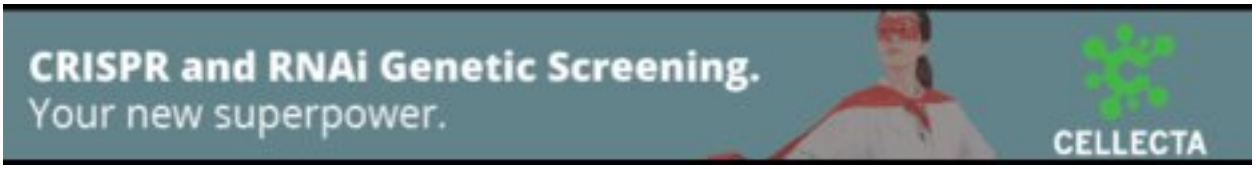
Factors Influencing the Identification of Transcription Factor Binding Sites by Cross-Species Comparison

Lee Ann McCue, William Thompson, C. Steven Carmack, et al.

Genome Res. 2002 12: 1523-1532
Access the most recent version at doi:10.1101/gr.323602



Letter

Factors Influencing the Identification of Transcription Factor Binding Sites by Cross-Species Comparison

Lee Ann McCue,[1,3] William Thompson,[1] C. Steven Carmack,[1] and Charles E. Lawrence[1,2]

[1]*The Wadsworth Center, New York State Department of Health, Albany, New York 12201-0509, USA;* [2]*Computer Science Department, Rensselaer Polytechnic Institute, Troy, New York 12180, USA*

As the number of sequenced genomes has grown, the questions of which species are most useful and how many genomes are sufficient for comparison have become increasingly important for comparative genomics studies. We have systematically addressed these questions with respect to phylogenetic footprinting of transcription factor (TF) binding sites in the γ-proteobacteria, and have evaluated the statistical significance of our motif predictions. We used a study set of 166 *Escherichia coli* genes that have experimentally identified TF binding sites upstream of the gene, with orthologous data from nine additional γ-proteobacteria for phylogenetic footprinting. Just three species were sufficient for ~74.0% of the motif predictions to correspond to the experimentally reported *E. coli* sites, and important characteristics to consider when choosing species were phylogenetic distance, genome size, and natural habitat. We also performed simulations using randomized data to determine the critical maximum a posteriori probability (MAP) values for statistical significance of our motif predictions ($P = 0.05$). Approximately 60% of motif predictions containing sites from just three species had average MAP values above these critical MAP values. The inclusion of a species very closely related to *E. coli* increased the number of statistically significant motif predictions, despite substantially increasing the critical MAP value.

[Supplemental material is available online at http://www.genome.org. In addition, our motif predictions for the study set and the entire *E. coli* genome are available online at http://www.wadsworth.org/resnres/bioinfo/.]

In this era of whole-genome sequencing and analysis, the importance of comparative genomics has become increasingly clear. The availability of genomic data has had a huge impact on our views of species evolution and genetic diversity (Huynen and Bork 1998; Snel et al. 1999; Eisen 2000; Tamames 2001) and provided a means for studying the molecular basis of pathogenicity at the genomic level, thereby accelerating drug and vaccine development (Behr et al. 1999; Pizza et al. 2000; Porcella and Schwan 2001). The methods for predicting gene location and function have also been revolutionized, as has the identification of probable regulatory signals (for review, see Hardison et al. 1997; Gelfand 1999; O'Brien et al. 1999; Galperin and Koonin 2000). As the methods for comparative genomics have developed and genomic data have accumulated, interest in the questions of which species are most useful and how many genomes are sufficient for comparison has increased. Although this is particularly true for vertebrate or mammalian species (Miller 2000; O'Brien et al. 2001), these questions are equally important for all taxonomic groups. Particularly now, as more genomes are being sequenced specifically for comparative studies, quantitative analyses of species selection are needed. Cliften and coworkers (2001) have addressed this question for *Saccharomyces* species, and their observations, although on a subgenomic level and nonquantitative, highlight the difficulty of species selection when predicting regulatory sequences by comparative genomics.

The availability of whole-genome sequence data for many prokaryotes has already provided the opportunity to identify probable transcription factor (TF) binding sites via cross-species comparison (i.e., phylogenetic footprinting), thus alleviating the need to experimentally identify sets of genes that are co-regulated within a species to identify common regulatory motifs (reviewed by Gelfand 1999). Such comparative genomics approaches have led to the identification of several probable regulatory sites and regulons in archaeal and eubacterial species (Mironov et al. 1999; Gelfand et al. 2000; McGuire et al. 2000; Rodionov et al. 2000, 2001a,b; Laikova et al. 2001; Makarova et al. 2001; McCue et al. 2001; Panina et al. 2001a,b; Tan et al. 2001; Terai et al. 2001; Rajewsky et al. 2002). Among these previous studies, the analyses range from the comparison of species representing very diverse phylogenetic groups (e.g., *E. coli* and *Bacillus subtilis* in McGuire et al. [2000]), to the comparison of very closely related organisms (e.g., enterobacteria in Rodionov et al. [2001a]). In a recent analysis of nine γ-proteobacterial species, we showed not only that a comparative genomics approach could predict sites with reasonable accuracy but also that the motif predictions could be used as the basis for identification of cognate TFs via affinity purification (McCue et al. 2001). Specifically, we showed that a probable TF, FabR (YijC), binds to predicted sites upstream of *fabA*, *fabB*, and *yqfA* in vitro. The physiological relevance of this interaction has been

[3]Corresponding author.
E-MAIL mccue@wadsworth.org; FAX (518) 473-2900.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.323602. Article published online before print in September 2002.

Table 1. Species Characteristics

Species	Family	Predominant habitat(s)	Genome size[a]	ORFs	%G + C content
Escherichia coli K-12, strain MG1655	Enterobacteriaceae	Mammalian intestine; contaminated food/water	4,639,221 bp (C)	4290	50.7
Salmonella enterica serovar Typhi CT-18	Enterobacteriaceae	Human intestine & bloodstream; contaminated food/water	4,809,037 bp (C)	4599	52.1
Yersinia pestis CO-92, Biovar Orientalis	Enterobacteriaceae	Human & rodent bloodstream	4,653,728 bp (C)	4012	47.6
Buchnera sp. APS	Enterobacteriaceae, aphid endosymbionts	bacteriocytes of aphid (obligate endosymbiont)	640,681 bp (C)	583	26.3
Haemophilus influenzae Rd, strain KW20	Pasteurellaceae	Human nasopharynx	1,830,138 bp (C)	1738	38.0
Vibrio cholerae El Tor, strain N16961	Vibrionaceae	Human small bowel	4,033,464 bp (C)	3890	47.5
Shewanella oneidensis MR-1	Alteromonadaceae	Fresh & marine water/sediments	~4.50 Mbp (P)	n.a.	~46
Pseudomonas aeruginosa PAO1	Pseudomonadaceae	Soil; human opportunistic pathogen	6,264,403 bp (C)	5570	66.6
Acidithiobacillus ferrooxidans ATCC 23270	Unclassified	Acidic water/soil	~2.90 Mbp (P)	n.a.	~59
Xylella fastidiosa strain 9a5c	Xanthomonas	Xylem of host plant	2,679,306 bp (C)	2782	52.7

ORFs indicates open reading frames; n.a., not available.
[a]Genome sizes are from the chromosome data only and do not include plasmid sequences; C indicates complete genome; P, genome sequencing in gap closure.

confirmed by the demonstration that FabR regulates *fabA* and *fabB* in vivo, and is important for controlling unsaturated fatty acid production in *E. coli* (Zhang et al. 2002).

In this report, we have identified 10 γ-proteobacterial genomes that are currently complete or in final gap closure (Table 1) with which to address the question of species selection for comparative genomics. Although the genome sequences of many prokaryotic species are becoming available, there are still no other groups of closely related species with as many sequenced genomes as the γ-proteobacteria. In addition, the wealth of experimental data available for *E. coli* allows for the validation of predictions from this phylogenetic group. The 10 species are of varying evolutionary distance to one another (Fig. 1) and inhabit diverse natural habitats (Table 1). We have focused on the questions of species selection and number of species required for comparative genomics through the application of phylogenetic footprinting using subsets of these γ-proteobacterial genomes. Simple species characteristics and the number of TF orthologs present were good predictors for which species were likely to have common regulatory mechanisms and, therefore, contribute substantially to phylogenetic footprinting. Good results were achieved with data from only three species. We have also developed a method with which to address the statistical significance of our motif predictions.

RESULTS

Description of the Study Set

We identified a set of 166 genes in the *E. coli* genome, for which 48 different TFs have been experimentally confirmed to activate or repress expression by sequence-specific binding to their cognate sites in the upstream intergenic region. With one exception (*hycB*), all 166 genes are the first gene of a transcription unit (i.e., operon), several of which are polycistronic; thus, 453 total genes are regulated as part of these transcription units in *E. coli*. The 166 genes (and their downstream, co-regulated genes) are part of 48 regulons, that is, they are regulated by 48 sequence-specific TFs binding to a total of 360 confirmed binding sites in the promoters of the 166 genes. In this study we chose to focus on the identification of binding sites for specific activators and repressors, proteins that likely could be identified by affinity chromatogra-

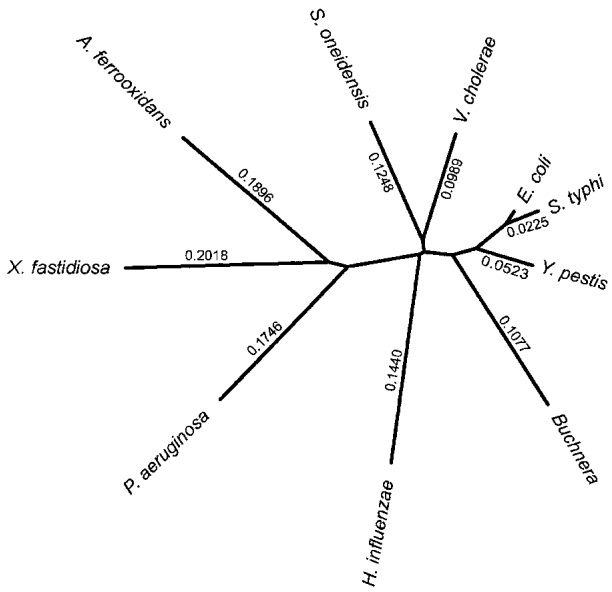


Figure 1 Phylogeny of the 10 γ-proteobacterial species inferred from 16S rRNA sequences (see Methods). Branch lengths are the calculated distance from *Escherichia coli* to each of the other species, measured as the number of expected nucleotide substitutions per site.

phy in the manner that FabR was identified (McCue et al. 2001). Therefore, we did not include in our study set genes for which the only known regulatory sites are RNA stem-loop structures or binding sites for nucleoid proteins (see Methods). We used this study set of 166 genes to systematically analyze which of the γ-proteobacterial species *Salmonella enterica* serovar typhi (*S. typhi*), *Yersinia pestis*, *Buchnera* sp. APS, *Haemophilus influenzae*, *Vibrio cholerae*, *Shewanella oneidensis*, *Pseudomonas aeruginosa*, *Acidithiobacillus ferrooxidans*, and *Xylella fastidiosa* (Table 1) would contribute substantially to phylogenetic footprinting with *E. coli*, with the express purpose of making predictions that would lead to affinity purification of novel TFs. The complete list of 166 genes included in the study set is available at our web site (http://www.wadsworth.org/resnres/bioinfo/) in two formats: (1) as a table of the 166 genes, indicating the confirmed TF binding sites for each, and (2) as a table of each of the 48 regulons, indicating the genes in the study set with confirmed binding sites for that TF. A list of the species in which probable orthologs were identified is also available for each of the genes.

Species Selection

First, we wished to determine which species were most useful for phylogenetic footprinting. Genomic-scale phylogenetic footprinting requires the identification of a set of related species that have orthologs to as many genes from the reference species as possible. Therefore, the first indication as to which species were likely to be useful was the simple observation of the number of probable orthologs that could be identified in each species for the 166 genes of our study set. In addition, the presence of orthologous TFs is a more specific indicator of similarity in transcription regulation; thus, we also tabulated the number of orthologs for the 48 TFs that regulate these 166 genes in *E. coli*. For each of the species in Table 1 and Figure 1, the number of orthologs that we identified correlated well with the phylogenetic distance and genome size relative to *E. coli* (Table 2). The low number of orthologs and, in particular, the low number of orthologous TFs identified in a few of the more distantly related species (*Buchnera*, *A. ferrooxidans*, *X. fastidiosa*), indicated that they were unlikely to contribute substantially to phylogenetic footprinting.

We further tested each of the nine species by systematically analyzing subsets of the complete study set via phylogenetic footprinting with the Gibbs sampler (see Methods), where the subsets contained the available orthologous data for each species paired with *E. coli* (Table 3), as well as a number of additional species combinations (see supplemental data online). Our Gibbs sampling method does not require that the motif prediction include a TF binding site from each sequence, but instead allows zero, one, or two sites per sequence in an orthologous data set (McCue et al. 2001). Therefore, in the analysis described below, we focused on those motif predictions that included at least one site from each of the species in the subset under analysis, to identify those species that contributed to motif predictions most frequently. In addition, when calculating the correspondence of predicted sites with known TF binding sites, we required that a predicted *E. coli* site overlap a known TF binding site by ≥10 bp. Consequently, an oligonucleotide of the *E. coli* site plus a few bases of flanking sequence would likely contain the entire TF recognition sequence, thereby allowing affinity purification of the TF. We did not count those motif predictions that corresponded to RNA stem-loop structures or nucleoid protein binding sites for genes that are in the study set owing to the presence of specific activator or repressor binding sites in their upstream region (e.g., *trpE*, which is regulated by specific binding of the TrpR repressor and by attenuator RNA stem-loops).

We identified probable orthologs for a large proportion of the study set genes and their TFs in *S. typhi*, *Y. pestis*, and *V. cholerae*. Also, the Gibbs sampling results from data sets pairing each of these species with the *E. coli* data (Table 3) revealed that for these combinations of two species, the majority of motif predictions included sites from both species (83.0% to 98.8%), and there was reasonable correspondence with known sites among those predictions (55.3% to 66.9%). These results indicate that *S. typhi*, *Y. pestis*, and *V. cholerae* have many regulatory mechanisms in common with *E. coli*, thus making these species useful for phylogenetic footprinting.

H. influenzae is considered a close relative of *E. coli* and has been used frequently for comparative genomics studies (Mironov et al. 1999; McGuire et al. 2000; Rodionov et al. 2000; Laikova et al. 2001; Makarova et al. 2001; Panina et al. 2001b; Tan et al. 2001). However, it has a significantly smaller genome than does *E. coli*, resulting in the identification of relatively few orthologs for the study set. But, among the available data sets, the Gibbs sampling results for *H. influenzae* paired with *E. coli* showed that the majority of motif predictions included sites from both species (81.9%) and had good correspondence with known sites (62.7%). The advantage—as well as the disadvantage—of using the *H. influenzae* genome became evident when we looked at the marginal contribution of data from this species added to the *E. coli–V. cholerae* data. The number of genes with orthologs in all three species (65 genes) was dramatically less than that in the *E. coli–V. cholerae* data (112 genes), and the percentage of motif predictions that included all of the species dropped from 83.0% to 50.8% when *H. influenzae* was added. However, the inclusion of *H. influenzae* increased the proportion of predictions that corresponded with known sites among those motif predictions that included sites from all the species (from 62.4% for *E. coli–V. cholerae* to 75.8% for *E. coli–V. cholerae–H. influenzae*). These effects were also observed with the addi-

Table 2. Orthologous Data for Each Species

Species	Study set orthologs	TF orthologs[a]	Mean % identity of intergenic regions[b]
Escherichia coli	166	48	100
Salmonella typhi	161	46	70.0 ± 12.4
Yersinia pestis	138	40	48.1 ± 7.9
Buchnera sp.	27	1	39.6 ± 3.4
Haemophilus influenzae	72	21	41.5 ± 4.9
Vibrio cholerae	112	31	41.6 ± 6.0
Shewanella oneidensis	91	24	41.1 ± 5.0
Pseudomonas aeruginosa	92	25	37.7 ± 3.7
Acidithiobacillus ferrooxidans	46	11	39.2 ± 3.7
Xylella fastidiosa	50	7	38.7 ± 3.4

[a]Transcription factors (TF) that regulate expression of the study set genes in *E. coli*.
[b]GAP (GCG Wisconsin Package version 10.2) was used to calculate the % identity (relative to *E. coli*) of the available orthologous intergenic sequences.

Table 3. Phylogenetic Footprinting Results for Selected Species Combinations[a]

Species[b]	Number of orthologous data sets	Motif predictions with all species included	Correspondence with known sites[c]
EC-ST	161	98.8% (159/161)	55.3% (88/159)
EC-YP	138	94.2% (130/138)	66.9% (87/130)
EC-VC	112	83.0% (93/112)	62.4% (58/93)
EC-HI	72	81.9% (59/72)	62.7% (37/59)
EC-SO	91	69.2% (63/91)	55.6% (35/63)
EC-PA	92	48.9% (45/92)	37.8% (17/45)
EC-BU	27	85.2% (23/27)	39.1% (9/23)
EC-AF	46	58.7% (27/46)	48.1% (13/27)
EC-XF	50	54.0% (27/50)	33.3% (9/27)
EC-VC-HI	65	50.8% (33/65)	75.8% (25/33)
EC-YP-SO	86	36.0% (31/86)	67.7% (21/31)
EC-YP-PA	86	19.8% (17/86)	70.6% (12/17)
EC-ST-YP-VC	103	42.7% (44/103)	86.4% (38/44)

[a]Results for all species combinations are in the Supplementary Data.
[b]Species abbreviations are as follows: EC, *Escherichia coli;* ST, *Salmonella typhi;* YP, *Yersinia pestis;* VC, *Vibrio cholerae;* HI, *Haemophilus influenzae;* SO, *Shewanella oneidensis;* PA, *Pseudomonas aeruginosa;* BU, *Buchnera* sp.; AF, *Acidithiobacillus ferrooxidans;* and XF, *Xylella fastidiosa.*
[c]Correspondence with known *E. coli* transcription factor binding sites was calculated only for those motif predictions that contained sites from all the species included in the data set.

tion of *H. influenzae* to other species pairs (see supplemental data online). Therefore, although relatively few orthologs were detected in the smaller genome of *H. influenzae*, inclusion of the *H. influenzae* data had a positive effect on the correspondence with known sites for those genes.

We identified a large number of orthologs in *S. oneidensis* despite the fact that this species is somewhat distantly related to, and has a natural habitat distinct from that of *E. coli*. This is perhaps not surprising, however, given that *E. coli* and *V. cholerae* show considerable overlap in gene content (Heidelberg et al. 2000), and that the preliminary results from the genome project for *S. oneidensis* indicate a significant similarity to *V. cholerae* in some metabolic pathways (Fraser 2000). And, in fact, the Gibbs sampling results from *S. oneidensis* paired with *E. coli* showed that 69.2% of the motif predictions included sites from both species, and the correspondence with known sites was 55.6%. However, we observed a similar reduction in the percentage of motif predictions with all species represented when we added *S. oneidensis* to existing species combinations such as *E. coli–Y. pestis* (from 94.2% to 36.0%; see Table 3 and supplemental data online), indicating that gene regulation of some metabolic pathways may have diverged during the adaptation to different environments. Nevertheless, as for *H. influenzae*, the correspondence with known sites benefited when data from *S. oneidensis* were included.

Although it is possible to find many orthologs in the relatively distantly related species *P. aeruginosa*, we observed that the *P. aeruginosa* data were frequently not represented in the motif predictions, and that correspondence with known sites was poor. When this species was paired with *E. coli*, only 45 of the 92 motif predictions included data from *P. aeruginosa* (48.9%), and of those predictions, only 17 corresponded to known regulatory sites in *E. coli* (37.8%). We also observed a marginal contribution from *P. aeruginosa* to existing paired species combinations, similar to that observed with *S. oneidensis*. Specifically, correspondence with known sites increased somewhat when data from all three species were included (e.g., 66.9% to 70.6% when *P. aeruginosa* was added to *E. coli–Y. pestis*), but few motif predictions included sites from all three species (only 17 for *E. coli–Y. pestis–P. aeruginosa*). These results likely reflect the fact that as phylogenetic distance increases between the species so does the chance that gene regulation of orthologs differs. For example, our study set contains several *E. coli* genes from the arginine biosynthetic pathway for which probable *P. aeruginosa* orthologs were identified (*argA*, *argC*, *argD*, *argG*, *carA*). However, *P. aeruginosa* data were not included in the motif predictions for these genes. In fact, the arginine repressor has undergone nonorthologous gene displacement (defined in Galperin and Koonin 2000) between these two species. Therefore, in this case the regulatory sites are not conserved, because the regulatory proteins are no longer similar (Park et al. 1997a,b).

The remaining three species had relatively few orthologs in the study set and contributed little to phylogenetic footprinting. *Buchnera*, which has undergone extreme genome reduction compared to our reference species *E. coli*, had only 27 identified orthologs in our study set, as well as only one ortholog for the TFs that regulate the study set genes in *E. coli*. These results were in agreement with the *Buchnera* genome annotation and analysis, which reported an almost complete lack of regulatory proteins in that genome (Shigenobu et al. 2000). When we applied our Gibbs sampling procedure to the data sets containing only *E. coli* and *Buchnera* data (27 data sets), we observed a poor correspondence of our motif predictions with the known *E. coli* regulatory sites (39.1%), confirming that the *Buchnera* genome contributed little to phylogenetic footprinting with *E. coli*. Additionally, relatively few study set or TF orthologs were identified for *A. ferrooxidans* or *X. fastidiosa*, which are relatively distantly related to, and also inhabit dramatically different environments from, *E. coli* (see Table 1). The results from Gibbs sampling with the data sets containing data from either of these species paired with *E. coli* indicated that sites from these species were not often included in the motif prediction. In other words, many motif predictions (19 of 46 predictions from the *E. coli–A. ferrooxidans* data sets, and 23 of 50 predictions from the *E. coli–X. fastidiosa* data sets) contained *E. coli* sites only. Among those motif predictions that did include sites from these species, correspondence with the known *E. coli* TF binding sites was poor: 48.1% in the *E. coli–A. ferrooxidans* data sets and 33.3% in the *E. coli–X. fastidiosa* data sets. These results may reflect uncertainty in the ortholog identification or that transcription regulation in these distant species frequently differs from that in *E. coli*.

Number of Species

We continued to analyze higher-level species combinations, performing phylogenetic footprinting for all combinations of three or four species that included *E. coli*, but excluding *Buchnera*, *A. ferrooxidans*, and *X. fastidiosa*. The three plots in Figure

2 represent the three columns of data in Table 3, expanded to include the combinations of three (15 combinations) or four (20 combinations) species. Figure 2A shows the steady decrease in the number of orthologous data sets with specific combinations of species, and Figure 2B shows the concomitant decrease in the percentage of motif predictions that included sites from all the species present in the combination, as the number of species increased. Conversely, the correspondence with known sites when all species were included in the motif prediction increased steadily as species were added, with medians of 59.0%, 75.0%, and 78.9% among these combinations of two, three, or four species, respectively (Fig. 2C).

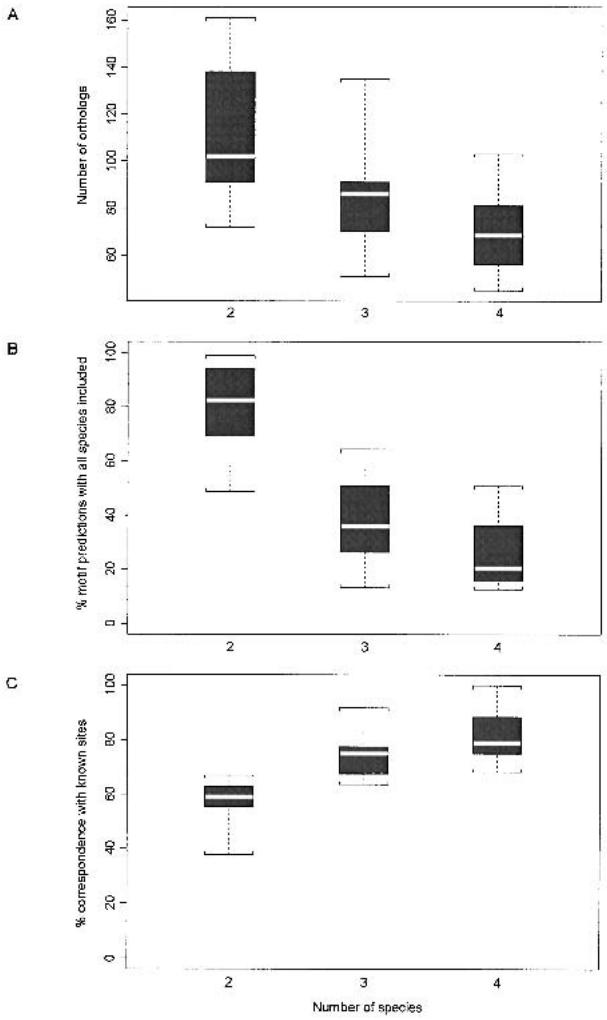


Figure 2 Boxplots representing the phylogenetic footprinting results of the study set for several species combinations: six combinations of two species, 15 combinations of three species, and 20 combinations of four species (see supplemental data for details). (*A*) The number of orthologous data sets. For combinations of two species, the upper boundary was the *Escherichia coli–Salmonella enterica* serovar typhi (*S. typhi*) combination, with 161 data sets, and the lower boundary was the *E. coli–Haemophilus influenzae* combination, with 72 data sets. (*B*) The percentage of motif predictions that included sites from all of the species in the data for each combination of species. For combinations of two species, the upper boundary was the *E. coli–S. typhi* combination, at 98.8%, and the lower boundary was the *E. coli–Pseudomonas aeruginosa* combination, at 48.9%. (*C*) The percent correspondence with known transcription factor binding sites for each combination of species. For combinations of two species, the upper boundary was the *E. coli–Yersinia pestis* combination, at 66.2%, and the lower boundary was the *E. coli–P. aeruginosa* combination, at 35.5%. The whiskers represent the species combinations with the highest and lowest numbers (*A*) or the highest and lowest percentages (*B*,*C*); the black boxes encompass the regions between the upper and lower quartiles, and the white lines indicate the medians.

The decrease in the number of orthologous data sets with specific combinations of species can be offset somewhat by using data sets that contain all of the available orthologous data from the following species set: *E. coli*, *S. typhi*, *Y. pestis*, *V. cholerae*, *H. influenzae*, *S. oneidensis*, and *P. aeruginosa*. For example, when all of the orthologous data in the study set were included, 146 data sets (88.0%) contained orthologous data from at least three species, more than for any single combination of three specific species (Fig. 2A). Phylogenetic footprinting of these 146 data sets showed that 116 (79.5%) yielded motif predictions that included sites from at least three species, and that correspondence with known sites remained high (74.1%). Analysis of the *E. coli* genome (2086 data sets) in this manner revealed that 1605 of the data sets (76.9%) included orthologous data from at least three species, and among those, 1076 of the motif predictions (67.0%) included sites from at least three species.

Determination of Critical MAP Values

In addition to determining which species were useful for phylogenetic footprinting, we wished to develop a method with which to evaluate the statistical significance of the Gibbs sampling motif predictions such that our results could be prioritized for TF affinity purification experiments. Our phylogenetic footprinting method generates multiple motif predictions for each data set (McCue et al. 2001), which in this report were ordered by their maximum a posteriori probability (MAP) values divided by the number of sites in the motif prediction, yielding what we refer to as the average MAP (see Methods). We performed simulations using randomized data to determine the critical MAP values for statistical significance of our motif predictions ($P = 0.05$) based on average MAP values. The data sets (derived from the 166 data sets in our study set) contained randomized *E. coli* sequence plus additional k random sequences ($1 \leq k \leq 9$), and the same Gibbs sampling strategy was used that had been used with the genuine data (see Methods). The average MAP values obtained with these randomized data were used to determine the 95% quantile, which we refer to as the critical MAP value. The critical MAP values obtained with these random uncorrelated data decreased only slightly as additional sequences were added (Fig. 3).

However, when the mean percent identity of the upstream intergenic sequence of the 166 genes from each of the species (as available) relative to *E. coli* was determined (Table 2), we observed that the mean percent identity between *E. coli* and *S. typhi* was considerably higher (70% ± 12.4%) than that for the other species. This indicated that there may be considerable sequence conservation owing to phylogenetic relatedness as opposed to functional constraints. For this species pair, the percent identity also varied widely (35.6% to 95.9%), as reflected by the high SD. Similarly, the mean percent identity between *E. coli* and *Y. pestis* was somewhat higher than that expected by chance (48% ± 7.9%). Because of this observed correlation in sequence identity within the data sets, we then performed additional simulations with randomized data, with the exception that one of the added "random"

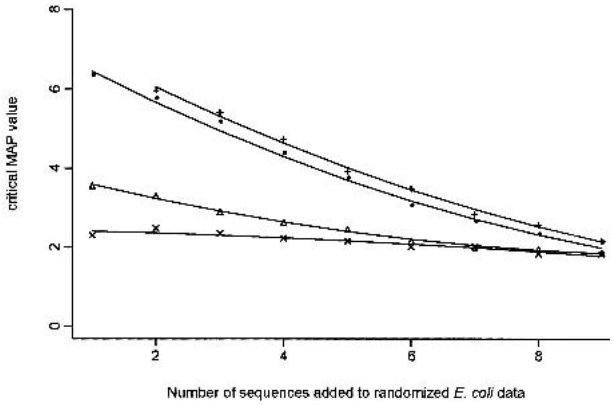


Figure 3 The critical maximum a posteriori probability (MAP) values for the 95% quantile ($P = 0.05$) calculated from the simulations for randomized *Escherichia coli* data plus k additional sequences ($1 \leq k \leq 9$): All additional sequences were randomized (*crosses*); one sequence was added at 48% identity (on average) to the randomized *E. coli* sequence, and additional sequences were randomized (*triangles*); one sequence was added at 70% identity (on average) to the randomized *E. coli* sequence, and additional sequences were randomized (*circles*); one sequence was added at 48% identity (on average), another sequence was added at 70% identity (on average) to the randomized *E. coli* sequence, and additional sequences were randomized (*plus symbols*).

sequences was forced to retain 70% or 48% identity on average to the randomized *E. coli* sequence to represent *S. typhi* and *Y. pestis* data, respectively (see Methods). These data sets were used to determine the respective critical MAP values, as described above, for data sets containing such correlated data. Our results from these simulations clearly show that considerably higher critical MAP values are the price that must be paid with correlated sequence data. However, the critical MAP values steadily decreased as up to nine additional random sequences were added, approaching the critical MAP value obtained with any number of random uncorrelated sequences (Fig. 3).

Analysis of the Data With Respect to the Critical MAP Values

We further analyzed our phylogenetic footprinting results for the study set with respect to the motif predictions above and below the critical MAP values, using data sets that contained all of the available orthologous data from several species (Table 4). Specifically, we analyzed data sets containing data from the seven species that we had previously determined to be useful, as well as data sets that excluded *S. typhi* data. With these combinations of species, the percentage of motif predictions with average MAP values above the critical MAP value increased steadily with the number of species that were included in the motif prediction, and the correspondence with known sites also showed a modest increase. Because of the dramatic increase in the critical MAP value caused by correlated sequence data (*E. coli* and *S. typhi*), we compared our results from data sets with and without *S. typhi* orthologous data. When *S. typhi* was excluded, we observed a decrease in the overall number of motif predictions that were made, as well as in the overall percent correspondence with known sites (Table 4, columns 2,3). Also, the stringency imposed by our critical MAP value calculations with correlated sequence data was shown by a significant increase in motif predictions that corresponded to known *E. coli* sites but were below the critical MAP value when *S. typhi* data were included (Table 4, column 6).

Of particular interest to our stated purpose were those motif predictions with an average MAP above the critical MAP value that do not correspond to known sites and therefore would be candidates for affinity chromatography to identify novel TFs. Results from phylogenetic footprinting with the set of seven species showed that among those motif predictions that included at least three species, there were 14 motif predictions with average MAPs above the critical MAP value that did not correspond to known TF binding sites (Table 4). Among these 14 motif predictions, chance false-positives were unlikely (less than one prediction), given our chosen level of significance ($P = 0.05$). Additionally, two of these 14 motifs correspond to site types that we were not scoring: one RNA stem-loop and one Ihf binding site. That the remaining 12 motif predictions were unlikely to be caused by chance and did not correspond to known sites indicates that they would be good candidates for synthesis of oligonucleotides for affinity chromatography. When these criteria were applied to the 2086 data sets representing the *E. coli* genome, 586 motif predictions were significant and 485 of these did not correspond to reported TF binding sites in *E. coli*.

Also of interest were the large number of motif predictions that corresponded with known sites that had an average

Table 4. Phylogenetic Footprinting Results for the Study Set Using Orthologous Data from Multiple Species

				Correspondence with known sites for:	
	Number of motif predictions	Correspondence with known sites	Motif predictions with avg MAP ≥ critical value	Motif predictions with avg MAP ≥ critical value	Motif predictions with avg MAP < critical value
EC-ST-YP-VC-HI-SO-PA:					
Predictions with ≥2 species	163	68.7% (112/163)	52.1% (85/163)	75.3% (64/85)	61.5% (48/78)
Predictions with ≥3 species	116	74.1% (86/116)	60.3% (70/116)	80.0% (56/70)	65.2% (30/46)
Predictions with ≥4 species	84	75.0% (63/84)	69.0% (58/84)	79.3% (46/58)	65.4% (17/26)
EC-YP-VC-HI-SO-PA:					
Predictions with ≥2 species	140	67.9% (95/140)	53.6% (75/140)	84.0% (63/75)	49.2% (32/65)
Predictions with ≥3 species	102	65.7% (67/102)	60.8% (62/102)	85.5% (53/62)	35.0% (14/40)
Predictions with ≥4 species	65	69.2% (45/65)	69.2% (45/65)	84.4% (38/45)	35.0% (7/20)

MAP indicates maximum a posteriori probability; EC, *Escherichia coli;* ST, *Salmonella typhi;* YP, *Yersinia pestis;* VC, *Vibrio cholerae;* HI, *Haemophilus influenzae;* SO, *Shewanella oneidensis;* PA, *Pseudomonas aeruginosa;* BU, *Buchnera* sp.; AF, *Acidithiobacillus ferrooxidans;* and XF, *Xylella fastidiosa.*

MAP value below the critical MAP value at $P = 0.05$, an indication of the stringency imposed by the critical MAP value cutoff. However, the choice of P value was subjective, and naturally, the number of motif predictions with an average MAP above the critical MAP value increased as the P value used for determining the critical MAP value was increased (Table 5). It is also worth noting that the numbers presented in Tables 4 and 5 reflect only the most probable motif prediction (the motif prediction with the highest average MAP value) for each orthologous data set, although multiple motif predictions from a data set may have average MAP values above the critical MAP value. For example, two motif predictions for the *nagB* data set had average MAP values above the critical MAP value; one corresponded to a known Crp site and the other corresponded to two known NagC sites. Adding up all of those motif predictions above the critical MAP value brought the total number of significant motif predictions to 105 for the study set (representing 85 genes) and 741 for the *E. coli* genome (representing 586 genes). Of these 741 motif predictions, 631 were unreported sites in *E. coli* and were therefore candidates for identification of novel TFs via affinity chromatography.

DISCUSSION

Species Selection for Phylogenetic Footprinting

Our results examining the question of which species are most useful for phylogenetic footprinting confirmed a number of natural assumptions. First, species that have few identifiable orthologs, especially TF orthologs, to the reference species are of relatively little value. Second, simple species characteristics can be used as predictors that this would be the case. Relative to our reference species *E. coli*, those species that have undergone extreme genome reduction (e.g., *Buchnera*), and species that are both distantly related phylogenetically and inhabit a different ecological niche (e.g., *A. ferrooxidans* and *X. fastidiosa*), are unlikely to encode many identifiable orthologs. Further, given their low number of orthologous TFs, these species are more likely to regulate orthologs differently than the reference species. In fact, the number of orthologs, and the frequency with which the data from these species were included in the motif predictions for the study set, overexaggerated the utility of these species at the genome level (data not shown). This was a result of the bias in the study set toward genes from common biochemical pathways (e.g., carbon and nitrogen metabolism, amino acid and nucleotide biosynthesis)—those functions that may be more likely to have common regulatory pathways among broad phylogenetic groups.

Our results from combinations of two, three, or four species indicate that a high degree of correspondence with known sites can be achieved with data from only three species, by adhering to the guidelines outlined above for the selection of species to use for phylogenetic footprinting. In fact, the results with four species showed only a modest improvement over the results with three species (Fig. 2C). These results indicate that our approach may be useful for studies of transcription regulation when applied to species in other branches of the prokaryotic phylogenetic tree for which few genomic sequences are available. Although good results are possible with only three or four species, there are good reasons for using as many species as possible. Given the diversity of prokaryotic species even within a narrow phylogenetic group such as the γ-proteobacteria, the orthologous gene sets among species are widely overlapping, and therefore, the number of genes from the reference species with at least three orthologs is maximized by the inclusion of as many species as possible. Also, there are common evolutionary phenomena that may result in a species contributing less to phylogenetic footprinting than might be expected, despite the presence of orthologous genes. For example, gene rearrangement within operons, and the splitting up of genes within an operon, can lead to a situation in which the promoter(s) are upstream of different genes in different species (for examples, see Gelfand 1999). In such a situation, the orthologous genes are present and may be regulated by a common mechanism; however, not all species will contribute to the motif prediction because the regulatory sequences are upstream of different genes in different species. Also, nonorthologous gene displacement, especially among TFs, can lead to disparate regulatory sequences among species, as was the case for *E. coli* and *P. aeruginosa* in this report.

Critical MAP Values

Ultimately, we want to prioritize our motif predictions for affinity purification experiments to identify the cognate TFs. Thus, to limit motif predictions that may result from chance alone, we devised a random simulation method to determine the critical MAP values for significance of our results. Throughout this report, we chose to present the results for $P = 0.05$, although this value is subjective. At this level of significance, we found that >75% of the significant motif predictions corresponded with known sites (see Table 4). We also observed that approximately two thirds of the motif predictions with average MAP values below the critical MAP values corresponded with reported sites. These findings reflect two effects: (1) The lower average MAP values indicate only that a motif prediction may be the result of chance, not that it is;

Table 5. Study Set Results at Different *P* Values[a]

Quantile	Number of motif predictions	Motif predictions with avg MAP ≥ critical value	Correspondence with known sites for: Motif predictions with avg MAP ≥ critical value	Correspondence with known sites for: Motif predictions with avg MAP < critical value
80% ($P = 0.20$)	163	129	93	19
90% ($P = 0.10$)	163	108	81	31
95% ($P = 0.05$)	163	85	64	48
99% ($P = 0.01$)	163	50	37	75

[a]Using orthologous data sets containing *Escherichia coli–Salmonella typhi–Yersinia pestis–Vibrio cholerae–Haemophilus influenzae–Shewanella oneidensis–Pseudomonas aeruginosa* and analyzing motif predictions that contain sites from at least two species.

and (2) many of the motif predictions that are below our chosen critical MAP value are at least moderately unlikely owing to chance alone ($0.05 < P < 0.2$). In Table 5 (column 3), for example, 44 motif predictions fall in this range.

Our results have also shown the importance of taxonomy to phylogenetic footprinting, in particular the level of conservation, or percent sequence identity, in the orthologous intergenic regions. Importantly, we observed a correlation in sequence conservation between some species that was likely caused by phylogenetic relatedness instead of function. This phenomenon was most dramatic with *S. typhi*, for which we observed a mean percent identity of 70% to the orthologous *E. coli* intergenic regions. Such a high degree of identity (on average) may be caused by several factors. *E. coli* and *S. typhi* are phylogenetically close, having diverged ~100 Mya (Parkhill et al. 2001). At this relatively close phylogenetic distance, we expect considerable sequence conservation, owing to an insufficient evolutionary time in which random mutations could accumulate. In addition, horizontal transfer between these species has likely been ongoing since speciation, leading to a high percent identity for those genes and upstream regions that have been recently transferred. Another possibility is that small intergenic regions containing densely packed regulatory regions may be expected to maintain relatively high sequence identity to retain regulatory function. However, for the study set genes, the logit of the percent identity of the *E. coli* and *S. typhi* orthologous intergenic sequences had only a weak negative correlation with sequence length (correlation coefficient = -0.16; $P = 0.042$). Also of interest were some intergenic regions with a significantly lower percent identity than the mean; this may be an indication that a paralog was identified in *S. typhi* instead of an orthologous gene, despite our seemingly rigorous orthology rules.

The fact that the *S. typhi* sequences were strongly correlated with the *E. coli* sequences, combined with the effect this correlation had on our critical MAP value calculations, prompted us to compare our phylogenetic footprinting results with and without inclusion of the *S. typhi* data. Overall, our results benefited when *S. typhi* was included; motif predictions were made for more genes, and therefore, more motif predictions had an average MAP value above the critical MAP value, despite the high threshold imposed by including *S. typhi*. The disadvantage of including *S. typhi*—more motif predictions that corresponded with known sites were below the critical MAP value—was a small price to pay, considering such benefits. Despite these overall benefits, those motif predictions involving only *E. coli* and *S. typhi* data (i.e., containing only correlated sequence data) should be regarded with caution, given that among these predictions, we observed only 55% correspondence with known sites, and only 38.5% had an average MAP value above the critical MAP value (see supplemental data online).

We chose to incorporate the information on sequence correlation into our simulations, adjusting for sequence identities above what was expected by chance, thereby adjusting the critical MAP values for statistical significance of our motif predictions. Another approach would be to incorporate a measure of phylogeny into the Gibbs sampling algorithm. This would effectively down-weight correlated sequence data, such as the *S. typhi* data in this report. Although progress along these lines has recently been reported (Holmes and Bruno 2001; Blanchette et al. 2002; Blanchette and Tompa 2002), the most suitable method by which to measure species phylogeny (i.e., genome phylogeny), as opposed to the phylogeny of a single gene, remains uncertain (Eisen 2000). In addition, given the current dogma of species evolution and the influence of horizontal transfer on bacterial evolution (Doolittle 1999), single measures of phylogenetic relatedness among the γ-proteobacteria would necessarily be estimates.

Conclusions

The identification of 741 statistically significant sites in the *E. coli* genome, along with our previous demonstration that cognate TFs can be identified based on these motif predictions (McCue et al. 2001), offers the opportunity to make substantial progress toward elucidating the transcription regulatory network of *E. coli*, as it could more than double the number of known binding sites with cognate TFs. However, high-throughput experimental approaches will be required for such an endeavor. The finding that only three or four related species may be required for credible motif predictions of TF binding sites indicates that this approach may be applicable to many more of the ~70 eubacterial species with genomes that have been sequenced. However, it is important to keep in mind that our findings are based exclusively on comparison with reported sites in *E. coli*; thus, application of this method to other species will require further study. Our results indicate that genome size, phylogenetic distance, and similarity of habitat are important factors in the selection of species for inclusion in such studies. Furthermore, the number of orthologous TFs identified in related species provides a reasonable guide for inclusion or exclusion of species. Unfortunately, because far fewer sites have been identified in any other species, prospective validation of most such motif predictions will be required. Finally, the many differences in eukaryotic transcription regulation, compared with that of prokaryotes, indicate that these findings should be considered with caution when similar approaches are considered for even unicellular eukaryotes.

METHODS

Identification and Analysis of Data Sets

We identified a study set of 166 genes in the *E. coli* genome for which 48 different TFs have been experimentally confirmed to activate or repress expression by sequence-specific binding to their cognate sites in the upstream intergenic region. Using stringent criteria to identify probable orthologous genes in nine additional γ-proteobacterial species (listed below; see Table 1), we extracted sets of probable orthologous upstream intergenic regions for the 166 genes by the method previously described (McCue et al. 2001). The species were chosen such that they were of varying evolutionary distance to *E. coli* (Fig. 1) and had complete, or nearly complete (in gap closure), genome sequence data available. From this master collection of sets of orthologous intergenic regions, we collected subsets that contained orthologous intergenic sequence data from all paired combinations of species and selected k-tuple ($k > 2$) combinations of species that included *E. coli*.

We applied the advanced Gibbs sampling algorithm previously described (McCue et al. 2001) to all of these subsets, with the following three modifications to our approach. (1) To better focus our study on identifying binding sites for specific activators and repressors, we did not include the following as "known" regulatory sites: sites for the nucleoid proteins Fis, Ihf, Lrp, DnaA, IciA, Hu, Hns, Hfq, CbpA, CbpB, Dps, and StpA (Azam and Ishihama 1999); sites for ArcA, which appears to form large oligomers on DNA (Jeon et al. 2001); or RNA stem-loop regulatory structures. (2) In our previous study, we showed the affinity purification of FabR (YijC) to predicted

sites upstream of *fabA*, *fabB*, and *yqfA* (McCue et al. 2001). We also observed that when the predicted *E. coli* sites overlapped known sites, the overlap was by ≥10 bp for the vast majority of predictions, although for 8.7%, the overlap was less than this. In this report, we wished to identify those motif predictions that had broad overlap with TF binding sites, such that oligonucleotides synthesized from the predicted sites plus 5- to 10-bp flanking would likely contain complete sites and would lead to successful affinity purification of TFs, in the manner that FabR was purified. Therefore, our criterion for correspondence of motif predictions with known sites required ≥10 bp of overlap. (3) To better estimate information content of the average site in a motif prediction, we normalized the MAP values of the Gibbs sampling motif predictions by calculating the average MAP, where average MAP = MAP/(number of predicted sites in a motif).

Random Data Simulations

For Gibbs sampling simulations using randomized sequence data, the *E. coli* sequence from each of the 166 data sets in the study set was randomized, with retention of the nucleotide composition. Data sets were assembled that contained a randomized *E. coli* sequence and additional k random sequences ($1 \leq k \leq 9$) of equal length. After global alignment (GAP program; GCG Wisconsin Package version 10.2), the percent identity between sequences in a randomized data set was ~40%. However, the mean percent identity between *E. coli* and *S. typhi* or *Y. pestis* in the real sequence data was significantly greater than 40% (Table 2). To address this correlation in sequence identity, we prepared additional data sets requiring that the percent identity between randomized *E. coli* and *S. typhi* sequences in each individual orthologous set remain approximately the same as that observed in the real data for that gene (therefore, the mean percent identity between *E. coli* and *S. typhi* in these randomized data remained ~70%). A similar procedure was applied for *Y. pestis* sequences. The same Gibbs sampling strategy was used as had been used with the genuine data, with 10 simulations performed on every data set. The maximum average MAP values from each simulation for each data set were combined to calculate 95% quantiles, which we refer to as critical MAP values.

Phylogenetic Tree

Aligned small subunit ribosomal RNA sequences for the 10 species were obtained from the rRNA WWW Server at the University of Antwerp (http://www-rrna.uia.ac.be/rrna/). These data were used to infer the phylogeny in Figure 1; specifically, Phylip (version 3.6) was used to calculate the DNA distances by maximum likelihood and to construct a tree by neighbor joining (J. Felsenstein. Department of Genetics, University of Washington, Seattle; http://evolution.genetics.washington.edu/phylip.html).

Genome Sequence Data

E. coli (U00096), *H. influenzae* (L42023), *Buchnera* sp. APS (AP000398), *P. aeruginosa* (AE004091), *S. enterica* serovar Typhi (AL513382), *V. cholerae* (AE003852 and AE003853), *X. fastidiosa* (AE003849), and *Yersinia pestis* (AL590842) genome sequence data were obtained from GenBank (ftp://ncbi.nlm.nih.gov/genbank/genomes/Bacteria/). Preliminary genome sequence data for *S. oneidensis* and *A. ferrooxidans* were obtained from the Institute for Genomic Research (http://www.tigr.org/).

ACKNOWLEDGMENTS

We thank the Computational Molecular Biology and Statistics Core Facility at the Wadsworth Center for assistance, Clarence Chan for Web site development, and the Institute for Genomic Research for making genome sequence data available before completion. This work was supported by National Institutes of Health grants RO1HG01257 and R21RR14036 to C.E.L. and by Department of Energy grant DE-FG02-01ER63204 to L.A.M. and C.E.L.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Azam, T.A. and Ishihama, A. 1999. Twelve species of the nucleoid-associated protein from *Escherichia coli*: Sequence recognition specificity and DNA binding affinity. *J. Biol. Chem.* **274:** 33105–33113.

Behr, M.A., Wilson, M.A., Gill, W.P., Salamon, H., Schoolnik, G.K., Rane, S., and Small, P.M. 1999. Comparative genomics of BCG vaccines by whole-genome DNA microarray. *Science* **284:** 1520–1523.

Blanchette, M. and Tompa, M. 2002. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.* **12:** 739–748.

Blanchette, M., Schwikowski, B., and Tompa, M. 2002. Algorithms for phylogenetic footprinting. *J. Comput. Biol.* **9:** 211–223.

Cliften, P.F., Hillier, L.W., Fulton, L., Graves, T., Miner, T., Gish, W.R., Waterston, R.H., and Johnston, M. 2001. Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.* **11:** 1175–1186.

Doolittle, W.F. 1999. Phylogenetic classification and the universal tree. *Science* **284:** 2124–2129.

Eisen, J.A. 2000. Assessing evolutionary relationships among microbes from whole-genome analysis. *Curr. Opin. Microbiol.* **3:** 475–480.

Fraser, C. 2000. DOE-funded microbial genome sequencing at the Institute for Genomic Research. Human Genome Program, U.S. Department of Energy, Microbial Genome Program Report. http://www.ornl.gov/hgmis/publicat/microbial/index.html.

Galperin, M.Y. and Koonin, E.V. 2000. Who's your neighbor?: New computational approaches for functional genomics. *Nat. Biotechnol.* **18:** 609–613.

Gelfand, M.S. 1999. Recognition of regulatory sites by genomic comparison. *Res. Microbiol.* **150:** 755–771.

Gelfand, M.S., Koonin, E.V., and Mironov, A.A. 2000. Prediction of transcription regulatory sites in archaea by a comparative genomic approach. *Nucleic Acids Res.* **28:** 695–705.

Hardison, R.C., Oeltjen, J., and Miller, W. 1997. Long human-mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. *Genome Res.* **7:** 959–966.

Heidelberg, J.F., Eisen, J.A., Nelson, W.C., Clayton, R.A., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Umayam, L., et al. 2000. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* **406:** 477–483.

Holmes, I. and Bruno, W.J. 2001. Evolutionary HMMs: A Bayesian approach to multiple alignment. *Bioinformatics* **17:** 803–820.

Huynen, M.A. and Bork, P. 1998. Measuring genome evolution. *Proc. Natl. Acad. Sci.* **95:** 5849–5856.

Jeon, Y., Lee, Y.S., Han, J.S., Kim, J.B., and Hwang, D.S. 2001. 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.

Laikova, O.N., Mironov, A.A., and Gelfand, M.S. 2001. Computational analysis of the transcriptional regulation of pentose utilization systems in the γ subdivision of proteobacteria. *FEMS Microbiol. Lett.* **205:** 315–322.

Makarova, K.S., Mironov, A.A., and Gelfand, M.S. 2001. Conservation of the binding site for the arginine repressor in all bacterial lineages. *Genome Biol.* **2:** RESEARCH0013.

McCue, L., Thompson, W., Carmack, C., Ryan, M.P., Liu, J.S., Derbyshire, V., and Lawrence, C.E. 2001. Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.* **29:** 774–782.

McGuire, A.M., Hughes, J.D., and Church, G.M. 2000. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.* **10:** 744–757.

Miller, W. 2000. So many genomes, so little time. *Nat. Biotechnol.* **18:** 148–149.

Mironov, A.A., Koonin, E.V., Roytberg, M.A., and Gelfand, M.S.

1999. Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucleic Acids Res.* **27:** 2981–2989.
O'Brien, S.J., Menotti-Raymond, M., Murphy, W.J., Nash, W.G., Wienberg, J., Stanyon, R., Copeland, N.G., Jenkins, N.A., Womack, J.E., and Marshall Graves, J.A. 1999. The promise of comparative genomics in mammals. *Science* **286:** 458–462, 479–481.
O'Brien, S.J., Eizirik, E., and Murphy, W.J. 2001. Genomics. On choosing mammalian genomes for sequencing. *Science* **292:** 2264–2266.
Panina, E.M., Mironov, A.A., and Gelfand, M.S. 2001a. Comparative analysis of FUR regulons in γ-proteobacteria. *Nucleic Acids Res.* **29:** 5195–5206.
Panina, E.M., Vitreschak, A.G., Mironov, A.A., and Gelfand, M.S. 2001b. Regulation of aromatic amino acid biosynthesis in γ-proteobacteria. *J. Mol. Microbiol. Biotechnol.* **3:** 529–543.
Park, S.M., Lu, C.D., and Abdelal, A.T. 1997a. Cloning and characterization of *argR*, a gene that participates in regulation of arginine biosynthesis and catabolism in *Pseudomonas aeruginosa* PAO1. *J. Bacteriol.* **179:** 5300–5308.
———. 1997b. Purification and characterization of an arginine regulatory protein, ArgR, from *Pseudomonas aeruginosa* and its interactions with the control regions for the *car*, *argF*, and *aru* operons. *J. Bacteriol.* **179:** 5309–5317.
Parkhill, J., Dougan, G., James, K.D., Thomson, N.R., Pickard, D., Wain, J., Churcher, C., Mungall, K.L., Bentley, S.D., Holden, M.T., et al. 2001. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. *Nature* **413:** 848–852.
Pizza, M., Scarlato, V., Masignani, V., Giuliani, M.M., Arico, B., Comanducci, M., Jennings, G.T., Baldi, L., Bartolini, E., Capecchi, B., et al. 2000. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* **287:** 1816–1820.
Porcella, S.F. and Schwan, T.G. 2001. *Borrelia burgdorferi* and *Treponema pallidum*: A comparison of functional genomics, environmental adaptations, and pathogenic mechanisms. *J. Clin. Invest.* **107:** 651–656.
Rajewsky, N., Socci, N.D., Zapotocky, M., and Siggia, E.D. 2002. The evolution of DNA regulatory regions for proteo-γ bacteria by interspecies comparisons. *Genome Res.* **12:** 298–308.
Rodionov, D.A., Mironov, A.A., Rakhmaninova, A.B., and Gelfand, M.S. 2000. Transcriptional regulation of transport and utilization systems for hexuronides, hexuronates and hexonates in γ purple bacteria. *Mol. Microbiol.* **38:** 673–683.
Rodionov, D.A., Gelfand, M.S., Mironov, A.A., and Rakhmaninova, A.B. 2001a. Comparative approach to analysis of regulation in complete genomes: multidrug resistance systems in γ-proteobacteria. *J. Mol. Microbiol. Biotechnol.* **3:** 319–324.
Rodionov, D.A., Mironov, A.A., and Gelfand, M.S. 2001b. Transcriptional regulation of pentose utilisation systems in the *Bacillus/Clostridium* group of bacteria. *FEMS Microbiol. Lett.* **205:** 305–314.
Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., and Ishikawa, H. 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* **407:** 81–86.
Snel, B., Bork, P., and Huynen, M.A. 1999. Genome phylogeny based on gene content. *Nat. Genet.* **21:** 108–110.
Tamames, J. 2001. Evolution of gene order conservation in prokaryotes. *Genome Biol.* **2:** RESEARCH0020.
Tan, K., Moreno-Hagelsieb, G., Collado-Vides, J., and Stormo, G.D. 2001. A comparative genomics approach to prediction of new members of regulons. *Genome Res.* **11:** 566–584.
Terai, G., Takagi, T., and Nakai, K. 2001. Prediction of co-regulated genes in *Bacillus subtilis* on the basis of upstream elements conserved across three closely related species. *Genome Biol.* **2:** RESEARCH0048.1.
Zhang, Y.M., Marrakchi, H., and Rock, C.O. 2002. The FabR (YijC) transcription factor regulates unsaturated fatty acid biosynthesis in *Escherichia coli*. *J. Biol. Chem.* **277:** 15558–15565.

WEB SITE REFERENCES

http://www.wadsworth.org/resnres/bioinfo/; Wadsworth Bioinformatics Web server.
http://www-rrna.uia.ac.be/rrna/; rRNA WWW Server at the University of Antwerp.
http://evolution.genetics.washington.edu/phylip.html; PHYLIP home page.
http://www.tigr.org/; The Institute for Genome Research.
ftp://ncbi.nlm.nih.gov/genbank/genomes/Bacteria/; GenBank.
http://www.ornl.gov/hgmis/publicat/microbial/index.html; United States Department of Energy.

Received March 29, 2002; accepted in revised form July 30, 2002.