Computational Inference of Regulatory Pathways in Microbes: an Application to Phosphorus Assimilation Pathways in *Synechococcus sp.* WH8102

Zhengchang Su[1] zhx@csbl.bmb.uga.edu
Phuongan Dam[1] phd@csbl.bmb.uga.edu
Xin Chen[2] xinchen@cs.ucr.edu
Victor Olman[1] olman@csbl.bmb.uga.edu
Tao Jiang[2] jiangt@cs.ucr.edu
Brian Palenik[3] palenikB@ucsd.edu
Ying Xu[1*] xyn@csbl.bmb.uga.edu

[1] Department of Biochemistry and Molecular Biology, University of Georgia at Athens, and Computational Biology Institute, Oak Ridge National Laboratory
[2] Department of Computer Science and Engineering, University of California at Riverside
[3] Scripps Institute of Oceanography, University of California at San Diego

Abstract

We present a computational protocol for inference of regulatory and signaling pathways in a microbial cell, through literature search, mining "high-throughput" biological data of various types, and computer-assisted human inference. This protocol consists of four key components: (a) construction of template pathways for microbial organisms related to the target genome, which either have been extensively studied and/or have a significant amount of (relevant) experimental data, (b) inference of initial pathway models for the target genome, through combining the template pathway models and target genome-specific information, (c) refinement and expansion of the initial pathway models through applications of various data mining tools, including phylogenetic profile analysis, inference of protein-protein interactions, and prediction of transcription factor binding sites, and (d) validation and refinement of the pathway models using pathway-specific experimental data or other information. To demonstrate the effectiveness of this procedure, we have applied it to the construction of the phosphorus assimilation pathways in cyanobacterium sp. WH8102. We present, in this paper, a model of the core components of this pathway.

Keywords: regulatory pathway, signaling pathway, computational inference

1 Introduction

In living systems, control of biological function occurs at the systematic, cellular and molecular levels [1]. Such control mechanisms are often coupled with signaling pathways that detect environmental changes and trigger the relevant components of the regulatory machinery, resulting in specific cellular responses. The complex machinery for transmitting and implementing the regulatory signals is made of a network of interacting proteins and/or other molecules like DNA, RNA and small molecules [1]. Characterization of these regulatory networks or pathways is essential to our understanding of biological functions at both molecular and cellular levels. Traditionally, interaction networks and signaling pathways have been characterized through *ad hoc* approaches, which could take a significant amount of time. With the advent of high-throughput measurement technologies, and the completion of sequences of a large number of microbial genomes, it is now feasible and desirable to develop new

*Correspondence: Ying Xu (xyn@bmb.uga.edu)

and effective computational protocols for tackling the challenge of characterization of regulatory and signaling pathways in a systematic manner.

Numerous computational methods have been developed to infer regulatory and signaling pathways, based on high-throughput biological data. These methods generally attempt to infer local networks based on microarray gene expression data [28], and/or protein-protein interaction map derived from two-hybrid data [10]. While such local networks might provide a rough picture of how a group of genes may relate to one another, they may not necessarily directly correspond to the physical interactions between the products of these genes. In addition, several computational strategies have been developed to infer transcriptional regulatory networks in microbes [25]. Though these methods are potentially capable of revealing the structure of a transcription regulatory network, they typically provide only partial information for a complete regulatory/signaling pathway/network responsible for a particular biological process in a cell.

Our experience has been that none of the data sources and methods aforementioned alone currently contain sufficient information for meaningful derivation of a relatively complete regulatory/signaling pathway in a systematic manner (in the rest of the paper, we simply use "pathways" to refer to regulatory and signaling pathways). However, meaningful and possibly accurate models could be derived based on multiple sources of data, especially when rough pathway templates exist, which may contain key components and/or local structures of the target pathway. For instance, in a microbe, proteins working in the same pathway are often encoded by genes in an *operon*, or encoded by genes in separate operons which share the same transcriptional binding site(s), called a *regulon*. Thus, genes identified to belong to the same operon/regulon provide good candidates for genes working in the same pathways. Protein-protein interaction predictions, based on two hybrid data [30] or computational techniques like protein fusion methods [17], could provide detailed information about how these genes (their products) might be wired together. We have recently developed a computational protocol for biological pathway inference using various sources of information. Our goal is to make such techniques applicable to a large class of microbial pathways, which have known or partially known homologous pathways in other related genomes.

We have chosen as our target pathway the phosphorus assimilation pathway in the marine cyanobacterium *Synechococcus sp.* WH8102 (WH8102), which has its genome sequenced, genes predicted and their functions annotated [21], and is one of the major primary producers in the large oligotrophic central gyres of the world's oceans. Because phosphorus is one of the significant limiting nutrients in the oceans, it was chosen as a starting point of our computational inference of pathways for WH8102. Currently, very little is known about this particular pathway in this species [22].

2 Methods

A pathway model can be represented as a set of molecules (proteins, DNA, RNA, and small molecules) connected by links representing physical or functional interactions. It could be represented either in its *generic* form (i.e., proteins without detailed gene assignments) or in its *specific* form with genes assigned to all proteins. Our computational procedure will first build a pathway model in its generic form through template-based pathway inference, and then attempt to assign (predicted) genes to individual proteins. This procedure consists of the following key steps: (I) construction of template pathways for related organisms, (II) inference of initial pathway modes for the target genome based on the template pathways, (III) expansion and refinement of the initial pathway models using various sources of information, and (IV) validation of pathway models. The information flow of the inference process is shown in Figure 1.

2.1 Construction of Template Pathways

Related organisms often employ similar biochemical processes to accomplish the same goal; hence, known biological pathways could possibly be used as templates for pathway inference in another organism though detailed differences among these pathways need to be carefully dealt with. For a

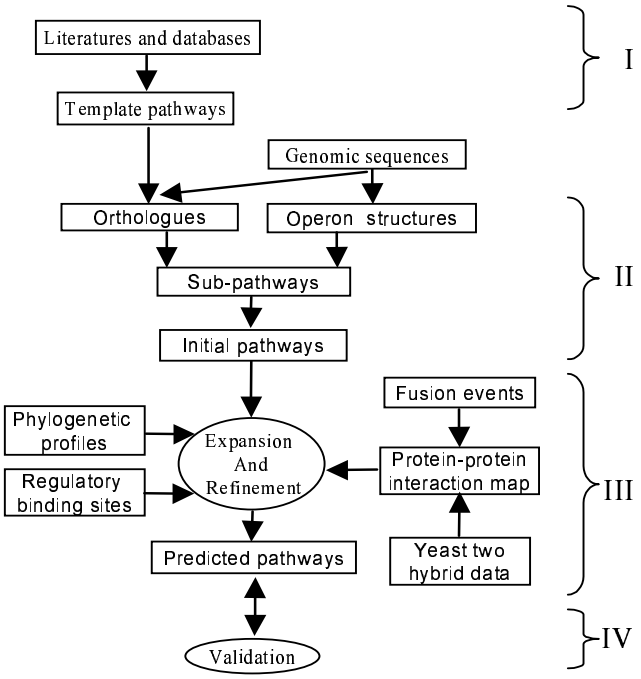

Figure 1: Information flowchart of the pathway inference protocol. A rectangular box represents an information source or/and a sink, and an oval denotes an information integration process.

specific target pathway P in organism X, we first attempt to build P's "homologous" pathways in X's related genomes for which a great amount of information and/or experimental data may exist. If such "homologous" pathways exist in one of the pathway databases, like KEGG [12], EcoCyc [14] and MetaCyc [13], we can use them directly as templates. Otherwise, we do extensive literature search to identify all components and interaction relationships known to be part of the pathway, and then piece together (often partial) pathway models based on our general knowledge of a particular class of pathways (see Section 3.1). In our template-based pathway inference, we typically use 3-5 related genomes as different organisms may share different common "sub-pathways" with the target pathway. Information derived from template pathways from different organisms may complement each other for our target pathway inference.

2.2 Construction of Initial Target Pathway

A key component in our template-based pathway inference is to build an initial target pathway by assigning genes to the proteins in the template pathways. The guiding principle in our gene assignments is to assign genes so the following criteria are met: (a) assigned genes should be homologous (more rigorously, orthologous) to their corresponding proteins in the template pathways; and (b) the operon structures of the assigned genes should be maximally consistent with the operon structures of their corresponding proteins (genes) in the template genome. We accomplish this goal in two phases.

a. Phase 1: Search for orthologues genes and prediction of operons

(1) Search for orthologous genes: Operationally, we define two proteins A and B, of two different genomes, as *orthologues* if they are the *bidirectional best hits* (BDBH) [20] against each other's genome. If no such BDBH is found, we relax the condition to find the closest homologue using BLASTP. In case that there is no detectable homologue by BLASTP, we run the threading program PROSPECT [34] for remote homology detection. A list of predicted homologues for all 2531 genes of WH8102 is provided at `http://compbio.ornl.gov/PROSPECT/`. Considering the possibilities of incorrect prediction of orthologues, we consider, for each protein, a few additional homologous genes as possible alternative

candidates with lower confidence scores.

(2) Prediction of operons: Our algorithm makes operon predictions, based on the assumption that operon structures are generally "conserved" among closely related genomes. Hence the input to the program is a set of (closely) related genomes. In this study, we used three cyanobacteria genomes: WH8102, *Prochlorococcus* MED4 and MIT9313, as the input genomes. The program first performs pair-wise genome comparisons to find pairs of genes from two different genomes that are deemed homologous by using the BLASTP program and the COG ID assigned by COGnitor program [36]. Two genes are considered *homologous* if (i) the E-value of their match given by BLASTP exceeds a pre-defined threshold (E-20 in the current experiment) in both directions and (ii) the two genes have the same COG ID. Then the program identifies conserved clusters of genes from at least two genomes that satisfy conditions (i) the genes of a cluster have the same direction on all involved genomes, (ii) consecutive genes are within 100 bases, and (iii) the genes from all involved genomes corresponding to the same position in a cluster have the same COG ID. These conserved gene clusters are then ranked by likelihood estimates based on (1) consistency of functional categories of the genes in a cluster as given by COGnitor, (2) existence and conservedness of key promoter motifs (such as the Pribnow box and sigma dependent promoters) in the promoter region of the first gene of a cluster as given by databases such as TRANSFAC and TFD [7] and computer programs such as SIGSCAN [24] and PromScan [29], and (3) the existence and conservedness of terminator motifs in the 3' (downstream) region of the last gene of a cluster as given by computer programs such as TransTerm [5]. The likelihood parameters and thresholds are estimated from a set of 237 experimentally confirmed operons from the *E. coli* K12 genome [27]. Conserved gene clusters that receive high likelihood scores above a certain threshold are finally output as putative operons. A more detailed description of the prediction algorithm will be reported elsewhere (Chen *et al.*, working paper). An up-to-date summary of the predictions on the three genomes can be found via WWW at `http://www.cs.ucr.edu/~xinchen/operons.htm`.

b. Phase 2: Construction of an initial pathway model through gene assignments
The essential idea in building the initial target pathway model is to find a set of gene assignments to each template pathway that best satisfy the criteria (a) and (b) listed in the beginning of Section 2.2, and to generate the union of the gene assignments against different template pathways after fixing possible gene-assignment conflicts. The following describes our current procedure for accomplishing this.

Step 1: identification of conserved sub-pathways in each template pathway
For each operons in each template pathway do the following:

If none of its proteins have orthologues (as defined by BDBH) in the target genome, we predict that the sub-pathway encoded by this operon does not exist in the target pathway;

If all of its proteins have orthologues in the target genome, we predict that the sub-pathway encoded by this operon exists in the target pathway, and use these orthologues as gene assignments;

If only a fraction of its proteins have orthologues in the target genomes (the so called "*missing gene problem*" [19]), do the following:

If all missing genes have detectable homologues (by our methods) in the target genome, which exist in the same operon of the identified orthologues, we predict that the sub-pathway encoded by this operon exists in the target pathway, and use these orthologues and homologues as gene assignments; Otherwise, we only predict that the target organism has a somewhat modified sub-pathway, without detailed structural predictions.

Step 2: initial model construction through merging sub-pathways
Do a *union* operation on all the sub-pathways identified in Step 1 and then resolve conflicts as follows. If different genes are assigned to the same protein in different sub-pathways of different templates, we resolve the problem by taking the gene assignment with the highest confidence score.

2.3 Expansion and Refinement of Initial Pathway Models

Considering the possible differences among "homologous" pathways across different organisms, some proteins in a template pathway may not exist and some additional elements may exist in the target pathway. Our inference protocol employs an "expansion and refinement" step to deal with such an issue. This is done through an application of predicted protein physical or functional associations. The basic idea is if protein A is in our pathway model but B is not, we may want to add B into the model if we predict A and B are interacting either physically or functionally - we can either predict detailed connections between B and the rest of the pathway, or simply indicate that B should be part of the pathway. Currently, we employ the following techniques for such predictions: (a) prediction of physical interactions of proteins through detecting protein fusion events and orthologues mapping, (b) prediction of co-regulated proteins through predictions of operons and regulatory binding sites, and (c) phylogenetic profile analysis for prediction of functionally related genes. Other data sources can also be used here, like microarray gene expression data if it available. We apply such "expansion" operations only on assigned genes with high confidence, i.e, they have respective orthologues in the template genome, and they are in a same operon if their respective orthologues in the template genome are in an operon.

a. Prediction of physical interactions of proteins

We have used two computational procedures for the prediction of physical interactions of proteins in WH8102.

(1) Prediction through orthologues mapping: There are a number of public databases of protein-protein interactions derived from large-scale two-hybrid experiments [30], mass spectrometry [6] and from individual experiments [33]. The most comprehensive such data sets are for *S. cerevisiae* [6, 30] and *H. pylori* [26]. Our orthologous gene mapping method predicts that two proteins interact if their orthologues are known to interact in *S. cerevisiae* or *H. pylori*, based on information stored in the DIP database, which contains 15099 and 1420 binary interactions for *S. cerevisiae* and *H. pylori,* respectively (release of January 5, 2003).

(2) Prediction through protein fusion analysis: The basic idea of gene-fusion method is that if two proteins A and B are homologous to different segments (domains) of the same protein chain C in another genome, then A and B are predicted to interact [17, 23]. We used BLASTP to search the all open reading frames (orfs) of WH8102 against the non-redundant protein (*nr*) database. Protein A and B in WH8102 are predicted to have physical interactions if they are homologous (E-values $<$ 0.0001) to different domains of the same protein in the *rn* database.

Two proteins are predicted to be interacting if one of the two approaches predicts the interaction. A total of 963 interactions involving 722 proteins are predicted for WH8102. Figure 1 shows the interaction map of this prediction.

b. Prediction of co-regulated proteins

Co-regulated proteins typically work in the same biological pathways in microbes. These are the proteins that are regulated by the same regulatory protein(s), and hence they share the same regulatory binding site(s) in their promoter regions. Our goal is to identify all operons that have the *same* binding site as that of the known co-regulated operons. Our approach is to first identify the conserved binding sites of possible co-regulated operons, using three computer programs CUBIC [18], MEME [3] and BIO-PROSPECTOR [16]; and then search all the promoter regions of the predicted operons in WH8102 for the same binding sites. Predicted binding sites by CUBIC, MEME and BIO-PROSPECTOR are analyzed manually. Predictions by multiple programs will receive higher confidence scores. After identifying the best set of promoters that contains a common motif, the profile of this motif is computed to search for similar motif present in the upstream regions of WH8102 orfs, using both publicly available MAST and an in-house program. For the in-house program, the search was constructed to maximize the information content score of the final motif. We have identified a 11-bp sequence motif in the regulatory region of pstBAC, pstS, phoBR, phnCED and phnX operons, which is equivalent to the *pho box* in *E coli* [4]. It is well known that these key components of this pathway, though not of

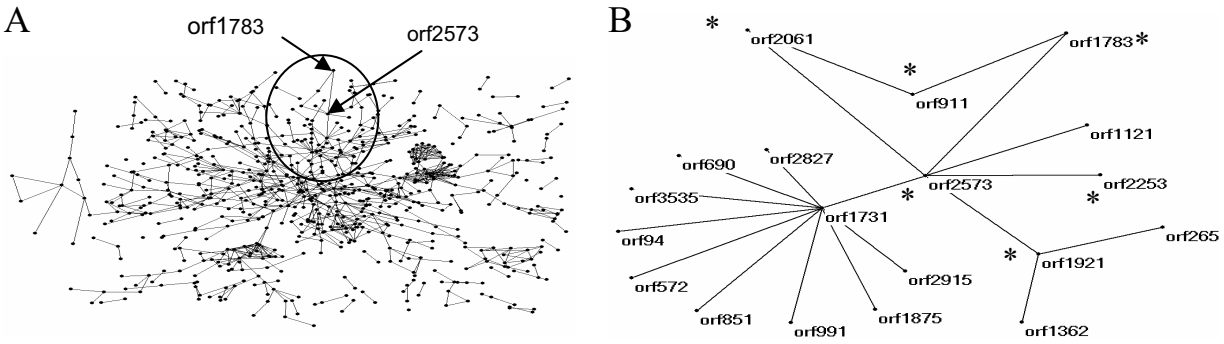

Figure 2: (A) Genome-scale protein-protein interactions for WH8102. Each vertex represents a protein and an edge linking two proteins represents a predicted interaction. (B) A blow-up of the portion encircled in part A shows the local interaction network of orf1783, the predicted phoR. Star-labeled orfs are also predicted by phylogenetic profile analysis.

the same operon, are co-regulated by the same transcription factor, phoB, which binds to the pho box motif [4]. Using this simple procedure, we have identified 81 candidate pho box sites in the regulatory regions WH8102 genome. Genes identified to be co-regulated with genes known to be in the phosphorus assimilation pathway are considered as candidates used for the "expansion" operation.

c. Prediction of functional associations of proteins through phylogenetic profile analysis Pioneered by D. Eisenberg and co-workers [23], the phylogenetic profile analysis predicts that two proteins are functionally related if they have highly similar phylogenetic profiles. When searched against a set of genomes $\{G_1, \ldots, G_k\}$, a protein*'s* phylogenetic profile is defined as a binary string $a_1 \ldots a_k$ with $a_i = 1$ if the protein has a detectable homologue in genome G_i(defined by BLASTP search using an e-value $< 10^{-6}$); otherwise 0. We have used 89 sequenced microbial genomes (see supplemental material at `http://compbio.ornl.gov/~xyn/pathways/` for the list of genomes) to compute the phylogenetic profiles for all genes of WH8102. We then clustered these genes into clusters based on the similarities of their phylogenetic profiles, using the EXCAVATOR program [35]. The clustering results for all predicted genes of WH8102 can be found at `http://compbio.ornl.gov/~xyn/patwhays`. Detailed description of this clustering procedure and the analysis of the clustering results will be published in a more expanded version of this paper.

2.4 Validation of Pathway Models

Experiments could be designed to validate predicted pathway models. This may include experiments to (a) check if two specific proteins interact; (b) check if the mRNA expression levels go up/down for certain genes under particular conditions, as predicted; and (c) check if a particular component in a pathway, e.g., a specific phosphorus transport system, exists – just to name a few. Validation results could be used to refine the pathway models, in an iterative manner.

3 Results and Discussion

3.1 Information Compilation through Literature Search

It is generally known that phosphorus assimilation pathways in microbial organisms have the following key components that are co-regulated in the so-called pho regulon: (a) A two-component sensor-regulator systems made up of phoR (sensor histidine kinase), phoB (response regulator, i.e. transcription factor), and related modulators. PhoR senses the inorganic phosphate (P_i) level and is activated when P_i level is low. Activated phoR in turn activates phoB by phosphorylating the latter. Phosphorylated phoB then either activates or suppresses the transcription of operons in the pho regulon. (b) Transporters of phosphorus-containing compounds of various types including P_i or C-P bond containing compounds called phosphonates (P_n). (c) Various metabolic enzymes that degrade

organic phosphorus compounds to P_i. (d) Other components that help the cell acclimate changes in phosphorus supply in the environment. However, proteins not encoded in the pho regulon might also be involved in the global responses induced by phosphorus stress. We have used such information as the general guidance in our construction of both template and target pathways.

3.2 Construction of Template Pathways

Through extensive literature and database searches, we found that the phosphorus assimilation pathway in *E. coli* [8, 31], *B. subtilis* [2] and *S. typhimurium* [11] have been extensively studied, which are all somewhat related to our target genome. Hence we have chosen these organisms to build the template pathways. For each of these template pathways, we start with the well characterized components and add the missing components/connections based on the information collected in Section 3.1.

a. Phosphorus assimilation pathway in *E. coli*: *E. coli* uses two low-affinity P_i transporters PitA and PitB [9] as its primary transporters when P_i is abundant in the environments. When P_i is low, the high affinity P_i transporter system Pst (PstSCAB-phoU) will be induced. In the absence of P_i and the presence of organic P_n, a P_n transporter system (phnCDE) as well as the metabolic enzyme complex C-P lyase (phnFGHIJKLMNOP) will be induced to transport a wide spectrum of P_n into the cell and break them down to P_i [8, 31], respectively. *E. coli* also utilizes sn-glycerol 3-phosphate by inducing the ugpBAECQ operon. All these proteins are co-regulated by the two-component signaling system PhoR and PhoB whose activity is negatively regulated by phoU.

b. Phosphorus assimilation pathway in *S. typhimurium* : This organism has similar P_i transport systems, a two-component regulatory system, to those of *E. coli*, except for that only one low-affinity P_i transporter pitA has been identified experimentally. Moreover, in the absence of P_i, *S. typhimurium* can only use a special form of P_n, i.e., the naturally existing 2-aminoethylphosphonate (2-AEP). This is because its P_n transporter (phnSTUV) and/or the metabolic enzymes (phnW and phnX) can only transport and break down 2-AEP into P_i [11], respectively.

c. Phosphorus assimilation pathway in *B. subtilis*: This microbe has similar P_i transporting systems and similar two-component signaling system to those of *S. typhymurium* [2]. Although more genes/operons in its pho regulon have been characterized experimentally, neither P_n transporter nor catabolic enzymes have been identified.

By piecing together the known partial pathway models of these three organisms using the above information to fill the gaps, we have built three pathway models for each of them, encoded by their respective pho regulons. The detailed models can be found at `http://compbio.ornl.gov/~xyn/pathways/`.

3.3 A Pathway Model for Phosphorus Assimilation and Its Evaluation

a. The initial model

Table 1 shows the identified WH8102 orthologues/homologues of the proteins in the three template pathways encoded by their respective pho regulons of *E. coli*, *S. typhimurium* and *B subtilis*. These assigned genes constitute the components in the initial model (not fully connected) of the phosphorus assimilation pathway in WH8102, and they are probably members of its pho regulon.

This predicted initial pathway model suggests several interesting possibilities about the phosphorus assimilation process in WH8102. The protocol predicted that orf1783 and orf1782 are the orthologues of phoR and phoB, respectively. Sequence homology search reveals that orf1783 has 31% sequence identity to sphS, which has been shown to be a phoR equivalent in *Synechococcus* PCC7942. Interestingly, both sphS and or1783 are predicted to be soluble proteins by three different approaches (TMHMM, SOSUI and DAS). This suggests that orf1783 can sense the low intracellular P_i level, or an associated protein can sense the low extracellular P_i level, which, if exists, is yet to be identified. In addition, our protocol did not find an orthologue of phoU, a negative regulator of phoR, in WH8102 just like in 33 out of 89 microbial genomes that we used for phylogenetic profile analysis, suggesting

Table 1: Predicted components of the initial phosphorus assimilation pathway in WH8102. Each row represents a protein in a template pathway. Consecutive rows under the same-colored shadow (blue or white) represent an operon in the template genome whose name is given at the last column of each row. The first column represents the names of proteins, the second column the functions, and the third column the assigned genes of WH8102 to the corresponding proteins. All orthologues are shown in red letters while homologues detected through other methods are shown in black or blue. A green-color shadow indicates an operon structure in WH8102.

Proteins	Functions	Orthologs	Binding site	Cluster coord.	Template pathway
pstB	ATP binding component	orf224		1897	*E. Coli, S.typhimurium.B.subtilis*
pstA	intergal membrane protein	orf223		1891	*E. Coli, S.typhimurium, B.subtilis*
pstC	intergal membrane protein	orf222	+	1892	*E. Coli, S.typhimurium, B.subtilis*
pstS	P_i binding protein	orf1096	+	1893	*E. Coli, S.typhimurium, B.subtilis*
phoB	response regulator	orf1782	+	1858	*E. Coli, S.typhimurium, B.subtilis*
phoR	sensor kinase	orf1783		1866	*E. Coli, S.typhimurium, B.subtilis*
phoH		orf3085		2085	*B. subtilis*
phoD	phosphodiesterase	orf1247		980	*B. subtilis*
phnE	channel protein	orf3427		2401	*E. Coli*
phnD	periplasmic binding protein	orf3425	+	2035	*E. Coli*
phnC	ATP binding component	orf3426		1596	*E. Coli*
ugpC	ATP-binding component	orf2670		1691	*E. Coli*
phnW	AEP animotransferase	orf86		1407	*S.typhimurium*
phnX	phosphonatase	orf191	+	2012	*S.typhimurium*
resE	sensor kinase	orf2061	+	1867	*B. subtilis*
resC	cytochrome c biogenesis	orf124		1966	*B. subtilis*
resB	cytochrome c biogenesis	orf2118	+	984	*B. subtilis*
tuaH	teichuronic acid biosynthesis	orf1228		1941	*B. subtilis*
tuaG	teichuronic acid biosynthesis	orf1994		1947	*B. subtilis*
tuaD	dehydrogenase	orf1250		1728	*B. subtilis*
tagA	teichoic acid biosynthesis	orf1271	+	2433	*B. subtilis*

the possibility that phoU was lost in some microbes including WH8102 during the evolution. It is intriguing to find out what mechanism has been evolved to compensate the role of phoU in these microbes.

No low-affinity P_i transporter was detected by our prediction in WH8102. This may reflect the fact that the P_i level in the seawater (100∼300 nM) is well below the K_m (25μM) of pitA/B transporters, and therefore the respective transporter genes may have been lost during the evolution. We also failed to find orthologues for phoE, an outer membrane pore, and for iciA, an inhibitor of DNA replication. The reason for this is not clear but might also be related to the ecological niche of WH8102.

Our procedure also did not predict any orthologue for the C-P lyase complex. This may suggest that WH8102 utilizes a different system to break down P_n's. Indeed, we have predicted that WH8102 has a hybrid P_n uptake and degradation system through the identification of orf3425, orf3426 and orf3427 as the orthologues of the $P_n$transporter phnCDE of *E. coli*, and orf86 and orf191 as the orthologues of phnW and phnX, which are responsible for breakdown of 2-AEP in two sequential enzymatic reactions in *S. Typhimurium* [15]. Since phnCDE is promiscuous P_n transporter [32] while phnWX is specific for 2-AEP [15], we could not predict the spectrum of P_n's in the predicted hybrid P_n uptake and degradation system of WH8102. However, our experimental result has demonstrated that WH8102 can utilize 2-AEP as well as ethylphosphonate as a sole phosphorus source (unpublished observation), suggesting that a novel enzymatic system might exit in WH8102 to break down ethylphosphonate, or alternatively, that the phnWX complex in WH8102 could also break down ethylphosphonate. Further experiments are needed to clear this cloud.

b. The expanded model

Based on our predicted protein-protein interaction map, we first expand this initial model by predicting a local interaction network involved the sensor kinase phoR (orf1783) (Figure 2B). In this local network, both orf1783 and orf2061, which are sensor kinases, are predicted to interact with response

regulators orf911 and orf2573. The latter is also predicted to interact with another kinase, orf1921. These predictions are consistent with the recent finding that cross-talk widely occurs in two-component systems in microbes. Interestingly, orf2253, predicted to interact with orf2573, is a regulatory subunit of cAMP-dependent protein kinase. It has been shown that phosphate-starvation-inducible psiE gene of the pho regulon in *E. coli* is dually regulated by the phoB/cAMP receptor protein complex. Moreover, orf1921, also predicted to interact with orf2573 is the σ-factor of RNA polymerase, which is recruited by the transcription factors during the transcription process. Although orf1731, another predicted interacting partner of orf2573, is a hypothetical protein, ten other orfs were predicted to interact with it, suggesting that it may be an important protein.

Next, through the prediction of phoB (or1782) binding sites, we added 81 operons to the pho regulon of WH8106. This set of proteins can be found at `http://compbio.ornl.gov/~xyn/pathways/`. Finally, through phylogenetic analysis, we predicted that 120 proteins are also likely to play a role in the phosphorus assimilation pathway of WH8102 (`http://compbio.ornl.gov/~xyn/pathways/`). Even though these two methods cannot define the precise positions and roles in the pathway of the most of predicted proteins, there are some interesting observations made on these predictions. For instance, the orf1888/orf1889 operon, predicted to be involved in the phosphorus pathway by phologenetic analysis, are homologous to phoR/phoB. It will be interesting to invetstigate their functions experimentally. Another two-component system predicted to be involved in this pathway is orf44/orf45, and orf44 was annotated as a P_i-binding protein. It has been shown that P_i limitation is correlated with up- or down-regulations of both specific phosphate-stress response proteins and general stress response proteins [2]. Thus, these predicted sensor/response systems might play an important role in the global responses during phosphate-starvation either through a general mechanism or through the interactions between these systems and the phoR/phoB system. Notably, orf2061, orf911, orf1121 orf2573, orf2253 and orf1921, predicted to be involved in the pathway based on the predicted protein-protein interaction map (Figure 2B), are also predicted to be involved in the pathway through phylogenetic profile analysis.

Based on the predicted initial pathway model, proteins added to the initial model through the Refinement and Extension Step, and our general knowledge about phosphorus assimilation processes, we have manually built a pathway model for the phosphorus assimilation pathway in WH8102, as shown in Figure 3. While experimental validation will be needed to build a final pathway models, the information provided by this computational protocol should be very useful in terms of target selection, and experimental design, especially for determination of a specific pathway of a less-studied species as WH8102.

In conclusion, we have developed a practical and effective computational protocol for deriving regulatory and signaling pathways in a microbial organism. Our initial application results on construction of phosphorus assimilation pathway are highly encouraging. The predicted pathway models provide highly useful information for rational design of experiments to fully characterize the pathways. We expect that such a combined computational-experimental procedure will prove to be generally applicable and highly effective for pathway inference of a large class of regulatory and signaling pathways in a systematic and efficient manner. While the front end of this protocol currently requires human involvement in information extraction from published literature and logical inference to piece together generic pathway models, we expect the amount of human involvement could be significantly decreased as the technology matures.

Acknowledgments

This work was funded in part by the US Department of Energy's Genomes to Life program (`www.doegenomestolife.org`) under project, "Carbon Sequestration in *Synechococcus sp.*: From Molecular Machines to Hierarchical Modeling" (`www.genomes-to-life.org`). We also want to thank Drs. Dong Xu, Frank Larimer, Loren Hauser and Mariam Land for helpful discussions and in helping us to retrieve the annotation data on. WH8102. VO and YX's work was also supported by the Office of Health

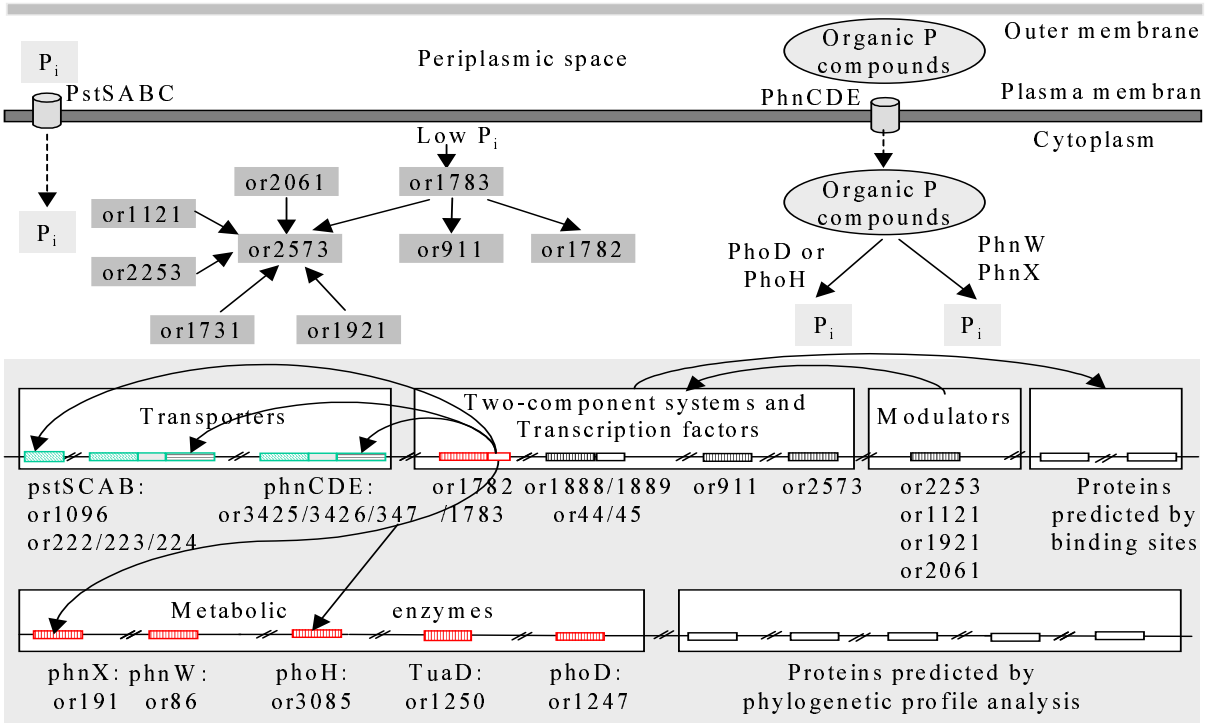

Figure 3: Predicted model of phosphorus assimilation pathway in WH8102. At low P_i levels, phoR is activated, which leads to the activation of phoB. The activated phoB regulates the transcription of genes whose promoters contain a phoB binding site. These proteins include pstSCAB, pstS, phnCDE, phnX, and phoBR operons and possibly others. Moreover, orf1888/orf1889 and orf44/45 are two-component regulatory systems that are probably involved in P_i limitation induced global responses. PhoR can also interact either directly or indirectly with other transcriptional regulators such as orf911 and orf2573, forming a hierarchical transcription regulatory network. For most of the proteins predicted to be involved in the phosphorus assimilation pathway through binding site predictions and phylogenetic profile analysis, we do not know their specific positions in the pathway at this moment. Genes in color: components in the initial model; genes in black: predicted by expansion step, one box may stand for multiple genes.

and Environmental Research, U.S. Department of Energy, under Contract No. DE-AC05-000R22725 managed by UT-Battelle, LLC.

References

[1] Akmaev, V.R., Kelley, S.T., and Stormo, G.D., Phylogenetically enhanced statistical tools for RNA structure prediction, *Bioinformatics*, 16:501–512, 2000.

[2] Antelmann, H., Scharf, C., and Hecker, M., Phosphate starvation-inducible proteins of *Bacillus subtilis*: proteomics and transcriptional analysis, *J Bacteriol*, 182:4478–4490, 2000.

[3] Bailey, T.L. and Gribskov, M., Combining evidence using p-values: application to sequence homology searches, *Bioinformatics*, 14:48–54, 1998.

[4] Blanco, A.G., Sola, M., Gomis-Ruth, F.X., and Coll M., Tandem DNA recognition by PhoB, a two-component signal transduction transcriptional activator, *Structure,* 10:701–713, 2002.

[5] Ermolaeva, M.D., Khalak, H.G., White, O., Smith, H.O., *et al.*, Prediction of transcription terminators in bacterial genomes, *J. Mol. Biol.,* 301:27–33, 2000.

[6] Gavin, A.C., Bosche, M., Krause, R., Grandi, P., *et al.*, Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature*, 415:141–147, 2002.

[7] Ghosh, D., New developments of a transcription factors database, *Trends Biochem. Sci.*, 16:445–447, 1991.

[8] Haldimann, A., Daniels, L.L., and Wanner, B.L., Use of new methods for construction of tightly regulated arabinose and rhamnose promoter fusions in studies of the *Escherichia coli* phosphate regulon, *J. Bacteriol.*, 180:1277–1286, 1998.

[9] Harris, R.M., Webb, D.C., Howitt, S.M., and Cox, G.B., Characterization of PitA and PitB from *Escherichia coli*, *J. Bacteriol*, 183:5008–5014, 2001.

[10] Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A.F., Discovering regulatory and signalling circuits in molecular interaction networks, *Bioinformatics*, 18 Suppl 1:S233–S240, 2002.
[11] Jiang, W., Metcalf, W.W., Lee, K.S., and Wanner, B.L., Molecular cloning, mapping, and regulation of Pho regulon genes for phosphonate breakdown by the phosphonatase pathway of *Salmonella typhimurium* LT2, *J. Bacteriol.*, 177:6411–6421, 1995.
[12] Kanehisa, M., The KEGG database, *Novartis Found. Symp.*, 247:91–101, 2002.
[13] Karp, P.D., Riley, M., Paley, S.M., and Pellegrini-Toole, A., The MetaCyc Database, *Nucleic Acids Res.*, 30:59–61, 2002.
[14] Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., *et al.*, The EcoCyc Database, *Nucleic Acids Res.*, 30:56–58, 2002.
[15] Kim, A.D., Baker, A.S., Dunaway-Mariano, D., Metcalf, W.W., *et al.*, The 2-aminoethylphosphonate-specific transaminase of the 2-aminoethylphosphonate degradation pathway, *J. Bacteriol.*, 184:4134–4140, 2002.
[16] Liu, X., Brutlag, D.L., and Liu, J.S., BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes, *Pac. Symp. Biocomput.*, 127–138, 2001.
[17] Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., *et al.*, Detecting protein function and protein-protein interactions from genome sequences, *Science*, 285:751–753, 1999.
[18] Olman, V., Xu, D., and Xu, Y., Identification of regulatory binding sites using minimum spanning trees, *Pac. Symp. Biocomput.*, 327–338, 2003.
[19] Osterman, A. and Overbeek, R., Missing genes in metabolic pathways: a comparative genomics approach, *Curr. Opin. Chem. Biol.*, 7:238–251, 2003.
[20] Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., *et al.*, The use of gene clusters to infer functional coupling, *Proc. Natl. Acad. Sci. USA*, 96:2896–2901, 1999.
[21] Palenik, B., Brahamsha, B., Larimer, F.W., Land, M., *et al.*, The genome of a motile marine Synechococcus, *Nature*, 424(6952):1037–1042, 2003.
[22] Palenik, B. and Dyhrman, S.T., Recent progress in understanding the regulation of marine primary productivity by phosphorus, In: *Phosphorus in Plant Biology: Regulatory Roles in Molecular, Cellular, Organismic, and Ecosystem Processes*, edited by Lynch, J.P. and Deikman, J., American Society of Plant Physiologists, 1998.
[23] Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., *et al.*, Assigning protein functions by comparative genome analysis: protein phylogenetic profiles, *Proc. Natl. Acad. Sci. USA*, 96:4285–4288, 1999.
[24] Prestridge, D.S., SIGNAL SCAN: a computer program that scans DNA sequences for eukaryotic transcriptional elements, *Comput. Appl. Biosci.*, 7:203–206, 1991.
[25] Qin, Z.S., McCue, L.A., Thompson, W., Mayerhofer, L., *et al.*, Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites, *Nat. Biotechnol.*, 21:435–439, 2003.
[26] Rain, J.C., Selig, L., De Reuse, H., Battaglia, V., *et al.*, The protein-protein interaction map of *Helicobacter pylori*, *Nature*, 409:211–215, 2001.
[27] Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Millan-Zarate, D., *et al.*, RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12, *Nucleic Acids Res.*, 29:72–74, 2001.
[28] Steffen, M., Petti, A., Aach, J., D'haeseleer, P., *et al.*, Automated modelling of signal transduction networks, *BMC Bioinformatics*, 3:34 2002.
[29] Studholme, D.J. and Dixon, R., Domain architectures of sigma54-dependent transcriptional activators, *J. Bacteriol.*, 185:1757–1767, 2003.
[30] Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., *et al.*, A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*, *Nature*, 403:623–627, 2000.
[31] VanBogelen, R.A., Olson, E.R., Wanner, B.L., and Neidhardt, F.C., Global analysis of proteins thesized during phosphorus restriction in *Escherichia coli*, *J. Bacteriol.*, 178:4344–4366, 1996.
[32] Wanner, B.L., Molecular genetics of carbon-phosphorus bond cleavage in bacteria, *Biodegradation*, 5:175–184, 1994.
[33] Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., *et al.*, DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions, *Nucleic Acids Res.*, 30:303–305, 2002.
[34] Xu, D., Crawford, O.H., LoCascio, P.F., and Xu, Y., Application of PROSPECT in CASP4: characterizing protein structures with new folds, *Proteins*, Suppl 5:140–148, 2001.
[35] Xu, Y., Olman, V., and Xu, D., Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees, *Bioinformatics*, 18:536–545, 2002.
[36] Yeh, I., Karp, P.D., Noy, N.F., and Altman, R.B., Knowledge acquisition, consistency checking and concurrency control for Gene Ontology (GO), *Bioinformatics*, 19:241–248, 2003.