Mapping of orthologous genes in the context of biological pathways: An application of integer programming

Fenglou Mao*, Zhengchang Su*†, Victor Olman*, Phuongan Dam*†, Zhijie Liu*, and Ying Xu*†‡

*Computational Systems Biology Laboratory, Biochemistry and Molecular Biology Department, University of Georgia, A110 Life Science Building, 120 Green Street, Athens, GA 30602; and †Computational Biology Institute, Oak Ridge National Laboratory, Oak Ridge, TN 37831

Communicated by Jeffrey L. Bennetzen, University of Georgia, Athens, GA, November 9, 2005 (received for review February 8, 2005)

Mapping biological pathways across microbial genomes is a highly important technique in functional studies of biological systems. Existing methods mainly rely on sequence-based orthologous gene mapping, which often leads to suboptimal mapping results because sequence-similarity information alone does not contain sufficient information for accurate identification of orthology relationship. Here we present an algorithm for pathway mapping across microbial genomes. The algorithm takes into account both sequence similarity and genomic structure information such as operons and regulons. One basic premise of our approach is that a microbial pathway could generally be decomposed into a few operons or regulons. We formulated the pathway-mapping problem to map genes across genomes to maximize their sequence similarity under the constraint that the mapped genes be grouped into a few operons, preferably coregulated in the target genome. We have developed an integer-programming algorithm for solving this constrained optimization problem and implemented the algorithm as a computer software program, P-MAP. We have tested P-MAP on a number of known homologous pathways. We conclude that using genomic structure information as constraints could greatly improve the pathway-mapping accuracy over methods that use sequence-similarity information alone.

genomic structure | operon | regulon | ortholog | pathway mapping

Comparative genome analysis represents a powerful technique for functional inference of genes. Its foundation is the ability to identify homologous (or more specifically, orthologous) genes. Here orthologous genes refer to isofunctional and heterospecic genes (1–3) for practical purposes. Several methods have been developed for mapping orthologous genes through sequence comparison. A popular approach is based on reciprocal BLAST searches, the so-called bidirectional best-hit (BDBH) approach (4). Based on this strategy, Wall *et al*. (5) recently designed a more sophisticated scheme for finding orthologous genes through BLAST searching followed by a more accurate sequence-alignment scheme (e.g., CLUSTALW and PAML). Koonin and co-workers (6) developed a popular method for orthologous gene identification based on the idea of clusters of orthologs groups (COG). Whereas all these approaches have provided useful practical tools for prediction of orthologous genes, their prediction accuracy has been less than optimal, as discussed below. One of the key issues with all these methods is their underlying assumption that sequence similarity alone contains sufficient information for prediction of orthologous gene relationship, which is probably far from being true.

One important piece of information for orthologous gene identification across microbial genomes could come from the genomic structures such as operons (7) and regulons (8). Using such information has proven to be very helpful in orthology mapping in our previous studies (9, 10). Based on this observation, we have developed an algorithm, called P-MAP, for mapping orthologous genes across microbial genomes, in the context of pathway mapping; it is generally known that genes working in the same biological pathway are typically organized in one or a few operons (11, 12) or regulons (13, 14). P-MAP maps orthologous genes by using both sequence similarity and genomic structure information. P-MAP is available at http://csbl.bmb.uga.edu/pmap2/PMAP2.htm for free download by academic users. In addition, a prediction server for P-MAP has been set up at http://csbl.bmb.uga.edu/pmap2 for users who prefer to use the server rather than install and run the code at their own site.

Problems with Existing Methods for Orthologous Gene Mapping

Although we intend to carry out a systematic study to evaluate the prediction accuracy of the popular methods for orthologous gene mapping such as COG, we had to contend with identifying specific examples of mispredictions caused by the lack of experimentally verified information of orthologous gene pairs or groups at a large scale. The examples we show here were collected mainly during our study of mapping phosphorus-assimilation pathways of *Bacillus subtilis* and *Escherichia coli* K12 to *Synechococcus* sp. WH8102 (9). During that study, we noticed that the existing methods, from time to time, give incorrect orthologous mapping results, which motivated the current work. Although we do not have a general estimate on the percentage of incorrect predictions by BDBH and COG, we believe that the level of such predictions is probably quite significant, because we found multiple such examples in just one pathway-mapping prediction. We show a few examples of false predictions collected in our mapping effort of the *pho* regulon across different genomes.

Missing (False-Negative) Prediction. One example of missing prediction involves the gene Bsu2906 (*phoR*) (15) of *B. subtilis* and the gene b0400 (*phoR*) (16) of *E. coli* K12, which are experimentally verified orthologs. However, they are not BDBHs of each other, because Bsu2906 is only the second-best hit of b0400 in *B. subtilis*.

False-Positive Prediction. The *B. subtilis* gene Bsu2530 (*phoH*) and *E. coli* K12 gene b0660 (*ybeZ*) are BDBHs of each other, but they are not orthologs. The correct ortholog of Bsu2530 is b0660 in *E. coli* K12, as reported in refs. 17 and 18. Although considering multidirectional best hit might possibly help to rectify such incorrect predictions, there is simply no generally sound reason to believe why multidirectional best hit should provide the correct prediction of orthologous gene pairs.

Uncertainty in Predictions. One of the issues with COG predictions is that it might have multiple predictions of possible orthologous

Conflict of interest statement: No conflicts declared.

Abbreviations: BDBH, bidirectional best hit; COG, clusters of orthologs groups; IP, integer programming; NCBI, National Center for Biotechnology Information.

[‡] To whom correspondence should be addressed. E-mail: xyn@bmb.uga.edu.

genes. For example, COG0642 has 15 *B. subtilis* genes [Bsu0202 (*ybdK*), Bsu0245 (*ycbA*), Bsu0257 (*ycbM*), Bsu0377 (*yclK*), Bsu1328 (*ykoH*), Bsu1355 (*ykrQ*), Bsu1368 (*ykvD*), Bsu2637 (*yrkQ*), Bsu2906 (*phoR*), Bsu3034 (*ytsB*), Bsu3140 (*kinB*), Bsu3299 (*yvqB*), Bsu3318 (*yvrG*), Bsu3468 (*yvcQ*), and Bsu3961 (*yxdK*)] and 16 *E. coli* K12 genes [b0400 (*phoR*), b0570 (*ybcZ*), b0993 (*torS*), b1129 (*phoQ*), b1609 (*rstB*), b1968 (*yedV*), b2078 (*baeS*), b2219 (*atoS*), b2556 (*yfhK*), b2786 (*barA*), b3026 (*qseC*), b3404 (*envZ*), b3911 (*cpxA*), b4003 (*hydH*), b4112 (*basS*), and b4399 (*creC*)]. These multiple predictions of COG genes make it difficult to use directly in orthologous gene mapping. Another source of uncertainty in orthologous gene predictions using sequence information alone comes from the small differences among the *P* values of the top hits when they are not statistically significant. We find that such examples occur quite often, making it scientifically challenging to figure out which hits represent the true orthologous gene. We believe that such uncertainty and other issues with the current methods could be reduced significantly through application of genomic structural information.

Pathways and Operons

We studied the relationship between genes in a pathway and their operon distributions through an analysis on all 131 known *E. coli* K12 pathways with at least 4 genes in each (19). This set of pathways consists of 758 genes. We have estimated the probability for two genes to be in the same operon by chance versus being from the same pathway. First, we randomly select two genes from the 1,540 *E. coli* K12 genes that are known to be part of the 678 known operons and determine whether they belong to the same operon. We repeated this procedure 1 million times to estimate the probability for it to happen by chance. We then repeated the whole procedure 50 times. The average probability is 0.0020 with an SD 3.78×10^{-5}. We then repeat the same procedure except that each pair of selected genes is from the same pathway. We obtain an average probability of 0.3314 with an SD 4.8×10^{-4}, which is 164.7 times higher than the probability for two arbitrary genes. This procedure clearly shows that genes in the same pathway have a significantly higher tendency to come from the same operons.

Results and Discussion

Based on the idea that we introduced above, we designed an integer-programming (IP) algorithm to resolve the pathway-mapping problem, and we implemented it as the software P-MAP.

Evaluation of P-MAP: Mapping *B. subtilis* Pathways to *E. coli* K12. We have tested the P-MAP program on a number of pathway-mapping problems to evaluate its effectiveness. We mapped experimentally verified pathway models of *B. subtilis* to *E. coli* K12, for which the corresponding *E. coli* K12 pathways and many of its operons have been verified experimentally. This process allows us to compare the mapped pathway models with the experimentally verified pathways. In our applications, we have used only experimentally verified operon and regulon information of *E. coli* K12 as constraints in P-MAP. We classify the mapped genes to three categories: "correctly mapped genes" (CM), "wrongly mapped genes" (WM), and "failed to find any orthologous genes" (FF).

Test 1: Mapping of the biotin biosynthesis I pathway. Experimental studies have shown that the biotin biosynthesis I pathway in *B. subtilis* consists of four genes: Bsu3018 (*bioA*), Bsu3015 (*bioB*), Bsu3016 (*bioD*), and Bsu3017 (*bioF*) (20). Based on the information from the Encyclopedia of *Escherichia coli* K12 Genes and Metabolism (EcoCyc, available at http://ecocyc.org), their orthologous genes in *E. coli* K12 are b0774 (*bioA*), b0775 (*bioB*), b0778 (*bioD*), and b0776 (*bioF*). In addition, we know from the EcoCyc that b0775 (*bioB*), b0776 (*bioF*), and b0778 (*bioD*) are in operon TU00117 (EcoCyc ID), and b0774 (*bioA*) is in operon TU00013 (EcoCyc ID), both of which are regulated by MONOMER-48 (EcoCyc ID) and hence belong to the same regulon.

Using a BDBH method, we can map three of four genes correctly: Bsu3018 (*bioA*), Bsu3015 (*bioB*), and Bsu3016 (*bioD*) to b0774, b0775, and b0778, respectively, by using 10^{-30} as the *e*-value cutoff. BDBH could not identify the orthologous gene of Bsu3017 (*bioF*) in *E. coli* K12 because its BLAST best hit is b3617 (*e* value, 6×10^{-63}), whereas the correct orthologous gene is b0776 (*e* value, 10^{-57}). Gene b3617 is 2-amino-3-ketobutyrate CoA ligase, and bioF is 8-amino-7-oxononanoate synthase. Interestingly, both genes b3617 and b0776 are assigned with the same COG number (COG0156), which indicates that COG cannot distinguish which one is the real orthologous gene. P-MAP maps all four genes correctly.

Test 2: Mapping of the menaquinone biosynthesis pathway. It is known that both *B. subtilis* and *E. coli* K12 have the menaquinone biosynthesis pathway (21, 22). Currently, it is known that the menaquinone biosynthesis pathway in *E. coli* K12 includes the following genes: b3930 (*menA*), b2262 (*menB*), b2261 (*menC*), b2264 (*menD*), b2260 (*menE*), b2265 (*menF*), and b3833 (*ubiE*). The menaquinone biosynthesis pathway in *B. subtilis* has not been as well characterized as the *E. coli* K12 pathway, but it is known (23–25) that Bsu3075, Bsu3077, Bsu3074, and Bsu3078 encode menB, menD, menE, and menF, respectively. No genes have been identified to encode menA, menC, and ubiE in *B. subtilis*. It is also known (21) that Bsu3196 (*dhbC*) is involved in the menaquinone biosynthesis pathway; dhbC is an isochorismate synthase, which is involved in both menaquinone and enterobactin biosynthesis. Based on the incomplete information of this *B. subtilis* pathway, we have mapped the five genes Bsu3075, Bsu3077, Bsu3074, Bsu3078, and Bsu3196 to the *E. coli* K12 genome.

By using a BDBH approach, Bsu3074 (*menE*) and Bsu3078 (menF) could not be mapped correctly to *E. coli* K12. The top five hits of Bsu3074 are b0037 (*e* value, 3×10^{-46}), b1805 (*e* value, 3×10^{-44}), b1701 (*e* value, 5×10^{-41}), b2260 (*e* value, 1×10^{-35}), and b2836 (*e* value, 3×10^{-34}). The best reciprocal BLAST hit is b0037, a putative crotonobetain gene, whereas the true ortholog is b2260, which is ranked number four. Among these five best hits, b1805, b2260, and b2836 are assigned with the same COG number (COG0318), whereas b0037 and b1701 do not have a COG number, which indicates that COG cannot identify the correct orthologous gene of Bsu3074 (*menE*) in *E. coli* K12. Bsu2073 (*menF*) cannot be mapped by the BDBH approach to any *E. coli* K12 gene. Note that Bsu3078 (*menF*), Bsu3196 (*dhbC*), and b0593 (*entC*) are all assigned the same COG number (COG1169), which indicates that COG is not capable of differentiating them, whereas b2265 has no COG number yet.

From EcoCyc we know that b2262 (*menB*), b2261 (*menC*), b2264 (*menD*), b2260 (*menE*), and b2265 (*menF*) are in the same operon (TU00298) in *E. coli* K12. By using this information, P-MAP mapped the *B. subtilis* genes Bsu3075, Bsu3077, Bsu3074, Bsu3078, and Bsu3196 to b2262 (*menB*), b2264 (*menD*), b2260 (*menE*), b2265 (*menF*), and b0593 (*entC*), respectively. Although the first four mapped genes are known to be correct, we don't have any independent information to verify whether b0593 (*entC*) is actually the orthologous gene of Bsu3196 (*dhbC*). However, the mapping seems to make sense because b0593 (*entC*), Bsu3196 (*dhbC*), and Bsu3078 (*menF*) have the same COG number (COG1169).

Test 3: Mapping of the phosphorus-assimilation pathway. The phosphorus-assimilation pathway in *E. coli* K12 is known (16, 26–30) to include the following operons: *phoBR*, *phoAEHU*, *pstABCS*, *phnCDEFGHIJKLMNOP*, and *glpQT*.

The details of the phosphorus-assimilation pathway in *B. subtilis* are not as well characterized as they are in *E. coli* K12. It is known that *phoBR* and *pstABC* genes are present and have been well studied (15). For this *B. subtilis* pathway, the National Center for Biotechnology Information (NCBI) annotation is quite confusing and not consistent with *E. coli* K12. In the NCBI annotation of *B. subtilis*, *yqgGHIJK* genes have the function of ABC transporter and they have been well studied. Recently, the operon was reannotated as *pstACS*, *pstB1*, and *pstB2*, with *pstB1/B2* as two transporters compared to one transporter in *E. coli* K12. *B. subtilis phoP* (NCBI annotation) has the same function of *phoB* in *E. coli* K12. *B. subtilis phoB* (NCBI annotation) is a paralog of the *E. coli* K12 *phoA* gene, which is completely different from *E. coli* K12 *phoB*, and *B. subtilis* also has its own *phoA*, which is the ortholog of *E. coli* K12 *phoA. B. subtilis phoD* doesn't have any homolog in *E. coli* K12, and *phnCDEFGHIJKLMNOP* is not studied in *B. subtilis*. On the basis of a literature search (15, 16, 26–30), we have reannotated *B. subtilis* genes to provide a more clear annotation: Bsu0941 (*phoA*), Bsu2530 (*phoH*), Bsu2906 (*phoR*), Bsu2907 (*phoB*), Bsu2492 (*pstB2*), Bsu2493 (*pstB1*), Bsu2494 (*pstA*), Bsu2495 (*pstC*), Bsu2496 (*pstS*), Bsu0214 (*glpQ*), and Bsu0215 (*glpT*); the corresponding orthologs in *E. coli* K12 are b0383 (*phoA*), b1020 (*phoH*), b0400 (*phoR*), b0399 (*phoB*), b3725 (*pstB*), b3726 (*pstA*), b3727 (*pstC*), b3728 (*pstS*), b2239 (*glpQ*), and b2240 (*glpT*).

By using a BDBH approach with an *e*-value cutoff of 10^{-30}, only three genes can be mapped correctly: Bsu0941 (*phoA*) ⇔ b0383 (*phoA*), Bsu2493 (*pstB1*) ⇔ b3725 (*pstB*), and Bsu0215 (*glpT*) ⇔ b2240 (*glpT*). Bsu2494 (*pstA*), Bsu2495 (*pstC*), Bsu2496 (*pstS*), and Bsu0214 (*glpQ*) cannot be mapped correctly because the *e* values are $>10^{-30}$; hence, the BDBH mappings are not very reliable with these high *e* values. Bsu2906 (*phoR*) cannot be mapped correctly by BDBH because the BLAST hits for b0400 (*phoR*) are Bsu2310 (*e* value, 4×10^{-40}), Bsu2906 (*e* value, 1×10^{-39}), Bsu4037 (*e* value, 1×10^{-38}), Bsu0377 (*e* value, 9×10^{-29}), and Bsu1238 (*e* value, 5×10^{-26}), and the best hit is not the correct ortholog. Bsu2530 (*phoH*) cannot be mapped correctly because the BLAST hits for this gene are b0660 (*e* value, 4×10^{-74}) and b1020 (*e* value, 2×10^{-47}). The correct ortholog is ranked number two. Bsu2907 (*phoB*) cannot be mapped correctly because the true ortholog is ranked number four, and its BLAST hits are b3912 (*e* value, 2×10^{-44}), b3405 (*e* value, 8×10^{-42}), b2079 (*e* value, 3×10^{-41}), b0399 (*e* value, 3×10^{-40}), and b0571 (*e* value, 3×10^{-48}). COG is also not capable of finding all of the orthologs. For example, genes Bsu2907 (*phoB*) and b0399 (*phoB*) are assigned to COG0745, which consists of 14 *E. coli* K12 genes. In *B. subtilis*, 13 genes are assigned to COG0745 also, which means that COG is not sensitive enough to find the correct ortholog.

By using P-MAP, all of these genes can be mapped correctly except Bsu2492 (*pstB2*), which has the same ortholog as Bsu2493 (*pstB1*). The reason for this failure is that P-MAP cannot deal with gene duplication, gene fusion, or gene-split events, but we intend to add this functionality to P-MAP soon. The genes Bsu2494 (*pstA*), Bsu2495 (*pstC*), and Bsu2496 (*pstS*), which are mapped incorrectly by the BDBH approach, are mapped correctly by P-MAP because their orthologs are in the same operon (EcoCyc ID TU00202) of b3725 (*pstB*). Bsu0214 (*glpQ*) cannot be mapped correctly by the BDBH method, whereas the gene *glpT*, which is in the same operon (EcoCyc ID TU00216) of *glpQ*, can be mapped correctly by P-MAP because of the operon information. The most difficult parts in this pathway mapping are phoH and phoB. Bsu2530 (*phoH*) has a wrong BDBH gene in *E. coli* K12 (b0660). b0660 is regulated by a transcriptional factor other than PHOSPHO-PHOB, which regulates many genes in this pathway, including b0400 (*phoR*), b0399 (*phoB*), b3725 (*pstB*), b3726 (*pstA*), b3727 (*pstC*), and b3728 (*pstS*), whereas the correct ortholog [b1020 (*phoH*)] is regulated by PHOSPHO-PHOB. This makes b1020 (*phoH*) more favorable than b0660, hence making the P-MAP mapping correct. It is the same for phoB. The real ortholog b0399 (*phoB*) is hidden in a paralog pool with 14 other genes with the same COG number. With the operon [b0399 (*phoB*) and b0400 (*phoR*) being in the same operon: TU00053] and regulon (PHOSPHO-PHOB) information, P-MAP can perform the correct mapping successfully.

For all of the 20 genes in the three pathways discussed above, the CM, WM, and FF values by BDBH are 9, 6, and 5 compared with the P-MAP values 19, 1, and 0, respectively. Clearly P-MAP outperformed BDBH by a significant margin. By comparing the performance of P-MAP and COG, we found that P-MAP consistently provides more specific and more accurate predictions than COG.

Prediction of the WH8012 Phosphorus-Assimilation Pathway Using Multiple Template Pathways. We now present a prediction of the phosphorus-assimilation pathway in *Synechococcus* sp. WH8102 (WH8102) through mapping pathway models from four other organisms: *B. subtilis* (15), *E. coli* K12 (26–30), *Salmonella typhimurium* (31, 32), and *Synecocystis* PCC 6803 (33). Before our work, very little was known about this particular pathway in this species, although some general knowledge about the phosphorus-assimilation process has been documented (34). To facilitate our pathway mapping, we predicted operons for WH8102 by using the JPOP program. The prediction result is available upon request. The four template pathway models are first constructed through piecing together information extracted from the Kyoto Encyclopedia of Genes and Genomes, EcoCyc, or the MetaCyc Encyclopedia of Metabolic Pathways databases, and then information is collected through a literature search.

Phosphorus-assimilation template pathway in* E. coli *K12. *E. coli* K12 uses two low-affinity P_i transporters, *pitA* and *pitB* (35). When P_i is low, the ABC-type high-affinity P_i transporter system *pst* (*pstSCAB-phoU*) is induced through the activation of the two-component system (*phoR/phoB*). In the absence of P_i and the presence of organic P_n, a P_n transporter system (*phnCDE*) and the relevant metabolic enzyme complex C-P lyase (*phnFGHIJKLMNOP*) is induced to transport a wide spectrum of P_n into the cell and to break them down to P_i (28, 29). *E. coli* K12 also utilizes *sn*-glycerol 3-phosphate by inducing the *ugpBAECQ* operon. *phoE* is a porin in the outer membrane to transport P_i-containing compounds into the periplasmic space (36), where P_i is released by alkaline phosphatase *phoA* (16). All these proteins are coregulated by the two-component signaling system *phoR* and *phoB*, the activity of which is negatively regulated by *phoU.*

***Phosphorus-assimilation template pathway in* S. typhimurium.** This organism also has P_i-transport systems and a two-component regulatory system, although it has only one low-affinity P_i transporter, *pitA* (37). Moreover, it uses only one special form of P_n, 2-aminoethylphosphonate (2-AEP), which is transported by *phnSTUV* and broken down by phnW and phnX (31, 32). In addition, the porin *phoE*, located in the outer membrane, is regulated by *phoB* also.

***Phosphorus-assimilation template pathway in* B. subtilis.** This microbe also has P_i transport system and two-component signaling system (15). Although more genes/operons in its pho regulon have been characterized experimentally (38, 39), neither the P_n transporter nor relevant catabolic enzymes have been identified.

Phosphorus-assimilation template pathway in* Synechosistis sp. *PCC 6803. This cyanobacterium has a two-component signaling system. It is made of the sensor kinase *SphS* and the response regulator *SphR*. Interestingly, the genome encodes two ABC-type P_i transporters, *sphX-pstS1C1A1B1B1′* and *pstS2C2A2B2*, both of which are induced after P_i limitation (33). P_i limitation also induces alkaline phosphatase *phoA* and extracellular nuclease *nucH* and represses the expression of a putative urea transporter, *urtA* (33).

We mapped these template pathways, in the form of collections of operons, to WH8102 and built an initial pathway model for this organism. The detailed mapping results are available from the authors upon request. Because of the way P-MAP uses the operon information, we found that the mapped genes are grouped into a few sets of (coregulated) operons. Although a detailed analysis of the mapped pathway components will be analyzed elsewhere, we highlight a few interesting mapped results here. By using P-MAP, SYNW0948 is mapped from *phoR* of *E. coli* K12 and *S. typhimurium*, although it is not a BDBH cognate of *phoR* in these two genomes. This mapping is clearly due to the fact that SYNW0947, the BDBH cognate of *phoB* in *S. typhimurium*, is in the same operon of SYNW0948. We believe that this mapping is correct, because SYNW0947 and SYNW0948 are BDBH cognates of *sphR* and *sphS* in PCC6803, which is phylogenetically very close to WH8102. This result suggests that even without a template from a closely related species, P-MAP still can correctly find orthologs between distantly related genomes. For a similar reason, SYNW1169 is predicted as the ortholog of *phnC*. In addition, SYNW1112 and SYNW0173 are mapped to *tagH* and *pstB2*, respectively, of *B. subtilis*, all of which were previously unknown.

Cross validation from pho regulon binding-site prediction. We don't have any direct experimental evidence to support our prediction yet. However, a simple prediction of conserved binding motifs across the promoters of the mapped WH8102 operons by P-MAP indicate that a majority of these regions contain consensus motif TTAACCTTXXXTTAACCAT, identified by both the CUBIC (40) and BIOPROSPECTOR (41) programs, which is highly consistent with a known binding site of pho regulon in *E. coli* (26, 27). Although not a direct validation of the predicted orthologous genes, it does provide good evidence for mapped genes of pho regulons from the four other genomes.

Map *E. coli* K12 Pathways to WH8012 Genome. We have mapped all 163 known pathways of *E. coli* K12 in EcoCyc to the genome of WH8102 by using P-MAP. The whole calculation is finished in half an hour on a 2.4-GHz Pentium IV CPU. All prediction results are available upon request. We also mapped these *E. coli* K12 pathways to 143 other genomes for which we have predicted operons; the prediction results are available on request. Detailed analyses of these mapping results will be published elsewhere.

Materials and Methods

IP Formulation of Pathway-Mapping Problem. The basic idea of our pathway-mapping algorithm is to first find all homologous genes in the target genome for each gene in a template pathway, if there are any; then, we find a one-to-one mapping among the homologous gene pairs so that the mapped genes are grouped into a few operons as much as possible (preferably coregulated operons) and the mapped gene pairs collectively have as high sequence similarity as possible. The gene pairs that optimize these two criteria are then predicted to be orthologous genes.

Mathematically, we formulate this pathway-mapping problem as an optimization problem as follows. Given are a template pathway consisting of *n* genes along with operons and regulons covering the whole target genome, predicted by the JPOP program (42, 43) (available at http://csbl.bmb.uga.edu/downloads/#jpop). Also given are a set of homologous gene pairs between the template pathway genes and genes in the target genome, predicted by BLAST (44). We first introduce a set of variables for the mathematical formulation of our pathway-mapping problem. Let x_{ij} represent the mapping of the *i*th template gene to its *j*th homologous gene in the target genome. $x_{ij} = 1$ if and only if gene *i* is mapped to the *j*th homologous gene in the target genome; otherwise, $x_{ij} = 0$. Let y_k represent the *k*th operon in the target genome. $y_k = 1$ if and only if one of the template genes is mapped to a gene in the *k*th operon; otherwise, $y_k = 0$. We call an operon with $y_k = 1$ an active operon, and we call a set of regulons active if each regulon has at least one active operon and no two active regulons share a common active operon. Let u_l be the *l*th regulon. $u_l = 1$ if and only if the *l*th regulon is active; otherwise, $u_l = 0$. Let Z_{lk} represent a link between the *k*th operon and the *l*th regulon [note that an operon could belong to multiple regulons (14)]. $Z_{lk} = 1$ if and only if both the *k*th operon and the *l*th regulon are active; otherwise, $Z_{lk} = 0$. We call such a link an active link. Our goal is to find an assignment of 0/1 to $\{x_{ij}\}$, $\{y_k\}$, $\{Z_{lk}\}$, and $\{u_l\}$, so the following objective function is minimized:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}H_{ij} + \sum_{k=1}^{p} y_k S_k + \sum_{l=1}^{r}\sum_{k=1}^{p} z_{lk}B_{lk} + \sum_{l=1}^{r} u_l A_l,$$

under the following constraints

$$\sum_{i=1}^{n} x_{ij} = 1, \text{ for } j = l \ldots m \qquad [1]$$

$$0 \le \sum_{j=1}^{m} x_{ij} \le 1, \text{ for } i = l \ldots n \qquad [2]$$

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{l_k} x_{ij}}{m \times n} \le y_k \le 1, \text{ for } k = l \ldots p \qquad [3]$$

$$\sum_{l=1}^{r} z_{lk} = y_k, \text{ for } k = l \ldots p \qquad [4]$$

$$\frac{\sum_{k=1}^{v_l} z_{lk}}{r \times p} \le u_l \le 1, \text{ for } l = l \ldots r, \qquad [5]$$

where *n*, *m*, *p*, and *r* denote the numbers of genes in the pathway, target genes, operons, and regulons, respectively; *tk* denotes the number of genes in the *k*th operon; *vl* denotes the number of operons in the *l*th regulon; Z_{ij}, a scaling factor, is log(*e* val*ue*) of the BLAST score between template gene *i* and its *j*th homologous gene in the target genome; and S_k is a scaling factor, currently set to be proportional to the prediction reliability for the *k*th operon; the scaling factor B_{lk} is set to be proportional to the predicted probability of the *k*th operon belonging to the *l*th regulon; and A_l is set to be proportional to the prediction reliability of the *l*th regulon. A short explanation of the linear constraints shown in Eqs. **1–5** is given as follows. Constraint **1** indicates that each gene in the template pathway must be mapped to exactly one gene; pathway genes without homologs are removed in a preprocessing step. Constraint **2** indicates that each target gene can be mapped from at most one template gene. Constraint **3** guarantees that if any gene of the *k*th operon is mapped from one of the template genes, y_k will have a value of 1. Constraint **4** guarantees that if an operon is active, then exactly one regulon containing the operon is active. Constraint **5** guarantees that if a regulon has an active link, then the regulon should be active. This formulation simultaneously guarantees that mapped gene pairs have as high sequence similarity as possible and the mapped genes are grouped into a few operons, preferably coregulated, as much as possible. Fig. 1 provides a schematic illustration of the relationships among the constraints.

This optimization problem can be solved as a linear IP problem. We have implemented a C code for solving the IP problem. The core of the code is an IP solver called COIN-OR (45). For a typical pathway-mapping problem with a few dozen

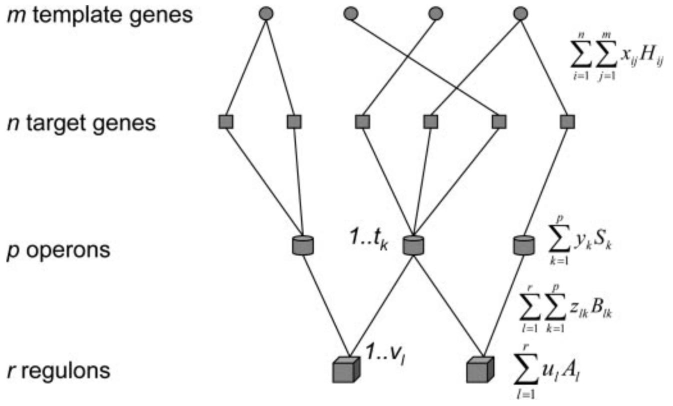

Fig. 1. IP formulation for the pathway-mapping problem. Circles represent genes in the template pathway; rectangles represent candidate genes in the target genome, where target candidate genes are obtained through BLAST search with a specific *e*-value cutoff; cylinders represent operons; and cubes represent regulons. A line between a template gene and target candidate gene represents a BLAST hit. A line between a target candidate gene and an operon indicates that the gene belongs to this operon, and a line between an operon and a regulon indicates that the operon belongs to the regulon.

of genes, the program takes at most a few seconds to minutes to find an optimal mapping on a personal computer (2.4 GHz).

Prediction of Operons: The Prerequisite for Application of P-MAP. We recently developed an effective method called JPOP (43) for operon prediction. The program uses a neural-network approach to find boundaries between operons on the basis of intergenic distances, COG gene functions, and phylogenetic profiles. Its overall prediction accuracy is 83.8%, based on test results on 236 known *E. Coli*. operons (42). The prediction accuracy is improved further when microarray gene-expression data are available and used. The refinement procedure, based on microarray data, is outlined as follows: We adjust the operon boundary (shifting, removing, and creating new ones) based on the consistency between predicted operons and available microarray data. Using this prediction tool, we have predicted operons in 145 sequenced bacterial genomes. These data are available on request.

Although having operon predictions is a prerequisite for running P-MAP, we do not require regulon prediction. If regulons are known through predictions or interpretation of available microarray data, it adds valuable information to our pathway mapping. In such a case, we set their weights accordingly to fully use such information. If regulons are not available, we simply set the corresponding weight factors to a constant that then will have no effect on the final prediction. We have found that one good way to get regulon information is through the prediction of über operons (46) or gene neighborhoods (47), which could be predicted through sequence-based methods. Basically, our framework is general enough to incorporate all such information as constraints of orthologous gene mapping if it is available.

Conclusion

Orthologous gene mapping is an important approach for gene-function prediction, but it remains an unsolved problem. Both the BDBH method and COG have their limitations in making accurate predictions. Through development and application of P-MAP, we have demonstrated that by using genomic structure information as well as sequence-similarity information we can greatly improve the orthology-mapping accuracy over previous methods. We expect that P-MAP will prove to be a highly useful tool for orthology gene mapping across microbial genomes.

This work was supported in part by the U.S. Department of Energy's Genomes to Life program under project "Carbon Sequestration in *Synechococcus* sp: From Molecular Machines to Hierarchical Modeling" (www.genomes-to-life.org). The work was also supported in part by National Science Foundation Grants NSF/DBI-0354771 and NSF/ITR-IIS-0407204.

1. Koonin, E. V. (2001) *Genome Biol*. 2001;2(4):COMMENT1005.
2. Petsko, G. A. (2001) *Genome Biol*. 2001;2(2):COMMENT1002.
3. Jensen, R. A. (2001) *Genome Biol*. 2001;2(8):INTERACTIONS1002.
4. Mushegian, A. R. & Koonin, E. V. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 10268–10273.
5. Wall, D. P., Fraser, H. B. & Hirsh, A. E. (2003) *Bioinformatics* **19,** 1710–1711.
6. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997) *Science* **278,** 631–637.
7. Jacob, F. & Monod, J. (1961) *J. Mol. Biol*. **3,** 318–356.
8. Stephanopoulos, G. N., Aristidou, A. A. & Nielsen, J. (1998) *Metabolic Engineering Principles and Methodologies* (Academic, San Diego).
9. Su, Z., Dam, A., Chen, X., Olman, V., Jiang, T., Palenik, B. & Xu, Y. (2003) *Genome Inform*. **14,** 3–13.
10. Dam, P., Su, Z., Olman, V. & Xu, Y. (2004) *J Biol. Syst*. **12,** 97–125.
11. Dandekar, T., Snel, B., Huynen, M. & Bork, P. (1998) *Trends Biochem. Sci*. **23,** 324–328.
12. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 2896–2901.
13. Eisenberg, M. A. (1973) *Adv. Enzymol. Relat. Areas Mol. Biol*. **38,** 317–372.
14. Chen, Y. M., Zhu, Y. & Lin, E. C. (1987) *Mol. Gen. Genet*. **210,** 331–337.
15. Antelmann, H., Scharf, C. & Hecker, M. (2000) *J. Bacteriol*. **182,** 4478–4490.
16. Wanner, B. L. (1996) in *Escherichia coli and Salmonella: Cellular and Molecular Biology*, eds. Neidhardt, F. C., Curtiss, R. I., III, Gross, C. A., Ingraham, J. L., Lin, E. C. C., Low, K. B., Jr., Magasanik, B., Reznikoff, W., Schaechter, M., Umbarger, H. E. & Riley, M. (Am. Soc. Microbiol., Washington, DC), 2nd ed.
17. Kim, S. K., Makino, K., Amemura, M., Shinagawa, H. & Nakata, A. (1993) *J. Bacteriol*. **175,** 1316–1324.
18. Kazakov, A. E., Vassieva, O., Gelfand, M. S., Osterman, A. & Overbeek, R. (2003) *In Silico Biol*. **3,** 3–15.
19. Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Collado-Vides, J., Paley, S. M., Pellegrini-Toole, A., Bonavides, C. & Gama-Castro, S. (2002) *Nucleic Acids Res*. **30,** 56–58.
20. Bower, S., Perkins, J. B., Yocum, R. R., Howitt, C. L., Rahaim, P. & Pero, J. (1996) *J. Bacteriol*. **178,** 4122–4130.
21. Rowland, B. M. & Taber, H. W. (1996) *J. Bacteriol*. **178,** 854–861.
22. Shineberg, B. & Young, I. G. (1976) *Biochemistry* **15,** 2754–2758.
23. Driscoll, J. R. & Taber, H. W. (1992) *J. Bacteriol*. **174,** 5063–5071.
24. Rowland, B., Hill, K., Miller, P., Driscoll, J. & Taber, H. (1995) *Gene* **167,** 105–109.
25. Lapidus, A., Galleron, N., Sorokin, A. & Ehrlich, S. D. (1997) *Microbiology* **143,** 3431–3441.
26. Makino, K., Shinagawa, H., Amemura, M. & Nakata, A. (1986) *J. Mol. Biol*. **190,** 37–44.
27. Makino, K., Shinagawa, H., Amemura, M., Kawamoto, T., Yamada, M. & Nakata, A. (1989) *J. Mol. Biol*. **210,** 551–559.
28. Haldimann, A., Daniels, L. L. & Wanner, B. L. (1998) *J. Bacteriol*. **180,** 1277–1286.
29. VanBogelen, R. A., Olson, E. R., Wanner, B. L. & Neidhardt, F. C. (1996) *J. Bacteriol*. **178,** 4344–4366.
30. Wanner, B. L. (1994) *Biodegradation* **5,** 175–184.
31. Jiang, W., Metcalf, W. W., Lee, K. S. & Wanner, B. L. (1995) *J. Bacteriol*. **177,** 6411–6421.
32. Kim, A. D., Baker, A. S., Dunaway-Mariano, D., Metcalf, W. W., Wanner, B. L. & Martin, B. M. (2002) *J. Bacteriol*. **184,** 4134–4140.
33. Suzuki, S., Ferjani, A., Suzuki, I. & Murata, N. (2004) *J. Biol. Chem*. **279,** 13234–13240.
34. Palenik, B. & Dyhrman, S. T. (1998) *Phosphorus in Plant Biology: Regulatory Roles in Molecular, Cellular, Organismic, and Ecosystem Processes*, eds. Lynch, J. P. & Deickman, J. (American Society of Plant Physiologists, Rockville, MD).
35. Harris, R. M., Webb, D. C., Howitt, S. M. & Cox, G. B. (2001) *J. Bacteriol*. **183,** 5008–5014.
36. de Cock, H., Overeem, W. & Tommassen, J. (1992) *J. Mol. Biol*. **224,** 369–379.
37. McClelland, M., Sanderson, K. E., Spieth, J., Clifton, S. W., Latreille, P., Courtney, L., Porwollik, S., Ali, J., Dante, M., Du, F., *et al*. (2001) *Nature* **413,** 852–856.
38. Birkey, S. M., Liu, W., Zhang, X., Duggan, M. F. & Hulett, F. M. (1998) *Mol. Microbiol*. **30,** 943–953.
39. Sun, G., Birkey, S. M. & Hulett, F. M. (1996) *Mol. Microbiol*. **19,** 941–948.
40. Olman, V., Xu, D. & Xu, Y. (2003) *J. Bioinform. Comput. Biol*. **1,** 21–40.
41. Liu, X., Brutlag, D. L. & Liu, J. S. (2001) *Pac. Symp. Biocomput.*, 127–138.

42. Chen, X., Su, Z., Xu, Y. & Jiang, T. (2004) *Genome Inform.* **15** (2), 211–222.
43. Chen, X., Su, Z., Dam, P., Palenik, B., Xu, Y. & Jiang, T. (2004) *Nucleic Acids Res.* **32,** 2147–2157.
44. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
45. Lougee-Heimer, R. (2003) *IBM J. Res. Dev.* **47,** 57–66.
46. Lathe, W. C., 3rd, Snel, B. & Bork, P. (2000) *Trends Biochem. Sci.* **25,** 474–479.
47. Rogozin, I. B., Makarova, K. S., Murvai, J., Czabarka, E., Wolf, Y. I., Tatusov, R. L., Szekely, L. A. & Koonin, E. V. (2002) *Nucleic Acids Res.* **30,** 2212–2223.