A Comparison Of Expectation Maximization and Gibbs Sampling Strategies for Motif Finding

Michele Banko

December 13, 2004

1 Introduction

A set of protein or nucleotide sequences may be found to share patterns reflecting biological structure, function and change. The task of identifying these patterns, known as motif finding, can be viewed as an instance of multiple sequence alignment. While it is possible to identify motifs using x-ray and magnetic resonance structures, biologists and computer scientists have developed several algorithms for automatic pattern discovery from unlabeled sequences. We will study two such well-known approaches, Expectation Maximization and Gibbs Sampling, by addressing the following questions:

Performance

- How well can each method find the optimal solution?
- How sensitive is each method to different initializations?
- How long does the algorithm take to converge to a solution?

Robustness

- How well can each method cope with noisy data?
- With small training sets?

1.1 Expectation Maximization

Expectation Maximization is a widely-used unsupervised learning technique that attempts to maximize the probability of a model given a series of observations. Lawrence and Reilly[2] derived an application of EM to motif finding in which we are given a set of biopolymer sequences each containing an instance of a motif of length w. Bailey and Elkan[1] later relaxed this assumption to allow for the possibility of zero or more occurrences.

The high-level idea behind the algorithm is the following: if we had a weight matrix model (WMM) representing the motif, then we could easily scan the sequences to find an alignment. Conversely, if we had an alignment, we could compute the WMM. By computing these models in alternation over a series of iterations, EM maximizes the expected log likelihood over the probability of a motif starting at a given position, given the observed sequences and the current estimate of the WMM. While with EM there is always the danger of converging to a sub-optimal alignment, the particular implementation we use heuristically tries to escape such optimum via systematic searches of possible starting points.

1.2 Gibbs Sampling

The Gibbs motif sampler described by Lawrence et al.[3] assesses potential alignments in an attempt to find the best alignment according to the maximum a posteriori log-likelihood ratio. Given an initial alignment of sequences $s_1, s_2, ..., s_n$, the algorithm proceeds to search for a motif of length w as follows, until it reaches convergence:

1. Choose a sequence s_i at random.

2. Build a WMM from the remaining $n - 1$ sequences.

3. For each position j in s_i, let m_j be the substring of length w that begins at j. Compute the likelihood that m_j was generated by the motif.

4. Weight each substring of length w according to this likelihood and choose a new sequence at random to replace s_i

The Gibbs sampler takes a stochastic approach, therefore is not guaranteed to converge to the same alignment with different random seeds. One may wish to run the sampler several times with the hope of eventually obtaining a satisfactory alignment.

2 Tools and Evaluation

In order to assess the performance of EM relative to the task of multiple sequence alignment, we made use of the MEME toolkit[1] that is available from SDSC [1]. The Gibbs Motif Sampler, which has been made available by Jun Liu[3],[2], was used in our experiments concerning Gibbs sampling. Both toolkits were freely available and no modifications were made to the source code.

We measure performance using site-level precision (P), recall (R) and F-measure (F), which are defined as $P = \frac{TP}{TP+FP}$, $R = \frac{TP}{P}$, $F = \frac{2*P*R}{P+R}$. Here, TP, is the number of correctly found instances of the motif (e.g. "true positives"), FP, is the number of patterns wrongfully classified as being an instance of the

[1] http://meme.sdsc.edu/meme/website/intro.html.

[2] http://www.people.fas.harvard.edu/ junliu/index1.html

Factor	Num Sequences (mouse/human)	Avg Seq Length	Motif Length	Swissprot Precision	Swissprot Recall
MYB 1	30 (12/18)	895	8	0.4144	0.8333
CYTOCHROME P450	90 (40/50)	535	9	0.9711	0.9492
ZINC PROTEASE	149 (62/87)	838	9	0.7625	0.8313
ZF RING 1	132 (52/80)	598	9	0.9921	0.5578

Table 1: Summary of datasets

motif (e.g. "false positives"), and P is number of known instances of the motif as designated by the Prosite corpus.

For each motif that is discovered automatically, we evaluate precision and recall based on the one possessing the highest recall. We compare each found motif relative to at most $\frac{w}{2}$ shifts to both the left and right of the site, w being the width of the motif we are looking for.

3 Data

For our experiments, we used PROSITE[4] to extract sequences containing 4 known transcription factors present in both the mouse and human species:

Myb 1, a retroviral oncogene, which has been implicated in regulation of the cell cycle.

Cytochrome P450, a group of enzymes involved in the metabolism of natural compounds such as steroids and fatty acids, as well as drugs, carcinogens and mutagens.

Zinc protease, a zinc-binding region signature, part of the family of neutral zinc metallopeptidases.

ZF Ring 1, a zinc finger RING-type signature.

These factors were chosen based on their ability to represent a variety of properties: small number of samples (MYB 1), large number of known false positives (MYB 1), large number of false negatives (ZF RING 1). We also chose several motifs having the same length so that we could later test the algorithms ability to simultaneously discover different motifs having the same width. Finally, we note that one limitation of the MEME implementation is its inability to find occurrences which may possibly contain insertions and deletions. Therefore, all four of the known motifs that were chosen do not contain any gaps. Table 1 summarizes the the data used for our experiments.

	Gibbs			MEME		
Dataset	Precision	Recall	Shift	Precision	Recall	Shift
MYB 1	0.9333	0.9333	0	0.9333	0.9333	0
CYTOCHROME P450	0.9778	0.9778	1	0.9778	0.9778	1
ZINC PROTEASE	0.9195	0.9195	3	0.9933	0.9933	0
ZF RING 1	0.9848	0.9848	0	0.9848	0.9848	0

Table 2: Best Results for Gibbs and MEME

4 Results

4.1 Performance Analysis

As a first round of evaluations, we ran each algorithm on each of our four datasets without modifying any of the default parameters. For both approaches we found that each algorithm's performance improved after shifting the site of the found patterns by up to three places. As expected, we found that the Gibbs sampler did not always converge to the same result, which will be further examined in the following section. Table 2 presents the best motif eventually found by each algorithm, as measured by precision and recall.

We also examined the time it took each algorithm to find the best alignment when the width of the pattern we seek to learn is known[3], and present the results in Table 3. While the Gibbs sampling method consistently converges faster than EM, the need to run several trials of the Gibbs sampler confounds this potential advantage. Since in our case, MEME is completes its run in a few minutes and not hours, the relative speed of Gibbs is not a clear advantage.

Initialization Running the Gibbs implementation off-the-shelf without any parameter tweaking, the Gibbs approach performed quite poorly over a variety of random starts for the ZINC PROTEASE motif, and was shown to be quite sensitive to initialization in general. As recommended by the authors, we found that increasing the number of independent searches from a given starting point improved our chances of picking a better alignment. Figure 1 shows results for

[3] The Gibbs Motif Sampler does not automatically detect motifs of arbitrary length.

Dataset	Gibbs	MEME
MYB 1	2	6.55
CYTOCHROME P450	5	33.04
ZINC PROTEASE	45	225.95
ZF RING 1	2	100.23

Table 3: Time to compute best alignment (in seconds) for Gibbs and MEME

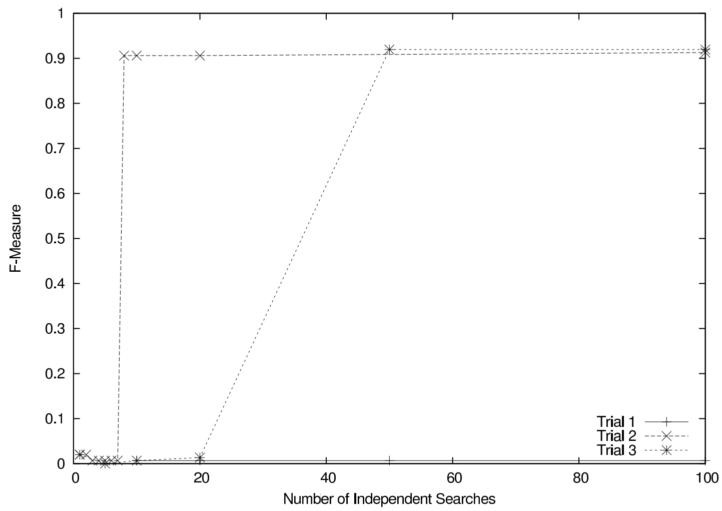

Figure 1: F-measure for the ZINC PROTEASE motif over 3 independent trials

three of several trials we performed, selecting a different random seed for each trial, and increasing the number of searches made by the algorithm.

On the other hand, MEME was shown to be quite stable with respect to initialization schemes. We took the case in which it performed the worst, and experimented with the starts that were possible using the toolkit. These include varying the fuzziness of the mapping function, and overriding the sampling of possible starting points by specifying a substring from which to start that is based on our knowledge of the actual motif. In all cases, modifying the initialization of the algorithm did not alter results.

4.2 Robustness Analysis

Simultaneous Motif-Finding In order to test each algorithm's ability to locate several motifs at the same time, we created one dataset consisting of the CYTOCHROME P450, ZINC PROTEASE and ZF RING 1 sequences, all of which contain motif instances of length nine. In each case, we guided the algorithms by specifying how many instances of each motif were to be expected.

We ran several trials of the Gibbs implementation, specifying 1, 10, 100 and 500 independent searches, and found that at best, it could model only one out of three possible motifs. MEME fared much better, finding three motifs at once, each with high precision and recall. Results are summarized in Tables 4 and 5.

Small Sample Sets Bailey and Elkan[1] claim that their experiments with MEME have shown that EM is capable of accurately detecting motifs even when as few as 20% of the sequences contain the pattern. To test how well

Searches	Found Motif	Known Motif	Precision	Recall
1	MOTIF C	cytochrome p45	0.0526	0.0111
	MOTIF C	zinc protease	0.0294	0.0076
10	MOTIF A	cytochrome p45	0.2308	0.0333
100	MOTIF A	zinc protease	0.4809	0.4773
500	MOTIF A	zinc protease	0.4809	0.4773

Table 4: Simultaneous Motif-Detection with Gibbs Sampling

Found Motif	Known Motif	Precision	Recall
MOTIF 1	zf ring 1	0.9847	0.9773
MOTIF 2	zinc protease	0.9851	0.8859
MOTIF 3	cytochrome p45	1.0000	0.9556

Table 5: Simultaneous Motif-Detection with MEME

each approach handled various datasets where only a portion of the sequences contained known instances of the zinc protease motif, we constructed datasets as follows. Randomly select 5% of the sequences containing occurrences of the motif. Select the remainder of the sequences at random from the mouse and human genome, keeping the entire size of the dataset fixed at the original size. Run the algorithm. To build a dataset containing only 10% known occurrences, select another 5% from the original sequences, ensuring no overlaps with the previous set, add it to the previous set of 5%, and select the remaining 80% at random from the total genomes. We carried out this procedure for percentages up to 20%.

We found that MEME was unable to find any instances of the motif, even when we specified how many sites we should expect to contain an instance of the motif. Using the best seed value from the previous 3 trials, Gibbs had at best a precision of 0.1250 and recall of 0.1429, which came when seeing only 5% of actual occurrences.

5 Conclusions

In general, both the EM and Gibbs-sampling approaches to motif finding made it possible to detect non-gapped motifs in a matter of minutes. For the known factors we assessed, both methods were able to obtain high levels of precision and recall. While the Gibbs approach has the advantage of being slightly faster, its susceptibility to its starting state and the number of searches from that initial state may be a cause for concern. In contrast, EM provided a more stable approach that was not much more time-consuming in terms of search time or required restarts. This method was also shown to be able to better learn more than more motif at a time.

References

[1] Timothy L. Bailey and Charles Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 28–36, Menlo Park, California, 1994. AAAI Press.

[2] C.E. Lawrence and A.A. Reilly. An expectation maximization(em) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *PROTEINS: Structure Function and Genetics*, 7:41–41, 1990.

[3] Charles E. Lawrence, Stephen F. Altschul, Mark S. Boguski, Jun S. Liu, Andrew F. Neuwald, and John C. Wootton. Detecting subtle sequence signals: A gibbs sampling strategy for multiple alignment. *Science, New Series*, 262(5131):208–214, 1993.

[4] C. Sigrist. Prosite: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform*, 3:265–274, 2002.