APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Feb. 2007, p. 846–854 Vol. 73, No. 3
0099-2240/07/$08.00+0 doi:10.1128/AEM.01686-06

Operon Prediction for Sequenced Bacterial Genomes without Experimental Information▽†

Nicholas H. Bergman,[1,2]* Karla D. Passalacqua,[2] Philip C. Hanna,[2] and Zhaohui S. Qin[3]*

Bioinformatics Program[1] and Department of Microbiology & Immunology,[2] University of Michigan Medical School, Ann Arbor, Michigan 48109, and Center for Statistical Genetics and Department of Biostatistics, University of Michigan School of Public Health, Ann Arbor, Michigan 48109[3]

Received 19 July 2006/Accepted 4 November 2006

Various computational approaches have been proposed for operon prediction, but most algorithms rely on experimental or functional data that are only available for a small subset of sequenced genomes. In this study, we explored the possibility of using phylogenetic information to aid in operon prediction, and we constructed a Bayesian hidden Markov model that incorporates comparative genomic data with traditional predictors, such as intergenic distances. The prediction algorithm performs as well as the best previously reported method, with several significant advantages. It uses fewer data sources and so it is easier to implement, and the method is more broadly applicable than previous methods—it can be applied to essentially every gene in any sequenced bacterial genome. Furthermore, we show that near-optimal performance is easily reached with a generic set of comparative genomes and does not depend on a specific relationship between the subject genome and the comparative set. We applied the algorithm to the *Bacillus anthracis* genome and found that it successfully predicted all previously verified *B. anthracis* operons. To further test its performance, we chose a predicted operon (BA1489-92) containing several genes with little apparent functional relatedness and tested their cotranscriptional nature. Experimental evidence shows that these genes are cotranscribed, and the data have interesting implications for *B. anthracis* biology. Overall, our findings show that this algorithm is capable of highly sensitive and accurate operon prediction in a wide range of bacterial genomes and that these predictions can lead to the rapid discovery of new functional relationships among genes.

A large number of bacterial genomes have been sequenced in the past decade, and with large-scale sequencing technologies becoming cheaper and more accessible, it is reasonable to expect that the rate at which new genomes are completed will continue to accelerate. Accurate tools for identifying the genes within a given genome have been developed (8, 29), but our understanding of how these genes are expressed and regulated depends also on knowledge of how they are organized into operons—sets of genes that are cotranscribed into a single mRNA sequence. Operons form the fundamental transcriptional units within a bacterial genome, and as a result, defining these structures is a key first step in examining transcriptional regulation. In addition, operons often contain genes that are functionally related and required by the cell for a certain process or pathway and, thus, they are highly predictive of biological networks. For these reasons, the definition of operon structure on a genome-wide level is an important starting point for microbial functional genomics.

Direct experimental approaches to operon identification, such as Northern blotting or primer extension, are usually costly and time-consuming, and so there is considerable interest in the development of computer algorithms that will accurately predict genome-wide operon structure. Given the rapid pace at which bacterial genomes are now being sequenced, there is a particular need for methods that are generally applicable across the bacterial domain. This requirement for "portability" places limits on the information that can be used in such algorithms and necessarily excludes experimental and detailed functional data, which are only available for a small subset of sequenced genomes (and often for only a subset of the genes within these genomes). A truly portable operon prediction algorithm must essentially rely on data inherent in the genome itself: the identity, spacing, and orientation of genes, as well as the sequence.

There has been a variety of prediction algorithms developed in recent years, including those that take advantage of experimental or functional data as well as examples that are more generalized and only require sequence information. Examples of the former include methods that rely on microarray-based expression data (3, 4, 7, 26) and others that use different forms of detailed functional annotation (5, 30, 35). Although these algorithms have shown great promise in terms of being able to predict operon structure with a high degree of specificity and sensitivity, the data they rely on are only available for a select subset of bacterial species, and this limits how widely they can be used.

Progress has also been made toward a more generalized method for operon prediction, and a number of groups have constructed algorithms based on a variety of diverse information sources, including codon usage statistics (3, 4) and the identification of promoter and terminator sequences (6, 28, 33, 34). Although these data have all proven to be valid predictors

* Corresponding author. Mailing address: University of Michigan Medical School, Bioinformatics Program and Department of Microbiology & Immunology, 6605H Medical Sciences Bldg. II, 1150 W. Medical Center Dr., Ann Arbor, MI 48109-0620. Phone for N. Bergman: (734) 615-2154. Fax: (734) 764-3562. E-mail: niber@umich.edu. Phone for Z. Qin: (734) 763-5965. Fax: (734) 763-2215. E-mail: qin@umich.edu.

† Supplemental material for this article may be found at http://aem.asm.org/.

▽ Published ahead of print on 22 November 2006.

of operon structure, it is striking that these studies have also consistently demonstrated that one of the most valuable predictors is simply intergenic distance. The distances between genes within an operon tend to be considerably shorter than the distances between genes that are not cotranscribed, and in several recently developed algorithms, intergenic distance was shown to be more informative than any other data source, including even microarray-based expression data (3, 7, 28, 34). In addition, this trend appears to be universal in bacterial genomes, making it a very attractive option for a generalized, portable prediction algorithm (7, 15, 33). Unfortunately, intergenic distance alone only allows for a specificity of ~65 to 70% when tested on a large set of experimentally verified operons from within the *Escherichia coli* genome, and so other sources of information must be added to bring the total accuracy to a more acceptable level (28).

Another promising generalized predictor of operon structure is the degree to which gene order is conserved across a variety of genomes, with the general idea that adjacent genes that are found in the same order in multiple genomes are more likely to be cotranscribed. This method has previously been used as a means of assessing functional relatedness among proteins (17), and several studies have shown that operons in a given bacterial genome could be identified with a very high degree of specificity using this approach (98%, as reported by Ermolaeva et al. [9]). The drawback, however, is that its accuracy is derived from genes being conserved in a relatively large number of species, and the method tends to miss operons containing genes that are unique or less conserved. In addition, a study by Itoh et al. showed that during evolution many operons undergo shuffling events that change the order of genes within (but not their overall content), and such operons are missed by an algorithm that requires conservation of order to make predictions (11). The result is that despite the high specificity of this algorithm, it is inherently insensitive and can only be applied to 30 to 50% of the genome being examined (9).

The extremely high specificity achieved using conserved gene pair information underscores the utility of these data in operon prediction, and we hypothesized that it might be possible to exploit phylogenetic information in a more general way, such that it could be applied to the entire genome. If so, we reasoned that these data, when combined with intergenic distance information in a rigorous statistical model, would allow for highly specific and sensitive operon prediction in any sequenced bacterial genome. With this in mind, we developed a generalized method for using phylogenetic data to predict operon structure. We adopted a Bayesian approach in constructing a hidden Markov model (HMM)-based algorithm that combines these data with intergenic distance statistics. Using a large set of experimentally verified operons from *E. coli*, we found that an optimized version of the algorithm predicted operon structure with specificity and sensitivity levels of >85%. We have also shown that the method can be generalized easily and applied to essentially any bacterial genome, regardless of the availability of any experimental or functional data.

We applied the algorithm in predicting the operon structure within the genome of *Bacillus anthracis* and successfully predicted all previously known *B. anthracis* operons. In addition, we identified a large number of putative operons that link apparently unrelated genes in cotranscriptional relationships, and we chose a particularly interesting example (BA1489-92) for further testing. This putative operon contains four genes that have little in common functionally, and they have not been predicted to be cotranscribed by any previously developed algorithm (23). Reverse transcription-PCR (RT-PCR) experiments confirmed that these genes are in fact cotranscribed, and targeted gene disruption data obtained in a related study (18) were also consistent with this finding. These results suggest a new functional link between the genes within this operon and have interesting implications for *B. anthracis* biology. In addition, they suggest that many other important functional and regulatory relationships may be identified in the same way and that the algorithm developed in this study may be a significant new tool for the field of bacterial genomics.

MATERIALS AND METHODS

Data sources. For a genome in which operon structure is being predicted (a subject genome), a tab-delimited *.ptt (Protein Table) file containing the location, length, orientation, and name of each gene was downloaded from the NCBI GenBank and used as a master reference file. For each comparative genome used, a multisequence FASTA file containing the complete set of protein sequences from that genome was obtained from the Comprehensive Microbial Resource at The Institute for Genomic Research (http://www.tigr.org/CMR) and used to build BLASTP databases and query files (21).

The set of 359 experimentally verified *E. coli* transcriptional units (257 operons and 102 singly transcribed genes) used in testing and developing the algorithm was obtained from the supplemental information provided by Sabatti et al. in their recent study (26). The list was originally compiled as part of the RegulonDB, a database of transcriptional regulation and organization for *E. coli* K-12, and further information regarding the experimental verification of the transcriptional status of each of these operons may be found there (http://www.cifn.unam.mx/Computational_Genomics/regulondb/) (27). The genes included in this set were mapped onto the *E. coli* reference file using scripts within MS Excel.

Phylogenetic distribution analyses. The peptide sequence files obtained for each genome were used to construct local BLASTP databases, and the complete set of peptide sequences from the subject genome was compared to each database using a locally installed copy of the NCBI BLAST search tool (software obtained from the NCBI website [ftp://ftp.ncbi.nlm.nih.gov/BLAST/]) run with default parameters (2). The results from these searches were parsed using a Perl script that identified potential orthologs as the best hit within a given comparative genome for each peptide sequence within the subject genome, with the additional provision that the expect value for potential orthologs was required to be less than our defined cutoff (10^{-4}). This method for identifying orthologs is intentionally more promiscuous than the commonly used "reciprocal best match" method. Our goal in this study was to construct an algorithm that was capable of predicting operon structure across an entire genome, and use of the reciprocal best match method was problematic because of the way this technique deals with paralogs within the subject genome. Briefly, if there are a number of paralogs within the subject genome that are all homologous to a single gene in the reference genome, the reciprocal best match method will only identify the most related of these as having an ortholog within the reference genome, even though all of them are actually related to that ortholog. Essentially, this means that the reciprocal best match method results in a number of paralogous genes within the subject genome for which we have inaccurate phylogenetic data, and the simpler "best match" method that we use here avoids this problem by considering each gene independently.

A binary vector, with each dimension corresponding to a given comparative genome and denoted by a 1 if an ortholog is present and a 0 if there is no ortholog within that genome. These vectors have a passing similarity to commercial barcodes, and we refer to them as phylogenetic barcodes. They can be written as follows: $(a_i^1, a_i^2, \ldots, a_i^m)$, where $a_i^j = 0$ or 1 and $j = 1, 2, \ldots, m$ and where each binary code a_i^j indicates whether an ortholog for gene i can be found in the jth related species. The final lists of potential orthologs from each comparative genome were combined into a single file from which the list of phylogenetic barcodes corresponding to each gene in the subject genome was compiled. To test the difference between the barcodes of two adjacent genes, and thus assess the difference in their phylogenetic distributions, we

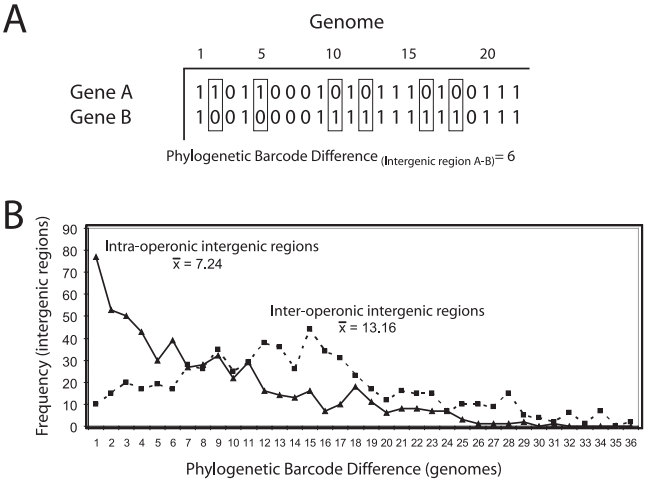

FIG. 1. Phylogenetic barcode differences in intra- and interoperonic pairs. (A) Illustration of the method for calculating the change in phylogenetic distribution (difference in phylogenetic barcode) for a given intergenic region. In the case shown, the phylogenetic barcode difference for intergenic region A-B is 6. (B) Phylogenetic barcode differences in experimentally verified transcriptional units. Data for both intraoperonic intergenic regions (squares connected by solid lines) and interoperonic intergenic regions (triangles connected by dashed lines) are shown.

compared the two vectors by counting how many differences exist between them. This was calculated as follows:

$$\Delta_i = \sum_{j=1}^{m} |a^j_{i+1} - a^j_i|$$

and is diagrammed schematically in Fig. 1A. All barcode compilations, as well as the calculations of barcode differences, were done using MS Excel.

General HMM framework. A given bacterial genome can be viewed as a series of adjacent gene pairs where the two genes in each pair are either within the same operon or transcribed separately. (Note that for the purposes of this study, two genes that are expressed as part of the same transcript under any condition are considered to be part of the same operon.) This scenario fits in the framework of the HMM. Whether a gene pair belongs to the same operon is the hidden state and the string of states assumed to follow a Markov chain.

Two sources of information are considered in this study: the distance between adjacent genes and the difference in their phylogenetic distribution. Both can be derived from the genome sequence alone, provided that the physical location of each gene is known. In this study, we adopted an HMM framework to accommodate all the information. Previous experience suggested that gamma distributions would fit the inter- and intraoperonic intergenic distances well. We modeled these two populations using two distinct gamma distributions and defined the transition probabilities from one state to the other as a function of the intergenic distance between each adjacent gene pair. In general, a small intergenic distance suggests that the two genes belong to the same operon, and a larger intergenic distance favors the opposite possibility. We also treated the phylogenic conservation barcode defined above as observed data generated from the hidden states. The differences in the barcodes between two adjacent genes are assumed to follow one of two binomial distributions, depending on whether or not they belong to the same operon. Due to the lack of good-quality training data, we adopted a Bayesian HMM scheme as described by Liu (12), where distribution parameters are used to calculate emission probabilities and path are inferred from the data, and two empirical distributions were used to calculate the distance-dependent transition probabilities. We applied the Gibbs sampler technique to iteratively sample from the conditional distributions of these unknown quantities. A detailed description of these methods can be found in the supplemental material.

For comparison purposes, we also constructed two homogeneous HMMs which model intergenic distances or phylogenic barcode differences exclusively. In these studies, all distribution parameters, path, and transition probabilities are inferred from the data. No additional training data are required. The HMM using intergenic distance alone performed better than the HMM using just the phylogenic barcode difference data and, not surprisingly, both were inferior relative to the aforementioned inhomogeneous HMM that combines the two sources of information. Detailed descriptions of the implementation of these HMMs can also be found in the supplemental material.

Algorithm testing and scoring of predictions. Predicted operons in the *E. coli* K-12 genome were scored relative to the set of known *E. coli* transcriptional units described in the Results section by using MS Excel. Statistical testing, including receiver operator characteristic (ROC) curve analysis, was done within Excel using the Analyze-It general statistics and clinical laboratory modules (Analyze-It Software, Ltd., Leeds, England).

Software availability. Software implementing the HMM-based prediction algorithm is available from our group upon request, as are the Perl scripts used to parse the BLASTP results. Potential users should note that the software tools developed for this study are generally quite fast (requiring less than 1 min on a typical desktop PC), and even the slowest step (the BLASTP comparisons) can be performed in less than an hour.

Cell growth conditions and RNA isolation. Brain heart infusion broth cultures of *Bacillus anthracis* strain Sterne (34F2) were grown overnight and then diluted 1:1,000 into nutrient-limiting (sporulation) modified G medium. At an optical density at 600 nm of 1.0, 5 ml of culture was collected and bacteria were pelleted by centrifugation. RNA isolation was performed using the Ambion RiboPure-Bacteria kit per the manufacturer's instructions with the following modifications: cell disruption with zirconia beads was done for 15 min, 450 μl of RNAwiz was used, bromochloropropane was used in place of chloroform, and 50 μl of RNase/DNase-free distilled water was added during extraction. A QIAGEN RNeasy mini kit with a DNase digestion step was used per the manufacturer's RNA cleanup protocol. RNA was quantitated via the A_{260}/A_{280} ratio on a Beckman DU530 spectrophotometer. One microgram of RNA was run on a denaturing formaldehyde gel to verify purity. The above procedures were carried out in three separate experiments utilizing two unique cultures each time.

RT-PCR. In three separate experiments, 500 ng to 1.0 μg of RNA was used to perform endpoint RT-PCR using the Invitrogen one-step RT-PCR with platinum *Taq* per the manufacturer's instructions. Briefly, reverse transcription was performed at 50°C for 30 min. PCR was performed with 0.25 pg of operon/gene-specific primers for 35 or 37 cycles with an elongation temperature of 70°C and extension time of 1 min 10 s. Five μl of endpoint PCR product was then run in 0.7% agarose gels and visualized with ethidium bromide. Negative controls omitting reverse transcriptase and positive controls with *B. anthracis* Sterne (34F2) genomic DNA were done with each experiment. Operon/gene-specific primers were designed to result in 0.6- to 1.0-kb products.

RESULTS AND DISCUSSION

General method for utilizing phylogenetic information in operon prediction. Several recent studies have shown that genes within bacterial operons tend to be related functionally and are often involved in the same pathway or process within the cell (10, 16, 17, 31, 35). Additionally, it has been demonstrated that functionally related genes tend to share a similar phylogenetic distribution; that is, they tend to be coinherited and to travel together along the phylogenetic tree (13, 14, 20). Given these facts, we reasoned that genes within an operon would be more likely to have similar phylogenetic distributions than genes that are not cotranscribed and that it may be possible to use phylogenetic distribution information to predict operon structure.

In order to test this possibility, we compared the *E. coli* K-12 genome to 35 other bacterial genomes (chosen arbitrarily as a diverse set of species, including both distant and close relatives of *E. coli*) (Table 1) and searched for the possible presence of orthologs to each gene from the K-12 genome in each of the other genomes. The phylogenetic distribution of each *E. coli* gene was then compiled from these searches and represented as a 35-digit phylogenetic barcode. We hypothesized that genes within an operon would be more likely to have similar phylo-

TABLE 1. Genomes used in the comparative analysis[a]

Phylum	Class	Order	Species
Actinobacteria	*Actinobacteridae*	*Actinomycetales*	*Mycobacterium tuberculosis* CDC1551
Chlamydiae	*Chlamydiales*	*Chlamydiaceae*	*Chlamydia pneumoniae* AR39
Firmicutes	*Bacillales*	*Bacillaceae*	*Bacillus subtilis* 168
			Bacillus anthracis Ames
			Bacillus cereus ATCC 14579
		Listeriaceae	*Listeria monocytogenes* 4b F2365
		Staphylococcus	*Staphylococcus aureus* MW2
	Clostridia	*Clostridiales*	*Clostridium perfringens* 13
			Clostridium tetani E88
	Lactobacillales	*Enterococcaceae*	*Enterococcus faecalis* V583
		Streptococcaceae	*Streptococcus pneumoniae* TIGR4
			Streptococcus pyogenes SF370 serotype M1
Proteobacteria	*Alphaproteobacteria*	*Rhizobiales*	*Brucella suis* 1330
	Betaproteobacteria	*Burkholderiales*	*Bordetella pertussis* Tohama I
		Neisseriales	*Neisseria meningitidis* MC58
	Gammaproteobacteria	*Alteromonadaceae*	*Shewanella oneidensis* MR-1
		Enterobacteriales	*Buchnera aphidicola* (*Baizongia pistaciae*)
			Escherichia coli O157:H7 EDL933
			Escherichia coli O157:H7 VT2-Sakai
			Escherichia coli CFT073
			Salmonella enterica serovar Typhimurium LT2 SGSC1412
			Salmonella enterica serovar Typhi Ty2
			Shigella flexneri 2a 2457T
			Yersinia pestis KIM
		Legionellaceae group	*Coxiella burnetii* RSA 493
		Pasteurellaceae	*Pasteurella multocida* M70
		Pasteurellales	*Haemophilus influenzae* KW20 Rd
		Pseudomonadales	*Pseudomonas aeruginosa* AO1
			Pseudomonas putida KT2440
		Vibrionales	*Vibrio cholerae* El Tor N16961
		Xanthomonadales	*Xylella fastidiosa* 9a5c
		Xanthomonas group	*Xanthomonas campestris* pv campestris ATCC 33913
	Epsilonproteobacteria	*Campylobacterales*	*Campylobacter jejuni* NCTC 11168
			Helicobacter pylori 26695
Spirochaetes	*Spirochaetales*	*Spirochaetaceae*	*Treponema pallidum* Nichols

[a] Comparative genomes used for phylogenetic distribution calculations are shown. Additional information related to the background, source, and phylogenetic classification of each can be found at The Institute for Genomic Research Comprehensive Microbial Resource (http://www.tigr.org/CMR).

genetic distributions and, therefore, that operon boundaries might be identifiable as comparatively large changes in barcode structure between two adjacent, codirectional genes (note that throughout this study, only codirectional gene pairs are examined). With this in mind, we calculated the difference between the barcodes of each adjacent gene pair in the *E. coli* K-12 genome by comparing each pair of vectors and counting how many differences exist between them (Fig. 1A). We then used a large set of experimentally verified *E. coli* transcriptional units (257 operons and 102 singly transcribed genes) to directly test our hypothesis. The 929 genes in this set provide us with 580 verified intraoperonic gene pairs and 626 verified interoperonic gene pairs, and the probability distribution of barcode differences for each of these two populations is shown in Fig. 1B. We found that the differences observed within known operons and those observed at operonic boundaries form significantly different populations, with intraoperon gene pairs generally having a smaller phylogenetic barcode difference than interoperon gene pairs. These results suggested that the phylogenetic barcode data might be a valuable predictor of operon structure throughout the entire genome.

Construction and testing of a hidden Markov model-based algorithm for predicting operon structure. Given the apparent utility of the phylogenetic barcode information in operon prediction, we next sought to develop a statistical framework in which the predictive value of the data could be tested and into which other information sources could also be added. We and others have observed that the operon prediction problem fits well into an HMM framework (4, 34), especially because an HMM framework allows us to estimate the confidence with which each prediction is made (as opposed to rule-based prediction frameworks, which generally do not). Accordingly, we adopted this approach in formulating our algorithm.

The HMM-based algorithm was constructed as described in Materials and Methods and applied to the *E. coli* K-12 genome. We found that, using only the phylogenetic barcode information derived from the 35 comparative genomes and scoring our predictions using the set of known transcriptional units, the algorithm was able to predict the operon status of adjacent gene pairs with 89.6% sensitivity [(true positives)/(true positives + false negatives)] and 61.6% specificity [(true negatives)/(true negatives + false positives)] when using a predicted probability cutoff value of 0.5. Because a full spectrum of prediction probabilities is possible, the performance of the algorithm is more completely described by an ROC curve. In this method of analysis, specificity and sensitivity are plotted for every possible prediction probability cutoff value, and the

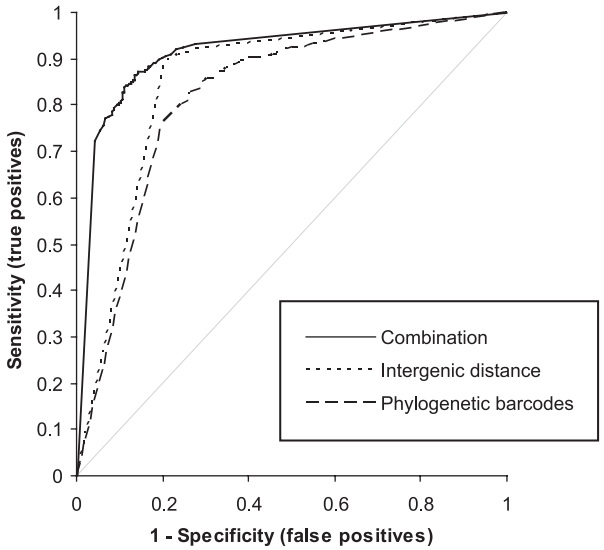

FIG. 2. ROC curves representing algorithm performance using different data sources. Dashed line, predictions generated using the phylogenetic barcode information; dotted line, predictions generated using the intergenic data alone; solid line, predictions generated using both sources.

area under the curve provides a combined measure of algorithm performance (with a maximal value of 1.00). The ROC curve corresponding to predictions generated using the phylogenetic barcode information is shown in Fig. 2 and the area under the curve is 0.819, indicating that this data source has substantial predictive value. Intergenic distance, when used alone in the HMM, yielded predictions that were somewhat different. At a probability cutoff of 0.5, this data source made the algorithm slightly more sensitive (91.1%) and also more specific (75.0%) than when the phylogenetic data were used alone (Fig. 2), and the area under the curve was slightly higher (0.852). When we combined the two sources in an inhomogeneous HMM, with transition probabilities estimated from intergenic distances, we found that a 0.5 cutoff value yielded predictions with a sensitivity of 87.5% and a specificity of 86.4% (Fig. 2) and an area under the ROC curve of 0.907. We also found that HMM helps improve the performance of the algorithm, as the Markov property captures the "clustering" characteristic of the operons. If the dependency among adjacent states were ignored (that is, if we fixed the transition probabilities at 0.5 in the HMM) and a naïve Bayes approach was used to combine the two sources of information, the area was decreased to 0.857. These data are summarized in Table 2, and they indicate that as we had hypothesized, the algorithm performs best when the two data sources are used in combination within the framework of the inhomogeneous HMM.

Testing possible refinements of the prediction algorithm. The performance of the algorithm in our initial experiments encouraged us to explore different ways in which it might be enhanced, especially as several aspects of how the phylogenetic data were compiled were set somewhat arbitrarily at the outset of the study. Perhaps the most obvious of these was the general composition of the set of comparative genomes; that is, the total number of genomes and their overall phylogenetic distribution. We initially explored this issue by increasing the number of comparative genomes to either 60 genomes that were widely distributed throughout the bacterial domain (see Table S1 in the supplemental material) or 49 genomes that were all from species of *Proteobacteria* and thus closely related to *E. coli* (see Table S2 in the supplemental material). In both cases, the performance of the algorithm was slightly worse as measured by area under the ROC curve (see Fig. S1A and summary in Table S3 in the supplemental material). In contrast, when we separated the original 35 genomes into two groups—22 species of *Proteobacteria* and 13 species which are more widely dispersed throughout the bacterial domain—we found that using either subset resulted in algorithm performance that was slightly enhanced relative to when the parent set of comparative genomes was used (see Fig. S1B in the supplemental material), suggesting that a large number of comparative genomes might be unnecessary for optimal algorithm performance. A possible explanation may be that a relatively small set of genomes provides all the necessary phylogenetic information and that additional comparative genomes merely duplicate this information (perhaps counterproductively).

These results also seemed to imply that the algorithm might be somewhat insensitive to the phylogenetic distribution of comparative genomes (relative to either the subject genome or to each other), a trait that would seem to be highly desirable in a method that is designed to be applicable to any bacterial genome, including those for which there are no sequenced close relatives. To further test this possibility, we examined algorithm performance under conditions in which we systematically varied either the relatedness of the comparative set to the subject genome (*E. coli*) or the diversity inherent within the comparative set itself. Consistent with our earlier results, we found that in both cases these changes had negligible effects on overall algorithm performance (see Fig. S1C and D in the

TABLE 2. Algorithm performance with different data sources[a]

Data source (area)	Probability cutoff value	Specificity (%)	Sensitivity (%)
A. Phylogenetic barcodes alone (0.819)	0.75	65.0	88.2
	0.50	61.6	89.6
	0.25	55.0	90.6
B. Intergenic distance alone (0.852)	0.75	73.1	91.7
	0.50	75.0	91.1
	0.25	77.8	90.6
C. Combination (0.907)	0.75	89.0	85.6
	0.50	86.4	87.5
	0.25	83.3	88.7
D. Combination (dependency ignored) (0.857)	0.75	78.1	90.7
	0.50	77.6	90.7
	0.25	76.7	91.1
E. Optimized combination (0.916)	0.75	89.3	83.4
	0.50	86.9	85.5
	0.25	85.2	87.2

[a] A summary of prediction algorithm performance using different data sources is shown. Specificity (the fraction of positive predictions that are correct) and sensitivity (the fraction of true operons recognized) levels are shown for different prediction probability cutoff values. Also noted are the area under the ROC curve measurements for each prediction set.

TABLE 3. Genomes used for *Bacillus anthracis* operon prediction[a]

Phylum	Class	Order	Species
Actinobacteria	*Actinobacteridae*	*Actinomycetales*	*Mycobacterium tuberculosis* CDC1551
Aquificae	*Aquificales*	*Aquificaceae*	*Aquifex aeolicus* VF5
Bacteroidetes	*Bacteroides* (*class*)	*Bacteroidales*	*Bacteroides thetaiotaomicron* VI-5482
Chlamydiae	*Chlamydiales*	*Chlamydiaceae*	*Chlamydia trachomatis* serovar D
Cyanobacteria	*Chroococcales*	*Synechocystis*	*Synechocystis* sp. strain CC6803
Deinococcus-Thermus	*Deinococci*	*Deinococcales*	*Deinococcus radiodurans* R1
Firmicutes	*Bacillales*	*Bacillaceae*	*Bacillus subtilis* 168
			Bacillus cereus ATCC 14579
		Listeriaceae	*Listeria monocytogenes* 4b F2365
		Staphylococcus	*Staphylococcus aureus* MW2
	Clostridia	*Clostridiales*	*Clostridium tetani* E88
	Lactobacillales	*Enterococcaceae*	*Enterococcus faecalis* V583
		Streptococcaceae	*Streptococcus pneumoniae* TIGR4
Fusobacteria	*Fusobacterales*	*Fusobacteriaceae*	*Fusobacterium nucleatum* ATCC 25586
Proteobacteria	*Alphaproteobacteria*	*Caulobacterales*	*Caulobacter crescentus* CB15
	Betaproteobacteria	*Burkholderiales*	*Bordetella bronchiseptica* RB50
	Gammaproteobacteria	*Enterobacteriales*	*Escherichia coli* K-12 MG1655
		Pseudomonadales	*Pseudomonas aeruginosa* AO1
	Epsilonproteobacteria	*Campylobacterales*	*Campylobacter jejuni* NCTC 11168
Spirochaetes	*Spirochaetales*	*Spirochaetaceae*	*Borrelia burgdorferi* B31

[a] Comparative genomes used for prediction of operon structure in *Bacillus anthracis* are shown. The set comprises 20 genomes that are broadly dispersed throughout the bacterial domain, with a slight bias toward the *Firmicutes* (close relatives).

supplemental material). Altogether, the data confirm the earlier indication that the algorithm is relatively insensitive to the phylogenetic distribution of comparative genomes, and they seem to suggest that near-optimal performance can be obtained using a variety of different comparative sets. This point should be stressed, because although it is not yet clear how to construct a perfectly optimal set of comparative genomes for a given bacterial genome, the very small differences we observed in algorithm performance even when relatively large changes were made to the comparative set suggest that a truly optimal comparative set may provide only a minimal improvement.

One other issue we were interested in testing was the effect of BLASTP stringency on the algorithm. The original comparative data were compiled with the requirement that potential orthologs were required to have a BLASTP expect value of less than 10^{-4}; this value is relatively permissive and is similar to the cutoff values used in earlier studies (9). We therefore tested whether making this cutoff value more stringent might affect the predictive value of the comparative data. We found that even when the expect value cutoff was changed to 10^{-8} there was essentially no change in the algorithm's performance (see Fig. S1E in the supplemental material). Given this, all other experiments were performed with the 10^{-4} cutoff value.

Altogether, we found that the algorithm was relatively insensitive to a variety of changes in how the phylogenetic data were compiled, and this was true even when the intergenic distance component was removed from the HMM (that is, the robust behavior of the algorithm to changes in the phylogenetic reference set was not due to the fact that the intergenic distance component was overwhelmingly dominant [data not shown]). The comparative set that provided the best prediction performance was the set containing the 22 species of *Proteobacteria* taken from the original 35 species in Table 1, and using the phylogenetic data compiled from this set together with the intergenic distance information in our algorithm resulted in an area under the ROC curve of 0.916. As shown in Table 2 (data source E), we found that with this method we were able to predict operon structure in *E. coli* with both sensitivity and specificity of >85% and that by choosing the appropriate cutoff value we could fairly easily obtain levels of 90% in either parameter (with a corresponding drop to ~80% in the other). The predictions generated in this set are available at http://www.sph.umich.edu/~qin/hmm/.

Operon prediction in *Bacillus anthracis*. We anticipate that predicting operon structure in previously uncharacterized genomes will provide a variety of clues in terms of possible functional and/or regulatory relationships. This is particularly significant in pathogenic bacteria, where these leads may be useful from the perspective of drug or vaccine development. To test this idea, we used a 20-genome comparative set (chosen as a relatively small, widely diverse group) (Table 3) to predict operon structure in the gram-positive pathogen *Bacillus anthracis*. Although this organism is now being given considerable attention because of its potential as a bioterror agent and several *B. anthracis* strains have been fully sequenced (19, 24, 25, 36), its operon structure remains essentially unknown. On the chromosome of *B. anthracis*, which contains 5,308 protein-coding genes, the algorithm predicted a total of 2,473 cooperonic gene pairs when we used a prediction probability cutoff of 0.5. These pairs form 1,121 multigene operons that contain between 2 and 32 genes, and the probability distribution of operon length is remarkably similar (Pearson's correlation, 0.9917) to that reported for *Bacillus subtilis* in a recent study (7). Although there are very few experimentally verified operons to use in testing the predictions, we note that the gene pairs within the operons that have been verified (i.e., *plcC-spmC*, *csaAB*, *rsbVW-sigB*, *asbABCDEF*, and *gerHABC*) were all predicted successfully.

One of the major benefits of operon prediction in organisms about which little is known is that in many cases we are able to link hypothetical genes to more-well-characterized loci and

TABLE 4. Predicted operon structure in the BA1488-94 region of the *B. anthracis* genome

Gene	Functional annotation	Intergenic distance (to gene$_{N+1}$)	Phylogenetic barcode difference (relative to gene$_{N+1}$)	Predicted probability that gene$_N$ and gene$_{N+1}$ are cooperonic
BA1488	Conserved hypothetical protein	106	18 (of a possible 20)	0.00
BA1489	Superoxide dismutase	113	6 of 20	0.98
BA1490	D-Alanyl-D-alanine carboxypeptidase	−7	12 of 20	1.00
BA1491	Spore maturation protein A	−3	0 of 20	1.00
BA1492	Spore maturation protein	1113	13 of 20	0.00

thus gain some insight into the possible function and regulation of the uncharacterized gene(s). There are a large number of such examples within the predicted *B. anthracis* operons; for instance, the gene BA1581 is a spore coat protein and is surrounded by several conserved hypothetical genes about which nothing is yet known. Spore coat proteins are often attractive options for vaccine and drug development, since they are frequently immunogenic and also tend to play a role in determining the resistance properties of the spore. The prediction algorithm estimated with a very high confidence ($\geq$0.99 in each case) that BA1580, BA1581, BA1582, and BA1583 make up a single operon, and thus by association we are able to propose not only that these uncharacterized genes might be somehow related to formation of the spore coat, but also that they could potentially be targets for new therapeutics.

Another relatively common finding in examining the predicted operon structure within the *B. anthracis* genome was that in many instances, regulatory genes (e.g., loci encoding transcription factors) appear to be cotranscribed with genes that have probable roles in sensing a particular environmental cue. One example is the predicted two-gene operon BA5371-2, which encodes an RNA polymerase sigma factor and a glutaredoxin family protein. This RNA polymerase sigma factor is one of many uncharacterized sigma factors encoded within the *B. anthracis* genome, and its apparent linkage to a glutaredoxin family member seems to suggest that its function may be related to the oxidative state of the environment. Another case in which a probable regulatory function is suggested by operon prediction is the putative three-gene operon BA5503-5, which encodes a sensor histidine kinase, a DNA-binding response regulator, and a UDP-glucose 4-epimerase, respectively. It is typically difficult to predict a priori the environmental signal that a two-component system responds to, and in this instance operon prediction provides a useful clue in suggesting that these genes may be associated with galactose utilization.

Perhaps even more useful are the instances in which prediction of operon structure links disparate biochemical functions and thus assists in our understanding of the organism's biology. A notable example of this is found in examining the genes BA1489 to -92, which encode a putative superoxide dismutase (*sod15*), a D-alanyl-D-alanine carboxypeptidase, and spore maturation proteins A and B, respectively (Table 4). The algorithm predicted that these four genes form a single operon, with a prediction probability of $\geq$0.99 for each of the three internal gene pairs and $\leq$0.01 for the two pairs on either side. Homologs of the three downstream genes have been shown to play roles in spore maturation (22), but there is no obvious function for superoxide dismutase (*sod15*) in this process. Since the finding that they appear to be part of the same operon may imply a possible functional or regulatory link between them, we sought to test the algorithm's prediction that these genes are cotranscribed. We isolated RNA from bacterial samples grown to late exponential phase and performed RT-PCR analyses as diagrammed in Fig. 3. Briefly, we designed primer pairs that would only amplify a product if two adjacent genes could be found on a single mRNA molecule, and we tested whether we could detect the presence of cotranscribed BA1489 and -90 (AB), BA1490 and -1 (CD), and BA1491 and -2 (EF) within the RNA pool. In each of these cases we detected an appropriately sized PCR product, indicating that these gene pairs are cotranscribed at least some of the time (our results do not rule out the possibility of multiple promoters and transcripts that may include some but not all of these genes). These data confirm the prediction made by our algorithm that these four genes are cotranscribed and point to

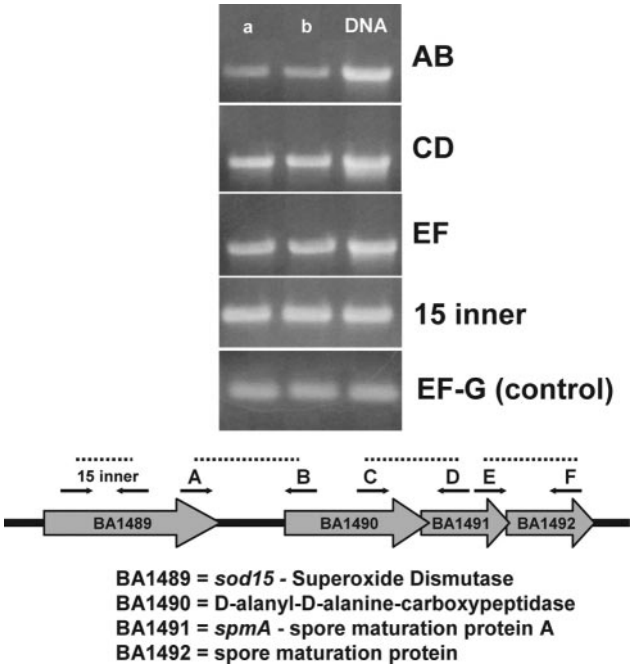

FIG. 3. RT-PCR analysis of the BA1489-92 region. Gels show the PCR products amplified by the designated primer pairs, which are located in the BA1489-92 region of the *B. anthracis* chromosome as shown at the bottom. Control RT-PCRs amplifying a section of the *sod15* (BA1489) locus or a portion of the elongation factor G mRNA sequence are also shown. Note that control reactions in which reverse transcriptase (but not DNA polymerase) was omitted were done for each set and showed in each case that product amplification was not due to DNA contamination.

a previously unseen link between the *sod15* gene and the process of sporulation (and perhaps a specialized role that distinguishes the Sod15 protein from the other three *B. anthracis* superoxide dismutases). Significantly, a related study showed that a strain of *B. anthracis* missing the *sod15* locus formed spores that had ultrastructural differences and a slightly higher sensitivity to heat relative to wild-type *B. anthracis*, confirming the idea that the *sod15* locus is likely involved in sporulation (18). It was also interesting that this operon would have been missed if the algorithm had used either the intergenic distance or the phylogenetic data alone; a model using intergenic distance alone predicts that the BA1489-90 transition is an operon border, and a model using phylogenetic data alone incorrectly predicts that the BA1490-1 pair is not cooperonic. The combined information allows the correct identification of all three intraoperon pairs, and this instance highlights the value of using combined data sources.

On a larger scale, our findings suggest that a large number of functional or regulatory relationships may be implied by the operon predictions within the *B. anthracis* genome, as there are a great many other examples of putative operons that contain seemingly unrelated genes. We anticipate that the predictions made here will lead to a number of potentially interesting experimental questions, and as with *E. coli* the complete set of predictions for *B. anthracis* is available at http://www.sph.umich.edu/~qin/hmm/.

Conclusions. In this study we have demonstrated that adjacent genes within the same operon tend to have a much more similar phylogenetic distribution than adjacent genes that are not cotranscribed, and we have developed a hidden Markov model-based algorithm in which these data can be used to predict operon structure in a newly sequenced bacterial genome. Furthermore, we have shown that when the phylogenetic data are combined with intergenic distance statistics in an inhomogeneous HMM, the algorithm is able to predict operon structure with a high degree of both sensitivity and specificity (Table 2, data source D). Significantly, we find that in general the algorithm appears to perform best using a relatively small group of comparative genomes, and it seems to be somewhat insensitive to the phylogenetic distribution of these genomes relative to each other and to the subject genome. Thus, it appears that the algorithm proposed here is easily portable to other completely sequenced bacterial species, including those for which there are little or no functional or experimental data available, and that operon prediction at or near the levels of specificity and sensitivity shown in Table 2 (data source D) should be attainable for these species as well.

This study was somewhat unique in aiming to construct an algorithm that does not rely on experimental data (e.g., gene expression data) or on detailed gene annotations (e.g., clusters of orthologous genes, or COG, family information) and in aiming to predict operon structure for any given bacterial genome. Perhaps the most similar study to date is that of Price et al., in which the authors proposed an algorithm that relies on intergenic distance, codon usage, and COG information for operon prediction in any bacterial species (23). Although our method relies on fewer data sources and assumes that the distributions of these data are generally species independent, we found that the algorithm's performance is almost identical to that described by Price et al. (areas under ROC curves of 0.916 and 0.917, respectively, when scored on an identical set of known *E. coli* operons) (Fig. 4). The fact that the algorithm presented here performs equivalently with a simpler set of input data is especially significant given that our method does not rely on detailed annotation information, which is unavailable for a significant fraction (20 to 30%) of most bacterial genomes (32), and is therefore able to predict operon structure throughout the entire genome.

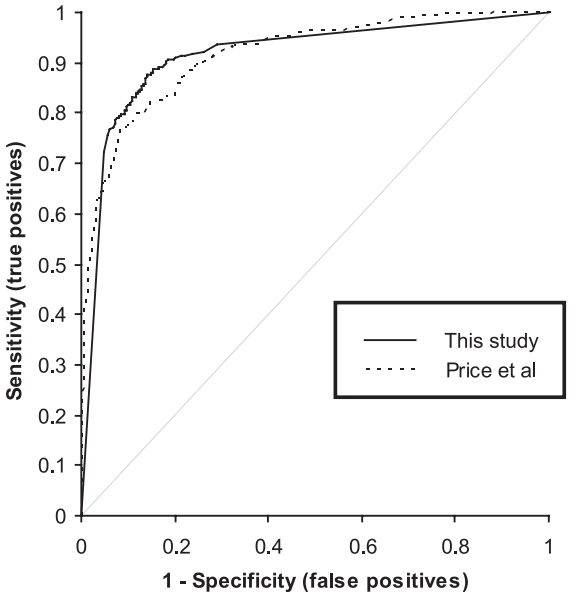

FIG. 4. ROC curves representing the performance of the optimized algorithm (Table 2) and the algorithm described recently by Price et al. (23) when tested using the same set of experimentally verified *E. coli* operons. The area under the curve is 0.916 for the algorithm reported here and 0.917 for that described by Price et al.

In general, it appears that there are several directions that could be taken in attempting to improve the performance of the algorithm described here. We are continuing to investigate different ways of optimizing the set of comparative genomes used in compiling the phylogenetic information, as well as ways to take a more sophisticated view of these data. One idea that seems particularly attractive would be to take advantage of the fact that conserved gene order is a very strong indicator of two genes being cooperonic, even if disruption of that order during evolution does not necessarily also disrupt their cotranscriptional status. It may be possible to improve prediction performance by developing a more complex model in which a measure of conserved gene order is included in the phylogenetic data, and we are currently exploring options for this sort of approach. Similarly, although we showed in this study that the algorithm is surprisingly insensitive to the distribution of reference genomes used, it seems intuitive that cooccurrence of a given gene pair in a more distantly related species is a stronger indicator of cotranscription than cooccurrence in a closely related reference species. Given this, it may be possible to improve performance by incorporating a weighting strategy such that evolutionary distance can be taken into account in compiling the phylogenetic data.

Going beyond phylogenetic and intergenic distance data, it will also be interesting to test the utility of other data sources

when added to the algorithm described here. These include both information available for essentially all bacterial species, such as codon usage or transcriptional terminators, as well as data that are only available for a smaller set of species, such as microarray and detailed functional classifications. Finally, we note that although different prediction algorithms often reach similar levels of accuracy, they typically do not make completely overlapping predictions. An equivalent problem is found in comparing gene prediction algorithms, and a recent study by Allen et al. showed that combining the results of multiple gene prediction models allowed for more accurate results than could be attained by any single algorithm alone (1). If the same trend holds true for operon prediction, it may be possible to reach unprecedented levels of accuracy by combining the predictions generated using several different statistical models.

ACKNOWLEDGMENTS

We gratefully acknowledge Nathan Fisher and other members of the Hanna lab for useful discussions.

This work was supported by DHHS contract N266200400059C/N01-AI-40059.

REFERENCES

1. **Allen, J., M. Pertea, and S. L. Salzberg.** 2004. Computational gene prediction using multiple sources of evidence. Genome Res. **14:**142–148.
2. **Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389–3402.
3. **Bockhorst, J., M. Craven, D. Page, J. Shavlik, and J. Glasner.** 2003. A Bayesian network approach to operon prediction. Bioinformatics **19:**1227–1235.
4. **Bockhorst, J., Y. Qiu, J. Glasner, M. Liu, F. Blattner, and M. Craven.** 2003. Predicting bacterial transcription units using sequence and expression data. Bioinformatics **19**(Suppl. 1)**:**i34–i43.
5. **Chen, X., Z. Su, P. Dam, B. Palenik, Y. Xu, and T. Jiang.** 2004. Operon prediction by comparative genomics: an application to the Synechococcus sp. WH8102 genome. Nucleic Acids Res. **32:**2147–2157.
6. **Craven, M., D. Page, J. Shavlik, J. Bockhorst, and J. Glasner.** 2000. A probabilistic learning approach to whole-genome operon prediction. Proc. Int. Conf. Intell. Syst. Mol. Biol. **8:**116–127.
7. **De Hoon, M. J., S. Imoto, K. Kobayashi, N. Ogasawara, and S. Miyano.** 2004. Predicting the operon structure of Bacillus subtilis using operon length, intergene distance, and gene expression information. Pac. Symp. Biocomput. **2004:**276–287.
8. **Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg.** 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. **27:**4636–4641.
9. **Ermolaeva, M. D., O. White, and S. L. Salzberg.** 2001. Prediction of operons in microbial genomes. Nucleic Acids Res. **29:**1216–1221.
10. **Ettema, T., J. van der Oost, and M. Huynen.** 2001. Modularity in the gain and loss of genes: applications for function prediction. Trends Genet. **17:**485–487.
11. **Itoh, T., K. Takemoto, H. Mori, and T. Gojobori.** 1999. Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. Mol. Biol Evol. **16:**332–346.
12. **Liu, J.** 2002. Bayesian modeling and computation in bioinformatics research, p. 11–44. *In* T. Jiang, Y. Xu, and M. Q. Zhang (ed.), Current topics in computational molecular biology. MIT Press, Cambridge, Mass.
13. **Marcotte, E. M., M. Pellegrini, H. L. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg.** 1999. Detecting protein function and protein-protein interactions from genome sequences. Science **285:**751–753.
14. **Marcotte, E. M., M. Pellegrini, M. J. Thompson, T. O. Yeates, and D. Eisenberg.** 1999. A combined algorithm for genome-wide prediction of protein function. Nature **402:**83–86.
15. **Moreno-Hagelsieb, G., and J. Collado-Vides.** 2002. A powerful non-homology method for the prediction of operons in prokaryotes. Bioinformatics **18**(Suppl. 1)**:**S329–S336.
16. **Ogata, H., W. Fujibuchi, S. Goto, and M. Kanehisa.** 2000. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. Nucleic Acids Res. **28:**4021–4028.
17. **Overbeek, R., M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev.** 1999. The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. USA **96:**2896–2901.
18. **Passalacqua, K. D., N. H. Bergman, A. Herring-Palmer, and P. Hanna.** 2006. The superoxide dismutases of *Bacillus anthracis* do not cooperatively protect against endogenous superoxide stress. J. Bacteriol. **188:**3837–3848.
19. **Pearson, T., J. D. Busch, J. Ravel, T. D. Read, S. D. Rhoton, J. M. U'Ren, T. S. Simonson, S. M. Kachur, R. R. Leadem, M. L. Cardon, M. N. Van Ert, L. Y. Huynh, C. M. Fraser, and P. Keim.** 2004. Phylogenetic discovery bias in Bacillus anthracis using single-nucleotide polymorphisms from whole-genome sequencing. Proc. Natl. Acad. Sci. USA **101:**13536–13541.
20. **Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates.** 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. USA **96:**4285–4288.
21. **Peterson, J. D., L. A. Umayam, T. Dickinson, E. K. Hickey, and O. White.** 2001. The Comprehensive Microbial Resource. Nucleic Acids Res. **29:**123–125.
22. **Popham, D. L., B. Illades-Aguiar, and P. Setlow.** 1995. The *Bacillus subtilis dacB* gene, encoding penicillin-binding protein 5*, is part of a three-gene operon required for proper spore cortex synthesis and spore core dehydration. J. Bacteriol. **177:**4721–4729.
23. **Price, M., K. Huang, E. Alm, and A. Arkin.** 2005. A novel method for accurate operon predictions in all sequenced prokaryotes. Nucleic Acids Res. **33:**880–892.
24. **Read, T. D., S. N. Peterson, N. Tourasse, L. W. Baillie, I. T. Paulsen, K. E. Nelson, H. Tettelin, D. E. Fouts, J. A. Eisen, S. R. Gill, E. Holtzapple, O. A. Okstad, E. Helgason, J. Rilstone, M. Wu, J. F. Kolonay, M. J. Beanan, R. J. Dodson, L. M. Brinkac, M. Gwinn, R. T. DeBoy, R. Madpu, S. C. Daugherty, A. S. Durkin, D. H. Haft, W. C. Nelson, J. D. Peterson, M. Pop, H. M. Khouri, D. Radune, J. L. Benton, Y. Mahamoud, L. Jiang, I. R. Hance, J. F. Weidman, K. J. Berry, R. D. Plaut, A. M. Wolf, K. L. Watkins, W. C. Nierman, A. Hazen, R. Cline, C. Redmond, J. E. Thwaite, O. White, S. L. Salzberg, B. Thomason, A. M. Friedlander, T. M. Koehler, P. C. Hanna, A. B. Kolsto, and C. M. Fraser.** 2003. The genome sequence of *Bacillus anthracis* Ames and comparison to closely-related bacteria. Nature **423:**81–86.
25. **Read, T. D., S. L. Salzberg, M. Pop, M. Shumway, L. Umayam, L. Jiang, E. Holtzapple, J. D. Busch, K. L. Smith, J. M. Schupp, D. Solomon, P. Keim, and C. M. Fraser.** 2002. Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. Science **296:**2028–2033.
26. **Sabatti, C., L. Rohlin, M. K. Oh, and J. C. Liao.** 2002. Co-expression pattern from DNA microarray experiments as a tool for operon prediction. Nucleic Acids Res. **30:**2886–2893.
27. **Salgado, H., S. Gama-Castro, A. Martinez-Antonio, E. Diaz-Peredo, F. Sanchez-Solano, M. Peralta-Gil, D. Garcia-Alonso, V. Jimenez-Jacinto, A. Santos-Zavaleta, C. Bonavides-Martinez, and J. Collado-Vides.** 2004. RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12. Nucleic Acids Res. **32**(database issue)**:**D303–D306.
28. **Salgado, H., G. Moreno-Hagelsieb, T. F. Smith, and J. Collado-Vides.** 2000. Operons in Escherichia coli: genomic analyses and predictions. Proc. Natl. Acad. Sci. USA **97:**6652–6657.
29. **Salzberg, S. L., A. L. Delcher, S. Kasif, and O. White.** 1998. Microbial gene identification using interpolated Markov models. Nucleic Acids Res. **26:**544–548.
30. **Strong, M., P. Mallick, M. Pellegrini, M. J. Thompson, and D. Eisenberg.** 2003. Inference of protein function and protein linkages in Mycobacterium tuberculosis based on prokaryotic genome organization: a combined computational approach. Genome Biol. **4:**R59.
31. **Tamames, J., G. Casari, C. Ouzounis, and A. Valencia.** 1997. Conserved clusters of functionally related genes in two bacterial genomes. J. Mol. Evol. **44:**66–73.
32. **Tatusov, R. L., M. Y. Galperin, D. A. Natale, and E. V. Koonin.** 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. **28:**33–36.
33. **Wang, L., J. D. Trawick, R. Yamamoto, and C. Zamudio.** 2004. Genome-wide operon prediction in Staphylococcus aureus. Nucleic Acids Res. **32:**3689–3702.
34. **Yada, T., M. Nakao, Y. Totoki, and K. Nakai.** 1999. Modeling and predicting transcriptional units of Escherichia coli genes using hidden Markov models. Bioinformatics **15:**987–993.
35. **Zheng, Y., J. D. Szustakowski, L. Fortnow, R. J. Roberts, and S. Kasif.** 2002. Computational identification of operons in microbial genomes. Genome Res. **12:**1221–1230.
36. **Zwick, M. E., F. McAfee, D. J. Cutler, T. D. Read, J. Ravel, G. R. Bowman, D. R. Galloway, and A. Mateczun.** 2005. Microarray-based resequencing of multiple Bacillus anthracis isolates. Genome Biol. **6:**R10.