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 minumum 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).

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

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 algorihm. 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.

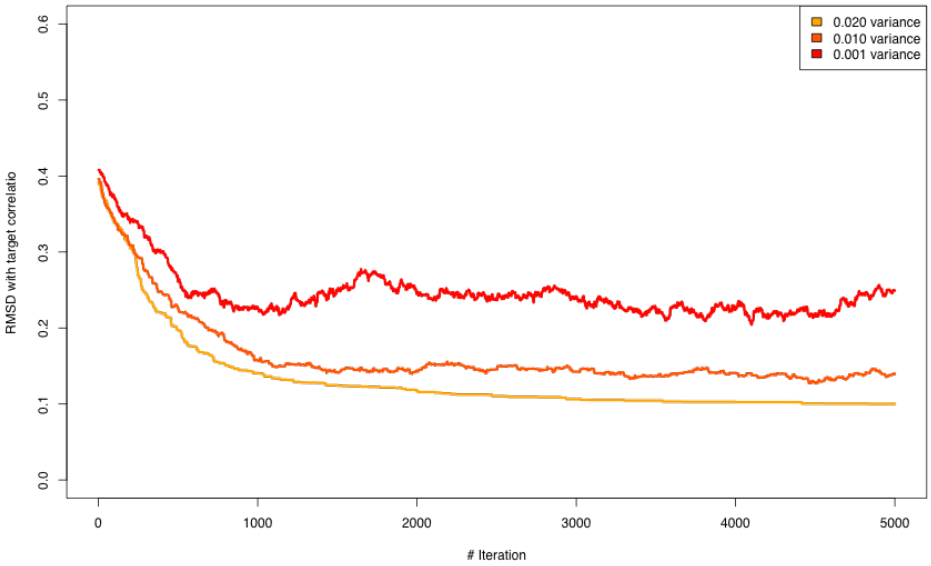

Figure 6.1: Three Markov chains aiming to draw a sample from $\mathbb{P}(\mathbf{M}|\mathbf{G}, p)$, each with a different p parameter.

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

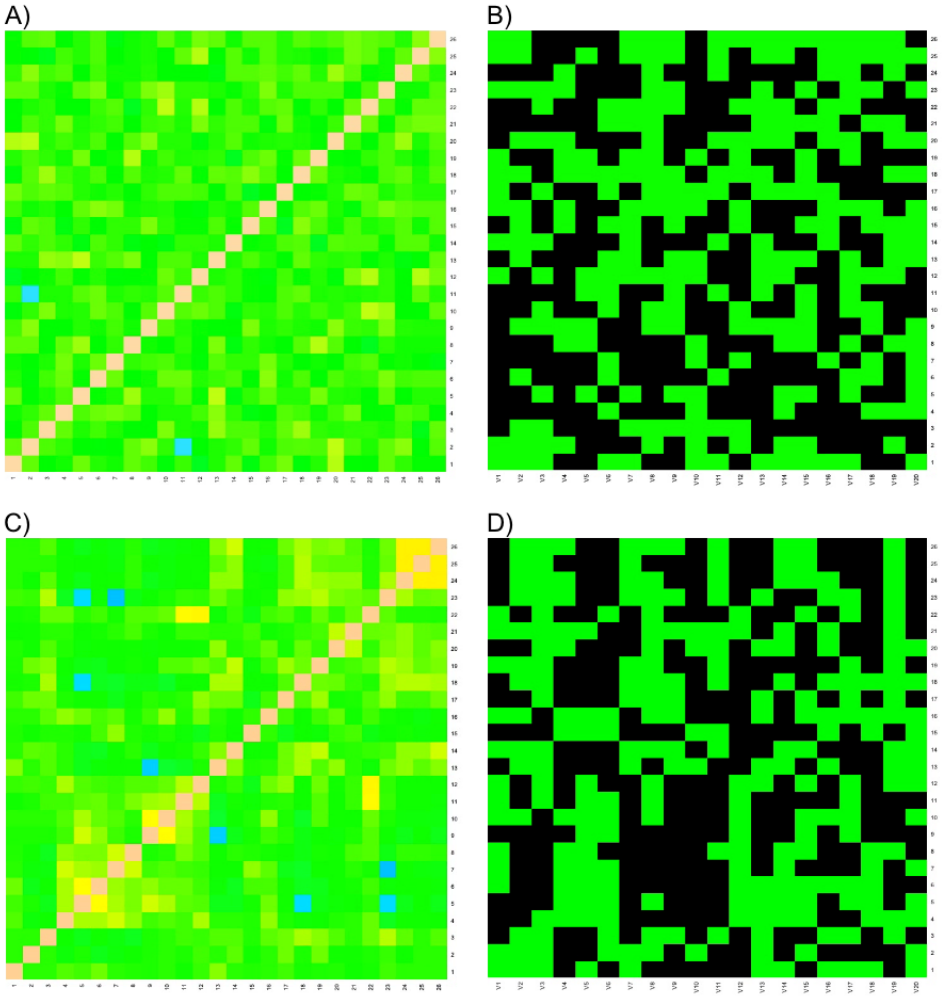

Figure 6.2: Mixing matrices and their correlations. The correlation matrices (panels A and C) of the start and end state of one of the 5000 step long MC chains from Figure 6.1. Panels B and D show the mixing matrices at the start (A) and end (D) of the sampling. Black states in panels B and D are mixing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

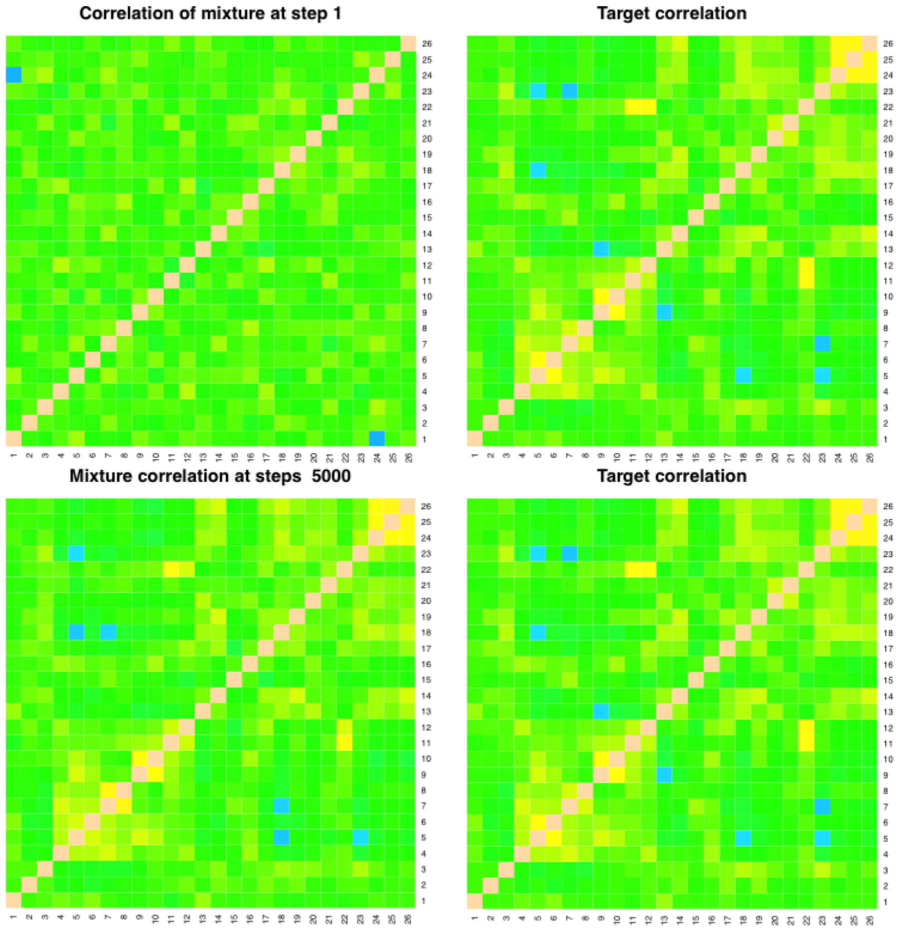

Figure 6.3: The sampling algorithm produces mixing matrices that are closely related in correlation pattern to the target (gene expression) correlation matrix. Gene expression correlations are shown on the right, and the mixture matrix correlations in the left. Whereas there are hardly any correlated 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.