BIOINFORMATICS ORIGINAL PAPER Vol. 21 no. 10 2005, pages 2240–2245
doi:10.1093/bioinformatics/bti336

Sequence analysis

A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length

A. V. Favorov[1,*], M. S. Gelfand[1,2], A. V. Gerasimova[1], D. A. Ravcheev[2,3], A. A. Mironov[1,3] and V. J. Makeev[1,4]

[1]State Scientific Centre 'GosNIIGenetika' Laboratory for Bioinformatics, 1st Dorozhny pr. 1, Moscow, 117545, Russia, [2]Institute of Information Transmission Problems, Russian Academy of Sciences, Bolshoi Karetny per. 19, Moscow 127994, Russia, [3]Department of Bioengineering and Bioinformatics, Moscow State University, Laboratory Building B, Vorobiovy Gory 1-73, Moscow 119992, Russia and [4]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Vavilova 32, Moscow 119991, Russia

Received on January 27, 2005; accepted on February 16, 2005
Advance Access publication February 22, 2005

ABSTRACT

Motivation: Transcription regulatory protein factors often bind DNA as homo-dimers or hetero-dimers. Thus they recognize structured DNA motifs that are inverted or direct repeats or spaced motif pairs. However, these motifs are often difficult to identify owing to their high divergence. The motif structure included explicitly into the motif recognition algorithm improves recognition efficiency for highly divergent motifs as well as estimation of motif geometric parameters.

Result: We present a modification of the Gibbs sampling motif extraction algorithm, SeSiMCMC (Sequence Similarities by Markov Chain Monte Carlo), which finds structured motifs of these types, as well as non-structured motifs, in a set of unaligned DNA sequences. It employs improved estimators of motif and spacer lengths. The probability that a sequence does not contain any motif is accounted for in a rigorous Bayesian manner. We have applied the algorithm to a set of upstream regions of genes from two *Escherichia coli* regulons involved in respiration. We have demonstrated that accounting for a symmetric motif structure allows the algorithm to identify weak motifs more accurately. In the examples studied, ArcA binding sites were demonstrated to have the structure of a direct spaced repeat, whereas NarP binding sites exhibited the palindromic structure.

Availability: The WWW interface of the program, its FreeBSD (4.0) and Windows 32 console executables are available at http://bioinform.genetika.ru/SeSiMCMC

Contact: favorov@sensi.org

Supplementary information: Supplementary material available at http://bioinform.genetika.ru/SeSiMCMC

1 INTRODUCTION

Extraction of a common motif from a set of unaligned sequence fragments (also known as the multiple local alignment or MLA problem) is often applied to identify DNA sites that are recognized by transcription factors. This approach is based on the assumption that DNA segments upstream of coregulated genes contain similar nucleotide subsequences.

Usually the analysis starts from a sample of DNA sequences, the majority of which are supposed to contain a protein-binding site (or some other specific segment). Therefore, these sequences include instances of the same motif. The objective is to classify all DNA sequence data into motif instances and the remaining background in an optimal manner. Different approaches to this problem were recently reviewed (Bulyk, 2003) and probabilistic methods based on Gibbs sampling (Roth *et al*., 1998; Hughes *et al*., 2000; Thijs *et al*., 2002; Thompson *et al*., 2003) appear to be the most efficient. Here we present a tool optimized to solve a specific variation of the MLA problem: identification of a motif exhibiting a double-box structure. Special cases of such structured motifs, the inverted and direct repeat, are often recognized by prokaryotic transcription factors, as demonstrated in the analysis and prediction of gene coregulation both in prokaryotes (Gelfand *et al*., 2000) and eukaryotes (Chiang *et al*., 2003). Usually, the motif, either symmetric or otherwise, is spaced, i.e. it contains several poorly conserved positions in the middle. Such dyad structures are recognized by factors binding in their dimer form (Pilpel *et al*., 2001) and this knowledge has been exploited in several powerful motif extraction tools (van Helden *et al*., 2000; Li *et al*., 2002; Robin *et al*., 2002; Mwangi and Siggia, 2003). Recent experiments on direct cross-linking of transcription regulatory proteins to DNA (Harbison *et al*., 2004) demonstrate that the pairs or clusters of binding sites, either identical or different, separated by spacers of approximately fixed length are also very common in eukaryotes. Identification of structured motifs by a general Gibbs sampling procedure appears to improve prediction. Since one usually does not know in advance the motif length and the spacer length, the program should estimate the optimal values for these two parameters during motif detection. In addition, the program will need to handle training sets that may contain biologically irrelevant sequences without a target site.

*To whom correspondence should be addressed.

2 SYSTEM AND METHODS

In order to create a specialized tool for finding weak motifs with spacers of unknown length, we designed a probabilistic model and an optimization procedure (Favorov *et al*., 2002), modifying the classic algorithm of Lawrence *et al*. (1993). Such a specialized tool might be more adequate for this particular task than a universal one (Hughes *et al*., 2000; Liu, 2001; Thijs *et al*., 2002).

Two probabilistic models, foreground (the motif) and background, are formulated. The optimal classification is the one most probable in the Bayesian sense (Sivia, 1996). The motif is represented by a positional probability matrix (Berg and von Hippel, 1987; Stormo and Hartzell, 1989; Lawrence *et al*., 1993; Bailey and Elkan, 1995; Hertz and Stormo, 1999); the background is modeled by independent symbols with fixed probabilities of DNA bases.

We maximize the posterior of the given foreground–background partition of the DNA sequence data as a function of the site positions in the sequences from the training set. Such a function may have many local maxima, so the Markov Chain Monte Carlo (MCMC) technique (Besag *et al*., 1996; Robert, 1998; Liu, 2001) is a natural algorithm for its optimization. The MCMC variant known as Gibbs sampling (Geman and Geman, 1984) has been applied to the MLA problem in Lawrence *et al*. (1993) and has become one of the most popular approaches to motif extraction in biological sequences (Roth *et al*., 1998; Hughes *et al*., 2000; Thijs *et al*., 2002; Liu *et al*., 2002). The algorithm implemented in the SeSiMCMC software additionally does not require specification of the length of the motif. The algorithm searches for either direct repeat or palindromic (two inverse complementary boxes) motif structures as specified by the user, possibly separated by a spacer of unknown length. Occurrences of non-palindromic motifs, both single and double boxes, can be searched for in one or both complementary DNA strands.

3 ALGORITHM

The probabilities $q(i,r)$ for the occurrence of nucleotide r at site position $i, i = 1..s$, where s is the site length and the background nucleotide probabilities $f(r)$ are estimated from the in-site and the background counters denoted by $c(i,r)$ and $g(r)$:

$$q(i,r) = \frac{c(i,r) + b(r)}{M + B} \tag{1}$$

and

$$f(r) = \frac{g(r) + b(r)}{K + B}, \tag{2}$$

where M is the number of sites in the set, from which the statistics are derived, K is the number of all non-site positions in the data. Pseudocounts $b(r)$ are proportional to the frequencies of nucleotides in the full dataset, while their sum

$$B = \sum_r b(r) \sim \sqrt{N},$$

where N is the number of data sequences (Lawrence *et al*., 1993).

For motifs that are supposed to be (imperfectly) symmetrical, the formula for $q(i,r)$ reflects the symmetry. For direct repeats it is given by

$$q(i,r) = \frac{c(i,r) + c\,(i + \mathrm{int}((s+1)/2), r) + 2b(r)}{2(M + B)}, \tag{3}$$

while for palindromes (inverted repeats) it is given by

$$q(i,r) = \frac{c(i,r) + c(s + 1 - i, \bar{r}) + b(r) + b(\bar{r})}{2(M + B)}, \tag{4}$$

where s is the motif length and $\bar{r}$ is the nucleotide complementary to r.

The core procedure for selection of a set of similar sites is as follows. We start with randomly scattered sites of a definite length, one per sequence. Then, we organize a cycle of one-by-one updates of site positions. At each step, we select only one sequence. For uniformity, we treat the site absence as a position of a specific type ('null'). At each step, we collect the nucleotide statistics for the internal site positions and for the background from all sequences except the one being updated. We estimate the positional nucleotide probabilities within the motif with formulae (1), (3), (4) and the background probabilities using formula (2). For each selected sequence $\mathrm{R} = r_1 r_2 \cdots r_{l-1} r_l$, the probability (likelihood) to obtain this sequence from a Bernoulli process (i.e. the site position likelihood) given the site position k is:

$$\begin{aligned} P(\mathrm{R}|[k], q, f) &= \prod_{i=1}^{k-1} f(r_i) \prod_{i=k}^{k+s-1} q(i-k+1, r_i) \\ &\quad \times \prod_{i=k+s}^{L-s+1} f(r_i) \quad k \neq 0 \\ P(\mathrm{R}|[0]) &= \prod_{i=1}^{L-s+1} f(r_i), \end{aligned} \tag{5}$$

where r_i is the i-th nucleotide in sequence R and $[k], k = 1..(L - s + 1)$ denotes the event 'the site starts at position k', [0] corresponds to the case where the site is absent ('null position'). The prior $P([0])$ is a user-defined probability for a sequence from the data to be noise (the sequence does not contain any site). All non-zero positions have equal probabilities a priori, thus the prior of the event $[k]$

$$P([k]) = \frac{1}{L - s + 1}(1 - P[0]) \quad k \neq 0. \tag{6}$$

The marginal probability of the sequence itself (evidence) is:

$$P(\mathrm{R}_{|q,f}) = \sum_{k=0}^{L-s+1} P(\mathrm{R}|[k], q, f) \cdot P([k]). \tag{7}$$

The probability (posterior) for a site to start at k is:

$$\begin{aligned} P([k]|\mathrm{R}, q, f) &= \frac{P(\mathrm{R}|[k], q, f) P([k])}{P(\mathrm{R}_{|q,f})} \\ &= \frac{P(\mathrm{R}|[k], q, f) P([k])}{P(\mathrm{R}_{|q,f})}. \end{aligned} \tag{8}$$

So combining the priors with the likelihoods in the usual Bayesian way, we obtain the posterior distribution for a site position in the current sequence and sample the new site position (possibly the 'null' one) from this distribution. The process is iterated until the chain comprising sets of site positions converges (i.e. the step-to-step changes become small). The algorithm is similar to the one described in Lawrence *et al*. (1993), but we include the possibility of the absence of a site in the Bayesian way at each update.

In fact, the algorithm optimizes the self-consistency of a set of site positions, making it very sensitive to changes in the mutual arrangement of the sites (i.e. it is quite tolerant to all as one shifts of the site position set). To overcome this problem, we adjust the results from time to time after the core algorithm converges satisfactorily and then restart the core. The adjustment is a deterministic search for the best solution among all possible cooperative shifts of the local alignment of sites.

At the adjustment step the best set of sites is defined by the highest information content per site position (ICP) in the signal. The information content is the sum of two components: the structural one and the spatial one. Both are related to the Kullback entropy distances. The structural component is the distance between the probability model for the nucleotide occurrence inside the motif (the position–probability matrix) and the background probability distribution:

$$I_{\text{struct}} = \sum_{i=1}^{s} \sum_{r=1}^{4} c(i,r) \cdot \log_2 \left(\frac{q(i,r)}{f(r)} \right). \tag{9}$$

Now, the counters $c(i,r)$ and the model parameters $q(i,r)$ and $f(r)$ are evaluated using all sequences. Formula (9) is different from the standard Kullback entropy distance in that we use $c(i,r)$ as the factor and $q(i,r)$ as the log argument. The distance between the estimated probability distribution of symbols in the alignment position $q(i,r)$ (which contains pseudocounts) and the background distribution $f(r)$ is calculated using the observed data $c(i,r)$. In standard Kullback distance measure, there are only two distributions and no observed data; in this case $q(i,r)$ would appear at both places. Note that the constant M in the denominator in Equation (11) given below, which provides for the correct normalization.

The spatial component is the distance between the distribution of the posterior of the site position in a sequence (including the 'null' position) given the known set of sites and the prior distribution of the site position.

$$\begin{aligned} I_{\text{spatial}} &= \sum_{\text{R}\subset\{\text{sequences}\}} \sum_{k=0}^{L_{\text{R}}-s+1} P([k]_{\text{R}}|\text{R}, q, f) \log_2 \left(\frac{P([k]_{\text{R}}|\text{R}, q, f)}{P([k]_{\text{R}})} \right) \\ &= \sum_{\text{R}\subset\{\text{sequences}\}} - \log_2 P(\text{R}_{|q,f}) + \frac{1}{P(\text{R}_{|q,f})} \\ &\quad \sum_{k=0}^{L_{\text{R}}-s+1} P(\text{R}|[k]_{\text{R}}, q, f) \times P([k]_{\text{R}}) \times \log_2 P(\text{R}|[k]_{\text{R}}, q, f) \end{aligned} \tag{10}$$

where $[k]_{\text{R}}$ denotes the event that a site is observed in position k of sequence R, L_{R} is the sequence length, and q and f are the same as in Equation (5).

Finally, the value of ICP equals

$$\left(\frac{I_{\text{spatial}} + I_{\text{struct}}}{s \cdot M} \right), \tag{11}$$

where s and M are the same as in Equations (1)–(4).

In fact, it is sufficient to maximize I_{struct} to find a best set among all shifts, although the spatial component is necessary to estimate the optimal motif length. Indeed, the structural component itself (e.g. the motif probability matrix information content) is not suitable as a value to be optimized with the motif length because it grows monotonically with the length. On the other hand, if the structural component is normalized for the motif length, the maximal value is attained at a single best position, creating a motif with length 1.

Thus, at every adjustment step, we take the motif for which the maximal ICP is attained in the preceding sampling chain and then vary the site length and the absolute position of the entire set as a whole in order to optimize the value of ICP as given by Equation (11). For spaced motifs at this stage we also estimate the length of the spacer separating two boxes of the same length; the background probabilistic model is adopted for the spacer. For each adjustment procedure, that is the cooperative shift of sites, the optimal spacer length is selected, which gives the (local) maximum of the ICP [Equation (11)].

This adjustment procedure is similar to the one described in Lawrence *et al*. (1993) with the following differences. The information content calculation [Equation (11)] has an improved spatial component. The adjustment stage is also used to evaluate the site length and the length of spacer, if spaced motifs are allowed. For every site length, the spacer length is chosen as the minimal value for which the local maximum for the ICP is attained. Since the spacer length can be zero, this procedure in effect is used to determine whether the motif is spaced.

4 IMPLEMENTATION

The SeSiMCMC software is written in C++ (gcc 3.x). Executable files for FreeBSD and Windows 32 console are available from the project site http://bioinform.genetika.ru/SeSiMCMC. This web page contains the program documentation describing the command line and the configuration file control interfaces. Also at this page one can find the web-based version of the program with input forms. Below, we describe the simple input form, which is used in the interface, from the user's viewpoint. The advanced form allows the user to control many program parameters, which are described in the documentation. All parameters except the obligatory input sequence data are originally set to their default values.

The sequence input data in the FastA format can be copied directly to the text window or submitted as a file from the user's computer. Each running task has its own unique identifier, allowing users to obtain the results of earlier computations, and the results are saved on the host computer for at least one month. There are several fields allowing the user to select the motif geometry and to set a priori information about the motif, i.e. the expected range of lengths as well as a reasonable length seed value. The expected fraction of sequence fragments that do not contain a site is also supplied in the 'motif absence prior' field. In addition, this parameter expresses the preference about the desired motif: an abundant weak motif or a strong, but rare one. The lower the parameter, the greater the attention paid to the motif population (i.e. abundant motifs).

Two protocols for the motif and spacer length optimization can be used. The default 'fast' mode performs the optimization at the stage of local alignment adjustment, as described above. In the 'slow mode', the motif length is changed stepwise and the full sampling procedure without the adjustment is executed at every step. The latter variation is similar to that described in Lawrence *et al*. (1993) and is rather slow. The final output of the slow mode contains results for each motif length with the motif geometry optimal for that length. Thus in the slow mode the user obtains more information about the possible motifs. Obviously, the ICP, maximized over all motif lengths provided by the algorithm working in the slow mode, can be greater than that obtained by maximization in the fast mode, where the motif geometry, i.e. the motif length and the spacer, are obtained simultaneously with the motif.

At the computational stage, the algorithm assumes that there is at most one site per sequence fragment. At the output stage, the program scans all sequences with the final positional probability matrix $q(i,r)$ and extracts all sites with a probability higher than

that of the least probable site identified at the computational stage in any sequence. This post-processing procedure is optional, and if this option is not selected, only the best local alignment is output. If this post-processing retrieves too many sites, it indicates that the site with the lowest probability fits the motif poorly. In this case it is advisable to repeat the calculations with an increased prior for the absence of a site. It is also possible to ask the program to set the threshold for the retrieved sites at the i-th lowest probability score of the initially identified sites. Like other tools (Thijs *et al*., 2002), the software can search for multiple motifs by restarting the process on the input data after masking the previously identified motif sites. There is a flag in the web form that requests such a masked output for further motif searches.

5 RESULTS

To test the performance of the algorithm we studied two factors whose binding sites are notoriously divergent and difficult for computer identification, ArcA and NarP, both of which are involved in the regulation of respiration. There are indications that binding signals of these proteins exhibit a symmetric structure. Experiments showed that the NarP signal might be palindromic (Darwin *et al*., 1997), whereas the ArcA signal was reported to be found in some regulatory regions with several copies on the same strand (McGuire *et al*., 1999; Liu and De Wulf, 2004), which yields the possibility that it is a direct repeat. It is important that these regulatory systems are vital for bacterial respiration and thus are well studied experimentally (Darwin *et al*., 1997; McGuire *et al*., 1999; Liu and De Wulf, 2004).

In both cases we combined the motif identification procedure with comparative genomic studies, analyzing sites found in the regions upstream of orthologous genes from several related genomes (see Gelfand *et al*., 2000 for a detailed review of the procedure). The parameters used for the motif search by SeSiMCMC are available as examples of the program runs at http://bioinform.genetika.ru/SeSiMCMC. All comparative genomics analyses were performed using the GenomeExplorer software tool (Mironov *et al*., 2000).

5.1 Phospho-ArcA

The Arc cascade regulates gene expression in response to aerobic/anaerobic environment changes. The cascade consists of the membrane-associated sensor kinase ArcB and the regulatory protein ArcA. When the cell lacks oxygen, ArcB phosphorylates itself and then catalyzes ArcA phosphorylation. In turn, phospho-ArcA (ArcA-P) represses some operons (e.g. *icd*, *lld*, *glt*, *glc*, *sdh* and *sodA*) and activates some others (e.g. *cyd* and *pfl*) (Lynch and Lin, 1996). Currently, there are about a dozen of operons that are known to be regulated by ArcA-P, although there are recent indications that ArcA-P regulation may be even more important with hundreds of genes involved directly and indirectly (Liu and De Wulf, 2004). Thus, identification of candidate ArcA binding sites is important for understanding bacterial metabolism.

We started with a set of regions upstream of *E.coli* genes, for which ArcA regulation was verified by various experimental techniques (Supplementary Table 1). SeSiMCMC was run for this set of sequence fragments using all possible parameter combinations, searching for: (1) an isolated motif, (2) a spaced generic motif, (3) a spaced direct repeat and (4) a spaced inverse complement repeat (a palindrome). The motif could be located on any DNA strand and its length could be from 6 to 22 bases. The best motif found had a structure of a spaced tandem repeat (Fig. 1). This 15 nt motif is better conserved and has more informative positions than the ArcA motif reported in McGuire *et al*. (1999). We believe that the refinement is a result of the defined structure of the motif. When a single-box motif is searched, a box with a stronger core is selected from two boxes of a double-box motif in the sequence. In so doing, the irrelevant positions flanking the cores of different boxes sometimes become aligned. In our case these positions were consistently assigned to the flanking positions or to the spacer, which allowed us to locate the core positions more precisely.

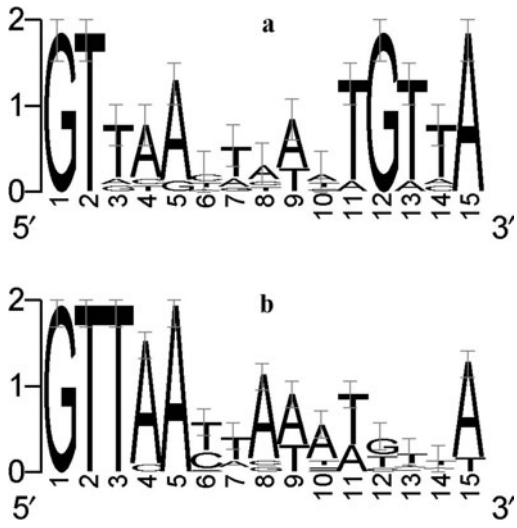

Fig. 1. ArcA regulatory motif variations. Horizontal axis, position in the signal; vertical axis, information content in bits. The height of each stack of letters is proportional to the positional information content in the given position; the height of each individual letter reflects its prevalence in the given position. The logos were created by WebLogo (Crooks *et al*., 2004; Schneider and Stephens, 1990; http://weblogo.berkeley.edu/). (a) The motif obtained from the alignment of sites identified by SeSiMCMC. (b) Logo for the motif from (McGuire *et al*., 1999) (according to http://arep.med.harvard.edu/ecoli_matrices/dat/arcA.dat).

The greater selectivity of the identified motif allowed us to use it in comparative genomic studies. The resulting set of sites was used to create a recognition profile (a variant of positional-weight matrix, Mironov *et al*., 1999). All sites for which that profile scored better than the worst site found by SeSiMCMC were accepted. With this rule at hand we scanned upstream regions of all orthologous genes of four gamma-proteobacteria: *E.coli*, *Yersinia pestis*, *Pasteurela multocida* and *Vibrio vulnificus*. Genes with the candidate site present in the regions upstream of an *E.coli* gene and at least one of its orthologs were selected as putative members of the ArcA-P regulon. As a result, we discovered a number of new genes putatively regulated by ArcA-P in *E.coli* and the other three gamma-proteobacteria, the majority of which had been reported in the literature as relevant to respiratory regulation (Supplementary Table 2) (Gerasimova *et al*., 2004).

5.2 NarP

The NarP regulatory system operates in anaerobic conditions. Nitrate and nitrite are the most efficient electron acceptors in this case. *E.coli* possesses a complex regulatory system for close monitoring of

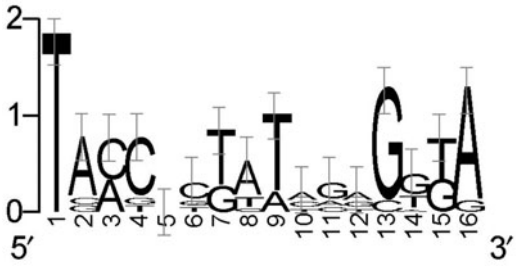

Fig. 2. Sequence logo for the NarP binding site. Horizontal axis, position in the signal; vertical axis, information content in bits. The height of each stack of letters is proportional to the positional information content in the given position; the height of each individual letter reflects its prevalence in the given position. The logos were created by WebLogo (Crooks *et al*., 2004; Schneider and Stephens, 1990; http://weblogo.berkeley.edu/).

and response to nitrate/nitrite concentration changes in the environment, which includes NarL and NarP transcription regulatory factors. These factors are activated by sensor kinases NarQ and NarX. In their active form NarL and NarP activate transcription of operons responsible for the nitrate/nitrite respiration (*narGHI*, *narK*, *nap*, *nir* and *nrf*) and operons coding for some respiratory dehydrogenases (*nuo*, *hya* and *fdnGHI*). They repress transcription of operons responsible for other forms of anaerobic respiration (*dms*, *focA-pflB*, *torCAD*, *dcuB-fumB* and *frd*).

NarP is believed to recognize a 16-nt palindromic site with the consensus TACYYMT-2-AKRRGTA (Darwin *et al*., 1997; Maris *et al*., 2002), whereas it is still unclear what the recognition site of NarL is; presumably this site also contains the TACYYMT boxes in different combinations (Darwin *et al*., 1997).

The training set included 16 regions upstream of *E.coli* operons, for which the NarP regulation was shown by different experimental methods (Supplementary Table 3). SeSiMCMC was run for this set using different parameter combinations, similar to the ArcA case above. The identification of the NarP binding motif proved to be a more difficult task than the identification of motifs for ArcA. The motif was found only when the inverse repeated (palindromic) structure was specified and the length of a candidate motif could vary between 10 and 20 nt with the starting length 16 (which is equal to the reported motif length). The prior for the absence of the motif in a sequence was evaluated as 0.5. In this case the motif with a characteristic NarP consensus was identified (Fig. 2).

Again, the obtained set of sites was used to create a recognition profile (Mironov *et al*., 1999). We selected the maximal profile cutoff value of 3.50, for which at least one NarP site was found in each sequence from the training set. We used comparative genomics to verify the signal and to find other candidate members of the NarP regulon in the four genomes (Supplementary Table 4).

Candidate NarP sites were found in other genomes upstream of most genes in the training set. In particular, we predicted NarP regulation for five operons from the training set (*narK*, *narG*, *dmsA*, *adhE* and *torC*) for which only NarL regulation had been shown.

Candidate NarP sites were also found for five new operons (*nirB*, *dcuB*, *ynfE*, *moaA* and *narXL*), three of which were previously reported as nitrate- or nitrite-regulated. Regulation of *narX* by NarP was experimentally demonstrated in *E.coli* (beta-galactosidase assays, Darwin and Stewart, 1995); and thus this site is probably functional despite the fact that no site was found upstream of *narX* in genomes other than *E.coli*. No candidate NarP sites were found upstream of *narQ* and *narP* in genomes other than *V.vulnificus*, where these two genes form an operon. This may indicate auto-regulation of NarPQ expression in *V.vulnificus*.

Other results of this study will be described in detail elsewhere. In brief, we have demonstrated conservation of the identified NarP sites and found several new conserved sites, thus identifying new members of the regulon. These members are *fdoG* (formate dehydrogenase isoenzyme) in *Y.pestis* and *P.multocida*, *nqr* (NADH-dehydrogenase) in *Y.pestis* and *P.multocida*, and *moaABCDE* (synthesis of molybdenum cofactor). Finally, candidate NarP sites were observed upstream of two homologous operons *ynfEFGHI* and *dmsABC* that encode dimethyl sulfoxide reductase in *E.coli*. These operons have no orthologs in other gamma-proteobacteria.

6 DISCUSSION AND CONCLUSIONS

All in all, SeSiMCMC is a tool for multiple local alignment of a set of DNA sequence fragments that is based on a modification of the Gibbs sampling algorithm (Lawrence *et al*., 1993). Our primary objective was to create a computationally efficient tool that uses user-defined motif symmetry and evaluates the motif length from the data. Sequence fragments in the training set can have arbitrary orientation, and there is a probability for a sequence to contain no sites.

In the recent assessment of different motif predictors by identification of binding sites in eukaryotic genomes (Tompa *et al*., 2005), SeSiMCMC, as a specialized tool, demonstrated a moderate performance over the general dataset, but was the best at *Drosophila* fly data and was among the 3 programs that gave positive results on this dataset out of 13.

SeSiMCMC testing on sets of bacterial regulatory regions, known to be difficult for signal identification with computational tools, allowed us to refine the binding motif for the global respiration regulator ArcA and demonstrate that it has a structure of a direct repeat. We also obtained the motif for the NarP regulator. This motif had the structure of an inverse complement repeat, a palindrome. The conserved signals were validated by means of comparative genomics, and a number of new members of the ArcA and the NarP regulons where identified. This in turn should lead to a better understanding of the important process of bacterial respiration.

The dramatic progress in experimental identification of transcription factor binding sites is now obvious. In this connection it is noteworthy that recent experiments on genome-wide cross-linking of transcription factor proteins to DNA in yeast (Harbison *et al*., 2004) included additional examination of the experimental results with site prediction tools, which allowed the authors to exclude experimental false positives. This important study also demonstrated the variety of patterns in the arrangement of transcription factor binding sites within regulatory regions in yeasts. A number of examples of site sequences for recurrent pairs of regulators, as well as multiple copies of the same binding signal with fixed or preferential spacing between site occurrences were observed. Thus, it is likely that the next generation of tools for signal identification would focus on programs predicting not individual sites but rather site combinations or more complex site arrangements (Li *et al*., 2002). SeSiMCMC provides only a limited step in this direction. It can be used to search for combinations of two different sites separated by a fixed spacer (when executed with the motif symmetry not specified) or for combinations

of two identical sites located on the same (direct repeat) or different DNA strands.

We are, however, fully aware that it is not yet sufficient even in bacterial studies for recovery of the regulatory region structure. In eukaryotes the site arrangements become very complex, including overlapping sites and sites with periodic positioning (Kel-Margoulis *et al*., 2002; Makeev *et al*., 2003; Frith *et al*., 2003; Qiu *et al*., 2003).

Another important area of application is the analysis of mass gene-expression data, e.g. in microarray experiments. Such experiments usually identify many genes with indirect regulation. In this case the option of an explicit prior allowing the absence of a motif is likely to be very useful.

ACKNOWLEDGEMENTS

We are grateful to D. Rodionov (IITP RAS, Moscow) for useful discussion, to L. Danilova (IITP RAS, Moscow) for assistance with the data and to Dr M. Ochs (Fox Chase Cancer Center, Philadelphia, PA) for helpful comments on the manuscript. This study was partially supported by grants from the Howard Hughes Medical Institute (55000309 to M.S.G.), the Russian Foundation for Basic Research (04-04-49601 to V.J.M.), the Russian Academy of Sciences (Programs 'Molecular and Cellular Biology', project No. 10 and 'Origin and Evolution of The Biosphere').

REFERENCES

Bailey,T. and Elkan,C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning J.*, **21**, 51–83.

Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical–mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723–750.

Besag,J. *et al*. (1996) Bayesian computation and stochastic systems. *Stat. Sci.*, **10**, 3–66.

Bulyk,M. (2003) Computational prediction of transcription-factor binding site locations. *Genome Biol.*, **5**, 201.

Chiang,D.Y. *et al*. (2003) Phylogenetically and spatially conserved word pairs associated with gene-expression changes in yeasts. *Genome Biol.*, **4**, R43.

Crooks,G.E. *et al*. (2004) Weblogo: a sequence logo generator. *Genome Res.*, 1188–1190.

Darwin,A.J. and Stewart,V. (1995) Expression of the *narX*, *narL*, *narP*, and *narQ* genes of *Escherichia coli* K-12: regulation of the regulators. *J. Bacteriol.*, **177**, 3865–3869.

Darwin,A.J. *et al*. (1997) Differential regulation by the homologous response regulators NarL and NarP of *Escherichia coli* K-12 depends on DNA binding site arrangement. *Mol. Microbiol.*, **25**, 583–595.

Favorov,A.V. *et al*. (2002) Yet another digging for DNA motifs Gibbs sampler. In *Proceedings of BGRS 2002*, Novisibisk, **1**, 31–33.

Frith,M.C. *et al*. (2003) Cluster-Buster: finding dense clusters of motifs in DNA sequences. *Nucleic Acids Res.*, **31**, 3666–3668.

Gelfand,M.S. *et al*. (2000) Comparative analysis of regulatory patterns in bacterial genomes. *Brief. Bioinformatics*, **1**, 357–371.

Geman,S. and Geman,D. (1984) Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. Patt. Anal. Mac. Intell.*, **6**, 621–641.

Gerasimova,A.V. *et al*. (2003) ArcA regulator of gamma-proteobacteria: identification of the binding signal and description of the regulon. *Biophysics* (*Moscow*), **48**, 21–25.

Harbison,C.T. *et al*. (2004) Transcription regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

Hughes,J.D. *et al*. (2000) Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.*, **296**, 1205–1214.

Kel-Margoulis,O.V. *et al*. (2002) TRANSCompel: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res.*, **30**, 332–334.

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

Li,H. *et al*. (2002) Identification of the binding sites of regulatory proteins in bacterial genomes. *Proc. Natl Acad. Sci. USA*, **99**, 11772–11777.

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

Liu,X. and De Wulf,P. (2004) Probing the ArcA-P modulon of *Escherichia coli* by whole genome transcriptional analysis and sequence recognition profiling. *J. Biol. Chem.*, **279**, 12588–12597.

Liu,X.S. *et al*. (2002) An algorithm for finding protein DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **20**, 835–839.

Lynch,A.S. and Lin,E.C. (1996) Transcriptional control mediated by the ArcA two-component response regulator protein of *Escherichia coli*: characterization of DNA binding at target promoters. *J. Bacteriol.*, **178**, 6238–6249.

Makeev,V.J. *et al*. (2003) Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. *Nucleic Acids Res.*, **31**, 6016–6026.

Maris,A.E. *et al*. (2002) Dimerization allows DNA target site recognition by the NarL response regulator. *Nat. Struct. Biol.*, **9**, 771–778.

McGuire,A.M. *et al*. (1999) A weight matrix for binding recognition by the redox-response regulator ArcA-P of *Escherichia coli*. *Mol. Microbiol.*, **32**, 219–221.

Mironov,A.A. *et al*. (1999) Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucleic Acids Res.*, **27**, 2981–2989.

Mironov,A.A. *et al*. (2000) Software for analyzing bacterial genomes. *Mol. Biol. (Mosk)*, **34**, 253–262.

Mwangi,M.M. and Siggia,E.D. (2003) Genome wide identification of regulatory motifs in *Bacillus subtilis*. *BMC Bioinformatics*, **4**, 18.

Pilpel,Y. *et al*. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, **29**, 153–159.

Qiu,P. *et al*. (2002) Computational analysis of composite regulatory elements. *Mamm. Genome*, **13**, 327–332.

Robert,C.P. (1998) *Discretization and MCMC Convergence Assessment*. Springer-Verlag, NY, Berlin, Heidelberg.

Robin,S. *et al*. (2002) Occurrence probability of structured motifs in random sequences. *J. Comp. Biol.*, **9**, 761–773

Roth,F.P. *et al*. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotech.*, **16**, 939–945.

Schneider,T.D. and Stephens,R.M. (1990) Sequence Logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Sivia,D.S. (1996) *Data Analysis. A Bayesian Tutorial*. Clarendon Press, Oxford.

Stormo,G.D. and Hartzell,G.W. III (1989) Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl Acad. Sci. USA*, **86**, 1183–11877.

Thijs,G. *et al*. (2002) A Gibbs sampling method to detect over-represented motifs in the upstream regions of coexpressed genes. *J. Comp. Biol.*, **9**, 447–464.

Thompson,W. *et al*. (2003) Gibbs recursive sampler: finding transcription factor binding sites. *Nucleic Acids Res.*, **31**, 3580–3585.

Tompa,M. *et al*. (2005) An assessment of computational tools for the discovery of transcription factor binding sites, *Nat. Biotech.*, **23**, 137–144.

van Helden,J. *et al*. (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.*, **28**, 1808–1818.