Stochastic algorithms for motif discovery: a comparison of sampling strategies

Alastair M. Kilpatrick and Stuart Aitken*
School of Informatics, University of Edinburgh

April 26, 2013

TFBS motifs are short DNA sequence patterns that have important roles in gene transcription and regulation. Discovery of these sequences remains an important task in the wider challenge of understanding the mechanisms of gene expression; consequently, there is much continuing interest in developing algorithms to computationally discover TFBS motifs.

The EM algorithm [4] is the basis of a number of algorithms for motif discovery (most notably the popular MEME algorithm [1]). However, it suffers from several well-known limitations: it is strongly dependent on its initial position and can converge to a saddle point of the likelihood function rather than a local maximum. A stochastic version of the EM algorithm has been shown to alleviate these limitations in theory [3] and has been implemented in a motif discovery context by Bi (using the OOPS, or One Occurrence Per Sequence model) as the SEAM algorithm [2]. In this study we compare a Metropolis independence sampler with the roulette wheel selection used in SEAM in order to evaluate the potential performance benefits and computational cost: the motivation for this study is to determine if it is possible to reduce the running time of the algorithm by designing a strategy where samples could be drawn from an input sequence without having to evaluate the probability of each possible motif start site being an occurrence of the motif. The correctness of the recovered motifs is assessed using the standard measures of site-level sensitivity (sSn) and positive predictive value ($sPPV$).

Background

The idea underlying SEAM is to replace the computation and maximisation of the expected complete-data log likelihood function by the much simpler estimation of the posterior distribution for each input sequence, simulating a 'pseudo-sample' from this distribution and updating the model parameters based on the pseudo-complete samples [2]. This method is equivalent to the weighted 'roulette wheel selection' (sometimes known as 'fitness proportionate selection') method in genetic algorithms. Having sampled each input sequence, a proposal model is constructed from the samples and the current model updated to the proposal model if the Metropolis ratio is satisfied [2].

Using this method, the probability that a given position j in input sequence i is a motif occurrence ($Z_{i,j}$) must be enumerated for every position in i at every EM iteration in order to calculate the density. This requires considerable computation and may be inefficient, especially at later EM iterations when the majority of $Z_{i,j}$ values are expected to be near zero. This motivates the current study: is it possible to sample from an input sequence without having to evaluate $Z_{i,j}$ at every position? One potential solution is to use Markov Chain Monte Carlo (MCMC) to sample from our input sequence.

Method and Results

The simplest MCMC strategy (Metropolis algorithm) uses an independence sampler as the proposal distribution; this simplifies the calculation of the acceptance probability. Clearly, this method is only an improvement on the roulette wheel selection method if the cost of drawing k samples is substantially smaller than the cost of evaluating $Z_{i,j}$ at every possible motif start site. It is well known that the

*AMK is supported by an EPSRC Doctoral Studentship. AMK and SA are funded by BBSRC grant BB/I023461/1 (Bayesian evidence analysis tools for systems biology).

Metropolis algorithm with independence sampler can be shown to converge to a target distribution when this distribution is well-behaved. While analysis of the posterior distribution for a given input sequence shows that this distribution is not well-behaved at all, we have shown that this general result holds true in the context of motif discovery for large k.

A modified version of SEAM was implemented, replacing the roulette wheel selection method with the Metropolis independence sampler for each input sequence. The Metropolis independence sampler was implemented within SEAM, replacing the roulette wheel selection method for each input sequence. Overall performance was assessed by running the modified SEAM algorithm with 1,000 random seeds, choosing the best result based on the motif energy function provided by Bi and calculating the site-level sensitivity (sSn) and site-level positive predictive value ($sPPV$) for the corresponding motif model. Bi's motif energy function is related to the sequence binding or structural configuration free energy, widely used in motif discovery algorithms [2].

Both the original roulette wheel and modified SEAM algorithms were tested on a small collection of datasets containing previously characterised *E. coli* TFBS motifs extracted from the RegulonDB database. Initial tests with $k = 1,000$ (around five times the number of possible motif start sites) returned similar results to the roulette wheel selection method, showing the Metropolis independence sampler converging to the target distribution. In some cases, the maximum value of the motif energy function was increased when using the independence sampler (i.e. the output motif model was stronger), giving a corresponding improvement in sSn and $sPPV$. While this improvement is encouraging, the main disadvantage of this result is that drawing 1,000 samples from each input sequence takes substantially longer than simply enumerating every position and drawing a sample from the roulette wheel. It is clear that the next step is to investigate whether this trend continues when k is decreased.

In tests on a single input sequence, the Metropolis independence sampler with smaller k shows relatively poor convergence. However, it may still give reasonable results when applied to the SEAM algorithm, as SEAM takes a sample from each input sequence and takes the consensus of all samples in order to form a new proposal model. It follows that even if the chosen sample for a single input sequence is relatively poor, this may be alleviated by the chosen samples from other input sequences. It is possible that the independence sampler still allows the stochastic EM algorithm at the heart of SEAM to converge, albeit at a slower rate than before.

Further tests were carried out with $k = 200$ (i.e. around the number of possible motif start sites) and $k = 20$ (i.e. around 0.1 of the number of possible motif start sites). In addition, the number of EM iterations was varied in order to determine whether increasing this would improve situations with fewer MC samples. The results of these tests show that, overall, as k is reduced, the maximum value of the motif energy function decreases (i.e. the output motif model becomes weaker), often reducing sSn and $sPPV$ as the number of true positive site predictions decreases.

Table 1 illustrates some of the results of the comparison of sampling strategies. In the case of the Ada motif, the Metropolis independence sampler improves the sSn and $sPPV$ of a motif which was not discovered well by the roulette wheel sampling method. This test also illustrates a slight increase in motif energy; this increase is also noted in other datasets. In the case of the MetR motif, while the sSn and $sPPV$ results for the Metropolis method with large k match those for the roulette wheel sampling method, this performance decreases as k is decreased. In both cases, as k decreases, the maximum motif energy also decreases. Our results also show that for k greater than the number of possible motif start sites, increasing the number of EM iterations may slightly increase the motif energy of the result. However, for small k, the overall result is poor and increasing the number of EM iterations makes little difference to the maximum motif energy (there is very little improvement over randomly choosing motif positions within the dataset). While increasing the number of EM iterations may lead to a small improvement in the mean motif energy over 1,000 random seeds, this improvement is not enough to offset the effect of reducing k.

Conclusions

Although the Metropolis algorithm with independence sampler is a relatively simple sampling strategy, this approach is shown to give surprisingly good recovery of motifs based on site-level sensitivity and

Motif	Ada				MetR			
Method	Roulette	Metropolis			Roulette	Metropolis		
MC samples	-	1000	200	20	-	1000	200	20
EM iterations	-	500	1500	2000	-	500	1500	2000
sSn	0.00	0.25	0.25	0.25	0.71	0.71	0.14	0.00
$sPPV$	0.00	0.25	0.25	0.25	0.71	0.71	0.14	0.00
Motif energy	-24.03	-23.17	-30.02	-45.80	-43.94	-47.52	-72.23	-84.72

Table 1: Results of sampling strategy comparison for two *E. coli* TFBS motifs.

positive predictive value. Implementing this approach and using large numbers of Monte Carlo samples is also shown to often return stronger motif models, based on Bi's motif energy function. We note the high computational cost of drawing large numbers of samples using the independence sampler, however its performance in this study indicates the potential in exploring alternative sampling strategies as replacements for the roulette wheel method.

References

[1] T. L. Bailey and C. Elkan. Unsupervised learning of multiple motifs in biopolymers using Expectation Maximization. *Machine Learning*, 21:51–80, 1995.

[2] C. Bi. SEAM: a stochastic EM-type algorithm for motif-finding in biopolymer sequences. *Journal of Bioinformatics and Computational Biology*, 5(1):47–77, 2007.

[3] G. Celeux, D. Chauveau, and J. Diebolt. On stochastic versions of the EM algorithm. *Rapport de Recherche-Institut National de Recherche en Informatique et en Automatique*, 1995.

[4] A. Dempster and N. Laird. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.