A Brief Overview of Gibbs Sampling

Eric C. Rouchka[1]
TR-ULBL-2008-02

March 24, 2008

[1]University of Louisville
Speed School of Engineering
Department of Computer Engineering and Computer Science
123 JB Speed Building
Louisville, Kentucky, USA 40292

eric.rouchka@louisville.edu

UNIVERSITY of LOUISVILLE dare to be great
Bioinformatics Laboratory Technical Report Series

Bioinformatics Review

A Brief Overview of Gibbs Sampling

Eric C. Rouchka[1,*]

[1]Department of Computer Engineering and Computer Science, University of Louisville, 123 JB Speed Building, Louisville, KY, USA

UNIVERSITY OF LOUISVILLE BIOINFORMATICS LABORATORY TECHNICAL REPORT SERIES REPORT NUMBER **TR-ULBL-2008-02**

ABSTRACT

Motivation: A number of techniques exist for the detection of short, subtle conserved regions within DNA and amino acid sequences. The purpose of this overview is to present the ideas of Gibbs sampling in terms of the data, parameters, model, and procedure both in a general sense and through an application of Gibbs sampling for multiple sequence alignment. This technical report was first presented at Washington University's Institute for Biomedical Computing Statistics Study Group in May of 1997. Since a number of individuals have found this overview helpful, it has been formatted as a technical report for further dissemination.

1 GIBBS SAMPLING

Gibbs sampling is a generalized probabilistic inference algorithm used to generate a sequence of samples from a joint probability distribution of two or more random variables (Casella and George, 1992). It is a variation of the Metropolis-Hastings algorithm (Hastings, 1970; Metropolis et al., 1953). The use of Gibbs sampling as a statistical technique was first described in 1984 (Geman and Geman, 1984). In the arena of bioinformatics, Gibbs sampling is one of several approaches to motif detection, including expectation-maximization approaches (Lawrence and Reilly, 1990). A number of modifications to the Gibbs sampler have been made (Newberg et al., 2007; Thompson et al., 2007a; Thompson et al., 2007b; Thompson et al., 2003; Neuwald et al., 1995; Liu et al., 1995; Lawrence et al., 1993) in order to more accurately detect subtle motif signals in multiple DNA and protein sequences. This technical report presents an overview of the Gibbs sampler with specific applications towards motif detection.

1.1 Gibbs Sampling Requirements

The first requirement for the Gibbs sampler is the observable data. The observed data will be denoted *Y*. In the general case of the Gibbs sampler, the observed data remains constant throughout. Gibbs sampling requires a vector of parameters of interest that are initially unknown. These parameters will be denoted by the vector Φ. Nuisance parameters, Θ, are also initially unknown. The goal of Gibbs sampling is to find estimates for the parameters of interest in order to determine how well the observable data fits the model of interest, and also whether or not data independent of the observed data fits the model described by the observed data.

Gibbs sampling requires an initial starting point for the parameters. Then, one at a time, a value for each parameter of interest is sampled given values for the other parameters and data. Once all of the parameters of interest have been sampled, the nuisance parameters are sampled given the parameters of interest and the observed data. At this point, the process is started over. The power of Gibbs sampling is that the joint distribution of the parameters will converge to the joint probability of the parameters given the observed data.

1.1 Explanation in Mathematical Terms

The Gibbs sampler requires a random starting point of parameters of interest, Φ, and nuisance parameters, Θ, with observed data *Y*, from which a converging distribution can be found. For the sampler, there is an initial starting point $(\Theta_1^{(0)}, \Theta_2^{(0)}, ..., \Theta_D^{(0)}, \Phi^{(0)})$. The steps a-d listed below are then repeatedly run:

a) Sample $\Theta_1^{(i+1)}$ from
$$p(\Theta_1 \mid \Theta_2^{(i)}, ..., \Theta_D^{(i)}, \Phi^{(i)}, Y).$$
b) Sample $\Theta_2^{(i+1)}$ from
$$p(\Theta_2 \mid \Theta_1^{(i+1)}, \Theta_3^{(i+1)} ..., \Theta_D^{(i)}, \Phi^{(i)}, Y)$$
c) Sample $\Theta_D^{(i+1)}$ from
$$p(\Theta_D \mid \Theta_1^{(i+1)}, ..., \Theta_{(D-1)}^{(i+1)}, \Phi^{(i)}, Y).$$
d) Sample $\Phi^{(i+1)}$ from
$$p(\Phi \mid \Theta_1^{(i+1)}, ..., \Theta_D^{(i+1)}, Y).$$

*To whom correspondence should be addressed.

The vectors $\Theta^{(0)}, \Theta^{(1)}, ..., \Theta^{(t)}$ represent the realization of a Markov chain, where the transition probability from Θ' to Θ is defined as:

$$K(\Theta', \Theta) = p(\Theta_1 \mid \Theta_2, ..., \Theta_D, \Phi, Y) * p(\Theta_2 \mid \Theta_1, \Theta_3 ..., \Theta_D, \Phi, Y) * ... p(\Theta_D \mid \Theta_1, ..., \Theta_{D-1}, \Phi, Y) *$$

The joint distribution of $(\Theta_1^{(i)}, ..., \Theta_D^{(i)}, \Phi^{(i)})$ converges geometrically to $p(\Theta_1, ..., \Theta_D, \Phi \mid Y)$ as $i \rightarrow \infty$.

The Gibbs sampler differs from the Metropolis algorithm because in each step only one parameter, Θ_D, is allowed to change.

2 MULITPLE ALIGNMENT USING GIBBS SAMPLING

One application of Gibbs sampling useful in computational molecular biology is the detection and alignment of locally conserved regions (motifs) in sequences of amino acids or nucleic acids assuming no prior information in the patterns or motifs (Thompson et al., 2003; Neuwald et al., 1995; Lawrence et al., 1993). Gibbs sampling strategies claim to be fast and sensitive, avoiding the problem that EM algorithms fall into as far as getting trapped by local optima. As an example, a set of 29 DNA sequences have been provided. These sequences contain sequences necessary for recognition by erythroid transcription factors, most notably a six nucleotide GATA binding site.

2.1 Basic Algorithm

First the basic multiple alignment strategy is examined where a single motif is desired. The most basic implementation, known as a site sampler, assumes that there is exactly one motif element located within each sequence.

2.1.1 Notation

- N : number of sequences
- $S_1 \ldots S_N$: set of sequences
- W: width of motif to be found in the sequences
- J: the number of residues in the alphabet. $J = 4$ for nucleic acid sequences and 20 for amino acid sequences.
- $c_{i,j,k}$: Observed counts of residue j in position i of motif k. j ranges from 1.. J, i ranges from 0..W where $c_{0,j}$ contains the counts of residue j in the background. If it is assumed that only a single motif is searched for, the k term can drop out.
- $q_{i,j}$: frequency of residue j occurring in position i of the motif. i ranges from 0..W as above. Note that in the literature, $q_{0,j}$ (the vector of background residue frequencies) is sometimes denoted as p_j. This is the parameter of interest, Φ.
- a_k: vector of starting positions of the motifs within the sequences. k ranges from 1..N. This is the nuisance parameter, Θ.
- b_j: pseudocounts for each residue – needed according to Bayesian statistical rules to eliminate problems with zero counts and overtraining.
- B: The total number of pseudocounts. $B = \sum_j b_j$

2.1.2 Initialization

Once the sequences are known, the counts for each residue can be calculated. Initially, $c_{0,j}$ will contain the total counts of residue j within all of the sequences and $c_{i,j}$ is initialized to 0 for all other values of i. This is a summary observed data. The site sampler is then initialized by randomly selecting a position for the motif within each sequence and recording these positions in a_k. The counts are updated according to this initial alignment. After the observed counts are set, $q_{i,j}$ can be calculated according to equations 1 and 2.

$$q_{i,j} = \frac{c_{i,j} + b_j}{N - 1 + B} \quad (1)$$

Motif Residue Frequencies

$$q_{0,j} = \frac{c_{0,j} + b_j}{\sum_{k=1}^{j} c_{0,k} + B} \quad (2)$$

Background Residue Frequencies

2.1.3 Predictive Update Step

The first step, known as the predictive update step, selects one of the sequences and places the motif within that sequence in the background and updates the residue counts. One of the N sequences, z, is chosen. The motif in sequence z is taken from the model and placed in the background. The observed counts $c_{i,j}$ are updated as are the frequencies $q_{i,j}$. The selection of z can be random or in a specified order.

2.1.4 Sampling Step

In the sampling step, a new motif position for the selected sequence is determined by sampling according to a weight distribution. All of the possible segments of width W within sequence z are considered. For each of these segments x, a

weight A_x is calculated according to the ratio $A_x = \frac{Q_x}{P_x}$

where $Q_x = \prod_{i=1}^{W} q_{i,r}$ is the model reside frequency according to equation 1 if segment x is the model, and $P_x = \prod_{i=1}^{W} q_{0,r}$ is the background residue frequency according to equation 2. r_i refers to the residue located at position i of segment x . Once A_x is calculated for every possible x, a new position a_z is chosen by randomly sampling over the set of weights A_x. Thus, possible starting positions with higher weights will be more likely to be chosen as the new motif position than those positions with lower weights. Since this is a stochastic process, the starting position with the highest weight is not guaranteed to be chosen.

Once the iterative predictive update and sampling steps have been performed for all of the sequences, a probable alignment is present. For this alignment, a maximum *a posteriori* (MAP) estimate can be calculated using equation 3.

$$F = \sum_{i=1}^{W}\sum_{j=1}^{J} c_{i,j} \log \frac{q_{i,j}}{q_{0,j}} \qquad (3)$$

Alignment conditional log-likelihood

The goal is to maximize F. This is accomplished using the following pseudocode:

```
globalMaxAlignmentProb = 0
For Iteration = 1 to 10:
  Initialize Random alignment
  localMaxAlignmentProb = 0;
  while (not in local maximum and
         innerloop < MAXLOOP) do
    for each sequence do {
      Predictive Update
      Sample
    }
    calculate AlignmentProb
    if (AlignmentProb > localMaxAlignmentProb)
    {
      localMaxAlignmentProb = AlignmentProb;
      not in local maximum = true;
    }
    innerloop++;
  }
  if (localMaxAlignmentProb ==
      globalMaxAlignmentProb)
    exit // max found twice
  else if (localMaxAlignmentProb >
           globalMaxAlignmentProb)
    globalMaxAlignmentProb = localMaxAlignmentProb
}
```

2.1.4 Explanation

The idea is that the more accurate the predictive update step is, the more accurate the sampling step will be since the background will be more distinguished from the motif description. Given random positions a_k in the sampling step, the pattern description $q_{i,j}$ will not favor any particular segment. Once some correct a_k has been selected by chance, the $q_{i,j}$ begins to favor a particular motif.

2.1.5 Details

There are a couple of problems that need to be addressed. First, it is possible that the correct pattern has not been chosen, but rather a shift of it has. This can be taken care of by shifting the alignment to the left and right by a specified number of columns and sampling from the values of F.

Another problem is that the pattern width W must also be specified. In order to decide what the width should be, the incomplete-data log-probability ratio as shown in Equation 4 can be implemented.

$$G = F - \sum_{i=1}^{N} (\log L_i' + \sum_{j=1}^{L_i'} Y_{i,j} \log Y_{i,j}) \qquad (4)$$

Incomplete-data log-probability ratio

In equation 4, L_i' is the number of the possible positions for the pattern within sequence i and $Y_{i,j}$ is the normalized weight of position j . Dividing G by the number of free parameters needed to specify the pattern (19 $*W$ for protein sequences, $3*W$ for nucleotide sequences) results in an information per parameter quantity. It is then desired to maximize the information per parameter to determine the value of W.

2.2 Algorithm Improvements

The method of determining motifs as described above requires multiple runs on the same data set with varying widths to find the correct pattern size. The *Protein Science* paper (Neuwald et al., 1995) discusses a method to determine the width of the motif in a single run of the program, while at the same time determining gaps within the motif. The Gibbs sampler described thus far also requires the existence of exactly one motif in each sequence. Another improvement made is to allow multiple motifs within sequences, and allow the possibility that a sequence does not have any motifs. The improvements made within the *Protein Science* paper describe a technique known as a motif sampler.

2.2.1 Allowing a Variable Number of Motif Sites

Assume that there are m different motif patterns that we are searching for in the sequences. Let n_k represent the number of sites matching motif k in the sequence. Initially, it is not known how many motif sites there are. To overcome this, a prior expectation e_k is made for each n_k . The new algorithm allows the prior expectations to become posterior expectations as it learns the number of sites for each motif. For the initialization step, e_k random starting points are selected for motif pattern k instead of selecting one starting point randomly within each of the N sequences. Now we can go through all possible motif starting locations in each sequence and decide if it is a motif starting site by using equation 5.

$$\frac{p_j}{1-p_j} A_x \qquad (5)$$

Current motif site probability

Where p_j is the posterior probability that any site belongs to the model (see the appendix of the *Protein Science* paper for the prior and posterior calculations of P_j), and A_x is the same as in the site sampler.

2.2.2 Width Optimization by Column Sampling

In order to help introduce gaps and include only the most informative positions of the motif, column sampling is introduced where only C columns out of a specified number of contiguous columns $w_{\max} \geq C$ are used for the residue frequency model. This is accomplished in a two step process. First, turn off one column either randomly or by selecting it proportional to how little information it provides. Then sample one of the columns that are turned off proportional to how information rich it is and turn it on. The column move operations need to be weighted in order to assure that there is not a bias to longer motif widths. A discussion is provided in the appendix of the *Protein Science* paper.

3 PROPERTIES OF GIBBS SAMPLING

While Gibbs sampling can be an effective means of determining common motif patterns, it is important to keep in mind some of its properties in order to ensure proper use and analysis of results. The Gibbs sampler requires relatively large sets (on the order of 15 or more sequences) for weakly conserved patterns to reach statistical significance. Gibbs sampling is a heuristic and not an exhaustive search, so you are not guaranteed to reach an optimal value. However, the sampling approach allows motif detection to move away from locally optimal solutions unlike expectation-maximization approaches. In order for motifs to be detected, the user must specify an estimate for the width of the motifs, and how many motifs should be detected for the algorithm to perform the best. Gibbs sampling allows the user to view suboptimal results which may in themselves be meaningful. This approach is fast and sensitive, generally finding an optimized local alignment model for N sequences in N-linear time.

4 RESULTS

4.1 Site Sampler

The site sampler is tested using a set of erythroid sequences. The set is tested for the presence of a GATA box, which should have a sequence (T/A)GATA(A/G), which in the reverse complement is (C/T)TATC(A/T). Since the width of the GATA box is shown, it is known that for this example $W = 6$. The process of determining the best alignment using the site sampler is described.

4.1.1 Initialization

The first step in the site sampler is to randomly assign an alignment to the set of sequences. Figure 1 indicates one such random alignment.

```
TCAGAACCAGTTATAAATTTATCATTTCCTTCTCCACTCCT
CCCACGCAGCCGCCCTCCTCCCCGGTCACTGACTGGTCCTG
TCGACCCTCTGAACCTATCAGGGACCACAGTCAGCCAGGCAAG
AAAACACTTGAGGGAGCAGATAACTGGGCCAACCATGACTC
GGGTGAATGGTACTGCTGATTACAACCTCTGGTGCTGC
AGCCTAGAGTGATGACTCCTATCTGGGTCCCCAGCAGGA
GCCTCAGGATCCAGCACACATTATCACAAACTTAGTGTCCA
CATTATCACAAACTTAGTGTCCATCCATCACTGCTGACCCT
TCGGAACAAGGCAAAGGCTATAAAAAAAATTAAGCAGC
GCCCCTTCCCCACACTATCTCAATGCAAATATCTGTCTGAAACGGTTCC
CATGCCCTCAAGTGTGCAGATTGGTCACAGCATTTCAAGG
GATTGGTCACAGCATTTCAAGGGAGAGACCTCATTGTAAG
TCCCCAACTCCCAACTGACCTTATCTGTGGGGGAGGCTTTTGA
CCTTATCTGTGGGGGAGGCTTTTGAAAAGTAATTAGGTTTAGC
ATTATTTTCCTTATCAGAAGCAGAGAGACAAGCCATTTCTCTTTCCTCCC
GGT
AGGCTATAAAAAAAATTAAGCAGCAGTATCCTCTTGGGGGCCCCTTC
CCAGCACACACACTTATCCAGTGGTAAATACACATCAT
TCAAATAGGTACGGATAAGTAGATATTGAAGTAAGGAT
ACTTGGGGTTCCAGTTTGATAAGAAAAGACTTCCTGTGGA
TGGCCGCAGGAAGGTGGGCCTGGAAGATAACAGCTAGTAGGCTAAGGCCA
G
CAACCACAACCTCTGTATCCGGTAGTGGCAGATGGAAA
CTGTATCCGGTAGTGGCAGATGGAAAGAGAAACGGTTAGAA
GAAAAAAAATAAATGAAGTCTGCCTATCTCCGGGCCAGAGCCCCT
TGCCTTGTCTGTTGTAGATAATGAATCTATCCTCCAGTGACT
GGCCAGGCTGATGGGCCTTATCTCTTTACCCACCTGGCTGT
CAACAGCAGGTCCTACTATCGCCTCCCTCTAGTCTCTG
CCAACCGTTAATGCTAGAGTTATCACTTTCTGTTATCAAGTGGCTTCAGC
TATGCA
GGGAGGGTGGGGCCCCTATCTCTCCTAGACTCTGTG
CTTTGTCACTGGATCTGATAAGAAACACCACCCCTGC
```

Fig. 1 Initial motif locations for site sampler.

The number of A's in all of the sequences combined is 327, the number of C's is 317, the number of G's is 272, and the number of T's is 304. In order to alleviate the issue of zero counts and overtraining of the data, pseudocounts are introduced to the observed counts. If information about the motif model is known a priori, pseudocounts can be incorporated according to the predicted model description. If not, one simple method, known as Laplace's rule, is to add a count of one to each known observed count. Thus the following information is known before any of the initial motif sites are set:

$$c_{0,1} = 327; c_{0,2} = 317; c_{0,3} = 272; c_{0,4} = 304;$$
$$\sum_{i=1}^{4} c_{0,i} = 1220$$

$$b_1 = 1.0; b_2 = 1.0; b_3 = 1.0; b_4 = 1.0; B = 4.0$$

If we assume we have the initial random alignment as described in figure 1, we can recalculate the counts and calculate the residue frequencies. Table I gives the results of these calculations, and Table II indicates the updated observations incorporating Laplace's rule.

Table I: Calculation of observed counts for initial motif alignment (taken from Fig 1)

	Motif Position (0 = Background)						
Nucleotide	0	1	2	3	4	5	6
A	279	6	12	6	6	11	7
C	280	8	3	5	7	7	7
G	225	9	8	10	7	5	8
T	262	6	6	8	9	6	7

Table II: Updated counts with Laplace pseudocounts

	Motif Position (0 = Background)						
Nucleotide	0	1	2	3	4	5	6
A	279	7	13	7	7	12	8
C	280	9	4	6	8	8	8
G	225	10	9	11	8	6	9
T	262	7	7	9	10	7	8

Based upon the updated counts, the frequency of each nucleotide in each position can be calculated based upon the observed number of A, C, G, T in each column divided by the total number of observations of A+C+G+T. Table III shows the frequency information. Using Table III as a guideline, an odds ratio can be calculated for each position as the frequency of a particular residue occurring in that location in the motif divided by the frequency of that residue occurring in the background, which is characterized as position 0 in the motif (Table IV). In order to handle small probabilities, a $\log_2$ transform of the odds ratio can be calculated, as is shown in Table V.

Table III: Residue frequencies

	Motif Position (0 = Background)						
Nt	0	1	2	3	4	5	6
A	0.267	0.212	0.394	0.212	0.212	0.364	0.242
C	0.268	0.273	0.121	0.182	0.242	0.242	0.242
G	0.215	0.303	0.273	0.333	0.242	0.182	0.273
T	0.250	0.212	0.212	0.272	0.303	0.212	0.242

Table IV: Odds ratios

	Motif Position					
Nucleo-tide	1	2	3	4	5	6
A	0.795	1.477	0.795	0.795	1.363	0.909
C	1.019	0.453	0.679	0.906	0.906	0.906
G	1.409	1.268	1.550	1.127	0.845	1.268
T	0.847	0.847	1.089	1.210	0.847	0.968

Table V: Initial log-odds ratios

	Motif Position					
Nucleotide	1	2	3	4	5	6
A	-0.33	0.56	-0.33	-0.33	0.45	-0.14
C	0.03	-1.14	-0.56	-0.14	-0.14	-0.14
G	0.49	0.34	0.63	0.17	-0.24	0.34
T	-0.24	-0.24	0.12	0.27	-0.24	-0.05

4.1.2 Predictive Update Step

Now that the initial random alignment for the site sampler is known, the predictive update step begins by choosing one of the sequences to update. For simplicity, choose the first sequence. In the predictive update stage, the motif for the selected sequence is placed in the background and the counts and frequencies are updated. Since the motif in the first sequence is ATTTAT, tables I-IV can be recalculated.

4.1.3 Sampling Step

Once the counts and frequencies have been updated in the predictive update step, the sampling step begins. In this step, all possible motif starting positions within the sequence selected from the predicted update are considered. The first sequence has a length of 41, and the width of the motif is 6. Therefore, there are 41 - 6 + 1 = 36 possible starting sites. The probability of each of these sites being in the model is calculated and then sampled from their weights. The normalized $\log_2$-odds scores for each of the possible motif locations for sequence 1, based on the log-odds ratios in Table V, are shown in Table VI.

Using the information in table VI, one of the segments will be sampled in according to the normalized value of A_x. The predictive update and sampling steps are repeated for each of the sequences. Once each of the sequences have been

Table VI: Weights for segments within sequence 1

X	Sequence	A_x	Normalized A_x
1	TCAGAA	5.494	0.027
2	**CAGAAC**	**7.110**	**0.034**
3	AGAACC	5.465	0.026
4	GAACCA	6.401	0.031
5	AACCAG	6.488	0.031
6	ACCAGT	4.536	0.022
7	CCAGTT	5.209	0.025
8	CAGTTA	7.011	0.034
9	AGTTAT	6.693	0.032
10	GTTATA	5.895	0.029
11	TTATAA	5.971	0.029
12	TATAAA	6.480	0.031
13	ATAAAT	5.564	0.027
14	TAAATT	5.729	0.028
15	AAATTT	6.092	0.029
16	AATTTA	6.327	0.031
17	ATTTAT	6.272	0.030
18	TTTATC	5.330	0.026
19	TTATCA	5.513	0.027
20	TATCAT	6.649	0.032
21	ATCATT	4.931	0.024
22	TCATTT	5.119	0.025
23	CATTTC	6.547	0.032
24	ATTTCC	5.752	0.028
25	TTTCCT	5.562	0.027
26	TTCCTT	5.093	0.025
27	TCCTTC	4.941	0.024
28	CCTTCT	5.644	0.027
29	CTTCTC	5.613	0.027
30	TTCTCC	5.394	0.026
31	TCTCCA	5.109	0.025
32	CTCCAC	5.719	0.028
33	TCCACT	4.648	0.023
34	CCACTC	4.925	0.024
35	CACTCC	6.196	0.030
36	ACTCCT	5.116	0.025

sampled, an alignment is present and the alignment probability is tested. This procedure is repeated until a plateau is reached. Then another initial random alignment is tested and the process begins again.

For the example used thus far, the final alignment is as shown in Fig. 2. This alignment yields the counts and $\log_2$-odds ratios described in tables VII and VIII.

If the predictive update/sampling stages were repeated with these results, the next motif position would be sampled from the log-odds scores shown in table IX.

```
TCAGAACCAGTTATAAATTTATCATTTCCTTCTCCACTCCT
CCCACGCAGCCGCCCTCCTCCCCGGTCACTGACTGGTCCTG
TCGACCCTCTGGAACCTATCAGGGACCACAGTCAGCCAGGCAAG
AAAACACTTGAGGGAGCAGATAACTGGGCCAACCATGACTC
GGGTGAATGGTACTGCTGATTACAACCTCTGGTGCTGC
AGCCTAGAGTGATGACTCCTATCTGGGTCCCCAGCAGGA
GCCTCAGGATCCAGCACACATTATCACAAACTTAGTGTCCA
CATTATCACAAACTTAGTGTCCATCCATCACTGCTGACCCT
TCGGAACAAGGCAAAGGCTATAAAAAAATTAAGCAGC
GCCCCTTCCCCACACTATCTCAATGCAAATATCTGTCTGAAACGGTTCC
CATGCCCTCAAGTGTGCAGATTGGTCACAGCATTTCAAGG
GATTGGTCACAGCATTTCAAGGGAGAGACCTCATTGTAAG
TCCCCAACTCCCAACTGACCTTATCTGTGGGGGAGGCTTTTGA
CCTTATCTGTGGGGGAGGCTTTTGAAAAGTAATTAGGTTTAGC
ATTATTTTCCTTATCAGAAGCAGAGAGACAAGCCATTTCTCTTTCCTCCC
GGT
AGGCTATAAAAAAAATTAAGCAGCAGTATCCTCTTGGGGGCCCCTTC
CCAGCACACACACTTATCCAGTGGTAAATACACATCAT
TCAAATAGGTACGGATAAGTAGATATTGAAGTAAGGAT
ACTTGGGGTTCCAGTTTGATAAGAAAAGACTTCCTGTGGA
TGGCCGCAGGAAGGTGGGCCTGGAAGATAACAGCTAGTAGGCTAAGGCCA
G
CAACCACAACCTCTGTATCCGGTAGTGGCAGATGGAAA
CTGTATCCGGTAGTGGCAGATGGAAAGAGAAACGGTTAGAA
GAAAAAAAATAAATGAAGTCTGCCTATCTCCGGGCCAGAGCCCCT
TGCCTTGTCTGTTGTAGATAATGAATCTATCCTCCAGTGACT
GGCCAGGCTGATGGGCCTTATCTCTTTACCCACCTGGCTGT
CAACAGCAGGTCCTACTATCGCCTCCCTCTAGTCTCTG
CCAACCGTTAATGCTAGAGTTATCACTTTCTGTTATCAAAGTGGCTTCAGC
TATGCA
GGGAGGGTGGGGCCCCTATCTCTCCTAGACTCTGTG
CTTTGTCACTGGATCTGATAAGAAACACCACCCCTGC
```

Fig. 2: Final alignment for site sampler

Table VII: Final observed counts from Fig. 2

	Motif Position (0 = Background)						
Nucleotide	0	1	2	3	4	5	6
A	276	3	1	21	0	11	15
C	287	10	0	0	0	18	2
G	256	0	8	8	0	0	0
T	227	16	20	0	29	0	12

Table VIII: Final site sampler log-odds ratios

	Motif Position					
Nucleotide	1	2	3	4	5	6
A	-0.33	0.56	-0.33	-0.33	0.45	-0.14
C	0.03	-1.14	-0.56	-0.14	-0.14	-0.14
G	0.49	0.34	0.63	0.17	-0.24	0.34
T	-0.24	-0.24	0.12	0.27	-0.24	-0.05

4.2 Comparison of site and motif samplers

The site sampler and motif sampler follow the same basic Gibbs techniques. The difference is that the motif sampler will allow for the detection of zero or more motif locations in each sequence, whereas the site sampler detects exactly one. Thus, for the initialization step with the motif sampler,

Table IX: Final weights for segments within sequence 1

X	Sequence	A_x	Normalized A_x
1	TCAGAA	8.350	0.030
2	CAGAAC	4.383	0.016
3	AGAACC	6.645	0.024
4	GAACCA	6.926	0.025
5	AACCAG	2.412	0.009
6	ACCAGT	2.734	0.010
7	CCAGTT	5.931	0.022
8	CAGTTA	8.725	0.032
9	AGTTAT	9.096	0.033
10	GTTATA	5.288	0.019
11	**TTATAA**	**15.237**	**0.056**
12	TATAAA	6.074	0.022
13	ATAAAT	9.226	0.034
14	TAAATT	7.200	0.026
15	AAATTT	9.360	0.034
16	AATTTA	6.995	0.026
17	ATTTAT	10.914	0.040
18	TTTATC	6.032	0.022
19	TTATCA	15.958	0.058
20	TATCAT	6.047	0.022
21	ATCATT	5.572	0.020
22	TCATTT	11.155	0.041
23	CATTTC	6.244	0.023
24	ATTTCC	10.150	0.037
25	TTTCCT	9.470	0.035
26	TTCCTT	7.482	0.027
27	TCCTTC	7.255	0.026
28	CCTTCT	9.568	0.035
29	CTTCTC	4.868	0.018
30	TTCTCC	12.035	0.044
31	TCTCCA	6.670	0.024
32	CTCCAC	6.078	0.022
33	TCCACT	6.623	0.024
34	CCACTC	4.433	0.016
35	CACTCC	8.174	0.030
36	ACTCCT	4.734	0.017

a random alignment is made according to an estimate as to how many motif sites exist in total. An example of an initial alignment is given in Fig. 3.

Note that the estimate does not need to be the exact number of motif positions to be found. This is just a starting number that will evolve within the motif sampler. Using the same data that is used with the site sampler, the maximal alignment using the motif sampler is given in Fig. 4. With the motif sampler, any given location is sampled into the model based on the ratio of the site being in the model to it being in the background.

```
TCAGAACCAGTTATAAATTTATCATTTCCTTCTCCACTCCT
CCCACGCAGCCGCCCTCCTCCCCGGTCACTGACTGGTCCTG
TCGACCCTCTGAACCTATCAGGGACCACAGTCAGCCAGGCAAG
AAAACACTTGAGGGAGCAGATAACTGGGCCAACCATGACTC
GGGTGAATGGTACTGCTGATTACAACCTCTGGTGCTGC
AGCCTAGAGTGATGACTCCTATCTGGGTCCCCAGCAGGA
GCCTCAGGATCCAGCACACATTATCACAAACTTAGTGTCCA
CATTATCACAAACTTAGTGTCCATCCATCACTGCTGACCCT
TCGGAACAAGGCAAAGGCTATAAAAAAAATTAAGCAGC
GCCCCTTCCCCACACTATCTCAATGCAAATATCTGTCTGAAACGGTTCC
CATGCCCTCAAGTGTGCAGATTGGTCACAGCATTTCAAGG
GATTGGTCACAGCATTTCAAGGGAGAGACCTCATTGTAAG
TCCCCAACTCCCAACTGACCTTATCTGTGGGGGAGGCTTTTGA
CCTTATCTGTGGGGGAGGCTTTTGAAAAGTAATTAGGTTTAGC
ATTATTTTCCTTATCAGAAGCAGAGAGACAAGCCATTTCTCTTTCCTCCC
GGT
AGGCTATAAAAAAAATTAAGCAGCAGTATCCTCTTGGGGGCCCCTTC
CCAGCACACACACTTATCCAGTGGTAAATACACATCAT
TCAAATAGGTACGGATAAGTAGATATTGAAGTAAGGAT
ACTTGGGGTTCCAGTTTGATAAGAAAAGACTTCCTGTGGA
TGGCCGCAGGAAGGTGGGCCTGGAAGATAACAGCTAGTAGGCTAAGGCCA
G
CAACCACAACCTCTGTATCCGGTAGTGGCAGATGGAAA
CTGTATCCGGTAGTGGCAGATGGAAAGAGAAACGGTTAGAA
GAAAAAAAATAAATGAAGTCTGCCTATCTCCGGGCCAGAGCCCCT
TGCCTTGTCTGTTGTAGATAATGAATCTATCCTCCAGTGACT
GGCCAGGCTGATGGGCCTTATCTCTTTACCCACCTGGCTGT
CAACAGCAGGTCCTACTATCGCCTCCCTCTAGTCTCTG
CCAACCGTTAATGCTAGAGTTATCACTTTCTGTTATCAAGTGGCTTCAGC
TATGCA
GGGAGGGTGGGGCCCCTATCTCTCCTAGACTCTGTG
CTTTGTCACTGGATCTGATAAGAAACACCACCCCTGC
```

Fig. 3: Initial alignment for motif sampler (e=30).

```
TCAGAACCAGTTATAAATTTATCATTTCCTTCTCCACTCCT
CCCACGCAGCCGCCCTCCTCCCCGGTCACTGACTGGTCCTG
TCGACCCTCTGGAACCTATCAGGGACCACAGTCAGCCAGGCAAG
AAAACACTTGAGGGAGCAGATAACTGGGCCAACCATGACTC
GGGTGAATGGTACTGCTGATTACAACCTCTGGTGCTGC
AGCCTAGAGTGATGACTCCTATCTGGGTCCCCAGCAGGA
GCCTCAGGATCCAGCACACATTATCACAAACTTAGTGTCCA
CATTATCACAAACTTAGTGTCCATCCATCACTGCTGACCCT
TCGGAACAAGGCAAAGGCTATAAAAAAAATTAAGCAGC
GCCCCTTCCCCACACTATCTCAATGCAAATATCTGTCTGAAACGGTTCC
CATGCCCTCAAGTGTGCAGATTGGTCACAGCATTTCAAGG
GATTGGTCACAGCATTTCAAGGGAGAGACCTCATTGTAAG
TCCCCAACTCCCAACTGACCTTATCTGTGGGGGAGGCTTTTGA
CCTTATCTGTGGGGGAGGCTTTTGAAAAGTAATTAGGTTTAGC
ATTATTTTCCTTATCAGAAGCAGAGAGACAAGCCATTTCTCTTTCCTCCC
GGT
AGGCTATAAAAAAAATTAAGCAGCAGTATCCTCTTGGGGGCCCCTTC
CCAGCACACACACTTATCCAGTGGTAAATACACATCAT
TCAAATAGGTACGGATAAGTAGATATTGAAGTAAGGAT
ACTTGGGGTTCCAGTTTGATAAGAAAAGACTTCCTGTGGA
TGGCCGCAGGAAGGTGGGCCTGGAAGATAACAGCTAGTAGGCTAAGGCCA
G
CAACCACAACCTCTGTATCCGGTAGTGGCAGATGGAAA
CTGTATCCGGTAGTGGCAGATGGAAAGAGAAACGGTTAGAA
GAAAAAAAATAAATGAAGTCTGCCTATCTCCGGGCCAGAGCCCCT
TGCCTTGTCTGTTGTAGATAATGAATCTATCCTCCAGTGACT
GGCCAGGCTGATGGGCCTTATCTCTTTACCCACCTGGCTGT
CAACAGCAGGTCCTACTATCGCCTCCCTCTAGTCTCTG
CCAACCGTTAATGCTAGAGTTATCACTTTCTGTTATCAAGTGGCTTCAGC
TATGCA
GGGAGGGTGGGGCCCCTATCTCTCCTAGACTCTGTG
CTTTGTCACTGGATCTGATAAGAAACACCACCCCTGC
```

Fig. 4: Initial alignment for motif sampler (e=30).

5 DISCUSSION

This technical report discusses the basics behind the Gibbs sampling algorithm. For further information, consult the following journal articles:

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. 1993. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment.*Science* **262**:208-214.

> *This paper describes a Gibbs sampling strategy where there is assumed to be a single occurrence of the motif within each sequence. No gaps are allowed within the alignment. The implementation is known in the literature as a site sampler.*

Neuwald AF, Liu JS, Lawrence CE. 1995. Gibbs motif sampling: detection of outer membrane repeats. *Protein Science* **4**:1618-1632.

> *This paper describes a Gibbs sampling strategy where the number of motifs is not known. This is the motif sampler. Examples are presented in the location of the immunoglobulin fold and hth motifs.*

Liu JS, Neuwald AF, Lawrence CE. 1995. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *Journal of the American Statistical Association* 90, 432:1156-1171.

> *This paper contains more of the derivations for implementing a Bayesian model in the Gibbs sampler. The details of this paper are not covered here, but if further research into the derivations of the various formulas sounds interesting, this should be a good place to start.*

Tanner, MA. 1993. Tools for Statistical Inference. Springer-Verlag.

ACKNOWLEDGEMENTS

The work reported here was originally reported in May of 1997 as part of a presentation for Washington University's Institute for Biomedical Computing (IBC) Statistics Study Group. It is published here in slightly modified form as a technical report in order to provide a widespread public dissemination of the work. ER is currently supported by NIH-NCRR grant P20RR16481 and NIH-NIEHS grant P30ES014443. The contents of this manuscript are solely the responsibility of the authors and do not represent the official views of NCRR, NIEHS, or NIH.

REFERENCES

Casella,G. and George,E.I. (1992) Explaining the Gibbs Sampler. *The American Statistician*, 46, 167-174.

Geman,S. and Geman,D. (1984) Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 6, 721-741.

Hastings,W.K. (1970) Monte Carlo Sampling Methods Using Markov Chains and their Applications. *Biometrika*, 57, 97-109.

Lawrence,C.E. et al. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262, 208-214.

Lawrence,C.E. and Reilly,A.A. (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, 7, 41-51.

Liu,J.S., Neuwald,A.F. and Lawrence,C.E. (1995) Bayesian Models for Multiple Sequence Alignment and Gibbs Sampling Strategies. *Journal of the American Statistical Association*, 90, 1156-1171.

Metropolis,N. et al. (1953) Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21, 1087-1092.

Neuwald,A.F., Liu,J.S. and Lawrence,C.E. (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.*, 4, 1618-1632.

Newberg,L.A. et al. (2007) A phylogenetic Gibbs sampler that yields centroid solutions for cis-regulatory site prediction. *Bioinformatics*, 23, 1718-1727.

Thompson,W. et al. (2007a) Using the Gibbs Motif Sampler for phylogenetic footprinting. *Methods Mol. Biol.*, 395, 403-424.

Thompson,W., Rouchka,E.C. and Lawrence,C.E. (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res.*, 31, 3580-3585.

Thompson,W.A. et al. (2007b) The Gibbs Centroid Sampler. *Nucleic Acids Res.*, 35, W232-W237.