744–752, 2010. doi:10.1016/j.cell.2010.01.044. [8](#), [176](#)
- Reddy, T.E., Delisi, C., and Shakhnovich, B.E. Binding Site Graphs: A New Graph Theoretical Framework for Prediction of Transcription Factor Binding Sites. *PLoS Comput Biol*, 3(5):e90, 2007. doi:10.1371/journal.pcbi.0030090. URL [http://dx.doi.org/10.1371/journal.pcbi.0030090](#). [108](#)

REFERENCES

- Reese, M.G., Hartzell, G., Harris, N.L., Ohler, U., et al. Genome annotation assessment in *Drosophila melanogaster*. *Genome Res*, 10(4):483–501, 2000. URL <http://genome.cshlp.org/content/10/4/483.long>. 106
- Reimand, J., Vaquerizas, J.M., Todd, A.E., Vilo, J., et al. Comprehensive re-analysis of transcription factor knockout expression data in *Saccharomyces cerevisiae* reveals many new targets. *Nucleic Acids Research*, pp. 1–10, 2010. doi:10.1093/nar/gkq232. 100, 118, 120, 143, 144, 145
- Reiner, A., Yekutieli, D., and Benjamini, Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19(3):368–75, 2003. URL <http://bioinformatics.oxfordjournals.org/cgi/reprint/19/3/368?view=long&pmid=12584122>. 120
- Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods*, 4(8):651–7, 2007. doi:10.1038/nmeth1068. 8, 12, 188
- Robinson, M., Sun, Y., Boekhorst, R.T., Kaye, P., et al. Improving computational predictions of cis-regulatory binding sites. *Biocomputing-Proceedings of the Pacific Symposium*, 2006. URL <http://homepages.feis.herts.ac.uk/~nngroup/pubs/papers/Robinson-PSB05.pdf>. 108
- Rolfes, R.J., Zhang, F., and Hinnebusch, A.G. The transcriptional activators BAS1, BAS2, and ABF1 bind positive regulatory sites as the critical elements for adenine regulation of ADE5,7. *J Biol Chem*, 272(20):13343–54, 1997. URL <http://www.jbc.org/content/272/20/13343.long>. 167
- Ross, E.D., Keating, A.M., and 3RD, M.L. DNA constraints on transcription activation in vitro. *J Mol Biol*, 297(2):321–34, 2000. doi:10.1006/jmbi.2000.3562. 123
- Roth, F.P., Hughes, J.D., Estep, P.W., and Church, G.M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature biotechnology*, 16(10):939–45, 1998. doi:10.1038/nbt1098-939. 15, 101, 102, 198

REFERENCES

- Sagot, M. Spelling approximate repeated or common motifs using a suffix tree. *Latin'98: Theoretical Informatics*, 1998. URL <http://www.springerlink.com/index/1469887m40070445.pdf>. 14
- Salgado, H., Gama-Castro, S., Peralta-Gil, M., Díaz-Peredo, E., et al. RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res*, 34(Database issue):D394–D397, 2006. doi:10.1093/nar/gkj156. URL <http://dx.doi.org/10.1093/nar/gkj156>. 6
- Saltzman, A.G. and Weinmann, R. Promoter specificity and modulation of RNA polymerase II transcription. *FASEB J*, 3(6):1723–33, 1989. URL <http://www.fasebj.org/cgi/reprint/3/6/1723>. 1
- Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., et al. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res*, 32(Database issue):D91–D94, 2004. doi:10.1093/nar/gkh012. URL <http://dx.doi.org/10.1093/nar/gkh012>. 34
- Sandelin, A. and Wasserman, W.W. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *J Mol Biol*, 338(2):207–215, 2004. doi:10.1016/j.jmb.2004.02.048. URL <http://dx.doi.org/10.1016/j.jmb.2004.02.048>. 39, 40, 42, 68, 77, 82, 91
- Sandve, G.K., Abul, O., and Drablos, F. Compo: composite motif discovery using discrete models. *BMC Bioinformatics*, 9(1):527, 2008. doi:10.1186/1471-2105-9-527. 108
- Sandve, G.K., Abul, O., Walseng, V., and Drabløs, F. Improved benchmarks for computational motif discovery. *BMC bioinformatics*, 8:193, 2007. doi:10.1186/1471-2105-8-193. 108
- Sandve, G.K. and Drabløs, F. A survey of motif discovery methods in an integrated framework. *Biol Direct*, 1:11, 2006. doi:10.1186/1745-6150-1-11. URL <http://www.biology-direct.com/content/1/1/11>. 13

REFERENCES

- Saxonov, S., Berg, P., and Brutlag, D.L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci USA*, 103(5):1412–7, 2006. doi:10.1073/pnas.0510310103. URL <http://www.pnas.org/cgi/content/full/103/5/1412>. 28
- Scharer, C.D., McCabe, C.D., Ali-Seyed, M., Berger, M.F., et al. Genome-wide promoter analysis of the SOX4 transcriptional network in prostate cancer cells. *Cancer Res*, 69(2):709–17, 2009. doi:10.1158/0008-5472.CAN-08-3415. URL <http://cancerres.aacrjournals.org/content/69/2/709.long>. 35
- Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., et al. Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. *Science*, 2010. doi:10.1126/science.1186176. 5, 112
- Schneider, T.D. and Stephens, R.M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, 18(20):6097–100, 1990. 14, 18, 45, 46
- Sharon, E., Lubliner, S., Segal, E., and Stormo, G. A Feature-Based Approach to Modeling Protein–DNA Interactions. *PLoS Comput Biol*, 4(8):e1000154, 2008. doi:10.1371/journal.pcbi.1000154. 18
- Shaw, J.R., Bridges, M., and Hobson, M.P. Efficient Bayesian inference for multimodal problems in cosmology. *arXiv*, astro-ph, 2007. doi:10.1111/j.1365-2966.2007.11871.x. URL <http://arxiv.org/abs/astro-ph/0701867v2>. 25
- Sherman, D., Durrens, P., Beyne, E., Nikolski, M., et al. Genolevures: comparative genomics and molecular evolution of hemiascomycetous yeasts. *Nucleic Acids Res*, 32(Database Issue):D315, 2004. 109
- Siddharthan, R. PhyloGibbs-MP: module prediction and discriminative motif-finding by Gibbs sampling. *PLoS Computational Biology*, 4(8):e1000156, 2008. doi:10.1371/journal.pcbi.1000156. URL <http://www.ploscompbiol.org/article/info252Fjournal.pcbi.1000156>. 16, 69, 77
- Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*,

REFERENCES

- 15(8):1034–50, 2005. doi:10.1101/gr.3715005. URL <http://genome.cshlp.org/content/15/8/1034.long>. 112, 121
- Sierro, N., Makita, Y., de Hoon, M., and Nakai, K. DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Research*, 36(Database issue):D93–6, 2008. doi:10.1093/nar/gkm910. URL http://nar.oxfordjournals.org/cgi/content/full/36/suppl_1/D93?view=long&pmid=17962296. 6
- Siggia, E.D. Computational methods for transcriptional regulation. *Current opinion in genetics & development*, 15(2):214–21, 2005. doi:10.1016/j.gde.2005.02.004. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VSO-4FK3PDM-2&_user=10&_coverDate=04rch&_sort=d&_docanchor=&view=c&_acct=C000050221&_version=1&_urlVersion=0&_userid=10&md5=c765281eea85c83bfed2db3fa697914e. 5
- Sikder, D. and Kodadek, T. Genomic studies of transcription factor-DNA interactions. *Curr Opin Chem Biol*, 9(1):38–45, 2005. doi:10.1016/j.cbpa.2004.12.008. URL <http://dx.doi.org/10.1016/j.cbpa.2004.12.008>. 11
- Silva, E.K.D., Gehrke, A.R., Olszewski, K., León, I., et al. Specific DNA-binding by apicomplexan AP2 transcription factors. *Proc Natl Acad Sci USA*, 105(24):8393–8, 2008. doi:10.1073/pnas.0801993105. URL <http://www.pnas.org/content/105/24/8393.long>. 35
- Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941, 2005. doi:10.1093/bioinformatics/bti623. URL <http://dx.doi.org/10.1093/bioinformatics/bti623>. 124
- Sinha, S., Blanchette, M., and Tompa, M. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, 5:170, 2004. doi:10.1186/1471-2105-5-170. URL <http://dx.doi.org/10.1186/1471-2105-5-170>. 16

REFERENCES

- Sinha, S. and Tompa, M. Performance Comparison of Algorithms for Finding Transcription Factor Binding Sites. 2003a. URL <http://citeseer.ist.psu.edu/592637>. Fetch Buhler & Tompa. 16
- Sinha, S. and Tompa, M. YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Research*, 31(13):3586–8, 2003b. URL <http://nar.oxfordjournals.org/cgi/content/full/31/13/3586?view=long&pmid=12824371>. 199
- Skilling, J. Nested Sampling for General Bayesian Computation. 2004. 23, 52
- Smith. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. 1987. 26
- Smith, T. Secret code. *Nature Structural Biology*, 5(2):100, 1998. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=9461070&dopt=abstractplus. 38
- Sonnhammer, E., Eddy, S., and Durbin, R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins: Structure, Function, and Bioinformatics*, 28(3):405–420, 1997. 6, 19
- Stabenau, A., McVicker, G., Melsopp, C., Proctor, G., et al. The Ensembl core software libraries. *Genome Res*, 14(5):929–33, 2004. doi:10.1101/gr.1857204. 32
- Stanke, M. and Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, 19(Suppl 2), 2003. 19
- Stark, A., Lin, M.F., Kheradpour, P., Pedersen, J.S., et al. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature*, 450(7167):219–32, 2007. doi:10.1038/nature06340. URL <http://www.nature.com/nature/journal/v450/n7167/full/nature06340.html>. 5
- Starr, D.B. and Hawley, D.K. TFIID binds in the minor groove of the TATA box. *Cell*, 67(6):1231–40, 1991. URL <http://www.cell.com/retrieve/pii/009286749190299E>. 165

REFERENCES

- Steinfeld, I., Shamir, R., and Kupiec, M. A genome-wide analysis in *Saccharomyces cerevisiae* demonstrates the influence of chromatin modifiers on transcription. *Nat Genet*, 39(3):303–309, 2007. doi:10.1038/ng1965. 2
- Stormo, G.D., Schneider, T.D., Gold, L., and Ehrenfeucht, A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. *Nucleic Acids Research*, 10(9):2997–3011, 1982. 39
- Sudarsanam, P., Pilpel, Y., and Church, G.M. Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in *Saccharomyces cerevisiae*. *Genome Res*, 12(11):1723–1731, 2002. doi:10.1101/gr.301202. URL <http://dx.doi.org/10.1101/gr.301202>. 165
- Tamada, Y., Kim, S., Bannai, H., Imoto, S., et al. Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, 19 Suppl 2:ii227–36, 2003. 167
- Tanaka, N., Kawakami, T., and Taniguchi, T. Recognition DNA sequences of interferon regulatory factor 1 (IRF-1) and IRF-2, regulators of cell growth and the interferon system. *Molecular and Cellular Biology*, 13(8):4531–8, 1993. URL <http://mcb.asm.org/cgi/reprint/13/8/4531?view=long&pmid=7687740>. 90
- Tang, M.H.E., Krogh, A., and Winther, O. BayesMD: Flexible Biological Modeling for Motif Discovery. *Journal of Computational Biology*, 15(10):1347–1363, 2008. doi:10.1089/cmb.2008.15.issue-10. 28, 56, 71, 111, 113, 174
- Team, R.D.C. *R: A Language and Environment for Statistical Computing*, 2007. URL <http://www.R-project.org>. ISBN 3-900051-07-0. 84
- Teixeira, M.C., Monteiro, P., Jain, P., Tenreiro, S., et al. The YEASTRACT database: a tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*. *Nucleic Acids Res*, 34(Database issue):D446–51, 2006. doi:10.1093/nar/gkj013. URL http://nar.oxfordjournals.org/cgi/content/full/34/suppl_1/D446?view=long&pmid=16381908. 100, 118, 119, 143

REFERENCES

- Thijs, G., Lescot, M., Marchal, K., Rombauts, S., et al. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, 17(12):1113–22, 2001. URL <http://bioinformatics.oxfordjournals.org/cgi/reprint/17/12/1113?view=long&pmid=11751219>. 117, 198
- Thijs, G., Marchal, K., Lescot, M., Rombauts, S., et al. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J Comput Biol*, 9(2):447–464, 2002. doi:10.1089/10665270252935566. URL <http://dx.doi.org/10.1089/10665270252935566>. 102
- Thomas-Chollier, M., Sand, O., Turatsinze, J.V., Janky, R., et al. RSAT: regulatory sequence analysis tools. *Nucleic Acids Research*, 36(Web Server issue):W119–27, 2008. doi:10.1093/nar/gkn304. URL http://nar.oxfordjournals.org/cgi/content/full/36/suppl_2/W119. 183, 201
- Thompson, W. and Rouchka, E. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic acids ...*, 2003. URL <http://nar.oxfordjournals.org/cgi/content/abstract/31/13/3580>. 27, 68
- Tice-Baldwin, K., Fink, G.R., and Arndt, K.T. BAS1 has a Myb motif and activates HIS4 transcription only in combination with BAS2. *Science*, 246(4932):931–5, 1989. 167
- Tompa, M., Li, N., Bailey, T.L., Church, G.M., et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23(1):137–144, 2005. doi:10.1038/nbt1053. URL <http://dx.doi.org/10.1038/nbt1053>. 13, 16, 102, 103, 104, 105, 107, 108, 111, 112, 153, 172, 174
- Valen, E., Pascarella, G., Chalk, A., Maeda, N., et al. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Res*, 19(2):255–265, 2008. doi:10.1101/gr.084541.108. 5
- Valen, E., Sandelin, A., Winther, O., and Krogh, A. Discovery of regulatory elements is improved by a discriminatory approach. *PLoS Comput Biol*, 5(11):e1000562, 2009. doi:10.1371/journal.pcbi.1000562. URL [http:](http://)

REFERENCES

- [//www.ploscompbiol.org/article/info252Fjournal.pcbi.1000562](http://www.ploscompbiol.org/article/info252Fjournal.pcbi.1000562). 111, 113
- van Dongen, S., Abreu-Goodger, C., and Enright, A.J. Detecting microRNA binding and siRNA off-target effects from expression data. *Nat Methods*, 5(12):1023–5, 2008. doi:10.1038/nmeth.1267. URL <http://www.nature.com/nmeth/journal/v5/n12/abs/nmeth.1267.html>. 13
- van Nimwegen, E. Scaling laws in the functional content of genomes. *Trends Genet*, 19(9):479–84, 2003. 7
- Vegetti, S. and Koopmans, L. Bayesian strong gravitational-lens modelling on adaptive grids: objective detection of mass substructure in Galaxies. *Monthly Notices of the Royal Astronomical ...*, 2009. URL <http://arxiv.org/pdf/0805.0201>. 25
- Venters, B.J. and Pugh, B.F. A canonical promoter organization of the transcription machinery and its regulators in the *Saccharomyces* genome. *Genome Res*, 19(3):360–371, 2008. doi:10.1101/gr.084970.108. 2, 3, 123
- Vilo, J. Discovering Frequent Patterns from Strings. *egeeninc.com*, 1998. URL <http://www.egeeninc.com/u/vilo/Publications/CS-Report-1998-9.ps>. 101
- Vilo, J., Brazma, A., Jonassen, I., Robinson, A., et al. Mining for putative regulatory elements in the yeast genome using gene expression data. *Proc Int Conf Intell Syst Mol Biol*, 8:384–94, 2000. 101
- Visel, A., Prabhakar, S., Akiyama, J.A., Shoukry, M., et al. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet*, 40(2):158–60, 2008. doi:10.1038/ng.2007.55. 2
- Visel, A., Rubin, E.M., and Pennacchio, L.A. Genomic views of distant-acting enhancers. *Nature*, 461(7261):199–205, 2009. doi:10.1038/nature08451. 2, 4
- von Neumann, J. Various techniques used in connection with random digits. Monte Carlo methods. *National Bureau of Standards AMS*, 12:36–38, 1951. 23

REFERENCES

- Walter, J., Dever, C., and Biggin, M. Two homeo domain proteins bind with similar specificity to a wide range of DNA sites in *Drosophila* embryos. *Genes & Development*, 1994. URL <http://genesdev.cshlp.org/content/8/14/1678.short>. 5, 8
- Wang, G. and Zhang, W. A steganalysis-based approach to comprehensive identification and characterization of functional regulatory elements. *Genome Biol*, 7(6):R49, 2006. doi:10.1186/gb-2006-7-6-r49. URL http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Abstract&list_uids=16787547. 108
- Wang, T. and Stormo, G.D. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, 19(18):2369–2380, 2003. 11, 16
- Warren, C.L., Kratochvil, N.C.S., Hauschild, K.E., Foister, S., et al. Defining the sequence-recognition profile of DNA-binding molecules. *Proc Natl Acad Sci USA*, 103(4):867–72, 2006. doi:10.1073/pnas.0509843102. URL <http://www.pnas.org/content/103/4/867.long>. 12, 114
- Wijaya, E., Yiu, S., Son, N., Kanagasabai, R., et al. MotifVoter: a novel ensemble method for fine-grained integration of generic motif finders. *Bioinformatics*, 2008. doi:10.1093/bioinformatics/btn420. 108
- Wilson, D., Charoensawan, V., Kummerfeld, S.K., and Teichmann, S.A. DBD—taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Research*, 36(Database issue):D88–92, 2008a. doi: 10.1093/nar/gkm964. 6, 7, 8, 9, 10, 35, 136, 155
- Wilson, D., Pethica, R., Zhou, Y., Talbot, C., et al. SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res*, 37(Database issue):D380–6, 2009. doi:10.1093/nar/gkn762. URL http://nar.oxfordjournals.org/cgi/content/full/37/suppl_1/D380?view=long&pmid=19036790. 6, 19
- Wilson, R.J., Goodman, J.L., Strelets, V.B., and Consortium, F. FlyBase: integration and improvements to query tools. *Nucleic*

REFERENCES

- Acids Research*, 36(Database issue):D588–93, 2008b. doi:10.1093/nar/gkm930. URL http://nar.oxfordjournals.org/cgi/content/full/36/suppl_1/D588?view=long&pmid=18160408. 6
- Wingender, E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Briefings in Bioinformatics*, 9(4):326–332, 2008. doi:10.1093/bib/bbn016. 34
- Wingender, E., Chen, X., Fricke, E., Geffers, R., et al. The TRANSFAC system on gene expression regulation. *Nucleic Acids Res*, 29(1):281–283, 2001. 33
- Wolfe, S.A., Nekludova, L., and Pabo, C.O. DNA recognition by Cys2His2 zinc finger proteins. *Annu Rev Biophys Biomol Struct*, 29:183–212, 2000. doi:10.1146/annurev.biophys.29.1.183. URL <http://dx.doi.org/10.1146/annurev.biophys.29.1.183>. 14, 39
- Workman, C. and Stormo, G. ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac Symp Biocomput*, 5:464–475, 2000. 68, 77, 103, 113
- Wunderlich, Z. and Mirny, L. Different gene regulation strategies revealed by analysis of binding motifs. *Trends in Genetics*, 25(10):434–440, 2009. 8
- Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., et al. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–345, 2005. doi:10.1038/nature03441. URL <http://dx.doi.org/10.1038/nature03441>. 13, 102, 111, 112
- Xie, X., Mikkelsen, T.S., Gnirke, A., Lindblad-Toh, K., et al. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc Natl Acad Sci USA*, 104(17):7145–50, 2007. doi:10.1073/pnas.0701811104. URL <http://www.pnas.org/cgi/content/full/104/17/7145>. 13, 102
- Xie, X., Rigor, P., and Baldi, P. MotifMap: a human genome-wide map of candidate regulatory motif sites. *Bioinformatics*, 25(2):167–74,

REFERENCES

2009. doi:10.1093/bioinformatics/btn605. URL <http://bioinformatics.oxfordjournals.org/cgi/content/full/25/2/167>. 5
- Xing, E., Jordan, M., Karp, R., and Russell, S. A hierarchical Bayesian Markovian model for motifs in biopolymer sequences. *Advances in Neural Information Processing Systems*, pp. 1513–1520, 2003a. 68
- Xing, E.P. and Karp, R.M. MotifPrototyper: a Bayesian profile model for motif families. *Proc Natl Acad Sci U S A*, 101(29):10523–10528, 2004. doi:10.1073/pnas.0403564101. URL <http://dx.doi.org/10.1073/pnas.0403564101>. 34, 39, 42, 67, 68, 81, 82, 83, 84, 85, 86, 87, 91
- Xing, E.P., Wu, W., Jordan, M.I., and Karp, R.M. LOGOS: a modular Bayesian model for de novo motif detection. *Proc IEEE Comput Soc Bioinform Conf*, 2:266–76, 2003b. 68
- yong Li, X., MacArthur, S., Bourgon, R., Nix, D., et al. Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. *PLoS Biol*, 6(2):e27, 2008. doi:10.1371/journal.pbio.0060027. URL <http://www.plosbiology.org/article/info52Fjournal.pbio.0060027>. 8
- Zare-Mirakabad, F., Ahrabian, H., Sadeghi, M., Nowzari-Dalini, A., et al. New scoring schema for finding motifs in DNA Sequences. *BMC Bioinformatics*, 10(1):93, 2009. doi:10.1186/1471-2105-10-93. 108
- Zeitlinger, J., Simon, I., Harbison, C., and Hannett, N. Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell*, 2003. URL <http://linkinghub.elsevier.com/retrieve/pii/S0092867403003015>. 144
- Zhang, J., Jiang, B., Li, M., Tromp, J., et al. Computing exact P-values for DNA motifs. *Bioinformatics*, 23(5):531, 2007. 117
- Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., et al. Model-based Analysis of ChIP-Seq (MACS). *Genome Biol*, 9(9):R137, 2008a. doi:10.1186/gb-2008-9-9-r137. 185

REFERENCES

- Zhang, Z.D., Rozowsky, J., Snyder, M., Chang, J., et al. Modeling ChIP Sequencing In Silico with Applications. *PLoS Comput Biol*, 4(8):e1000158, 2008b. doi:10.1371/journal.pcbi.1000158. [189](#)
- Zheng, W., Zhao, H., Mancera, E., Steinmetz, L.M., et al. Genetic analysis of variation in transcription factor binding in yeast. *Nature*, 464(7292):1187–91, 2010. doi:10.1038/nature08934. [112](#)
- Zhu, C., Byers, K.J., Mccord, R.P., Shi, Z., et al. High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res*, 19(4):556–566, 2009. doi:10.1101/gr.090233.108. [XII](#), [12](#), [35](#), [36](#), [99](#), [114](#), [115](#), [128](#), [129](#), [130](#), [131](#), [132](#), [134](#), [139](#), [140](#)
- Zhu, J. and Zhang, M.Q. SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, 15(7-8):607–611, 1999. [101](#), [114](#)
- Zhu, Q. and Halfon, M.S. Complex organizational structure of the genome revealed by genome-wide analysis of single and alternative promoters in *Drosophila melanogaster*. *BMC Genomics*, 10(1):9, 2009. doi:10.1186/1471-2164-10-9. [3](#)

Appendix A - iMotifs

Motivation

¹ Short sequence motifs are an important class of models in molecular biology, used most commonly for describing transcription factor binding site specificity patterns. High-throughput methods have been recently developed for detecting regulatory factor binding sites *in vivo* and *in vitro* and consequently high-quality binding site motif data are becoming available for increasing number of organisms and regulatory factors. Development of intuitive tools for the study of sequence motifs is therefore important.

iMotifs is a graphical motif analysis environment that allows visualisation of annotated sequence motifs and scored motif hits in sequences. It also offers motif inference with the sensitive NestedMICA algorithm, as well as overrepresentation and pairwise motif matching capabilities. All of the analysis functionality is provided without the need to convert between file formats or learn different command line interfaces.

The application includes a bundled and graphically integrated version of the NestedMICA motif inference suite that has no outside dependencies. Problems associated with local deployment of software are therefore avoided.

¹The following manuscript is published in [Piipari et al. \(2010b\)](#) and is a result of collaborative work between the author of this thesis (MP), Dr Thomas Down (TD) and my PhD thesis supervisor Dr Tim Hubbard. The authors' contributions are as follows: MP conceived the work, wrote the software and the manuscript. TD and TH provided feedback. All authors read the manuscript and provided feedback.

Availability

iMotifs is licensed with the GNU Lesser General Public License v2.0 (LGPL 2.0). The software and its source is available at <http://wiki.github.com/mz2/imotifs> and can be run on Mac OS X Leopard (Intel/PowerPC). I also provide a cross-platform (Linux, OS X, Windows) LGPL 2.0 licensed library `libxms` for the Perl, Ruby, R and Objective-C programming languages for input, output of XMS formatted annotated sequence motif set files.

Introduction

Until recent years, studying sequence specificity of transcription factors systematically has been limited to a relatively small number of organisms and transcription factors. High throughput protein-DNA interaction assays such as protein binding microarrays (Berger et al., 2006), bacterial one-hybrid screens (Meng et al., 2005), large ChIP-chip studies and advances in motif inference algorithms and tools has however caused an expansion of motif databases such as UNI-PROBE (Newburger and Bulyk, 2009), TRANSFAC (Matys et al., 2006) and JASPAR (Bryne et al., 2008).

Sequence motif analysis tools can be hard to deploy and use locally. Many commonly used software packages have therefore been made available as web applications (Mahony and Benos, 2007; Thomas-Chollier et al., 2008). Public servers can however be limited in the CPU time given to users which can rule out their use for large scale studies. Data exchange and usability can also be a challenge. Therefore I have created an OS X based desktop software package for sequence motif analysis that is easy to install and update. Compared to previously published desktop based *cis*-regulatory sequence analysis tools such as TOUCAN (Aerts et al., 2003) or Sockeye (Montgomery et al., 2004), iMotifs is more focused on visualisation and computation of sequence motifs, although it also supports visualising scored motif matches in sequences.

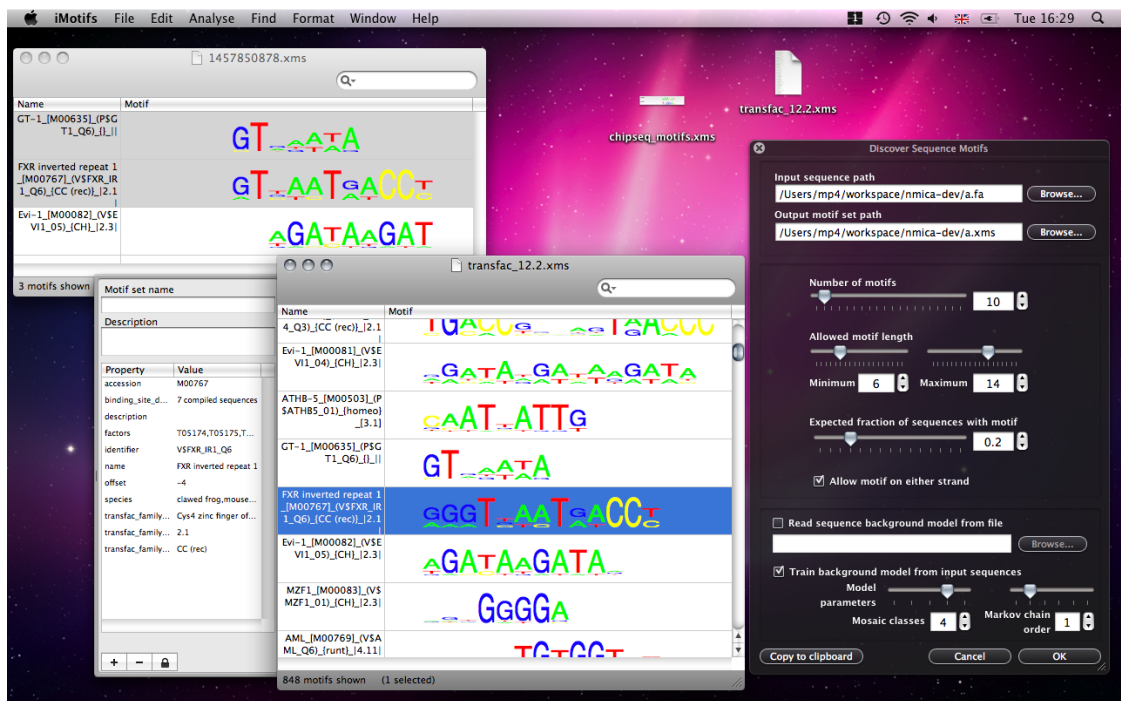


Figure A1: iMotifs can present motif sets and alignments. It integrates with the OS X desktop's previewing functionality and includes a number of analysis tools including an integrated NestedMICA motif inference tool.

Features

iMotifs is designed for visualisation and analysis of *cis*-regulatory motifs and sequences. It can be used to retrieve sequences (for example for a coregulated group of genes), infer *cis*-regulatory motifs from them and score sequences with motif models, visualise them and their scored matches, and compare them against other motifs (Fig. 1 shows the core functionality). A tutorial is included on the website for common tasks (see Availability). Motifs can be manipulated and moved between sets by dragging and dropping, and filtered using keyword searches. Summary statistics such as entropy, column count or distance from closest pair can also be shown alongside. Free form key-value pair metadata such as database identifiers, species or notes can be viewed and edited. PDF export and printing is available. Import and export of TRANSFAC formatted motif files is also possible.

iMotifs can be used to retrieve sequences from the Ensembl database ([Hubbard et al., 2009](#)). The retrieved sequences can be aligned either to transcription start sites (putative promoter sequence) or ends (e.g. for micro-RNA seed finding), and they can be filtered by gene identifiers. The retrieval tool can fetch specific sequence regions using GFF formatted annotation files, and includes specific support for ranking and retrieving regions of interest based on ChIP-seq ‘peaks’: MACS ([Zhang et al., 2008a](#)), FindPeaks ([Fejes et al., 2008](#)) and SWEMBL formats are supported. Sequences are optionally processed to mask repeats and translated sequence.

iMotifs supports the quick previewing and thumbnailing service native to OS X (QuickLook). Previewing is especially useful for browsing sequence motif sets stored remotely (e.g. on a remote cluster) as no manual transfer or file opening is needed. An automated software update mechanism is included.

Many common motif analysis tasks are supported. These include finding closest matching and reciprocally matching motif pairs between two motif sets with the distance metric and algorithm described in [Down et al. \(2007\)](#). Motif multiple alignments can be visualised and computed with a greedy gapless motif multiple alignment algorithm. Motif inference experiments can be run with the integrated NestedMICA ([Down and Hubbard, 2005](#)) tool simply by dragging FASTA for-

matted sequence files to iMotifs. Downstream analyses such as motif scanning, overrepresentation analysis, and motif hit score cutoff assignment as described in [Down et al. \(2007\)](#) is also possible. Analysis tasks are run in parallel without blocking the user interacting with the application.

Interoperability

Although iMotifs itself works only on computers running Mac OS X, the analysis tools developed for and included in iMotifs are cross-platform (Java based) and depend only on libraries included with the package. Most analysis functions are implemented by stand-alone command-line programs. This makes it possible to rapidly integrate unmodified tools into iMotifs. The included analysis tools can also be run on any UNIX system without iMotifs.

I feel that the use of a standard format for exchanging sequence motif data is beneficial for the research community, given the literally hundreds of motif inference tools and databases that are available (reviewed in [Das and Dai \(2007\)](#)). To encourage the take up of a standard file format for motifs, I provide a programming interface for the input and output of the annotated motif file format XMS for the Perl, Ruby, R and Objective-C languages. The Perl and R libraries can also be used to visualise sequence logos.

Conclusions

I have created an integrated desktop application for short sequence motif analysis. It incorporates visualisation, inference, alignment and comparison tools. The application widens the user base of sequence motif analysis tools and can improve the productivity of researchers working with sequence motif data. I aim to integrate with more sequence motif analysis tools and web services and to develop further the already included basic protein motif visualisation and inference support.

I also encourage the introduction of a standard format for exchange of sequence motif data by providing conversion utilities and an API for input and

output of XMS motif set files for a number of common bioinformatics programming languages.

Appendix B - The motif inference tutorial

Introduction

¹The tutorial below is aimed to introduce a researcher new to regulatory genomics to taking use of the NestedMICA and NMICA-extra motif inference tools to identify and analyze sequence motifs from noncoding genomic sequence. We demonstrate uses of the NMICA-extra package with a short sequence analysis project where NestedMICA is first used to recreate the STAT1 transcription factor binding motif from [Robertson et al. \(2007\)](#).

The first step is retrieving input genomic sequences corresponding to the ChIP-seq peak regions. To ease the retrieval and importantly preprocessing of input sequence (repeat masking and exclusion of translated sequences), NestedMICA has been enhanced with a number of tools for retrieving sequence from the Ensembl database (Flicek et al. 2008): `nmensemblseq`, `nmensemblfeat` and `nmensemblpeakseq`.

1. `nmensemblseq`: retrieves sequences around transcription start sites or 3' UTRs or introns.

¹The following manuscript is a result of collaboration between the author of this thesis (MP), Dr. Thomas Down (TD), and MP's thesis supervisor Dr. Tim Hubbard (TH). The work is published in ([Piipari et al., 2011](#)). Authors' contributions are as follows: MP wrote the manuscript, all authors read it and provided feedback.

-
2. `nmensemblfeat`: retrieves specific sequence regions using GFF formatted annotation files as input.
 3. `nmensemblpeakseq`: retrieves sequence regions close to ChIP-seq peaks
 - MACS ([Zhang et al., 2008b](#))
 - FindPeaks ([Fejes et al., 2008](#))
 - SWEMBL (<http://www.ebi.ac.uk/~swilder/SWEMBL/>)

Two more generic sequence feature formats are also supported:

1. BED (<https://cgwb.nci.nih.gov/goldenPath/help/customTrack.html>)
2. GFF (<http://www.sanger.ac.uk/resources/software/gff/spec.html>)

We will use `nmensemblpeakseq` to retrieve sequence windows corresponding to 50 base long sequence windows around ranked ChIP-sequencing peak maximum positions of the 500 top-ranking peaks.

```
nmensemblpeakseq -database homo_sapiens_core_52_36n \  
-host ensembl.db.ensembl.org \  
-user anonymous -port 5306 \  
-inputFormat peaks \  
-peaks STAT1_IFNGstim_hg18_xset200_dupsN_ht10.sub.peaks \  
-maxCount 500 \  
-aroundPeak 50 \  
-minLength 50 \  
-minNonN 80 \  
-repeatMask \  
-excludeTranslations \  
-chunkLength 100 > stat1-stimulated-50bp-around-max.fasta
```

The regions included in the dataset have been mapped to the NCBI36 human genome assembly (Ensembl release 52). We therefore request sequences relative to the same release of the Ensembl database (`homo_sapiens_core_52_36n`). The reason for choosing the database, hostname and port combination above is that at the time of writing the publicly available Ensembl instance that serves the Ensembl release 52 is the port 5306 on `ensembl.db.ensembl.org`.

Sequence background model estimation

Before motif inference from the retrieved sequences, it is advisable to estimate a NestedMICA sequence background model as a separate step. This can be done with the command `nmmakebg`, which requires two input parameters: Markov chain order and the number of mosaic classes. The Markov chain parameter is usually set to 1st order because some of the DNA motif specific downstream analysis tools require this. The class count parameter that yields best performance tends to be 4 (Down and Hubbard, 2005), but it is best to evaluate different mosaic class parameters before the potentially long-running motif inference analysis. Background models can be evaluated using the command `nmevaluatebg`

```
nmevaluatebg -order 1 \  
-minClasses 1 -maxClasses 8 \  
-seqs stat1-stimulated-500bp-around-max.fasta \  
-testSeqs stat1-stimulated.fasta \  
> min1classes-max8classes-eval-bg.eval
```

The output of `nmevaluatebg` can be used to find the mosaic order parameters at which the background model performance, as measured by sequence likelihood given the background model, shows little increase or drops. These parameter values are then taken as the optimal ones. The easiest way to interpret the results is to plot them using R with the `nmica` R package (<http://github.com/mz2/r-utilities>).

```
>library(nmica)  
>eval.results <-  
  read.nmevaluatebg(  
    min1classes-max8classes-eval-bg.eval )  
>plot(eval.results$classes ~ eval.results$likelihood)
```

This evaluation (Figure A2) suggest a suitable order parameter as 4. We can now commence with the background model estimation:

```
nmmakebg -classes 4 -order 1 \  
-seqs stat1-stimulated-500bp-around-max.fasta \  
-out seqs-4classes-1storder.bg
```

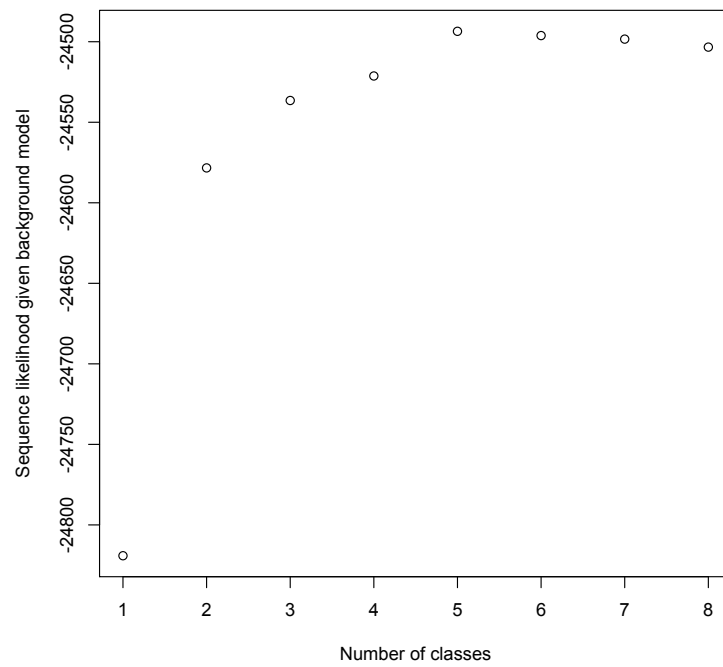


Figure A2: Output of the `nmevaluatebg` command plotted in R.

Motif inference

After retrieving the input sequences and determining class and order parameters with `nmmakebg`, we can now run the NestedMICA motif inference with the command `nminfer`.

```
nminfer -seqs stat1-stimulated.fasta \  
-numMotifs 1 \  
-backgroundModel stat1-stimulated-4classes.bg \  
-minLength 6 -maxLength 14 \  
-minSeqLength 50 \  
-maxCycles 1000000 \  
-revComp \  
-expectedUsageFraction 0.70 \  
-checkpoint stat1-stimulated-checkpoint \  
-sampleFile stat1-stimulated-sample \  
-sampleInterval 10000 \  
-checkpointInterval 10000 \  
-logInterval 100 \  
-distributed -port 5001 -threads 4 \  
-out motifs.xml > nminfer.log 2> nminfer.err
```

Note that the above command line instructs periodic output of checkpoint files that can be used to restart the computation, as well as sample motif set files (preliminary motif set solutions that can be visualised whilst the computation is still running). The above `nminfer` command line also demonstrates distributed computing with NestedMICA: the `-distributed` and `-port 5001` instruct `nminfer` to act as a server that responds at port 5001 to distribute its work load to separate worker nodes (each of which would typically correspond to one computer in a computational cluster). Worker nodes that connect to a server can be created with the command `nmworker`.

```
nmworker -server nmica_server_hostname -port 5001 -threads 4
```

The actual host name given above depends on the host name of the computer where `nminfer` was set to run.

Motif overrepresentation

When interpreting the output of NestedMICA, it is important to note that the algorithm does not rank its output motifs relative to each other or predict hit positions for them. A common way of assessing computationally inferred motifs is through a motif overrepresentation analysis. By overrepresentation analysis we mean a statistical exercise where sequences with the motif (the positive set) are discriminated from those assumed to be devoid of it (the negative set). The approach taken in NMICA-extra for computing the degree of overrepresentation in a set of sequences is the ROC-AUC (Receiver-Operator Characteristic Area Under the Curve) statistic, computed with the tool `nmrocauc`. In short, sequences are labelled as positive or negative and the maximum motif bit score is used to predict if any given sequence is part of the positive or the negative sequence set – the maximum motif hit score is used to classify the sequences. The AUC statistic that is reported by this analysis is a measure of how often a randomly chosen positive sequence is ranked above a randomly chosen negative sequence. It therefore provides a measure of separation of maximum motif hit score distribution of the positive examples from the negative examples. To estimate the null distribution of scores with the length distribution and sequence composition used, the negative sequences are shuffled and the randomly generated sequences are then scored according to the same criterion. The shuffling conducted as part of this method accounts for the fact that the maximum hit score distributions of sequences can vary based on nucleotide composition.

```
#Retrieve 1000 random core promoter sequences:  
#900bp upstream of TSS and up to 100bp downstream  
#Exclude any repeats and translated sequence  
nmensemblseq \  
-sampleRandomGenes 1000 \  
-fivePrimeUTR 900 100 \  
-proteinCoding \  
-repeatMask \  
-excludeTranslations \  
-database homo_sapiens_core_52_36n \  

```

```

-host ensembl.ensembl.org \
-port 5306 \
-user anonymous > \
1000-random-human-promoters_900bp-upstream-100bp5utr.fasta

#Sample 1000 random sequences of length 50
#The sequence window length
#is the same as that of the peak sequence windows
nmrandomseq \
-count 500 \
-length 50 \
-seqs 500-random-human-promoters_900bp-upstream-100bp5utr.fasta > \
100bp-windows-from-random-human-promoters.fasta

nmrocauc \
-positiveSeqs stat1_chip_peaks.fasta \
-negativeSeqs \
50bp-windows-from-random-human-promoters.fasta \
-motifs stat1_human.xml
#Output:
#motif2  0.992880      0.00000

```

The above analysis shows that the discovered motif is strongly over-represented in the ChIP-sequencing peaks when compared to random noncoding sequence regions of the same genome (the empirical p -value, which is the second value in the `nmrocauc` output, is below 10^{-5}).

The STAT transcription factors and DNA binding motif have therefore been deposited to publicly available databases such as TRANSFAC (Matys et al., 2006). This makes it possible to validate the sequence motif we have inferred from the ChIP-seq data with NestedMICA by searching it against motif databases with the reciprocal matching procedure described above. Reciprocal matching of motifs is implemented in the tool `nmshuffle` that is distributed as part of NestedMICA.

```

nmshuffle -bootstraps 100000 \
transfac_12.2.xms stat1-human.xms
#Output:
#motif0  STAT5A_[M00457]  0.531520      0      0.00000

```

A statistically significant match is identified for the NestedMICA STAT1 motif in the TRANSFAC database (the empirical p -value which is the last column in the nmshuffle output above, is below 10^{-5}). An inspection of the closest matching motifs makes it clear that NestedMICA infers a very similar binding specificity pattern for STAT1 as has been previously reported for members of the STAT family transcription factors (Figure A3).

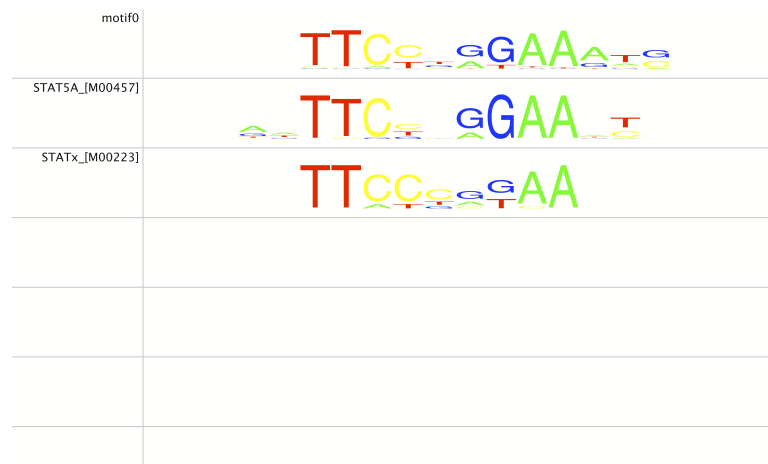


Figure A3: The predicted motif alongside known STAT motifs from the TRANSFAC database.

Appendix C - Motif inference algorithm assessment parameters

The parameters given for each of the motif inference methods tested in Chapter 5 are given below.

NestedMICA

The NestedMICA algorithm was run with the following parameters:

```
nminfer -numMotifs 200 \  
-minLength 6 -maxLength 14 \  
-expectedUsageFraction 0.2 \  
-backgroundModel sc_4classes_1order.bg \  
-seqs orthologs-sc-1000.fa
```

Sequence background model parameters were evaluated with `nmevaluatebg` using a randomly chosen half of the input sequence for model learning (`-trainSeqs`) and the remaining half for model evaluation (`-testSeqs`). As suggested in the NestedMICA manual, the Markov chain order was kept constant at 1 (`-order 1`) and the mosaic class parameter was varied between 1 and 8 (`-minClasses 1 -maxClasses 8`). The sequence likelihood values achieved with each of these parameter settings are shown in Figure A4.

Mosaic class count 4 was chosen based on the above evaluation because it presents an acceptable compromise between a descriptiveness and complexity of

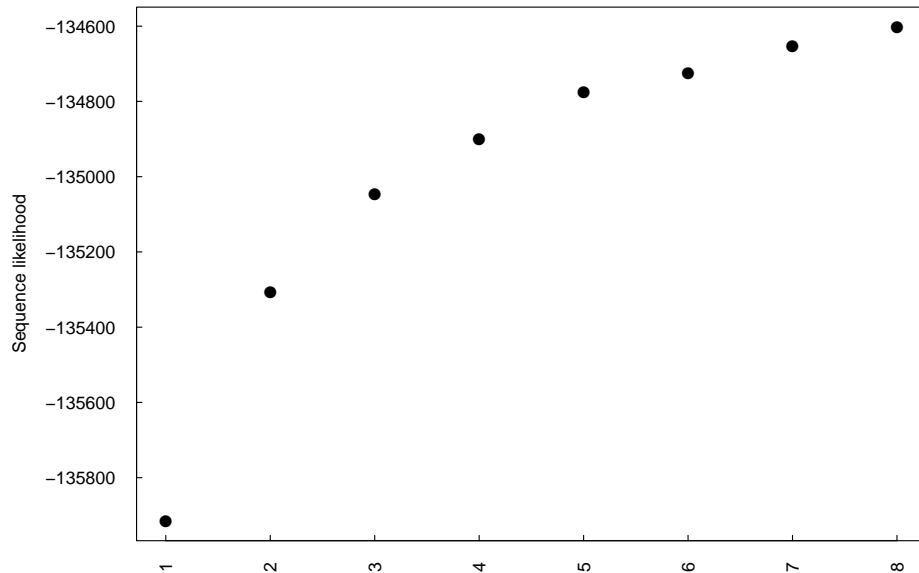


Figure A4: Evaluation of sequence background model class counts at Markov chain order 1.

the model; Increasing the class count beyond 4 results in diminishing gains in the sequence likelihood. The runtime of the application also increases.

Weeder

The weeder algorithm ([Pavesi et al., 2001](#)) was run with the `weederlauncher.out` driver script distributed with the program. The ‘large’ settings were used to search for motifs between 6 and 12 nucleotides long, and motifs were allowed to be present on either strand:

```
weederlauncher.out orthologs-sc-1000.fa SC large S M T200
```

For all downstream analyses, the motif output by the program were trimmed to the top 200 reported motifs.

AlignACE

The parameters used for running AlignACE (Roth et al., 1998) are described below:

```
AlignACE -numCols 10 -gcback 0.38 -i orthologs-sc-1000.fa
```

The sequence background model used by AlignACE is a 0th order Markov chain, simply parameterised by the overall GC content of the yeast genome (Goffeau et al., 1996). The motif length (number of columns) was set to 10. Length of 10 was chosen because it is the median motif length in the JASPAR motif database which the predicted motifs are primarily compared with.

MEME

MEME version 4.3.0 (Bailey et al., 2006) was run with the following parameters:

```
meme.bin orthologs-sc-1000.fa \  
-dna -mod anr \  
-nmotifs 100 -minw 6 -maxw 14 \  
-bfile ~/meme_4.3.0/tests/common/yeast.nc.6.freq
```

The motifs were constrained to lengths between 6 and 14, similarly as done with NestedMICA. The background model used was the 6th order Markov chain background model trained from *S. cerevisiae* intergenic sequences which is supplied with MEME 4.3.0 (motif finding with a 3rd order background was also attempted). The sequence-motif model used was the “any number of repeats” model (-mod anr). Number of motifs was set to 100 – it was the largest number of motifs that MEME allows.

MotifSampler

MotifSampler (Thijs et al., 2001) was run with the following parameters:

```
MotifSampler -f orthologs-sc-1000.fa \  

```

```
-b orthologs-sc-1000.motifsamplerbg \  
-r 50 -s 1 -M 1 -n 50 -w 10 \  
-o orthologs-sc-1000.motifsamplerout \  
-m orthologs-sc-1000.motifs
```

The motif count parameter 50 (`-n 50`) was used because the program did not report motifs when large numbers of motifs were requested. The motif width 10 was chosen as it was the maximum allowed by the program, and the median motif length in the JASPAR database. Before the motif inference program was run, a 2nd order background model was trained from the input sequences using the `CreateBackgroundModel` tool supplied with `MotifSampler`, with the following parameters:

```
CreateBackgroundModel \  
-f ../orthologs-sc-1000.fa \  
-b orthologs-sc-1000.motifsamplerbg \  
-o 2 -n SC
```

YMF

YMF ([Sinha and Tompa, 2003b](#)) was run with the following parameters::

```
./stats stats.config 200 8 \  
ymftables/yeast -sort orthologs-sc-1000.fa
```

Two hundred 8-mers were inferred, using the yeast background nucleotide frequencies from the table supplied with the program (`./ymftables/yeast`). The output of YMF was post-processed another program, `FindExplanators` ([Blanchette and Sinha, 2001](#)), which removes redundancy amongst the consensus strings, outputting supposedly independent motifs.

```
find_explanators \  
ymftables/yeast_powersGeneralized.3.bin \  
orthologs-sc-1000.fa stats/results 5
```

FindExplanators reported a single motif AAARNRAAA regardless of the final explainer motif count parameter, which was varied. An inspection of the YMF results, which were given to FindExplanators as input, shows that the YMF output indeed only contains consensus strings that closely fit either AAARNRAAA or its reverse complement TTTYNYTTT. An excerpt with the first ten motifs from the set of 200 are given below.

```
2 AAARNRAAA 1529 48.93 345.6754 584.8017
3 TTTYNTTTY 1582 48.37 365.7588 632.3148
4 AAAANRAAA 1223 48.17 242.6953 414.1873
5 AAAANAAAA 994 47.53 167.9777 302.0721
6 AAARNAAAA 1202 47.46 239.8017 411.0152
7 ARAANRAAA 1478 47.30 354.0071 564.6885
8 TTTTNTTTY 1258 47.03 253.1523 456.4886
9 TTYTNTTTY 1514 47.00 360.8600 602.0605
10 AAAANRRRAA 1493 46.94 351.1163 591.8228
```

As one can see, motifs output by YMF with these parameters are a largely redundant set. I chose to still analyse these motifs alongside the other predictions further, to see how a highly redundant motif set would perform in my assessment.

SOMBRERO

SOMBRERO ([Mahony et al., 2005b](#)) was run with the following parameters:

```
SOMBRERO -t orthologs -sc -1000.fa \
-b /nfs/users/nfs_m/mp4/sombrero/yeast.back \
-lm 6 14 \
-time 200 \
-out results.sombrero
```

The 2^{nd} order sequence background model of the yeast genome was downloaded from <http://bioinf.nuigalway.ie/sombrero/binaries/backgrounds.zip>. The training iteration count was set to 1000 (ten times larger value than the default, to reflect the large nature of the problem). The minimum and maximum motif

lengths were set to 6 and 14 respectively. The program output was cut to 200 motifs by ranking motifs by the z -score which SOMBRERO reports.

Oligoanalysis

Oligo-analysis (Thomas-Chollier et al., 2008) was run with the web form included in the RSA Tools web server at http://rsat.ulb.ac.be/rsat/oligo-analysis_form.cgi, with the parameters shown in Figure A5, to discover a total of 50 over-represented sequence words.

Analysis of oligomer occurrences in nucleotidic of peptidic sequences
Reference: van Helden, J., André, B. and Collado-Vides, J. (1998). . J Mol Biol 281, 827-42.

Sequence **Format** Paste your sequence in the box below

Or select a file to upload
 No file chosen

Mask

Sequence type

purge sequences (highly recommended)

Oligomer counting mode

Oligomer length **prevent overlapping matches**

Count on **return reverse complements together in the output**

Background model

Genome subset Sequence type

Organism

Taxon

Estimate from input sequence

Markov model (higher order dependencies) order

Equiprobable residues (usually NOT recommended)

Figure A5: Parameter choices used with Oligo-analysis.