W232–W237 *Nucleic Acids Research, 2007, Vol. 35, Web Server issue*
doi:10.1093/nar/gkm265

The Gibbs Centroid Sampler

William A. Thompson[1,*], Lee A. Newberg[2,3], Sean Conlan[4], Lee Ann McCue[5] and Charles E. Lawrence[1]

[1]Center for Computational Molecular Biology and the Division of Applied Mathematics, Brown University, Providence, RI 02912, USA, [2]The Wadsworth Center, New York State Department of Health, Albany, NY 12201, USA, [3]Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180, USA, [4]Columbia University, New York, NY 10032, USA and [5]Pacific Northwest National Laboratory, Richland, WA 99352, USA

Received January 31, 2007; Revised March 27, 2007; Accepted April 8, 2007

ABSTRACT

The Gibbs Centroid Sampler is a software package designed for locating conserved elements in biopolymer sequences. The Gibbs Centroid Sampler reports a centroid alignment, i.e. an alignment that has the minimum total distance to the set of samples chosen from the *a posteriori* probability distribution of transcription factor binding-site alignments. In so doing, it garners information from the full ensemble of solutions, rather than only the single most probable point that is the target of many motif-finding algorithms, including its predecessor, the Gibbs Recursive Sampler. Centroid estimators have been shown to yield substantial improvements, in both sensitivity and positive predictive values, to the prediction of RNA secondary structure and motif finding. The Gibbs Centroid Sampler, along with interactive tutorials, an online user manual, and information on downloading the software, is available at: http://bayesweb.wadsworth.org/gibbs/gibbs.html.

INTRODUCTION

The identification of transcription factor binding sites (TFBSs) in the promoters of genes is a critical step in the delineation of the genetic regulatory network of an organism. A number of motif discovery algorithms have been developed over the past decade and a half, for the detection of *cis*-regulatory sites (1). Most of these algorithms depend, in one way or another, on finding an optimal alignment of motif sites. In this article, we describe the web server for an improved motif discovery algorithm, the Gibbs Centroid Sampler, which finds a centroid alignment. The centroid alignment is the alignment that has the minimum total distance to the set of samples chosen from the *a posteriori* probability distribution of TFBS alignments. By focusing on the region of solution space containing the most posterior probability, rather than on the single solution that is most probable, this approach significantly enhances the predictive power of the algorithm. In computational experiments using simulated proteobacterial and yeast data (2), the centroid sampler showed improved specificity and positive predictive value over algorithms that report an optimal solution.

The Gibbs Centroid Sampler is an improved version of the Gibbs Recursive Sampler (3), which has been used extensively in the identification of TFBSs (4–8), and has been available at our Web site for some time (3,9). The software currently available at the Web site retains all of the features of the previous versions, including searches for multiple motif types, multiple instances (sites) of a motif, palindromic motifs, motifs of varying widths and a heterogeneous background frequency model (see (3) for descriptions of these and other features). The users' choices of options are entered through a web form, described below, and the output is returned to the user via e-mail. In addition to the new algorithmic features, the Web site has been updated to include extensive tutorials on the use of the Gibbs sampling software for prokaryotic phylogenetic footprinting and for the analysis of prokaryotic co-expression data.

The Gibbs Centroid Sampler

A key feature of most sequence-based Gibbs sampling and expectation maximization algorithms (10,11), is the use of a probabilistic score that is maximized. Typically, the alignment that has the maximum of this score is reported to the user. Previous versions of the Gibbs Sampler used the posterior probability of the alignment, called the MAP (maximum *a posteriori* probability) (12), as a measure of the quality of the alignment, and thus the alignment that produced the highest posterior probability (i.e. the MAP alignment) was returned. The reported MAP was calculated as the logarithm of the alignment probability minus

*To whom correspondence should be addressed. Tel: +1-401-863-2048; Email: william_thompson_1@brown.edu the logarithm of an empty or background alignment. Thus, the reported value was a measure of the extent to which a particular alignment was better than background.

The use of methods such as this, which seek to obtain global or local optimal solutions to inference problems, is common in computational biology. Typically, however, the probability of even the best arrangement of motif sites is extremely small. That is, since motif detection is a high-dimensional problem, from a Bayesian viewpoint, the data likelihood will contain an immense number of terms, of which the optimal solution is simply one. From this perspective, the question arises, 'How representative is the optimum when its probability is very small compared to the overall probability mass?'

It has been shown in RNA secondary structure prediction (13) and TFBS discovery algorithms (2,14) that reliance on the optimal solution can be misleading and can adversely affect prediction accuracy. Specifically, Ding *et al.* (13,15) showed that centroid estimates reduced errors in RNA secondary structure prediction by 30%, while simultaneously improving sensitivity, and Newberg *et al.* (2) showed similar substantial improvements over algorithms finding local optima for TFBS discovery in sequences from phylogenetically closely related species. Centroid solutions garner information from the full ensemble of solutions, while MAP solutions focus exclusively on the single most probable point.

The centroid sampling algorithm

The user supplies to the algorithm a collection of sequences in FASTA format and enters several parameters, such as motif widths, as described below. The centroid algorithm begins in a manner similar to previous Gibbs sampling algorithms. It is initialized with a, typically random, alignment. From this alignment, motif models are calculated (12). The sampling procedure then proceeds through the following steps:

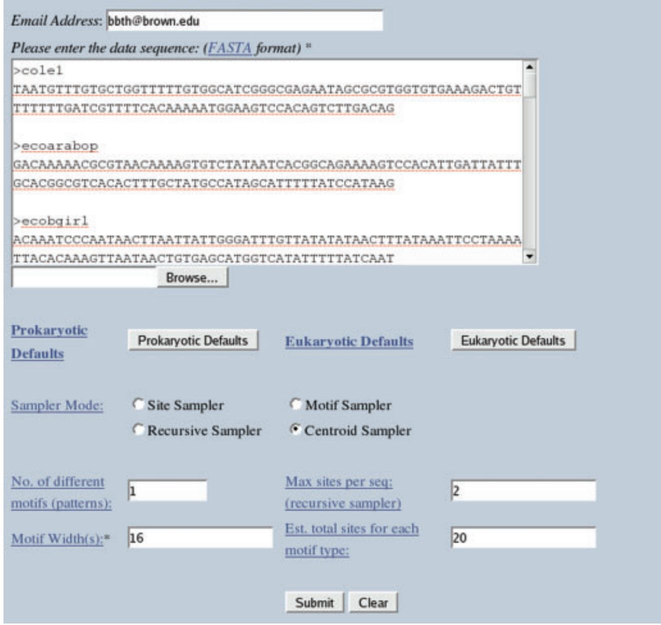
(i) A sequence is selected, and the probability of each possible number of sites, up to the maximum specified by the user, is calculated based on the current model;
(ii) the number of sites is sampled;
(iii) the predicted positions and types of the sites are sampled based on their probabilities, calculated as described by Thompson *et al.* (3);
(iv) the motif models are updated based on the sampled sites in all sequences.

An iteration of the algorithm consists of the completion of Steps 1–4 for each sequence. 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 repeat the above procedure through a burn-in period, typically 2000 iterations. The burn-in period is required for the sampler to move away from transient effects of the particular initial conditions. After the burn-in period, the sampler proceeds, again through a fixed number of iterations (typically 8000). During this sampling process, the algorithm tracks each sampled position. The entire process (burn-in and sampling iterations) is repeated with a number of different random starting alignments called 'seeds'. By default, 20 seeds are used. The samples from each seed are accumulated, and a centroid alignment solution is obtained from the accumulated samples; the centroid is the alignment that minimizes the sum of the pair-wise distances between it and each of the alignments in the collection. Thus, the centroid is defined in terms of a distance measure between pairs of proposed alignments. The centroid alignment is calculated via a dynamic programming algorithm.

In previous versions of the sampler, the model update step (Step 4 above) was accomplished using the predictive update method (12). The centroid sampler performs the model update step by sampling a new model from the posterior Dirichlet distribution of motif or background models. Starting with the existing model Θ, the algorithm draws a new model, Θ_p, using the motif or background counts from *Dir* $(c + \beta)$, where *Dir* is the Dirichlet distribution, and $\mathbf{c}$ and β are the current count and pseudo-count vectors. While predictive update works when at most one new binding site is chosen between motif model updates, it is not entirely appropriate in the present context, where multiple binding sites are chosen between model updates. This new model update method is of greatest value in the identification of sites among aligned sequences derived from multiple phylogenetically related species (2).

The Gibbs Sampler Web Site

The Gibbs Sampler Web site consists of three layers, each offering an increasing number of options for control of the sampling process. The first page, shown in Figure 1, allows the user to input sequences, select the version of the Gibbs Sampler, and control the basic motif parameters (16).


Email Address: bbth@brown.edu

Please enter the data sequence: (FASTA format) *

```
>cole1
TAATGTTTGTGCTGGTTTTTGTGGCATCGGGCGAGAATAGCGCGTGGTGTGAAAGACTGT
TTTTTTGATCGTTTTCACAAAAATGGAAGTCCACAGTCTTGACAG

>ecoarabop
GACAAAAACGCGTAACAAAAGTGTCTATAATCACGGCAGAAAAGTCCACATTGATTATTT
GCACGGCGTCACACTTTGCTATGCCATAGCATTTTTATCCATAAG

>ecobgirl
ACAAATCCCAATAACTTAATTATTGGGATTTGTTATATATAACTTTATAAATTCCTAAAA
TTACACAAAGTTAATAACTGTGAGCATGGTCATATTTTTATCAAT
```

Sampler Mode: ○ Site Sampler ○ Motif Sampler ○ Recursive Sampler ● Centroid Sampler

No. of different motifs (patterns): 1

Max sites per seq: (recursive sampler) 2

Motif Width(s):* 16

Est. total sites for each motif type: 20

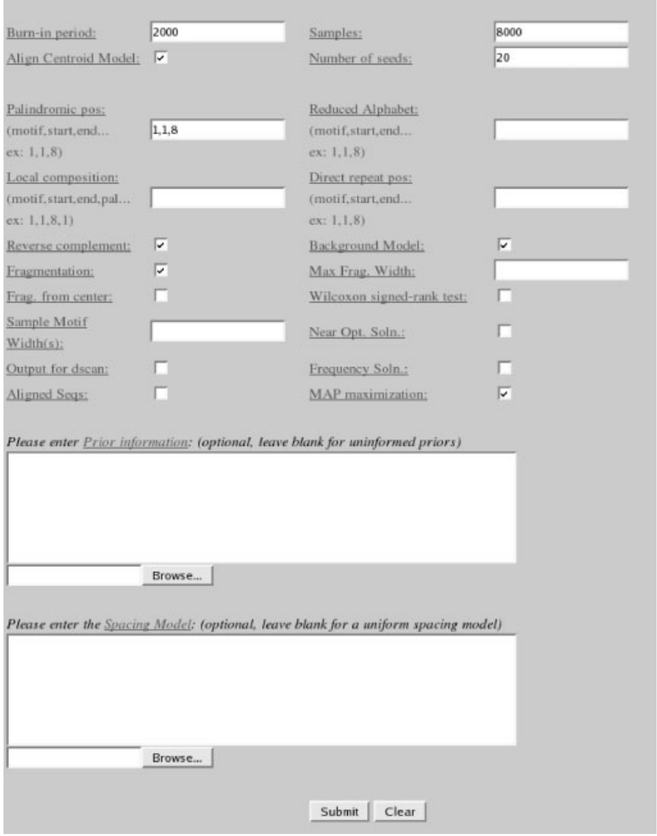
Figure 1. The basic Gibbs Centroid Sampler entry screen.

While we continue to make earlier versions available for selection on this page, in most circumstances the centroid sampler should return better results (2). An e-mail address, a set of sequences in FASTA format, an optional initial guess of the total number of sites, the number of conserved positions in the motif sites, and the maximum allowable number of sites in any one sequence are entered on this page. The estimate of the number of sites affects the initial starting solution for the burn-in process. If it is not supplied, the default of one site for each motif type for each sequence is used. We have found this default adequate for most datasets, and the centroid sampler is relatively insensitive to reasonably small changes in this value. The number of conserved positions in the motif model(s) is a required parameter. This value sets the minimum width of the predicted sites, although sites may fragment to a greater width by the inclusion of non-conserved positions (12). Motif widths for multiple models can be entered, although it is best to use no more motif models than is reasonable given the number of expected TFBS types. Increasing the number of motif models beyond the number of relevant site types should not adversely affect the solutions, if the number of burn-in and sample iterations is adequate (described below), because extra models will not sample sites sufficiently to be included in the centroid. However, as the number of models increases, the program runtime increases (described below). The maximum number of sites in a single sequence is also a required parameter for the centroid sampler. The value entered for this parameter should be based on knowledge of the biological system under study. For example, when analyzing bacterial intergenic sequences for TFBSs, a value of two or three is typically used, whereas for eukaryotic data, this number is typically set higher. This parameter sets the maximum for the total sum of all motif sites in any one sequence. The sequence data can be pasted into the entry window or uploaded from a file. Each entry field has an associated hyperlink, which leads to a page describing the required data format. From this entry screen, default options will be automatically selected for the sampling parameters. The defaults for the centroid sampler include the use of a heterogeneous background model (16), 20 random seeds, a burn-in period of 2000 iterations and a sampling period of 8000 iterations.

Control of sampling parameters

Selection of the 'Show Advanced Options' link opens a page with several more options (Figure 2). Most of these, such as options for palindromic models, fragmentation, the Wilcoxon signed-rank test and the number of random seeds, are available for all sampling modes (site, motif, recursive and centroid) and have been described earlier (3,9) New options for controlling the behavior of the centroid sampler are now also presented on this page. The 'Burn-in Period' and 'Samples' fields control the numbers of burn-in and sampling iterations for each seed; these fields are disabled when non-centroid sampling modes are selected. Initially, when the centroid sampler is selected, the 'Burn-in Period' and 'Samples' fields contain default values. We have found the defaults of 2000 iterations for burn-in and 8000 sampling iterations to be broadly applicable for prokaryotic or eukaryotic data of modest size. However, for small datasets, in the order of 10 to 20 sequences, each of <200 nucleotides, our experience has shown that the burn-in and sample iterations can be reduced (to 1000 and 4000, respectively) without adversely affecting the results. Conversely, for large datasets (>50 sequences, each of 5000 to 10,000 nucleotides) where the TFBS are likely short and not well conserved, as is common in eukaryotic sequences, the number of iterations should be increased for both parameters.

It is important to note a difficulty that can arise when the centroid sampler is used with multiple motif models; specifically, the non-indentifiability of models from finite mixtures, stemming from label switching (17) among the various restarts of the algorithm. Gibbs sampling is inherently a stochastic procedure; in order to avoid being trapped in regions of low probability, the sampling process is restarted a number of different times with different starting seeds. When multiple motif models are used, the separate seeds can converge to similar solutions, with different orderings of the motif models. For example, in the case of two motif models, a particular seed may converge to a set of sites for model A and sites for model B. Another seed may converge to the same overall collection of sites, but with the sites previously labeled as



Burn-in period:	2000	Samples:	8000
Align Centroid Model:	☑	Number of seeds:	20
Palindromic pos: (motif,start,end... ex: 1,1,8)	1,1,8	Reduced Alphabet: (motif,start,end... ex: 1,1,8)	
Local composition: (motif,start,end,pal... ex: 1,1,8,1)		Direct repeat pos: (motif,start,end... ex: 1,1,8)	
Reverse complement:	☑	Background Model:	☑
Fragmentation:	☑	Max Frag. Width:	
Frag. from center:	□	Wilcoxon signed-rank test:	□
Sample Motif Width(s):		Near Opt. Soln.:	□
Output for dscan:	□	Frequency Soln.:	□
Aligned Seqs:	□	MAP maximization:	☑

Please enter Prior information: (optional, leave blank for uninformed priors)

Please enter the Spacing Model: (optional, leave blank for a uniform spacing model)

Figure 2. The Gibbs Centroid 'Advanced options' entry page.

model A now labeled as model B, and sites previously labeled as model B now labeled as model A. The centroid solution is obtained by summing the number of times a given position (i.e. site) is sampled across all restarts and models, which means that sites from multiple models are not separated in the output. Furthermore, different fragmentation models (12) can be generated among the different seed runs, giving rise to a collection of centroid sites that differ in length, and making it difficult to visualize the TFBSs in a more traditional probability matrix representation.

To address these two difficulties, the selection of the 'Align Centroid Model' option causes the Gibbs Centroid Sampler to use the Gibbs Recursive Sampler to align the collection of centroid sites. In the case of multiple models, this process will separate the sites into related groups, and thus aid identification of the different site types. This process can also give the user insight into which positions in the models are highly conserved. It is important to note that the resulting alignment is neither a MAP alignment nor a centroid alignment of the complete set of data sequences. It is provided only to lend additional insight into the centroid solution.

Program output

Program output is returned via e-mail. The initial portion of the Gibbs Centroid Sampler output is identical to that of the other versions of the sampler, simply providing a list of the options used for the current run, followed by a list of the FASTA headings for the input sequences (see (3) for an example). Following these is the list of the sites making up the centroid model. Figure 3 shows the results for a set of 18 *Escherichia coli* sequences; these sequences are well studied, known to contain binding sites for the cyclic AMP receptor protein (Crp) (11), and are provided as a test dataset when the Gibbs Sampler software is downloaded. The results in Figure 3 were generated using the centroid sampler with a motif width of 16, a palindromic motif model requirement, a maximum number of sites per sequence of two, heterogeneous background composition, the default number of restarts (20 seeds), the default burn-in (2000 iterations) and the default centroid sampling periods (8000 iterations). The motif models were allowed to fragment to a width of 24 bases.

At the top of Figure 3 is the set of sites making up the centroid; the centroid sites are listed in upper case, and flanking positions are in lower case. The sites correspond well with the DNaseI footprinted sites for these sequences (11). The variation in the length of the sites is a result of different fragmentation models generated during the sampling periods (mentioned above). The dynamic program that calculates the centroid can be found elsewhere [see the supplementary material for (2)]. The legend below the list of sites identifies the various columns of the output. The probability column shows the sampling frequencies for these sites. These sampling frequencies are an estimate of the probabilities that the cognate transcription factors bind at the predicted sites.

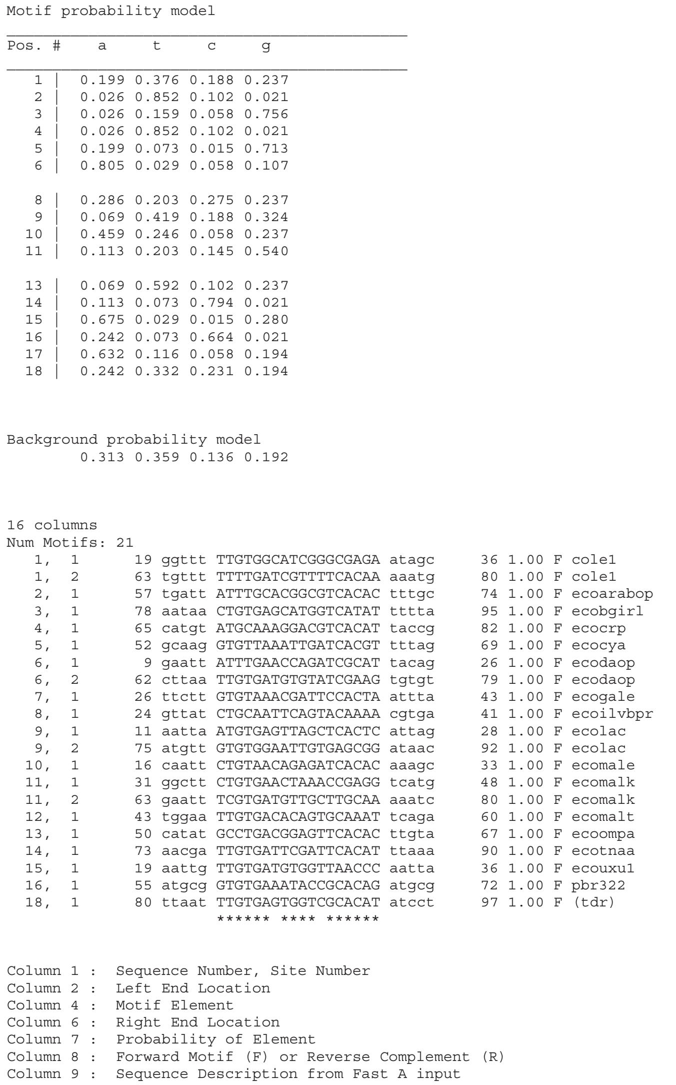
The second part of Figure 3 shows an alignment of the centroid sites. The program generates this alignment by taking the collection of sites in the centroid, plus their flanking sequences, and using the Gibbs Recursive Sampler to find the best alignment among this set of sites, with at most one site in each sequence. As such, this is neither a centroid nor an optimal alignment. It is provided simply to allow the user to identify different site types (when multiple motif models were used) and to visualize which positions are highly conserved in the centroid sites. The format of this alignment is identical to that of the Gibbs Recursive Sampler previously described in (3).

```
======================================================================
========================== CENTROID RESULTS ==========================
======================================================================

    1, 1      11 tttgt GCTGGTTTTTGTGGCATCGGGCG  agaat      33  0.64 cole1
    1, 2      54 gtgaa AGACTGTTTTTTTGATCGTTTTC  acaaa      76  0.98 cole1
    2, 1      56 ttgat TATTTGCACGGCGTCACAC      tttgc      74  0.98 ecoarabop
    3, 1      78 aataa CTGTGAGCATGGTCATATTTTTA  tcaat     100  0.89 ecobgirl
    4, 1      55 tgatg TACTGCATGTATGCAAAGGACGT  cacat      77  0.82 ecocrp
    5, 1      48 atcag CAAGGTGTTAAATTGATCACGTT  ttaga      70  0.72 ecocya
    6, 1       5  agtg AATTATTTGAACCAGATCGCATTA cagtg      28  0.97 ecodaop
    6, 2      67 ttgtg ATGTGTATCGAAGTGTGTTGCGG  agtag      89  0.83 ecodaop
    7, 1      31 gtgta AACGATTCCACTAATTTATTCCA  tgtca      53  0.89 ecogale
    8, 1      29 ctgca ATTCAGTACAAAACGTGATCAAC  ccctc      51  0.89 ecoilvbpr
    9, 1       8 cgcaa TTAATGTGAGTTAGCTCACTC    attag      28  0.97 ecolac
    9, 2      73 gtatg TTGTGTGGAATTGTGAGCGGATA  acaat      95  0.66 ecolac
   10, 1      11 accgc CAATTCTGTAACAGAGATCAC    acaaa      31  0.97 ecomale
   11, 1      31 ggctt CTGTGAACTAAACCGAGGTCATG  taagg      53  0.50 ecomalk
   11, 2      56 atgta AGGAATTTCGTGATGTTGCTT    gcaaa      76  0.78 ecomalk
   12, 1      41 tttgg AATTGTGACACAGTGCAAATTCA  gacac      63  0.93 ecomalt
   13, 1      48 ttcat ATGCCTGACGGAGTTCACACTTG  taagt      70  0.79 ecoompa
   14, 1      78 ttgtg ATTCGATTCACATTTAAACAA    tttca      98  0.89 ecotnaa
   15, 1      15 gtgaa ATTGTTGTGATGTGGTTAACCCA  attag      37  0.53 ecouxu1
   16, 1      53 atatg CGGTGTGAAATACCGCACAGATG  cgtaa      75  0.83 pbr322
   18, 1      75 gaaag TTAATTTGTGAGTGGTCGCACAT  atcct      97  0.99 (tdr)
Num Sites: 21

Column 1 :  Sequence Number, Site Number
Column 2 :  Left End Location
Column 4 :  Motif Element
Column 6 :  Right End Location
Column 7 :  Probability of Element
Column 8 :  Sequence Description from FastA input

======================================================================
======================= Aligned Centroid Sites =======================
======================================================================

-------------------------------------------------------------------------
                             MOTIF a

Motif model (residue frequency x 100)
____________________________________________
Pos. #     a   t   c   g  Info
_______________________________
   1 |    19  38  19  23  0.0
   2 |     .  90   9   .  1.0
   3 |     .  14   4  80  1.0
   4 |     .  90   9   .  1.0
   5 |    19   4   .  76  0.9
   6 |    85   .   4   9  0.9

   8 |    28  19  28  23  0.1
   9 |     4  42  19  33  0.2
  10 |    47  23   4  23  0.1
  11 |     9  19  14  57  0.4

  13 |     4  61   9  23  0.3
  14 |     9   4  85   .  1.4
  15 |    71   .   .  28  0.8
  16 |    23   4  71   .  1.1
  17 |    66   9   4  19  0.4
  18 |    23  33  23  19  0.0

nonsite   28  32  16  22
site      25  28  19  26
```

Figure 3. Output from the Gibbs Centroid Sampler.

Performance

The underlying algorithm for the Gibbs Centroid Sampler and the Gibbs Recursive Sampler is a forward–backward algorithm (7). The forward step is the most compute intensive part of the algorithm, with runtime increasing as the square of the length of the individual sequences; thus, the most important factor affecting runtime is the length of the individual sequences. Other parameters, such as



```
Motif probability model
______________________________________________
Pos. #    a     t     c     g
______________________________________________
   1 |  0.199 0.376 0.188 0.237
   2 |  0.026 0.852 0.102 0.021
   3 |  0.026 0.159 0.058 0.756
   4 |  0.026 0.852 0.102 0.021
   5 |  0.199 0.073 0.015 0.713
   6 |  0.805 0.029 0.058 0.107

   8 |  0.286 0.203 0.275 0.237
   9 |  0.069 0.419 0.188 0.324
  10 |  0.459 0.246 0.058 0.237
  11 |  0.113 0.203 0.145 0.540

  13 |  0.069 0.592 0.102 0.237
  14 |  0.113 0.073 0.794 0.021
  15 |  0.675 0.029 0.015 0.280
  16 |  0.242 0.073 0.664 0.021
  17 |  0.632 0.116 0.058 0.194
  18 |  0.242 0.332 0.231 0.194


Background probability model
        0.313 0.359 0.136 0.192


16 columns
Num Motifs: 21
   1,  1      19 ggttt TTGTGGCATCGGGCGAGA atagc     36 1.00 F cole1
   1,  2      63 tgttt TTTTGATCGTTTTCACAA aaatg     80 1.00 F cole1
   2,  1      57 tgatt ATTTGCACGGCGTCACAC tttgc     74 1.00 F ecoarabop
   3,  1      78 aataa CTGTGAGCATGGTCATAT tttta     95 1.00 F ecobgirl
   4,  1      65 catgt ATGCAAAGGACGTCACAT taccg     82 1.00 F ecocrp
   5,  1      52 gcaag GTGTTAAATTGATCACGT tttag     69 1.00 F ecocya
   6,  1       9 gaatt ATTTGAACCAGATCGCAT tacag     26 1.00 F ecodaop
   6,  2      62 cttaa TTGTGATGTGTATCGAAG tgtgt     79 1.00 F ecodaop
   7,  1      26 ttctt GTGTAAACGATTCCACTA attta     43 1.00 F ecogale
   8,  1      24 gttat CTGCAATTCAGTACAAAA cgtga     41 1.00 F ecoilvbpr
   9,  1      11 aatta ATGTGAGTTAGCTCACTC attag     28 1.00 F ecolac
   9,  2      75 atgtt GTGTGGAATTGTGAGCGG ataac     92 1.00 F ecolac
  10,  1      16 caatt CTGTAACAGAGATCACAC aaagc     33 1.00 F ecomale
  11,  1      31 ggctt CTGTGAACTAAACCGAGG tcatg     48 1.00 F ecomalk
  11,  2      63 gaatt TCGTGATGTTGCTTGCAA aaatc     80 1.00 F ecomalk
  12,  1      43 tggaa TTGTGACACAGTGCAAAT tcaga     60 1.00 F ecomalt
  13,  1      50 catat GCCTGACGGAGTTCACAC ttgta     67 1.00 F ecoompa
  14,  1      73 aacga TTGTGATTCGATTCACAT ttaaa     90 1.00 F ecotnaa
  15,  1      19 aattg TTGTGATGTGGTTAACCC aatta     36 1.00 F ecouxu1
  16,  1      55 atgcg GTGTGAAATACCGCACAG atgcg     72 1.00 F pbr322
  18,  1      80 ttaat TTGTGAGTGGTCGCACAT atcct     97 1.00 F (tdr)
                       ****** **** ******


Column 1 :  Sequence Number, Site Number
Column 2 :  Left End Location
Column 4 :  Motif Element
Column 6 :  Right End Location
Column 7 :  Probability of Element
Column 8 :  Forward Motif (F) or Reverse Complement (R)
Column 9 :  Sequence Description from Fast A input
```

Figure 3. Continued.

the number of sequences, the number of motif models, the number of seeds and the number of iterations, affect the runtime linearly. Therefore, due to the increased number of iterations for burn-in and sampling, the runtime of the centroid sampler is somewhat greater than that of the Gibbs Recursive Sampler. Additional parameters, such as the use of palindromic or direct repeat models, while not directly affecting the runtime of the centroid sampler, greatly improve its ability to discover realistic TFBS by taking into account the biological characteristics of the system under study. The program lists the total execution time for the program as the last line of the output.

Web-based tutorials

The Gibbs Sampler Web site contains tutorials for prokaryotic phylogenetic footprinting (http://bayesweb.wadsworth.org/web_help.PF.html) and for analysis of prokaryotic co-expression data from microarray and promoter fusion experiments (http://bayesweb.wadsworth.org/web_help_text.CE.html). Links to these pages are provided on the main Gibbs entry pages. The tutorials provide guidance to users for all the sampling modes available (site, motif, recursive and centroid), and for both the Gibbs Sampler web server and the stand-alone version of Gibbs. Specifically, the Gibbs Sampler offers a large array of options, some of which are used to model-specific aspects of biological sequences, while others are meant to control details of how the sampling is done. The tutorials focus on the options that are useful in modeling the biology of transcription regulation. The particular examples presented in the tutorials are drawn from the studies presented in (5,6,8,18). Each tutorial gives the command line used to run the analysis, a description of each parameter and why its particular value was chosen, and a link that will automatically run the data on the Gibbs Web site with the Gibbs Centroid Sampler or with the Gibbs Recursive Sampler. The data from the examples can also be downloaded to be run with the stand-alone version. It is important to note that Gibbs sampling is a stochastic process, and thus results run from the links may differ slightly from the examples. In addition, although the examples in these interactive tutorials use prokaryotic sequence data, the principles described and the reasoning behind how to choose parameters are species-independent; all sampling modes, including the Gibbs Centroid Sampler, can be readily applied to the analysis of eukaryotic sequences.

The tutorials, besides presenting detailed examples of the use of the Gibbs software, provide insights into the interpretation of, and biological reasoning behind, the computational experiments. The tutorial examples illustrate how solutions from MAP-based samplers sometimes include low probability sites in the solution. These sites increase the MAP slightly but may be false positive predictions. The centroid sampler avoids these low probability predictions and is thus less likely to make false positive predictions (2). This is illustrated in the tutorial example, ‘Co-expression data from a microarray study of *M. tuberculosis* genes’, where the data comes from microarray results (18) that report a set of co-expressed genes, a subset of which are likely co-regulated by a common transcription factor. When the Gibbs Recursive Sampler is used on the upstream sequences from these co-expressed genes, the results include several sites with low probability in the MAP solution, whereas the Gibbs Centroid Sampler avoids these low probability sites. The fully Bayesian sampling process that is performed by the Gibbs Centroid Sampler is more robust at eliminating these likely false-positive predictions (2) than the process employed in previous versions of the sampler, where, once a MAP solution was found, the sampler was allowed (as an option) to sample among high probability sites in order to find sites which were sampled reproducibly (i.e. the frequency solution) (3). Since we began using centroid estimates, we have discovered that the inclusion of steps that even partially increase focus on MAP (or near MAP) solutions have a detrimental impact on the correct identification of sites.

Additional features

The Gibbs Centroid Sampler can be used for the analysis of amino-acid sequences. The link from the main Gibbs Web site page leads to a page allowing the entry of amino-acid sequences. The Web site also contains a link to an online user guide, which describes the various parameters and their input formats, has detailed descriptions of the output and lists possible error messages and their causes. The Gibbs Sampler Web site allows a maximum of 1000 sequences of no longer than 10,000 nucleotides in length. Users with larger datasets are directed to use the stand-alone version of the Gibbs Sampler.

ACKNOWLEDGEMENTS

The research is supported by the United States Department of Energy grant DE-FG02-04ER63942 to C.E.L. and L.A.M., and the United States National Institutes of Health grants R01-HG01257 to CEL, 2P20-RR01-5578-06 to Walter Atwood, and K25-HG003291 to L.A.N. 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. Funding to pay the Open Access publication charges for this article was provided by Department of Energy grant DE-FG02-04ER63942.

Conflict of interest statement. None declared.

REFERENCES

1. Sandve,G. and Drablos,F. (2006) A survey of motif discovery methods in an integrated framework. *Biology Direct*., **1**, 11.
2. Newberg,L., Thompson,W.A., Conlan,S.P., Smith,T.M., McCue,L.A. and Lawrence,C.E. (2007) A phylogenetic Gibbs sampler that yields centroid solutions for *cis* regulatory site prediction. *Bioinformatics*, Accepted.
3. Thompson,W., Rouchka,E.C. and Lawrence,C.E. (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res*., **31**, 3580–3585.
4. Conlan,S., Lawrence,C. and McCue,L.A. (2005) Rhodopseudomonas palustris Regulons Detected by Cross-Species Analysis of Alphaproteobacterial Genomes. *Appl. Environ. Microbiol*., **71**, 7442–7452.
5. McCue,L., Thompson,W., Carmack,C., Ryan,M.P., Liu,J.S., Derbyshire,V. and Lawrence,C.E. (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res*., **29**, 774–782.
6. McCue,L.A., Thompson,W., Carmack,C.S. and Lawrence,C.E. (2002) Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res*., **12**, 1523–1532.
7. Thompson,W., Palumbo,M.J., Wasserman,W.W., Liu,J.S. and Lawrence,C.E. (2004) Decoding Human Regulatory Circuits. *Genome Res*., **14**, 1967–1974.
8. Florczyk,M.A., McCue,L.A., Purkayastha,A., Currenti,E., Wolin,M.J. and McDonough,K.A. (2003) A Family of acr-Coregulated Mycobacterium tuberculosis Genes Shares a Common DNA Motif and Requires Rv3133c (dosR or devR) for Expression. *Infect. Immun*., **71**, 5332–5343.
9. Thompson,W., McCue,L.A. and Lawrence,C.E. (2005) In Baxevanis,A.D., Davison,D.B., Page,R.D.M., Petsko,G.A., Stein,L.D. and Stormo,G.D. (eds), *Current Protocols in Bioinformatics*, John Wiley & Sons, Inc., New York, NY, pp. 2.8.1–2.8.38.
10. Bailey,T.L. and Elkan,C. (1995) Unsupervised Learning of Multiple Motifs in Biopolymers using EM. *Mach Learn*, **21**, 51–80.
11. 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: Struct. Funct. Genet*., **7**, 41–51.
12. Liu,J., Neuwald,A. and Lawrence,C. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *JASA*, **90, 432**, 1156–1170.
13. Ding,Y.E., Chan,C.Y. and Lawrence,C.E. (2005) RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA*, **11**, 1157–1166.
14. Thompson,W., Conlan,S., McCue,L.A. and Lawrence,C.E. (2007) In Bergman,N. (ed.), *Methods in Molecular Biology, Comparative Genomics*, Humana Press, **1,** 403–423.
15. Ding,Y., Chan,C.Y. and Lawrence,C.E. (2006) Clustering of RNA Secondary Structures with Application to Messenger RNAs. *J Mol. Biol*., **359**, 554.
16. Liu,J. and Lawrence,C. (1999) Bayesian inference on biopolymer models. *Bioinformatics*, **15**, 38–52.
17. Stephens,M. (2000) Dealing with label switching in mixture models. *J. R. Stat. Soc.: Series B (Statistical Methodology)*, **62**, 795–809.
18. Sherman,D.R., Voskuil,M., Schnappinger,D., Liao,R., Harrell,M.I. and Schoolnik,G.K. (2001) Regulation of the Mycobacterium tuberculosis hypoxic response gene encoding alpha -crystallin. *PNAS*, **98**, 7534–7539.