Published online August 7, 2006

Nucleic Acids Research, 2006, Vol. 34, No. 13 ***3687–3697***
doi:10.1093/nar/gkl438

The outcomes of pathway database computations depend on pathway ontology

M. L. Green* and P. D. Karp*

Bioinformatics Research Group, Artificial Intelligence Center, SRI International, Menlo Park, CA 94025, USA

Received March 10, 2006; Revised June 3, 2006; Accepted June 5, 2006

ABSTRACT

Different biological notions of pathways are used in different pathway databases. Those pathway ontologies significantly impact pathway computations. Computational users of pathway databases will obtain different results depending on the pathway ontology used by the databases they employ, and different pathway ontologies are preferable for different end uses. We explore differences in pathway ontologies by comparing the BioCyc and KEGG ontologies. The BioCyc ontology defines a pathway as a conserved, atomic module of the metabolic network of a single organism, i.e. often regulated as a unit, whose boundaries are defined at high-connectivity stable metabolites. KEGG pathways are on average 4.2 times larger than BioCyc pathways, and combine multiple biological processes from different organisms to produce a substrate-centered reaction mosaic. We compared KEGG and BioCyc pathways using genome context methods, which determine the functional relatedness of pairs of genes. For each method we employed, a pair of genes randomly selected from a BioCyc pathway is more likely to be related by that method than is a pair of genes randomly selected from a KEGG pathway, supporting the conclusion that the BioCyc pathway conceptualization is closer to a single conserved biological process than is that of KEGG.

INTRODUCTION

Pathway bioinformatics has become an increasingly active area of research in the past 5 years, yet different researchers mean very different things by the word 'pathway'. For example, 'pathway' can denote a metabolic pathway involving a sequence of enzyme-catalyzed reactions of small molecules, or a signaling pathway involving a set of protein-phosphorylation reactions and gene regulation events.

We argue that understanding the notion of pathways used by each pathway database (DB) is a critical part of understanding that resource. Furthermore, we argue that the use of different pathway conceptualizations can lead to different outcomes from a computational study that relies on pathway databases. By analogy, just as a study of associations between single-nucleotide polymorphisms and disease prevalence must pay extremely careful attention to the human populations from which the sequence data were drawn, studies using pathway data, or computational tools based on pathway data, must ensure that the conceptualization of pathway data that they choose is appropriate for the question or task at hand. Different pathway ontologies are best suited for different tasks, and can often be complementary.

We explore these issues in the context of two well-known metabolic pathway databases, KEGG (1) and BioCyc (2). We show how their pathway ontologies differ, and we investigate the suitability of different conceptualizations for different computational uses of pathway data.

METHODS

Datasets utilized

Our experiments employed Lisp queries against version 9.0 of the BioCyc DBs within the software/database bundle that combines BioCyc DBs with the Pathway Tools software. BioCyc is a collection of >200 databases (2) where each database describes one organism; e.g. EcoCyc describes *E.coli*. EcoCyc is a manually curated DB, whereas the metabolic networks in the other BioCyc DBs used in this study were predicted computationally based on the MetaCyc pathway DB, followed in some cases by limited addition of pathways from the literature for a given organism (3). To identify the genes in each KEGG pathway for each organism, we used the KEGG KGML v.0.4 dataset of metabolic pathways. KEGG metabolic networks are generated by a combined computational and manual approach (1). Table 1 summarizes the datasets used in our experiments including the number of metabolic genes and pathways in each organism.

Each genome-context experiment was performed against a single dataset, where a dataset is considered to be a database description of the metabolic map of a single organism as

*To whom correspondence should be addressed. Tel: +1 650 859 4358; Fax: +1 650 859 3735; Email: pkarp@ai.sri.com
*Correspondence may also be addressed to M. L. Green. Tel: +1 650 859 5669; Fax: +1 650 859 3735; Email: green@ai.sri.com

Table 1. Summary of each organism's metabolic network as defined in KEGG and BioCyc

Organism	Number of genes in organism	Number of KEGG metabolic genes	Number of BioCyc metabolic genes	Number of KEGG maps/ number of BioCyc pathways
E.coli K12	4475	833	684	98/162
C.crescentus	3818	595	475	88/138
M.tuberculosis	3966	702	525	96/154
H.pylori	1609	323	259	66/96
V.cholerae	3950	623	499	84/171

derived from KEGG or BioCyc, such as the EcoCyc dataset, the CauloCyc dataset for *Caulobacter crescentus*, the KEGG *E.coli* dataset, or the KEGG dataset for *C.crescentus*. Each experiment considered either chromosomal proximity, phylogenetic profiles, gene clusters or gene fusions. And each experiment considered pairs of genes from different regions of the metabolic map, such as pairs of genes within the same EcoCyc pathway or pairs of genes within the same KEGG map, or pairs of genes within the same KEGG map that are not in the same EcoCyc pathway.

Genome context experiments

We used data from Prolinks (4) to determine if two genes were related by one of the four genome context methods. Each method uses a different metric to compute functional relatedness as described by Bowers *et al*. (4). In each experiment our program chose 10 000 pairs of genes A and B at random from a region of a metabolic pathway or map, and counted the occurrence of one of the following:

- The number of times that genes A and B are conserved gene neighbors, meaning that the two genes are found in close proximity across multiple genomes. In the Prolinks database, the probability that the genes are functionally related (*P*-value) is computed based on the number of intervening genes in each organism and the number of organisms including homologs to both genes. Bowers *et al*. (4) have shown that conserved gene neighbors provided the most accurate and extensive coverage in recovering protein pairs assigned to the same COG pathway when compared with phylogenetic profiles, gene clusters and gene fusion events.
- The number of times that genes A and B have similar phylogenetic profiles. In the Prolinks dataset, the probability that the two genes have coevolved has been computed based on the numbers of genomes containing homologs of each protein and the number of genomes containing homologs of both proteins.
- The number of times that genes A and B appear in the same gene cluster. In the Prolinks dataset, the gene cluster method computes the probability that a pair of adjacent, closely spaced genes is part of an operon based on the probability of finding a smaller gap distance.
- The number of times that genes A and B, expressed as separate proteins in the organism of interest, exist as a fused gene in another organism. In the Prolinks dataset, these pairs are ranked by the probability that the two genes might be found linked by chance, based on the number of homologs to genes A and B in the database.

For each organism we retrieved from Prolinks a list of gene pairs, where each pair is related by one of the above methods (gene neighbor, gene cluster, phylogenetic profile and gene fusion). In each experiment we performed, we deemed all retrieved pairs as related by the stated relationship regardless of the confidence level of the association as scored in the Prolinks dataset (confidence levels reflect the recovery of COG pathway assignments at a certain *P*-value). Hence, a gene pair has similar phylogenetic profiles if the pair is related by the phylogenetic profile method in the Prolinks dataset for the organism. Note that a pair of genes may be related by multiple genome context methods. The Supplementary Data file for each organism includes two tables showing all pairs of genes from the same KEGG or BioCyc pathway for which we found data in the Prolinks dataset (i.e. all gene pairs occurring in the same KEGG or BioCyc pathway that are related according to one or more of the Prolinks methods).

Data availability

Virtually all BioCyc DBs, including EcoCyc, are freely and openly available to all users, and may be redistributed. To further facilitate use of EcoCyc as a training set for genome context methods, we have created a new flat file that contains all of EcoCyc's functional associations among *E.coli* genes. The file (func-associations.col) is available through the Software/Data Download section of the EcoCyc website (http://BioCyc.org/download.shtml). This file includes genes grouped by metabolic pathway and protein complex, as well as pairs of transcription factor genes and regulated genes.

RESULTS

Conceptualizations of metabolic pathways in KEGG and BioCyc

To understand the ontologies of metabolic pathways used by a given database project, we must understand the rules or principles used by that project to determine what will and will not be defined as a metabolic pathway in their database. The more precise and rigorous those rules are, the more clearly articulated is the conceptualization. We note that most pathway databases do not provide a clear statement of the pathway conceptualization that they use.

Pathways of small-molecule metabolism are generally considered to have the following properties. These pathways involve two types of chemical entities: the small molecular-weight chemical compounds that are transformed by the pathway, and the enzymes that accelerate the biochemical reaction steps within the pathway but that are unchanged by the pathway. There are rare exceptions to these rules, for example, in some cases protein substrates to metabolic

pathways are found (often as carriers of small molecules), and a small number of reactions within small-molecule metabolic pathways occur spontaneously (with no enzyme).

Metabolic pathways consist of linked sequences of enzyme-catalyzed reactions—linked in the sense that the small-molecule product (output) of one reaction in the pathway becomes the reactant (input) of the next reaction in the pathway. Metabolic pathways are not always linear sequences—some pathways exhibit a branching tree structure, whereas other pathways contain cycles.

A critical notion in defining pathways concerns how to define their start and end points. Put another way, how do we choose to divide the complete metabolic reaction network of the cell, which for a typical bacterium will include 500–1000 reaction steps, into fragments that we call pathways? Pathways after all connect to one another to form a large biochemical network, but what rules are used to separate the connection points from the interior nodes of pathways? Do hard-and-fast rules exist, which are firmly grounded in biological principles, or are pathway boundaries purely arbitrary human constructs? One argument that these boundaries are not arbitrary is that in our experience, the same pathway start and end points are often identified by multiple criteria for delimiting pathways.

Recent advances have resulted in mathematical methods for defining pathway boundaries, termed extreme pathways and elementary modes (5). We are unaware of any analysis of how the pathway boundaries inferred from these methods correspond to experimentally elucidated pathways or to the BioCyc or KEGG conceptualizations; therefore, it is unclear whether these methods can contribute to the problem of delimiting pathway boundaries.

Here we present the rules used by the BioCyc family of databases to conceptualize metabolic pathways (the rules are used most actively in curation of the EcoCyc and MetaCyc DBs, which are the most intensively curated DBs in the BioCyc collection.) The goal of our pathway curation efforts is to define metabolic pathways that correspond to individual biological processes, and to conserved, functional, atomic modules of the metabolic network. We note that few if any of these rules are absolute within biology (or BioCyc), and that some of these rules are new to the BioCyc curation methodology, and are therefore not reflected in the definitions of some older BioCyc pathways.

BioCyc Rule 1: Find a common biological process

BioCyc pathways consist of a linked set of reaction steps that participate in a single biological process, such as biosynthesis of tryptophan, or degradation of arginine. For a set of reactions to be part of the same process, most or all of those reactions must be active simultaneously, i.e. we cannot refer to a pathway as a single biological process if different parts of the pathway are expressed or activated at very different points in time, or under different growth conditions, or are mutually exclusive, or do not occur together in the same organism.

In contrast, KEGG maps contain large segments that cannot be part of a single process because they are mutually exclusive (e.g. biosynthesis and degradation of the same metabolite), and because different segments of a KEGG map are found in different organisms. That is, KEGG *reference maps* are mosaics that combine pathway information from multiple organisms. KEGG uses the term *reference map* to refer to a single pathway diagram on a single KEGG web page, which, again, integrates information from multiple organisms. Note that KEGG maps can be colored to show which reactions within a given reference map occur in a given organism, based on the set of enzymes identified in its genome.

BioCyc Rule 2: Define pathway boundaries at high-connectivity substrates

BioCyc defines pathway boundaries at high-connectivity metabolites—branching points within the metabolic network. Such branching points typically correspond to decision points within the cell, or to junctions between different processes. The most common high-connectivity metabolites are 13 metabolites of central metabolism within glycolysis, the TCA cycle and the pentose phosphate pathway. The 13 metabolites are glucose-6-phosphate, fructose-6-phosphate, ribose-5-phosphate, erythrose-4-phosphate, triose phosphate, 3-phosphoglycerate, phosphoenolpyruvate, pyruvate, acetyl CoA, alpha-oxoglutarate, succinyl CoA, oxaloacetate and sedoheptulose-7-phosphate.

Because those 13 metabolites supply the carbon skeletons of all cellular components manufactured by the cell, BioCyc biosynthetic pathways typically begin with these intermediates and end with a building block of macromolecules, a cofactor or a coenzyme. BioCyc catabolic pathways typically end at one of the 13 central metabolites.

BioCyc Rule 3: Define pathway boundaries at stable substrates rather than transient intermediates

BioCyc defines pathway boundaries at stable metabolites rather than at metabolites that decay quickly.

BioCyc Rule 4: Pathway steps share common regulation

BioCyc pathways are often regulated as a unit, through a variety of biological mechanisms. For example, the five reactions of tryptophan biosynthesis in *E.coli* are regulated as a unit at the level of substrate-level enzyme inhibition. In addition, the genes that code for the enzymes within the pathway are regulated within a single operon by both repression and attenuation. All three regulatory mechanisms delimit the same pathway boundaries.

Note that a hierarchy of levels of genetic regulation exists, including the operon level (contiguous genes that are expressed within a single transcript), the regulon level (multiple operons regulated by the same transcription factor) and the stimulon level (all responses to a common stimulus). There is no consistent rule as to which of these levels maps to a single pathway. For example, sometimes all genes within a pathway are within a single operon, sometimes they are within a single regulon, and sometimes they are within a single stimulon.

Entire KEGG maps are rarely if ever regulated as a unit because they combine segments from different biological processes, and segments that occur in different organisms.

BioCyc Rule 5: Pathways exhibit evolutionary conservation

Another criterion for delimiting pathway boundaries comes from evolutionary conservation of pathways, whereby the same set of reactions is either present or not present in multiple organisms as an atomic unit. Evolutionary considerations shape BioCyc pathway boundaries, and our experiences in pathway prediction across many genomes sometimes lead our curators to partition pathways into smaller units when those units are found to occur independently in different organisms. For example, our curators had originally defined a single pathway for the synthesis of coenzyme A from 2-keto-isovalerate, with pantothenate as an intermediate in the pathway (Figure 1A and B). Although that entire pathway was present in many organisms, we noticed some organisms that contained only the first half of the pathway producing pantothenate (such as *Helicobacter pylori*), and other organisms that contained only the second half of the pathway starting from pantothenate (such as *Homo sapiens*). These observations led us to split this pathway into two new pathways—pantothenate biosynthesis and coenzyme A biosynthesis—and to define a superpathway that combines the two pathways in organisms that have both. This approach allows more accurate prediction of which individual pathways are present in a given organism.

As a second example, consider the pathway teichoic acid (poly-glycerol) biosynthesis in Figure 2. Our Pathway Tools software predicts (2) this pathway to be present in *Sinorhizobium meliloti* because enzymes catalyzing two steps in this pathway are present in *S.meliloti*. However, those two steps form a branch of the teichoic acid biosynthesis pathway that synthesizes UDP-D-glucose, which is one of several inputs in the formation of teichoic acid. Because UDP-D-glucose is involved in many other metabolic pathways (it is a reactant in 49 MetaCyc pathways), and because none of the enzymes that synthesize compounds along the main backbone of the teichoic acid biosynthesis pathway is present, we consider it incorrect to state that *S.meliloti* has a pathway for biosynthesis of teichoic acid. Biosynthesis of UDP-D--glucose should be treated as a separate biochemical module, and therefore in the latest version of MetaCyc we have defined a separate pathway for its biosynthesis, and removed this branch from the pathway for biosynthesis of teichoic acid. These changes should in the future prevent the false-positive prediction that teichoic acid biosynthesis is present in *S.meliloti*.

Summary of rules of pathway conceptualization

BioCyc pathways, like pathways defined in the experimental literature, correspond to single biological processes that take place in a single organism. They are evolutionarily conserved and are regulated as a unit. Their boundaries are defined at stable, high-connectivity chemical substrates.

As will become clear from the examples that follow, the KEGG conceptualization of pathways shares few if any of these principles. We have not been able to find a statement of the principles underlying the KEGG pathway ontology in KEGG publications or on the KEGG website. Therefore, the following statements are based on our inferences of the KEGG pathway ontology.

We infer that KEGG metabolic pathways are designed to be mosaics of related groups of reactions from multiple organisms that accomplish similar functions or that accomplish the biosynthesis, degradation or interconversion of related substrates; i.e. KEGG pathways are substrate-centric, and are aggregated from multiple organisms, thus integrating all possible transformations of a given substrate in all known organisms.

Global properties of KEGG and BioCyc pathways

We explore the effects of these different conceptualizations by examining different global properties of KEGG and BioCyc pathways that we believe stem in large part from the different conceptualizations that these databases employ.

KEGG maps tend to be much larger than the BioCyc pathways (Figure 3). For example, if we compare the *E.coli* KEGG maps with the EcoCyc DB, we see that KEGG contains 102 maps with at least one reaction marked as present in *E.coli*, whereas EcoCyc contains 162 metabolic pathways. The average KEGG reference map (i.e. the organism-independent maps used for pathway prediction) contains 21 reactions; the average KEGG *E.coli* map contains 9.1 reactions (i.e. KEGG finds *E.coli* enzymes for on average 9.1 reactions in each reference map, because a reference map usually combines reactions from many organisms, not all of which occur in *E.coli*). In contrast, the average EcoCyc pathway contains 5.0 reactions. EcoCyc partitions the *E.coli* metabolic network into a larger number of smaller pathways than does KEGG.

KEGG maps tend to be larger because many KEGG maps combine reactions from multiple biological processes. For example, the single KEGG map entitled 'arginine and proline metabolism' contains reactions found in the following EcoCyc pathways: 'proline biosynthesis I', 'proline degradation I', 'arginine biosynthesis II', 'arginine degradation VI' and 'arginine degradation XII'. It also contains reactions for charging of prolyl- and arginyl-tRNAs. Similarly, the KEGG map entitled 'cysteine metabolism' contains reactions from the following EcoCyc pathways: 'cysteine biosynthesis I' and 'L-cysteine degradation II'. It also contains a reaction for charging of cysteinyl-tRNA. These observations reinforce our conjecture that KEGG pathways are designed to be mosaics that integrate information from many organisms about those biological processes that impinge upon a given substrate, or a combination of substrates.

Comparison of example KEGG and BioCyc pathway

Here we illustrate the different pathway conceptualizations of KEGG and BioCyc by comparing the *E.coli* version of the 'Aminosugarsmetabolism'map from KEGG (eco00530) with the corresponding EcoCyc pathways. The KEGG Aminosugars metabolism map includes 21 genes for *E.coli*. These 21 genes occur in six different EcoCyc pathways, as shown in Table 2. Considering genes from EcoCyc pathways that contain more than one of the genes in the list of genes in Table 2, we computed the number of pairs of conserved gene neighbors (Methods) and genes with similar phylogenetic profiles within the 'Aminosugars metabolism' KEGG map, and within the individual EcoCyc pathways. The EcoCyc pathways considered were 'UDP-*N*-acetylglucosamine

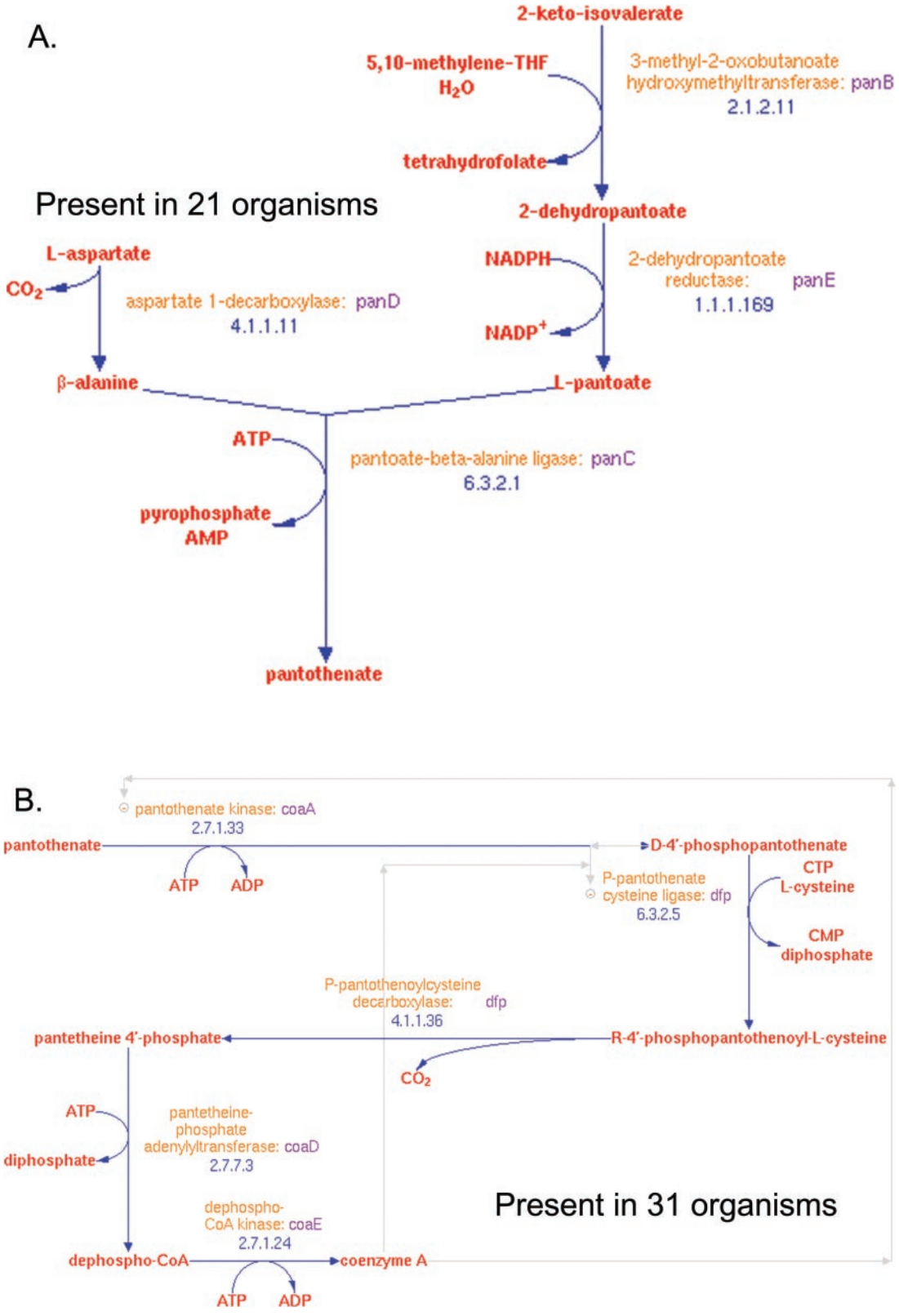

Figure 1. (**A**) Pathway for biosynthesis of pantothenate from 2-keto-isovalerate. (**B**) Pathway for biosynthesis of coenzyme A from pantothenate. Of the 204 BioCyc organisms, 31 include the pathway for biosynthesis of coenzyme A, but lack the pantothenate biosynthesis pathway. Another 21 organisms include the pathway for pantothenate biosynthesis, but lack the coenzyme A biosynthesis pathway. A total of 108 organisms include both pathways, while 44 lack both pathways.

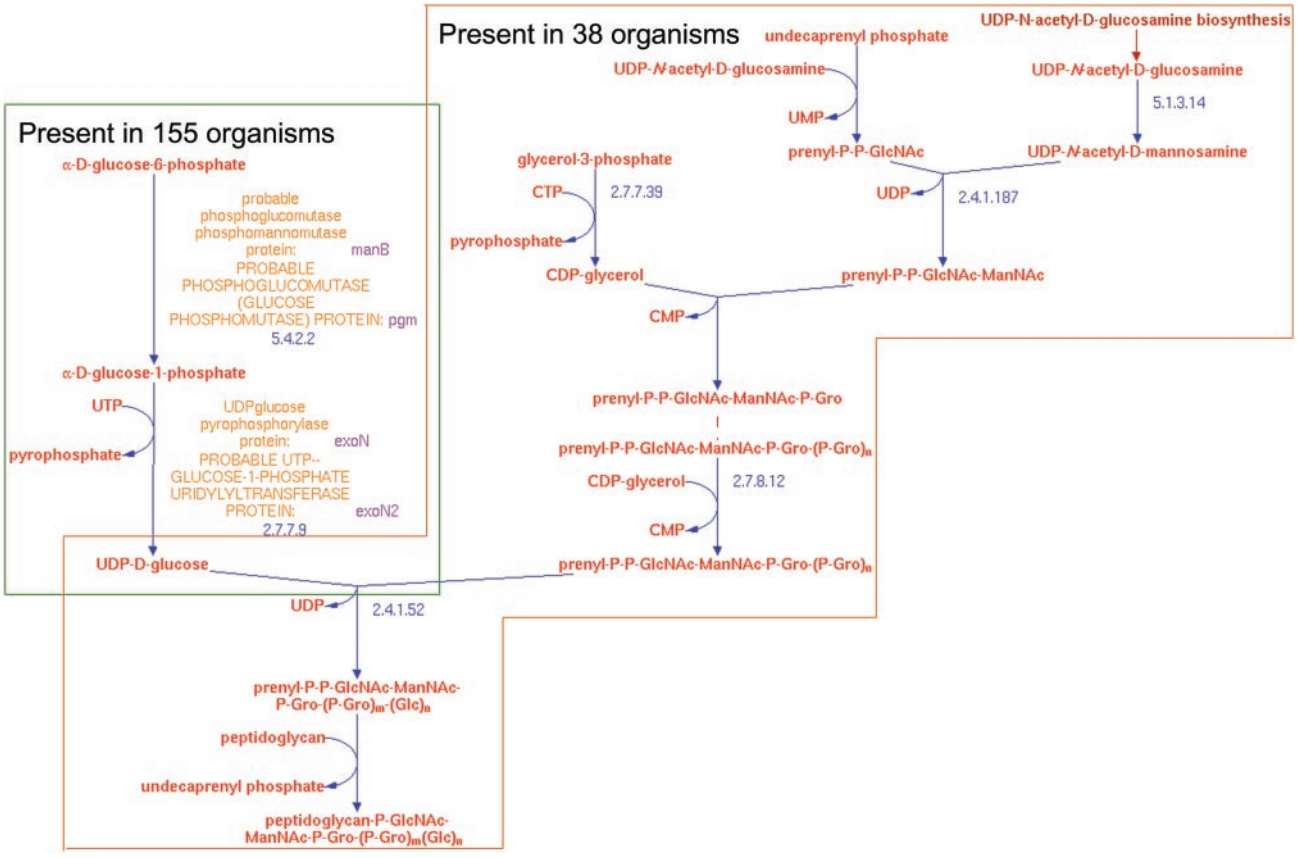

Figure 2. Predicted pathway for teichoic acid biosynthesis in *S.meliloti*. Only where enzyme names and gene names are shown has an enzyme catalyzing a reaction in this pathway been identified in *S.meliloti*. Of the 204 BioCyc organisms, 155 include the branch of the teichoic acid biosynthesis pathway that synthesizes UDP-D-glucose. Only 38 organisms include any of the remaining reactions in the pathway, and only 31 of these include reactions from both branches.

biosynthesis', 'peptidoglycan biosynthesis' and 'glucosamine degradation'.

Within the KEGG map, out of a total of 210 gene pairs there are 5 conserved neighbor pairs, 3 pairs with similar phylogenetic profiles, 1 pair related by gene fusion and 2 pairs related by gene clusters. Among the set of EcoCyc pathways, there are a total of 70 gene pairs. From these 70 pairs, 29 pairs are conserved gene neighbors, 39 pairs have similar phylogenetic profiles, 4 pairs are related by gene fusion and 8 pairs are related by gene clusters. Clearly, in this case EcoCyc's pathways, each corresponding to a single biological process, include a greater number of gene pairs that are deemed as related by one of these genome context methods. (Note that the EcoCyc pathways include 10 genes not found in this KEGG map, thus permitting the higher number of gene pairs related by genome context methods.)

Functional relatedness of genes within KEGG and BioCyc pathways

Consider the question of which conceptualization of pathways corresponds more closely to a functional biological unit of the cell? And which conceptualization corresponds more closely to atomic biological processes that are conserved as a unit through evolution? One way to address these questions is to choose pairs of genes at random, within a single KEGG pathway, or within a single BioCyc pathway, and to evaluate the relative frequency with which those genes are functionally related.

Genome-context methods (4,6–22) infer that two genes A and B are functionally related based on evidence from patterns conserved across many genomes. Two example genome context methods are the conserved chromosomal proximity method (4,8,10,20), which infers that genes A and B are functionally related if orthologs of A and B across many organisms are nearby on the chromosome; and the phylogenetic profile method (4,12,19), which infers that genes A and B are functionally related if orthologs of A and B have similar patterns of presence and absence across many genomes. We use genome context methods to evaluate the degree to which randomly chosen genes within a KEGG or a BioCyc pathway are functionally related, as detected by these methods.

For the KEGG *E.coli* dataset and the EcoCyc pathway-genome database (PGDB), we compared the number of times genes A and B appeared to be conserved chromosomal neighbors out of 10 000 random selections of pairs of genes occurring in the same EcoCyc pathway, or in the same KEGG *E.coli* metabolic map. Similarly, we assessed whether gene pairs chosen from the same pathway/map had similar phylogenetic profiles, appear in a gene cluster, or are related by gene fusion as defined in the methods section. As shown in Figure 4, two genes chosen at random from the same EcoCyc pathway were 3.8 times more likely to be conserved

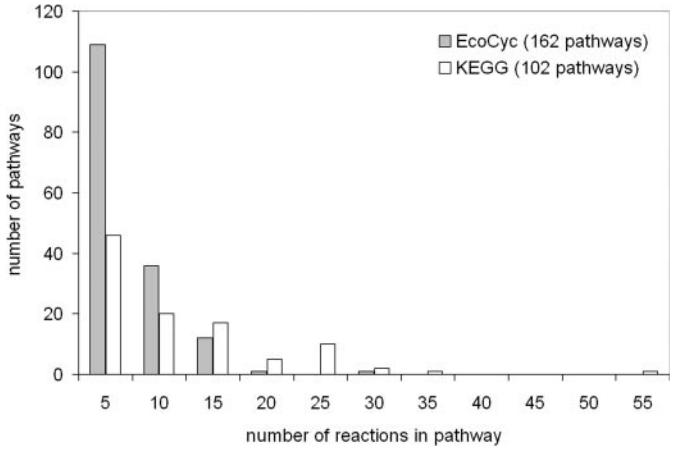

Figure 3. Size distribution of the KEGG and EcoCyc metabolic pathways.

Table 2. Genes from the 'Aminosugars metabolism' map in KEGG and the EcoCyc pathway(s) that include each gene

Gene	KEGG EC number	EcoCyc pathway(s) that include the gene (gene name shown when not in a pathway, with EC number in EcoCyc when available)
b0211	3.2.1.-	mltD
b0271	3.2.1.-	yagH (3.2.1.37)
b0677	3.5.1.25	Glucosamine degradation
b0678	3.5.99.6	Glucosamine degradation
b0679	2.7.1.69	nagE (transporter)
b1084	3.1.4.-	rne
b1107	3.2.1.52	nagZ
b1193	3.2.1.-	emtA
b2340	3.1.3.-	ArcAB two-component signal transduction system
b2701	3.2.1.-	mltB
b2813	3.2.1.-	mltA
b2963	3.2.1.-	mltC
b3189	2.5.1.7	Peptidoglycan biosynthesis
b3223	5.1.3.9	*N*-acetylglucosamine, *N*-acetylmannosamine and *N*-acetylneuraminic acid dissimilation (superpathway)
b3225	4.1.3.3	*N*-acetylglucosamine, *N*-acetylmannosamine and *N*-acetylneuraminic acid dissimilation (superpathway)
b3247	3.1.4.-	Rng
b3729	2.6.1.16	UDP-*N*-acetylglucosamine biosynthesis
b3730	2.7.7.23	Peptidoglycan biosynthesis, UDP-*N*-acetylglucosamine biosynthesis
b3786	5.1.3.14	Enterobacterial common antigen biosynthesis
b3972	1.1.1.158	Peptidoglycan biosynthesis
b4392	3.2.1.-	Slt

chromosomal neighbors than two genes chosen at random from the same *E.coli* KEGG map. Figure 5 shows that two genes chosen at random from the same EcoCyc pathway were 2.3 times more likely to have similar phylogenetic profiles than two genes chosen at random from the same *E.coli* KEGG map. Figures 6 and 7 show the results from the gene cluster and the gene fusion methods. Two genes from the same EcoCyc pathway were 4.8 or 3.8 times more likely to be in a conserved gene cluster or to be conserved gene neighbors.

The black bars in Figures 4–7 display a third group of gene pairs, namely, pairs that lie in the same KEGG map but that do not appear in the same EcoCyc pathway. This third

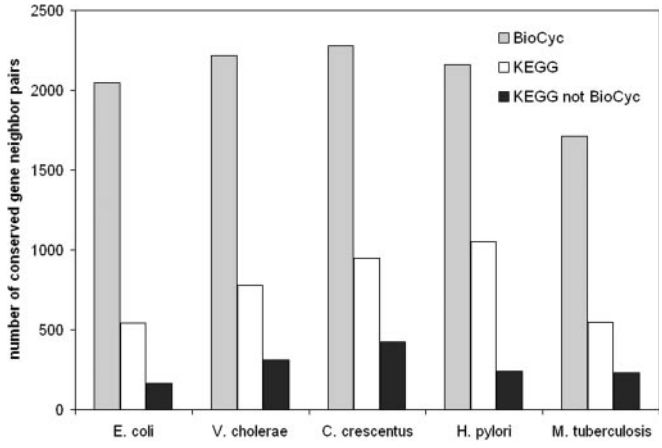

Figure 4. Number of conserved gene neighbors. Each pair of genes is selected randomly from a single KEGG metabolic map or from a single BioCyc metabolic pathway.

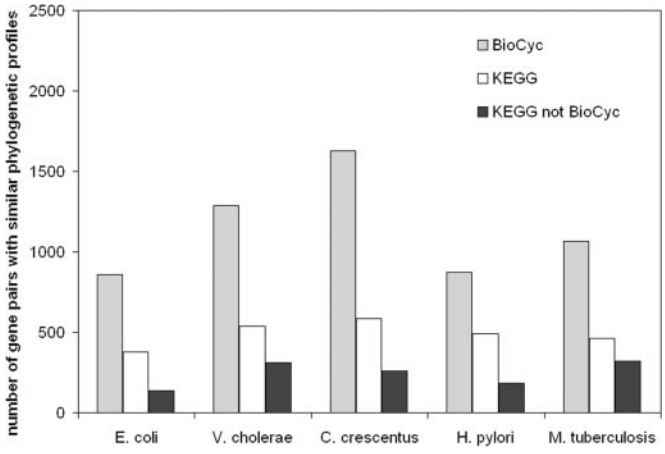

Figure 5. Number of gene pairs with similar phylogenetic profiles. Each pair of genes is selected randomly from a single KEGG metabolic map or from a single BioCyc metabolic pathway.

group essentially represents those gene pairs for which the KEGG map asserts a functional relationship but no EcoCyc pathway supports that relationship. The genome context methods show a very small degree of functional associations for these genes.

For several organisms in addition to *E.coli*, we repeated the same comparisons using data from each organism's respective KEGG or BioCyc dataset, and Prolinks data. The results for these organisms are also shown in Figures 4–7. In each case, we observe the same trend that was observed for *E.coli*. The greatest number of conserved neighbor gene pairs is found in sets of genes drawn from the same BioCyc pathway. Considerably fewer pairs of conserved neighbor genes are selected from a set of genes drawn from the same KEGG pathway map and even fewer from the set of genes in the same KEGG pathway map that do not occur together in a BioCyc pathway. Two genes chosen at random from pathways in BioCyc DBs other than EcoCyc were 2.6 times (compared with 3.8 times for EcoCyc) more likely to be conserved gene neighbors than two genes chosen at random from the KEGG map for that organism. This lower number may result

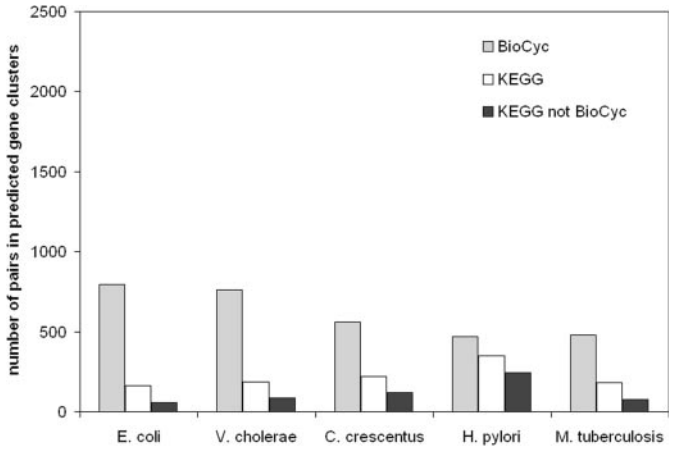

Figure 6. Number of gene pairs occurring in a predicted gene cluster. Each pair of genes is selected randomly from a single KEGG metabolic map or from a single BioCyc metabolic pathway.

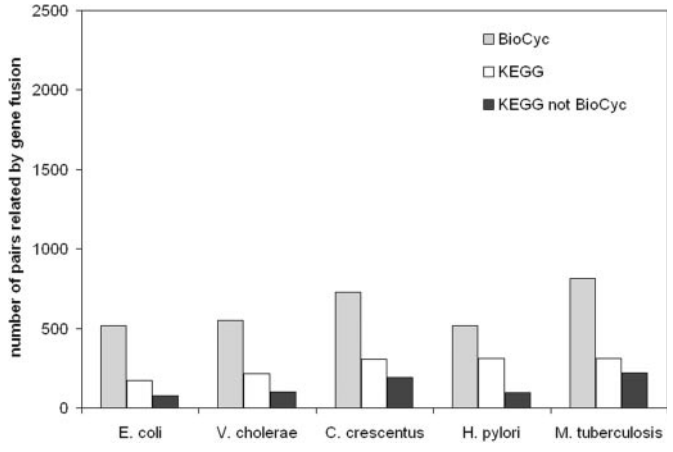

Figure 7. Number of gene pairs related by gene fusion events. Each pair of genes is selected randomly from a single KEGG metabolic map or from a single BioCyc metabolic pathway.

from the fact that the other BioCyc DBs have undergone much less literature-based curation than has EcoCyc.

How do EcoCyc 'superpathways' compare?

Could the differing size of KEGG and EcoCyc pathways alone be responsible for the preceding results?

Several of the metabolic pathways in the EcoCyc database are organized into entities referred to as 'superpathways'. A superpathway is an aggregation of two or more base EcoCyc pathways that are related in some way, usually because the pathways are connected through common substrates. Superpathways are intended to convey relationships among base pathways to the user. EcoCyc contains 31 superpathways. The average number of reactions per superpathway is 14.4, compared to 5.1 reactions per regular EcoCyc pathway and 9.1 reactions for the average KEGG pathway for *E.coli*.

We performed the same type of experiment on all of EcoCyc's 31 superpathways, counting the number of conserved gene neighbor pairs in 10 000 randomly selected gene pairs, where each pair of genes was drawn from the same superpathway. Figure 8 displays the data for EcoCyc's superpathways, EcoCyc's standard pathways and KEGG's metabolic maps. The number of gene neighbor pairs randomly selected from the same EcoCyc pathway is about twice the number of gene neighbor pairs randomly selected from the same EcoCyc superpathway, which in turn is also approximately twice the number selected from the same KEGG pathway. Therefore, although EcoCyc superpathways are on average larger than KEGG pathways, more gene pairs in each pathway are related by genome context methods than are genes within a KEGG map.

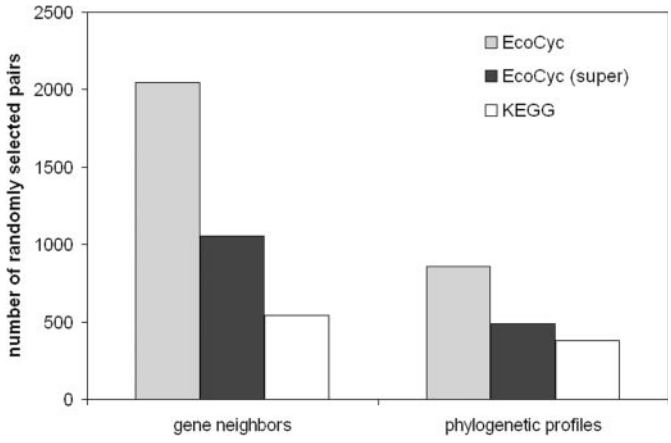

Figure 8. Number of conserved gene neighbors and similar phylogenetic profiles randomly selected from EcoCyc superpathways compared to standard EcoCyc pathways and KEGG *E.coli* metabolic maps.

DISCUSSION

The preceding statistics clearly show that KEGG pathways are on average significantly larger than BioCyc pathways. Our analysis of example pathways indicates that the reason for this size difference is that individual KEGG pathways are substrate-centric—they combine multiple biological processes that impinge on a single substrate, such as the biosynthesis, catabolism and tRNA-charging of an amino acid. Individual KEGG pathways can contain alternative routes of biosynthesis (or catabolism) of a substrate, either from one organism or from multiple organisms.

Our experiments with genome-context methods have shown that the likelihood that a pair of randomly chosen genes from the same BioCyc pathway is functionally related is much greater than that of a pair of randomly chosen genes from the same KEGG pathway map. This result is consistent with the notion of KEGG pathway maps as mosaics that combine multiple biological processes.

What do these differences in pathway conceptualization imply about the suitability of BioCyc or KEGG pathways for particular computational purposes in bioinformatics? Here, we consider several uses of pathway databases in bioinformatics and the possible utility of these conceptualizations for each.

Purpose: encyclopedia of distinct metabolic processes present in a given organism

One goal for a pathway DB is to precisely encode all metabolic processes present in a given organism, as determined

either experimentally or computationally. We posit that the BioCyc pathway conceptualization is better suited to this task, because each BioCyc pathway database object represents as closely as possible the set of reactions occurring in one biological process in one organism. In contrast, a KEGG reference map by its very nature integrates reactions from many organisms, and from many biological processes, and thereby blurs the boundaries between those processes. A user viewing an uncolored reference map has no way to ascertain which parts of the map were originally elucidated in which organism(s), nor to ascertain which parts of the map work together as a single process. In contrast, every BioCyc pathway clearly indicates the organism to which it pertains. For example, pathways in the MetaCyc DB are each tagged with the names of the one or more species in which they were experimentally elucidated.

Although the organism-specific coloring of KEGG maps does paint a suggestive picture of what pathways might be present in the organism, the KEGG framework does not allow a pathway map to be precisely customized to the metabolism of the organism. For example, a pathway that is in fact absent from the organism, despite the presence in the organism of several enzymes from the pathway, cannot be removed (manually or otherwise) from the KEGG map. Nor can a KEGG map indicate how its component reactions are assigned to distinct biological processes that are regulated as a unit, that operate as a functional unit, and that evolve as a functional unit.

Purpose: encyclopedia of processes impinging on a given substrate

An alternative goal for a pathway DB is to communicate to the user all processes that contribute to the metabolism of a single substrate, or a set of related substrates. The KEGG pathway conceptualization is well suited to this task, because each KEGG map combines many processes related to a given substrate within one diagram.

BioCyc is also well suited to this task, but takes different approaches, under the philosophy that such relationships can be detected computationally in a pathway DB. A BioCyc user can understand the multiple processes impinging on one substrate by (i) viewing a software-generated compound page that shows all reactions and pathways that consume or produce that compound; (ii) viewing a super pathway that combines multiple pathways in which that compound is a substrate; or (iii) requesting that all occurrences of that compound be highlighted on the cellular overview diagram for that organism (e.g. http://biocyc.org/ECOLI/new-image?type=OVERVIEW) to depict all cellular pathways that metabolize that compound (such highlighting is supported in the desktop version of the software, but not through the website).

Purpose: pathway prediction or reconstruction

Pathway databases are commonly used to predict the pathway complement of a newly sequenced organism by analogy to known pathways from other organisms. In our opinion, pathway prediction will be more accurate if the pathway units used for prediction correspond as closely as possible to those biological processes that are conserved as units in evolution. When predictions are considered in terms of distinct biological processes, it is easier to identify cases where there is insufficient evidence to support the prediction of a pathway, e.g. when only a few enzymes for a pathway are present and those enzymes also catalyze reactions in other, more completely predicted pathways. If the pathways used for prediction include multiple biological processes, the probability of false-positive prediction of pathway components will be increased because the presence of enzymes for one process will incorrectly be interpreted as evidence for another process.

For example, consider the KEGG map 'Methane metabolism' as projected into *E.coli* K-12 MG1655 (http://www.genome.ad.jp/dbget-bin/get_pathway?org_name=eco&mapno=00680). The KEGG coloring of this map indicates that six of the enzymes of this pathway are present in *E.coli*. But is KEGG in fact predicting the presence of this pathway? One interpretation is to say yes, KEGG predicts the pathway as present in *E.coli* because the pathway map is selectable for *E.coli*, in contrast to the KEGG 'Photosynthesis' pathway, which is selectable only for 11 photosynthetic organisms in KEGG version 35.1. This interpretation is clearly the correct one for computational users of KEGG, who perform computations across all KEGG maps that are provided for a given organism, because programs do not have the biological wisdom that *E.coli* does not metabolize methane.

Another interpretation is that the inference as to whether this pathway is present in *E.coli* is left to the user—that the purpose of the KEGG site is simply to aid the user in making this final assessment.

E.coli is not known to produce or catabolize methane, and most of the six enzymes of this pathway are actually used in other *E.coli* pathways, supporting the notion that using such large pathways as predictive units increases the probability that some enzymes in the pathway will be present because of their role in other pathways, leading to false positive pathway inferences by both end users who visually inspect KEGG pathways, and by end users who compute with KEGG pathways.

A final point related to pathway prediction is that KEGG maps may have an advantage when the user is trying to distinguish which of several alternative pathway variants is present in a given organism. For example, a number of pathways of arginine catabolism are present in MetaCyc and in KEGG. The KEGG arginine catabolism map shows all these variants simultaneously, and by coloring all enzymes present in the organism on that map, the user can easily compare the evidence for each variant. In contrast, the BioCyc approach is to computationally compare the evidence for each variant, and to predict multiple alternative variants in a given organism if the pathway predictor finds significant evidence for more than one. The user must display each variant separately to manually compare the evidence for them.

Purpose: detection of missing pathway components

Another use of pathway DBs is to detect missing pathway components, or 'pathway holes' as we have called them in past work (23). Consider the KEGG pathway for 'Phenylalanine, tyrosine and tryptophan biosynthesis' (http://www.genome.ad.jp/dbget-bin/get_pathway?org_name=hsa&mapno=00400) as projected into *H.sapiens*. That KEGG page shows that 8 of

the 44 enzymes in this map are present in humans. Straightforward application of the pathway hole filling approach would instruct an algorithm to search the human genome for enzymes that catalyze all the other 36 steps in this KEGG map, even though it is well known that humans cannot synthesize phenylalanine, and even though, as a mosaic, this KEGG pathway probably contains reactions known from a wide variety of other organisms that would not be expected to be found in humans. One view is that large KEGG pathways are well suited to this task because it does not hurt to cast a wide net when looking for missing components. An opposing view is that the wider a net is cast, the more likely it is that we will find false positives, and one should search only for components of the specific biological processes that are predicted to be present.

Purpose: gold standard for developing methods that predict pathways

Developers of computational methods for predicting pathways train and evaluate their methods with respect to known pathways in pathway DBs. Developers of such methods should choose the pathway conceptualization that is closer to the type of pathway they want to predict. A learning program that is exposed to instances of the wrong concept will usually learn the wrong concept.

Purpose: gold standard for developing genome context methods

Many developers of genome context methods train and evaluate their methods against pathway DBs, under the rationale that all genes within a single pathway must be functionally related, and therefore the more frequently their genome-context method predicts that two genes within the same pathway are functionally related, the better is their method. Again, developers of such methods should choose a pathway DB that uses a conceptualization that is closest to the notion of 'functionally related' that they want to predict. A researcher who considers 'functionally related' to mean that two genes are involved in separate biological processes that impinge on a common substrate with no other evolutionary or regulatory constraints should choose KEGG pathways. A researcher who considers genes to be functionally related if they are involved in a single biological process, are regulated as a unit, and are conserved evolutionarily should choose BioCyc pathways.

Because these methods are likely to play a significant role in genome annotation in the future (24,25), accurate evaluation of the methods is critical in order to maximize their accuracy. We note that the definition of a functional association has always been vague in genome-context research, and that few publications in the field attempt to define the term precisely. Clarification of the goals of genome-context methods would aid in selecting an optimal evaluation strategy.

Purpose: analysis of omics datasets

Some scientists use pathway DBs to analyze large-scale data such as gene-expression measurements by viewing them within a pathway context (26,27). Various application programs perform this analysis by coloring steps within pathway diagrams with colors corresponding to gene expression levels. It is not clear that either of the two pathway conceptualizations is particularly better suited to this task. KEGG pathways alone allow the scientist to view data that encompasses several related biological processes, and are thus broader than a single BioCyc pathway. On the other hand, the BioCyc Omics Viewer allows omics data to be projected on a pathway map of the entire cell, for a broader view yet, with the ability to zoom in to see omics data on individual pathways or clusters of pathways. It is not clear that either ontology is better suited to this task.

Complementation of pathway databases

Given the complementary strengths and weaknesses of pathway DBs discussed in this article, it may be that for certain applications, different pathway DBs will complement each other and should be combined. Three projects that would support such complementation are BioWarehouse (28), which allows BioCyc and KEGG to be loaded side-by-side into a single relational DB system; BioPAX, a single format into which both BioCyc and KEGG can be converted (29,30); and SBML, another common data format for systems biology (30,31).

Limitations of our approach

We cannot say for certain that there does not exist some functional relationship between genes in the same KEGG pathway, but not in the same BioCyc pathway, that remains undetected using genome context methods. However, the bulk of available evidence suggests a tighter functional cohesiveness among genes within BioCyc than KEGG pathways.

'Membership in the same metabolic pathway' carries with it implications for regulatory and evolutionary relationships. Our results suggest that the extent to which these relationships hold can be impacted by the demarcation of the pathways themselves. Despite its limitations, the results from our analysis warrant more rigorous consideration of the types of relationships and degrees of relatedness predicted by methods developed using the variety of available metabolic pathway databases.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

The BioCyc pathway conceptualization and this paper benefited from discussions with Dr John Ingraham. We are also grateful to Drs Frederique Lisacek and Anne Morgat for comments on this manuscript. This work was supported by grants RR07861 from the NIH National Center for Research Resources, GM70065 from the NIH National Institute of General Medical Sciences, and DE-FG03-01ER63219 from the US Department of Energy. The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health or the Department of Energy. Funding to pay the Open Access publication charges for this article was provided by GM70065 from the NIH NIGMS.

Conflict of interest statement. None declared.

REFERENCES

1. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.
2. Karp,P.D., Ouzounis,C.A., Moore-Kochlacs,C., Goldovsky,L., Kaipa,P., Ahren,D., Tsoka,S., Darzentas,N., Kunin,V. and Lopez-Bigas,N. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res*., **33**, 6083–6089.
3. Karp,P.D., Paley,S. and Romero,P. (2002) The pathway tools software. *Bioinformatics*, **18**, S225–S232.
4. Bowers,P., Pellegrini,M., Thompson,M., Fierro,J., Yeates,T. and Eisenberg,D. (2004) Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol*., **5**, R35.
5. Papin,J.A., Stelling,J., Price,N.D., Klamt,S., Schuster,S. and Palsson,B.O. (2004) Comparison of network-based pathway analysis methods. *Trends Biotechnol*., **22**, 400–405.
6. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
7. Tsoka,S. and Ouzounis,C.A. (2000) Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion. *Nature Genet*., **26**, 141–142.
8. Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci*., **23**, 324–328.
9. Gabaldon,T. and Huynen,M.A. (2004) Prediction of protein function and pathways in the genome era. *Cell Mol. Life Sci*., **61**, 930–944.
10. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
11. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1998) Use of contiguity on the chromosome to predict functional coupling. *In Silico Biol*., **1**, 93–108.
12. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
13. von Mering,C., Jensen,L.J., Snel,B., Hooper,S.D., Krupp,M., Foglierini,M., Jouffre,N., Huynen,M.A. and Bork,P. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res*., **33**, D433–D437.
14. von Mering,C., Zdobnov,E.M., Tsoka,S., Ciccarelli,F.D., Pereira-Leal,J.B., Ouzounis,C.A. and Bork,P. (2003) Genome evolution reveals biochemical networks and functional modules. *Proc. Natl Acad. Sci. USA*, **100**, 15428–15433.
15. Wu,J., Kasif,S. and DeLisi,C. (2003) Identification of functional links between genes using phylogenetic profiles. *Bioinformatics*, **19**, 1524–1530.
16. Yanai,I. and DeLisi,C. (2002) The society of genes: networks of functional links between genes from comparative genomics. *Genome Biol*., **3**, Research 0064.
17. Yanai,I., Mellor,J.C. and DeLisi,C. (2002) Identifying functional links between genes using conserved chromosomal proximity. *Trends Genet*., **18**, 176–179.
18. Yanai,I., Wolf,Y.I. and Koonin,E.V. (2002) Evolution of gene fusions: horizontal transfer versus independent events. *Genome Biol*., **3**, Research 0024.
19. Gaasterland,T. and Ragan,M.A. (1998) Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb. Comp. Genomics*, **3**, 199–217.
20. Pellegrini,M., Thompson,M., Fierro,J. and Bowers,P. (2001) Computational method to assign microbial genes to pathways. *J. Cell Biochem.,* (Suppl. 37), 106–109.
21. Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.
22. Marcotte,E.M., Pellegrini,M., Thompson,M.J., Yeates,T.O. and Eisenberg,D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.
23. Green,M.L. and Karp,P.D. (2004) A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, **5**, 76.
24. Karp,P.D. (2004) Call for an enzyme genomics initiative. *Genome Biol*., **5**, 401.
25. Roberts,R.J. (2004) Identifying protein function—a call for community action. *PLoS Biol*., **2**, E42.
26. 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.
27. Dahlquist,K.D., Salomonis,N., Vranizan,K., Lawlor,S.C. and Conklin,B.R. (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genet*., **31**, 19–20.
28. Lee,T.J., Pouliot,Y., Wagner,V., Gupta,P., Stringer-Calvert,D.W., Tenenbaum,J.D. and Karp,P.D. (2006) BioWarehouse: a bioinformatics database warehouse toolkit. *BMC Bioinformatics*, **7**, 170.
29. Luciano,J.S. (2005) PAX of mind for pathway researchers. *Drug Discov. Today*, **10**, 937–942.
30. Stromback,L. and Lambrix,P. (2005) Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. *Bioinformatics*, **21**, 4401–4407.
31. Finney,A. and Hucka,M. (2003) Systems biology markup language: Level 2 and beyond. *Biochem. Soc. Trans*., **31**, 1472–1473.