

## Sequence analysis

# A phylogenetic Gibbs sampler that yields centroid solutions for *cis*-regulatory site prediction

Lee A. Newberg<sup>1,2,†</sup>, William A. Thompson<sup>3,\*,†</sup>, Sean Conlan<sup>1</sup>, Thomas M. Smith<sup>1,2,\*</sup>, Lee Ann McCue<sup>1,4</sup> and Charles E. Lawrence<sup>3</sup>

<sup>1</sup>The Wadsworth Center, New York State Department of Health, Albany, NY 12201, <sup>2</sup>Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180, <sup>3</sup>Center for Computational Molecular Biology and the Division of Applied Mathematics, Brown University, Providence, RI 02912 and <sup>4</sup>Pacific Northwest National Laboratory, Richland, WA 99352, USA

Received on January 10, 2007; revised on April 27, 2007; accepted on April 28, 2007

Advance Access publication May 8, 2007

Associate Editor: Dmitrij Frishman

**ABSTRACT**

**Motivation:** Identification of functionally conserved regulatory elements in sequence data from closely related organisms is becoming feasible, due to the rapid growth of public sequence databases. Closely related organisms are most likely to have common regulatory motifs; however, the recent speciation of such organisms results in the high degree of correlation in their genome sequences, confounding the detection of functional elements. Additionally, alignment algorithms that use optimization techniques are limited to the detection of a single alignment that may not be representative. Comparative-genomics studies must be able to address the phylogenetic correlation in the data and efficiently explore the alignment space, in order to make specific and biologically relevant predictions.

**Results:** We describe here a Gibbs sampler that employs a full phylogenetic model and reports an ensemble centroid solution. We describe regulatory motif detection using both simulated and real data, and demonstrate that this approach achieves improved specificity, sensitivity, and positive predictive value over non-phylogenetic algorithms, and over phylogenetic algorithms that report a maximum likelihood solution.

**Availability:** The software is freely available at <http://bayesweb.wadsworth.org/gibbs/gibbs.html>

**Contact:** William\_Thompson\_1@brown.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 INTRODUCTION

We are entering an unprecedented time in the evolution of the field of comparative genomics. High-throughput sequencing facilities can produce billions of bases of sequenced DNA per month. This capability has enabled ambitious sequencing projects, including the sequencing of many closely related genomes (Achtman *et al.*, 1999; Chain *et al.*, 2004; Clark *et al.*, 2003; ENCODE, 2004; Liu *et al.*, 2002). Analysis of the resulting sequence data will contribute to answering diverse

scientific questions throughout the tree of life, including the computational detection of functionally conserved regions in protein and nucleic acid sequences. In particular, the identification of transcription factor binding sites and *cis*-regulatory modules in the promoters of genes is critical in the delineation of the transcription regulatory network of an organism. Closely related species are most likely to have common transcription factors, and therefore common *cis*-regulatory elements, a fact which makes their inclusion in comparative studies attractive. Unfortunately, the recent speciation of closely related species results in correlation among their genomic sequences, due to insufficient evolutionary time over which mutations could have occurred. This problem confounds the detection of functionally conserved motifs via non-phylogenetic methods, because motif-detection algorithms typically attempt to find conserved motifs against a background sequence that is less conserved. When the background sequence is highly correlated, as occurs for sequences from closely related species, the algorithms have difficulty in identifying the functionally conserved regions.

To address this challenge, a number of algorithms have been developed recently, which incorporate phylogenetic models into consensus (Wang and Stormo, 2003), expectation maximization (Moses *et al.*, 2004; Prakash, *et al.*, 2004; Sinha, *et al.*, 2004), or Gibbs sampling methods (Li and Wong, 2005; Liu *et al.*, 2004a; Liu *et al.*, 2004b; Siddharthan *et al.*, 2005) (see Supplementary Material for more information on these methods). Most of these algorithms were designed with the goal of identifying regulatory motifs in sequence data that consist of orthologous data from a group of co-regulated genes, although the Gibbs sampling-based PhyloGibbs (Siddharthan *et al.*, 2005) will function on purely orthologous data (i.e. the orthologous regulatory regions for a single gene). PhyloGibbs uses aligned sequence data, and an evolutionary model that assumes that all positions in a binding site evolve independently (at equal rates), that the probability of fixation of a mutation at a position is proportional to the weight matrix entry at that position, and a star-topology decomposition to approximate a posterior-probability calculation. PhyloGibbs is the algorithm most similar to the phylogenetic Gibbs sampler that we describe in the following.

\*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

To facilitate high-resolution comparative genomics studies, we have developed a version of the Gibbs recursive sampler (Thompson *et al.*, 2003, 2005) that incorporates phylogeny of the input sequences, through the use of an evolutionary model (Felsenstein, 1981), and calculates an ensemble centroid motif solution (Ding *et al.*, 2005). Our phylogenetic Gibbs sampler accepts (globally) pre-aligned as well as unaligned orthologous sequence data; these may be orthologous sequences from a single gene or orthologous sequences for a group of co-regulated genes. The algorithm also requires a user-supplied tree describing the phylogenetic relationship of the aligned orthologous sequences. For each alignment of orthologous sequences, the algorithm traverses this phylogenetic tree and calculates the joint probability of each nucleotide at each position, ultimately describing a motif as a product phylogeny model, i.e. the product across motif positions of the joint probabilities calculated by the Felsenstein algorithm (Felsenstein, 1981). The phylogenetic Gibbs sampler employs a fully phylogenetic model and avoids star-topology approximations. Furthermore, in studies of RNA secondary structure, Ding and colleagues (Ding *et al.*, 2005) recently showed the limitations of optimization-based estimators such as the maximum *a posteriori* (MAP); as an alternative, they described ensemble-based ‘centroid’ estimators. Following their lead, we employ an ‘ensemble centroid’ motif estimator, i.e. the solution that is the set of sites having the minimum total distance to the set of sites sampled from the posterior weighted ensemble of sites. The results below indicate that these two extensions provide significant additional power to the existing capabilities of the Gibbs recursive sampler. Building on the previous work of our group, this algorithm also searches for multiple motif types, multiple instances (sites) of a motif, palindromic motifs and motifs of varying widths, and it employs a heterogeneous background frequency model (see Thompson *et al.*, 2005 for descriptions of these features).

We demonstrate the advanced features of the phylogenetic Gibbs sampler by applying it to the challenging problem of predicting motifs in orthologous data from a single gene. Using data that simulate this scenario, we show that false-positive predictions, arising from correlation among the sequences, are dramatically reduced by the features described here. We also compare our results from the phylogenetic Gibbs sampler with results from PhyloGibbs, because, among the available phylogenetic algorithms (Li and Wong, 2005; Liu *et al.*, 2004a; Liu *et al.*, 2004b; Moses *et al.*, 2004; Prakash *et al.*, 2004; Siddharthan *et al.*, 2005; Sinha *et al.*, 2004; Wang and Stormo, 2003), PhyloGibbs permits analysis of orthologous data from a single gene. Specifically, we show that the use of a phylogenetic model and ensemble centroid solutions yields improved specificity (the ability to avoid false positives) and positive predictive value (PPV) relative to (1) a non-phylogenetic version of the Gibbs recursive sampler, (2) a phylogenetic version of the Gibbs sampler that predicts maximum-likelihood alignments and (3) PhyloGibbs (Siddharthan *et al.*, 2005). We further demonstrate the ability of the phylogenetic Gibbs sampler to detect transcription factor binding sites in real orthologous sequence data from eight  $\gamma$ -proteobacterial genomes, the majority of which are enterobacterial.

## 2 METHODS

### 2.1 Algorithm

**2.1.1 The phylogenetic model** In previous work, we analyzed aligned pairs of human and mouse sequences, treating the data from the aligned sequence pairs as independent (Thompson *et al.*, 2004). To accomplish this, we modified the basic predictive inference distribution for a single position (Liu *et al.*, 1995), to incorporate data from aligned pairs of bases treated as if they were independent. In the current model, phylogenetic correlation among the species in the aligned set is accounted for by modeling the joint probability of an aligned set of nucleotides using the Felsenstein tree-likelihood algorithm (Durbin *et al.*, 1998; Felsenstein, 1981). This phylogenetic algorithm is an efficient recursive algorithm, which traverses a phylogenetic tree to marginalize over its interior nodes, so as to obtain the joint probability of the nucleotides on its leaves. In previous Gibbs sampling algorithms, we used a product multinomial model to describe the motif model (Liu *et al.*, 1995):

$$P(R_{i,j,\dots,j+w-1}|\Theta) = \prod_{l=1}^w P(R_{i,j+l-1}|\Theta_l) = \prod_{l=1}^w \prod_{b \in \{A,T,C,G\}} \Theta_{lb}^{c_{j+l-1,b}}, \quad (1)$$

where  $R_{ij}$  represents the nucleotide at the  $j$ -th position of the  $i$ -th sequence,  $\Theta$  is the position weight matrix  $\Theta_{lb}$  is the model probability for base  $b$  at motif position  $l$ , and  $c_{j+l-1,b}$  is an indicator variable, where  $c_{j+l-1,b} = 1$  if the base at position  $j+l-1$  is type  $b$ . Here, as described below, we use a product phylogeny model to combine values across positions within a motif.

The input to the algorithm is a collection of multiply aligned sequence sets (MASSes), where each set consists of a global multiple alignment of orthologous sequences related by a known (or user-estimated) phylogenetic tree. Each unalignable sequence is treated as a (degenerate) MASS of one sequence. Our model treats the MASSes as statistically independent of one another. Notationally, a MASS consists of  $N$  aligned sequences, each of length  $L$ , including any gaps.  $R_{ij}$  represents the nucleotide, or gap, at the  $j$ -th position of the  $i$ -th sequence. Using the Felsenstein tree-likelihood algorithm, we can calculate  $P(R_{1\dots N,j}\theta, T)$ , the joint probability of the nucleotides at the aligned position  $j$  of the MASS, where  $\theta$  is the current motif position, or background nucleotide probability model as applicable, and  $T$  is the phylogenetic tree. We assume independence of the individual positions in a transcription factor binding site; thus, we model a motif site of width  $w$  as a product of the probabilities of the individual positions:

$$P(R_{1\dots N,j\dots j+w-1}|\Theta, T) = \prod_{l=1}^w P(R_{1\dots N,j+l-1}|\Theta_l, T). \quad (2)$$

We assume that all nucleotides not in motif positions are drawn from a (possibly position-dependent) background model  $\Theta_0$ . When there is more than one motif model, we assume that each model has its own nucleotide probability model, where positions of the  $k$ -th motif are described by a vector of independent models  $\Theta^{(k)} = (\Theta_1^{(k)}, \Theta_2^{(k)}, \dots, \Theta_w^{(k)})$ .

The algorithm and model used here are extensions of those previously described (Liu *et al.*, 1995, 2004a). The algorithm has two sampling steps, the site-sampling step and the model-update step. This two-step approach is essential in the analysis of a single MASS. In the site-sampling step, the probability of a site from motif model  $k$  at position  $j$  in the aligned sequences of a MASS is proportional to

$$\frac{P(R_{1\dots N,j\dots j+w-1}|\Theta^{(k)}, T)}{P(R_{1\dots N,j\dots j+w-1}|\Theta_0, T)} = \prod_{l=1}^w \frac{P(R_{1\dots N,j+l-1}|\Theta_l^{(k)}, T)}{P(R_{1\dots N,j+l-1}|\Theta_0, T)}. \quad (3)$$

The multiple sites from multiple motifs are sampled using the recursive algorithm described previously (Thompson *et al.*, 2004). This recursive algorithm infers the total number of sites in each MASS,

the motif types of these sites, and the neighboring relationship between pairs of motif types. For the model-update step, we use a Metropolis–Hastings algorithm (Hastings, 1970). Starting with an existing model  $\Theta$ , the algorithm first draws a proposed next model,  $\hat{\Theta}$ , using sequence-weighted counts (described below) from  $Dir(c + \beta)$ , where  $Dir$  is the Dirichlet distribution, and  $c$  and  $\beta$  are the current weighted count and pseudocount vectors. Then, the proposed model is accepted with a probability based on a Metropolis–Hastings ratio:

$$\min \left\{ 1, \prod_{l=1}^w \frac{\left( Fels(\hat{\Theta}_l) / \prod_{b \in \{A,T,C,G\}} \hat{\Theta}_{lb}^{c_{lb}} \right)}{\left( Fels(\Theta_l) / \prod_{b \in \{A,T,C,G\}} (\Theta_{lb}^{c_{lb}}) \right)} \right\}, \quad (4)$$

where  $Fels(\hat{\Theta}_l)$  is the result of the Felsenstein calculation for position  $l$  of the appropriate motif sites, and  $c_{lb}$  denote the sequence-weighted counts for each base at those positions.

**2.1.2 Sequence weights** In previous versions of the Gibbs sampling algorithm, sites were sampled according to the nucleotide probability distribution:

$$\Theta_{lb} = \frac{c_{lb} + \alpha_{lb}}{\sum_{b' \in \{A,T,C,G\}} (c_{lb'} + \alpha_{lb'})}, \quad (5)$$

where the  $c_{lb}$  values denote the sums of the counts of nucleotide type  $b$  at position  $l$  in the aligned motif sites, and the  $\alpha_{lb}$  values denote the corresponding pseudocounts (Liu *et al.*, 1995). A similar equation applies for the non-site, background positions. This expression assumes independence among the sequences. Here, to obtain the proposal distribution for the Metropolis–Hastings ratio (described above), we have employed optimal sequence weights as input, to obtain the mean values of a modified posterior Dirichlet distribution. Specifically, in the equation above, the counts,  $c_{lb}$ , are replaced by the weighted counts computed from the optimal sequence weights, generated from the phylogenetic tree (Newberg *et al.*, 2005). These optimal sequence weights are those weights that minimize the sum of the variances of the estimators of base frequency parameters, for sequences related by the phylogenetic tree. However, even an optimal weighting scheme yields parameter estimates whose variances are unnecessarily large, as shown previously (Newberg *et al.*, 2005). The above-described Metropolis–Hastings feature addresses this weakness, by drawing samples from the full-likelihood model.

**2.1.3 MAP calculation** Previous versions of the Gibbs sampler used the posterior probability of the alignment, called the MAP (Liu *et al.*, 1995), as a measure of the quality of the alignment, and thus the alignment that produced the highest MAP (i.e. the MAP alignment) was returned. The reported MAP was calculated as the logarithm of the alignment probability minus the logarithm of an empty or background alignment. Thus, it was a measure of the extent to which a particular alignment was better than background. To calculate the MAP, all parameters except the alignment were integrated from the expression for the posterior probability. This was possible because the motif and background models were represented as a product multinomial distribution, and the prior distribution was modeled as its conjugate Dirichlet distribution (Liu *et al.*, 1995). For example, for the motif prior probability distribution:

$$P(\Theta) = \prod_{l=1}^w \frac{\Gamma\left(\sum_b \alpha_{lb}\right)}{\prod_b \Gamma(\alpha_{lb})} \prod_b \Theta_{lb}^{\alpha_{lb}-1}, \quad (6)$$

where  $\Gamma$  is the Gamma function,  $b$  ranges over the set  $\{A, T, C, G\}$  and  $\alpha_{lb} > 0$  is the number of pseudocounts for nucleotide  $b$  at motif position  $l$ , the foreground integral can be computed exactly.

$$\int \prod_{l=1}^w \prod_b \Theta_{lb}^{c_{lb}} P(\Theta_{lb}) d\Theta = \prod_{l=1}^w \frac{\Gamma\left(\sum_b \alpha_{lb}\right)}{\Gamma\left(\sum_b (\alpha_{lb} + c_{lb})\right)} \prod_b \frac{\Gamma(\alpha_{lb} + c_{lb})}{\Gamma(\alpha_{lb})}, \quad (7)$$

where, as before,  $c_{lb}$  is the observed number of occurrences of nucleotide  $b$  at motif position  $l$ , and  $\alpha_{lb}$  is the corresponding pseudocounts.

In the phylogenetic algorithm, the likelihoods of the motif and background models are the output of the Felsenstein tree-likelihood algorithm (Felsenstein, 1981). Specifically, the probability of the alignment is given by the expression:

$$P(R, \Theta, \Theta_0 | A) \propto \left( \prod_{j \in \text{background}} Fels(R_{1 \dots N_j} | \Theta_0) \prod_b \Theta_{0b}^{\alpha_b - 1} \right) \times \prod_{l=1}^w \left[ \left( \prod_{j \in \text{motif}(l)} Fels(R_{1 \dots N_j} | \Theta_l) \right) \left( \prod_b \Theta_{lb}^{\beta_{lb} - 1} \right) \right], \quad (8)$$

where  $b \in \{A, T, C, G\}$ ;  $A$  is an indicator variable, where  $a_j = 1$  if a motif site starts at position  $j$  and 0 otherwise;  $R$  is the sequence data;  $\Theta$ ,  $\Theta_0$  are the motif and the background probability models (possibly position-dependent background probability model); and  $\alpha$  and  $\beta$  are the background and motif pseudocounts, respectively.  $Fels(R_{1 \dots N_j} | \theta)$  is the result of the Felsenstein algorithm for a particular motif or background position with nucleotide equilibrium probability distribution  $\theta$ , and is the joint probability of the nucleotides at the aligned position  $j$  of the MASS.

Since we know of no method to analytically integrate  $P(R|A) \propto \int P(R, \Theta, \Theta_0 | A) d\Theta_0 d\Theta$ , we instead resort to importance sampling (Liu, 2001), a form of numerical integration. We found the importance sampling procedure to be computationally intensive (see the Supplementary Material for details). In PhyloGibbs (Siddharthan *et al.*, 2005), by comparison, a star-topology approximation is used for estimation of this integral. Fortunately, as shown below, centroid estimators provide more accurate estimates, and thus the computationally intensive calculation of the MAP can be avoided completely.

**2.1.4 Ensemble centroid** Once initialized, usually with a random alignment, the Gibbs sampling procedure proceeds through the following steps: (1) the probability of each possible number of sites for each MASS is calculated based on the current model; (2) the number of sites for each MASS is sampled; (3) the predicted sites of each of the current motif models are sampled in each MASS; and (4) the motif models are updated from the sampled sites from all MASSes. In previous versions, this process repeated until the MAP failed to increase for a fixed number of iterations. To obtain a sampling solution, we allow the algorithm to sample through a burn-in period, typically 2000–3000 iterations. Next, the sampler proceeds, again through a fixed number of iterations (typically 8000–10000), tracking each sampled position. A centroid alignment solution is obtained from these samples through identification of the alignment that possesses the minimum total distance to the other alignments in the set. Thus, the centroid is defined in terms of a distance measure between pairs of proposed alignments.

In the computation of the distance between two alignments, each site in each of the two alignments is considered separately, and the distance between the two alignments is the sum of the values, computed as follows. If the site exactly overlaps a site in the other alignment, then this is ‘perfect overlap’ and the contribution to the distance sum is zero. If the site overlaps a site in the other alignment by more than half of the larger of the widths of each of the two sites, then this is ‘sufficient



phylogenetic Gibbs sampler achieves a PPV of 0.67 with one planted site, and the PPV improves significantly as the number of planted sites increases. The centroid version is highly resistant to false predictions, as indicated by a PPV of 1.0 throughout. The tendency of PhyloGibbs to over-identify sites in background data (Table 1) has a carryover negative effect on the algorithm's PPV; not surprisingly, this effect is greatest with only one or two planted sites, where PhyloGibbs returns a PPV of under 0.25, and the effect diminishes as the number of planted sites increases (Fig. 2A).

### 3.4 Sensitivity

We also evaluated the sensitivity of each of the three phylogenetic algorithms on the simulated data with one to four planted sites, to ask what proportion of the planted sites (i.e. the true positives) were detected. Figure 2B shows that the two optimization-based procedures, PhyloGibbs and the MAP version of the phylogenetic Gibbs sampler, detect fewer than 10% of the planted sites, when there is only one planted site per sequence. As the number of planted sites increases, these two algorithms show gradual improvement in sensitivity. However, of particular interest in these results is the marked improvement in sensitivity demonstrated by the centroid version of the phylogenetic Gibbs sampler, over the optimization based procedures. Our results show that for this simulated data set, the centroid alignment has distinct advantage over optimization-based algorithms. These results further suggest that both phylogenetic optimization algorithms focus only on a subset of the true sites, and in so doing derive an overly-focused model that fails to generalize sufficiently to capture the remaining sites.

### 3.5 Evaluation of algorithms using simulated yeast data

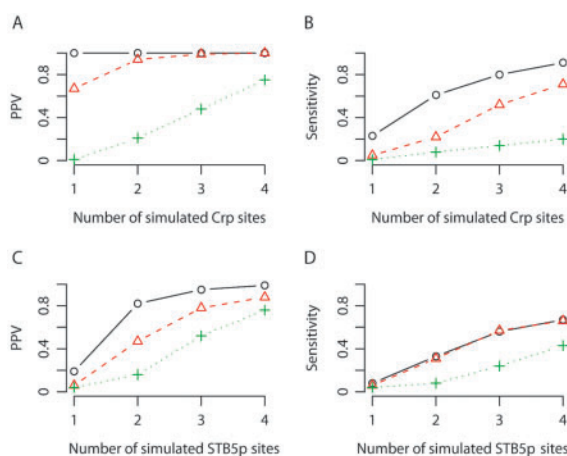
When compared to the MAP version of the phylogenetic Gibbs sampler (algorithm #2), the relatively poor performance of

PhyloGibbs (algorithm #4) was surprising, given the similarities between the two algorithms. One significant difference between the algorithms is that PhyloGibbs uses an approximate calculation based upon a star-topology decomposition of the phylogenetic tree, whereas the phylogenetic Gibbs sampler employs a full phylogenetic model and a fairly accurate, though computationally intensive, integration over possible motif models (see Methods). To investigate whether this contributes to the observed difference in performance on the simulated proteobacterial data, we evaluated the algorithms on simulated yeast data described by a tree with a star topology (Fig. 1B) from the PhyloGibbs paper (Siddharthan *et al.*, 2005). The data were simulated as described in the Supplementary Materials, with planted 10-column-wide, non-palindromic motif sites (0–4 per simulated sequence), and analyzed with PhyloGibbs and both the MAP and centroid versions of the phylogenetic Gibbs sampler. The specificity of the algorithms was somewhat poorer on this simulated yeast data, than on the simulated proteobacterial data (Table 1), likely because the algorithms were searching for a shorter non-palindromic model. The centroid-based phylogenetic Gibbs sampler (#3) made 26.0 false-positive predictions while the MAP version of Gibbs (#2) and PhyloGibbs (#4) make similar numbers of false-positive predictions (93.3 and 100.3, respectively). The PPV and sensitivity (Fig. 2) showed trends similar to that observed with the simulated proteobacterial data; that is, all three algorithms show improved PPV and sensitivity as the number of planted sites increases. The MAP and centroid versions of the phylogenetic Gibbs sampler show similar sensitivity, though the centroid version exhibits consistently higher PPV, due to its tendency to avoid false predictions. When compared to the results from the proteobacterial simulation, the PPV of PhyloGibbs on the yeast data showed little change, though the overall sensitivity was improved, supporting our expectation that PhyloGibbs is more suited to motif finding when sequences are related by a star phylogeny. The overall difference in performance between the phylogenetic Gibbs sampler and PhyloGibbs cannot, however, be attributed solely to the choice of tree topology.

### 3.6 Uncertainty of motif estimators

Uncertainty is an inescapable component of prediction and inference, yet it has been addressed in only very limited ways in computational molecular biology. Here, to examine the uncertainty associated with the motif estimators employed by the phylogenetic Gibbs sampler, we measured the Hamming distance between the predicted sites (in the centroid or MAP solutions) and those in the samples drawn by the sampler, where the Hamming distance gives the number of differences in predicted sites between two alignments (sampled/predicted sites in a sequence or MASS). Table 2 gives the mean, taken over the samples, of the Hamming distances, as well as the average number of sites predicted by both estimators in these simulations.

Table 2 shows that the mean distances from MAP solutions to the samples are consistently larger than those between the centroids and the samples. This is not surprising, since the centroid estimator minimizes these differences.



**Fig. 2.** (A) PPV and (B) sensitivity as a function of the number of planted simulated Crp sites in each simulated  $\gamma$ -proteobacterial sequence; (C) PPV and (D) sensitivity as a function of the number of planted simulated STB5p sites in each simulated yeast sequence. Open circle, results for the Centroid+Phylogeny; open triangle, results for MAP+Phylogeny; plus symbol, results for PhyloGibbs.

What is perhaps surprising is the size of the differences, especially when two or more sites were planted. Most striking are the differences for the Crp data sets with two or more sites, showing that the sites in MAP solutions differ from the sites in a representative sample by more than one site, on average. Thus, in these cases, the MAP estimator is not well recommended by the data. This failure to follow the data's recommendations leads to the loss of sensitivity shown in Figure 2B. In these data sets (Crp data sets with two or more sites), the fact that the corresponding distances from the centroids to the samples have about half the average number of differences than the MAP to samples distances, indicates that in these cases, the MAP estimates are not centered in the posterior distribution. Furthermore, since by design centroid estimates are centered in the posterior, the large distances from the samples to the centroid for STB5 indicate high variance in the posterior space, reflecting high uncertainty in these predictions. For the rows corresponding to zero and one

planted site for STB5, we see that the centroid predicts  $< 0.5$  sites, yet the distances to the samples are, on average, over two sites. These results indicate that sampling procedure often samples many sites in these sequences, but their predicted locations are highly inconsistent, indicating high uncertainty in the locations of the sites and producing a high mean distance. By comparison, the posterior space for the sampled solutions in the Crp data is gathered much more tightly around the centroid. The greater uncertainty of the STB5 predictions likely stems from the fact that the STB5 motifs are shorter than the Crp motifs, and not palindromic.

### 3.7 $\gamma$ -Proteobacterial study set

Based on the results from analysis of the simulated data, we conclude that the phylogenetic Gibbs sampler centroid alignment (algorithm #3) provides the best balance of PPV and sensitivity. Therefore, we used that algorithm on a study set of real proteobacterial data (see Supplementary Materials). These data present challenges not contained in the simulated data, and our approach employed features of our Gibbs sampling algorithm (Thompson *et al.*, 2005) to adjust accordingly. Specifically, because some of these real promoters contain more than one type of site (i.e. multiple motif types), we ran the algorithm on all the data sets using one, two, and three different motif models. Furthermore, the known binding sites vary in width (e.g. FruR sites are 16 bp wide, and Crp sites are 22 bp wide), and, while many are palindromic, a few are not (e.g. DnaA sites). We employed the following parameters, because we have shown (McCue *et al.*, 2001) that they perform well with such proteobacterial data: each motif model was palindromic and contained 16 conserved positions that were allowed to fragment (Thompson *et al.*, 2003) to a maximum width of 24 positions. Table 3 summarizes the results of the phylogenetic Gibbs sampler centroid alignments for the study set data, averaged over three runs for each of the three configurations (one, two or three motif models). These results are consistent with the results obtained from simulated data, based on a calculation of the weighted average of the PPV (0.74) and sensitivity (0.45) for the simulation results (see Supplementary Material).

We also utilized data from the real  $\gamma$ -proteobacterial data set to test performance of the phylogenetic Gibbs sampler centroid algorithm on sets of co-regulated genes with orthologous data. We assembled three collections of orthologous promoter

**Table 2.** Uncertainty of motif estimators

# Sites <sup>a</sup>	Mean distance <sup>b</sup>		Number of predicted sites <sup>c</sup>	
	centroid-to-samples	MAP-to-samples	centroid	MAP
<b>Crp</b>				
0	0.14 ± 0.15	0.16 ± 0.23	0.01 ± 0.10	0.02 ± 0.19
1	0.30 ± 0.19	0.38 ± 0.29	0.23 ± 0.42	0.08 ± 0.28
2	0.57 ± 0.29	1.02 ± 0.52	1.22 ± 0.85	0.46 ± 0.81
3	0.71 ± 0.42	1.30 ± 0.82	2.41 ± 0.90	1.57 ± 1.44
4	0.78 ± 0.46	1.42 ± 1.10	3.64 ± 0.76	2.85 ± 1.73
<b>STB5</b>				
0	2.23 ± 0.55	2.67 ± 1.15	0.26 ± 0.60	0.93 ± 1.55
1	2.25 ± 0.50	2.66 ± 1.03	0.42 ± 0.82	0.96 ± 1.55
2	2.10 ± 0.60	2.63 ± 1.17	0.80 ± 0.92	1.32 ± 1.42
3	1.90 ± 0.76	2.45 ± 1.28	1.78 ± 1.26	2.18 ± 1.41
4	1.70 ± 0.80	2.25 ± 1.36	2.73 ± 1.21	3.03 ± 1.22

<sup>a</sup>The number of planted, simulated transcription factor binding sites.

<sup>b</sup>The mean Hamming distance from the sampled alignments in the simulated data sets to the centroid solution (using algorithm #3) or the MAP solution (using algorithm #2).

<sup>c</sup>The average number of predicted sites per data set, when using the centroid estimator (algorithm #3) or the MAP estimator (algorithm #2).

**Table 3.** Results<sup>a</sup> for  $\gamma$ -proteobacterial data sets (real data)

Motif models	Possible sites <sup>b</sup>	Total predictions	True positives	False positives	False negatives	PPV	Sensitivity	Mean distance <sup>c</sup>
1	103	57.7	47.3	10.3	55.7	0.82	0.46	0.20
2	128	74.3	57.3	17.0	70.7	0.77	0.45	0.25
3	132	79.3	61.3	18.0	70.7	0.77	0.46	0.29

<sup>a</sup>The results, averaged over three runs, of the phylogenetic Gibbs sampler with the centroid solution, for the set of sequences of 72 orthologous  $\gamma$ -proteobacterial promoter regions.

<sup>b</sup>The maximum total number of reported sites (TP+FN) that could possibly be located by the algorithm varies with the number of motif models; for example, in a data set that contains seven reported sites—two of type A, one of type B and four of type C—the algorithm could at most locate four sites using one motif model, six sites using two motif models and seven sites using three motif models.

<sup>c</sup>The mean Hamming distance from the centroid to the sampled alignments.

sequence data from the study set, containing binding sites for the transcription factors Crp (25 genes), LexA (eight genes) and TyrR (five genes). Each set contained multiple MASSes from the eight  $\gamma$ -proteobacterial species (Fig. 1A). We ran the algorithm on each of the three data sets using a single palindromic motif model, with the 16 conserved positions allowed to fragment to a maximum width of 24 positions. Table 4 summarizes the results of the phylogenetic Gibbs sampler centroid alignments for the three data sets, averaged over three runs. As expected, these results show a significant improvement in sensitivity, over results obtained from analysis of data from a single gene.

#### 4 DISCUSSION

The ability to predict functional sites among sequences from closely related species will become increasingly important, as more species' genomes are sequenced. Motif discovery algorithms will need to compensate for the high degree of conservation across homologous regions among these evolutionarily related sequences. We have introduced a number of extensions to the recursive Gibbs sampling algorithm (Thompson *et al.*, 2003) that specifically address this issue.

We have incorporated phylogenetic information, based on a user-supplied phylogenetic tree that relates those species contributing the sequence data. This tree is used to draw samples from a fully phylogenetic model in both steps (site sampling and model update) of the algorithm, thus eliminating the need for approximations. The power of incorporating phylogeny was demonstrated by the results in Table 1, which show that in the absence of transcription factor binding sites, all the phylogenetic algorithms tested showed improved specificity over the 'plain' recursive Gibbs sampling algorithm. Further, the centroid version of the phylogenetic Gibbs sampler made far fewer false positive predictions in either simulated data set than did the other versions of the Gibbs sampler, and substantially fewer than did PhyloGibbs. These simulations on null data sets that contain no planted sites aptly illustrate the hazards of failing to address phylogenetic effects when studying closely related species.

Our results also suggest that the algorithms using a full phylogenetic model (Felsenstein, 1981) and the Metropolis–Hastings model update step (Hastings, 1970) display an advantage over algorithms, such as PhyloGibbs (Siddharthan *et al.*, 2005), that use approximations.

Specifically, the phylogenetic Gibbs sampler MAP algorithm exhibited higher sensitivity (Fig. 2) and higher specificity (Table 1) than PhyloGibbs on the simulated proteobacterial data. These results were somewhat surprising, given the fact that both algorithms use an optimization-based approach. One possible explanation for this difference is that the star phylogeny-based model of PhyloGibbs does not accurately describe the relationship among the  $\gamma$ -proteobacterial species. In line with this explanation, the sensitivity of PhyloGibbs did improve on the simulated star-topology yeast data set. However, PhyloGibbs exhibited poor specificity on this data, and therefore the PPV did not improve over the results observed for the simulated proteobacterial data. Thus, some effect on performance of the algorithms may be attributed to the tree topology. While a star topology is appropriate in circumstances where species are somewhat uniformly distant from one another, there exist a number of cases in which the sequencing of related species has produced trees that are not well modeled by a star topology, for example, yeast (Rokas *et al.*, 2003), *Drosophila* (Kopp, 2006), mammals (ENCODE, 2004) and the  $\gamma$ -proteobacterial species under study here.

Furthermore, we believe that a significant portion of the improvement exhibited by the phylogenetic Gibbs sampler over PhyloGibbs is due to the improved model update step. The Metropolis–Hastings model update step ensures statistical 'detailed balance', in that motif models that would otherwise be selected more frequently than is justified are subjected to a probabilistic rejection test. This enables an accurate walk through the search space, where formerly there was approximation. For instance, when we disable the probabilistic rejection test in runs of the phylogenetic Gibbs centroid algorithm on the simulated proteobacterial data, we observe sensitivities of 0, 4, 19 and 49% for data with 1, 2, 3 and 4 planted sites, respectively (data not shown). This is notably poorer than the fully enabled algorithm that yields sensitivities of 23, 61, 80 and 91% for these cases (Supplementary Material, Table S1), and is a strong indication that traditional 'sequence weight' approaches to this task are insufficient.

In analysis of the simulated yeast data, the phylogenetic Gibbs sampler algorithms showed greater improvement in sensitivity than did PhyloGibbs as the number of simulated sites increased, and although the MAP version of the phylogenetic Gibbs sampler had a sensitivity equal to that of the centroid version, the superior specificity (and hence PPV) of the centroid version yielded a better overall performance.

**Table 4.** Results<sup>a</sup> on co-regulated sets of  $\gamma$ -proteobacterial sequence data

Regulon	Genes	TF sites <sup>b</sup>	Total predictions	True positives	False positives	False negatives	PPV	Sensitivity	Mean distance <sup>c</sup>
Crp	25	29	19.7	13.7	6.0	15.3	0.69	0.47	0.30
LexA	8	11	9.0	9.0	0.0	2.0	1.00	0.82	0.00
TyrR	5	8	5.0	5.0	0.0	3.0	1.00	0.62	0.72

<sup>a</sup>The results, averaged over three runs, of applying the phylogenetic Gibbs sampler with the centroid solution, to collections of sequences of orthologous promoter regions from multiple genes known to be regulated by specific transcription factors.

<sup>b</sup>The total number of reported sites for the transcription factor listed in the first column. The counts of true positives, false positives and false negatives are based on this number.

<sup>c</sup>The mean Hamming distance from the centroid to the sampled alignments.

All the phylogenetic algorithms tested, however, showed poorer specificity on the simulated yeast data than they did on the simulated proteobacterial data (Table 1), and in particular, poor sensitivity when there was only one planted 10-mer site in each simulated sequence (Fig. 2D). This latter goal, the location of a single non-palindromic site per sequence, in data sets of the size and phylogenetic relatedness that are common in current studies, has traditionally been nearly impossible to achieve with a reasonable balance of sensitivity and specificity, and thus we do not find these poor sensitivity numbers to be surprising. Though difficult, this scenario (a single, non-palindromic site per sequence) is common in real data; thus, many studies focus on combining sequence data from closely related species with co-expression data, in order to increase sensitivity. Many phylogenetic motif finders were designed with such studies in mind (Li and Wong, 2005; Liu *et al.*, 2004b; Moses *et al.*, 2004; Prakash *et al.*, 2004; Sinha *et al.*, 2004; Wang and Stormo, 2003), and require orthologous promoter sequence data for more than one gene as input. However, such approaches are not appropriate when co-expression or co-regulation data are unavailable, or for those genes and pathways that are the targets of specific regulation, such as LacI regulation of the *lac* operon among enterobacteria. The phylogenetic Gibbs sampler provides the flexibility to study cases such as this, with high positive predictive value.

In addition to the incorporation of phylogeny and the Metropolis–Hastings model update step, we have extended the Gibbs sampler to provide a centroid alignment solution. Sampling methods in RNA secondary structure prediction (Ding *et al.*, 2005) have shown that in high-dimensional discrete inferences, the most probable solution often has low probability, indicating that no single solution can represent the posterior space well. It has been shown in the prediction of RNA secondary structure (Ding *et al.*, 2005) that centroid solutions better represent the full posterior weighted ensemble of solutions, and yield more specific predictions. In line with those observations, we have demonstrated that use of the ensemble centroid alignment yields good sensitivity while avoiding false positives. We interpret these findings in the following way. The MAP versions of the Gibbs sampling algorithm perform additional sampling steps after convergence to a near-optimal solution, and include measures intended to focus sampling in the region near this solution. That is, once a MAP solution is found, the sampler may (as an option) sample among high probability sites, in order to find sites which are sampled reproducibly (i.e. the frequency solution) (Thompson *et al.*, 2003, 2005). In this research, we discovered that inclusion of such steps, that tend to draw samples near the MAP solution, adversely constrained sampling and negatively impacted the sampler's ability to correctly identify sites (data not shown). Thus, when we employed this sample-after-convergence strategy in trials of the phylogenetic algorithm, it almost always returned more false-positive results, indicating that the algorithm was improving the MAP by adding sites that marginally improved the overall posterior probability but had low individual probabilities. Only when we allowed our sampling algorithm to fully explore the posterior space, did accuracy improve. This interpretation is consistent with the findings of Ding and colleagues (Ding *et al.*, 2005) that

show that the MAP is, at times, distant from the high posterior-probability region of the solution space. While we have not made any comparisons with expectation–maximization algorithms, such maximum-likelihood procedures will suffer the same disadvantages as do the optimization-based Gibbs sampling algorithms examined here. Taken together, these findings have raised serious doubts in our minds about the utility of optimization-based approaches for drawing inferences in high-dimensional discrete spaces.

While this research has focused on sets of sequences derived from species with close phylogenetic relationships, it is important to point out that the centroid sampling technique is effective for sequence data from more distantly related species, and for sequence data from co-expression studies in the absence of orthologous data. In cases where the sequences can be treated as independent, the Gibbs centroid sampler shows performance equivalent to, or better than, the Gibbs recursive sampler (Thompson *et al.*, 2007 and data not shown).

It is also important to note that one problem that arises, when centroid solutions and multiple motif models are used, is the non-identifiability of models from finite mixtures, stemming from label switching (Stephens, 2000) among the various restarts of the algorithm. Typically, multiple Gibbs sampling chains are run, in order to avoid the problem of being trapped in a local optimum. The samples from the individual chains are summed to obtain the centroid solution. Summing is performed across all chains and models, for a given position. When multiple motif models are used, it is not uncommon for separate chains to converge to very similar solutions, with motif models labeled differently. Such overlapping sites are likely represented by alternative variations of the same motif model, and we capitalize on this feature to address non-identifiability. However, this approach can break down when sites from two different motif models are closely spaced in sequence. In such a case, it may be necessary to cluster motif models in order to determine the sites to which a particular transcription factor binds (Jensen *et al.*, 2005; Qin *et al.*, 2003).

Our application of the phylogenetic Gibbs sampler centroid algorithm to 72 real  $\gamma$ -proteobacterial promoter sequences, with each promoter analyzed individually using only the orthologous data, showed results consistent with the results from the earlier simulations: PPVs > 75%, and sensitivities > 45% (Table 3). However, one difficulty in evaluating results from these data is that some promoters have regulatory sites that have not yet been discovered or experimentally validated, and thus some of the apparent false positives may actually be unreported genuine sites. This is likely the case in at least a few examples in our data, for which gene expression or promoter fusion experiments have provided evidence for regulation by an identified transcription factor, although the binding site has not been validated by DNaseI footprinting. Specifically, we predict a site upstream of the *edd* gene that is likely a GntR binding site (Murray and Conway, 2005), and a site upstream of the *glyA* gene that is likely an NsrR binding site (Bodenmiller and Spiro, 2006). Another difficulty in evaluating these results is the possible presence of additional regulatory signals in the sequences, such as  $\sigma$ -factor binding sites, ribosome binding sites, attenuators, and terminators of transcription for the upstream gene. We are not scoring predictions that correspond

to such regulatory signals, almost none of which could meet the criterion we used for inclusion as a validated site (DNaseI footprint identification), although some predictions very likely do correspond to such signals. For example, our prediction that is upstream of the *pyrD* gene is a likely translational attenuator (Neuhard and Kelln, 1996). Accordingly, we are likely underestimating the value of the PPV in Table 3.

Naturally, in the situation when sequence data from a set of several co-regulated genes are analyzed, the sensitivity of the phylogenetic Gibbs sampler improves (Table 4). As noted above however, it is likely that many predicted sites reported in Table 4 as false positives are instead not-yet-reported transcription factor binding sites. In particular, we observe false-positive predictions for the Crp regulon. However, there are often multiple Crp binding sites in a promoter region (Kolb et al., 1993), and the presence of additional binding sites within some of the 25 promoter regions used here is not unlikely.

By extending the recursive Gibbs sampler so as to incorporate phylogeny and provide a centroid alignment solution, we now have in hand the tools necessary to enable analysis of promoter data (for a single gene or a group of co-regulated genes) from closely related species. The centroid-finding phylogenetic Gibbs sampler described here provides the flexibility to analyze orthologous promoter data from a single gene, so as to identify multiple sites and multiple motif types with higher positive predictive value and sensitivity, relative to previous versions of the Gibbs sampler (Thompson et al., 2003, 2005). The ability to examine the regulatory features of a single orthologous gene has already led to the discovery of new and interesting regulatory motifs (McCue et al., 2001; Zhang et al., 2002) that have been experimentally validated, and the phylogenetic Gibbs sampler provides the capability to examine highly correlated sequences toward this same goal.

## ACKNOWLEDGEMENTS

The research is supported by the United States Department of Energy grant DE-FG02-04ER63942 to CEL and LAM, and the United States National Institutes of Health grants K25-HG003291 to LAN, R01-HG01257 to CEL, and 2P20-RR01-5578-06 to Walter Atwood. We thank the Wellcome Trust Sanger Institute and the Washington University Genome Sequencing Center for making preliminary genome sequence data available. The assistance of the Wadsworth Center Bioinformatics Core Facility and the Brown University Center for Computational Molecular Biology is greatly appreciated. PNNL is operated by Battelle for the US Department of Energy under Contract DE-AC06-76RLO 1830.

*Conflict of Interest:* none declared.

## REFERENCES

Achtman, M. et al. (1999) *Yersinia pestis*, the cause of plague, is a recently emerged clone of *Yersinia pseudotuberculosis*. *Proc. Natl Acad. Sci. USA*, **96**, 14043–14048.

Bodenmiller, D.M. and Spiro, S. (2006) The *yjeB* (*nsrR*) gene of *Escherichia coli* encodes a nitric oxide-sensitive transcriptional regulator. *J. Bacteriol.*, **188**, 874–881.

Chain, P.S. et al. (2004) Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. *Proc. Natl Acad. Sci. USA*, **101**, 13826–13831.

Clark, A. et al. (2003) Proposal for *Drosophila* as a model system for comparative genomics. <http://flybase.net/.data/docs/CommunityWhitePapers/GenomesWP2003.html>.

Ding, Y. et al. (2005) RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA*, **11**, 1157–1166.

Durbin, R. et al. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK.

ENCODE Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.

Felsenstein, J. (2002) *PHYMLIP (Phylogeny Inference Package) version 3.6a3*. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Hastings, W.K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Jensen, S.T. et al. (2005) Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes. *Bioinformatics*, **21**, 3832–3839.

Kolb, A. et al. (1993) Transcriptional regulation by cAMP and its receptor protein. *Annu. Rev. Biochem.*, **62**, 749–795.

Kopp, A. (2006) Basal relationships in the *Drosophila melanogaster* species group. *Mol. Phylogenet. Evol.*, **39**, 787–798.

Li, X. and Wong, W.H. (2005) Sampling motifs on phylogenetic trees. *Proc. Natl Acad. Sci. USA*, **102**, 9481–9486.

Liu, C. et al. (2002) Reduction kinetics of Fe(III), Co(III), U(VI), Cr(VI), and Tc(VII) in cultures of dissimilatory metal-reducing bacteria. *Biotechnol. Bioeng.*, **80**, 637–649.

Liu, J.S. (2001) *Monte Carlo Strategies in Scientific Computing*. Springer-Verlag, New York, NY.

Liu, J.S. et al. (2004a) Statistical models for biological sequence motif discovery. In: *Case Studies in Bayesian Statistics*. Carnegie Mellon University, Pittsburgh, PA, pp. 1–18.

Liu, J.S. et al. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Amer. Stat. Assoc.*, **90**, 1156–1170.

Liu, Y. et al. (2004b) Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Res.*, **14**, 451–458.

McCue, L. et al. (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.*, **29**, 774–782.

Moses, A.M. et al. (2004) Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pac. Symp. Biocomput.*, 324–335.

Murray, E.L. and Conway, T. (2005) Multiple regulators control expression of the Entner-Doudoroff aldolase (Eda) of *Escherichia coli*. *J. Bacteriol.*, **187**, 991–1000.

Neuhard, J. and Kelln, R.A. (1996) Biosynthesis and conversions of pyrimidines. In: Neidhardt, F.C. (ed.) *Escherichia coli and Salmonella, Cellular and Molecular Biology*. ASM Press, Washington DC, pp. 580–599.

Newberg, L.A. et al. (2005) The relative inefficiency of sequence weights approaches in determining a nucleotide position weight matrix. *Stat. Appl. Genet. Mol. Biol.*, **4**, Article13.

Prakash, A. et al. (2004) Motif discovery in heterogeneous sequence data. *Pac. Symp. Biocomput.*, **9**, 348–359.

Qin, Z.S. et al. (2003) Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nat. Biotechnol.*, **21**, 435–439.

Rokas, A. et al. (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, **425**, 798–804.

Siddharthan, R. et al. (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.*, **1**, e67.

Sinha, S. et al. (2004) PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinform.*, **5**, 170.

Stephens, M. (2000) Dealing with label switching in mixture models. *J. Royal Stat. Soc.*, **62**, 795–809.

Thompson, W. et al. (2005) Using the Gibbs motif sampler to find conserved domains in DNA and protein sequences. In: Baxevanis, A.D. et al. (ed.)

- Current Protocols in Bioinformatics*. John Wiley & Sons, Inc., New York, NY, pp. 2.8.1–2.8.38.
- Thompson,W. *et al.* (2004) Decoding human regulatory circuits. *Genome Res.*, **14**, 1967–1974.
- Thompson,W. *et al.* (2003) Gibbs recursive sampler: finding transcription factor binding sites. *Nucleic Acids Res.*, **31**, 3580–3585.
- Thompson,W.A. *et al.* (2007) The Gibbs centroid sampler. *Nucleic Acids Res.*, (web server issue), **35**, W232–W237.
- Wang,T. and Stormo,G.D. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, **19**, 2369–2380.
- Zhang,Y.M. *et al.* (2002) The FabR (YijC) transcription factor regulates unsaturated fatty acid biosynthesis in *Escherichia coli*. *J. Biol. Chem.*, **277**, 15558–15565.