



BioOptimizer: a Bayesian scoring function approach to motif discovery

Shane T. Jensen* and Jun S. Liu*

Department of Statistics, Harvard University, Cambridge, MA 02138-2901, USA

Received on September 21, 2003; revised on December 10, 2003; accepted on January 3, 2004

Advance Access publication February 12, 2004

ABSTRACT

Motivation: Transcription factors (TFs) bind directly to short segments on the genome, often within hundreds to thousands of base pairs upstream of gene transcription start sites, to regulate gene expression. The experimental determination of TFs binding sites is expensive and time-consuming. Many motif-finding programs have been developed, but no program is clearly superior in all situations. Practitioners often find it difficult to judge which of the motifs predicted by these algorithms are more likely to be biologically relevant.

Results: We derive a comprehensive scoring function based on a full Bayesian model that can handle unknown site abundance, unknown motif width and two-block motifs with variable-length gaps. An algorithm called BioOptimizer is proposed to optimize this scoring function so as to reduce noise in the motif signal found by any motif-finding program. The accuracy of BioOptimizer, which can be used in conjunction with several existing programs, is shown to be superior to using any of these motif-finding programs alone when evaluated by both simulation studies and application to sets of co-regulated genes in bacteria. In addition, this scoring function formulation enables us to compare objectively different predicted motifs and select the optimal ones, effectively combining the strengths of existing programs.

Availability: BioOptimizer is available for download at www.fas.harvard.edu/~junliu/BioOptimizer/

Contact: jensen@stat.harvard.edu

1 INTRODUCTION

Transcription factors (TFs), proteins specializing in the regulation of transcription, can bind directly onto the DNA double-helix in close proximity to the transcription start site. While bound to the DNA, a TF either interferes with or promotes the transcription process, thereby leading to either a lower or higher concentration of the protein product. A particular TF binds to DNA only at specific short sequences of nucleotides, often called the recognition (or binding) sites. These sites must be specific enough so that the TF protein does not bind to many random locations throughout the genome, but the specificity cannot be absolute in that varying binding affinities between

the TF and its target sites are required for different genes. Thus, different sites bound by the same TF are not necessarily exact matches. This natural variability between different binding sites of the same TF suggests that probability models, such as the popular position-specific weight (PSW) matrix model, are needed to describe the TF binding motif.

Popular PSW-based motif-finding algorithms include AlignACE (Roth *et al.*, 1998), BioProspector (Liu *et al.*, 2001), Consensus (Hertz and Stormo, 1999), Gibbs Motif Sampler (Liu *et al.*, 1995; Neuwald *et al.*, 1995), MDscan (Liu *et al.*, 2002) and MEME (Bailey and Elkan, 1994), all using procedures that are more or less statistically formulated. AlignACE, BioProspector and Gibbs Motif Sampler are all stochastic Gibbs sampling implementations of models similar to the Bayesian model presented in our Methods section. MEME uses an EM-algorithm to find the maximum-likelihood estimates of parameters of a similar statistical model. Consensus uses an iterative procedure that builds up motif sites one sequence at a time based on the statistical principle of information content. More details about each of these programs is given in Jensen *et al.* (2004).

These algorithms are all fairly fast, easy to use and reasonably accurate, although their relative performances may vary depending on the real-data situation. However, each of them has certain limitations, e.g. the need to input a site abundance parameter, restrictions on the number of sites per sequence or a fixed motif width. More seriously, when these algorithms give different motif predictions, a practitioner will typically have difficulty deciding which one(s) are 'real'.

We present here a scoring function optimization algorithm, BioOptimizer, that builds upon these previous motif-finding algorithms to predict motif site locations. The scoring function approach has the advantage of being simple to understand as well as easy to implement and extend to incorporate more scientific elements. Our procedure can be used in conjunction with any motif-finding program currently available to compare and improve different prediction results. We demonstrate improved motif-finding accuracy for BioOptimizer over other motif-finding programs in both simulation and real-data studies. We also show that BioOptimizer can provide extra flexibility compared with other motif-finding programs, e.g. inferring the motif site abundance parameter and the

*To whom correspondence should be addressed.

motif width. A two-block version of BioOptimizer easily handles the situation where a motif consists of two conserved blocks separated by a variable-length gap of non-conserved nucleotides.

2 METHODS

Our data, \mathbf{S} , consist of m upstream sequences, each of length l_i , where S_{ij} is the nucleotide in position j of sequence i . Our hypothesis is that several similar-looking motif sites are present within \mathbf{S} , and we indicate the location of each motif by a matrix \mathbf{A} where each $A_{ij} = 1$ if a motif site starts in position j of sequence i and 0 otherwise. The problem is, of course, that these locations are unknown, and so we consider each A_{ij} as an independent random indicator variable with an a priori probability p_0 of equaling 1. This parameter is referred to as the motif site abundance parameter. Since each A_{ij} is independent, we allow for the possibility that some sequences will have multiple motif sites (i.e. several $A_{ij} = 1$ in sequence i) as well as the possibility that some sequences may have no motif sites (i.e. all $A_{ij} = 0$ in sequence i).

The composition of the motif is represented by the frequency matrix, Θ , where θ_{jk} is the frequency of nucleotide k in column j of the motif. The nucleotide composition of the background (portions of the sequences that are not motif sites) is represented by the vector θ_0 , where θ_{0k} is the frequency of nucleotide k in the background. This vector is treated as known since it can be estimated a priori. A full Bayesian probability model that links our observed data, \mathbf{S} , to our missing data, \mathbf{A} , unknown motif matrix, Θ , and known parameters θ_0 and p_0 can be written succinctly as

$$\begin{aligned}
 & p(\Theta, \mathbf{A} \mid \mathbf{S}, \theta_0, p_0) \\
 & \propto p(\mathbf{S} \mid \theta_0, \Theta, \mathbf{A}) \times p(\mathbf{A} \mid p_0) \times p(\Theta) \\
 & \propto \prod_{k=1}^4 \theta_{0k}^{n_{0k}} \times \prod_{j=1}^w \prod_{k=1}^4 \theta_{jk}^{n_{jk}} \times p_0^{|\mathbf{A}|} (1 - p_0)^{N^* - |\mathbf{A}|} \times p(\Theta),
 \end{aligned}$$

where n_{jk} is the count of nucleotide k in column j of the current motif matrix, n_{0k} is the count of nucleotide k in the background, $|\mathbf{A}|$ is the number of predicted sites and $N^* = N - m(w - 1)$ is the number of valid start positions, which is the total number of entries N in \mathbf{A} minus the last $w - 1$ positions in each of m sequences since motifs are not allowed to overlap the end of a sequence.

A Markovian background model can also be accommodated easily by our model. With a Dirichlet prior distribution for Θ , one can obtain the marginal posterior distribution for \mathbf{A} alone:

$$p(\mathbf{A} \mid \mathbf{S}, \theta_0, p_0) \propto \int p(\Theta, \mathbf{A} \mid \mathbf{S}, \theta_0, p_0) d\Theta. \quad (1)$$

More details can be found in Jensen *et al.* (2004). We focus on this marginal posterior distribution since the motif locations

($A_{ij} = 1$) are the parameters of direct interest and we avoid having to estimate the entries of the $w \times 4$ matrix Θ .

2.1 The Bayesian scoring function

Under our Bayesian model, we define an ‘optimal’ configuration of start sites \mathbf{A} as a mode of the posterior distribution (1). For those more comfortable with the likelihood framework, this posterior mode is equivalent to the maximum-likelihood estimate under vague prior information. There are advantages to using the Bayesian framework, however, since it allows for the easy incorporation of prior information and for removing of nuisance parameters. Maximizing the posterior distribution (1) is equivalent to maximizing the log-posterior distribution:

$$\begin{aligned}
 \psi_{\text{exact}}(\mathbf{A}) = & K + |\mathbf{A}| \text{logit}(p_0) + \sum_{k=1}^4 n_{0k} \log \theta_{0k} \\
 & + \sum_{j=1}^w \log \left(\frac{\prod_{k=1}^4 \Gamma(n_{jk} + \beta_{jk})}{\Gamma(|\mathbf{A}| + |\boldsymbol{\beta}_j|)} \right), \quad (2)
 \end{aligned}$$

where β_{jk} is the prior count of nucleotide k in column j of the current motif matrix, $|\boldsymbol{\beta}_j|$ is the total prior counts in column j and K collects other terms that are constant with respect to \mathbf{A} . The function logit denotes the log-odds of p_0 , and Γ denotes the gamma function, which is $\Gamma(x + 1) = x!$ for integer x . We call (2) the exact scoring function because it corresponds to the exact form of the log-posterior distribution. Although this exact scoring function may not appear very intuitive to the reader, it is closely related to the following intuitive scoring function through a series of approximations including Stirling’s approximation (Stirling, 1730),

$$\psi_{\text{ent}}(\mathbf{A}) = K + |\mathbf{A}| \left[\text{logit}(p_0) + \sum_{j=1}^w \sum_{k=1}^4 \hat{\theta}_{jk} \log \left(\frac{\hat{\theta}_{jk}}{\theta_{0k}} \right) \right], \quad (3)$$

with details given in Jensen *et al.* (2004). The term $\sum \sum \hat{\theta}_{jk} \log(\hat{\theta}_{jk}/\theta_{0k})$ is often referred to as the entropy distance between the current estimate of our motif frequency matrix entries, $\hat{\theta}_{jk}$, and the fixed background frequencies, θ_{0k} . The entropy distance is also called the Kullback–Leibler information (for discrete measures) in the statistics literature (Kullback and Leibler, 1951) and the information content (Stormo and Hartzell, 1989). Many current motif-finding programs, although not necessarily explicitly based upon the statistical model given above, do in fact make use of scoring functions similar to (3). This scoring function formulation enables us to quantify the ‘goodness’ of different configurations of \mathbf{A} in terms of their fit to our full probability model. As an important additional benefit, we will use this formulation to introduce the following model extensions.

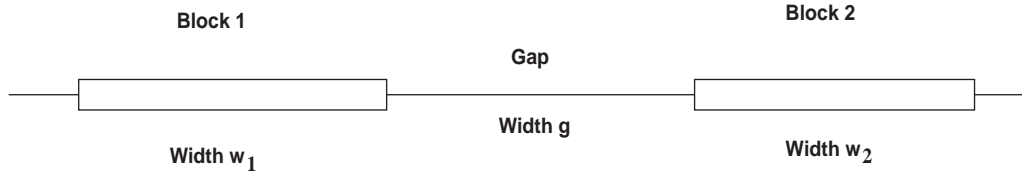


Fig. 1. Two-block motif.

2.2 Extension to unknown motif abundance

The statistical model summarized by (1) assumes known motif site abundance, p_0 . However, in practice one might not have a very good idea of how many motif sites to expect in a given dataset. Other motif-finding algorithms often use *ad hoc* estimates of p_0 , such as assuming a particular number of sites per sequence. With our continued focus on full Bayesian modeling, we instead consider p_0 as a random variable with a Uniform $(0, 1)$ prior distribution. This prior distribution is non-informative in the sense that it will have very little influence on the results compared with the influence of the observed sequence data. We can then mathematically integrate the random variable p_0 out of our model, which will leave us with a posterior distribution that no longer depends on pre-specified site abundance:

$$\begin{aligned} \psi'_{\text{exact}}(\mathbf{A}) &= K + \log B_{1,1}(|\mathbf{A}|, N - |\mathbf{A}|) + \sum_{k=1}^4 n_{0k} \log \theta_{0k} \\ &+ \sum_{j=1}^w \log \left(\frac{\prod_{k=1}^4 \Gamma(n_{jk} + \beta_{jk})}{\Gamma(|\mathbf{A}| + |\boldsymbol{\beta}_j|)} \right), \end{aligned} \quad (4)$$

where $B_{a,b}(c, d)$ is the Beta function $\int_0^1 x^{a+c-1} (1-x)^{b+d-1} dx / \int_0^1 x^{a-1} (1-x)^{b-1} dx$. This exact scoring function can again be simplified using a series of approximations,

$$\psi'_{\text{ent}}(\mathbf{A}) = K + |\mathbf{A}| \left[\text{logit}(\hat{p}_0) - 1 + \sum_{j=1}^w \sum_{k=1}^4 \hat{\theta}_{jk} \log \frac{\hat{\theta}_{jk}}{\theta_{0k}} \right], \quad (5)$$

where $\hat{p}_0 = |\mathbf{A}|/L$ is the estimated motif abundance out of L total possible locations. Despite their more complicated mathematical form, these scoring functions are easy to compute for any \mathbf{A} .

2.3 Extension to variable motif width

We can also consider extending our model to allow the width of our unknown motif to vary. This extension is useful since, in real datasets, there is often very little known about the motif width a priori. Current motif-finding programs, such as BioProspector, Consensus or AlignACE, force the user

to input a motif width that is fixed for the entire run of the program. MEME is the lone exception that allows the motif width to vary.

We can instead let the motif width w be a random variable that has a prior distribution $p(w)$. For example, we could let w have a Poisson(w_0) distribution, where w_0 is an a priori expected motif width. With a prior distribution on w , we obtain a new log-posterior scoring function:

$$\begin{aligned} \psi_{\text{exact}}(\mathbf{A}, w) &= K + \log p(w) + \log B_{1,1}(|\mathbf{A}|, N - |\mathbf{A}|) + \sum_{k=1}^4 n_{0k} \log \theta_{0k} \\ &+ \sum_{j=1}^w \log \left(\frac{\Gamma(|\boldsymbol{\beta}_j|)}{\prod_{k=1}^4 \Gamma(\beta_{jk})} \cdot \frac{\prod_{k=1}^4 \Gamma(n_{jk} + \beta_{jk})}{\Gamma(|\mathbf{A}| + |\boldsymbol{\beta}_j|)} \right). \end{aligned} \quad (6)$$

An intuitive approximation can again be derived, the details of which are given in Jensen *et al.* (2004).

2.4 Extension to two-block motifs

We consider a final extension for the possibility that a particular transcription factor binds to the DNA strand in two places instead of just one. In this case, the binding motif can be summarized by two conserved blocks that are separated by a gap of non-conserved nucleotides that can vary slightly in length, as depicted in Figure 1.

We let Θ_1 and Θ_2 , with width w_1 and w_2 , be the frequency matrices of the two motif blocks, respectively. If we assume that the nucleotide compositions of both blocks are independent of each other, it is not difficult to write out a complete Bayesian model to accommodate the two-block motifs (for more details, see Jensen *et al.*, 2004). The only complication is that we must account for the gap between the two blocks, which can be of different length between different sites. If our current configuration of \mathbf{A} has m sites, the gap lengths of these two-block motif sites are denoted as $\mathbf{G} = (g_1, \dots, g_m)$. We assume a priori that $p(g_i) \sim \text{Discrete Uniform}(G_1, G_2)$. In other words, each gap length can be anywhere from a minimum of G_1 to a maximum of G_2 . Due to the rotation of the DNA double-helix, in many studies G_1 and G_2 are typically separated by about 3 nt. We now have a model where \mathbf{A} , \mathbf{G} and the motif widths w_1 and w_2 are all allowed to vary. The

resulting ‘exact’ scoring function is

$$\begin{aligned} \psi_{\text{exact}}(\mathbf{A}, \mathbf{G}, w_1, w_2) &= K + \log p(w_1) + \log p(w_2) + \log B_{1,1}(|\mathbf{A}|, N - |\mathbf{A}|) \\ &+ \sum_{k=1}^4 n_{0k} \log \theta_{0k} \\ &+ \sum_{j=1}^{w_1+w_2} \log \left(\frac{\Gamma(|\boldsymbol{\beta}_j|)}{\prod_{k=1}^4 \Gamma(\beta_{jk})} \cdot \frac{\prod_{k=1}^4 \Gamma(n_{jk} + \beta_{jk})}{\Gamma(|\mathbf{A}| + |\boldsymbol{\beta}_j|)} \right), \end{aligned} \quad (7)$$

with the implicit restriction that each g_i lies within the interval $[G_1, G_2]$.

2.5 BioOptimizer

To optimize the scoring functions outlined in the previous section, we developed a software package called BioOptimizer, which is currently available for Unix platforms. BioOptimizer takes as input both the sequence data and a starting set of motif sites, such as those provided by BioProspector, Consensus or AlignACE. The output of BioOptimizer is a new set of predicted motif sites that are the best possible fit to our model for the given input.

BioOptimizer systematically scans through every element of the matrix \mathbf{A} and changes the indicator variable at the position A_{ij} to its opposite value only if the value of the scoring function is improved. In addition to the basic algorithm for improving the fit of \mathbf{A} , we can now also propose small changes to the motif width, w , and accept these changes if they improve the score. Specifically, we consider either adding or deleting a position from the current motif and seeing if such a change can make the score higher. The above procedures are repeated until no further changes to \mathbf{A} are accepted, at which point we consider \mathbf{A} to be ‘optimized’.

Those with experience in the fields of statistical computing or physics must have recognized that our procedure is just a local hill-climbing method and can be viewed as a special case of simulated annealing (with immediate freezing). Thus, BioOptimizer only introduces small changes to \mathbf{A} and only accepts changes that immediately improve the score, ψ , and so if the algorithm is started near a inferior local mode, then it will converge only to that inferior local mode.

Although BioOptimizer only does local optimization, it has two basic advantages: (a) it can compare motifs predicted by different motif-finding algorithms and find the best one among them; and (b) it further improves the motif prediction resulting from any of the current algorithms we have tested, e.g. BioProspector, Consensus, AlignACE and MEME. These existing algorithms are proficient at finding good configurations of \mathbf{A} , and each program has its own advantages in real-data situations. In the Results section, we demonstrate

that BioOptimizer has improved the motif site predictions in almost all cases.

3 RESULTS

3.1 One-block motifs

We examined two sequence datasets, each of which contains a one-block transcription factor binding motif. The first dataset is for the transcription factor Spo0A in the bacterium *Bacillus subtilis*. The Spo0A sequence dataset consists of the 200 bp upstream regions of 70 genes that showed preferential hybridization to the Spo0A protein in chromatin immunoprecipitation experiments (Molle *et al.*, 2003). We have 20 Spo0A binding sites that have been confirmed experimentally and can be used to validate our strategy. There is some prior information about the Spo0A binding motif. The literature consensus (Strauch *et al.*, 1990) is thought to be a 7-mer, although the true width of the motif has not been firmly established. Also, it is not known whether or not the orientation of the bound protein (relative to the gene) is relevant, and so we need to look for sites in both the forward ($5' \rightarrow 3'$) and the reverse complement strands.

The second dataset consists of 18 sequences from *Escherichia coli* that contain cyclic-AMP receptor protein (CRP) binding sites. Each sequence is 105 bp long, and each contains at least one motif site, with a total of 23 sites that have been experimentally determined via the footprinting method (Stormo and Hartzell, 1989). This dataset has been previously analyzed by Stormo and Hartzell (1989), Lawrence and Reilly (1990) and Liu (1994). Their analyses focused on detecting sites for a motif with a width of 22 bp, which we will use as our prior expectation, w_0 , but we will let the true width be inferred by the algorithm.

As outlined in the Methods section, our basic strategy is to use a current motif-finding program, such as AlignACE, BioProspector, Consensus or MEME, to find a good configuration of motif start sites \mathbf{A} and then use our optimization program, BioOptimizer, to improve the fit of the motif to our full probability model. Since the performances of these motif-finding programs vary between datasets, BioOptimizer has the advantage of being able to build upon motif results from all these different first-pass programs.

In most cases, the motif width is not known a priori but must be fixed when using a first-pass program, such as BioProspector, Consensus or AlignACE. Our strategy is to collect the motif results from each first-pass program using each of several different motif widths and then apply our optimization program, BioOptimizer, to each result separately. BioOptimizer will then optimize each motif result with respect to both the predicted sites and the unknown motif width as well as provide an optimal score for each motif result that can be used to compare between motif results. The ‘best’ motif would then be the motif result with the greatest BioOptimizer score.

Table 1. Motif site predictions resulting from BioOptimizer

TF	No. of sequences	Results from first-pass program				Best BioOptimizer result			Consensus
		Program	w	$ A $	numTrue	w	$ A $	numTrue	
spo0A	70	BioProspector	12	40	7/20	12	50	12/20	TTGTGCGAAaaa
		Consensus	11	38	10/20	11	47	10/20	TTGTGCGAAaaa
		AlignACE	9	28	6/20	12	49	12/20	tTTGTGCGAAaaa
		MEME	15	38	7/20	14	50	11/20	TTGTGCGAAaaatg
CRP	18	BioProspector	22	11	10/23	24	13	13/23	AtttaTgTGAtcgaggTCACActt
		Consensus	24	13	13/23	24	13	13/23	AtttaTgTGAtcgaggTCACActt
		AlignACE	24	10	10/23	24	13	13/23	AtttaTgTGAtcgaggTCACActt
		MEME	20	18	16/23	19	18	16/23	TGTgAacgagttCACAttt

Four first-pass programs, BioProspector, Consensus, MEME and AlignACE, have been used that provide starting points for running BioOptimizer. Motif predictions resulting from the first-pass programs are also shown. In addition to the motif width, w , consensus sequence and number of predicted sites $|A|$, we also provide ‘numTrue’, which is the proportion of experimentally confirmed sites in each dataset that was found by each algorithm.

For the spo0A dataset, we ran BioProspector separately for motif widths varying from 7 to 12 bp, each time collecting the top five motif predictions (a total of 30 motifs). We also ran Consensus and AlignACE for each of these motif widths and collected the top five motif results. For the CRP dataset, we collected the top five motif results from BioProspector, Consensus and AlignACE for each fixed motif width between 20 and 24 bp. MEME has a built-in capability to try different motif widths, and so we collected the top five motifs from MEME directly. BioOptimizer was then applied to each of these motif results, giving us a total of 30 optimized spo0A motifs (6 widths \times top 5) and 25 optimized CRP motifs (5 widths \times top 5) for each of our first-pass programs. Table 1 shows the BioOptimizer results for these two datasets.

We see from the table that the identical optimal CRP motif resulted from three different starting configurations in terms of both motif width and actual binding sites. However, as noted in the Methods section, different starting points are not guaranteed to converge to exactly the same optimal configuration, as we see in the Spo0A results, where very different starting configurations led to very similar but not identical optimal motifs. In general, BioOptimizer leads to more consistent results even when started from BioProspector, AlignACE, MEME or Consensus results that differ in both motif width and consensus sequence. This is a reassuring result since there are many cases in practice where little is known a priori about a binding motif, including its width.

For both TFs, the optimal motif width seems to be longer than our prior expectations. It also appears that the binding motif of CRP actually consists of two highly conserved blocks with a gap of less-conserved nucleotides, and so we will revisit the CRP dataset in the two-block motif section below. For both datasets, the use of BioOptimizer increased the proportion of true sites found compared with the motif results from one of the first-pass programs alone, suggesting that BioOptimizer has improved the accuracy of the motif results for both CRP and spo0A.

3.2 Two-block motifs

We examined datasets for four two-block transcription factors, σ^E , σ^F , σ^H and σ^K , in the bacterium *B.subtilis*. Given the results in the previous section, we also re-examined the CRP dataset to see if we could find the CRP-binding motif in two short blocks instead of one long block.

Microarray experiments (Eichenberger *et al.*, 2003) comparing wild-type *B.subtilis* cells with cells where the gene for σ^E had been inactivated and with cells where σ^E was overexpressed were used to identify 155 transcriptional units (operons) as direct targets of the σ^E binding protein. Our σ^E dataset consisted of the 200 bp upstream regions from these 155 transcriptional units. Our σ^F dataset (S.Wang, P.Eichenberger and R.Losick, personal communication), σ^H dataset (Britton *et al.*, 2002) and σ^K dataset (P.Eichenberger and R.Losick, personal communication) consisted of 38, 46 and 76 upstream regions, respectively, each found by a similar set of experiments.

Some prior information is available for each of these two-block binding motifs. Helmann and Moran (2002) give the consensus of σ^E as ATa (block 1) and cATACanT (block 2) with a gap of 16–18 bp, the consensus of σ^F binding motif as GywTA and GgnrAnAnTw with a gap of 15 bp, the consensus of σ^H as RnAGGAawWW and RnnGAAT with a gap of 11–12 bp and the consensus of σ^K as AC and CATAnnnT with a gap of 16–18 bp.

Since AlignACE, Consensus or MEME cannot be used to find a two-block motif, we used only BioProspector as a first-pass program. For each σ dataset, BioProspector was used to find good starting configurations under a variety of fixed block widths ranging from 5 to 9 bp and several different gap ranges (11–13 bp, 12–14 bp, 13–15 bp). For the CRP dataset, we specified fixed block widths from 5 to 7 bp with shorter gap ranges (4–6 bp, 5–7 bp, 6–8 bp). For all five datasets, we used a prior expected width of 7 bp for both blocks in the BioOptimizer runs. Table 2 shows both the initial results

Table 2. The two-block motif prediction by BioOptimizer

TF	No. of sequences	Results from first-pass program				Best BioOptimizer result				Consensus
		Program	Dim	A	numTrue	Dim	A	numTrue		
σ^E	155	BioPros.	8-(11-13)-8	106	27/59	11-(10-12)-11	145	47/59	ttgtcaTattt	ttcATAtaatg
σ^F	38	BioPros.	9-(11-13)-9	25	7/11	7-(10-12)-11	38	10/11	GtaTaaa	tGgcaAtAcTa
σ^H	46	BioPros.	7-(13-15)-7	39	13/19	6-(13-15)-8	80	14/19	aaAGGa	tagaGAAt
σ^K	76	BioPros.	7-(13-15)-7	58	6/35	5-(14-16)-11	58	20/35	gcACa	gcATAtgaTaa
CRP	18	BioPros.	6-(6-8)-6	17	16/23	5-(7-9)-7	27	21/23	tGTcA	CAcattt

For each σ -factor data set, BioProspector was first applied with different motif widths and gap length combinations, resulting in 75 predicted motifs (the top five motifs for each of five different widths and three different gap ranges). For the CRP dataset, only three different widths were examined, and a total of 45 motifs were predicted by BioProspector. Then, BioOptimizer was used to optimize and choose the best motif among all the BioProspector motif predictions for a dataset. The BioProspector motif result that was the starting point for this 'best' BioOptimizer motif is also shown. Column 'numTrue' indicates the proportion of experimentally confirmed sites found by that motif result. The consensus and the number of predicted sites are given along with the dimension attribute, 'Dim', of the motif, defined as w_1 -(gap range)- w_2 .

obtained from BioProspector and the optimized predictions by BioOptimizer.

The BioOptimizer motif results for all four σ datasets resemble their prior consensus sequence. In all five datasets, the use of BioOptimizer increased the proportion of confirmed sites found when compared with the BioProspector motif result alone. This improvement in accuracy is especially dramatic in the larger datasets of σ^E and σ^K as well as in the CRP dataset.

In addition to this improved accuracy, BioOptimizer also has the important feature that the motif width is treated as an unknown quantity that can vary. In the datasets studied here, the optimal motif width found by BioOptimizer was often substantially different from our a priori expectations.

3.3 Simulation evaluation of BioOptimizer

In order to further validate the superior performance of BioOptimizer in motif site prediction, we designed the following simulation study. A total of 200 sequence datasets were generated under each combination of several conditions:

- (1) Number of sequences: small (20 sequences) or large (100 sequences).
- (2) Width of motif: short (8 bp) or long (16 bp).
- (3) Degree of motif conservation: high or low.

In each dataset, a true motif site was placed in each sequence. High conservation means that each column of the true motif matrix had a dominant nucleotide with 91% probability (all others 3% equally). Low conservation means that each motif position had a dominant nucleotide with 70% probability (all others 10% equally).

For each simulated dataset, we applied the motif-finding programs BioProspector, Consensus, AlignACE and MEME and compared the results in terms of predicted sites and the true site locations. BioOptimizer was then applied separately to each set of BioProspector, Consensus, AlignACE and MEME results, and the optimized results were also compared

with the true site locations. Unlike the examples presented previously, we did not allow BioOptimizer to vary the motif width in this simulation study so as to examine the benefit of the most basic optimization algorithm. We compared the performances (Table 3) of these algorithms in terms of accuracy of predicted sites, which is the percentage of true sites found in each simulated dataset averaged over all simulated datasets, and total number of predicted sites (|A|).

As shown in Table 3, this simulation study demonstrates again that BioOptimizer has improved the accuracy of the motif site prediction over AlignACE, BioProspector, Consensus and MEME alone for all combinations of motif length, conservation level and number of sequences. The number of predicted sites is generally closer to the truth for BioOptimizer over any of the motif-finding programs alone. In addition to a clear gain in accuracy from using BioOptimizer, it is also worth noting that the accuracy seems to be generally best when using BioProspector or MEME as a starting point compared with Consensus and AlignACE. For the cases with short motifs and low conservation, the performance of all motif-finding programs was very poor. In many of these cases, none of the first-pass algorithms was able to detect the true motif signal, and BioOptimizer did not improve upon these results. In many of these weak signal cases, it was observed that the BioOptimizer algorithm would start from the incorrect signal (based completely on false positive motif sites) found by a first-pass algorithm and converge to a motif configuration, **A**, with no sites. This may be an added benefit of BioOptimizer over other motif-finding programs in that BioOptimizer will not tend to give the false impression of a real motif signal when in fact the correct motif signal has not been found.

4 DISCUSSION

We have introduced a scoring function formulation, implemented in the software BioOptimizer, designed to improve the prediction of regulatory binding motifs. The advantage of scoring functions is that they give us an intuitive means by

Table 3. Average motif prediction accuracies of AlignACE, BioProspector, Consensus and MEME for 200×8 simulated sequence datasets, in comparison with those post-processed by BioOptimizer

Motif width	Conservation	True no. of sites	First-pass program	Average percentage of true sites found (average A)	
				First-pass	BioOptimizer
Short	High	20	AlignACE	59 (17)	64 (15)
Short	High	20	BioProspector	79 (18)	81 (19)
Short	High	20	Consensus	79 (17)	81 (18)
Short	High	20	MEME	81 (18)	81 (18)
Short	High	100	AlignACE	50 (55)	70 (76)
Short	High	100	BioProspector	68 (74)	81 (88)
Short	High	100	Consensus	17 (24)	27 (30)
Short	High	100	MEME	49 (50)	80 (87)
Long	High	20	AlignACE	90 (18)	93 (19)
Long	High	20	BioProspector	85 (17)	92 (19)
Long	High	20	Consensus	90 (18)	92 (19)
Long	High	20	MEME	91 (18)	92 (19)
Long	High	100	AlignACE	89 (90)	91 (92)
Long	High	100	BioProspector	85 (86)	91 (92)
Long	High	100	Consensus	50 (50)	91 (92)
Long	High	100	MEME	50 (50)	91 (92)
Long	Low	20	AlignACE	27 (14)	30 (8)
Long	Low	20	BioProspector	39 (11)	46 (12)
Long	Low	20	Consensus	37 (9)	44 (11)
Long	Low	20	MEME	45 (11)	48 (12)
Long	Low	100	AlignACE	34 (44)	44 (48)
Long	Low	100	BioProspector	38 (41)	54 (59)
Long	Low	100	Consensus	45 (48)	54 (59)
Long	Low	100	MEME	45 (48)	54 (58)

which to compare different possible configurations of motif locations and can serve as a framework for the comparative use of several motif-finding programs, thereby benefiting from the advantages that different motif-finding programs may offer in different situations. This general approach of using multiple methods to obtain different estimates of an unknown quantity that are subsequently compared and improved can be useful beyond models for motif discovery.

This usefulness of BioOptimizer was demonstrated by the increased accuracy of predicted sites across the board compared with BioProspector, Consensus, MEME and AlignACE. Although BioOptimizer is not guaranteed to find a global best fit to our model, there is still a significant gain resulting from its use with very little extra computational time.

BioOptimizer also allows for unknown motif abundance, unknown motif width and two-block motifs with variable-length gaps between the blocks. Allowing the motif width to be inferred from the data has led to non-conventional results when applied to datasets for the *spo0A* binding motif in *B.subtilis* and the CRP-binding motif in *E.coli*. The two-block version of BioOptimizer provided interesting results when applied to the search for binding motifs for several σ -factors in *B.subtilis* as well as the CRP-binding motif. It is seen that the optimal motif width found by BioOptimizer was often substantially different from our a priori expectations.

There are still many open questions in the field of motif discovery. Morphological features of the DNA strand in 3D space are presumably of importance in the process of protein binding, and these features are not adequately captured by modeling only the primary sequence information. (Keles *et al.*, 2003, <http://www.bepress.com/ucbbiostat/paper131>) propose a supervised motif detection method, COMODE, that takes into account structural information about the DNA-binding protein by constraining the motif search to be similar to previously known information content profiles.

Additionally, in eukaryotic species, the DNA strand is wrapped around histone proteins, making some regions of the DNA strand more accessible to binding proteins than other regions, but this information has not been built into a formal motif-finding model.

Finally, although several promising algorithms have been proposed to take advantage of the microarray information (Bussemaker *et al.*, 2001; Keles *et al.*, 2002; Liu *et al.*, 2002; Conlon *et al.*, 2003), a formal joint probability model for both sequence motif and expression microarray information may lead to a more efficient way of utilizing the available information.

ACKNOWLEDGEMENTS

We are grateful to Richard Losick and members in his laboratory, P. Eichenberger and V. Molle, for their scientific advice

regarding *B.subtilis* and their enthusiastic support during the development of BioOptimizer, to Xiaole Liu for her help in our application of BioProspector and to Lei Shen for his bug report. J.S.L is supported in part by the NSF grant DMS-0204674 and the NIH grant HG02518-01.

REFERENCES

- Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, Menlo Park, California. AAAI Press, pp. 28–36.
- Britton,R.A., Eichenberger,P., Gonzalez-Pastor,J.E., Fawcett,P., Monson,R., Losick,R. and Grossman,A.D. (2002) Genome-wide analysis of the stationary-phase sigma factor (σ^H) regulon of *Bacillus subtilis*. *J. Bacteriol.*, **184**, 4881–4890.
- Bussemaker,H.J., Li,H. and Siggia,E.D. (2001) Regulatory element detection using correlation with expression. *Nat. Genet.*, **27**, 167–171.
- Conlon,E.M., Liu,X.S., Lieb,J.D. and Liu,J.S. (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl Acad. Sci., USA*, **100**, 3339–3344.
- Eichenberger,P., Jensen,S.T., Conlon,E.M., van Ooij,C., Silvaggi,J., Gonzalez-Pastor,J., Fujita,M., Ben-Yehuda,S., Stragier,P., Liu,J.S. and Losick,R. (2003) The σ^E regulon and the identification of additional sporulation genes in *Bacillus subtilis*. *J. Mol. Biol.*, **327**, 945–972.
- Helmann,J.D. and Moran,C.P.,Jr (2002) RNA polymerase and sigma factors. In Sonenshein,A.L., Hoch,J.A., Losick,R. (eds) *Bacillus subtilis and Its Closest Relatives*, Chapter 21. ASM Press, Washington, D.C.
- Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
- Jensen,S.T., Liu,X.S., Zhou,Q. and Liu,J.S. (2004) Computational discovery of gene regulatory binding motifs: a Bayesian perspective. *Stat. Sci.*, in press.
- Keles,S., van der Laan,M. and Eisen,M.B. (2002) Identification of regulatory elements using a feature selection method. *Bioinformatics*, **18**, 1167–1175.
- Keles,S., van der Laan,M., Dudoit,S., Xing,B. and Eisen,M.B. (2003) Supervised detection of regulatory motifs in DNA sequences. *U.C. Berkeley Division of Biostatistics Working Paper Series*, Working Paper 131, May 23.
- Kullback,S. and Leibler,R.A. (1951) On information and sufficiency. *Ann. Math. Stat.*, **22**, 79–86.
- 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. (1994) The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Am. Stat. Assoc.*, **94**, 958–966.
- Liu,J.S., Neuwald,A.N. and Lawrence,C.E. (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.*, **90**, 1156–1170.
- Liu,X., Brutlag,D.L. and Liu,J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, **6**, 127–138.
- Liu,X.S., Brutlag,D.L. and Liu,J.S. (2002) An algorithm for finding protein–DNA interaction sites with applications to chromatin immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **20**, 835–839.
- Molle,V., Fujita,M., Jensen,S.T., Liu,J.S. and Losick,R. (2003) The spo0A regulon in *Bacillus subtilis*. *Mol. Microbiol.*, **50**, 1683–1701.
- 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.
- Roth,F.R., Hughes,J.D., Estep,P.E. and Church,G.M. (1998) Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.
- Stirling,J. (1730) *Methodus differentialis*. William Bowyer, London.
- Stormo,G.D. and Hartzell,G.W. (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl Acad. Sci., USA*, **86**, 1183–1187.
- Strauch,M., Webb,V., Spiegelman,G. and Hoch,J.A. (1990) The Spo0A protein of *Bacillus subtilis* is a repressor of the abrB gene. *Proc. Natl Acad. Sci., USA*, **87**, 1801–1805.