

# Mapping of Microbial Pathways through Constrained Mapping of Orthologous Genes

Victor Olman<sup>1</sup>, Hanchuan Peng<sup>1,2</sup>, Zhengchang Su<sup>1,2</sup> and Ying Xu<sup>1,2</sup>

<sup>1</sup>Computational Systems Biology Laboratory

Biochemistry and Molecular Biology Department, University of Georgia, and

<sup>2</sup>Computational Biology Institute, Oak Ridge National Laboratory

Corresponding author: [xyn@bmb.uga.edu](mailto:xyn@bmb.uga.edu)

## Abstract

*We present a novel computer algorithm for mapping biological pathways from one prokaryotic genome to another. The algorithm maps genes in a known pathway to their homologous genes (if any) in a target genome that is most consistent with (a) predicted orthologous gene relationship, (b) predicted operon structures, and (c) predicted co-regulation relationship of operons. Mathematically, we have formulated this problem as a constrained minimum spanning tree problem (called a Steiner network problem), and demonstrated that this formulation has the desired property through applications. We have solved this mapping problem using a combinatorial optimization algorithm, with guaranteed global optimality. We have implemented this algorithm as a computer program, called P-MAP. Our test results on pathway mapping are highly encouraging -- we have mapped a number of pathways of *H. influenzae*, *B. subtilis*, *H. pylori*, and *M. tuberculosis* to *E. coli* using P-MAP, whose homologous pathways in *E. coli* are known and hence the mapping accuracy could be checked. We have then mapped known *E. coli* pathways in the EcoCyc database to the newly sequenced organism *Synechococcus sp* WH8102, and predicted 158 *Synechococcus* pathways. Detailed analyses on the predicted pathways indicate that P-MAP's mapping results are consistent with our general knowledge about (local) pathways. We believe that P-MAP will be a useful tool for microbial genome*

*annotation projects and inference of individual microbial pathways.*

## 1. Introduction

Comparative genome analysis [9] is a powerful tool for inference of gene functions for newly sequenced genomes. This is typically done through identification of "orthologous" genes with known functions in another genome. Using such a method, a significant fraction of the predicted genes in a newly sequenced genome could be assigned biological functions, providing a great amount of information about the new genome [17]. Currently orthology (or homology) mapping is typically carried out at the individual gene level. However it should be noted that "homology" relationships exist at higher functional-unit levels as well, say at the pathway level, across related organisms [19]. For example, different organisms may employ (highly) similar biological pathways to uptake and process a particular nutrient. These "homologous" pathways generally employ orthologous genes to implement each of their corresponding functional steps in the pathways. Hence by identifying orthologous genes across related genomes, one could map a known biological pathway from one genome to another.

Existing methods for pathway mapping are typically done through finding orthologous genes of a known pathway (template pathway) in a target genome, based solely on sequence similarity. For example, Ogata et al [12] used a heuristic graph comparison algorithm to map a generic metabolic pathway to a bacterial genome. One popular approach for orthologous gene mapping is done using reciprocal BLAST search – e.g. the bi-directional-best-hit (BDBH) scheme [11], where two genes are regarded as *orthologous* if they are the best hit in both directions of the BLAST search. Other methods for orthologous gene identification include (i) Clusters of Orthologous Groups (COG) [18], (ii) multiple sequence alignment-based methods [20], (iii) phylogeny tree based methods [1], etc. While these methods have demonstrated their effectiveness

through applications, we have observed in our studies that sequence-based methods for orthologous gene identification alone may not always give the *correct* answer. For example, we found that there are a number of cases where the BDBH genes (for orthologues) of a mapped pathway, using such methods, do not fall into a few co-regulated operons (predicted) while lower ranked genes do. Since a (local) pathway in microbes is typically encoded as a regulon, i.e., a group of co-regulated operons, we believe that additional information such as operons and their co-regulated relationships (predicted) could help improve the mapping accuracy of biological pathways across genomes. In this paper, we present a computer algorithm/program called P-MAP for mapping a known (template) pathway to a target genome, which employs both homologous gene predictions and predicted operons and their co-regulation relationship.

P-MAP maps a group of genes in a template pathway to a target genome in such a way that the overall sequence similarity between mapped gene pairs is as high as possible while attempting to preserve the operon and regulon structures as much as possible. That is, the mapped homologous genes should generally belong to a minimal set of operons sharing similar (common) regulatory binding sites. We formulate this problem as a combinatorial optimization problem called a *Steiner network* problem or a constrained minimum spanning tree problem [2]. We have then developed a rigorous algorithm for solving the problem. To the best of our knowledge, this is the only rigorous algorithm for the pathway-mapping problem.

To demonstrate the capability of P-MAP, we have mapped a number of pathways of *H. influenzae*, *B. subtilis*, and *H. pylori* to *E. coli*, for which the corresponding *E. coli* pathways are known. The results indicate that P-MAP's mappings are highly accurate. We have also mapped a number of *E. coli* pathways to *Synechococcus sp.* WH8102, and compared these mapped results with our manually mapped results. Overall our pathway mapping results are highly encouraging.

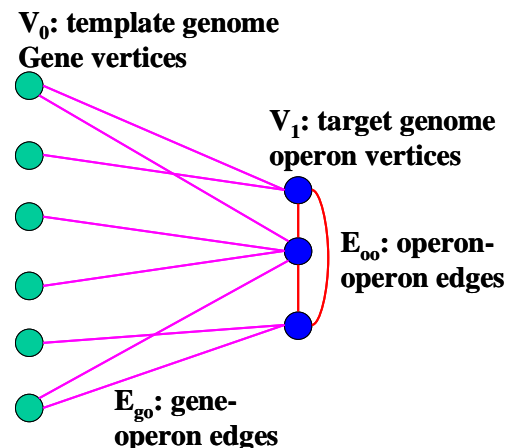
## 2. Methods

### 2.1. Problem Formulation

We consider two microbial genomes: a template genome (TMG) and a target genome (TRG). Let  $P$  be a template pathway in TMG, whose genes we want to map to their "orthologous" genes in TRG. A basic assumption of our pathway-mapping scheme is that genes of a (local) pathway are encoded as a group of co-regulated operons, or a regulon (This constraint can

be relaxed to a small group of regulons, without major changes in our algorithm). Operationally, co-regulated operons are defined as operons sharing a common (similar) regulatory binding site in their promoter regions. In our formulation of a pathway-mapping problem, we intend to include both homology information and organizational information of genes in a pathway, which is a key distinguishing feature of our method, compared to previous pathway-mapping methods.

In a *pathway mapping* problem, we want to have the mapped genes scattered in a minimum number of TRG operons; In addition, among all mappings with a minimum number of operons, we want to have the overall sequence similarity between the homologous gene pairs as high as possible, and have the involved operons share as many similar regulatory binding sites as possible. We now give a formal definition of the problem, capturing this intuition. We define a graph  $G = (V, E)$  as follows. Each gene of the template pathway  $P$  is represented as a vertex, collectively the set denoted as  $V_0$ . Each operon in TRG is represented as a vertex, collectively the set represented as  $V_1$ . We have  $V = V_0 \cup V_1$ . Let the sizes of the two sets be  $m$  and  $n$ , respectively. We connect two vertices,  $u \in V_0$  and  $v \in V_1$ , by an edge  $(u, v)$  if  $u$  has a homologous gene in operon  $v$ . Homology is defined by BLAST search. Specifically, two genes are considered as homologous if the E-value of their BLAST sequence comparison is  $< 10^{-2}$  -- we use a rather high E-value in order not to miss any remote orthologous genes. For the simplicity of discussion, we assume that each gene in  $P$  has at least one homologue in TRG, and each operon contains at least one homologous gene to one of  $P$ 's genes (otherwise we remove the isolated genes and operons from consideration). We also connect each pair of operons by an edge. We denote the whole set of edges as  $E$ , and call the whole graph  $G = (V, E)$  as a *mapping graph*. Figure 1 shows a schematic of a mapping graph.



## Figure 1: A schematic of a mapping graph

For each edge of  $G$ , we assign a positive weight as follows. First all gene-operon edges have larger weights than operon-operon edges. All operon-operon edges have a (positive) constant weight unless the two operons share common (similar) regulatory binding sites. In such cases, we assign a smaller weight -- the more confident we are that the two operons have common (predicted) regulatory binding sites, the lower the weight is. Gene-operon edges are assigned in a similar fashion -- the more similar two gene sequences are, the lower the weight is, without violating the aforementioned relationship between the two types of edges. In the following subsection we will give a detailed description of how the edge weights are assigned.

For a given mapping graph  $G = (V_0 \cup V_1, E)$ , we want to find a subtree of  $G$  spanning all vertices of  $V_0$  and possibly some vertices of  $V_1$ , that has the minimum weight among all such spanning trees. This is a *Steiner network* problem [2], which has been proved to be a NP-hard problem [10]. In the following subsection, we discuss the relationship between our Steiner network formulation and the pathway-mapping problem, and present an algorithm for solving the problem.

### 2.2. An Algorithm for Pathway Mapping

We first define a few notations. Let  $W_{go}$  and  $W_{oo}$  be the maximum weights among all gene-operon and operon-operon edges of  $G$ , respectively. Similarly, let  $w_{go}$  and  $w_{oo}$  be the minimum edge weights, respectively. The following Theorem links an optimum solution to a Steiner network problem to our pathway-mapping problem.

**Theorem 1:** For a template pathway  $P$ , a target genome TRG, and  $P$ 's mapping graph  $G_P = (V_0 \cup V_1, E)$ , an optimum solution to the Steiner network problem (spanning at least all vertices of  $V_0$ ) gives a mapping of  $P$  to TRG such that (a) every gene of  $P$  is mapped to one vertex of  $V_1$  (i.e., one operon in TRG), and (b) the number of  $V_1$  vertices included into the solution is minimal among all feasible mappings if the following two conditions are satisfied:

(i)  $W_{oo} < w_{go}$ , and (ii)

$$W_{go} - w_{go} < [w_{oo} - (n - 2) \cdot (W_{oo} - w_{oo})] / m.$$

□

Hence, if we could assign the edge weights of  $G$  to satisfy the conditions (i) and (ii), we will obtain a mapping of genes of  $P$  to their homologous genes in the target genome, that covers the minimum number of

operons, through solving the Steiner network problem. An outline of a proof of the Theorem is given in the Appendix. In addition, it is clear that the total sequence similarity between the mapped homologous genes is maximized, calculated through solving the Steiner Network problem, among possible mappings satisfying (i) and (ii).

We now present how we assign the edge weights that satisfy the conditions (i) and (ii) in Theorem 1. For each gene-operon edge  $(u, v)$ , we assign as its edge weight the p-value of the BLAST score between gene  $u$  and its homologue in operon  $v$  (we assume that  $u$  has only one homologue in operon  $v$ ). For each operon-operon edge  $(u, v)$ , we initially assign a positive constant  $C$  as its weight. We adjust the edge weight as follows if operons  $u$  and  $v$  are predicted to be co-regulated. For each operon, we predict its transcription regulatory (TF) binding sites using our program CUBIC [14], by identifying conserved sequence motifs across predicted orthologous genes (and operons) of closely related genomes (Su et al, unpublished results, 2004). Two operons are considered *co-regulated* if they have common (similar) predicted TF binding motifs. A p-value is assigned based on the similarity of the two similar TF binding sites (Su et al, unpublished results, 2004). Each such operon-operon edge is assigned a weight the inverse of the corresponding p-value. The  $C$  value assigned to non-co-regulated operons is chosen so that it is greater than any of the above operon-operon edge weights. The following theorem shows how to transform the initial edge-weight assignments so that the transformed edge weights will satisfy conditions (i) and (ii) in **Theorem 1** without violating the relative order (degree of similarity) in BLAST and CUBIC results.

**Theorem 2.** Let  $w(E)$  be the weight of edge  $E$  in  $G$ . Let's choose  $c_o, c_{g0}, c_{g1}$  such that

$$c_o > \max(0, (n - 2) * W_{oo} - (n - 1) * w_{oo}),$$

$$0 < c_{g0} < \frac{w_o - (n - 2) * (W_{oo} - w_{oo}) + c_o}{m(W_{go} - w_{go})} \text{ and}$$

$c_{g1} > \max(0, W_{oo} - c_{g0} * w_{go} + c_o)$ . Then the Steiner network problem with newly assigned weights for each operon-operon edge  $E_{oo}$

as  $w^*(E_{oo}) = w(E_{oo}) + c_o$ , and for each gene-operon edge  $E_{go}$  as

$w^*(E_{go}) = c_{g0} * w(E_{go}) + c_{g1}$ , will satisfy both conditions of **Theorem 1**. □

A proof of the theorem is given in the Appendix.

For our applications to *E coli* and *Synechococcus* sp. WH8102 (see Section 3), we have used the *E coli* operons in RegulonDB and our own prediction of operons in *Synechococcus* sp WH8102 (<http://www.cs.ucr.edu/~xinchen/operons.htm>).

**Algorithm:** For each subset  $S$  of  $V_I$ , we find a minimum spanning tree for the graph  $G' = (V_0 \cup S, E')$ , where  $E' \subset E$  is the set of edges incident to  $V_0 \cup S$ . The global optimum solution of Steiner network problem corresponds to the minimum spanning tree with the smallest weight, among all  $2^{|V_I|}$  subsets of  $V_I$ . Typically in an application,  $|V_I| < |V_0|$ . In case  $|V_0| < |V_I|$ , we will switch between  $V_0$  and  $V_I$  in our algorithm to minimize the computing time. Since the minimum spanning tree problem can be solved in polynomial time, this algorithm runs in polynomial time of  $|V_0|$  and an exponential time of  $|V_I|$ . For pathways with 10~30 gene-vertices, the algorithm typically runs in a few seconds on a PC (2.5 GHz).

**Software Availability:** The software package P-MAP (including the source codes) can be downloaded at the following web site: <http://csbl.bmb.uga.edu/WH8102>.

### 3. Results

#### 3.1 Mapping pathways to *E. coli*

Our first test is on two pathways, KDO<sub>2</sub>-lipid A and peptidoglycan biosynthesis and Chorismate superpathways. These two pathways represent two of the largest pathway models in the EcoCyc database (12), and they have been well characterized in *H. influenza*, *B. subtilis*, *H. pylori*, *M. tuberculosis* and *E. coli*, respectively. Our tests involve mapping the pathways of *H. influenza*, *B. subtilis*, *H. pylori*, *M. tuberculosis* to *E. coli*, and comparing the mapped pathways with the known *E. coli* pathways in the EcoCyc database [3].

**KDO<sub>2</sub>-lipid A and peptidoglycan biosynthesis superpathway.** The KDO<sub>2</sub>-lipid A and peptidoglycan biosynthesis (super)pathway is a metabolic pathway for lipopolysaccharide and glycan biosynthesis [6] with over 20 genes (the number varies for different organisms). We have mapped the pathway models of *H. influenza*, *B. subtilis*, *H. pylori* and *M. tuberculosis*, stored in MetaCyc database [5], to *E. coli* using P-MAP, and the mapped results are given in Table 1.

As shown in Table 1, 20 out of the 24 genes (83.3%) in the *E. coli* pathway are mapped correctly when mapping the genes of pathways of *H. influenza*,

*B. subtilis*, *H. pylori* and *M. tuberculosis* to the *E. coli* genome, and there is no inconsistency among mappings from the four template pathways for the 20 correctly mapped genes. A few interesting results are worth noting. First, we find that operon structure information does help to make the pathway mapping more accurate. For example, if we consider the best BLAST hits only, genes Bsu0050 and Bsu0178 (marked in yellow) of *B. subtilis* are incorrectly mapped to WH8102. But P-MAP maps both genes to the correct genes (and they belong to the same operon) in *E. coli*, using the structural constraints of operons. Second, considering the operon structures in our mapping also allows us to recruit additional genes, unmapped genes in the reliably mapped operons (defined in terms of the number of mapped genes in an operon) into the pathway models. For example, gene b0096 is correctly added to the predicted pathway though it does not have a counter part in the template pathways. Third, one mapped gene (b1094) does not appear in the annotated *E. coli* pathway, which could be a false positive in our prediction or a missing annotation in *E. coli* pathway database. Regardless of the true or false prediction, it does point out that our prediction needs an additional post-processing step to assess such predictions using additional information, say predicted functions. Fourth, all false negative predictions are due to the fact that the templates do not have the corresponding genes. This suggests that more template pathways may benefit the pathway prediction, particularly knowing the low false positive rate of our pathway mapping.

**Chorismate pathway.** Chorismate pathway involves the synthesis of three amino acids: tyrosine, phenylalanine and tryptophan [4], with over 30 genes (34 in *E. coli*). We have mapped the pathway models of *H. influenza*, *B. subtilis*, *H. pylori* and *M. tuberculosis*, stored in MetaCyc database [8], to *E. coli* using P-MAP, and the mapped results are given in Table 2.

As shown in Table 2, 23 out of the 34 genes (67.6%) in the *E. coli* pathway are mapped correctly when mapping the genes of pathways of *H. influenza*, *B. subtilis*, *H. pylori* and *M. tuberculosis* to the *E. coli* genome, and there is no inconsistency among mappings from the four template pathways for the 26 correctly mapped genes. As in the case of Table 1, a few interesting results are worth noting. First, three new genes (b1264, b2599, b2600) are predicted to be in the pathway model of *E. coli*, which we believe are probably true because they exist in the template pathways, have the relevant function, and form an operon. Second, all false negative predictions (11) are due to the small coverage of the template pathways (i.e., the templates do not have the corresponding

genes). Third, four “false” positives are predicted. However it is difficult to know at this point if these are

real false positives or true false under disguise (caused by the missing parts in the annotated *E coli* pathway).

**Table 1.** Mapping results of the "KDO<sub>2</sub>-lipid A and peptidoglycan biosynthesis superpathway". Light shadow: the correct operon is found for both the template and target organisms; *italics*: genes in different operons in the template genomes are mapped to one operon in target genome; underline: the genes/operons are highly likely to be in the actual *E coli* pathway, based on additional information we have though they are not included in EcoCyc database; dark shadow: a mapped gene is not in the true pathway, or a gene in the true pathway is not predicted.

Template Pathways used				<i>E coli</i> genes in the mapped or "true" pathway	
<i>H. influenzae</i>	<i>B. subtilis</i>	<i>H. pylori</i>	<i>M. tuberculosis</i>	Mapped	True
HI1133	Bsu0457	HP0493	MT2211	b0085	b0085 (murE)
HI1134	Bsu1520	HP0494	MT2212	b0086	b0086 (murF)
HI1135	Bsu1522	HP0623	MT2214	b0087	b0087 (mraY)
HI1136	Bsu2975	HP0740	MT2216	b0088	b0088 (murD)
HI1138		HP1052	MT2217	b0090	b0090 (murG)
HI1139		HP1155		b0091	b0091 (murC)
HI1140		HP1494		b0092	b0092 (ddlB)
				b0096	b0096 (lpxC)
	Bsu0456	HP0738	MT3059	b0381	b0381 (ddlA)
HI1060		<i>HP0196</i>		b0179	b0179 (lpxD)
HI1061		<i>HP0867</i>		b0181	b0181 (lpxA)
		<i>HP1375</i>		b0182	b0182 (lpxB)
					b0914 (msbA)
HI0058		HP0230		b0918	b0918 (kdsB)
					b1054 (htrB)
		HP0962		b1094	
HI1557		HP0003		b1215	b1215 (kdsA)
					b1855 (msbB)
HI1081	Bsu3674	HP0648	MT1355	b3189	b3189 (murA)
					b3197 (yrbH)
HI0652		HP0957		b3633	b3633 (kdtA)
<i>HI0429</i>	<i>Bsu0050</i>	<i>HP0683</i>	<i>MT1046</i>	b3729	b3729 (glmS)
<i>HI0642</i>	<i>Bsu0178</i>	<i>HP1532</i>	<i>MT3542</i>	b3730	b3730 (glmU)
	Bsu2676	HP0549	MT1379	b3967	b3967 (murl)
HI0268	Bsu1525	HP1418	MT0500	b3972	b3972 (murB)

### 3.2 Mapping 201 *E coli* pathways to *Synechococcus sp.* WH8102

After the initial testing on P-MAP, we have applied the program to mapping 201 *E coli* pathways in the Ecocyc database [3], to *Synechococcus sp.* WH8102 (WH8102), a newly sequenced cyanobacteria genome [16] (we have an on-going project to study the regulatory networks of WH8102). Using a particular set of thresholds involving the percentage of genes mapped and the percentage of genes sharing common operons, P-MAP mapped 151 of these 201 pathways to WH8102. The mapping results can be found at <http://csbl.bmb.uga.edu/WH8102/>. While detailed analysis of these prediction results will be published elsewhere, we provide a detailed analysis on one particular

pathway, the phosphorus assimilation pathway encoded by the pho regulon. Table 3 shows the result of mapping the phosphorus assimilation pathway of *E coli* to WH8102.

As shown in table 3, all 9 genes in the template are mapped into 4 operons in WH8102. By examining the functions, operon structures and predicted regulatory binding sites, we believe that these mapped results make biological sense. First, a two-component system encoded in operon {SYNW0947,SYNW0948} was mapped and predicted, which might be responsible for controlling of the activities of pho regulon in response to changes in phosphorus level in the environment. Second, two phosphorus uptake systems are mapped and predicted: a high affinity inorganic phosphate uptake system is predicted to be encoded in operons {SYNW1270-1272} and {SYN2507}, and an organic

**Table 2.** Mapping results of the "chorismate pathway". Code conventions are the same as used in Table 1.

Template pathways used				<i>E. coli</i> genes in the mapped or "true" pathway	
<i>H. influenzae</i>	<i>B. subtilis</i>	<i>H. pylori</i>	<i>M. tuberculosis</i>	Mapped	True
HI0899	Bsu2182		MT2833	b0048	b0048 (folA)
HI0064	Bsu0079	HP1036	MT3711	b0142	b0142 (folK)
					b0388 (aroL)
HI0609	Bsu2429	HP0577	MT3464	b0529	b0529 (folD)
HI1547				b0754	b0754 (aroG)
HI1589	Bsu2259	HP0401	MT3324	b0908	b0908 (aroA)
	Bsu0076	HP0587		b1096	b1096 (pabC)
		HP0587		b1097	
	Bsu2971			b1215	
HI1389	Bsu2262	HP1277	MT1646	b1260	b1260 (trpA)
HI1431	Bsu2263	HP1278	MT1647	b1261	b1261 (trpB)
HI1432	Bsu2264	HP1279	MT1648	b1262	b1262 (trpC)
	Bsu2265	HP1281	MT2248	b1263	b1263 (trpD)
	Bsu2266				
HI1387	Bsu2267	HP1282	MT1644	b1264	b1264 (trpE)
					b1323 (tyrR)
	Bsu2307		MT2629	b1692	
	Bsu2562			b1693	b1693 (aroD)
					b1658 (purR)
					b1704 (aroH)
		HP0293	MT1034	b1812	b1812 (pabB)
HI1447	Bsu2277	HP0928	MT3713	b2153	b2153 (folE)
					b2232 (ubiG)
					b2311 (ubiX)
HI1261	Bsu2804	HP1545	MT2523	b2315	b2315 (folC)
HI0196	Bsu2270	HP0663	MT2615	b2329	b2329 (aroC)
HI0889	Bsu3688	HP0183	MT0076	b2551	b2551 (glyA)
HI1145	Bsu2786			b2599	b2599 (pheA)
		HP1380		b2600	b2600 (tryA)
					b2601 (aroF)
	Bsu0078		MT3712.1	b3058	
					b2907 (ubiH)
HI1336	Bsu0077	HP1232	MT1245	b3177	b3177 (folP)
HI0655		HP1249		b3281	b3281 (aroE)
	Bsu0075			b3360	b3360 (pabA)
HI0208	Bsu2269	HP0283	MT2613	b3389	b3389 (aroB)
HI0207	Bsu0316	HP0157	MT2614	b3390	b3390 (aroK)
		HP1468		b3770	
			MT0584	b3833	b3833 (ubiE)
					b3844 (ubiB)
					b4039 (ubiC)
		HP1360		b4040	b4040 (ubiA)
					b4054 (tyrB)
					b4393 (trpR)

phosphorus-containing compound uptake system is predicted to be encoded by operon {SYNW1168,SYNW1169,SYNW1170}, suggesting that WH8102 can utilize both source of phosphorus. Both predictions have been verified experimentally by B Palenik's lab of UCSD (unpublished results). One interesting observation we made is that if we use BDBH (bi-direction best hit) to map this pathway, *phoR* of *E. coli* is not mapped to SYNNW0948 as it should as there is no BDBH pair for

*phoR* between *E. coli* and WH8102. The same is true for mapping *E. coli* proteins *phnC* and *phnE*. However by considering the operon structure as well as binding sites of the operons (operons {SYNW1270-1271}, {SYNW2507}, {SYNW0947,SYNW0948} and {SYNW1168-1170} have similar binding sites (14)), P-MAP has successfully mapped these proteins. This example demonstrates the power of using genomic structural information in addition to sequence similarity information.

**Table 3.** Mapping pho regulon of *E coli* to WH8102. Each shaded block represents an operon. The mapped genes of WH8102 in *italics* represent genes that could not be mapped correctly using BDBH method, a popular method for "orthologous" gene mapping.

E coli template	Functions	Orthologs in WH8102
pstB	ATP binding component	SYNW1272
pstA	intergal membrane protein	SYNW1271
pstC	intergal membrane protein	SYNW1270
pstS	P <sub>i</sub> binding protein	SYNW2507
phoB	response regulator	SYNW0947
phoR	sensor kinase	SYNW0948
phnE	channel protein	SYNW1168
phnD	periplasmic binding protein	SYNW1170
phnC	ATP binding component	SYNW1169

#### 4. Discussion

Pathway mapping from one genome to another is a fundamental problem in comparative genomic studies. For one thing, it is very useful for genome annotation projects as has been demonstrated in KEGG [3] and MetaCyc database [7]. Several algorithms [13] have been proposed to solve this problem, however, none of them provides a rigorous solution to utilize not only the homologous information, but also genomic structural information of operons and regulons. By formulating the pathway mapping problem as a special Steiner network problem, our algorithm takes full advantages of both types of information aforementioned.

As demonstrated in this paper, our algorithm performs well in pathway mapping for both closely related genomes (Table 1 and 2) and remotely related genomes (Table 3). Some remote orthologs might not be found by conventional ortholog finding methods, such as BDBH. For instance, BDBH method failed to map E coli genes phoR, phnC and phnE to WH8102 correctly, but our algorithm gave correct results. We believe that one of the merits of our algorithm is its ability to map pathways between two remotely related genomes. This makes it very useful tool for genome annotation projects and researchers when a specific pathway data for a closely related genome is lacking as in the case of phosphorus assimilation pathway in WH8102 [15].

Apparently, the ability of our algorithm to map correctly the remote orthologs in pathway mapping is due to the incorporation of operon and regulon structures in our algorithm. Other data that imply pathway information about the target genome, such as microarray data from the target genome, can also be incorporated in our formulation to increase the accuracy and robustness of the algorithm when they become available.

In conclusion, we have developed and implemented a novel computer algorithm called P-MAP, for mapping biological pathways/networks from one prokaryotic genome to another. P-MAP is a very useful tool for microbial genome annotation projects and inference of individual microbial pathways.

#### Appendix:

##### A proof of Theorem 1.

The inequality (i) guarantees that  $T(V_0)$ , a tree that solves Steiner network problem, will have exactly one edge coming out of any  $V_0$  - vertex. Let's assume it's not a true, and there is  $v_0 \in V_0$  such that  $T(V_0)$  contains two edges  $e_1, e_2$  from  $v_0$  to two operons  $O_1, O_2$ . Replacing  $e_1$  by the edge connecting  $O_1$  and  $O_2$ , we will still keep a tree structure and reduce the total weight of a tree, which contradicts the optimality of  $T(V_0)$ .

The condition (ii) guarantees that a total weight of any spanning tree with T operons is greater than the one with  $t < T$  operons. For proving that it's sufficient to check that for  $t < T$

$$n * W_{go} + (t - 1)W_{oo} < n * w_{go} + (T - 1)w_{oo} ,$$

or after simple transformation

$$W_{go} - w_{go} < \frac{1}{n} ((T - 1)w_{oo} - (t - 1) * W_{oo}) ,$$

and

this is a true because the right side of the last inequality is decreasing in  $t$ , and for  $t = T - 1$  is decreasing in  $T$ , while for  $T = n$  it is a condition (ii). This completes the proof.

##### A proof of Theorem 2.

From the inequality for  $C_o$  we can easily check that  $C_{g_0} > 0$ , i.e. the transformation keeps the rank of gene-operon edge weights. The inequality for  $C_{g_0}$  guarantees the condition (ii) in the Theorem 1. After choosing  $C_o$  and  $C_{g_0}$  we pick up

$C_{g1}$  in order to satisfy the condition (i). Hence the theorem is proved.

## Acknowledgements

This work was funded in part by the US Department of Energy's Genomes to Life Program under project "Carbon Sequestration in *Synechococcus sp*: From Molecular Machines to Hierarchical Modeling" (<<http://www.genomes-to-life.org>>). The work is also supported, in part, by the NSF grant ITR-325386.

## References

- [1] Arvestad L., Berglund A. C., Lagergren J. and Sennblad B., Bayesian gene/species tree reconciliation and orthology analysis using MCMC, *Bioinformatics*, 2003, 19 Suppl 1: pp. I7-I15.
- [2] Dreufus E. and Wagner R., The Steiner Problem in Graphs, *Networks*, 1971, 1: pp. 195-207.
- [3] Kanehisa M., The KEGG database, *Novartis Found Symp*, 2002, 247: pp. 91-101.
- [4] Karp P. D., Riley M., Paley S. M. and Pellegrini-Toole A., The MetaCyc Database, *Nucleic Acids Res*, 2002, 30: pp. 59-61.
- [5] Karp P. D., Riley M., Paley S. M. and Pellegrini-Toole A., The MetaCyc Database, *Nucleic Acids Res*, 2002, 30: pp. 59-61.
- [6] Karp P. D., Riley M., Paley S. M. and Pellegrini-Toole A., The MetaCyc Database, *Nucleic Acids Res*, 2002, 30: pp. 59-61.
- [7] Karp P. D., Riley M., Paley S. M. and Pellegrini-Toole A., The MetaCyc Database, *Nucleic Acids Res*, 2002, 30: pp. 59-61.
- [8] Karp P. D., Riley M., Paley S. M. and Pellegrini-Toole A., The MetaCyc Database, *Nucleic Acids Res*, 2002, 30: pp. 59-61.
- [9] Koonin E. V., Aravind L. and Kondrashov A. S., The impact of comparative genomics on our understanding of evolution, *Cell*, 2000, 101: pp. 573-576.
- [10] Lawler E, *Combinatorial Optimization: Networks and matroids*, Saunders College Publishing, 2001.
- [11] Mushegian A. R. and Koonin E. V., A minimal gene set for cellular life derived by comparison of complete bacterial genomes, *Proc Natl Acad Sci U S A*, 1996, 93: pp. 10268-10273.
- [12] Ogata H., Fujibuchi W., Goto S. and Kanehisa M., A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters, *Nucleic Acids Res*, 2000, 28: pp. 4021-4028.
- [13] Ogata H., Fujibuchi W., Goto S. and Kanehisa M., A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters, *Nucleic Acids Res*, 2000, 28: pp. 4021-4028.
- [14] Olman V., Xu D. and Xu Y., Identification of regulatory binding sites using minimum spanning trees, *Pac Symp Biocomput*, 2003, pp. 327-338.
- [15] Palenik B., Brahamsha B., Larimer F. W., Land M. and et.al, The genome of a motile marine *Synechococcus*, *Nature*, 2003, 424: pp. 1037-1042.
- [16] Palenik B., Brahamsha B., Larimer F. W., Land M. and et.al, The genome of a motile marine *Synechococcus*, *Nature*, 2003, 424: pp. 1037-1042.
- [17] Palenik B., Brahamsha B., Larimer F. W., Land M. and et.al, The genome of a motile marine *Synechococcus*, *Nature*, 2003, 424: pp. 1037-1042.
- [18] Tatusov R. L., Koonin E. V. and Lipman D. J., A genomic perspective on protein families, *Science*, 1997, 278: pp. 631-637.
- [19] Tsoka S. and Ouzounis C. A., Functional versatility and molecular diversity of the metabolic map of *Escherichia coli*, *Genome Res*, 2001, 11: pp. 1503-1510.
- [20] Wall D. P., Fraser H. B. and Hirsh A. E., Detecting putative orthologs, *Bioinformatics*, 2003, 19: pp. 1710-1711.