2418–2427 Nucleic Acids Research, 2006, Vol. 34, No. 8
doi:10.1093/nar/gkl294

Detecting uber-operons in prokaryotic genomes

Dongsheng Che[2], Guojun Li[1,3], Fenglou Mao[1], Hongwei Wu[1] and Ying Xu[1,2,*]

[1]Department of Biochemistry and Molecular Biology, [2]Department of Computer Science, University of Georgia, USA and [3]School of Mathematics and System Sciences, Shandong University, China

Received January 6, 2006; Revised February 15, 2006; Accepted April 6, 2006

ABSTRACT

We present a study on computational identification of uber-operons in a prokaryotic genome, each of which represents a group of operons that are evolutionarily or functionally associated through operons in other (reference) genomes. Uber-operons represent a rich set of footprints of operon evolution, whose full utilization could lead to new and more powerful tools for elucidation of biological pathways and networks than what operons have provided, and a better understanding of prokaryotic genome structures and evolution. Our prediction algorithm predicts uber-operons through identifying groups of functionally or transcriptionally related operons, whose gene sets are conserved across the target and multiple reference genomes. Using this algorithm, we have predicted uber-operons for each of a group of 91 genomes, using the other 90 genomes as references. In particular, we predicted 158 uber-operons in *Escherichia coli* K12 covering 1830 genes, and found that many of the uber-operons correspond to parts of known regulons or biological pathways or are involved in highly related biological processes based on their Gene Ontology (GO) assignments. For some of the predicted uber-operons that are not parts of known regulons or pathways, our analyses indicate that their genes are highly likely to work together in the same biological processes, suggesting the possibility of new regulons and pathways. We believe that our uber-operon prediction provides a highly useful capability and a rich information source for elucidation of complex biological processes, such as pathways in microbes. All the prediction results are available at our Uber-Operon Database: http://csbl.bmb.uga.edu/uber, the first of its kind.

INTRODUCTION

The rapidly expanding pool of sequenced microbial genomes provides a very rich source of information for deciphering the hidden information encoded in a genome and the organizational structures of the encoded information. One powerful tool for decoding such information is the so-called comparative genome analysis, which attempts to derive the encoded information through directly comparing the genomes against one another. Through such comparisons, 'conserved' genomic structures at various organizational levels could possibly be detected (1,2). Then by linking these identified genomic structures to already well-established biological entities, one could possibly infer their biological meanings (3–5). For situations where such links are not clearly identifiable yet, the significance of the uncovered 'conserved' genomic structures could possibly be established through statistical means. Comparative genome analyses have been used to predict operon structures, a layer of well-established genomic structure, at a whole genome scale (2,6–9). As more powerful comparative genome analysis tools become available, we expect that more genomic structures, previously understood or new, will be revealed.

As we understand now, there are a number of well-established higher-level genomic structures beyond operons in a microbial genome, which include regulons, modulons and stimulons. A regulon (10,11) is a group of operons which are co-regulated by the same transcriptional machinery, while a modulon (10,11) is a group of regulons that are controlled by more global regulators and respond to more general physiological states of a cell. At an even higher level is a set of stimulons (10,11), each of which consists of a collection of operons, regulons and/or modulons that respond to a common environmental stimulus. Each of these genomic structures generally encodes a biological pathway or a complex network (or possibly portions of a pathway/network). Hence identification and characterization of these genomic structures has direct implications in deciphering biological pathways and networks in a systematic manner, which represents one of the key tasks in the study of an organism at a systems level.

*To whom correspondence should be addressed. Tel: 1 706 542 9779; Fax: 1 706 542 9751; Email: Ying Xu xyn@bmb.uga.edu

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors

It is known that in bacteria, genes are transcribed using operons (including single-gene operons) as the basic units, while in eukaryotes genes are transcribed individually. While the exact reason for this phenomenon requires more investigation, we suspect that one possible reason might be that as organisms evolve to become more complex, they might have the tendency to use each of their genes in more biological processes, which requires the flexibility of different gene associations to efficiently handle different needs for co-transcription. This, in turn, might have led to the breakup of the fixed gene associations enforced by the (large) operon structures in the ancient and simple organisms to possibly smaller transcriptional units in more complex organisms. We have recently carried out a systematic study on the tryptophan synthesis operons. We found that these operons are fairly larger (average operon size is 6.4 for 24 archaeabacteria genomes) in some archaeabacteria while their sizes are in general smaller (average operon size is 1.4 for 17 cyanobacteria genomes) in some cyanobacteria (P. Wan, F. Mao and Y. Xu, manuscript in preparation). This observation seems to suggest that some of these operons may have experienced the fission process during the evolution. To the extreme along this discussion, all eukaryotic genomes have each of their genes individually regulated transcriptionally, i.e. all their 'operons' are singletons.

By identifying operons that used to be associated with some ancient organisms (e.g. two whole operons or parts of them belong to the same operon in an ancient organism) or other organisms in general, we may detect the footprints of operon evolution. This footprint of operon evolution might provide useful information leading to not only better understanding about genomic structures and their organization, but also possibly a new set of tools for studying biological machineries in a prokaryotic cell, just like the powerful tool that operons have provided to biological pathway prediction (3–5,12,13). In this study, we focus on the identification of the footprint of a particular class of operon evolution, uber-operons, a concept introduced by Lathe *et al*. (14). The essential idea of a uber-operon is that during evolution, larger operons might have broken into smaller operons in different ways along different evolutionary lineages. Hence by studying conservations among groups of operons ('uber-operons') rather than individual operons, it may help to uncover the 'lost' association relationships among operons that used to work together constitutively. By the very nature of the uber-operon definition, it requires reference genomes to uncover the 'lost' association relationship among of the uber-operons. In particular, it requires the knowledge of orthologous genes across genomes. Lathe *et al*. (14) proposed an iterative procedure for identification of uber-operons, assuming that orthologous gene relationships are given, which has limited the practical value of their method. In this paper, we present a novel algorithm to simultaneously identify uber-operons in a target genome and orthologous gene relationships between the target and reference genomes. We first give a revised definition of uber-operons, which we believe is more precise and better captures the association relationship among operons as outlined above. A uber-operon, U, is a group of operons in a genome whose component operons are transcriptionally or functionally related, and U is conserved across multiple (reference) genomes in the following sense: the orthologous genes of U's genes in each of these reference genomes form a group of operons, which (approximately) contain these orthologous genes only (i.e. these operons approximately do not contain other genes nor miss genes). Here 'transcriptionally related' refers to that operons are transcriptionally co-regulated (15); 'functionally related' refers to that operons include genes of the same pathway (15) or with highly similar Gene Ontology (GO) annotations (16); and orthologous genes (or simply, orthologs) refer to isofunctional and heterospecic genes (17–19) throughout this paper. Another concept used repeatedly throughout the paper is linker genes, each of which refers to a pair of genes in a genome that each gene is in a different operon and their orthologous genes are in the same operon in a reference genome.

MATERIALS AND METHODS

In our two-layer algorithm, the lower layer predicts the initial candidate uber-operons through identifying a set of linker genes using a single reference genome. The higher layer fuses all the uber-operon predictions provided by the lower layer against each set of reference genomes to give the final prediction. The purpose of using multiple reference genomes is to increase the prediction reliability by reducing accidental false prediction or missing linker genes, which may occur by using a single reference genome.

Data preparation

By selecting one complete genome in each genus, we have obtained 115 genomes from 224 complete bacterial genomes at the NCBI website (release of 03/05/2005). Operon prediction results for these genomes were downloaded from http://www.microbesonline.org/operons/, denoted as *VIMSS* operons (7). We have also applied our in-house program, JPOP (2,6), for operon prediction. The average operon size predicted by JPOP is slightly smaller than that of the *VIMSS* operons, although the two programs have similar prediction accuracy (F. Mao and Y. Xu, unpublished data). The *VIMSS* operons are used for our study, because their slightly larger operon size should in principle lead to lower false negative rate in linker gene identification. Since *VIMSS* has operon predictions for only 91 out of the 115 genomes (including *Escherichia coli* K12), we have removed the remaining 24 genomes from further consideration (see Supplementary Table S1).

Another dataset needed for our uber-operon prediction is the homologous genes in the reference genomes for each gene in our target genome. We have carried out a homologous gene mapping for each of the 91 genomes against the remaining 90 genomes, using BLAST search with an E-value cutoff at 10^{-3}. Both the predicted operons and the homologous genes are provided at our Uber-Operon Database: http://csbl.bmb.uga.edu/uber.

Uber-operon prediction against one reference genome

We first formulate the problem of uber-operon identification based on one reference genome, and then outline an algorithm for solving the problem. The main and fundamental difference between our algorithm and the algorithm of (14) is that we do not assume that the orthologous gene relationship is given; instead orthologous gene relationship is detected simultaneously with uber-operon prediction.

Consider a target genome G_1 and a reference genome G_2. We assume that each gene in G_1 has at most one ortholog in G_2, and vice versa. Intuitively, a uber-operon is modeled as a maximal group of transcriptionally or functionally related operons that are linked through linker genes; and there is no overlap between any two uber-operons (unlike regulons). One challenging issue in identifying uber-operons is to accurately identify orthologous genes between two genomes. Our previous study has demonstrated that existing methods, such as BDBH (20), its variations (21) and COG (22) are not adequate for highly specific and accurate identification of orthologous genes at a large scale, since these algorithms all attempt to predict orthology based mainly on sequence similarity information, and sequence similarity information alone does not imply orthology (12). This problem has been partially overcome by a new strategy employed in our recent work on orthologous gene mapping by using both sequence similarity and genomic structure information (12,23). The basic idea is as follows. If a pair of genes g_1, g_2 are in the same operon of G_1 and their homologous genes ${g_1}'$ and ${g_2}'$ are also in the same operon in $G_{2,}$ then the probability for g_1 and ${g_1}'$ and g_2 and ${g_2}'$, respectively, to be orthologous is high (23). So our uber-operon identification algorithm is to find such mappings in the context of finding uber-operons, which maximizes the overall probability for all the mapped gene pairs to be orthologous.

Formally, we define a bipartite graph $B = (U, V, E)$ for genomes G_1 and G_2 as follows. Let $U = \bigcup_{i=1}^{m} U_i$ and $V = \bigcup_{j=1}^{n} V_j$ be the two vertex sets, with $U_i = \{u_{i,s} \mid s = 1, 2, \ldots, p_i\}$ and $V_j = \{v_{j,t} \mid t = 1, 2, \ldots, q_j\}$ representing the gene list of the ith operon of G_1 and the gene list of the jth operon of G_2, respectively; $u_{i,s}$ and $v_{j,t}$ representing the sth gene in U_i and the tth gene in V_i, respectively; p_i and q_j being the numbers of genes in U_i and V_j, respectively; and m and n being the numbers of operons in G_1 and G_2, respectively. E is the edge set connecting vertices of U and V such that an edge exists if and only if the two corresponding genes are homologous defined by BLAST with an E-value cutoff 10^{-3}. A matching (24) of B is defined as a subset of E such that no two edges in the subset share a common vertex. Intuitively a matching represents a one-to-one correspondence between genes in subsets of U and V. For any matching M of B, we define a multigraph $A_M = (O, M)$, with $O = \{U_i; V_j \mid 1 \leqslant i \leqslant m; 1 \leqslant j \leqslant n\}$ being the vertex set and M being the edge set. It should be noted that the edge set of B and A_M are the same. In B the vertices are genes and in A_M the vertices are operons, thus there can be multiple edges between two vertices in A_M, so A_M is a multigraph. Define $c(M)$ to be the number of connected components of A_M. An uber-operon identification problem is defined as to find the maximum matching M of B that maximizes $c(M)$. While a detailed discussion on the rationale for our objective function is given in the Supplementary Data, intuitively this formulation attempts to partition B into as highly densely linked (through homologous relationships) operons across the two genomes as possible, particularly to maximize the number of orthologous gene pairs as defined above.

For a general bipartite graph without any constraint, finding the maximum matching can be solved efficiently (24). However, it is computationally highly challenging to solve the constrained maximum matching problem. We have proved that the uber-operon identification problem, formulated as above, is NP-hard (the proof is given in the Supplementary Data), indicating that there is no fast and rigorous algorithm for solving this problem. So we present a heuristic algorithm for this problem. The basic idea is as follows: we first find non-overlapping individual operon pairs (no operon pairs share the same operon) across U and V that give the highest total matching size among all such operon pairs. This can be achieved by first finding one pair of operons that has highest matching size between any possible operon pairs across U and V; and then remove this pair from B and repeat this procedure on the updated B until no more operon pairs can be found. We then merge operon pairs (or operon-group pairs) into operon-group pairs if such merging can lead to the increase of the overall matching size, or more specifically the objective value. This merge operation is repeated until the objective value cannot be increased any more. The resulting operon groups in U and V are the predicted uber-operons in the two genomes, respectively. Although this heuristic algorithm does not guarantee to reach the globally optimal solution, the following property can always lead this algorithm to reach a good solution: the orthologous gene pairs in two conserved uber-operons (of two genomes) are always denser than that in two unrelated uber-operons.

We now provide a formal description of the algorithm on how to merge the connected groups. We first construct a dynamic auxiliary weighted graph $G(M) = [V(M), E(M)]$ for a given matching M, where the vertex set $V(M)$ consists of all the connected components of $A_M = (O, M)$, and the edge set $E(M)$ is created dynamically by connecting any two connected components of A_M. The weight of the edge e, which is created by connecting two connected components, say C_1 and C_2, is defined as follows: Let M_1 and M_2 be the current maximum matching of C_1 and C_2, respectively, and $M_{1,2}$ be the maximum matching of the subgraph C_{12}, which is created by combining C_1 and C_2, then the weight of edge $e = (C_1, C_2)$ is defined as $w(e) = |M_{1,2}| - |M_1| - |M_2|$. In fact, the weight of an edge is the number of augmenting paths (24) related to M in the subgraph C_{12}. A schematic diagram of our algorithm is shown in Figure 1.

Initially, $M = \phi$ (the empty set) and $G(M) = [V(\phi), E(\phi)]$. The algorithm starts to find and merge two connected components where the edge between them has the maximum weight among all edges in $E(M)$ (Figure 1); then the algorithm updates the auxiliary graph and the connected components, and repeats the merge operation. The iterative process stops when the maximum weight of edges in $E(M)$ reaches zero. At this point, the final matching M and the final connected components are reached. Though the algorithm does not guarantee to find the globally optimal matching, we found that in practice, the maximal matching M identified by this algorithm is often the globally optimal solution (data not shown). Our algorithm outputs M, which gives orthologous gene pairs between G_1 and G_2, and the connected components

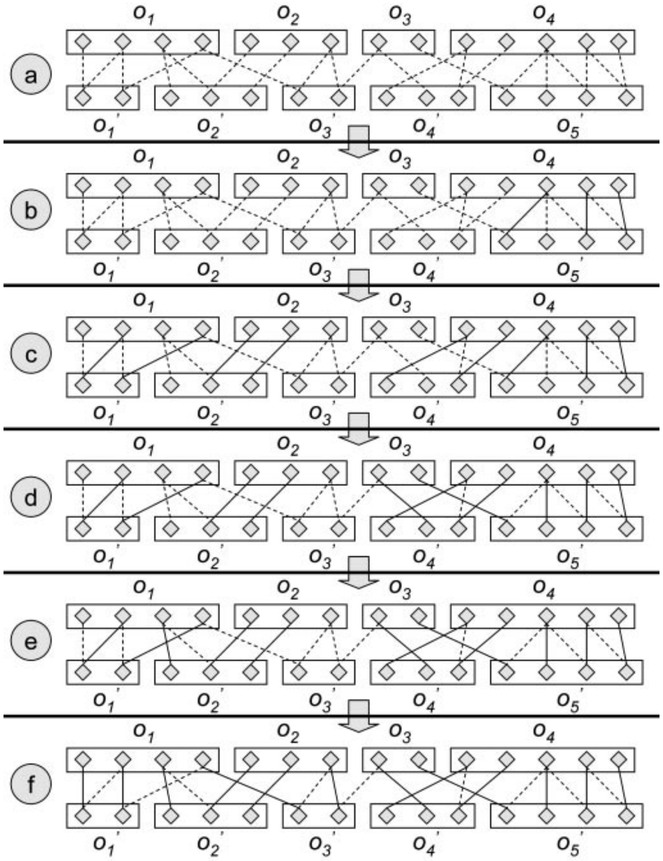

Figure 1. A schematic diagram showing how our algorithm works. In each (**a**, **b**, **c**, **d**, **e** and **f**), the first row represents genes and operons in one genome, and the second row represents genes and operons in another genome. (a) The initial homologous relationship (dashed lines) between the two genomes; each operon is considered as a vertex; (b) the weight of O_4-O_5' is 3 (because the maximum mapping between them is 3), and it is the maximal among all the weights, so they are merged to one operon group, where the solid lines represent orthologous relationship, and this operon group becomes a new vertex; (c) the weights of O_1-O_1' , O_2-O_2' and O_4'-O_4O_5' are 2; they are merged to operon groups and become the new vertices; (d) the weight of O_3-$O_4'O_4O_5'$ are 2; they are merged into one operon group; it should be noted that when the maximum mapping is re-calculated, one pair of orthologues between O_4 and O_5' has been re-predicted; the new prediction is more accurate when all the four operons are considered, which represents a correcting mechanism in this algorithm; (e) O_1O_1' and O_2O_2' are merged into one operon group; (f) O_3' and $O_1O_1'O_2O_2'$ are merged into one operon group; it should be noted that when the maximum mapping is re-calculated, some of the predicted orthologous relationships could be different from that by the previous iteration. At the end two uber-operons in each genome are generated.

determined by M correspond to (part of) uber-operons to be detected. A pseudo code of the algorithm is given in the Supplementary Data.

Uber-operon prediction using multiple reference genomes

For a target genome, our higher-layer algorithm makes the final uber-operon prediction, which is 'maximally' consistent with all the initial predictions by the lower-layer algorithm based on all reference genomes. Generally, the uber-operons predicted based on different reference genomes may be different, because each reference genome might provide different 'reference' information. By effectively combining all these predictions, we could possibly (i) eliminate accidental false predictions due to various reasons, such as false operon prediction in a particular reference genome and (ii) reduce false negative predictions due to the incomplete (reference) information given by any specific reference genome. While more sophisticated 'integration' strategies could be employed, our strategy is to capture the consensus of the initial predictions. This is achieved through a clustering algorithm, described as follows.

For N ($N = 90$ in our study) sets of uber-operon predictions based on N reference genomes, we define a weighted graph G as follows: (i) each predicted operon in the target genome is represented as a vertex; (ii) two vertices have an edge between them if and only if the two corresponding operons are predicted to be in the same uber-operon by at least one of the N predictions; and (iii) the weight of an edge is defined to be the number of times that the two corresponding operons are in the same uber-operon among all the N predictions. In general, G consists of a number of connected sub-graphs. A naïve prediction might predict each such connected sub-graph as a uber-operon. However, we have observed that many of these connected sub-graphs are only intra-linked through 'thin' edges (e.g. edges with weight 1), which we suspect to be accidental predictions due to various reasons (e.g. false operon predictions). To uncover the 'true' uber-operons (with dense linkages), we have used the Markov cluster algorithm (MCL) [http://micans.org/mcl/] (25) to partition G into a set of non-overlapping subgraphs (or clusters) whose vertices are densely intra-linked. MCL is used because of its previous successes in graph partitioning with similar characteristics to ours [http://micans.org/mcl/lit/#3party] (26–28).

The MCL algorithm simulates random walks on a graph using Markov matrices to determine the transition probabilities among the vertices of the graph (25). By alternating expansion and inflation steps in random walks iteratively, MCL eventually separates a graph into unconnected or loosely connected subgraphs, each of which is densely intra-connected among its vertices. Using a parameter that controls the inflation rate, the MCL algorithm partitions a graph into 'densely' intra-connected subgraphs at different levels of granularity. The inflation rate in MCL varies from 2.0 to 5.0. We have applied the algorithm using four different inflation rates (2.0, 3.0, 4.0 and 5.0) and obtained graph partitions with different levels of granularity. For any fixed inflation rate, we predict the vertices of each partitioned subgraph as a uber-operon. The detailed procedure is given in Figure 2.

In our prediction for *E.coli* K12, we have compared our predicted uber-operons using different inflation rates, 2.0, 3.0, 4.0 and 5.0 (Table 1) with KEGG pathways, EcoCyc regulons and GO annotations (see Results and Discussion), and found that the difference between uber-operon predictions by using different inflation rates is small. There are two possible reasons: (i) MCL has indeed captured some intrinsic 'cluster' information in the graph, so it is not very sensitive to the inflation rates. A similar observation is also made for a recent study on accurate orthologs predictions, using MCL (23). (ii) Our comparison is against biological processes at different levels, including pathways, super- and sub-pathways (15). Hence a slight over- or under-prediction of uber-operons may not quite be reflected by such comparisons. We have chosen uber-operon predictions at the inflation rate = 5.0 as our default prediction.

The time required to make uber-operon prediction in one genome against one reference genome is quite short, about 30–60 s on a 2.4 GHz Xeon processor. The time to make uber-operon prediction by using multiple reference genomes depends on the number of reference genomes, the number of operons in the target genome, and the complexity of the graph used as the input to the MCL clustering program. In our case study using 90 reference genomes, it took about 90 min to make the uber-operon prediction for all the 90 genomes on the same Xeon processor (note that the time for running BLAST is not included). The software can be downloaded at http://csbl.bmb.uga.edu/uber/UBER_v0.1.tar.gz.

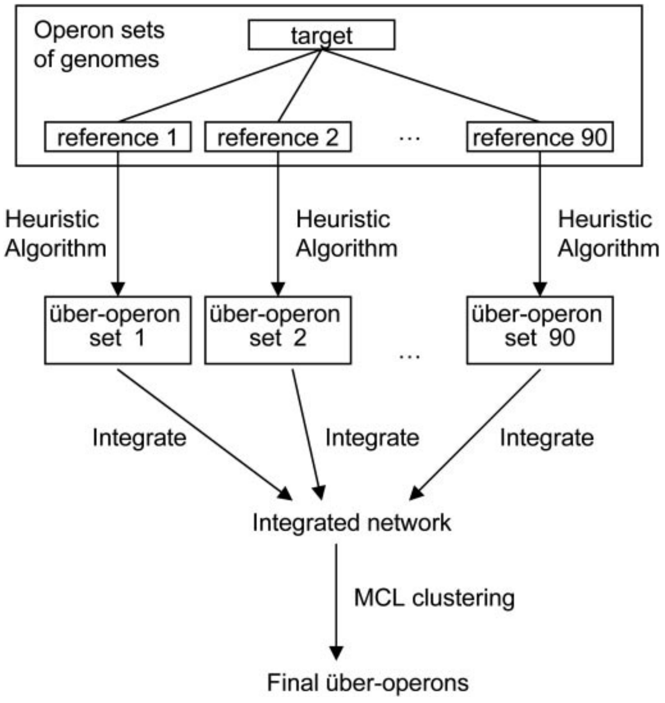

Figure 2. An overview of the uber-operon prediction procedure that consists of preparing operon data, identifying candidate uber-operons using a heuristic algorithm (the lower-layer algorithm), and clustering (the higher-layer algorithm).

RESULTS AND DISCUSSION

We have predicted uber-operons for 91 genomes using the method described in Materials and Methods. For prediction for each genome, we use the other 90 genomes as the references. To evaluate these predictions, we have performed a detailed analysis on the uber-operon predictions in *E.coli*, and assessed the prediction reliability based on known information about *E.coli*. For this genome, we have predicted 158 uber-operons, covering 578 operons and 1830 genes. The size distribution of all the predicted uber-operons in *E.coli* in terms of the number of included operons is given in Figure 3. As we can see, most of the predicted uber-operons contain two or three operons, though a few uber-operons have more than ten operons. It can be checked that this distribution follows a power law distribution.

Analysis of predicted *E.coli* uber-operons

Because there is no dataset of experimentally verified uber-operons, we have used four types of information to assess the soundness of our predicted uber-operons, in terms of both biology and statistics. They are (i) experimentally verified regulons of *E.coli* (15), (ii) experimentally verified pathways in *E.coli* (29), (iii) GO assignments for *E.coli* genes (16), and (iv) bacteria taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html). The first three types of information are used to evaluate the 158 predicted *E.coli* uber-operons, while the bacteria taxonomy data are used to evaluate uber-operon conservation across genomes.

Comparison between predicted uber-operons and regulons. We have collected 153 *E.coli* regulons from the EcoCyc database (15). Our hypothesis is that many of the predicted uber-operons each belong to a regulon. So we use the following approach to compare the consistency between the predicted uber-operons and the known regulons in *E.coli*. We

Table 1. AHMDs of all predicted uber-operons, means and standard deviations (SD) of AHMDs of randomly combined operons, and their corresponding *Z*-scores, for the known pathways and regulons, and ASgo of all predicted uber-operons, means and SD of *ASgo* of randomly combined operons and their corresponding *Z*-scores, for the known GO terms

IR[a]	PathAHMD[b]	RandAHMD(sd)[c]	*Z*-score[d]	RegAHMD[e]	RandAHMD(sd)[f]	*Z*-score[g]	ASgo[h]	RandASgo(sd)[i]	*Z*-score[j]
2.0	0.098	0.066 (0.0058)	5.637	0.125	0.078 (0.0093)	4.979	3.419	2.861 (0.079)	7.102
3.0	0.107	0.082 (0.0063)	3.991	0.166	0.104 (0.011)	5.76	3.511	2.864 (0.069)	9.378
4.0	0.112	0.085 (0.0069)	3.93	0.166	0.107 (0.011)	5.173	3.509	2.850 (0.068)	9.691
5.0	0.115	0.085 (0.0071)	4.091	0.159	0.110 (0.012)	4.145	3.561	2.855 (0.074)	9.579

Four different inflation values (2.0, 3.0, 4.0 and 5.0) in MCL were tested in our experiments.

[a] Inflation rate.

[b] $\mathrm{AHMD}_p(U)$, calculated using formula (3).

[c] Average $\mathrm{AHMD}_p(U')$ for 100 sets of pseudo uber-operons, and its SD, calculated by formula (3) for randomly generated uber-operons.

[d] Z-score for $\mathrm{AHMD}_p(U)$, calculated using formula (5).

[e] $\mathrm{AHMD}_R(U)$, calculated using formula (3).

[f] Average $\mathrm{AHMD}_R(U')$ for 100 sets of pseudo uber-operons, and its standard deviation, calculated by formula (3) for randomly generated uber-operons.

[g] Z-score for $\mathrm{AHMD}_R(U)$, calculated using formula (5).

[h] *ASgo* for the predicted uber-operons, calculated using formula (6).

[i] Average *ASgo* for 100 sets of pseudo uber-operons, and its SD.

[j] Z-score for *ASgo*, calculated using formula (5).

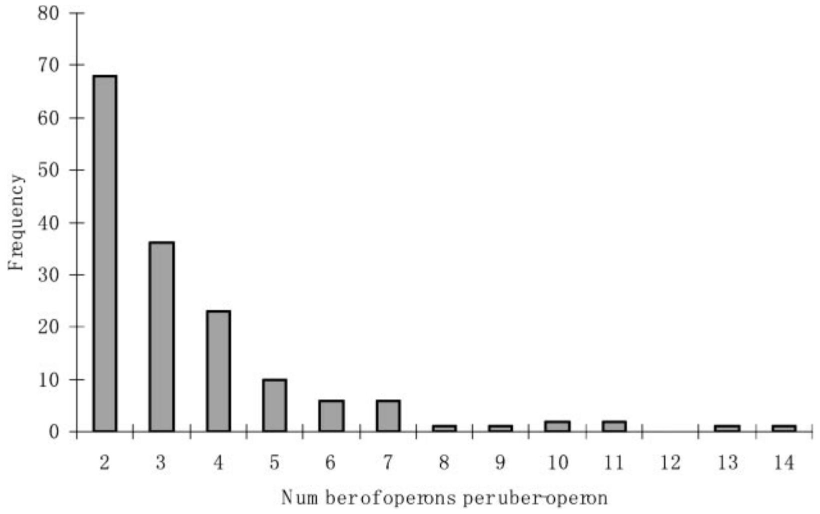

Figure 3. Frequency distribution of the number of operons in a uber-operon in *E.coli*. A total of 157 predicted uber-operons in *E.coli* were used. One uber-operon containing 28 operons was not included.

understand that both the uber-operon predictions and known regulons represent only a fraction of all the uber-operons and regulons in the genome, due to the possible incompleteness of our prediction and experiments. So we have taken this into consideration in our analysis. The basic idea of our comparison is given as follows [some of the ideas have been used for a different application (1)].

Let $A = \{a_i\}$ be a gene list and $P = \{p_i, 1 \leq i \leq m\}$ and $Q = \{q_j, 1 \leq j \leq n\}$ be its two partitions. The matching degree ($\mathrm{MD}_{i,j}$) between p_i and q_j is defined as:

$$\mathrm{MD}_{i,j} = \frac{|\, p_i \cap q_j \,|}{|\, p_i \cup q_j \,|} \qquad \mathbf{1}$$

and the highest matching degree (HMD) achieved by Q for p_i is defined as:

$$\mathrm{HMD}_{p_i} = \max_{j=1}^{n} \left(\frac{|\, p_i \cap q_j \,|}{|\, p_i \cup q_j \,|} \right) \qquad \mathbf{2}$$

The average highest matching degree (AHMD) achieved by Q for P is defined as:

$$\mathrm{AHMD}_P = \frac{\sum_{i=1}^{m} \mathrm{HMD}_{p_i}}{m} \qquad \mathbf{3}$$

The matching degree ($\mathrm{MD}_{i,j}$) gives the similarity between two subsets: p_i and q_j. The HMD for p_i (HMD_{p_i}) gives the subset in Q that achieves the highest similarity with p_i. The AHMD measures the similarity between P and Q. In our analysis, P represents the available regulons or pathways while Q is the predicted uber-operons. Though some of regulons/pathways may have overlaps, it should not have serious effects on our overall evaluation because of the overlaps in general are small compared to the size of the gene list. We have found that when both P and Q are fully available, we can use a more accurate formula as given in definition (4) to more accurately measure the similarity between P and Q.

$$\mathrm{AHMD} = \frac{\sum_{i=1}^{m} \mathrm{HMD}_{p_i} + \sum_{j=1}^{n} \mathrm{HMD}_{q_j}}{m+n} \qquad \mathbf{4}$$

Note that in this definition (4), AHMD is symmetrical with respect to P and Q.

For each set of the predicted uber-operons U, we calculated the $\mathrm{AHMD}_R(U)$ between U and the known regulons R using definition (3). The $\mathrm{AHMD}_R(U)$ value is 0.159 (Table 1). To assess the statistical significance of this obtained $\mathrm{AHMD}_R(U)$ value, we have calculated its Z-score as follows. We first constructed a set of pseudo uber-operons U', by randomly combining the predicted operons such that the ith pseudo uber-operon has the same number of operons as the ith uber-operon in U. We constructed 100 such sets of pseudo uber-operons, and calculated their $\mathrm{AHMD}_R(U')$ values. The Z-score of $\mathrm{AHMD}_R(U)$ is computed as

$$Z_R(U) = \frac{|\, \mathrm{AHMD}_R(U) - \overline{\mathrm{AHMD}_R(U')} \,|}{\sigma_{\mathrm{AHMD}_R(U')}} \qquad \mathbf{5}$$

with $\overline{\mathrm{AHMD}_R(U')}$ being the average $\mathrm{AHMD}_R(U')$ value and $\sigma_{\mathrm{AHMD}_R(U')}$ the standard deviation. We obtain a Z-score 4.091 for $\mathrm{AHMD}_R(U) = 0.159$, indicating that the matching between the predicted uber-operons and the known regulons is highly significant.

Comparison between predicted uber-operons and pathways. We have carried out a similar comparison between the predicted uber-operons (denoted as U) and all the known pathways (denoted as P) in *E.coli* as given in KEGG (29), and calculated the $\mathrm{AHMD}_P(U)$ value and its Z-score, using the same procedures outlined in (A). The value of $\mathrm{AHMD}_P(U)$ is 0.115, and its Z-score is 4.091 (Table 1). This result again suggests that the matching between the predicted uber-operons and known pathways is highly significant.

We have also assessed the matching between known regulons and pathways by calculating $\mathrm{AHMD}_P(R)$.

The AHMD$_P(R)$ is 0.090, slightly smaller than AHMD$_P(U)$. This seems to make good sense because the uber-operons cover not only the operons that are co-regulated but also the operons that work together, say, in the same pathway, while being regulated possibly by different mechanisms. These two similar AHMD$_P$'s indicate the relationship between genes in the same uber-operon is at least as tight as genes in the same regulon.

Relationship between GO assignments and predicted uber-operons. We have assessed the statistical significance of the predicted uber-operons in terms of their GO assignments. Among the three levels of GO functionalities, namely, molecular function, biological process and cellular component, we have used GO's biological processes to compare genes assigned to the same uber-operon. The GO term assignments for *E.coli* were retrieved from Integr8 (http://www.ebi.ac.uk/integr8/EBI-Integr8-HomePage.do). We have previously developed a method for comparing two GO biological processes (1). For two genes g_1 and g_2, we define d_{g_1,g_2} as the similarity score between their GO biological processes, as defined in (1). We then measured the overall consistency of GO assignments for the genes in a predicted uber-operon using the following formula.

$$S_{go} = \frac{1}{L} \sum_{i=1}^{r} \sum_{j=i+1}^{r} \sum_{k=1}^{s_i} \sum_{l=1}^{s_j} d_{k,l} \qquad \mathbf{6}$$

where L is the total number of gene pairs across operons in the uber-operon, r is the number of operons in the uber-operon, s_i and s_j are the numbers of genes in the ith operon and jth operon of the uber-operon, respectively, and $d_{k,l}$ is the similarity score for the kth genes in ith operon and lth gene in jth operon. We have calculated the average S_{go} for all the predicted uber-operons in *E.coli*, denoted as AS_{go}, as,

$$AS_{go} = \frac{1}{n} \sum_{i=1}^{n} S_{go}(U_i), \qquad \mathbf{7}$$

where n is the number of uber-operons in the genome, and $S_{go}(U_i)$ is S_{go} for ith uber-operon. We have obtained $AS_{go} = 3.561$. For Z-score estimation, we have calculated the AS_{go} values for 100 pseudo uber-operons defined in (A), and obtained a Z-score 9.579 for $AS_{go} = 3.561$, indicating that the similarity among the functionalities of genes from the same uber-operons across all predicted uber-operons are highly significant. As a reference, we have also calculated AS_{go} for all known *E.coli* regulons, and obtained $AS_{go} = 4.32$. The similar values between the two AS_{go} indicate that the functional similarity among genes from the same uber-operon is quite comparable to that among genes from the same regulon.

Comparison between predicted uber-operons and bacteria taxonomy. Note that for each predicted set of uber-operons between two genomes, we also get a (possibly incomplete) mapping of orthologous genes between the two genomes. Clearly, the correctness of the orthologous gene mapping reflects the 'correctness' of our predicted uber-operons. Here we demonstrate that the genomic distances measured by uber-operons and their associated orthologous genes are generally consistent with the bacteria taxonomy, which provides an indirect evidence that our predicted orthologous genes are largely correct, and our predicted uber-operons are generally in agreement with our knowledge about the evolutionary history of the involved bacterial genomes.

We have calculated the AHMD between two genomes using definitions (1–4), in which $|p_i \cap q_j|$ in (1) represents the number of orthologous gene pairs between uber-operons p_i of genome G_1 and q_j of genome G_2, and $|p_i \cup q_j| = |p_i| + |q_j| - |p_i \cap q_j|$. We have used this AHMD, denoted as AHMD$_{i,j}$, for the ith and jth genomes, as a similarity measure between genomes, reflected by the predicted orthologous genes.

To compare AHMD$_{i,j}$ with the bacteria taxonomy data from NCBI, we have defined a similarity score based on the taxonomy tree $S_{i,j}$ to be the depth of the nearest common parent of ith and jth genomes in the taxonomy tree. Hence the bigger $S_{i,j}$ is, the closer the two bacteria. We then assessed the relationship between these two similarity measures using a linear regression approach, and obtained the following regression equation based on the 4095 (91∗90/2) similarity measures for all pairs of genomes.

$$\text{AHMD} = 0.0128{*}\text{S} + 0.0246 \qquad \mathbf{8}$$

We have obtained the following statistics for this derived relationship: the squared correlation coefficient (coefficient of determination) $r^2 = 0.38$, F statistics F-ratio $= 2503$, and the corresponding significance level P-value $<10^{-308}$. All these numbers indicate that the correlation between these two similarities is highly significant.

We have also assessed the statistical significance of this correlation measure between the uber-operon sets and the taxonomies. We have first constructed a pseudo uber-operon set for each genome using the approach outlined in (A), calculated the AHMD values for each pair of genomes based on their pseudo uber-operon sets, and then applied the linear regression analysis. We have repeated this procedure for 100 times, and obtained the average squared correlation coefficient $r^2 = 0.258$, and a Z-score $= 9.28$ for the squared correlation coefficient ($r^2 = 0.38$). We have noticed that the squared correlation coefficients for the pseudo uber-operon sets are not very low. This can be partially explained by that the matching degree between two uber-operons is mainly contributed by their large-sized operons. For example, let o_i be the operons in the ith genome whose size dominates all the other operons, (e.g. the operon consisting of ribosome genes), then the mapping degree for the (true or pseudo) uber-operon containing o_i is mainly determined by o_i; and if o_i is well conserved so that its corresponding operons also dominate (in terms of the operon size) in other genomes, then the highest mapping degree for the (true or pseudo) uber-operon containing o_i is always achieved by the (true or pseudo) uber-operon containing o_i's corresponding operon. This indicates that the operons are also conserved across genomes, but the higher squared correlation coefficient for our predicted uber-operons and the high Z-score strongly suggest that the predicted uber-operons have captured a higher-level conserved genomic structure than operons.

In summary, these analyses have shown that our predicted uber-operons are biologically and statistically meaningful.

A detailed list of 158 predicted uber-operons in *E.coli*, in terms of its component operons, genes and their functions, is provided in Supplementary Table S2. As we can see from Supplementary Table S2, numerous predicted uber-operons contain ABC transporter systems, which is consistent with the KEGG where many pathways contain ABC transporter systems. Some of the predicted uber-operons are not associated with any known KEGG pathways, which might indicate that they belong to pathways that are yet to be elucidated. For instance, two operons containing *csg*A, *csgB*, *csgD*, *csgE*, *csgF* and *csgG* have been predicted to form a uber-operon. These genes have not been previously reported to be involved in any known KEGG pathway, but their genes are known to belong to the same regulon based on known EcoCyc regulons. We expect such predicted uber-operons, particularly the ones not known to belong to the same regulons, will provide a highly useful information source for discovery of novel pathways and regulons.

Case studies: detailed analyses of three examples of uber-operons

We now further demonstrate the quality of the predictions by providing detailed analysis of three predicted uber-operons, which are involved in the flagellar system, tricarboxylic acid (TCA) cycle and sulfur metabolism, respectively. These examples highlight the possibility of using uber-operon prediction for elucidation of regulons and the component genes of pathways.

Flagellar assembly. The bacteria flagellum is the motor organelle for propulsion, driven by the transmembrane proton motive force. The full function of flagella requires the expression of more than 50 genes, including structural genes, chemotaxis-related genes, and possibly other related genes (30). We have predicted one uber-operon consisting of 54 genes from 10 operons.

Among the 54 genes, 30 genes (*flgB*, *flgC*, *flgD*, *flgE*, *flgF*, *flgG*, *flgH*, *flgI*, *flgJ*, *flgK*, *flgL*, *flhE*, *flhA*, *flhB*, *fliA*, *fliD*, *fliS*, *fliF*, *fliG*, *fliH*, *fliI*, *fliJ*, *fliK*, *fliL*, *fliM*, *fliN*, *fliO*, *fliP*, *fliQ*, *fliR*) are known to be in the pathway of the flagellar assembly according to the KEGG database, 12 genes (*cheZ*, *cheY*, *cheB*, *cheR*, *tap*, *tar*, *cheW*, *cheA*, *motB*, *motA*, *flhC*, *flhD*) are known to be in the chemotaxis pathway, and the remaining 12 genes are involved in cell division and other biological processes, based on their GO annotation. In (14), Lathe *et al*. used four reference genomes to predict uber-operons and identified flagellar uber-operon genes. While we found some level of agreement between our uber-operon prediction and the corresponding uber-operon in (14), we noticed that a number of genes in our uber-operon, annotated as possibly flagellar-related by GO, are not reported by Lathe *et al*., such as *fliZ* and *fliT*. For instance, *fliZ* is annotated as the putative regulatory gene on *fliA*. Interestingly, a cell division related gene, *minD*, seemingly not related to the flagellar system, is found both in our predicted uber-operon and in the Nebulon system (31). The association of the cell division and the flagellar system clearly warrants further experimental investigation.

TCA cycle. TCA cycle is a common pathway in mitochondria. It starts with oxidizing acetyl CoA, which is the product from the oxidative decarboxylation of pyruvate, and goes through a ten-step reaction process that yields energy and CO2. We have predicted one uber-operon consisting of three operons covering nine genes: *sdhC*, *sdhD*, *sdhA*, *sdhB*, *b0275*, *sucA*, *sucB*, *sucC*, *sucD*, eight of which, except for *b0275*, are known to be involved in the TCA cycle pathway, as reported in KEGG. Further analysis indicates that these eight genes are predicted to be in one operon in six other genomes, i.e. *Candidatus Blochmannia floridanus*, *Coxiella burnetii* RSA 493, *Legionella pneumophila* str. Paris, *Neisseria meningitidis* MC58, *Photobacterium profundum* SS9 and *Vibrio vulnificus* YJ016. The functionality of gene *b0275* is unknown at this point. Our BLAST search did not reveal any homologous genes in other genomes, suggesting that it might represent a unique gene involved in the *E.coli* TCA process. This uber-operon does not include other genes known to be in the pathway, such as *frdD*, which encodes fumarate redutase. This indicates that the gene rearrangement might have occurred locally, i.e. within succinate related genes.

Sulfur metabolism. Sulfur metabolism is one of the most important components in energy metabolism in *E.coli*, which consists of synthesis and catabolism of the sulfur-containing amino acids, such as cysteine and methionine (32). Our predicted uber-operon contains two operons covering seven genes, six of which are involved in the sulfur metabolism pathway, i.e. *cysC*, *cysN*, *cysD*, *cysH*, *cysI* and *cysJ*. *ygbE* is annotated as a putative cytochrome oxidase subunit. We have not been able to find its homologous gene in the corresponding uber-operons in other genomes, and so far no literatures have suggested that cytochrome oxidase is involved in the process of this metabolism. This seemingly displaced gene could possibly be explained by the 'selfish operon' (33) hypothesis. In the 'selfish operon' model, an operon deletes its un-used genes through horizontal transfers, and only useful genes are retained. The gene *ygbE* may represent a trace of incomplete evolution, and the *cysCNDHIJ* genes may represent the 'useful' genes suggested by the 'selfish operon' hypothesis. This in turn indicates that our method could tolerate some level of noise, i.e. irrelevant genes in some operons.

Novel uber-operons?

Our prediction includes a set of putative uber-operons, which haven not been confirmed by any known pathways or regulons, though they are highly statistically significant. GO assignments cannot reveal much clue about the biological processes in which the involved genes participate, either. Supplementary Table S3 summarizes this set of predicted uber-operons and the possible biological processes in which they might be involved, according to individual gene annotations. To show what possible biological processes these putative uber-operons might suggest biologically, we provide two examples to show how the uber-operon prediction could possibly be further explored.

Membrane proteins. One of the predicted uber-operons contains six genes (*yqjA*, *yqjB*, *yqjC*, *yqjD*, *yqjE* and *yqjK*) from two operons in *E.coli*, and has its corresponding uber-operons predicted in a few reference genomes, including

Erwinia carotovora subsp. atroseptica SCRI1043 and *Yersinia pestis biovar Medievalis* str 91 001. All these six genes in *E.coli* have their orthologous genes belong to one operon in *Y.pestis biovar Medievalis* str. 91 001, and have orthologous genes that belong to two operons in *Erwinia carotovora subsp. atroseptica* SCRI1043. The conservation of these genes indicates the significance of this novel uber-operon. The genes of this uber-operon encode integral membrane proteins, although their detailed collective functionality is unknown to date (Supplementary Table S4).

Rhs-family related proteins. The *Rhs* family consists of at least five *Rhs* elements in *E.coli*, with the most prominent *Rhs* component containing extended repeated regions and often participating in ligand-binding processes in the cell surface (34). Our uber-operon prediction indicates that gene b0499 and gene b1456, belonging to two different operons in *E.coli,* have their predicted orthologous genes SF0267 and SF0268 belong to the same operon in *Shigella flexneri 2a* str. 301. We have also observed that gene b0499 and gene b3428, belonging to two different operons in *E.coli,* have their predicted orthologous genes CV1238 and CV1239 belong to the same operon in *Chromobacterium violaceum* ATCC 12 472. All these genes are annotated as *Rhs*-family proteins or putative *Rhs*-family proteins in the NCBI microbial genome database. The initial prediction of two uber-operons, one based on *S.flexneri 2a* str. 301 and the other based on *C.violaceum* ATCC 12 472, respectively, ultimately leads to the final prediction of a combined uber-operon, which contains three operons including three genes b0499, b1456 and b3428. Unlike the previous example, this putative uber-operon does not seem to have a corresponding prediction in other genomes. While we do not rule out the possibility of a false prediction, we do suspect that these genes work together as a unit as their proteins are mostly annotated as the *Rhs* family related (Supplementary Table S5). We believe that this prediction warrants further experimental investigation.

The accuracy of the predicted uber-operons

A number of factors contribute the accuracy of our uber-operon prediction. For one, since our uber-operon prediction relies on the operon prediction, the accuracy of our uber-operon prediction depends on the accuracy of operon prediction. Fortunately, the best operon prediction programs have reached a prediction accuracy level at 80–90% (2,6–9) and this prediction accuracy will definitely continue to improve. Another key contributing factor is how well the reference genomes are selected in a non-biased way. In the clustering step of our higher-level prediction algorithm, each reference genome is considered to contribute equally to our final prediction of uber-operons; hence over- or under-representation of any group of genomes could possibly bias our prediction results towards some direction. Another complicating factor that could contribute to the prediction accuracy comes from horizontal gene transfers, which may change the composition of an operon in an 'random' way, and hence possibly affect the accuracy of our uber-operon predictions. So it is necessary to make additional studies to identify such gene transfers and remove them from consideration during our uber-operon prediction.

CONCLUDING REMARKS

We have developed a new framework for identification of uber-operons, which represent a class of genomic structure yet to be fully investigated, and record the footprints of operon evolution. Uber-operons may prove to be highly useful for elucidation of biological pathways. Our analyses on the predicted uber-operons, in terms of the statistical significance, evolutionary conservation and functional relatedness among their component genes, have indicated that this concept is well founded, though further investigation and refinement might be needed. We can see a number of important applications of our uber-operon prediction capability. (i) The component genes of a predicted uber-operon could suggest possible candidate genes in a particular biological process, such as a pathway, which has higher gene coverage than operons. (ii) Many of the predicted uber-operons seem to be parts or even whole regulons, based on our analyses. Hence, this could possibly lead to an effective way for regulon prediction. As of today there is no publicly available computer program for regulon prediction. Uber-operon-based approach could become the first general approach to regulon prediction. (iii) If we consider genes in an operon as tightly coupled working unit in a biological process, uber-operons might provide lists of genes that are less tightly coupled, possibly including genes responsible for different biological functions in a complex biological network. Specifically, a uber-operon might include genes involved in both metabolic and regulatory functions, providing richer information for elucidation of complex biological networks.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online

ACKNOWLEDGEMENTS

The authors thank Drs L. Cai, Z. Su and P. Wan for helpful discussions during the course of this project. This work was supported 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). The work is also supported, in part, by National Science Foundation (NSF/DBI-0354771, NSF/ITR-IIS-0407204). G. Li's work is also supported by NSFC of China under Grant No. 10271065 and No. 60373025. Funding to pay the Open Access publication charges for this article was provided by NSF/ITR-IIS-0407204.

Conflict of interest statement. None declared.

REFERENCES

1. Wu,H., Su,Z., Mao,F., Olman,V. and Xu,Y. (2005) Prediction of functional modules based on comparative genome analysis and Gene Ontology application. *Nucleic Acids Res.*, **33**, 2822–2837.
2. Chen,X., Su,Z., Dam,P., Palenik,B., Xu,Y. and Jiang,T. (2004) Operon prediction by comparative genomics: an application to the *Synechococcus* sp. WH8102 genome. *Nucleic Acids Res.*, **32**, 2147–2157.
3. Su,Z., Olman,V., Mao,F. and Xu,Y. (2005) Comparative genomics analysis of NtcA regulons in cyanobacteria: regulation of nitrogen

assimilation and its coupling to photosynthesis. *Nucleic Acids Res.*, **33**, 5156–5171.

4. Su,Z., Dam,A., Chen,X., Olman,V., Jiang,T., Palenik,B. and Xu,Y. (2003) Computational inference of regulatory pathways in microbes: an application to the construction of phosphorus assimilation pathways in *Synechococcus* WH8102. *Genome Inform Ser Workshop Genome Inform.*, **14**, 3–13.
5. Dam,P., Su,Z., Olman,V. and Xu,Y. (2004) *In silico* construction of the carbon fixation pathway in *Synechococcus* sp. WH8102. *J. Biol. Syst.*, **12**, 97–125.
6. Chen,X., Su,Z., Xu,Y. and Jiang,T. (2004) Computational prediction of operons in *Synechococcus* sp. WH8102. *Genome Inform Ser Workshop Genome Inform.*, **15**, 211–222.
7. Price,M.N., Huang,K.H., Alm,E.J. and Arkin,A.P. (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res.*, **33**, 880–892.
8. Westover,B.P., Buhler,J.D., Sonnenburg,J.L. and Gordon,J.I. (2005) Operon prediction without a training set. *Bioinformatics*, **21**, 880–888.
9. Romero,P.R. and Karp,P.D. (2004) Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases. *Bioinformatics*, **20**, 709–717.
10. Kremling,A., Jahreis,K., Lengeler,J.W. and Gilles,E.D. (2000) The organization of metabolic reaction networks: a signal-oriented approach to cellular models. *Metab. Eng.*, **2**, 190–200.
11. Wagner,R. (2000) *Transcription Regulation in Prokaryotes.* Oxford University Press, Oxford, UK.
12. Mao,F., Su,Z., Olman,V., Dam,P., Liu,Z. and Xu,Y. (2006) Mapping of orthologous genes in the context of biological pathways: an application of integer programming. *Proc. Natl Acad. Sci. USA*, **103**, 129–134.
13. Olman,V., Peng,H., Su,Z. and Xu,Y. (2004) Mapping of microbial pathways through constrained mapping of orthologous genes. *Proc IEEE Comput Syst Bioinform Conf.*, pp. 363–370.
14. Lathe,W.C.,IIIrd, Snel,B. and Bork,P. (2000) Gene context conservation of a higher order than operons. *Trends. Biochem. Sci.*, **25**, 474–479.
15. Keseler,I.M., Collado-Vides,J., Gama-Castro,S., Ingraham,J., Paley,S., Paulsen,I.T., Peralta-Gil,M. and Karp,P.D. (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res.*, **33**, D334–D337.
16. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
17. Jensen,R.A. (2001) Orthologs and paralogs—we need to get it right. *Genome Biol.*, 2(4): INTERACTIONS1002.
18. Koonin,E.V. (2001) An apology for orthologs—or brave new memes. *Genome Biol.*, 2(4): COMMENT1005.
19. Petsko,G.A. (2001) Homologuephobia. *Genome Biol.*, 2(2): COMMENT1002.
20. Mushegian,A.R. and Koonin,E.V. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl Acad. Sci. USA*, **93**, 10268–10273.
21. Wall,D.P., Fraser,H.B. and Hirsh,A.E. (2003) Detecting putative orthologs. *Bioinformatics*, **19**, 1710–1711.
22. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
23. Wu,H., Mao,F., Olman,V. and Xu,Y. (2005) Accurate prediction of orthologous gene groups in microbes. *Proc. IEEE Comput. Syst. Bioinform. Conf.*, pp. 73–79.
24. Bondy,J.A. and Murty,U.S.R. (1976) *Graph Theory with Applications.* Macmilliam Press Ltd., London, UK & Elsevier Science Publishing Co., Inc., NY, US.
25. Dongen,S.v. (2000) A cluster algorithm for graphs. Technical Report INS-R0010, National Research Institute for Mathematics and Computer Science in the Netherlands. Amsterdam, Netherlands.
26. Enright,A.J., Kunin,V. and Ouzounis,C.A. (2003) Protein families and TRIBES in genome sequence space. *Nucleic Acids Res.*, **31**, 4632–4638.
27. Li,L., Stoeckert,C.J.,Jr and , Roos,D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.
28. Pereira-Leal,J.B., Enright,A.J. and Ouzounis,C.A. (2004) Detection of functional modules from protein interaction networks. *Proteins*, **54**, 49–57.
29. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
30. Chilcott,G.S. and Hughes,K.T. (2000) Coupling of flagellar gene expression to flagellar assembly in *Salmonella enterica* serovar typhimurium and *Escherichia coli*. *Microbiol. Mol. Biol. Rev.*, **64**, 694–708.
31. Janga,S.C., Collado-Vides,J. and Moreno-Hagelsieb,G. (2005) Nebulon: a system for the inference of functional relationships of gene products from the rearrangement of predicted operons. *Nucleic Acids Res.*, **33**, 2521–2530.
32. Sekowska,A., Kung,H.F. and Danchin,A. (2000) Sulfur metabolism in *Escherichia coli* and related bacteria: facts and fiction. *J. Mol. Microbiol. Biotechnol.*, **2**, 145–177.
33. Lawrence,J. (1999) Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Curr. Opin. Genet. Dev.*, **9**, 642–648.
34. Hill,C.W., Sandt,C.H. and Vlazny,D.A. (1994) Rhs elements of *Escherichia coli*: a family of genetic composites each encoding a large mosaic protein. *Mol. Microbiol.*, **12**, 865–871.