Published online April 19, 2004

Nucleic Acids Research, 2004, Vol. 32, No. 7 ***2147–2157***
DOI: 10.1093/nar/gkh510

Operon prediction by comparative genomics: an application to the *Synechococcus* sp. WH8102 genome

X. Chen*, Z. Su[1,2], P. Dam[1,2], B. Palenik[3], Y. Xu[1,2] and T. Jiang

Department of Computer Science and Engineering, University of California at Riverside, 900 University Avenue, Riverside, CA 92521, USA, [1]Department of Biochemistry and Molecular Biology, University of Georgia at Athens, GA, [2]Institute of Computational Biology, Oak Ridge National Laboratory, TN, USA and [3]Scripps Institution of Oceanography, University of California at San Diego, CA, USA

Received November 7, 2003; Revised January 30, 2004; Accepted March 8, 2004

ABSTRACT

We present a computational method for operon prediction based on a comparative genomics approach. A group of consecutive genes is considered as a candidate operon if both their gene sequences and functions are conserved across several phylogenetically related genomes. In addition, various supporting data for operons are also collected through the application of public domain computer programs, and used in our prediction method. These include the prediction of conserved gene functions, promoter motifs and terminators. An apparent advantage of our approach over other operon prediction methods is that it does not require many experimental data (such as gene expression data and pathway data) as input. This feature makes it applicable to many newly sequenced genomes that do not have extensive experimental information. In order to validate our prediction, we have tested the method on *Escherichia coli* K12, in which operon structures have been extensively studied, through a comparative analysis against *Haemophilus influenzae* Rd and *Salmonella typhimurium* LT2. Our method successfully predicted most of the 237 known operons. After this initial validation, we then applied the method to a newly sequenced and annotated microbial genome, *Synechococcus* sp. WH8102, through a comparative genome analysis with two other cyanobacterial genomes, *Prochlorococcus marinus s*p. MED4 and *P.marinus* sp. MIT9313. Our results are consistent with previously reported results and statistics on operons in the literature.

INTRODUCTION

Operons represent a basic organizational unit of genes in the complex hierarchical structure of biological processes in a cell of prokaryotes. They are mainly used to facilitate efficient implementation of transcriptional regulation in microbial genomes (1). Operons provide highly useful information for the characterization and construction of biological pathways and networks, at a large scale. Therefore, the prediction of operons at the whole-genome level is one of the most fundamental and challenging computational problems in microbial functional genomics.

Generally, genes in an operon are arranged in tandem in the genomic sequence and delimited by an upstream promoter and a downstream terminator. Operationally, an operon has the following properties, which can be used for their prediction. (i) An operon consists of one or more genes, arranged in tandem on the same strand of a genomic sequence. (ii) Intergenic distances within an operon are generally shorter than the inter-operon distances; hence genes within an operon generally form clusters along the genomic sequence. (iii) Genes in an operon have a common (upstream) promoter and a (downstream) terminator, while the intergenic regions within an operon usually do not contain any promoter or terminator. (iv) Genes in an operon tend to have related functions, which are expected to belong to the same functional category. (v) As a functional unit, the neighborhood relationship of genes in an operon tend to be well conserved across phylogenetically related species; this makes comparative genomics a plausible approach to operon prediction. (vi) If microarray gene expression data are available, genes of the same operon exhibit highly correlated expression patterns.

Because of the importance of operons in functional studies of a microbe, in the past decade, a great amount of effort has been devoted to investigation of effective methods for predicting operons, and a number of algorithms have been developed. These algorithms differ mainly in the operon characteristics that they use, and are summarized here. (i) Overbeek *et al.* (2) proposed a method to search for gene pairs of close bi-directional best hits (PCBBHs) across genomes. These PCBBH pairs form conserved gene clusters that can be used to infer functional coupling. Apparently, the functional coupling information is useful in operon prediction, although the authors did not specify the connection between PCBBH pairs and operons explicitly in their paper. (ii)

*To whom correspondence should be addressed. Tel: +1 909 787 2882; Fax: +1 909 787 2991; Email: xinchen@cs.ucr.edu

Salgado *et al.* (3) introduced a method using intergenic distance distributions and gene functional annotations to predict operons in prokaryotes. However, accurate functional annotations are usually not available for newly sequenced genomes. (iii) Ermolaeva *et al.* (4) proposed a computational method to estimate the likelihood that a set of conserved genes forms an operon. This is based on a mechanism of validating conserved gene pairs by the frequencies of their appearances in multiple genomes. The method also requires that neighboring genes in an operon are within a certain distance and all genes in an operon are located on the same strand. However, it does not consider (conserved) functions of the genes or promoters and terminators in their upstream and downstream regions, respectively. (v) Craven *et al.* (5) developed a probabilistic learning approach to whole-genome operon prediction based on statistical characteristics derived from a variety of operon data including nucleotide sequences (e.g. intergenic distance distribution), transcription control signals (e.g. existence of promoters/terminators), gene expression data and functional annotations associated with genes. This approach first learns a model for estimating the probability that an arbitrary sequence of genes constitutes an operon. It then uses a dynamic programming method to assign each gene in a given genome to its most probable operon, including operons consisting of single genes (although such singletons will not be considered as operons herein). This approach is based on information present in a single genome and is only applicable to well-studied genomes. (vi) Sabatti *et al.* (6) applied a Bayesian classification scheme to gene microarray expression data and then validated the resulting potential operon pairs (POPs), which are simply pairs of adjacent open reading frames (ORFs) in a genome. This approach again relies on experimental data from a single genome. (vii) Zheng *et al.* (7) developed an algorithm that relies on the fact that genes in an operon tend to encode enzymes that catalyze successive reactions in metabolic pathways. Although yielding a high prediction sensitivity as well as specificity, this approach has an apparent limitation that metabolic pathways of studied genes must be given in advance, which is usually not available for a newly sequenced genome. In fact, operon information can be essential for the construction of pathways (8).

Clearly, most of these algorithms require a significant amount of experimental data as input (in addition to gene annotation), which considerably limits their applications. In this study, we present an operon prediction method that employs a comparative genomics approach and incorporates several popular computer programs from the public domain, including BLASTp (9) for homology search, COGnitor and the COG database (10,11) for assigning COG IDs and functional categories, SIGSCAN (12) and the TFD database for identifying promoters, and TransTerm (13) for finding rho-independent terminators. The supporting information given by the last three programs is combined to define a likelihood score for a predicated operon to be a true one. The parameters used in the likelihood score are estimated by examining the 237 known operons in *Escherichia coli* K12 (14). This method requires only a sequenced genome with gene annotation and can thus be applied to recently sequenced genomes that have not been under extensive experimental investigation.

In order to evaluate the performance of our prediction, we tested it on genome *E.coli* K12, in comparison with genomes *Haemophilus influenzae* Rd and *Salmonella typhimurium* LT2, and successfully predicted 178 of the 237 known operons. After the initial validation, we then applied the method to a newly sequenced and annotated microbial genome, *Synechococcus* sp. WH8102 (15), based on comparisons to the genomes of two other closely related microbes, *Prochlorococcus marinus* MED4 and *P.marinus* MIT9313 (16). Our results are consistent with the known operons in *Synechococcus* and previously reported statistics concerning the distribution of operons in a genome. The detailed prediction results can be found at the website http://www.cs.ucr.edu/~xinchen/operons.htm.

A COMPARATIVE GENOMICS APPROACH TO OPERON PREDICTION

With the availability of a large number of complete genome sequences (e.g. about 149 complete microbial genomes have been sequenced so far) for a variety of organisms, comparative analysis is becoming an invaluable tool for understanding genomes and has proven to be more powerful than methods that utilize information from single genomes. Comparative analysis is based on a common belief that functional segments tend to co-evolve at slower rates than non-functional segments, which makes well-conserved regions likely to be of high interest (2). Comparative genomics methods are especially useful for analyzing genomes that have been recently sequenced without extensive experimental studies, as a large amount of information could be derived through identifying corresponding genes or other functional elements between a newly sequenced genome and previously well studied ones.

Our comparative genomics approach for predicting operons is outlined in Figure 1. Here, we use *Synechococcus* sp. WH8102 as the target genome and *P.marinus* MED4 and *P.marinus* MIT9313 as the reference genomes to illustrate the basic idea of our method, though our approach is applicable to any microbial genome and related reference genomes. For simplicity, the names of these genomes are shortened respectively to WH8102, MED4 and MIT9313 in the rest of this paper. There are four major steps in our method. Given a target genome (e.g. WH8102) and some related genomes (e.g. MED4 and MIT9313) with annotated genes/ORFs, we first conduct pairwise genome comparison through finding corresponding (orthologous) genes for each pair of genomes, using BLASTp. The program COGnitor is then used to assign each gene a COG ID. The resulting pairs of matching homologous genes (i.e. with the same COG IDs) are used to build a multistage graph (called a gene-matching graph). The second step is to cluster neighboring genes in the target genome that are conserved across the reference genomes into gene groups and to check if they satisfy several constraints that usually hold for operons, resulting in candidate operons. The third step is to produce for each candidate operon a likelihood score that takes into account several pieces of supporting information such as consistency of function categories among the genes in the operon and the existence of conserved promoters and terminators. Finally, predicted candidate operons are sorted based on their likelihood scores. The details of our method are described below.

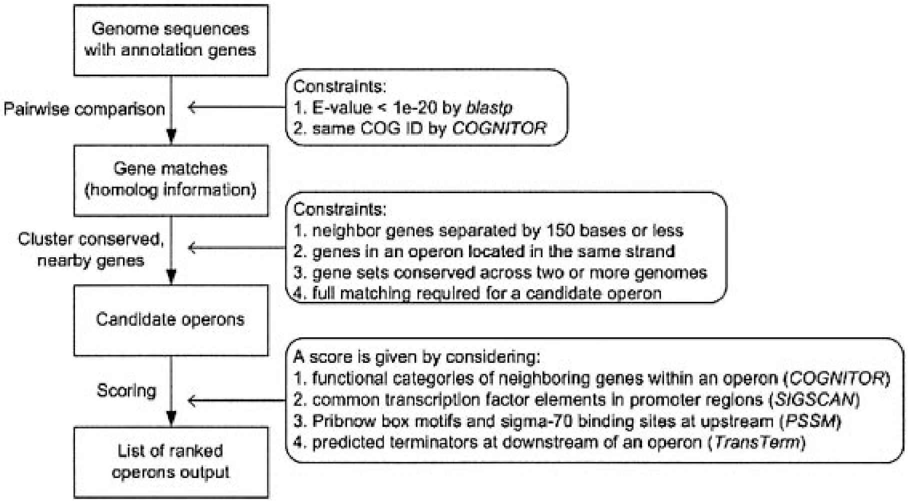
Genome sequences with annotation genes
Pairwise comparison
Constraints:
1. E-value < 1e-20 by blastp
2. same COG ID by COGNITOR
Gene matches (homolog information)
Cluster conserved, nearby genes
Constraints:
1. neighbor genes separated by 150 bases or less
2. genes in an operon located in the same strand
3. gene sets conserved across two or more genomes
4. full matching required for a candidate operon
Candidate operons
Scoring
A score is given by considering:
1. functional categories of neighboring genes within an operon (COGNITOR)
2. common transcription factor elements in promoter regions (SIGSCAN)
3. Pribnow box motifs and sigma-70 binding sites at upstream (PSSM)
4. predicted terminators at downstream of an operon (TransTerm)
List of ranked operons output

Figure 1. An outline of the comparative genomics method for operon prediction.

An interesting question in the above approach is how to choose the reference genomes. It is known that gene order is not conserved in microbes (17), but the composition of the genes in an operon is often conserved in closely related species. In distantly related species, operon structures have undergone extensive shuffling during evolution and hence their gene compositions are generally not conserved (17,18). Based on these observations, we propose to use genomes that are closely related to the target genome as the reference genomes in this comparative genomics method. Of course, such a preference could potentially lead to more erroneously predicted operons (i.e. false positives). However, the probability of getting such false positives should be small, since it is known that genes in very closely related microbial genomes such as different strains of *E.coli* have been extensively shuffled except for the orthologs belonging to conserved operons (19,20). Moreover, besides the conservation of genes and their proximity, our approach also considers several other issues such as gene functions, promoters and terminators. These additional considerations should help eliminate some of the false positives.

Pairwise genome comparison

We match a pair of genes based on two pieces of information: homology measured by the BLASTp program (9) and COG ID generated by the COGnitor program (11). The program BLASTp is widely used to identify homologous genes based on the local alignment of their protein sequences. An *E*-value is provided for each BLASTp, representing the expected frequency of such an alignment appearing in two random genomes. The smaller an *E*-value is, the more probable it is that the two genes are homologous. The COG ID of a gene assigned by the COGnitor program refers to a cluster of orthologous groups (i.e. COG) of genes in the COG database (10). A gene is given a COG ID if it has homologous hits from at least three lineages recorded in the COG database. Genes with the same COG ID are predicted to be orthologous or in an orthologous set of paralogs, and typically share the same function. We have observed that a pair of genes satisfying one of the following conditions may not always satisfy the other one: (i) the *E*-value of BLASTp for a pair of genes is smaller than a pre-defined threshold (e.g. 1e – 20 as used in our experiment on WH8102); and (ii) the two genes have an identical COG ID as assigned by COGnitor.

Therefore, a pair of genes is considered to be a match only when they satisfy both of the two conditions. In other words, we expect matching genes to have the same functions as defined in the COG database. In the case that a gene is not characterized by COG, we would assume that it has the same COG ID as any other genes since its function is unknown as yet. Among the 2517 annotated genes in WH8102, 1343 genes have been assigned COG IDs. The numbers of genes with assigned COG IDs in MED4 and MIT9313 are 1060 (out of 1712) and 1241 (out of 2265), respectively; i.e. an average of 56.1% of genes have been characterized by the COG database. Once all possible pairwise comparisons among all genes in the target genome are made, we build a multistage gene-matching graph. Figure 2 shows a simple three-stage gene-matching graph for three genomes. Notice that, if there are k genomes under consideration, then a total of $k(k-1)/2$ pairwise genome comparisons by BLASTp and COGnitor will be performed. The time required by BLASTp for comparing a pair of genomes is proportional to the total length of the two genomes (9), and the time required by COGnitor is proportional to the total number of genes.

Gene clustering

Given a multistage graph of matching genes, a simple and exhaustive search algorithm is used to cluster genes in the target genome into groups, each of which satisfies the following conditions. (i) If the genes are listed in their sequential order (as given by their locations on the target genome), then each pair of consecutive genes in a group $\{g_1,...,g_k\}$ should be separated by no more than some predefined distance. In our experiment on WH8102, we set the distance threshold as 150 bp. This threshold was chosen based on the average distance between all consecutive genes in WH8102, which is ~96 bp, and the intergenic distance information from some well characterized operons of WH8102 and closely related microbes. (ii) All genes in a

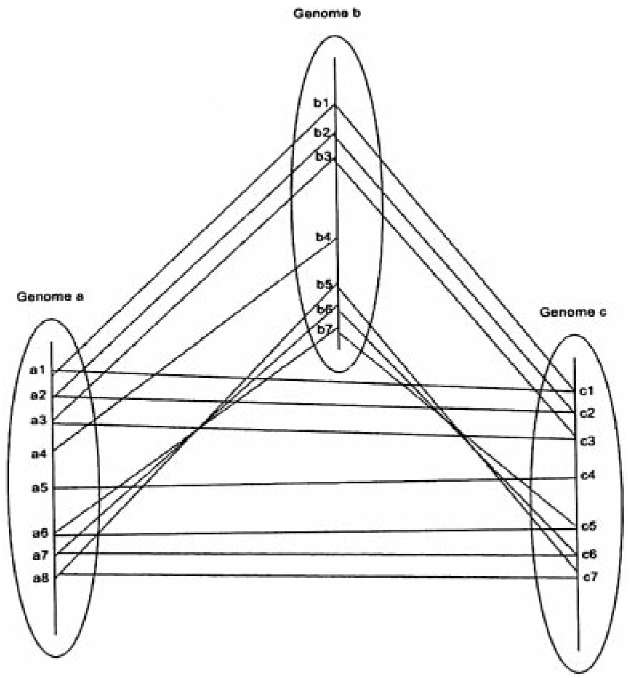

Figure 2. A simple illustration of a three-stage gene-matching graph. Each oval represents a genome, and a link between two genomes represents a pair of matched genes.

clustered group have the same transcription direction, i.e. they are located on the same strand of the target genome. (iii) There exists at least one other (i.e. reference) genome that contains a corresponding gene group $\{g'_1,...,g'_k\}$ so that g_i and g'_i are matched (i.e. there exists an edge between g_i and g'_i in the gene-matching graph) for every index i. The corresponding gene group $\{g'_1,...,g'_k\}$ should also satisfy the two conditions above in its own genome. Observe that the genes in two matching groups do not necessarily have the same sequential order in terms of their locations in their respective genomes. (iv) If there exist two or more reference genomes with matching gene groups, then these corresponding gene groups are also required to match each other.

Our prediction method outputs gene groups from the target genome that satisfy all the above conditions as candidate operons. As an example, for genome *a* in Figure 2, two candidate operons {a1, a2, a3} and {a6, a7, a8} could be predicted if they satisfy all the conditions listed above. The detailed procedure for clustering genes is given below. (i) Sort the list of genes of the target genome in the increasing order of gene locations and denote the sorted list as $g_1g_2...g_n$. (ii) Scan the sorted gene list from the leftmost, i.e. set $i = 1$, and find the longest consecutive subsequence $g_i...g_j$ beginning from gene g_i that satisfies the first two conditions described above. (iii) For each subsequence of $g_i...g_j$, retrieve all the matching gene groups in the reference genomes from the gene-matching graph, and check whether there exists one (or more) such corresponding gene group satisfying the first two of the above conditions and the fourth condition if it applies. If yes, output such a gene group as a candidate operon in the target genome. (iv) Set $i = j + 1$. If $i \geqslant n$, all genes in the target genome have been processed and the algorithm is thus terminated; otherwise go to step (ii).

Although the above procedure is an exhaustive search algorithm because it has to consider all possible gene groups in step (ii), it is in practice very fast and requires only a little memory. This is because a candidate gene group usually consists of a small number of genes. For example, clustering genes for the three genomes (WH8102, MED4 and MIT9313), each with about 2000 genes, takes only 53.4 s on a PC (2.4 GHz).

A likelihood score based on supporting information

In order to provide a quantitative measure on how likely it is that a predicted operon is a true one, we assign a score to each candidate operon as follows. First, we collect a 'profile' for each candidate operon using several public domain programs. The profile includes information about the consistency of gene functions as given by COGnitor and the existence of conserved promoters and terminators as predicted by SIGSCAN and TransTerm. Secondly, a likelihood value is calculated for each item in the profile. The sum of all these likelihood values gives an overall likelihood score for a candidate operon to be a true one. Finally, all the candidate operons are sorted based on their overall likelihood scores. The details are explained below.

Functional categories from COGnitor

Genes in an operon are involved in a specific functional process or pathway. Therefore, such genes tend to share the same (or related) functional category. Based on this observation, a confidence value can be calculated for a predicted operon by considering whether its genes have the same functional category. A key issue here is how to derive gene functions for a newly sequenced genome.

The COGnitor program (11) gives a simple and effective method to infer gene functional categories without requiring much experimental knowledge. Recall that the program was used earlier to classify each gene into a COG. The COGs comprise a framework for the analysis of evolutionary and functional relationships among homologous genes from multiple genomes (10). In particular, the genes belonging to a COG are likely to be involved in the same or similar functional processes.

In the COG database, there are four functional categories at the first level and a total of 16 categories at the second level (see the Appendix). Intuitively, we would expect the genes in an operon to share the same first-level COG functional category and perhaps also the same second-level COG functional category with high probability. In order to test this hypothesis, we studied the 237 known operons from *E.coli* K12 that have been experimentally verified (14). We collected the COG functional categories for the genes in these operons using COGnitor. Because genes in the fourth (first-level) functional category are poorly characterized and their functions are generally unknown, these genes were ignored in the following consistency checking. Out of the 237 known operons, only 39 (i.e. 16.5%) consist of genes from different first-level functional categories. In other words, in each of the remaining 83.5% operons, the genes either belong to the same first-level functional category or fall into the (fourth) poorly characterized category. We also considered the question of how many pairs of neighboring genes share the same second-level functional category. Again, gene pairs that contain at least one gene from the poorly characterized category were ignored. We found that, for the known operons that are conserved in only one reference genome, 67.7% (132 out of 195, denoted as p_0) of the pairs of neighboring genes share the same second-level functional category. As a comparison, among the neighboring gene pairs across the borders, only 24.2% (eight out of 33, denoted as p_1) share the same second-level category. For the known operons that are conserved in two reference genomes, the corresponding percentages are 84.5% (120 out of 142, denoted as p_2) and 23.1% (three out of 13, denoted as p_3), respectively. Therefore, the log-likelihood of a neighboring gene pair being in an operon is defined as follows.

When a putative operon is conserved in only one reference genome,
$L'_0 = \ln(p_0/p_1) \approx 1.029$, if two genes are in the same second-level category
$L'_1 = \ln(1 - p_0/1 - p_1) \approx -0.853$, if two genes are in different second-level categories.

When a putative operon is conserved in two reference genomes,
$L'_0 = \ln(p_2/p_3) \approx 1.297$, if two genes are in the same second-level category
$L'_1 = \ln(1 - p_2/1 - p_3) \approx -1.602$, if two genes are in different second-level categories.

The above likelihood values suggest that, if a pair of neighboring genes have the same second-level functional category, a putative operon conserved in three genomes is more likely to be true than one conserved in two genomes. The overall likelihood for a predicted candidate operon to be a true one, considering only the functional category information, can thus be defined as:

$$L_0 = (m_0 L'_0 + m_1 L'_1)/m$$

where m_0 (or m_1) denotes the number of neighboring gene pairs in the operon belonging to the same second-level category (or different second-level categories, respectively) and m denotes the total number of neighboring gene pairs. Note that the actual values of L'_0 and L'_1 depend on the number of reference genomes across which the candidate operon is conserved, and gene pairs with an uncharacterized gene are ignored in the above likelihood calculation.

Conserved promoters from SIGSCAN

The promoter region of an operon typically consists of several regulatory elements located upstream of the first gene of the operon. Such regulatory elements play a key role in the transcription of the genes of an operon. They are usually conserved across related genomes and their existence could thus provide strong support for a predicted operon. Unfortunately, the promoter information is usually unavailable for a newly sequenced genome and they are hard to predict reliably based on a computational approach.

Here, we adopt a hybrid approach to promoter prediction. We consider a promoter region as a series of transcription factor-binding sites (TFBSs). These transcription factor (TF)-binding motifs may vary from operon to operon, but they are likely to be conserved in the upstream regions of matching operons from different species. Under this assumption, our prediction method makes use of a public domain program, SIGSCAN (12), to find TFBSs in the upstream regions of predicted operons. SIGSCAN searches the transcription factor database (TFD) developed by Ghosh (21), which has a large collection of well-characterized TF-binding motifs in both eukaryotes and prokaryotes, for matching TFBSs. If the found (prokaryotic) TFBSs are conserved in the upstream regions of corresponding candidate operons from several genomes, then we should have more confidence in the predicted operons. Moreover, one would expect that the second gene in an operon is less likely to have such conserved TFBSs.

Because the TFBSs in TFD were mainly extracted from *E.coli* K12, we first performed two experiments to test the applicability of SIGSCAN to our target genome. The first experiment involved (whole) regions between neighboring convergently transcribed genes and those between divergently transcribed genes. The former regions are naturally expected to have more TFBSs than the latter. In the following statistical calculations, overlapping gene pairs are excluded since there are no intergenic (non-coding) regions between them. Among all intergenic regions of two convergent genes in the WH8102 genome, 112 were found to have TFBSs while 169 regions did not. As a comparison, in all the intergenic regions of two divergent genes, 232 regions were found to have TFBSs while 186 did not. In the second experiment, we looked at the promoter region of homologous genes between WH8102 and *E.coli* K12. Sixty-two percent of such homologous genes were found to either have conserved TFBSs or no TFBSs at all. These statistics demonstrate that the SIGSCAN program with TFD does work for cyanobacteria genomes (at least for WH8102 in our experiment), though its effectiveness and reliability may not be very high.

We then ran SIGSCAN on the promoter regions of the known operons from *E.coli* K12. In order to increase the amount of training data, 125 verified transcription units each containing a single gene (14) were also included in the test (these transcription units are also called single-gene operons in some studies). Because very few TFBSs were found to be conserved among the three involved genomes (i.e. *E.coli* K12, *H.influenzae* Rd and *S.typhimurium* LT2), we did not treat TFBSs conserved in three genomes and those conserved in two genomes separately. Operons that are not conserved in any reference genomes were ignored in the statistics (there are 44 such operons among the 237 known operons and 92 from the 125 single-gene transcription units). As a result, the first genes of 43.8% (99 out of 226 = 237 + 125 – 44 – 92, denoted as p_0) of the operons have TFBSs conserved in the 100 bp upstream regions of at least two of the genomes, while 37.6% (79 out of 210 known operons with conserved second genes, denoted as p_1) of the operons have conserved TFBSs in the upstream region of their second genes. (Although some TFBSs are known to appear >100 bp upstream of their respective coding regions, most key TFBSs that are conserved across microbes seem to concentrate in the 100 bp upstream region and we focus on this small window for simplicity herein.) This only poses a weak discrimination; however, the information could still be useful and it will be enhanced when more known operons become available. We define a log-likelihood value based on conserved TFBSs as:

$$L'_2 = \ln(p_0/p_1) \approx 0.152$$

For operons that have no conserved TFBSs across the genomes, we define $L'_2 = 0$.

We further considered two specific TF-binding motifs that exist in most of the promoter regions of prokaryotes: the Pribnow box located at –10 bp from the transcription start site and the motif at –35 bp where the σ70 transcription factor usually binds. The consensus sequence for Pribnow boxes is TATAAT and the consensus for the σ70-binding motif is TTGACA (22). Our algorithm scans the upstream regions of candidate operons using a position-specific score matrix (PSSM) (23), trained on the known Pribnow boxes and σ70-binding sites, to detect pairs of such motifs separated by an almost constant distance (~17 bp). A likelihood value based on the existence of these two TF-binding motifs can be defined in the same way as before: 79.6% (180 out of the 226 known operons with conserved first genes, denoted as p_0) of the known operons from *E.coli* K12 were found to have these motifs conserved in at least two genomes in the 100 bp upstream region of their first genes, while 43.8% (98 out of the 210 known operons with conserved second genes, denoted as p_1) of the operons have motifs conserved in the upstream regions of their second genes. Hence, the log-likelihood score for a candidate operon that has both of the motifs conserved in its 100 bp upstream region is:

$$L'_3 = \ln(p_0/p_1) = 0.620$$

For simplicity, the overall likelihood score for a candidate operon based on its predicted promoter information is then defined as the average of L'_2 and L'_3, i.e.

$$L_1 = (L'_2 + L'_3)/2$$

Conserved terminators from TransTerm

Another key signal that can be used to validate a predicted operon is the terminator of an operon. A terminator usually marks the termination of a transcriptional process downstream of an operon structure. Therefore, we can use knowledge of terminators to infer operons that appear in upstream regions. Unfortunately, the prediction of terminator sites reliably is not a trivial task either.

There are two types of terminators: rho-dependent and rho-independent transcription terminators. The rho-independent terminator usually consists of a hairpin structure followed by a short uracil-rich region. The structure of rho-dependent terminators varies greatly so they are very hard to predict. Therefore, most proposed algorithms in the literature focus on rho-independent terminators. Although there has been some suspicion that many prokaryotes do not have many rho-independent terminators for transcription termination (24), the TransTerm program (13) was able to successfully identify 214 (73 and 244) rho-independent terminators in WH8102 (MED4 and MIT9313, respectively). This indicates that these microbes do use the hairpin structure in transcription termination. Therefore, in our algorithm, we use TransTerm to predict rho-independent terminators in the downstream regions of candidate operons. The TransTerm algorithm scans the input genomes to find hairpins with adjacent uracil-rich stretches and calculates a confidence value for each one. The predicted terminators from TransTerm are then compared across genomes to identify operons with conserved terminators.

In order to formulate a likelihood score based on predicted terminators, we ran TransTerm on the data of the 237 known operons and 125 single-gene transcription units in *E.coli* K12 as we did for promoters. We did not consider terminators conserved in three genomes separately. The results of TransTerm show that 34.9% (78 out of 223 known oeprons with conserved last genes, denoted as p_0) of the operons have terminators conserved in at least two genomes in their 100 bp downstream regions, while only 11.3% (23 out of 203 operons with conserved second last genes, denoted as p_1) have conserved terminators in the 100 bp downstream regions of their second last genes. We hence define the log-likelihood value for a candidate operon that has a terminator conserved in at least two genomes as:

$$L_2 = l \cdot \ln(p_0/p_1) \approx 1.127 \cdot l$$

where l is the confidence score of the terminator from TransTerm (as a measure of the reliability of its prediction). If the operon has no conserved terminator, then $L_2 = 0$.

Our application results of SIGSCAN and TransTerm above show that neither of them could reliably predict promoter elements or terminators, partly due to insufficient studies on prokaryotic promoters and terminators (especially rho-dependent terminators). The overall likelihood score for a candidate operon, combining all the above information, is then given by simply summing up values from each piece of the supporting information:

$$L = L_0 + L_1 + L_2.$$

The summation is based on a simple assumption that all the supporting pieces of evidence for a candidate operon are independent of each other. The final step of our prediction method sorts the candidate operons in the decreasing order of their likelihood scores L. Roughly speaking, a positive L would suggest that the involved candidate operon is likely to be a true one, and a negative L would suggest the opposite. Note that our program outputs all candidate operons regardless of their likelihood scores, i.e. no threshold is applied here. This could be convenient for the users because they could then choose any threshold and apply it on the ranked list of candidate operons.

APPLICATION RESULTS

We have implemented the above method using public domain programs as well as our own algorithms described above and applied it to predict operons in two bacteria, *E.coli* K12 and WH8102, which have completely sequenced and annotated genomes. The first application serves as a controlled test, while the second could provide useful information for a US DoE Genomes to Life (GtL) project (see the website http://www.genomes2life.org/) and to the research community on cyanobacteria studies.

Application to *E.coli* K12

In order to evaluate the performance of the operon prediction method described above, we applied it to the well-studied bacterial genome, *E.coli* K12 (data from GenBank under

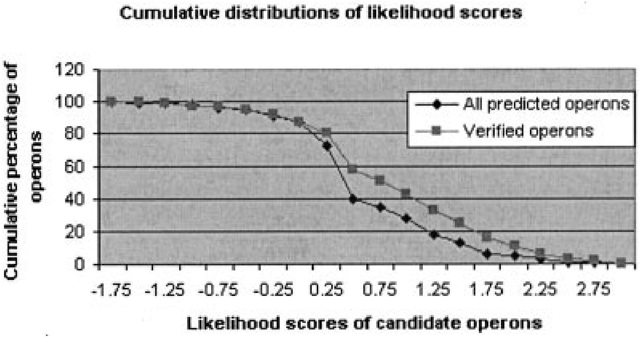

Figure 3. Cumulative distributions of the likelihood scores of predicted and verified operons.

accession number NC_000913). By comparing the genome with another microbial genome, *H.influenzae* Rd (data from GenBank under accession number NC_000907), and using an intergenic distance threshold of 300 bp (chosen based on the intergenic distance information in the 237 known), our method predicted a total of 237 candidate operons shared by these two genomes. A further examination shows that these include 61 out of the 237 experimentally verified operons of *E.coli* K12 described in Salgado *et al.* (14). The result seems quite promising since only one reference genome was considered and some of the 237 verified operons might not be conserved in the reference genome. By comparing these with more phylogenetically related genomes, we could identify more of these known operons. For instance, using both *H.influenzae* Rd and *S.typhimurium* LT2 (data from GenBank under accession number NC_003197) as the reference genomes, our method predicted 853 candidate operons in *E.coli* K12, including 178 of the 237 verified operons.

Since a set of predicted operons may typically include many false positives, we further look at the cumulative distributions of the likelihood scores of the 853 predicted operons and the 178 verified operons, which are depicted in Figure 3. The figure shows that close to 80% of the predicted operons have positive likelihood scores, and the verified operons generally have higher likelihood scores than the predicted ones. For example, 28% of the predicted operons have likelihood scores exceeding 1, while 43% of the verified operons have likelihood scores above 1. This indicates that among the predicted operons, true positives tend to have higher likelihood scores than false positives.

Application to *Synechococcus* sp. WH8102

After the initial validation study, we applied our method to a newly sequenced microbial genome, WH8102 (15), which is the focus of an ongoing US DoE GtL project (see the website http://www.genomes2life.org/). The genome was downloaded from GenBank (under accession number NC_005070), and was compared with two closely related cyanobacterial genomes, MED4 and MIT9313 (data from GenBank under accession number NC_005072 and NC_005071, respectively). These genomes contain 2517, 1712 and 2265 annotated ORFs, respectively. In the step of pairwise comparison, 1448 pairs of matching genes were found between WH8102 and MED4, 1976 pairs were found between WH8102 and MIT9313, and 1422 pairs were found between MED4 and MIT9313. The three-stage gene-matching graph obtained in this step is displayed in Figure 4.

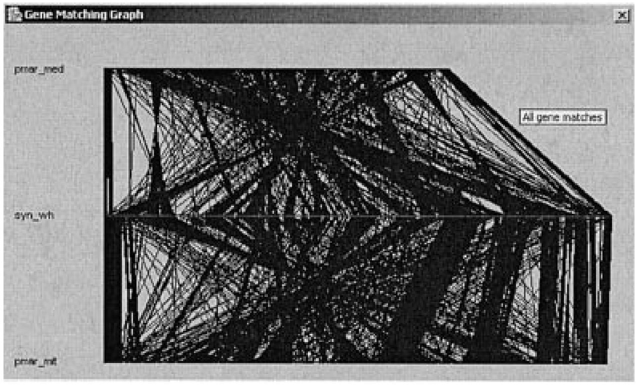

Figure 4. The three-stage gene-matching graph for the three microbial genomes. Each link between two genomes represents a pair of matched genes.

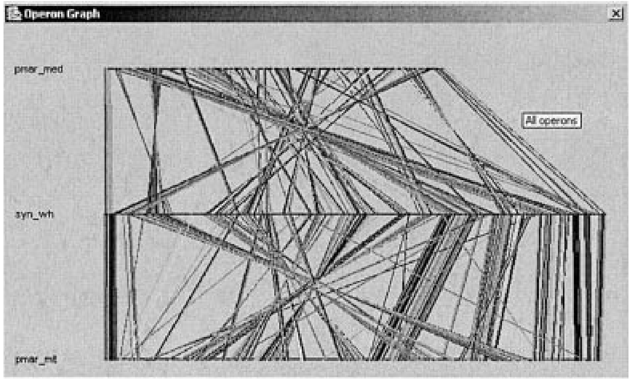

Figure 5. The 446 predicted operons in WH8102 and their distributions in MED4 and MIT9313. Each link represents a candidate operon conserved between two genomes, and different operons are shown using different colors.

The gene clustering process described in the above was then performed to predict all probable operons in WH8102, using an intergenic distance threshold of 150 bp that was chosen based on the average distance between all consecutive genes in WH8102 (i.e. 96 bp) and the maximum intergenic distance in some well characterized operons of WH8102 and closely related microbes. This results in a total of 446 candidate operons. The conservation of the operons is as follows: (i) 206 of the operons are shared with both reference genomes; (ii) 55 operons are shared with MED4 only; and (iii) 185 operons are shared with MIT9313 only.

The distribution of the operons in the three genomes is illustrated in Figure 5. Among these operons, 242 are located on the positive strand of WH8102 and 204 are on the negative strand. These candidate operons were finally sorted based on the supporting information collected from COGnitor, SIGSCAN and TransTerm, and the likelihood score. The detailed prediction results can be found at http://www.cs.ucr.edu/~xinchen/operons.htm.

Although very little is known in the literature about the (true) operons in WH8102, we have been able to identify

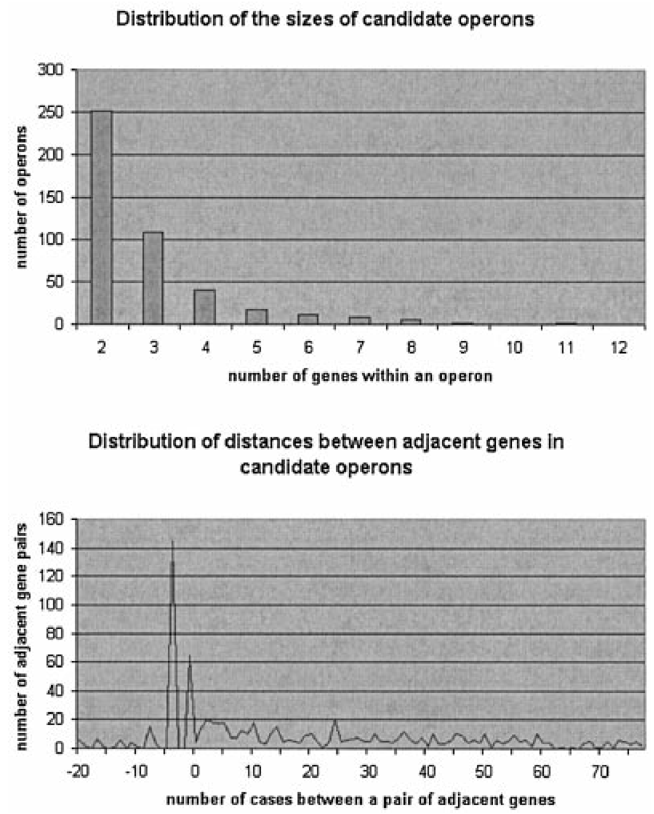

Figure 6. Distributions of the putative operon sizes and intergenic distances.

several known operons from WH8102 or other cyanobacteria in the list of the predicted operons, such as the well-known carboxysome operon whose gene products participate in the carbon reduction cycle (Calvin–Benson–Bassham cycle) and some transporter operons (25). Moreover, we have made several observations about these putative operons.

The average size of the predicted operons (i.e. the average number of genes in each operon) is 2.89, which is very close to the average size reported in the literature. For example, Zheng *et al.* (7) observed that the average size of operons remains as a constant at around 3 in most of the genomes. A graphical illustration of the distribution of the sizes of our predicted operons is given in Figure 6.

The average distance between neighboring genes in an operon is 21.8 bp. Salgado *et al.* (14) analyzed a data set of operons from RegulonDB and found that intergenic distances within an operon peak at some very small values. The two most frequent distances are –4 bp and –1 bp (i.e. many neighboring genes overlap by a few base pairs). This phenomenon is also observed in our predicted operons, as shown in Figure 6.

Among the 446 predicted operons, 345 consist of genes with the same first-level functional categories as assigned by COGnitor, disregarding genes that fall into the poorly characterized category. A total of 139 operons were found to have conserved TFBSs obtained by SIGSCAN in their upstream regions, while 143 operons have conserved pairs of Pribnow box and σ70-binding motifs. Only 21 operons have conserved rho-independent terminators obtained by TransTerm in their downstream regions, which perhaps manifests the difficulty of terminator prediction.

Table 1. The functional categories of the predicted operons in WH8102

Functional categories	Number of operons
Information storage and processing	67
Cellular processes	64
Metabolism	141

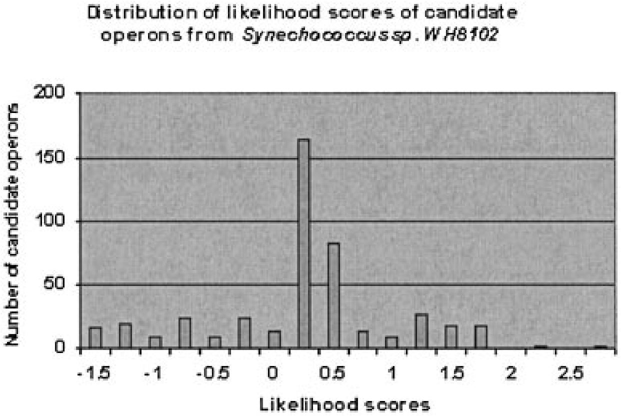

Figure 7. Distribution of likelihood scores of the putative operons in WH8102.

We have classified the putative operons into functional categories as follows. An operon is classified into a functional category if each of its genes is in this functional category or the poorly characterized category, and at least one gene must be in this category. A total of 272 operons were classified, as shown in Table 1.

The distribution of the likelihood scores of these putative operons is shown in Figure 7. The bar diagram shows that most of the operons have positive scores. More precisely, more than half of the operons have likelihood scores around 0.25. Furthermore, there are 74 operons with scores above 1.0, making us greatly confident that they are perhaps true operons.

Table 2 shows 10 putative operons with the highest likelihood scores. The functional information inferred from the COG database indicates that the genes in each operon are involved in the same biological process and thus the operons are likely to be true. The information will be useful in a genome-wide functional analysis (e.g. pathway construction) for WH8102.

Although not many operons are known in WH8102, we have looked at a few known operons such as the well-known carboxysome operon consisting of eight genes (data not shown) and 19 ABC transporter operons. ABC transporters are composed of multiple subunits often found on genes in operons (25). They are the major family of transporters in WH8102. The individual genes were identified as in Palenik *et al.* (15) and collected manually into operons. Our method was able to predict the carboxysome operon successfully. Table 3 exhibits the prediction results on the ABC transporter

Table 2. The top 10 putative operons in WH8102

	WH8102	MED4	MIT9313	Function inferred from COG	Likelihood score
1	SYNW0513	PMM1437	PMT1450	Co-chaperonin GroES (HSP10)	2.72
	SYNW0514	PMM1436	PMT1449	Chaperonin GroEL (HSP60 family)	
2	SYNW2341	PMM0202	PMT2090	Ribosomal protein L10	2.65
	SYNW2340	PMM0201	PMT2091	Ribosomal protein L7/L12	
3	SYNW1967	PMM0325	PMT1649	Cytochrome b subunit of the bc complex	2.06
	SYNW1966	PMM0326	PMT1648	Cytochrome b subunit of the bc complex	
4	SYNW0492	PMM1453	PMT1469	F_0F_1-type ATP synthase β-subunit	2.05
	SYNW0493	PMM1452	PMT1468	F_0F_1-type ATP synthase δ-subunit (mitochondrial oligomycin sensitivity protein)	
	SYNW0494	PMM1451	PMT1467	F0F1-type ATP synthase α-subunit	
	SYNW0495	PMM1450	PMT1466	F0F1-type ATP synthase γ-subunit	
5	SYNW2093	PMM1532	PMT1758	Ribosomal protein L13	1.68
	SYNW2094	PMM1531	PMT1759	Ribosomal protein S9	
	SYNW2095	PMM1530	PMT1760	Ribosomal protein L31	
	SYNW2096	PMM1529	PMT1761	Protein chain release factor A	
6	SYNW0709	PMM1049	PMT1145	ABC-type dipeptide/oligopeptide/nickel transport systems, periplasmic components	1.68
	SYNW0708	PMM1048	PMT1146	ABC-type dipeptide/oligopeptide/nickel transport systems, permease components	
7	SYNW1341	PMM0601	PMT0418	ABC-type Mn/Zn transport system, periplasmic Mn/Zn-binding (lipo) protein (surface adhesin A)	1.68
	SYNW1340	PMM0602	PMT0417	ABC-type Mn/Zn transport systems, ATPase component	
8	SYNW1090	PMM0753	PMT0584	Ribosomal protein S2	1.68
	SYNW1091	PMM0754	PMT0583	Translation elongation factor Ts	
9	SYNW1239	PMM0878	PMT728	Branched-chain amino acid aminotransferase/ 4-amino-4-deoxychorismate lyase	1.68
	SYNW1238	PMM0877	PMT729	Methionine synthase I, cobalamin-binding domain	
10	SYNW0917	PMM1103	PMT1086	ATP-dependent exoDNase (exonuclease V) β-subunit (contains helicase and exonuclease domains)	1.61
	SYNW0918	PMM1102	PMT1085	ATP-dependent exoDNase (exonuclease V), α-subunit—helicase superfamily I member	

operons. Close to 50% of these operons were predicted exactly by our method, and most of the remaining operons were partially predicted. Only two operons were completely missed. One of these is a likely transporter operon found only in WH8102, but not in the other genomes. The other WH8102 operon missed has genes that do not seem to be in an operon in the other two genomes. Moreover, almost all of these predictions had positive likelihood scores.

DISCUSSION

The prediction of operons is an important and highly challenging problem in computational biology. The problem becomes even more difficult when not many experimental data are available. Here, we have presented an approach based on comparative genomics that incorporates several public domain programs such as BLASTp, COGnitor, SIGSCAN and TransTerm. The approach has been tested on two bacterial genomes, *E.coli* K12 and WH8102, with very promising results. The candidate operons predicted for WH8102 are being used in the study of functional pathways in the organism, such as the phosphorus assimilation pathway (8), and our future effort will include applying the prediction method to more genomes (e.g. we may include *P.marinus* sp. SS120 as an additional reference genome in the analysis of WH8102) and extending the method to allow gene insertions and deletions in an operon structure.

Our above prediction on WH8102 was based on the intergenic distance threshold of 150 bp. It would be interesting to know the sensitivity of the prediction results to this threshold (when it is sufficiently large). We have also considered two other thresholds, 100 and 200 bp. Using these thresholds, our method predicted 446 and 451 operons, respectively, in WH8102. These numbers are the same as or a little bit higher than the number of operons predicted using the threshold 150 bp, although the actual predicted operons are slightly different (data not shown) in all three cases because some operons could be merged and some new operons could be added when the distance threshold increases. This suggests that our overall prediction results on WH8102 were not very sensitive to the intergenic distance threshold, as long as it is larger than the average distance between all consecutive genes in the genome. We also plan to consider replacing the hard intergenic distance threshold by a more flexible constraint that takes into account the common distribution of intergenic distances within known operons.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous referees for many valuable suggestions in the revision. This work was

Table 3. The ABC transporter operons and our prediction results

ABC transporter operon	Prediction result	Likelihood score
(SYNW0211, 0212)	Exactly found	0.31
(SYNW0319, 0320, 0321)	(SYNW0319, 0320, 0321, 0322)	0.39
(SYNW0708, 0709)	Exactly found	1.68
(SYNW0840, 0841, 0842, 0843)	(SYNW0840, 0841); (SYNW0842, 0843)	1.34; 1.03
(SYNW0969, 0970)	Exactly found	1.03
(SYNW1086, 1087)	(SYNW1084, 1085, 1086, 1087)	0.31
(SYNW1111, 1112)	(SYNW1109, 1110, 1111, 1112, 1113, 1114)	0.00
(SYNW1168, 1169, 1170)	(SYNW1167, 1168); (SYNW1170, 1171)	0.08; 0.39
(SYNW1270, 1271, 1272)	Exactly found	1.61
(SYNW1283, 1284, 1285)	(SYNW1282, 1283, 1284, 1285)	1.03
(SYNW1340, 1341)	Exactly found	1.68
(SYNW1415, 1416, 1417)	Not found	–
(SYNW1797, 1798)	Not found	–
(SYNW1857, 1858)	(SYNW1857, 1858, 1859, 1860)	–0.22
(SYNW1915, 1916, 1917)	Exactly found	1.42
(SYNW2393, 2394, 2395)	Exactly found	0.08
(SYNW2438, 2439, 2440, 2441, 2442)	(SYNW2438, 2439, 2440, 2441, 2442, 2443)	1.04
(SYNW2479, 2480, 2481)	(SYNW2479, 2480)	1.03
(SYNW2485, 2486, 2487)	Exactly found	0.31

funded in part by the US Department of Energy's Genomes to Life Program under project 'Carbon Sequestration in *Synechococcus* sp: From Molecular Machines to Hierarchical Modeling' (http://www.genomes-to-life.org). Y.X.'s work is also supported, in part, by the NSF grant ITR-325386. T.J.'s research is also supported, in part, by NSF grant ITR-0085910.

REFERENCES

1. Xu,Y. (2004) Computational genome annotation. In Zhou,J., Thompson,D., Xu,Y. and Tidge,J. (eds), *Microbial Functional Genomics*. Wiley-LISS, Hoboken, NJ, pp. 41–66.
2. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
3. Salgado,H., Moreno-Hagelsieb,G., Smith,T. and Collado-Vides,J. (2000) Operons in *Escherichia coli*: genomic analyses and predictions. *Proc. Natl Acad. Sci. USA*, **97**, 6652–6657.
4. Ermolaeva,M.D., White,O. and Salzberg,S.L. (2001) Prediction of operons in microbial genomes. *Nucleic Acids Res.*, **29**, 1216–1221.
5. Craven,M., Page,D., Shavlik,J., Bockhorst,J. and Glasner,J. (2000) A probabilistic learning approach to whole-genome operon prediction. *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, San Diego, CA, pp. 116–127.
6. Sabatti,C., Rohlin,L., Oh,M.K. and Liao,J.C. (2002) Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res.*, **30**, 2886–2893.
7. Zheng,Y., Szustakowski,J.D., Fortnow,L., Roberts,R.J. and Kasif,S. (2002) Computational identification of operons in microbial genomes. *Genome Res.*, **12**, 1221–1230.
8. Su,Z., Dam,P., Chen,X., Olman,V., Jiang,T., Palenik,B. and Xu,Y. (2003) Computational inference of regulatory pathways in microbes: an application to phosphorus assimilation pathways in *Synechococcus* WH8102. *Proceedings of 14th Conference on Genome Informatics (GIW)*, Universal Academy Press, Tokyo, pp. 3–13.
9. Altschul,S., Madden,T., Schaffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
10. Tatusov,R., Koonin,E. and Lipman,D. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
11. Tatusov,R., Natale,D., Garkavtsev,I., Tatusova,T., Shankavaram,U., Rao,B., Kiryutin,B., Galperin,M., Fedorova,N. and Koonin,E. (2001) The COG databases: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
12. Prestridge,D. (1991) SIGNAL SCAN: a computer program that scans DNA sequences for eukaryotic transcriptional elements. *CABIOS*, **7**, 203–206.
13. Ermolaeva,M., Khalak,H, White,O., Smith,H. and Salzberg,S. (2000) Prediction of transcription terminators in bacterial genomes. *J. Mol. Biol.* **301**, 27–33.
14. Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Millan-Zarate,D., Blattner,F.R. and Collado-Vides,J. (2001) RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.*, **29**, 72–74.
15. Palenik,B., Brahamsha,B., Larimer,F.W., Land,M., Hauser,L., Chain,P., Lamerdin,J., Regala,W., Allen,E.E., McCarren,J. *et al.* (2003) The genome of a motile marine *Synechococcus*. *Nature*, **424**, 1037–1042.
16. Rocap,G., Larimer,F.W., Lamerdin,J., Malfatti,S., Chain,P., Ahlgren,N.A., Arellano,A., Coleman,M., Hauser,L., Hess,W.R. *et al.* (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature*, **424**, 1042–1047.
17. Mushegian,A.R. and Koonin,E.V. (1996) Gene order is not conserved in bacterial evolution. *Trends Genet.*, **12**, 289–290.
18. Gelfand,M.S. (1999) Recognition of regulatory sites by genomic comparison. *Res. Microbiol.*, **150**, 755–771.
19. Hayashi,T., Makino,K., Ohnishi,M., Kurokawa,K., Ishii,K., Yokoyama,K., Han,C.G., Ohtsubo,E., Nakayama,K., Murata,T. *et al.* (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.*, **8**, 11–22.
20. Welch,R.A., Burland,V., Plunkett,G.,III, Redford,P., Roesch,P., Rasho,D., Buckles.E.L., Liou,S.R., Boutin,A., Hackett,J. *et al.* (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **99**, 17020–17024.
21. Ghosh,D. (1991) New developments of a transcription factors database. *Trends Biochem. Sci.*, **16**, 445–447.
22. Lewin,B. (2000) *Genes VII*. Oxford University Press, New York, NY.
23. Stormo,G. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
24. Washio,T., Sasayama,J. and Tomita,M. (1998) Analysis of complete genomes suggests that many prokaryotes do not rely on hairpin formation in transcription termination. *Nucleic Acids Res.*, **26**, 5456–5463.
25. Higgins,C.F. (1992) ABC transporters: from microorganisms to man. *Annu. Rev. Cell. Biol.*, **8**, 67–113.

APPENDIX

Functional categories in the COG database (http://www.ncbi.nlm.nih.gov/COG/):

1. Information storage and processing
   a. Translation, ribosomal structure and biogenesis (J)
   b. Transcription (K)
   c. DNA replication, recombination and repair (L)
2. Cellular processes
   a. Cell division and chromosome partitioning (D)
   b. Post-translational modification, protein turnover (O)
   c. Cell envelope biogenesis, outer membrane (M)
   d. Cell motility and secretion (N)
   e. Inorganic ion transport and metabolism (P)
   f. Signal transduction mechanisms (T)
3. Metabolism
   a. Energy production and conversion (C)
   b. Carbohydrate transport and metabolism (G)
   c. Amino acid transport and metabolism (E)
   d. Nucleotide transport and metabolism (F)
   e. Coenzyme metabolism (H)
   f. Lipid metabolism (I)
   g. Secondary metabolites biosynthesis, transport and catabolism (Q)
4. Poorly characterized
   a. General function prediction only (R)
   b. Function unknown (S)