Empirical results on initialization schemes for MEME

Kush Dubey (kd356)

December 17, 2019

This project reimplements the MEME algorithm for motif finding, and studies how different initialization schemes affect biological accuracy, statistical accuracy, and computational performance on small datasets.

1 Introduction

Motifs are short subsequences of biopolymers that repeatedly appear across sequences of a genome. Their higher-than-expected frequency usually implies biological function, such as the regulation of gene expression. Detecting motifs among a given set of sequences is made computationally challenging by the fact that motif instances are often degenerate—meaning their exact biopolymer sequence is indeterminate. For this reason, geneticists commonly employ probabilistic tools that model the random motif generation process given an unlabeled set of sequences.

Two of these tools include Gibbs sampling and the mixture model. Gibbs sampling is a Markov Chain Monte Carlo algorithm that samples from a probability distribution that increasingly resembles that of the motif after sufficient iterations. Mixture models, on the other hand, directly imitate the underlying process by learning parameters of a probability distribution that maximize the likelihood of the given dataset. While both techniques have been shown to discover biological structure (Bailey et al. 1994, Lawrence et al. 1993) and each have their advantages, this project focuses on the mixture model.

The Multiple EM for Motif Elicitation (MEME) implementation of the mixture model framework is particularly practical due to the fact that it doesn't require each input sequence to contain a motif instance (Bailey et al. 1994). It works by assuming that all bases in a W-mer (where W is a user-specified argument determining the length of motifs to search for) are generated by a *mixture* of a motif model and a background (i.e. non-motif) model rather than one or the other.

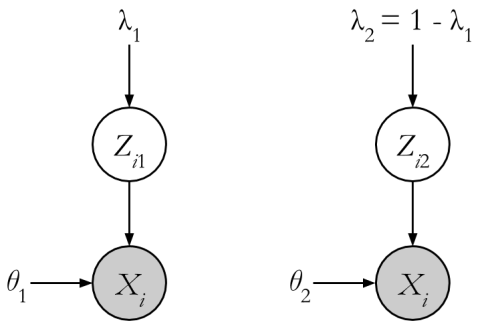

Figure 1: Two-component mixture of multinomial distributions. Random variables X_i (observed) and Z_{ij} (hidden) are parameterized by θ_j and λ_j respectively. Left: motif model. Right: background model.

1.1 Mixture Model (MM)

The motif and background models are multinomial distributions each with their own parameterization. $\theta_1 \in \mathbb{R}^{4W}$ corresponds to the motif model, as it contains probabilities f_{wa} for each position w in the W-mer[1] and each base $a \in \{A, C, G, T\}$. $\theta_2 \in \mathbb{R}^4$ corresponds to the background model, as it contains probabilities f_{0a} for each base a that uniformly apply to all positions in the W-mer.

Z_{ij} for $j \in \{1, 2\}$ is the hidden variable denoting whether or not the i'th W-mer was generated by the motif model, $Z_{i1} = 1 \equiv Z_{i2} = 0$, or the background model, $Z_{i1} = 0 \equiv Z_{i2} = 1$. The mixture model density function and its equivalent graphical interpretation in Figure 1 are given below for arbitrary W-mer X_i, as these expressions aren't explicitly provided in Bailey et al. (1994).

$$\begin{aligned}
\Pr(X_i \mid \theta, \lambda) &= \sum_{j=1}^{2} \Pr(X_i, Z_{ij} = 1 \mid \theta, \lambda) && \text{total probability} \\
&= \sum_{j=1}^{2} \Pr(Z_{ij} = 1 \mid \theta, \lambda) \Pr(X_i \mid Z_{ij} = 1, \theta, \lambda) && \text{Bayes rule} \\
&= \lambda_1 p(X_i \mid \theta_1) + \lambda_2 p(X_i \mid \theta_2) && \text{defns} \\
&= \lambda_1 \prod_{w=1}^{W} \prod_{a} {f_{wa}}^{I(X_{iw}=a)} + (1 - \lambda_1) \prod_{w=1}^{W} \prod_{a} {f_{0a}}^{I(X_{iw}=a)} && \text{defn of multinomial}
\end{aligned}$$

The final estimates of interest are the posterior probabilities $\Pr(Z_{i1} = 1 \mid X_i, \theta, \lambda)$, as these inform us of the degree that X_i is a motif. Deriving these estimates

[1] the term W-mer is used interchangeably with "subsequence"

requires that multinomial distribution parameters $\theta = (\theta_1, \theta_2)$ and prior probabilities (i.e. mixing proportion parameters) $\lambda = (\lambda_1, \lambda_2)$, be learned based on all of the subsequence data $X = (X_1, X_2, \ldots, X_n)$. Bailey et al. (1994) formulate the learning problem by solving the usual likelihood objective:

$$\text{maximize} \quad L(\theta, \lambda \mid X, Z) \tag{1}$$

(1) is solved via expectation-maximization (EM) due to the presence of hidden variables Z. In short, the Expectation-step or E-step computes posterior probabilities based on an initial set of parameters, and the Maximization or M-step locally maximizes the expected log-likelihood (ELL) of these parameters by setting derivatives to 0. These steps form an iterative cycle that is set to terminate when the ELL does not increase enough. Mathematical expressions for the E-step and M-step are given in Bailey et al. (1994) and are implemented in this project.

1.2 Goal

A neat property of the EM algorithm, like all gradient descent algorithms, is that the ELL will never decrease. In other words, progress towards termination is guaranteed. This is visually confirmed in Figure 3.

From an optimization standpoint, the pitfall of this strategy is that only convergence to a stationary point (where the gradient is 0) is guaranteed. These points are indistinguishable from saddle points, which fit and generalize very poorly in the high-dimensional, non-convex setting (Dauphin et al. 2014).

This is a fundamental limitation to probabilistically modeling motif generation according to Bailey et al. (1994) and Lawrence et al. (1993). In fact, important extensions of MM that aim to estimate the number of unique motifs in the dataset induce an even higher-dimensional search space. This makes avoiding the pitfall even more relevant.

The observations above motivate closer study of improving the convergence of MM through probabilistic and biological heuristics. This project specifically assesses how well three different initialization schemes for MM perform in terms of biological accuracy, statistical accuracy, and computational cost.

2 Experiments

The first experiment reimplements the precise initialization scheme for the motif model parameters θ_1 from Bailey et al. (1994), and compares it to a naive but efficient baseline. The second experiment simply initializes θ_2 by estimating base probabilities using the whole dataset.

Two small vertebrate profiles with JASPAR IDs MA0006.1 and MA0259.1 were retrieved (Fornes et al. 2019). Next, sequences were split into two datasets corresponding to two unique motifs for each profile, which made for a total of four sets of sequences and motifs.

Biological accuracy was indicated by whether or not the consensus motif—computed based on the top 10 motifs found by MM in the dataset—matched the biological motif labeled in the dataset. Statistical accuracy was measured using the ELL, and computational cost was measured by the total wall-clock time to train MM under each initialization scheme.

But before delving into the experiments, it's important to clarify which parts of MEME this project does not implement and why. First, it doesn't implement erasure because the goal is to broadly explore the per-iteration (i.e. per-motif finding) capabilities of MM. This limitation is ineffectual due to the preprocessing method described above; each dataset is a set of sequences sharing only one biological, recurring motif. Second, re-normalization (which handles overlapping, statistically dependent subsequences) turned out not to improve either the biological or statistical accuracy through experimentation due to the minimal presence of repeating sequences. So this method was left out to improve run-time. Third, the hyperparameter W (which determines the length of motifs to search for) was pre-determined to be the length of true motifs in order to allow for simpler comparisons. A fourth and minor note is that this project's EM implementation terminates when ELL is close (within 10^{-2} units) to that in the previous iteration. This differs from the one described in Bailey et al. (1994) where convergence is determined by distance between parameters rather than the objective value. This discrepancy is presumably ineffectual.

The big, messy module `meme.py` contains all the code used to generate summary statistics in Tables 2 and 3 and Figure 3. The script is run using commands of the form:

```
python meme.py -f data/MA0006.1/MA0006.1-motif1.sites
               -W 6
               -init plain
               -backg unif
```

All experimental results in this project can be reproduced by running commands listed at the bottom of the module.

2.1 Experiment 1: θ_1 (motif) initialization

This section explains in some detail how the "smart" θ_1 initialization described in Bailey et al. (1994) is performed. Results are presented in Table 2 and Figure 3, and analyzed in the Discussion section.

The underlying assumption of this mixture model is that motifs can be discriminated from the background because they're generated from a "different enough" multinomial model. Close-to-optimal solutions for datasets containing distinguishable motifs will roughly satisfy $f_{wa} \neq 1/4$, where $1/4$ is the uniform probability of observing a base. So a good initialization for the motif base probabilities, θ_1, is one that makes it sufficiently different from the uniform distribution. Relative entropy (RE) is one way to formalize this difference.

But the θ_1 initialization should not only be different enough from $1/4$; it

θ_1	A	C	G	T
1	m	m'	m'	m'
2	m'	m'	m	m'

Table 1: Initialization for $W = 2$ and $X_i = (A, G)$.

should also faithfully represent the probabilities of observing an actual motif in the dataset. Otherwise, θ_1 will start far from the optimum that discriminates true, frequent motifs from the background.

Bailey et al. (1994) simplify the estimation problem by (1) uniformly estimating entries f_{wa} for $w \in [W]$ and $a \in X_i$ (i.e. the probabilities of the W observed bases in a presumed motif X_i) as one parameter m, and (2) assuming the rest of the bases are equiprobable with parameter $m' = (1 - m)/3$. This makes the problem easy, as the degrees of freedom are reduced from $3W$ to 1.

The toy example in Table 1 more concretely illustrates the initialization scheme of θ_1. The goal of is to estimate m. m should ideally be greater than $1/4$ if X_i is a motif, and should be closer to 1 for motifs that are less degenerate.

m is found by solving the root-finding problem in (2) below, substituting $f_{wa} = m$ if $a = X_{iw}$, otherwise $f_{wa} = m'$. $\gamma \in [0, 1]$ is a relatively sensitive hyperparameter reflecting the user's prior belief on the degeneracy of motifs in their dataset. If motifs are believed to be non-degenerate, then γ should be set high to reflect the motif model's high divergence from the uniform distribution. But setting γ too high prevents truly degenerate motifs from being discovered.

$$\underbrace{\sum_a f_{wa} \log \frac{f_{wa}}{1/4}}_{\text{RE wrt uniform}} = \gamma \overbrace{\left(-\log \frac{1}{4} \right)}^{\text{max RE}} \tag{2}$$

All of the analysis above assumes that X_i is a representative motif. But how do we know a subsequence is a motif in the unsupervised learning setting? Bailey et al. (1994) resolve this by double-dipping into the assumption that parameters initialized according to true motifs will optimally discriminate (i.e. have high ELL) even at the start of EM.

So now we can roughly determine what is and isn't a motif. But how do we quickly search for and encounter motifs in our dataset? Here Bailey et al. (1994) provide a bound that a motif will be selected with at least α probability if at least $Q = \frac{\log(1-\alpha)}{\log(1-\lambda_1)}$ subsequences are randomly drawn from the dataset without replacement. The quality of the bound depends on the estimation quality of λ_1—far more samples than necessary will be drawn if λ_1 is an underestimate of the true proportion of motif instances in the dataset and vice versa.

The overall algorithm implemented in this project is depicted in Figure 2. It starts with a search over values for λ_1 and then initializes θ_1 using the method described above.

The baseline (equivalently referred to as "plain" initialization) sets $\lambda_1 = 1/N$

for $\lambda^{(0)} = \frac{\sqrt{N}}{n}$ **to** $\frac{1}{2W}$ **by** $\times 2$ **do**
  for $j = 1$ **to** Q **do**
    Randomly (w/o replacement) select a subsequence X_j
    from dataset Y.
    Derive $\theta^{(0)}$ from subsequence X_j.
    Estimate goodness of ($\theta^{(0)}$, $\lambda^{(0)}$) as starting point for MM.
  end
  Run MM to convergence from best starting point found above.
end
Print best motif found above: $\hat{\theta}$, $\hat{\lambda}$.

Figure 2: Subroutine of MEME+ algorithm that is implemented in this project. Taken from Figure 1 in Bailey et al. (1994).

where N is the number of sequences, $\lambda_2 = 1 - \lambda_1$, and $f_{wa} = f_{0a} = 1/4$ for all w and all a. In other words, the baseline assumes that (1) every sequence contains one motif and (2) θ_1 and θ_2 are no different, as all prior knowledge is ignored. Table 2 contains the results for the motif initialization experiment.

2.2 Experiment 2: θ_2 (background) initialization

The promising results in Experiment 1 suggest initializing background probabilities in a more data-based way. One way to do this is to assume that all observed bases are sampled from a multinomial distribution. Then the maximum likelihood estimate (MLE) of parameter f_{0a} for all a is just:

$$f_{0a} = \frac{1}{M} \sum_{m=1}^{N} \sum_{p=1}^{l_m} I(Y_{mp} = a)$$

This step requires $O(M)$ time where M is the total number of bases in dataset Y. Clearly $M < n$, so this preprocessing step has no asymptotic effect on the MEME algorithm.

The multinomial model is a reasonable assumption because of two observations. The first is just that the θ_2 that EM converged to for MA0006.1-motif1 was $(0.11, 0.27, 0.34, 0.29)$ while the MLE estimates are $(0.12, 0.25, 0.37, 0.26)$. The second is that nucleotide frequencies are usually non-uniform. For example, the human genome consists of 42% GC content (Lander et al. 2001).

The estimation procedure based on the full dataset is unbiased only if it contains no motifs. Although it's close enough as long as motifs are sparse. In practice this may be a good assumption for large sequence data. Table 3 contains the results for the background initialization experiment.

Dataset	W	N	Instances	Plain found	Smart found	Plain ELL	Smart ELL	Plain time	Smart time
MA0006.1-motif1	6	8	8	×	×	-463.5	-466.5	0.9	2.8
MA0006.1-motif2	6	11	11		×	-702.5	-629.0	0.4	8.5
MA0259.1-motif1	8	12	6		×	-1437.8	-1355.9	0.3	17.2
MA0259.1-motif2	8	12	6		×	-1413.6	-1323.3	0.3	16.5

Table 2: Results of θ_1 motif initialization scheme described in Experiment 1 above. The "found" columns indicate biological accuracy. The ELL columns measure statistical accuracy. The "time" columns measure computational cost. W = length of subsequence. N = number of sequences. "Instances" denotes number of motif instances. "Plain found" denotes that the consensus motif found by the plain initialization of θ_1 was the same as the true motif. ELL denotes expected log-likelihood. Time is measured in seconds. γ was set to 0.4 for the first three datasets, and hand-tuned to 0.3 for the last.

Dataset	W	N	Instances	Plain found	Smart found	Plain ELL	Smart ELL	Plain time	Smart time
MA0006.1-motif1	6	8	8	×	×	-477.9	-466.3	0.7	3.0
MA0006.1-motif2	6	11	11		×	**-692.3**	-629.0	0.4	8.7
MA0259.1-motif1	8	12	6	×	×	**-1355.7**	**-1350.9**	0.4	18.1
MA0259.1-motif2	8	12	6		×	**-1407.6**	-1322.5	0.3	17.8

Table 3: Results of θ_2 background initialization scheme described in Experiment 2 above. Entries where accuracy increased from Table 2 are bold. Note that a 1 unit increase in ELL corresponds to an e-fold increase in expected likelihood.

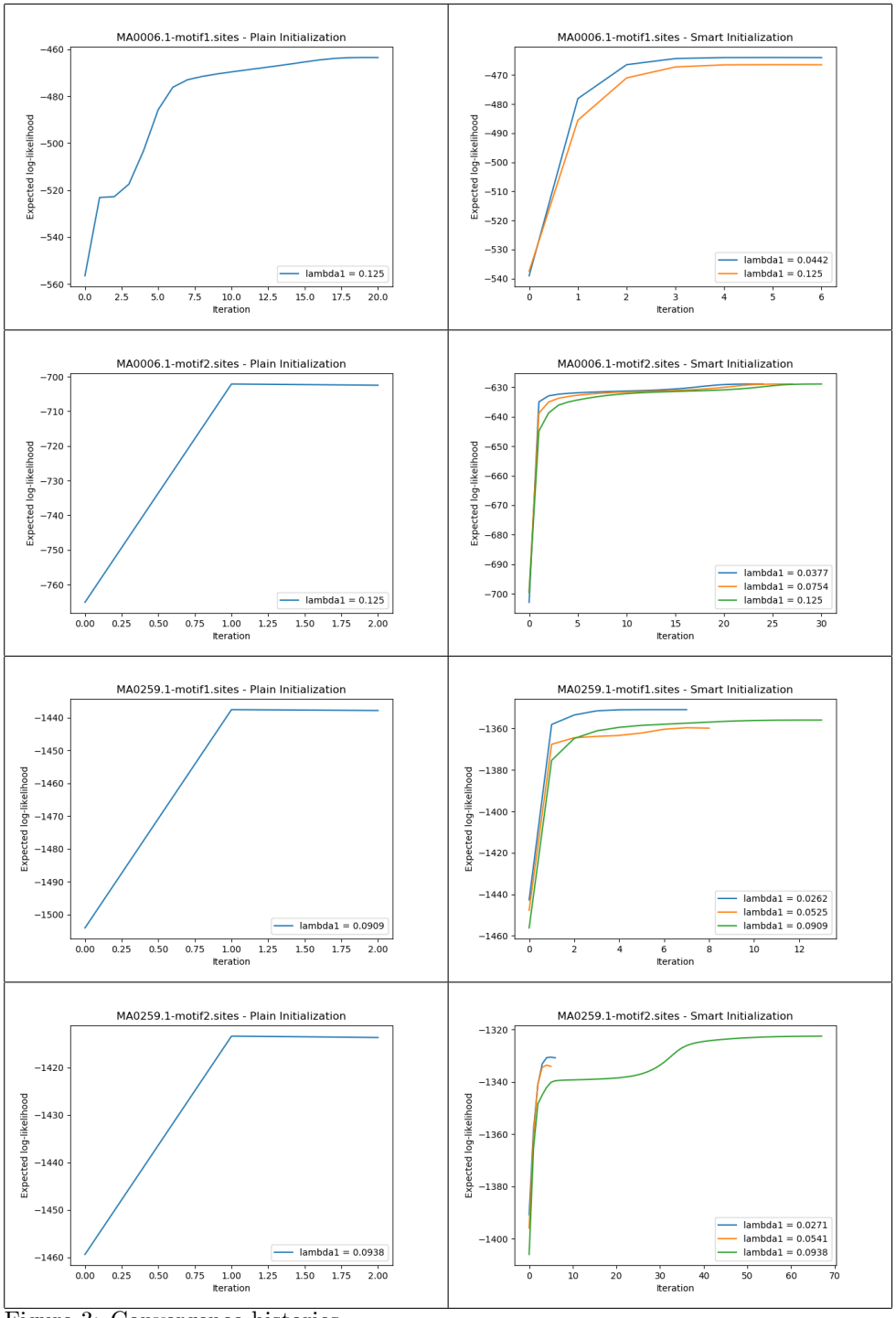

Figure 3: Convergence histories.

3 Discussion

A natural question to address is whether or not the sophisticated θ_1 initialization of trying many, specially selected initial parameters is worth the extra computational burden. The facts that (1) the EM objective is high-dimensional and non-convex (2) λ_1 is observed to be relatively high in all four datasets, and (3) randomly sampling from the training set will likely result in a representative initial θ_1 all suggest that the tradeoff between accuracy and efficiency will be positive. The results in Table 2 validate this hypothesis. There's only one case where the plain initialization leads to higher ELL. Investigations into the MA0006.1-motif1 dataset showed that while the smart initialization scheme successfully sampled instances of the true motif, the initial ELL of parameters based on that instance (-480.3) was less than that of other sampled subsequences (-473.4). This demonstrates the pitfall of the smart heuristic because forcing the worse initialization led to a solution with an ELL of -462.57, which is better than those in Table 2.

In all other cases, the final ELL of parameters found using the smart initialization are significantly higher[2] than those found by the plain one. Furthermore, the smart initialization correctly determines all four motifs, while the plain one only does so once[3]. However, the smart initialization could take 57 times as long even on these small datasets, as seen for MA0006.1-motif2.

Figure 3 validates that the data-based θ_1 initialization is better than the uniform θ_1 initialization: the initial ELL (i.e. at iteration 0) for every smart initialization is always greater than that of the plain one.

For the small datasets in this experiment, the background model was still expected to improve the statistical accuracy at little cost during preprocessing. Table 2 shows that these hypotheses are mostly confirmed. The background initialization unexpectedly worsened the performance for the plain θ_1 initialization in MA0006.1-motif1, despite the closeness between the plain MLE solution and the converged EM one. Other datasets mainly benefited from the θ_2 initialization. MM with the θ_2 initialization now finds the biological motif for the MA0259.1-motif1 dataset. Half of the ELLs are mostly unchanged or decreased slightly, which means the background initialization step is not as universally beneficial as the motif one.

3.1 Incorporating CpG island posterior probabilities

The work above ultimately demonstrates that estimating initial parameters allows MM to obtain slightly better statistical and biological results. These initialization schemes make little use of a biologist's prior knowledge on where motifs might be found, and instead rely on the dataset and inherent assumptions of the mixture model. This section briefly explores one situation in which biologi-

[2] not necessarily statistically significantly higher as the likelihood ratio test doesn't apply to ELLs

[3] for MA0006.1-motif1, the sequence TCGCGT also occurred 8 times and was also counted as a biological motif for this experiment

cal knowledge can be encoded into the model—specifically, the knowledge that CpG motifs are specifically being searched for or are common in some species (Hartmann and Krieg 2000).

The Hidden Markov Model presented in class provides posterior probabilities that a given position belongs to a CpG island. These probabilities could be treated as known, fixed parameters

$$z_{iw} = \Pr(\text{position } w \text{ in subsequence } i \text{ is in a CpG island} \mid Y)$$

and computed as a preprocessing step to MM in $O(4M)$ time. The subsequence density functions function would change to:

$$p(X_i \mid \theta_1, z_i) = \prod_{w=1}^{W} z_{iw} \prod_a {f_{wa}}^{I(X_{iw}=a)}$$
$$p(X_i \mid \theta_2, z_i) = \prod_{w=1}^{W} (1 - z_{iw}) \prod_a {f_{0a}}^{I(X_{iw}=a)}$$

The corresponding graphical interpretation would involve adding z_i parameters to node X_i in Figure 1.

In words, the multinomial distribution parameters for each position in X_i are weighed so that subsequences with high GC content probabilities will be more likely under the motif model and less likely under the background model. This would effectively limit the search space of motifs to areas of high GC content, and may result in faster convergence by enforcing greater initial discrimination between the motif and background models. However, no actual results are presented here. So claims about improvements are not yet valid.

References

Timothy L Bailey, Charles Elkan, et al. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. 1994.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pages 2933–2941, 2014.

Oriol Fornes, Jaime A Castro-Mondragon, Aziz Khan, Robin van der Lee, Xi Zhang, Phillip A Richmond, Bhavi P Modi, Solenne Correard, Marius Gheorghe, Damir Baranašić, et al. Jaspar 2020: update of the open-access database of transcription factor binding profiles. *Nucleic acids research*, 2019.

Gunther Hartmann and Arthur M Krieg. Mechanism and function of a newly identified cpg dna motif in human primary b cells. *The Journal of Immunology*, 164(2):944–953, 2000.

Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al. Initial sequencing and analysis of the human genome. 2001.

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*, 262(5131):208–214, 1993.