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

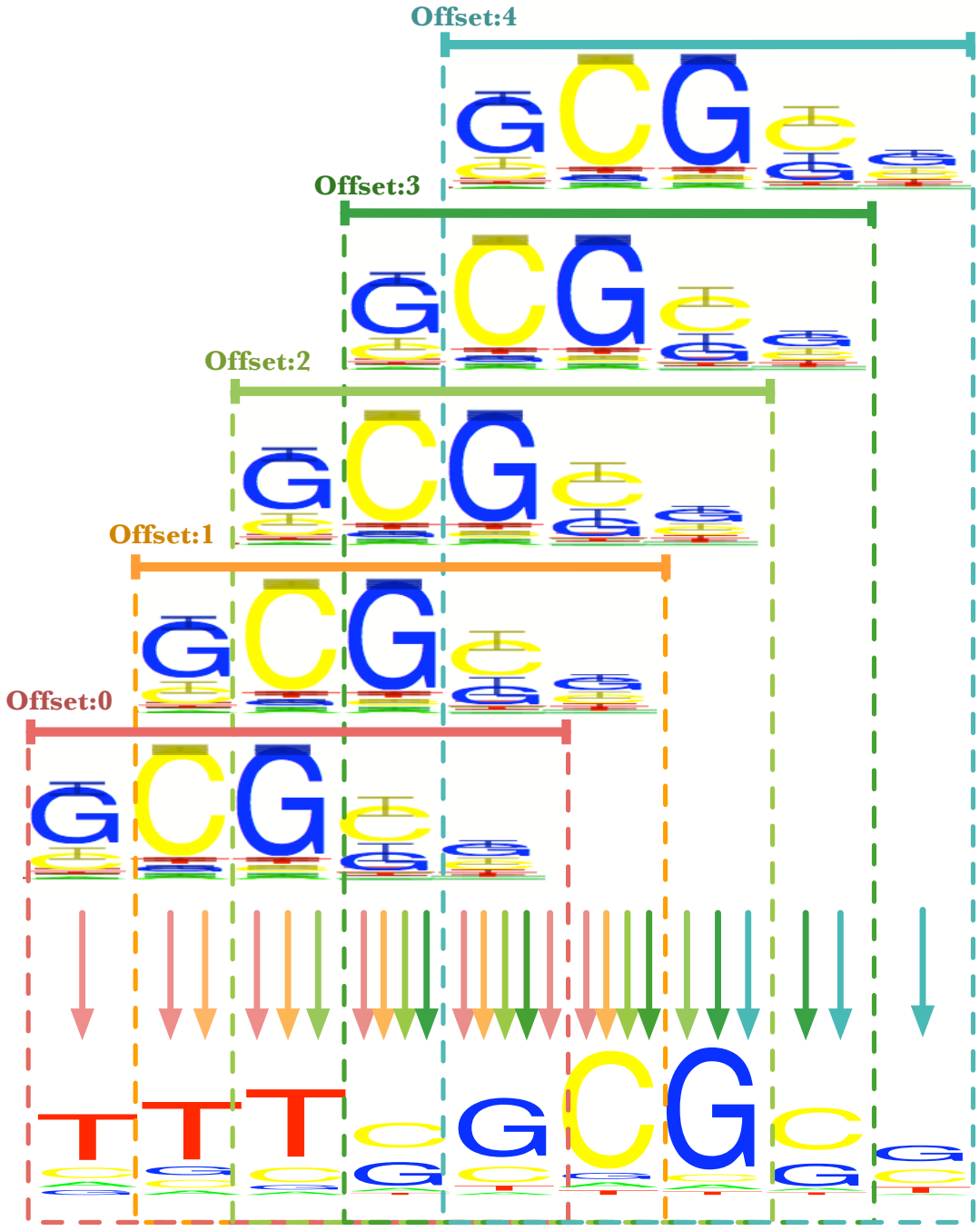

Figure 3.1: Metamotif densities with all offsets of the metamotif (shown above the PWM) are summed over the length of the motif (five different offsets shown, with different colours).

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-κB 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 1^{st} order trained from the 2000nt long intronic sequences with `nmmakebg`.

The source motifs (ZAP1, HIF1, TBX5, TAL1, NF-κB) 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.

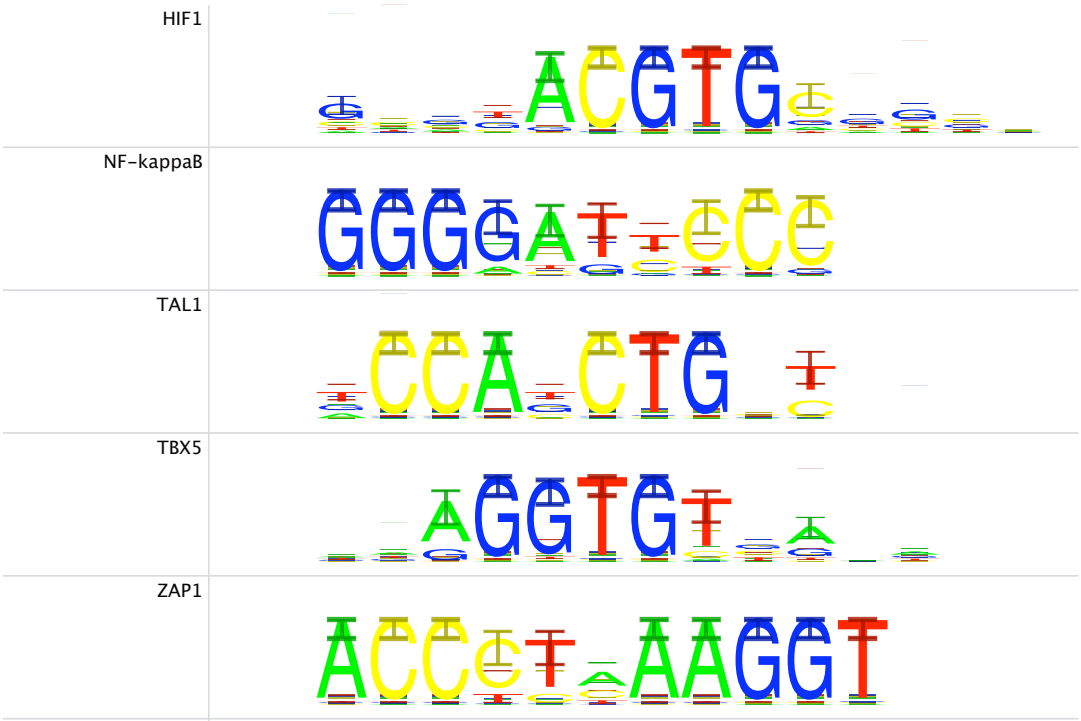

Figure 3.2: Synthetic metamotifs contributing to the motif prior functions used in the assessment. Error bars represent 95% confidence intervals.

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

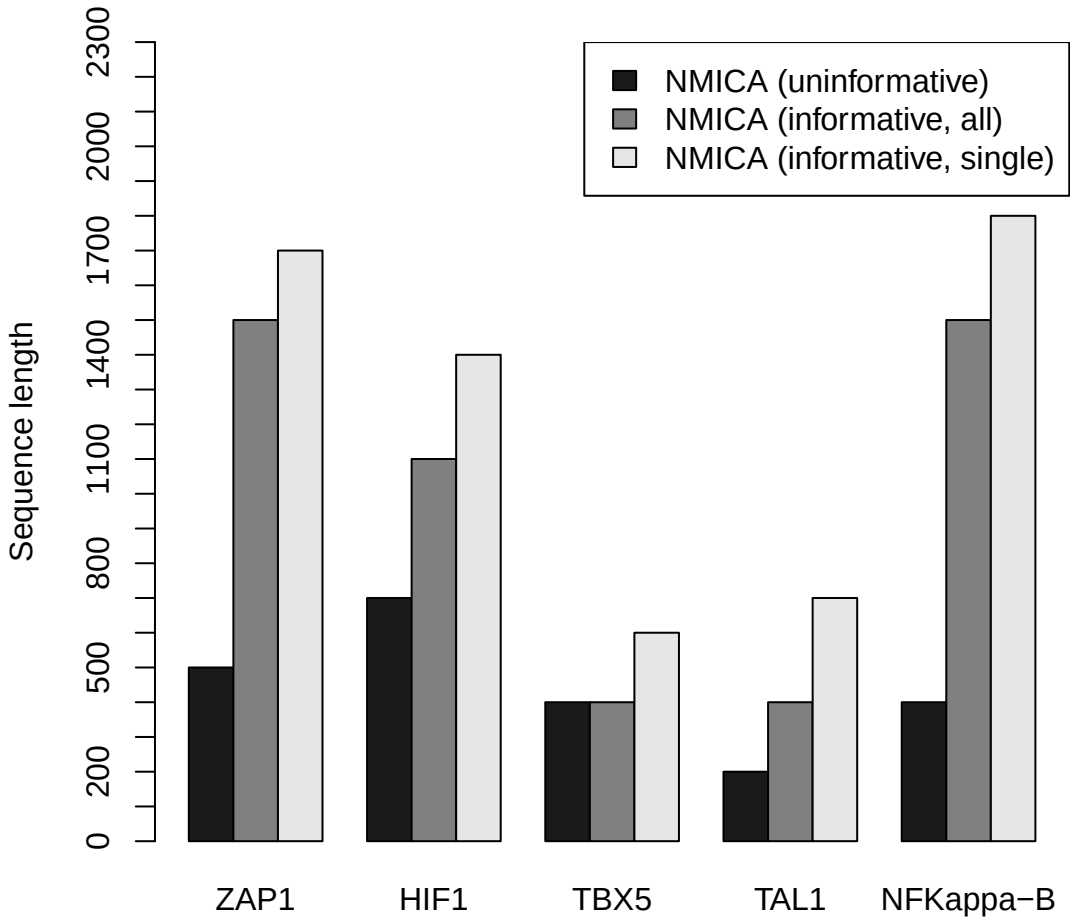

Figure 3.3: Informative weight matrix prior improves NMICA's sensitivity to resolve motifs present in human intronic sequence in low frequency (0.2 frequency). The bars represent the sequence length at which a motif closely similar to the input motif was successfully recovered ($p < 0.05$, empirical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

spiking intronic sequence with the NF-κB 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).

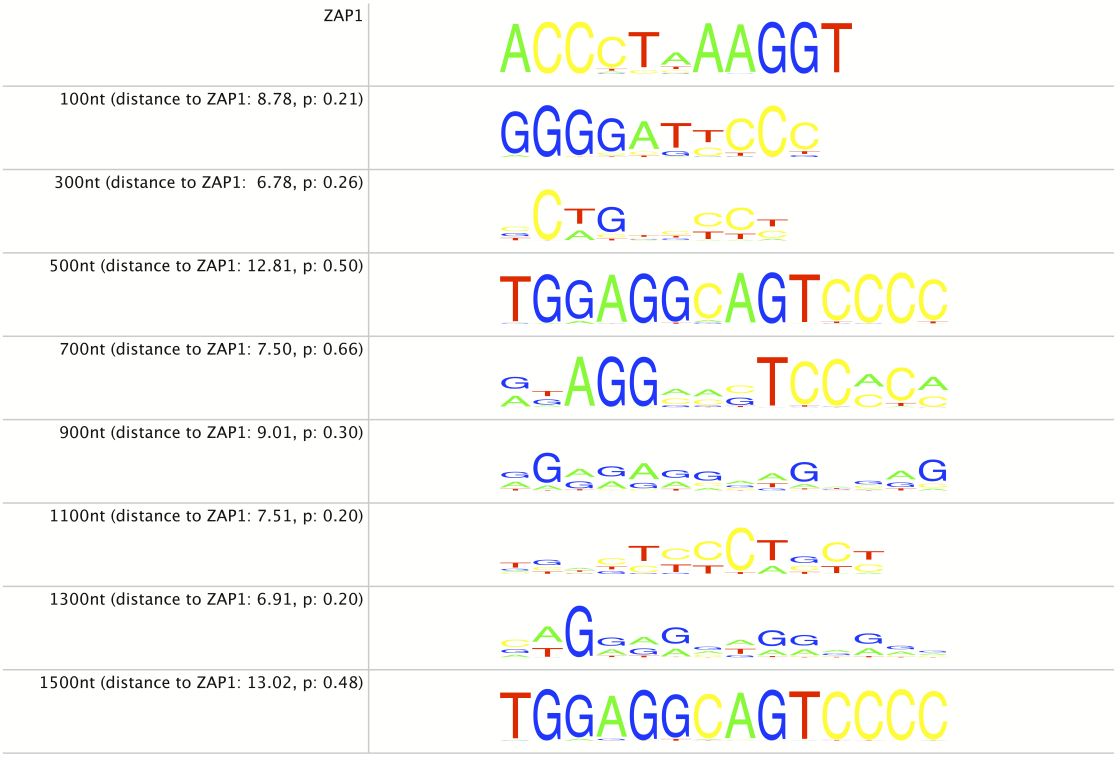

Figure 3.4: The closest motif match to the invalid motif pattern (ZAP1) shown alongside the ZAP1 motif. No pattern like ZAP1 should be seen, and indeed is not seen. Five motifs were inferred at each sequence length (100nt, . . .,1500nt).

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 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

[1] The 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 -minLenth 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.

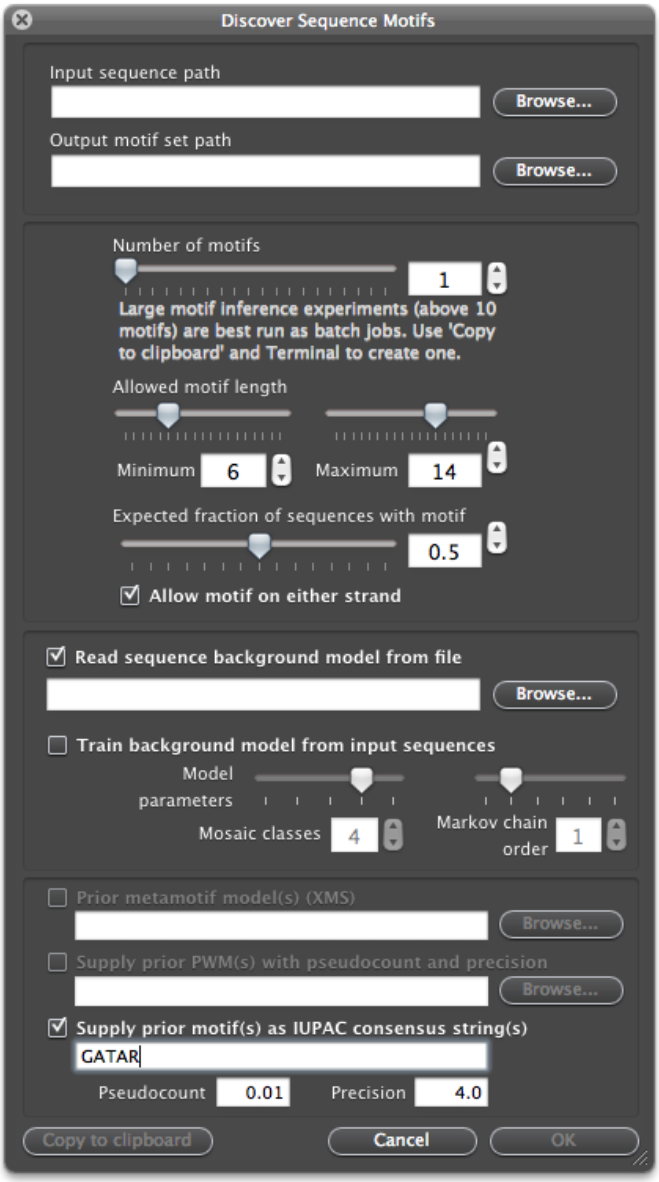

Figure 3.5: A NestedMICA motif inference run can be configured and run directly in iMotifs. Alternatively the NestedMICA run can be configured in iMotifs (Analysis >Discover Motifs from Sequence) and executed in the terminal after using the 'Copy to clipboard' function. A metamotif prior with one or 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.