Published online May 8, 2006

2438–2444 *Nucleic Acids Research, 2006, Vol. 34, No. 8*
doi:10.1093/nar/gkl310

Evidence supporting predicted metabolic pathways for *Vibrio cholerae*: gene expression data and clinical tests

Jing Shi[1,*], Pedro R. Romero[4], Gary K. Schoolnik[2], Alfred M. Spormann[3] and Peter D. Karp[5]

[1]Biomedical Informatics Program, MC 5429, [2]Department of Microbiology and Immunology, MC 5107 and [3]Department of Civil and Environmental Engineering, MC 5429, Stanford University, Stanford, CA 94305, USA, [4]School of Informatics, Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202, USA and [5]Bioinformatics Research Group, Artificial Intelligence Center, SRI International Menlo Park, CA 94025, USA

Received December 16, 2005; Revised January 8, 2006; Accepted April 11, 2006

ABSTRACT

***Vibrio cholerae*, the etiological agent of the diarrheal illness cholera, can kill an infected adult in 24 h. *V.cholerae* lives as an autochthonous microbe in estuaries, rivers and coastal waters. A better understanding of its metabolic pathways will assist the development of more effective treatments and will provide a deeper understanding of how this bacterium persists in natural aquatic habitats. Using the completed *V.cholerae* genome sequence and PathoLogic software, we created VchoCyc, a pathway-genome database that predicted 171 likely metabolic pathways in the bacterium. We report here experimental evidence supporting the computationally predicted pathways. The evidence comes from microarray gene expression studies of *V.cholerae* in the stools of three cholera patients [D. S. Merrell, S. M. Butler, F. Qadri, N. A. Dolganov, A. Alam, M. B. Cohen, S. B. Calderwood, G. K. Schoolnik and A. Camilli (2002) Nature, 417, 642–645.], from gene expression studies in minimal growth conditions and LB rich medium, and from clinical tests that identify *V.cholerae*. Expression data provide evidence supporting 92 (53%) of the 171 pathways. The clinical tests provide evidence supporting seven pathways, with six pathways supported by both methods. VchoCyc provides biologists with a useful tool for analyzing this organism's metabolic and genomic information, which could lead to potential insights into new anti-bacterial agents. VchoCyc is available in the BioCyc database collection (http://BioCyc.org).**

INTRODUCTION

Vibrio cholerae causes cholera, a diarrheal disease that occurs most frequently in the form of severe epidemics (1). There have been seven pandemics worldwide since 1817; the current pandemic started in 1961 and affects six continents (2). When untreated, an infected adult can be killed in 24 h (3). *V.cholerae* is a Gram-negative, facultative anaerobic bacterium, and its most unusual feature is that it has two distinctive lifestyles: as a pathogen living in the human intestine and as a bacterium well-adapted to survive in various niches in aquatic habitats, including symbiotic and commensal associations with phytoplankton and zooplankton, respectively (2). As a result, it is logical to speculate that its metabolic repertoire has a capacity to transport and assimilate nutrients from these diverse environments. We are interested in predicting the metabolic pathways of this organism to better understand its biochemistry and obtain potential insights or targets for new anti-bacterial agents. The genome sequence of *V.cholerae* (strain N16961) provides a new foundation for the study of this organism's pathological and environmental characteristics (4). Using the *V.cholerae* genome sequence and the PathoLogic software (5,6) we created VchoCyc (pronounced Vee-koh-sike), a pathway-genome database (PGDB) that predicts *V.cholerae*'s likely metabolic pathways. Here, we describe VchoCyc and report experimental evidence supporting the computationally predicted pathways. Experimental results come from microarray gene expression studies performed with *V.cholerae* isolated from

*To whom correspondence should be addressed. Tel: +1 650 724 5013; Fax: +1 650 725 1536; Email: js2400@stanford.edu

three patients, as well as the same type strain grown in minimal media or LB rich medium, and from known nutrient requirements employed in clinical tests used to distinguish *V.cholerae* from other species.

MATERIALS AND METHODS

VchoCyc Generation

The PathoLogic software that we used to generate VchoCyc is described in previous papers (5–9). We describe here only the additional steps in the creation of VchoCyc.

Since EC numbers were not available in the GenBank annotation file used (4), PathoLogic matched the *V.cholerae* protein names in a table that contains enzyme names from the MetaCyc and Enzyme databases (9). In those cases where there is an unambiguous match of the enzyme name, we accept the PathoLogic assignment of the enzyme to a reaction. In those cases where there is an ambiguous match or a probable enzyme, we manually perform searches in additional online sources such as UniProt and Medline to determine the reaction(s) that the enzyme catalyzes. After the that the enzyme catalyzes. After the name matching process is complete, PathoLogic will import metabolic pathways from MetaCyc into the new PGDB by matching the biochemical reactions for which enzymes have been identified against MetaCyc pathways (9). PathoLogic is intended to provide a first-pass pathway analysis. It is conservative in that it errs more on the side of false-positive pathway predictions than on the side of false-negative predictions (7). After a manual review of predicted pathways by the user to eliminate suspected false-positive pathways, VchoCyc is initially generated.

After generating the VchoCyc database, not all reactions in the predicted pathways have enzymes assigned. When no enzyme is assigned to a reaction, we have a missing enzyme or pathway hole in VchoCyc. The pathway hole filler program, described in an earlier paper (10), identifies candidate enzymes for filling pathway holes. The program uses a set of sequences encoding the required activity in other genomes to identify candidate proteins in the genome of interest, in addition to genomic context (such as the candidate enzyme being in the same operon as another gene in the pathway) to determine the probability that a candidate enzyme has the required function.

Literature on VchoCyc pathways

A comparison of known pathways of *V.cholerae* documented in the literature with those computationally predicted in VchoCyc would provide a measure of the validity of the predictions. Unfortunately, there are almost no fully characterized pathways in *V.cholerae*. The best studied *V.cholerae* pathway is the vibriobactin biosynthesis pathway, which is also not completely characterized. Part of this pathway was predicted in VchoCyc under the name 'enterobactin pathway', and more details are described in the results section.

Validation of VchoCyc with the clinical tests of *V.cholerae*

Clinical tests used to diagnose whether a patient is infected by *V.cholerae* specify known clinical tests for this organism (11). These tests can be used to evaluate VchoCyc's predictions. Each nutrient that can support *V.cholerae* growth must enter the cell either by diffusion or by an active transport mechanism. Once inside the cell it must either be degraded by catabolic pathways/reactions, or used as a substrate to synthesize other metabolites. Thus, the corresponding transporters and metabolic pathways/reactions should be present in VchoCyc. In contrast, if *V.cholerae* cannot use a nutrient, either the transporter system for the uptake of this nutrient and/or the corresponding pathways/reactions should be missing from VchoCyc. We compared clinical tests data and VchoCyc to determine whether VchoCyc's predictions are consistent with the known data.

V.cholerae microarray gene expression data: patient samples

Microarray gene expression data for *V.cholerae* isolated from the stools of three cholera patients were obtained from glass-spotted DNA amplicon arrays encompassing ~90% of the identified open reading frames of this organism (4), and they are available from the Stanford Microarray Database (http://genome-www5.stanford.edu/MicroArray/SMD/). These expression data document the metabolic repertoire of the organism during the last phase of the infectious process in man, in which cells that have detached from the ileal mucosal surface pass through the colon in the typical rice water stool produced by cholera patients. A detailed description of the patient dataset is available in Ref. (12). For each of the three patients four replicate microarray analyses were performed, giving a total of 12 arrays. For the analyses described here, we used the average expression of each gene across the 12 arrays.

V.cholerae microarray gene expression data: minimal media and LB medium

Microarray gene expression data for *V.cholerae* grown in three different growth media were obtained from the Stanford Microarray Database. Data are available for M9 minimal medium supplemented with lactate, M9 minimal medium supplemented with maltose and LB complex medium. These data came from two-channel arrays, where one channel contains the gene expression data for the organism grown in a supplemented minimal medium while the second channel contains data for the gene expression of the organism grown in LB (enriched) medium—a growth condition where genes coding for certain biosynthetic pathways should be silent. All assays were performed with cells during mid-exponential growth phase. Each growth condition was assayed in four replicates; for the analyses described here, we used the average expression of each gene in the four replicate arrays.

Determination of a threshold for considering a gene to be expressed

When working with the gene expression data, one needs to determine when a gene is considered 'expressed'. We determined a threshold for stating a gene as 'expressed' that is based on the minimum expression levels of a set of essential biosynthetic genes. Under minimal growth conditions, essential biosynthetic pathways such as those for amino acids,

nucleotides, coenzymes, vitamins and cofactors must operate, necessitating the expression of the encoding enzymes involved (although not necessarily simultaneously). We examined a set of 36 such essential biosynthetic pathways. For each microarray, we selected a threshold value of absolute expression such that a certain percentage of the reactions in each essential biosynthetic pathway have corresponding 'expressed' genes (based on the requirement of 2-fold greater expression in minimal media than in the enriched medium). Using this threshold, an average of 42% the genes in the 36 essential biosynthetic pathways were 'expressed' in the four datasets, compared with only 21% of the genes in the remaining non-essential pathways.

Definition of evidence that a reaction is catalyzed and that a pathway exists in *V.cholerae*

A metabolic pathway consists of a set of reactions catalyzed by enzymes. In the simplest case, a single reaction (in a pathway) is catalyzed by a single enzyme. For the analyses reported here, we say that there is evidence supporting a reaction when the gene encoding an enzyme catalyzing that reaction is expressed (as defined above). If multiple enzymes can catalyze a single reaction, expression of any one is sufficient. Similarly, we say that there is evidence supporting the existence of a pathway when a specified fraction of the reactions in that pathway are active simultaneously (i.e. when the genes encoding the enzymes catalyzing those reactions are expressed simultaneously). In the results section we discuss how modifying the specification of this fraction affects the results.

RESULTS

Generation of the VchoCyc PGDB

Using PathoLogic with the annotated genome of *V.cholerae*, strain N16961, we have created the VchoCyc PGDB, which is available at the BioCyc website (http://biocyc.org).

Table 1 shows the results of automatic enzyme name matching made by PathoLogic. Curated EC numbers have not been assigned for *V.cholerae*, so the matches were made by comparing protein names. A total of 3828 proteins are predicted in the GenBank annotation for *V.cholerae*. Of the 3828 proteins, 601 (16%) matched reactions in the MetaCyc database. Of these 601 reactions 573 (95%) were unambiguous matches. A total of 28 (5%) were ambiguous matches (i.e. the protein was matched to more than one reaction in MetaCyc) and were resolved manually. Of the remaining 3227 unmatched proteins, 284 (9%) were labeled by PathoLogic as probable enzymes, 269 (95%) of which were resolved manually: 102 were assigned to their corresponding reactions, 167 were recognized as not metabolic enzymes, and the functions of the remaining 15 probable enzymes could not be determined.

Table 1. Results of automatic matching of *V.cholerae* enzymes to reactions by PathoLogic

Type of match	Number of proteins
Matched by EC number	0
Matched by name	601
Ambiguous	28
Unmatched	3227
Probable enzymes	284

Table 2 is a statistical summary of version 1.0 of VchoCyc. VchoCyc contains 912 enzymatic reactions, 654 of which have been assigned to one or more of the 639 enzymes. A total of 258 reactions in the 171 VchoCyc pathways have no assigned enzyme and are called pathway holes. We then applied the pathway hole filler program, which assigned enzymes to 59 pathway holes (http://www.biomedcentral.com/content/supplementary/1471-2105-5-76-S4.html) and completed 14 out of the 114 incomplete pathways. For example, the glycogen degradation pathway was completed by the assignment of gene VCA0250, encoding alpha-amylase, to reaction EC3.2.2.1; the *de novo* biosynthesis of pyrimidine ribonucleotides pathway was completed by the assignment of gene VC1491, encoding dihydroorotate oxidase, to reaction EC1.3.3.1.

Figure 1 shows the distribution of the remaining pathway holes in VchoCyc. Note that 35% of the pathways are complete and ~22% of the pathways have only one missing reaction.

Table 2. VchoCyc statistics

PGDB objects	Quantity
Chromosomes	2
Size (bp)	4 033 464
Genes	3950
Protein genes	3828
Enzyme genes	703
RNA genes	122
Transfer RNAs	98
Compounds	656
Polypeptides	3853
Protein complexes	39
Enzymes	639
Enzymatic Reactions	912
With enzymes in VchoCyc	654
Pathways	171

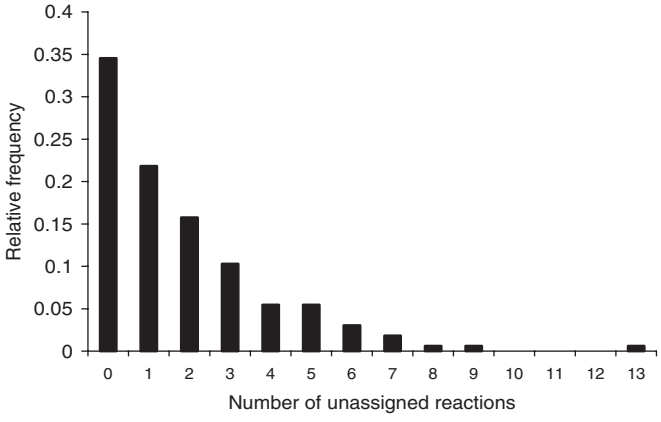

Figure 1. Distribution of the completeness of pathways in VchoCyc, i.e. the fraction of pathways lacking a certain number of missing reactions as a function of that number. For example, ~35% of the pathways in VchoCyc are complete (i.e. lacking zero reactions), meaning that all the enzymes participating in these pathways have been identified in the genome.

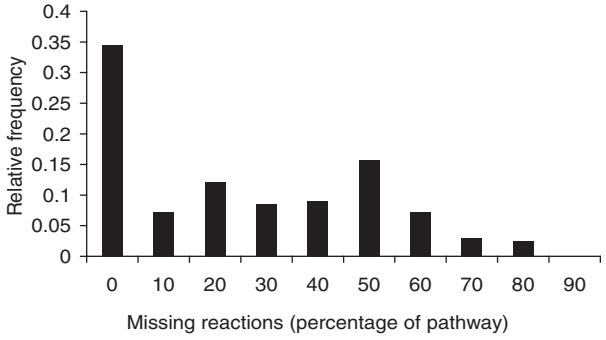

Figure 2. Distribution of incomplete pathways in VchoCyc based on the percentage of the reactions in a pathway that do not have assigned enzymes. For example, ~16% of the pathways in VchoCyc lack enzymes for 50% of their reactions.

Figure 2 shows the distribution of pathway holes as a percentage of a pathway, i.e. the percentage of the reactions in a pathway that do not have assigned enzymes. About 16% of the pathways in VchoCyc have 50% of the reactions without enzymes assigned. The VchoCyc pathways are relatively complete compared with other organisms that we have worked with. This might be because the genome of *V.cholerae* is very similar to that of *Escherichia coli* K-12 (4), and many pathways in MetaCyc are from *E.coli*.

Evidence from published literature supporting predicted pathways

We performed a literature search to find experimentally verified pathways in *V.cholerae* to assess the agreement between predicted and known pathways. Surprisingly, we were not able to find any complete *V.cholerae* pathways that have been experimentally confirmed. The best studied pathway is most probably the vibriobactin biosynthesis pathway. Vibriobactin is an iron-chelating catechol siderophore produced by *V.cholerae* when iron is limited in the environment (13–15). Other catechol siderophores include enterobactin from *E.coli,* yersiniabactin from *Yersinia pestis* and *protochelin* from *Azotobacer vinelandii* (13–15).

Figure 3 shows the biosynthetic pathways of enterobactin and vibriobactin (13,14). The precursor for both enterobactin and vibriobactin is 2,3-dihydroxybenzoic acid (DHBA). Thus, the first committed steps in the synthesis of enterobactin or vibriobactin lead to the synthesis of DHBA from chorismate. The pathway for DHBA synthesis appears to be the same in *V.cholerae* and in *E.coli* (15), and PathoLogic identified the *V.cholerae* enzymes corresponding to the *E.coli* DHBA-synthesizing enzymes (Figure 3). Although DHBA is the precursor for both enterobactin and vibriobactin, their structures differ, and their biosynthetic pathways diverge after DHBA. Both involve multi-enzyme complexes that are not yet fully understood (15). Enterobactin is synthesized from DHBA and serine. In *E.coli,* among the genes involved in this pathway are *entB, entD, entE* and *entF* (15), but the enzymatic steps in this pathway are not well characterized.

Vibriobactin is synthesized from DHBA, threonine and norspermidine. Several genes have been proposed to be involved in this process, including *vibB, vibD,vibE, vibF*

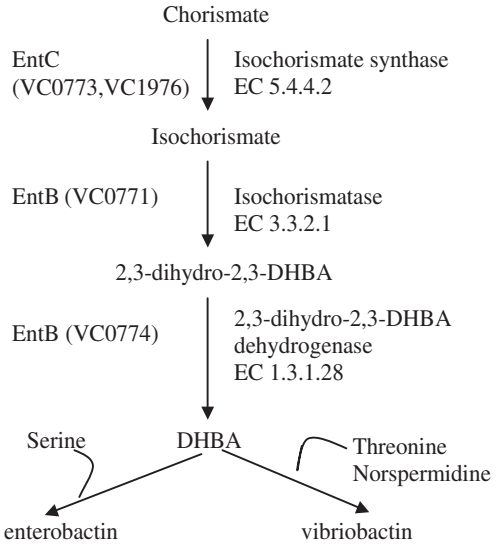

Figure 3. Biosynthetic pathways of the iron-chelating catechol siderophores enterobactin and vibriobactin (15).

and *vibH* (15). *vibB, vibD* and *vibE* are located in a genetic cluster (15). Evidence that these genes participate in vibriobactin synthesis comes from mutation studies. For example, a *vibD* mutant strain was able to synthesize DHBA, but not vibriobactin; this vibriobactin defect was complemented by *vibD* in a plasmid (15). Sequence comparison indicates that *vibD* encodes phosphopantetheinyl transferase and that *vibH* encodes a novel non-ribosomal peptide synthetases (15).

PathoLogic identified the pathway and genes that catalyze DHBA synthesis from chorismate in *V.cholerae*. However, it labeled this as the enterobactin pathway instead of vibriobactin pathway because of the similarity between their biosynthesis pathways, and because there is no vibriobactin pathway in MetaCyc. Since the specific steps from DHBA to enterobactin are not known in *E.coli* and are not in MetaCyc (the example pathway database used to generate VchoCyc), PathoLogic could not predict those specific steps for *V.cholerae*.

Evidence from clinical tests supporting predicted pathways

We compared VchoCyc's metabolic predictions with 33 known clinical tests used in clinical diagnostic tests (11). Two of these tests were excluded because different strains give conflicting results for them. Of the remaining 31 *V.cholerae* clinical tests, VchoCyc includes a reaction or a pathway that utilizes the substrate for 17 of them. For 16 of these 17 clinical tests, VchoCyc identified an appropriate transporter of the substrate and identifies an enzyme or set of enzymes to catalyze the reaction(s) utilizing the substrate and generating the appropriate product(s), and thus is consistent with the data from the clinical tests. The only inconsistent case is D-mannose with acid production—the reaction to catalyze mannose was found but not the transporter for it.

Table 3 shows the results of the comparison. We will discuss two examples in greater detail. Clinical test data

Table 3. Consistency between clinical tests and the computational predictions in VchoCyc

Test[a]	*V.cholerae* (clinical tests)[b]	VchoCyc (predicted)[c]	Evidence on transporter and reaction in VchoCyc[d]
Urea hydrolysis	−[e]	−	Transported through simple diffusion; no reaction found
Phenylalanine deaminase	−	−	No transporter; no reaction found
Arginine, Moeller	−	−	Transported by arginine ABC transporter; no reaction found
Lysine, Moeller	+	+	Transported by Cadaverine/Lysine antiporter; EC4.1.1.18(VC0281)
Ornithine, Moeller	+	+	Transported by Putrescine-Orthinine antiporter; EC4.1.1.17 (VCA1068) (ornithine degradation pwy)
Malonate utilization	−	−	No transporter; no reaction found
D-Glucose, acid production	+	+	Transported by PTS; too many pwys can do this
D-Glucose, gas production	−	−	Transported by PTS; no reaction found
D-Adonitol, acid production	−	N/A[f]	N/A
L-Arabinose, acid production	−	N/A	N/A
Cellobiose, acid production	+/−	−	Transported by PTS; no reaction found
Dulcitol, acid production	−	N/A	N/A
Erythritol, acid production	−	−	No transporter; no reaction found
D-Galactose, acid production	+	+	Transported by Galactose ABC transporter; EC2.7.1.6 (VC1595) (galactose degradation pwy)
Glycerol, acid production	+/−	+	Transported by simple diffusion; EC2.7.1.30 (VCA0774) (many pwys)
myo-Inositol, acid production	−	N/A	N/A
Lactose, acid production	+/−	−	No transporter; By EC3.2.1.23 (no enzyme assigned), lactose becomes to β-D-galactose and β-D-glucose, which further produce acid (lactose degradation IV pwy and glactose, galactoside, and glucose catabolism
Maltose, acid production	+	+	Transported by Maltose ABC transporter; By EC2.4.1.25 (VCA0014), maltose becomes β-D-glucose (Glycogen biosynthesis pwy), which further produces acid.
D-Mannitol, acid production	+	+	Transported by PTS system; EC 1.1.1.17 (Mannitol degradation pathway)
D-Mannose, acid production	+	−	No transporter for Mannose; EC2.7.1.7 (GDP-mannose matabolism pwy)
Melibiose, acid production	−	−	No transporter for Melibiose; EC3.2.1.22 (VC1690) (galactose, galactoside and glucose catabolism)
A-Methyl-D-glucoside, acid production	−	N/A	N/A
Raffinose, acid production	−	N/A	N/A
L-Rhamnose, acid production	−	N/A	N/A
Salicin, acid production	−	N/A	N/A
D-Sorbitol, acid production	−	N/A	N/A
Sucrose, acid production	+	+	Transported by PTS; By EC2.4.1.13, sucrose can become fructose, which then further produces acid.
Trehalose, acid production	+	+	Transported by PTS; by EC3.2.1.28 , threhalose can become β-D-glucose, which further produces acid
D-Xylose, acid production	−	N/A	N/A
Mucate, acid production	−	N/A	N/A
Esculin hydrolysis	−	N/A	N/A
Acetate utilization	+	+	Acetate can be transported by simple diffusion; Acetate utilization pwy
Nitrate->Nitrite	+	+	Transported by Nitrate Reductase; EC1.7.99.4 (VC1690) (anaerobic respiration, electron acceptors reaction list)

[a]The names of the tests.
[b]The results of the tests from clinical tests.
[c]The results of the tests from checking the VchoCyc database.
[d]The reactions/pathways that are actually present in VchoCyc.
[e]Symbols: +, most strains (generally about 90–100%) positive; −, most strains negative (generally about 0–10% positive); +/−, strains could be positive or negative (generally about 11–89% positive); N/A, the substrate in the test cannot be found in VchoCyc.
[f]N/A means the substrate in a clinical test is not found in VchoCyc.

suggest that *V.cholerae* produces acid when grown on the sugar Trehalose. In VchoCyc, the transporter for Trehalose (phosphoenolpyruvate-dependant phosphotransferase system, Trehalose-specific IIBC component) is present and enzymes that catalyze the reaction to produce acid from Trehalose (EC 3.2.1.28 and EC 2.7.1.2) are also present. The test of D-sorbitol with acid production is negative clinically and in VchoCyc neither the transporter nor the reaction is present, which is consistent with the clinical finding.

Evidence from microarray expression data supporting predicted pathways

PathoLogic predicted 171 metabolic pathways in *V.cholerae*. As described in the Materials and Methods section, for the purpose of this analysis, we say that experimental evidence supports the existence of a predicted pathway when a specified percent of the reactions in that pathway are active simultaneously (i.e. the genes encoding the enzymes catalyzing those reactions are expressed simultaneously). Below, we consider the effect of the fraction of the reactions necessary to determine that there is evidence supporting the existence of a pathway.

Supplementary Table 4 shows the percent of reactions supported by expression evidence for each predicted pathway in each dataset. At the top of the table we see several pathways for which there is expression evidence supporting 100% of the component reactions under all growth conditions; these pathways are aspartate biosynthesis and

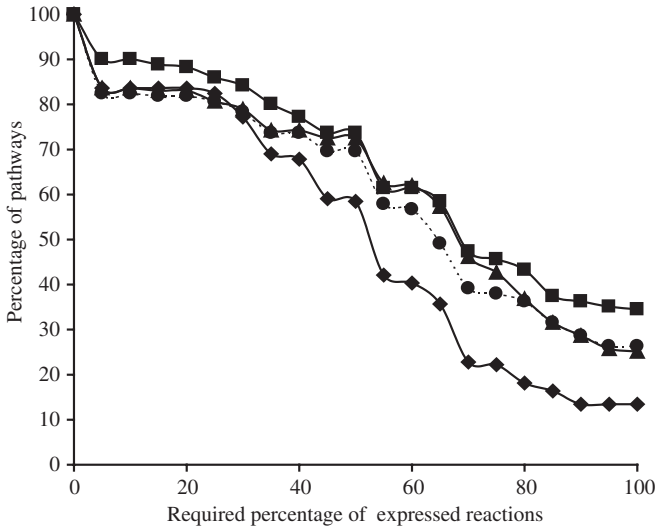

Figure 4. Percentage of pathways for which there is supporting evidence as a function of the percent of the reactions in that pathway that we require to be active simultaneously (i.e. enzymes catalyzing those reactions are expressed simultaneously). M9 with lactate (diamond), M9 with maltose (square), patient stool sample (triangle) and LB medium (circle with dotted line).

degradation, saturated fatty acid elongation, unsaturated fatty acid elongation, glycine biosynthesis I, glycine cleavage and the malate/aspartate shuttle pathway. At the bottom of the table we see several pathways for which there is no expression evidence supporting any of the component reactions; these pathways are 4-hydroxyproline degradation, Entner-Doudoroff pathway, L-idonate degradation and tyrosine biosynthesis II.

We now consider the effect of specifying a series of different percents of the reactions necessary to determine that there is evidence supporting a pathway. Figure 4 shows the percent of pathways for which there is supporting evidence as a function of the percent of the reactions in that pathway that we require to be active simultaneously (genes encoding enzymes catalyzing those reactions are expressed simultaneously). For example, if we require at least 60% of the reactions active simultaneously, then out of the 171 pathways included in VchoCyc, 40% (69) of the pathways have supporting evidence from the microarray gene expression data for growth in M9 + lactate medium, 61% (105) of the pathways have supporting evidence from growth in M9 + maltose medium, 70% (97) of the pathways have supporting evidence from data for growth in LB medium and 62% (106) of the pathways have supporting evidence from clinical samples from patients. Using a value of 100% (meaning that all the reactions in a pathway must be active simultaneously), the percents of pathways having supporting evidence are 13% (29), 34% (59), 26% (45) and 25% (43) in M9 + lactate medium, M9 + maltose medium, LB medium and patient samples, respectively.

Are the reactions supported by expression evidence under the different growth conditions, as summarized in Supplementary Table 4, consistent with the expected metabolic activity of each condition? Merrell and colleagues previously reported on the genes that are differentially expressed in patient samples versus LB medium (12); they found that in patient samples there is an increase in the expression of genes required for amino acid biosynthesis. In our analysis of the expression evidence supporting the predicted pathways (Supplementary Table 4), we found in both patient samples and minimal media the evidence supporting the existence of biosynthesis pathways for amino acids such as Valine, Leucine, Glutamate and Alanine, as well as for nucleotides such as *de novo* biosynthesis of Purine nucleotides and *de novo* biosynthesis of Pyrimidine deoxyribonucleotides. In LB rich medium, we found evidence supporting the existence of the degradation pathways for Glutamine, Ornithine, Threonine and so on. In patients with cholera, there is extreme fluid loss; the intestines become a nutrient-poor environment that is similar to minimal media, so it is reasonable that expression evidence supporting the existence of these biosynthesis pathways comes from these two conditions. In contrast, LB medium is extremely rich in amino acids and nucleotides; we would expect that the degradation pathways are up regulated and that the biosynthesis pathways are down regulated, so it is reasonable that we see expression evidence supporting the existence of the degradation pathways from the LB rich medium condition.

DISCUSSION

Using Pathway Tools software, we constructed a pathway-genome database for *Vibrio cholerae* from its annotated genome. After running PathoLogic to create VchoCyc, 114 pathways had reactions that lacked one or more assigned enzymes, for a total of 258 pathways holes. We used the pathway hole filler tool to assign 59 candidate enzymes out of the total 258 pathway holes, which completed 14 of the 114 pathways.

In order to validate the predicted pathways, we used several sources of evidence: (i) previously known pathways in literature; (ii) consistency with the clinical tests; (iii) consistency with microarray expression data. Vibriobactin pathway, perhaps the best studied pathway in *V.cholerae*, is predicted to be present. Among the 17 clinical tests used for the diagnosis of *V.cholerae*, 16 of them are consistent with VchoCyc.

We examined the expression evidence supporting the presence of each pathway. Using the criterion that at least 60% of the reactions in a pathway are supported by expression evidence, the expression data under the conditions in these experiments provide the evidence supporting 53% (92) of the 171 pathways. We found that the reactions supported by expression evidence under the different growth conditions (patient stool, minimal media, rich medium) are consistent with the expected metabolic activity of each condition: in both (nutrient poor) patient samples and minimal media, the evidence supports the existence of nucleic acid and amino acid biosynthesis pathways, while in rich LB medium we see expression evidence supporting the existence of the corresponding degradation pathways.

There are several explanations as to why a predicted pathway would not be supported by expression evidence: (i) It is not present in *V.cholerae*. (ii) It is present in *V.cholerae* but

is not expressed under the conditions examined. (iii) It is expressed under the conditions examined, but the pathway does not meet our evidence criteria because an array experiment measures a snapshot of a cell, while the genes encoding the different enzymes in a pathway might be turned on and off in a temporal manner.

Since the genome of *V.cholerae* is most similar to that of *E.coli* K-12 (4), it is interesting to determine which *V.cholerae* pathways are not shared by the two organisms. They are both Gram-negative bacteria, but their lifestyles and pathogenicities are quite different: *E.coli* mainly lives in the large intestine of a human host and rarely causes diseases for its host. In contrast, *V.cholerae* cannot only live in the small intestine of its human host but also live in estuarine community; it causes severe diarrhea to its host. We compared the pathways from these organisms using the VchoCyc and EcoCyc databases available through the BioCyc Database collection. In some cases, we can identify the role that these unique pathways play. For example, *V.cholerae* uses ectoine, the product of the ectoine synthesis pathway, to adjust to changes in the ionic composition and osmolarity of its estuarine environment (16). D-rhamnose, the product of the GDP-D-rhamnose biosynthesis pathway, is a rare 6-deoxy monosaccharide that occurs primarily in the lipopolysaccharide of pathogenic bacteria; it is involved in host-bacterium interactions and the establishment of infection (17).

Little pathway information is available for *V.cholerae*, despite it being a serious pathogen. VchoCyc can help integrate the currently available knowledge, and it can help by pointing to omissions in our knowledge (such as missing pathways, as well as pathway holes). We hope and encourage the cholerae community to adopt the VchoCyc database as a tool for sharing knowledge and for suggesting promising directions for research.

The query, visualization and editing facilities of Pathway Tools provide extensive analysis capabilities to researchers. The metabolic overview expression viewer allows researchers to visualize expression data in a metabolic context, and thus differential metabolic gene expression can be discovered easily under different conditions. Therefore, we envision that PGDBs such as VchoCyc generated by Pathway Tools provide not only a useful platform to integrate and disseminate knowledge about this organism, but also provide a tool that facilitates analysis and discovery.

One might argue that we do not need computer software like Pathway Tools to predict pathways and that a biochemist could predict pathways just by looking at the list of annotated genes. Prediction by hand is possible, but the question is how accurate and complete it is. It is difficult to predict pathways accurately because some reactions are present in many pathways, and it is difficult to be complete by hand because there are so many known metabolic pathways to consider (see our analysis of *Helicobacter pylori* metabolism in which PathoLogic predicted many pathways whose presence was not recognized by *H. pylori* experts). Also, manual prediction requires so much time that it is prohibitive to perform multiple manual pathway predictions for evolving annotations of a single genome. Thus, a tool to automatically evaluate experimental evidence supporting predicted pathways is of practical value.

SUPPLEMENTARY DATA

Supplementary data are available at *NAR* Online.

ACKNOWLEDGEMENTS

The authors thank Drs E. Baron, R. Caspi, M. Green, N. Dolganov, S. Cohen, M. Walker, Y. Feng, C. Miller and J. Huang for helpful discussions and advice. This research was supported by DARPA contract N66001-01-C-8011. J.S. was supported by Stanford Graduate Fellowship. Funding to pay the Open Access publication charges for this article was provided by National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Lobitz,B., Beck,L., Huq,A., Wood,B., Fuchs,G., Faruque,A.S. and Colwell,R. (2000) Climate and infectious disease: use of remote sensing for detection of *Vibrio cholerae* by indirect measurement. *Proc. Natl Acad. Sci. USA*, **97**, 1438–1443.
2. Colwell,R.R. (1996) Global climate and infectious disease: the cholera paradigm. *Science*, **274**, 2025–2031.
3. Schoolnik,G.K. and Yildiz,F.H. (2000) The complete genome sequence of *Vibrio cholerae*: a tale of two chromosomes and of two lifestyles. *Genome Biol.*, **1**, REVIEWS1016.
4. Heidelberg,J.F., Eisen,J.A., Nelson,W.C., Clayton,R.A., Gwinn,M.L., Dodson,R.J., Haft,D.H., Hickey,E.K., Peterson,J.D., Umayam,L. *et al.* (2000) DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae. *Nature*, **406**, 477–483.
5. Karp,P.D., Krummenacker,M., Paley,S. and Wagg,J. (1999) Integrated pathway-genome databases and their role in drug discovery. *Trends Biotechnol*, **17**, 275–281.
6. Karp,P.D. (2000) An ontology for biological function based on molecular interactions. *Bioinformatics*, **16**, 269–285.
7. Karp,P.D. (2001) Pathway databases: a case study in computational symbolic theories. *Science*, **293**, 2040–2044.
8. Paley,S.M. and Karp,P.D. (2002) Evaluation of computational metabolic-pathway predictions for Helicobacter pylori. *Bioinformatics*, **18**, 715–724.
9. Caspi,R., Foerster,H., Fulcher,C.A., Hopkinson,R., Ingraham,J., Kaipa,P., Krummenacker,M., Paley,S., Pick,J., Rhee,S.Y. *et al.* (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.*, **34**, D511–D516.
10. Green,M.L. and Karp,P.D. (2004) A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, **5**, 76.
11. Murray,P.R., Baron,E.J., Jorgensen,J.H., Pfaller,M.A. and Yolken,R.H. (2003) *Manual of Clinical Microbiology*, *8th edn. Vols. 1 and 2*. American Society Microbiology, Washington D.C., USA.
12. Merrell,D.S., Butler,S.M., Qadri,F., Dolganov,N.A., Alam,A., Cohen,M.B., Calderwood,S.B., Schoolnik,G.K. and Camilli,A. (2002) Host-induced epidemic spread of the cholera bacterium. *Nature*, **417**, 642–645.
13. Griffiths,G.L., Sigel,S.P., Payne,S.M. and Neilands,J.B. (1984) Vibriobactin, a siderophore from *Vibrio cholerae*. *J. Biol. Chem.*, **259**, 383–385.
14. Marshall,C.G., Hillson,N.J. and Walsh,C.T. (2002) Catalytic mapping of the vibriobactin biosynthetic enzyme VibF. *Biochemistry*, **41**, 244–250.
15. Wyckoff,E.E., Smith,S.L. and Payne,S.M. (2001) VibD and VibH are required for late steps in vibriobactin biosynthesis in *Vibrio cholerae*. *J. Bacteriol*, **183**, 1830–1834.
16. Pflughoeft,K.J., Kierek,K. and Watnick,P.I. (2003) Role of ectoine in *Vibrio cholerae* osmoadaptation. *Appl. Environ. Microbiol.*, **69**, 5919–5927.
17. Webb,N.A., Mulichak,A.M., Lam,J.S., Rocchetta,H.L. and Garavito,R.M. (2004) Crystal structure of a tetrameric GDP-D-mannose 4,6-dehydratase from a bacterial GDP-D-rhamnose biosynthetic pathway. *Protein Sci.*, **13**, 529–539.