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 nucleotide 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 *Saccharomyces cerevisiae* 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 *et al.* (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 *et al.* (2010) TF knockout and expression data based target set 120
5.2.6.3 Harbison *et al.* (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 *cerevisiae* 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 *de novo* 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 (ZOOPS). 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 *nmalign*. 59

2.8 Metamotifs estimated with the metamotif nested sampler algorithm with varying relative frequency of metamotif samples. 61

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 crossvalidation). 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

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 *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). 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

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

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

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 TRANSFAC database. 195

A4 Evaluation of sequence background model class counts at Markov chain order 1. 197

A5 Parameter choices used with Oligo-analysis. 201