

Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites

Zhaohui S. Qin¹, Lee Ann McCue², William Thompson², Linda Mayerhofer², Charles E. Lawrence^{2,3}, and Jun S. Liu*¹

Published online 10 March 2003; doi:10.1038/nbt802

The identification of co-regulated genes and their transcription-factor binding sites (TFBS) are key steps toward understanding transcription regulation. In addition to effective laboratory assays, various computational approaches for the detection of TFBS in promoter regions of coexpressed genes have been developed. The availability of complete genome sequences combined with the likelihood that transcription factors and their cognate sites are often conserved during evolution has led to the development of phylogenetic footprinting^{1,2}. The *modus operandi* of this technique is to search for conserved motifs upstream of orthologous genes from closely related species^{1,2}. The method can identify hundreds of TFBS without prior knowledge of co-regulation or coexpression. Because many of these predicted sites are likely to be bound by the same transcription factor, motifs with similar patterns can be put into clusters so as to infer the sets of co-regulated genes, that is, the regulons. This strategy utilizes only genome sequence information and is complementary to and confirmative of gene expression data generated by microarray experiments. However, the limited data available to characterize individual binding patterns, the variation in motif alignment, motif width, and base conservation, and the lack of knowledge of the number and sizes of regulons make this inference problem difficult. We have developed a Gibbs sampling-based³ Bayesian motif clustering (BMC) algorithm to address these challenges. Tests on simulated data sets show that BMC produces many fewer errors than hierarchical and K-means clustering methods⁴. The application of BMC to hundreds of predicted γ -proteobacterial motifs² correctly identified many experimentally reported regulons, inferred the existence of previously unreported members of these regulons, and suggested novel regulons.

Cluster analysis has been the subject of active research for many years⁴, yet only two strategies have emerged as the main approaches for sequence motif clustering: a hierarchical clustering algorithm^{5,6} and a two-stage method based on statistical mechanics models⁷. As the latter algorithm was not publicly available, we evaluated the performance of BMC by comparing it with hierarchical clustering (implemented by algorithms CompareACE and Tree⁶ (<http://atlas.med.harvard.edu/>) and another popular clustering procedure, the K-means method (using function *kmeans* from the

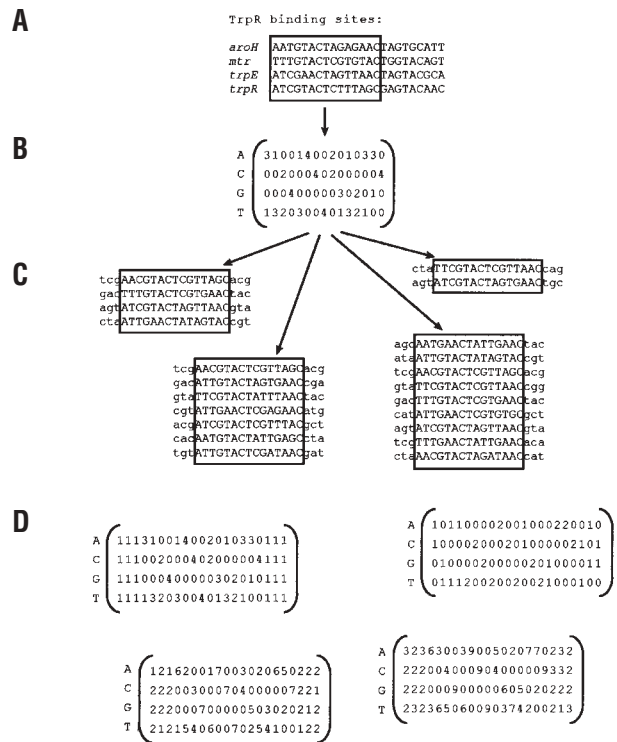


Figure 1. The procedure for simulating motif data. (A) The aligned TFBS of four genes regulated by TrpR (from DPInteract⁸). (B) The weight matrix for the first 15 positions of the alignment in (A). The j^{th} column records the numbers of nucleotides A, C, G, and T, respectively, occurring in the j^{th} position of the motif. (C) Four motifs with sizes ranging from two to ten sequences generated from the weight matrix in (B) with three positions of random nucleotides added to each end of the sequences. (D) The weight matrices of the four simulated 'TrpR' motifs.

statistical computing language R, <http://www.r-project.org/>), on simulated data sets with features that mirror data derived from a genome-scale phylogenetic footprinting study. To obtain realistic motif models, we extracted 43 position-specific weight matrices resulting from 356 experimentally determined *Escherichia coli* TFBS from DPInteract⁸ (http://arep.med.harvard.edu/ecoli_matrices/). We then simulated a motif consisting of a small and variable number of aligned sequences (henceforth, motif sizes) for each of the 356 TFBS, pretending that these motifs were discovered by cross-species comparisons. The resulting simulated regulons have the same distribution of regulon sizes as in DPInteract, which ranges from 2 to 49 motifs, with motif sizes varying over three ranges (see Experimental Protocol and Fig. 1). As both K-means and hierarchical clustering lack a mechanism to determine the number of clusters, we gave them the advantage of knowing the correct number (that is, 43). We also tested hierarchical clustering at a recommended cutoff value⁶ of 0.7. The overall performance of each clustering method is assessed by the percent clustering errors averaged over 100 trials:

$$\left[1 - \frac{\# \text{ motifs correctly assigned to regulons}}{\text{total number of motifs}} \right] \times 100$$

(see Supplementary Experimental Protocol online for definition). BMC yielded fewer errors under all tested conditions (Table 1). In particular, BMC performed better by nearly threefold, as compared to its closest competitor, for the small-motif size range, a critical aspect and common feature of phylogenetic footprinting studies.

¹Department of Statistics, Harvard University, Cambridge, MA 02138. ²The Wadsworth Center, New York State Department of Health, Albany, NY 12201. ³Computer Science Department, Rensselaer Polytechnic Institute, Troy, NY 12180. Corresponding author (jliu@stat.harvard.edu).

Table 1. Comparison of the average error rates of the BMC, K-means, and hierarchical clustering approaches for the simulated datasets^a

Motif size ^b	Item ^c	KM ^d	HC-1 ^e	HC-2 ^f	BMC-1 ^g	BMC-2 ^h
2–4 rows	Error rate	32.9 (5.2)	26.3 (6.1)	63.7 (1.7)	9.0 (1.7)	5.8 (0.2)
	No. clusters	43	43	250.0 (4.2)	34.4 (1.5)	38.0 (1.3)
2–10 rows	Error rate	33.4 (3.9)	14.1 (3.1)	25.7 (2.4)	3.3 (1.0)	2.5 (0.1)
	No. clusters	43	43	118.9 (6.7)	40.6 (1.2)	41.6 (0.6)
5–10 rows	Error rate	31.6 (4.6)	3.9 (1.1)	11.0 (1.5)	2.6 (0.4)	2.2 (0.0)
	No. clusters	43	43	66.0 (3.9)	41.4 (0.7)	42.0 (0.1)

^aThere were 356 simulated motifs belonging to 43 clusters. The consensus weight matrix for each cluster was downloaded from DPInteract⁸. Binding sites for the following DNA-binding proteins present in DPInteract were not included: RNA polymerase σ -subunits (RpoD, RpoE, RpoH, RpoN, RpoS); terminus DNA-binding protein (Tus); nucleoid proteins (DnaA, Fis, IHF, Lrp, H-NS, and HU), which bind to DNA nonspecifically or with weak specificity¹⁸; ArcA, which binds to DNA as a large oligomer¹⁹; and transcription factors with recognition sites shorter than 15 nt (FarR, MatI, TorR).

^bMotif size refers to the number of sequences in a motif.

^cError rate, averaged over 100 trials and reported as a percentage, is defined as the number of incorrectly assigned motifs divided by the total number of motifs (356) clustered. The average number of clusters generated in the 100 trials is also listed. Numbers in parenthesis are the corresponding s.d. of the value to the left.

^dK-means clustering was performed using the *kmeans* function contained in the Splus software package. Kullback-Leibler distance was used when calculating the distance matrix. Cluster number $k = 43$ was assumed known.

^eHierarchical clustering was performed by applying the programs CompareACE and Tree⁶. The cutoff value was selected such that the correct number of clusters (43) was generated.

^fHierarchical clustering was performed by the same method, but the cutoff value was fixed at the recommended value⁶ of 0.7.

^gBMC was performed without selection of motif widths. The motif width is fixed at 15 nt, and each sequence has up to 7 shifting alignment positions.

^hThe full version of BMC was performed. Each sequence can select a motif width of 13–17 nt, with up to five shifting alignment positions.

McCue *et al.*¹ have applied phylogenetic footprinting to the promoter regions upstream of over 2,000 orthologous genes from *E. coli* and six closely related γ -proteobacteria and identified 741 statistically significant motifs from these regions using palindromic models with the motif width ranging from 16 to 24 positions. Many of *E. coli*'s 4,289 genes were excluded because they were part of operons. The predicted motifs, of which two-thirds consist of

no more than five sequences, contain at least one *E. coli* site and one site from a comparison species. We identified a subset of 438 motifs for clustering by eliminating (i) one of two motifs found for divergently transcribed genes that contain the same *E. coli* sites, and (ii) motifs composed of sites exclusively from *E. coli* and *Salmonella enterica* serovar *Typhi*². The current implementation of the algorithm requires the user to choose a palindromic or nonpalindromic

Table 2. Most significant clusters from the *E. coli* phylogenetic footprints

Cluster ID	TF ^a	No. motifs ^b	No. sites ^c	Strength ^d	Genes ^e
Even-2	PurR	9	44	18.55	<i>cvpA, glyA, purE, purH, purL/yfhD, purM/upp, purR, yicE, yjcD</i>
Even-1	LexA	9	59	15.96	<i>lexA, recA, recN, ruvA, sulA/b0959, uvrD, ftsK, dinI, dinP</i>
Odd-9	PdhR	4	25	14.98	<i>pdhR, glcC/glcD, lldP, ung/yfiD</i>
Even-3	MetJ	7	47	14.58	<i>metA, metB/metJ, metE, metF, metR, abc/yaed</i>
Even-6	Fnr	6	33	13.98	<i>focA, ndh, nirB, ung/yfiD, fnr, yciD</i>
Even-41	FabR	2	14	13.22	<i>fabA, b2899</i>
Even-21	TrpR	3	12	12.59	<i>mtr, trpE/yjiV, trpR</i>
Even-4	Crp	7	48	12.34	<i>cdd, cyaA/hemC, glpT, mtlA, nagB/nagE, udp, mgIB</i>
Odd-29	?	3	22	12.19	<i>muth/ygdP, uvrB</i>
Even-35	?	2	23	12.07	<i>nrdA, nrdD</i>
Odd-50	NtrC	2	14	11.70	<i>glnA/yihK, glnL</i>
Odd-30	Fur, SoxS, OxyR	3	18	9.89	<i>sodA</i> (Fur), <i>ydiC</i> (OxyR), <i>b1821</i>
Odd-5	?	5	14	9.69	<i>fecl, gpmA, nupC/b2392, ygiD/ygiE, b2295</i>
Even-53	FruR	2	15	9.42	<i>epd, fruB</i>
Even-16	ArgR	4	17	9.03	<i>argG, hemN, pepQ/fadB, rpsP</i>
Even-50	TyrR	2	19	8.40	<i>aroF, tyrP</i>
Even-23	ArgR	3	19	8.30	<i>argR/mdh, yhbC, yifE</i>
Odd-2	OmpR, XylR	6	23	8.27	<i>fadL</i> (OmpR), <i>xylF</i> (XylR), <i>trpE/yjiV, ygaG, ykgM, yneJ</i>
Odd-28	Fur, OxyR	3	18	8.01	<i>Fur</i> (Fur), <i>gcvA, ispB/rpiU</i>

^a The transcription factors that have been experimentally determined to bind to *E. coli* sites in the motifs listed in column 6.

^b Number of motifs (genes) in the cluster.

^c Total number of binding sites (known and putative) in the cluster.

^d This is defined as the log-Bayes ratio normalized by the number of motifs in the cluster. The Bayes ratio is the ratio of the probability of the data belonging to one cluster to the probability of the data as separate motifs. For example, the data in the PurR cluster (Even-2) is $\exp(18.55) \approx 10^8$ times more likely (per motif) in a cluster than as nine separate motifs.

^e The names of the *E. coli* genes corresponding to the motifs in each cluster (that is, the motifs were identified in the promoter regions of these genes; many of these genes are known to be the first gene of an operon, but the downstream genes of those operons are not listed). Gene names not in bold face indicate that the *E. coli* site has been experimentally confirmed as a binding site for the transcription factor in column 2, unless more than one type of transcription factor is listed. Gene names in bold face either have not been experimentally confirmed as a binding site, or there is more than one type of transcription factor listed in this cluster, in which case the corresponding TF's name is given in parenthesis. The names of divergently transcribed genes are shown separated by a "/", (for example, *purM/upp*). Frequently the motif identified upstream of divergent genes contained the same *E. coli* sites; while only one of these motifs was used for clustering, both genes are indicated above (for example, *trpE/yjiV*). However in some cases, unique motifs were identified for divergent genes, and both motifs were used for clustering (for example, *muth/ygdP*; these genes have unique motifs, both of which are in cluster Odd-29).

model a priori. It is possible to modify the algorithm so that it can make this decision by itself based on the data for each individual motif. However, in this study we elected to continue using the palindromic models of McCue *et al.*¹. Because the odd-width motifs include a central unmatched position between the palindromic half-sites whereas the even ones do not, the even- and odd-width motifs were separated and clustered independently. BMC formed 85 clusters from the 224 even motifs and 96 clusters from the 214 odd motifs. Table 2 lists the most significant 19 clusters ranked by Bayes ratio for all clusters with a ratio above 8 (see Supplementary Tables 1 and 2 online; the full clustering results are available on the author's website (see URL in Experimental Protocol)). For convenience, each motif is referred to by the name of the corresponding *E. coli* gene.

The cognate transcription factors for 76 of the 438 input motifs are known^{1,2}, which allowed us to examine the ability of BMC to cluster these phylogenetic footprinting motif data. Frequently, a single cluster contained all of the confirmed motifs for a transcription factor (for example, PurR, LexA, MetJ, Crp, TrpR, NtrC, FabR, Mlc, and ModE) as well as putative motifs representing potentially unreported members of that regulon. For example, cluster Even-1 contained all of the known LexA sites in our data as well as the *dinP* motif. Although the transcription factor binding site has not been experimentally determined in the case of *dinP*, the *E. coli dinP* gene has been shown to be regulated *in vivo* by LexA⁹. Several of the clusters in Table 2 contain genes of unknown function; the inclusion of these genes together with known regulons may provide clues to their functions in *E. coli*. Table 2 also shows that some transcription factor motifs do not cluster well. For example, the known motifs for Fur, OxyR, and SoxS were clustered together (Table 2, clusters Odd-30 and Odd-28), which may reflect the fact that these TFBSs sometimes overlap (e.g., the Fur and SoxS sites overlap in the *sodA* promoter; see http://bayesweb.wadsworth.org/binding_sites/index.html). That these three regulons overlap substantially in the set of genes that they regulate, and thus may compete for binding at overlapping binding sites, makes biological sense considering that the oxidative stress response (OxyR and SoxS regulons) is linked to the availability of iron (Fur regulon)¹⁰. In addition to Odd-30 and Odd-28, Odd-5 may also represent a Fur binding motif; the input *gpmA* motif is predicted to be a Fur binding site¹¹. The fact that Fur binding sites can overlap binding sites for other transcription factors, combined with the alternate interpretations of Fur binding sites as either 19 bp palindromes or a concatenation of at least three copies of a 6 bp repeat¹², may have led to the clustering results observed here (see Supplementary Experimental Protocol online for more details).

Also of potential interest are those clusters in Table 2 for which the cognate transcription factor was not known. These clusters suggest groups of genes that are likely co-regulated by an as yet undetermined transcription factor. For example, Even-35 contains the *nrdA* and *nrdD* motifs. Both the *nrdAB* and *nrdDG* operons encode reductases that function to rectify depletion in the deoxyribonucleotide concentrations by reducing ribonucleotides to deoxyribonucleotides. The NrdAB reductase is active during aerobic growth, whereas the NrdDG reductase is active during anaerobic growth. Transcription of *nrdAB* is regulated by DnaA, Fis, and IciA^{13,14}. Microaerophilic conditions cause the upregulation of *nrdDG* expression, which is likely to be regulated by Fnr, and the downregulation of *nrdAB* expression¹³. However, *nrdAB* is expressed under microaerophilic conditions in a *nrdDG* mutant¹³. Thus, it is possible that the sites in cluster Even-35 may be related to a previously unreported regulatory mechanism of both these operons that is responsive to the relative ribonucleotide/deoxyribonucleotide concentrations in the cell. Odd-29 contains the *uvrB*, *mutH*, and *ygdp* motifs. UvrB is a member of the UvrABC

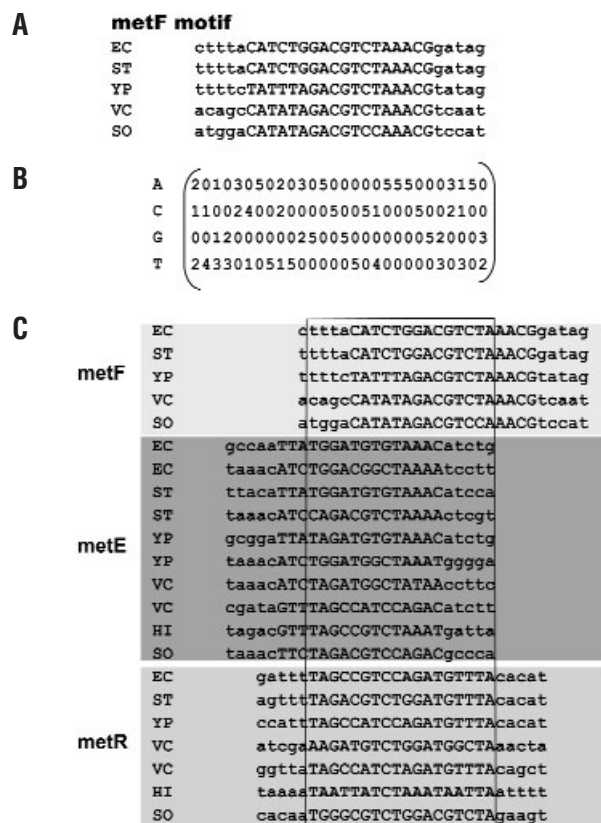


Figure 2. MetJ motif alignment. (A) A sample input motif from the phylogenetic footprint analysis of the orthologous promoter regions of the *metF* gene from *E. coli* and four additional species (ST, *Salmonella enterica* serovar Typhi; YP, *Yersinia pestis*; VC, *Vibrio cholerae*; SO, *Shewanella oneidensis*). Capital letters indicate the sequence included in the cross-species motif; lower-case letters are additional flanking nucleotides. (B) Weight matrix of the motif and flanking sequence in (A). (C) Sequence alignment of three of the seven total motifs that formed the cluster weight matrix of the Even-3 cluster (MetJ) listed in Table 2. The columns within the box were those columns included in the cluster weight matrix by BMC. Note that the alignment in the cluster of two of these three motifs has been shifted relative to their input alignment. MetJ binds to DNA containing two to five tandem repeats of an 8 nt recognition sequence (which is internally palindromic)¹⁶. Thus these input motifs for clustering were not centered on a common palindromic position, and demonstrated the importance of allowing shifting of the motifs during the clustering procedure.

nucleotide excision repair mechanism, which is induced by the SOS response in a LexA-dependent manner⁹. UvrABC is also active during normal DNA replication. The *uvrB* motif corresponds to a region upstream of the *E. coli uvrB* gene, which may contain sites for DnaA binding¹⁵. The MutH endonuclease is part of the MutHLS complex, which repairs mismatched nucleotide bases during DNA replication. YgdP is a pyrophosphatase that catalyzes the hydrolysis of diadenosine oligophosphates, which are cellular 'alarmones' formed during stress responses such as heat shock, oxidative stress, and pathogenic invasion. The motif model for this cluster exhibits partial similarity to DnaA binding sites (whose consensus sequence¹⁶ is 5'-TTATCCACA). In summary, our particular use of palindromic models may have identified a cluster of misaligned DnaA sites or, alternatively, a motif that describes the truly palindromic binding sites for an unknown transcription factor that coordinates DNA replication with regulation of DNA damage repair, DNA mismatch repair, and environmental stress.

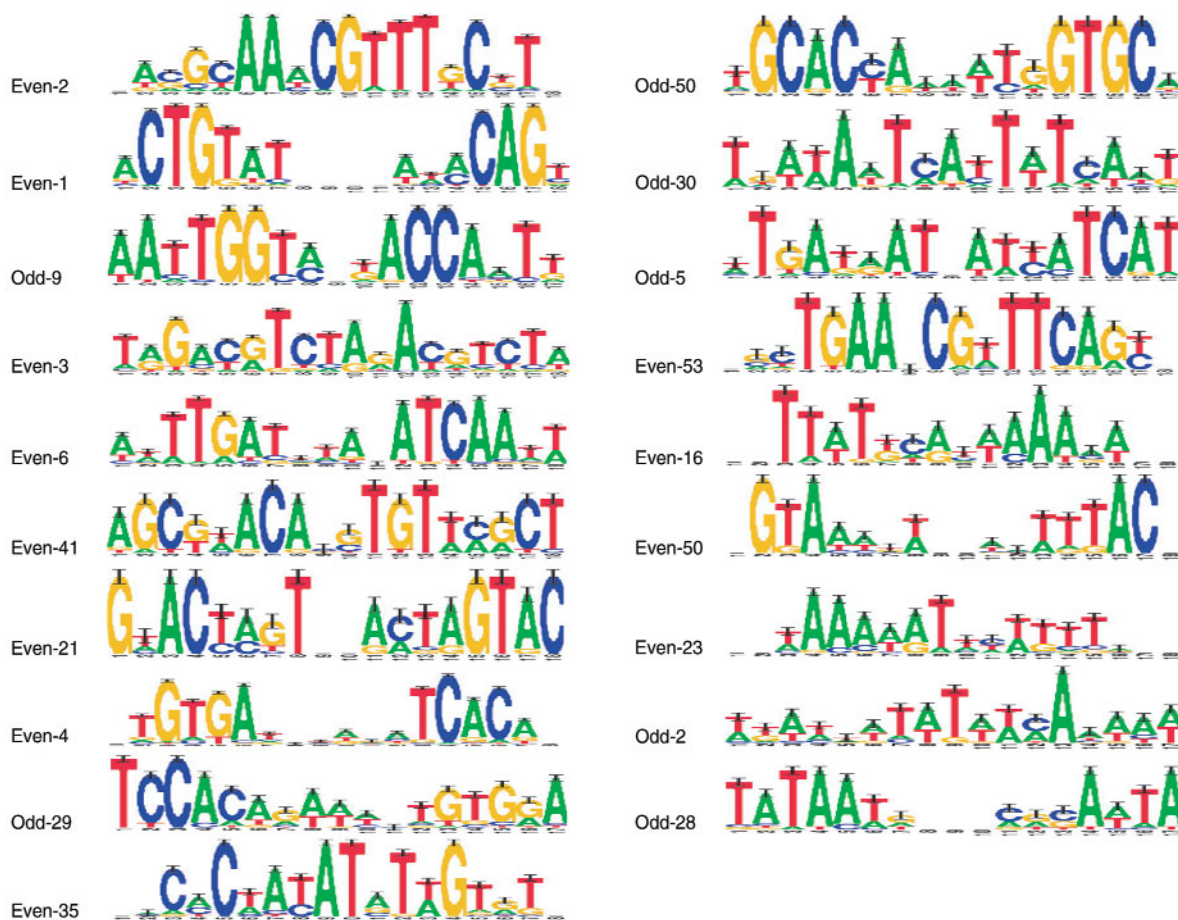


Figure 3. Motif patterns for the clusters reported in Table 2. Sequence logos²⁰ created using the software from <http://www.lecb.ncicrf.gov/~toms/logoprograms.html>.

Because the motif sizes in phylogenetic footprinting studies are small, there is a considerable uncertainty concerning their estimated binding patterns. Perhaps the most important reason for the improvements seen in the simulation study is that the model-based approach of BMC employs a ‘pooling’ strategy to combine similar motif weight matrices and thus dynamically adjusts the distances measured among them. Our observation that the clustering performance with small size motifs is substantially improved compared to other approaches is consistent with the fact that the BMC strategy tends to average out noise to capture weak signals. The incorporation of sampling steps that address variations in motif widths, misalignment among motifs, and the presence of noninformative motif positions may also have contributed to the improvements. Because the Bayes ratio ranking of predicted clusters brought nearly all of the experimentally determined regulons to the top of the cluster list, this measure may be of value in ranking clusters from phylogenetic footprinting studies of other species where few regulons have been experimentally determined. Although it was applied only to sets of binding motif patterns found by phylogenetic footprinting, the BMC algorithm should be equally effective in clustering motif patterns discovered by other means.

Experimental protocol

The *E. coli* input data consisted of the aligned binding sites predicted by the Gibbs sampler in the promoters of orthologous genes from a number of species². Fig. 2A illustrates one such motif, resulting from the alignment of five putative regulatory binding sites found upstream of the *metF* gene in five

γ -proteobacterial species. Each motif is rendered as a $4 \times w$ weight matrix, whose column j records the numbers of occurrences of each nucleotide at the aligned position j (Fig. 2B). To account for variable widths of the predicted motifs, we also include 5 nt of flanking sequence (lower-case letters in Fig. 2).

The motif data for the simulation study were generated according to the statistical model illustrated in Figure 1, which may offer some advantages to BMC. For each of the 43 transcription factors, a weight matrix was constructed based on the known sites in DPIInteract⁸. From each weight matrix Θ_k , M_k independent motifs of width 15 nt were generated, where M_k is the number of sites from DPIInteract that were used to derive Θ_k . Each motif consisted of s sites, each generated by sampling a base at each of the 15 positions in proportion to fractions in the corresponding column of Θ_k . To reflect the fact that the number of sites contributing to a motif varies, the motif size (referred to as s) was chosen uniformly over three different ranges: 2–4, 2–10, and 5–10. We also added three noninformative columns, with the four nucleotides evenly distributed, to each end of the motif. Thus, 356 motifs (of width 21) were generated, which form 43 distinct regulons containing from 2 to 49 motifs. We generated 100 such motif data sets for each of the three motif size ranges, and applied the K-means method, hierarchical clustering, and the BMC to them.

The algorithm begins by randomly assigning the n motifs into an arbitrary number of $N \leq n$ clusters. Let $E = (e_1, \dots, e_n)$ denote the clustering information, in which $e_i = k$ means that the i^{th} motif belongs to the k^{th} cluster. On each iteration, the algorithm considers the reassignment of one of the motifs, say the i^{th} one, given the current assignment of all the other motifs. First, the probability that this motif belongs to one of the current N clusters, or begins a new cluster by itself, is calculated: $Q_l = P(e_i = l \mid e_0, \dots, e_{i-1}, e_{i+1}, \dots, e_n, \text{data})$, $l = 0, \dots, N$, where $l = 0$ indicates the formation of a new cluster, and $l > 0$ indicates assignment to an existing cluster. Before considering the data, we assume that the motif belongs to an existing cluster ($l > 0$) with probability $(q + N)^{-1}$ and

forms a new cluster ($l = 0$) with probability $q(q + N)^{-1}$, where q is a free parameter (see Supplementary Table 3 online). Second, we update e_i by a value drawn from $\{0, \dots, N\}$ with probability proportional to (Q_0, Q_1, \dots, Q_N) . Cluster number N is also updated according to whether a new cluster is formed or an existing cluster with a single member is removed.

The widths of the motifs are not always the same and the motifs (see logos in Figure 3) are not necessarily aligned with each other (see Fig. 2C). In addition, some positions within a binding site are not well conserved and should not be used for clustering. To overcome these potential problems, we take the following steps: (i) a procedure similar to the Gibbs sampler⁸ is employed to re-align the motifs within a cluster (in each iteration step, the motif under reassignment consideration is allowed to shift up to k nt in either direction, and the alignment is sampled according to its posterior probability), and (ii) a 'fragmentation' procedure¹⁷ is employed to select informative motif positions or widths (the columns of a cluster's weight matrix are considered one by one, being turned on or off according to the ratio of the probability of the data in the column as described by the clustered motif model over its probability in individual motifs).

BMC took ~150 minutes to complete the analysis (150 iteration cycles) of the γ -proteobacteria phylogenetic footprinting data, for even- or odd-width motifs, on a Dell OptiPlex GX110 PC with a single Pentium 800 MHz CPU. A more detailed account of the model, prior, posterior, and mathematical definitions can be found in the Supplementary Experimental protocol online.

URL. The BMC algorithm and more clustering results (with Bayes ration between 6 and 8) can be found on authors' website: <http://www.fas.harvard.edu/~junliu/clust/>.

Note: Supplementary information is available on the Nature Biotechnology website.

Acknowledgments

This research was partly supported by US National Institutes of Health (NIH) grant R01HG02518-01 and National Science Foundation grants DMS-0104129 and DMS-0204674 to J.S.L., NIH grant R01HG01257 to C.E.L., and Department of Energy grant DEFG0201ER63204 to C.E.L. and L.A.M.

Competing interests statement

The authors declare that they have no competing financial interests.

Received 6 May 2002; accepted 7 December 2002

- McCue, L.A. *et al.* Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.* **29**, 774–782 (2001).
- McCue, L.A., Thompson, W., Carmack, C.S. & Lawrence, C.E. Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res.* **12**, 1523–1532 (2002).
- Liu, J.S. Monte Carlo Strategies in Scientific Computing (Springer, New York, 2001).
- Everitt, B.S., Landau, S. & Leese, M. *Cluster Analysis*, edn. 4 (Arnold, London, 2001).
- Petrokovski, S. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.* **24**, 3836–3845 (1996).
- Hughes, J.D., Estep, P.W., Tarazoie, S. & Church, G.M. Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* **296**, 1205–1214 (2000).
- van Nimwegen, E., Zavolan, M., Rajewsky, N. & Siggia, E.D. Probabilistic clustering of sequences: inferring new bacterial regulons by comparative genomics. *Proc. Nat. Acad. Sci. USA.* **99**, 7323–7328 (2002).
- Robison, K., McGuire, A.M. & Church, G.M. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol.* **284**, 241–254 (1998).
- Courcelle, J., Khodursky, A., Peter, B., Brown, P.O. & Hanawalt, P.C. Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient *Escherichia coli*. *Genetics* **158**, 41–64 (2001).
- Hantke, K. Iron and metal regulation in bacteria. *Curr. Opin. Microbiol.* **4**, 172–177 (2001).
- Vassinova, N. & Kozyrev, D. A method for direct cloning of Fur-regulated genes: identification of seven new Fur-regulated loci in *Escherichia coli*. *Microbiology* **146**, 3171–3182 (2000).
- Escobar, L., Perez-Martin, J. & de Lorenzo, V. Opening the iron box: transcriptional metalloregulation by the Fur protein. *J. Bacteriology* **181**, 6223–6229 (1999).
- Jordan, A. & Reichard, P. Ribonucleotide reductases. *Annu. Rev. Biochem.* **67**, 71–98 (1998).
- Han, J.S., Kwon, H.S., Yim, J.-B. & Hwang, D.S. Effect of IciA protein on the expression of the *nrd* gene encoding ribonucleoside diphosphate reductase in *E. coli*. *Mol. Gen. Genet.* **259**, 610–614 (1998).
- van den Berg, E.A., Geerse, R.H., Memelink, J., Bovenberg, R.A., Magnee, F.A. &

- van der Putte, P. Analysis of regulatory sequences upstream of the *E. coli* *uvrB* gene; involvement of the DnaA protein. *Nucleic Acids Res.* **13**, 1829–1840 (1985).
- Neidhardt, F.C. *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology (ASM Press, Washington, DC, 1996).
- Liu, J.S., Neuwald, A.F. & Lawrence, C.E. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Amer. Stat. Assoc.* **90**, 1156–1170 (1995).
- Azam, T.A. & Ishihama, A. Twelve species of the nucleoid-associated protein from *Escherichia coli*. Sequence recognition specificity and DNA binding affinity. *J. Biol. Chem.* **274**, 33105–33113 (1999).
- Jeon, Y., Lee, Y.S., Han, J.S., Kim, J.B. & Hwang, D.S. Multimerization of phosphorylated and non-phosphorylated ArcA is necessary for the response regulator function of the Arc two-component signal transduction system. *J. Biol. Chem.* **276**, 40873–40879 (2001).
- Schneider, T.D. & Stephens, R.M. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100 (1990).