Use of a global metabolic network to curate organismal metabolic networks

A. R. Pah[1,2], R. Guimerà[4,5,1], A. M. Mustoe[1], L. A. N. Amaral[1,3,6]
1 Department of Chemical and Biological Engineering,
2 Interdepartmental Biological Sciences (IBIS), and
3 Northwestern Institute on Complex Systems, (NICO)
Northwestern University, Evanston, IL 60208, USA
4 Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, ES
5 Departament d'Enginyeria Química, Universitat Rovira i Virgili, Tarragona 43007, ES
6 Howard Hughes Medical Institute, Northwestern University, Evanston, IL 60208, USA
∗ **E-mail: amaral@northwestern.edu**

Abstract

Correctly annotating biological data is a significant problem in all cellular networks. While the amount of available genomic, proteomic, and metabolomic data has increased dramatically, we have struggled to keep pace in curating high-fidelity cellular organismal networks. Here we tackle the challenge of curating metabolic network reconstructions. We predict organismal metabolic networks using a global metabolic network as a reference along with sequence homology. While sequence homology has been a *de facto* standard to annotate metabolic networks in the past it has been widely faulted for its lack of predictive power. Here, however, we show that when homology is used in conjunction with the global metabolic network as a reference scaffold, then one is able to predict organismal metabolic networks that have enhanced network connectivity. We also compare the annotation behavior of current database curation efforts with our predictions and find that changes in the database are biased towards adding reactions to organismal metabolic networks.

Synopsis

Understanding the evolution of cellular processes is a significant task, not only will it give us greater insights into the origin of life but it should also increase our ability to selectively target diseased states and disorders such as cancer. However, to study a topic so broad we must approach the problem in a systematic manner. Key to this is having correct information about what reactions and metabolic processes an organism is capable of performing. While there has been extensive work in this field to predict metabolic networks, ranging from simple techniques to more exotic and complicated ones, we believe that this problem can benefit greatly by focusing on what metabolic network should be analyzed. We predict organismal metabolic networks by examining a global network of the reactions that any organism has. Using this reference network we are able to predict metabolic networks that are more connected, and thus better able to take in nutrients and convert them to growth, than the current, publicly available metabolic networks.

Introduction

Advances in high-throughput experimental biology have greatly advanced our knowledge and made it possible to interrogate cellular processes in a systematic manner. However, this data deluge is only as useful as our ability to interpret it [1]. The current lack of data reliability hampers our efforts to understand which network topologies fulfill physical, chemical and biological constraints [2–6], which is essential knowledge if we hope to obtain a better grasp of the function and evolution of cellular networks.

Improving organismal metabolic networks has been an area of intense interest, especially owing to their functionality in assessing organismal fitness *in silico*. There have been numerous methods proposed that attempt to solve the incompleteness of metabolic network reconstructions, ranging from gap-filling the organismal network based on what other known organismal networks possess [7–9] to methods that rely on multiple sources of annotation information to provide an assessment of enzyme presence [10].

However, even with these methods we are still far from achieving consensus on the correct metabolic network for a given organism, even one as well-studied as *Escherichia coli* (Fig. 1a). Indeed, there are dramatic differences in both the size and degree of overlap of the metabolic network for *E. coli* recorded in (i) different databases and (ii) different time snapshots over time (Fig. 1b). The magnitude of the changes in annotation is further exacerbated for organisms that have their genomes sequenced (Fig. 1c). This last example is a perfect demonstration of both our lack of knowledge and the problem of developing computational analyses that perfectly recapitulate the known network at a given time.

Data reliability is a real and pressing problem for experimental and computational researchers alike. Lately, there has been a push in research to consider the analysis of metabolic networks from the perspec-

tive of a global network. This framework has been applied to good measure in phylometabolism to assess the emergence biological carbon-fixation [11] and to understand the regulation of metabolism [12]. A global network has also been recently used in conjunction with probabilistic methods to predict metabolic networks on a small scale with experimental verification [13]. While the motivation for the global network approach has been mostly pragmatic, it reminds us of the "*Res Potentia*" framework proposed by Whitehead [14]. Wherein he proposes that which does exist, termed the *Res Extenta*, springs forth as a set, specific realization from the realm of possibilities in the *Res Potentia*.

Furthermore, We contend that using a global network approach to the study of metabolism is comparable to what epidemiologists do when studying worldwide propagation of infection. In building the worldwide air transportation network [15] all carrier flights are aggregated into a single network and analyzed. This is an important feature of its construction because it aids in the identification of and distinction between international and regional hubs. As an example, if we were to only consider US Airways (a North American carrier) alone we would not have fully grasped the importance of London, since this carrier will have more flights that go to Los Angeles or even San Diego instead of London. Even if we were to pick a group of carriers based on similarity (such as operating primarily in North America) and assess the ensemble of their individual networks, it would be difficult to assess the relative importance of the individual hub airports. Reframing the analysis of metabolic networks to a global network is an appropriate method to both assess our current network annotations and to gain an understanding of what evolutionary dynamics shape organismal metabolism into the structures that we currently know.

In the following we predict entire organismal metabolic networks using only the global network as a reference and sequence homology. We show that using an appropriate reference set allows for more insight to be obtained with sequence homology and the assessment of how curation behavior in a metabolic database affects known organismal metabolic networks.

Methods

Data Acquisition

We downloaded multiple instances of the Kyoto Encyclopedia of Genes and Genomes (KEGG) LIGAND database [16–18]; the first instance on June 24, 2009 and the last on February 22, 2011. We also downloaded enzyme protein sequences from KEGG on five occasions, all between July 2010 and February 2011. All possible, unique sequences for each enzyme were used, based on the associations to reactions from KEGG. We downloaded bioreaction databases for Ma 2003 and Zeng 2011 [19, 20] from http://www.tu-harburg.de and the iAF1260 *Escherichia coli* reconstruction [21] from the BiGG database [22].

We considered 998 organisms listed in the KEGG database to construct the global network. We constructed protein databases and predict metabolic networks for 874 of these 998 organisms. We did not predict the networks for 125 organisms due to a lack of sequence availability or because the time necessary to run a complete analysis for larger organisms was prohibitively long.

The bacterial domain dominates in representation due to the breakdown of organisms in KEGG itself. However, the network is not influenced by this over-representation because each reaction is only counted once in the construction of the global network. We include the domain and clade breakdown of the organisms that we tested and predicted metabolic networks for in Table 1.

Organismal and Global Network Construction

We constructed individual metabolic networks for 998 organisms using a 2009 snapshot of the KEGG database. In these networks, each node represents a metabolite, and two metabolites i and j are connected by an edge if there is a chemical reaction in which i is a substrate and j is its product, or vice versa. We established these relationships using the main reaction pair designations on KEGG and, as in prior

studies [23, 24], excluded transfer ions, co-factors, and energy carrier molecules so as to maintain the focus on the biomass transfer through the networks (Fig. 2a).

We constructed a global network by performing the graph union of all organismal networks (Fig. 2b). The 3,467 distinct reactions listed for the 998 organisms in the KEGG database yielded a global metabolic network comprising 6,656 metabolites and 3,328 unique edges. These metabolites are organized into a giant component comprising 2,023 metabolites and 2,729 edges, and 333 smaller components typically comprising only a few metabolites each. We focused our analyses on the giant component of the global network because it contains the most reliable data, as its metabolites are more conserved and have more pathway annotations.

Metabolic networks for *E. coli* based on other databases were constructed in the same manner as the organismal metabolic networks constructed using the KEGG database. For the Ma 2003 and Zeng 2011 datasets the main pairs designation was included in the original dataset and it is used in lieu of the KEGG main pairs designation, while we used the main pairs designation from KEGG for the iAF1260 reconstruction.

Organismal Network Prediction

To predict individual organismal metabolic networks we assumed that a given reaction can be catalyzed within an organism if, and only if, the organism synthesizes a protein that is sufficiently similar to the known enzymes for the reaction. We collected 5.94×10^6 known enzyme amino acid sequences from the KEGG database that are associated with the 3,467 reactions in the global network and prepared databases of all known proteins for 874 organisms from the `nr` database (downloaded February 23, 2011) in accordance with the BLAST user manual [25] in order to test sequence homology. We then evaluated each reaction in the global network for its possibility of existence in any individual organism.

We used `blastp` [26, 27], version 2.2.24, to align the enzyme sequences associated with each reaction to each organism's protein database (Fig. 3a) and determined the expectation value (E-value) of the alignment. The E-value is a measure of the number of times the match between the sequences would be expected to occur by chance; E-value = 0.0 indicates a perfect match between the queried enzyme sequence and a protein in the database, while E-value > 1.0 is interpreted as a sequence match that is not indicative of biological homology. We obtain a total of 2.6×10^{10} BLAST alignments subsequently used in our analysis.

For clarity, we define several additional terms. A reaction predicted to be catalyzed in a certain organism by a certain database curation team is "annotated" in that database. Otherwise, the reaction is "unannotated" — it exists in the global network but not in the organismal one. To make our predictions of annotation status we separated the alignments associated with a reaction r into two categories, hits and poor matches, based on the magnitude of the E-values obtained. If an alignment has E-value ≤ 0.01 then we classify the alignment as a hit; otherwise, if $0.01 <$ E-value ≤ 10, we classify it as a poor match (Fig. 3b).

We find that the distribution of the fraction of hits has a peak at greater values for KEGG annotated reactions when all of the reactions are considered. Furthermore, we see that this behaviour holds no matter which domain of organisms is considered (Fig. 3c), indicating this is a robust behavior that is preserved across all organismal networks. Unsurprisingly there is not a perfect separation between the annotated and unannotated distributions, given that not all reactions are correct in the database annotations. We know for certain that some unannotated reactions should be annotated and vice-versa, given the changes to the database over time (Fig. 1).

We use the fraction of alignments f_r^i that are classified as hits as a predictor of whether reaction r can be catalyzed within organism i. To determine whether a given f_r^i is large enough to be considered a reaction that can be catalyzed we must set a threshold value $f_\times$; if $f_r^i \geq f_\times$, then we predict reaction r to exist in organism i. To identify an appropriate value for $f_\times$, we calculated the receiver operator characteristic (ROC) curve, accuracy, and false discovery rate statistics [28]. The ROC curve analysis

demonstrates that f_r^i can discriminate between annotated and unannotated reactions (Fig. 4a). We thus use the accuracy and false discovery rates to determine a good threshold value for f_r^i and set $f_\times = 0.14$ (Figs. 4b). In summary, we predict organismal metabolic networks by checking whether a reaction r for organism i has a value $f_r^i > 0.14$ (Fig. 4c). This approach allows us to predict entire organismal networks using only the global metabolic network and the associated organismal BLAST alignments.

Validation

In validating our approach we face the problem of data reliability in KEGG and other databases, as detailed previously in the introduction and Figure 1. The disparity in organismal annotations is a substantial problem with the number and breadth of organisms for which we predict metabolic networks. Thus we validate our method in two separate manners: comparing our predicted network for *E.coli* against those from several databases and evaluating the connectedness of the organismal metabolic networks.

Consensus Network Construction

In order to calculate the accuracy of our predictions reliably it is necessary to compare it to a known answer. However, we lack such a known answer since the organismal metabolic networks are in a state of flux with significant differences in their content across different databases and time points (Fig. 3). In an effort to estimate the true accuracy of our predictions we consider the metabolic network reconstructions of *E.coli*, a well-studied organism, from three different sources, with two of the sources having network data at two separate time points.

We thus create a consensus network using a majority rule, similar to other work [29]. A set of the networks is selected and every edge in all of the networks is evaluated. If the edge appears in the majority of the networks in the set then it is added to the consensus network, otherwise it is not added. We calculate the statistics associated with the ROC analysis for the metabolic networks against the consensus network that were not used in its construction.

Network Connectivity

Given the challenge traditional validation and our aim to predict an organism's true metabolic network instead of simply recapitulating the annotations in KEGG, we also use a validation scheme focused on the expected properties of metabolic networks. Specifically, we surmise that organismal metabolic networks must have a bias toward connectedness. Indeed FBA metabolic reconstructions assume that metabolic networks act as "transportation" networks that carry mass from external nutrients to biomass [7,8]. The possession of fewer network components implies a greater ability of the organism to exploit a broad range of incoming nutrients for disparate cellular roles, and thus offers a fitness advantage over topologies where each network component must be individually fed.

To assess network connectivity we examine two quantities, the probability of a reaction addition closing a gap between two network components and a reaction removal creating an additional network component. For the random filling of gaps, we use the intersection of additional reactions between our predicted network and the KEGG 2011 network for an organism as the number of reactions that should be added. We then compare the observed number of gaps versus the random chance expectation of completing a gap of a given size with the available number of additional reactions. For the creation of additional network components we removed every edge individually in all organismal networks and determine if an additional network component is created. We then average the fractional number of additional components added across all edges tested.

Results

Comparison with Consensus Networks

We constructed ten separate consensus networks as detailed in Methods and evaluated the accuracy of the networks that were left out of the construction, with the results shown in Table 2. While our predicted network is not 100% accurate with respect to the consensus network, we find that our predictions range in accuracy between 70 and 71% while the database networks range in accuracy between 66 and 93%.

In an effort to understand what types of reactions our method incorrectly predicts we characterize the smallest set of false positives. First, we examine the pathway annotations associated with the metabolites in this set of false positive edges and reactions (Fig. 5b). We find that the majority of the metabolites are either unclassified or classified in pathways that are not central to metabolism (that is, they do not belong to carbohydrate, amino acid, nucleotide, and lipid metabolism pathways). However, even being associated with a central pathway does not mean that all of the metabolites are specifically involved in central or essential processes and these characterizations could be due only to their presence as a byproduct in a reaction.

Second, we examine the distribution of the false positive reactions according to conservation (Fig. 5a). Conservation is calculated as the fraction of times that the edge appears in an organismal network in comparison to the total number of organismal networks. We calculate the conservation for all edges in all organismal networks in KEGG and define three bins in the distribution (lower, middle, and upper thirds of the distribution). When we bin the false positive edges into these bins we find that the overwhelming majority are in the lowest third of conservation values. If we consider the lower and middle thirds together these groups accounts for more than 90% of all edges in the false positive set.

The abundance of low conservation reactions in the "false positive" set of our method could plausibly be interpreted as suggesting that these reactions may not actually be false positives. It is likely that a majority of the edges in this set do actually exist, they just have not been incorporated into a majority of the databases due to poor characterization and understanding of the reactions themselves.

Network Connectivity

We find that both our predictions and the changes made in KEGG in the period 2009-2011 close more gaps between network components than would be expected if new reactions were added at random (Fig. 6a). When we consider gaps of size one, our predictions fill almost twice as many gaps as the KEGG changes. Remarkably, we also find that our predictions introduce fewer new components than random removals. The changes in KEGG actually cause the creation of more additional network components than would be expected if reactions were randomly removed from organisms (Fig. 6b). The fact that so many gaps are closed by both our method and by KEGG curation in the period 2009-2011 lends credence to our original hypothesis that metabolic networks should be evolutionarily biased towards minimizing the number of network components.

It is important to note that our method takes in no information from the global network concerning reactions other than the possibility of their existence. This method is no more biased towards closing gaps or preserving network structure than the actual changes could or should be.

Biases in Database Curation

When we examine how our predictions compare to the corrections made to the KEGG database over time we find that there is a distinct bias towards adding instead of removing reactions to organismal networks (Fig. 7). It would be simple to assume that our method under-predicts in comparison to the reference dataset; however, we do not observe this trend when we examine the set of well-studied organisms used in Fig. 1b and c. This suggests that the curation teams are more aggressive in adding

reactions than removing them, despite the fact that both errors (spurious or lacking information) are equally detrimental. Large-scale comparison and tracking of database changes could influence curation teams' actions and help attenuate this problem.

Discussion

There are several distinct advantages to reframing the study of metabolic networks and, more broadly, metabolism to the organismal usages of the global network. As demonstrated in this study we are able to extract substantially more predictive power from sequence homology when it is used in conjunction with the global network. While most studies have moved beyond homology due to a lack of predictive power to more complicated and time consuming methods (such as Bayesian or multiple information methods), we are able to predict metabolic networks that compare favorably to the known database data and exceed them in producing connected networks. We are also able to easily increase the efficacy of our method by including additional network information such as whether a reaction completes a gap or not, which would be trivial to calculate and consider.

The global network also enables community detection and other graphical analyses that are unchanging in the face of organismal usage, facilitating an understanding of the true importance of a metabolite. Comparing the differences in organismal usage of metabolites and reactions can then be used to more robustly characterize the evolutionary forces that have optimized an organismal network in this manner. Specifically, when studying an organismal network we cannot fully comprehend the importance of a given metabolite because we do not have access to all the manners in which that metabolite could potentially connect to other metabolites in the network. Thus, we cannot accurately determine, for example, the centrality of the metabolite within metabolism or ascertain its true importance from an evolutionary standpoint. In contrast, the global network makes apparent these possibilities because it includes all available organismal knowledge. An increased understanding of why an organism develops certain "solutions" for its metabolic needs will aid in predicting unique features of the organism's metabolite and reaction usage that can be specifically targeted by drugs or other therapeutics and metabolic engineering.

More importantly, the global metabolism also allows us to view the metabolite and reaction usage of organisms in a general framework providing a means to identify metabolic "devices", small groups of metabolites and reactions that have a functional purpose, and other features that become apparent only when considering intermediate scales within the network [30–32]. This enables us to give greater insight into both metabolic evolution as well as ways to design synthetic metabolic "circuits" from these devices [33, 34].

Acknowledgments

We thank I. Sirer, P.D. McMullen, E.N. Sawardecker, S.M.D. Seaver, and P.B. Winter for comments and suggestions. A.R.P. acknowledges the support of a Northwestern Predoctoral Biotechnology Training Grant and from the Chicago Biomedical Consortium with support from The Searle Funds at the Chicago Community Trust. R.G. acknowledges the support of the James S. McDonnell Foundation, of the Spanish Ministerio de Ciencia e Innovación grant FIS2010-18639, and of the European Union Grant PIRG-GA-2010-277166. L.A.N.A. acknowledges the support of NSF award SBE 0624318 Foundation and the W.M. Keck Foundation.

References

1. Pennisi E (2005) How will big pictures emerge from a sea of biological data? Science 309: 94.

2. Vitkup D, Kharchenko P, Wagner A (2006) Influence of metabolic network structure and function on enzyme evolution. Genome Biology 7: R39+.

3. Yus E, Maier T, Michalodimitrakis K, van Noort V, Yamada T, et al. (2009) Impact of genome reduction on bacterial metabolism and its regulation. Science 326: 1263–1268.

4. Sommer MO, Church GM, Dantas G (2010) A functional metagenomic approach for expanding the synthetic biology toolbox for biomass conversion. Molecular systems biology 6.

5. Raymond J, Segrè D (2006) The effect of oxygen on biochemical networks and the evolution of complex life. Science 311: 1764–1767.

6. Zhao J, Ding GHH, Tao L, Yu H, Yu ZHH, et al. (2007) Modular co-evolution of metabolic networks. BMC bioinformatics 8: 311+.

7. Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, et al. (2006) Systems approach to refining genome annotation. Proceedings of the National Academy of Sciences 103: 17480–17484.

8. Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, et al. (2010) High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotech 28: 977–982.

9. Christian N, May P, Kempa S, Handorf T, Ebenhoh O (2009) An integrative approach towards completing genome-scale metabolic networks. Mol BioSyst 5: 1889–1903.

10. Kharchenko P, Chen L, Freund Y, Vitkup D, Church GM (2006) Identifying metabolic enzymes with multiple types of association evidence. BMC bioinformatics 7: 177+.

11. Braakman R, Smith E (2012) The emergence and early evolution of biological Carbon-Fixation. PLoS Comput Biol 8: e1002455.

12. Maslov S, Krishna S, Pang TY, Sneppen K (2009) Toolbox model of evolution of prokaryotic metabolic networks and their regulation. Proceedings of the National Academy of Sciences 106: 9743–9748.

13. Plata GA, Fuhrer T, Hsiao TL, Sauer U, Vitkup D (2012) Global probabilistic annotation of metabolic networks enables enzyme discovery. Nat Chem Biol 8: 848–854.

14. Whitehead AN (1979) Process and Reality (Gifford Lectures Delivered in the University of Edinburgh During the Session 1927-28). Free Press, 2nd edition.

15. Guimerà R, Sales-Pardo M, Amaral L (2007) Classes of complex networks defined by role-to-role connectivity profiles. Nature Phys 3: 63-69.

16. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38: D355-D360.

17. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, et al. (2006) From genomics to chemical genomics: New developments in KEGG. Nucleic Acids Res 34: D354–D357.

18. Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28: 27–30.

19. Ma H, Zeng AP (2003) Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. Bioinformatics 19: 270-277.

20. Sun J, Kamphans T, Fekete SP, Zeng AP (2011) An extended bioreaction database that significantly improves reconstruction and analysis of genome-scale metabolic networks. Integr Biol 3: 1071–1086.

21. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, et al. (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 3: 121.

22. Schellenberger J, Park JO, Conrad TM, Palsson BØ (2010) BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions. BMC Bioinformatics 11: 213.

23. Guimerà R, Amaral L (2005) Cartography of complex networks: modules and universal roles. J Stat Mech: Theor Exp : art. no. P02001.

24. Guimerà R, Amaral L (2005) Functional cartography of complex metabolic networks. Nature 433: 895–900.

25. Camacho C, Madden T, Coulouris G, Ma N, Tao T, et al. (2008). Blast command line applications user manual. Internet.

26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.

27. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

28. Fawcett T (2006) An introduction to ROC analysis. Pattern Recognition Letters 27: 861-874.

29. Osterman A (2003) Missing genes in metabolic pathways: a comparative genomics approach. Current Opinion in Chemical Biology 7: 238–251.

30. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, et al. (2002) Network motifs: Simple building blocks of complex networks. Science 298: 824–827.

31. Alon U (2007) Network motifs: theory and experimental approaches. Nature Rev Gen 8: 450–461.

32. Spirin V, Gelfand MS, Mironov AA, Mirny LA (2006) A metabolic network in the evolutionary context: Multiscale structure and modularity. Proceedings of the National Academy of Sciences 103: 8774–8779.

33. Papin JA, Reed JL, Palsson BØ (2004) Hierarchical thinking in network biology: The unbiased modularization of biochemical networks. Trends Biochem Sci 29: 641–647.

34. Lazebnik Y (2002) Can a biologist fix a radio?—Or, what I learned while studying apoptosis. Cancer Cell 2: 179 - 182.

Figure Legends

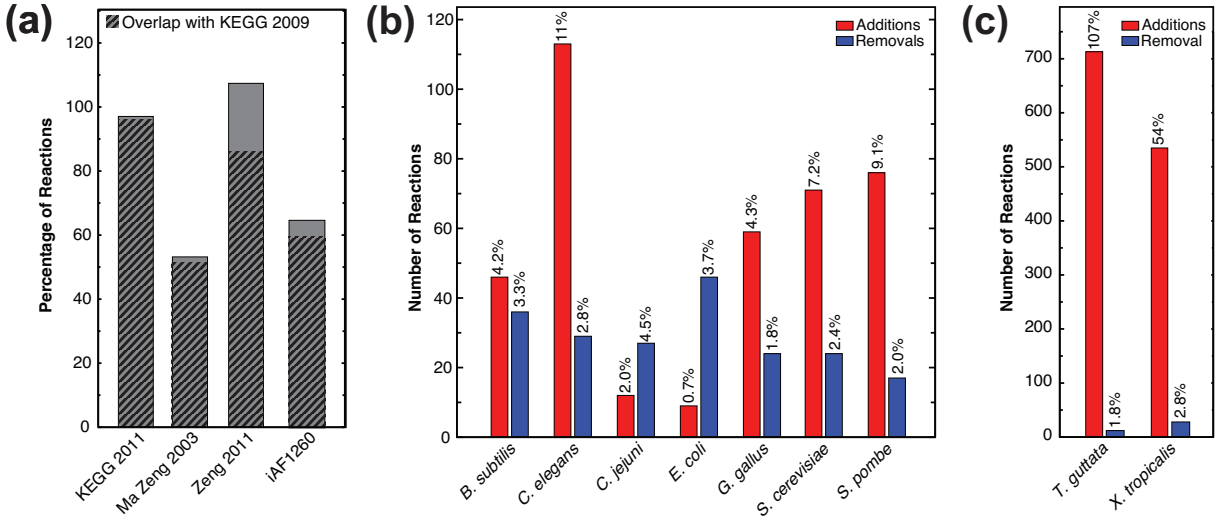

Figure 1. Metabolic networks are still being actively revised. We focus here on the metabolites and reactions belonging to the core metabolism and show network size and changes in metabolic networks for selected organisms. **a**, Relative size of the *E. coli* metabolic network for four databases. Note the dramatic differences in both the size of the networks and the percentage of overlap with the KEGG 2009 network. **b**, Number of reactions KEGG added or removed for selected organismal networks over the course of two years (2009 to 2011). Even though the networks are again restricted to the intersection with the core metabolism there is appreciable turnover. **c**, Number of reactions KEGG added or removed for selected organismal networks over the course of the same period for two recently sequenced organisms. To illustrate the effect of these revisions, if a model was to perfectly predict the 2009 organismal network for *Taeniopygia guttata* it would be less than 50% accurate when it was compared to the 2011 network.

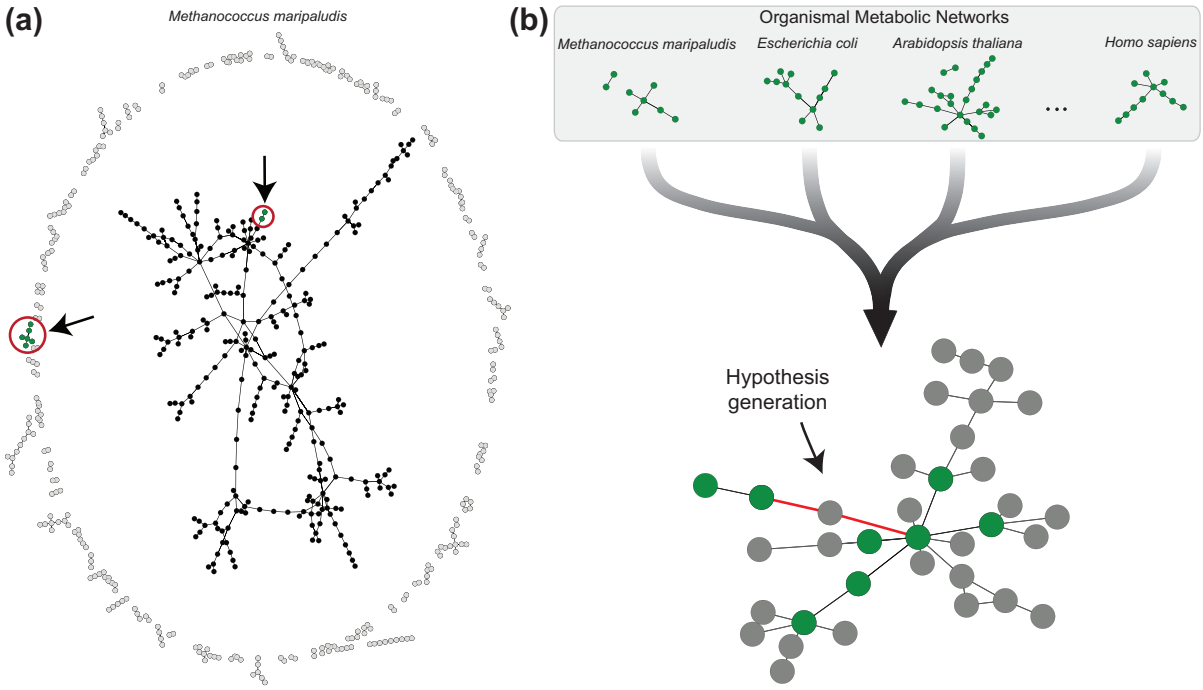

Figure 2. Construction of a global metabolic network from KEGG data. a, Construction of the *Methanococcus maripaludis* metabolic network from KEGG. We show the giant component in black and the 122 small components in grey. We highlight eight metabolites (shown in green, highlighted with red circles) that will be considered in the other panel. **b**, The set of all organismal networks is the extent of our current knowledge from KEGG. We construct the global metabolic network as the union of all known networks. We superimpose the eight highlighted metabolites from *M. maripaludis* (shown in green) on the global network making their network proximity apparent. This superimposition elucidates potential additional reactions that can bridge gaps between different components of the network. Such superimpositions can dramatically decrease the search space for new reactions that can accomplish this bridge.

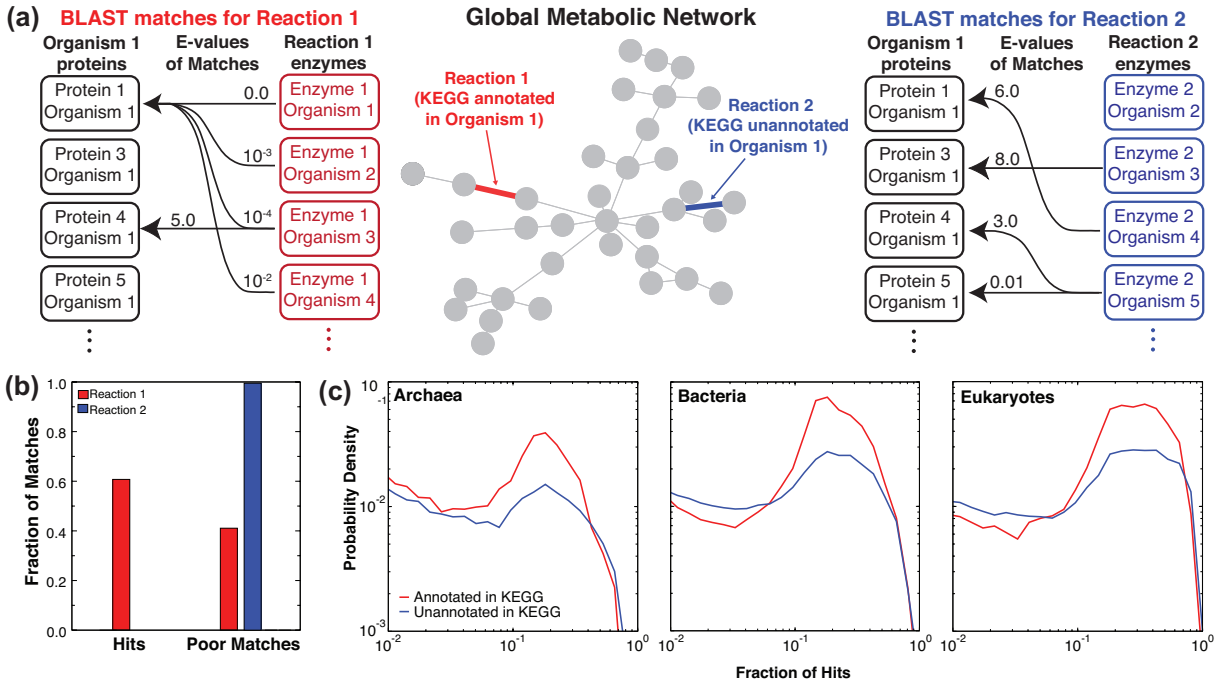

Figure 3. Organismal network prediction using the global metabolic network. **a**, We BLAST the enzyme sequences associated with Reaction 1 from the global metabolism in all organisms against the known protein sequences from open reading frames in Organism 1 in order to determine whether Organism 1 possesses an enzyme that can catalyze Reaction 1. We repeat this same procedure for Reaction 2 from the global metabolism. **b**, We categorize each alignment as a hit or poor match based on the magnitude of its E-value. If an alignment has an E-value ≤ 0.01 then we classify the alignment as a hit; otherwise, if $0.01 <$ E-value ≤ 10, we classify it as a poor match. It is visually apparent in this example that the annotated reaction has a much larger fraction of hits. **c**, The distributions of the fraction of hits for annotated and unannotated reactions show a clear difference no matter which domain of organisms is considered.

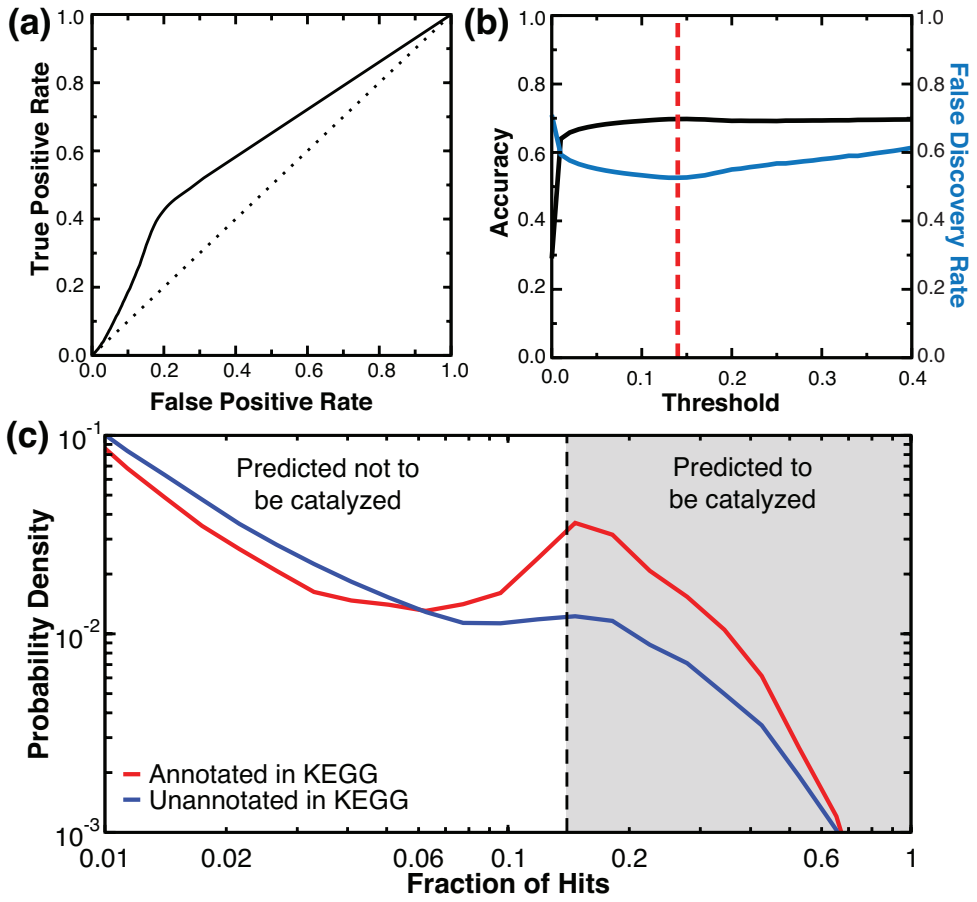

Figure 4. Proposing a good threshold to predict reactions that will be catalyzed. We use multiple statistics for judging the optimal threshold to differentiate between annotated and unannotated groups. The receiver operator characteristic curve demonstrates that all of the thresholds possess discriminatory power in comparison to random chance (**a**). To choose a precise threshold we combine this with the false discovery rate and accuracy statistics (**b**). Balancing the two statistics results in a threshold of 0.14 as the optimal point to separate the two groups. This value corresponds to the peak seen in the annotated distribution for all organisms (**c**). We use this threshold to predict whether an organism can catalyze a given reaction between two metabolites when we predict organismal metabolic networks *de novo* from the global metabolic network.

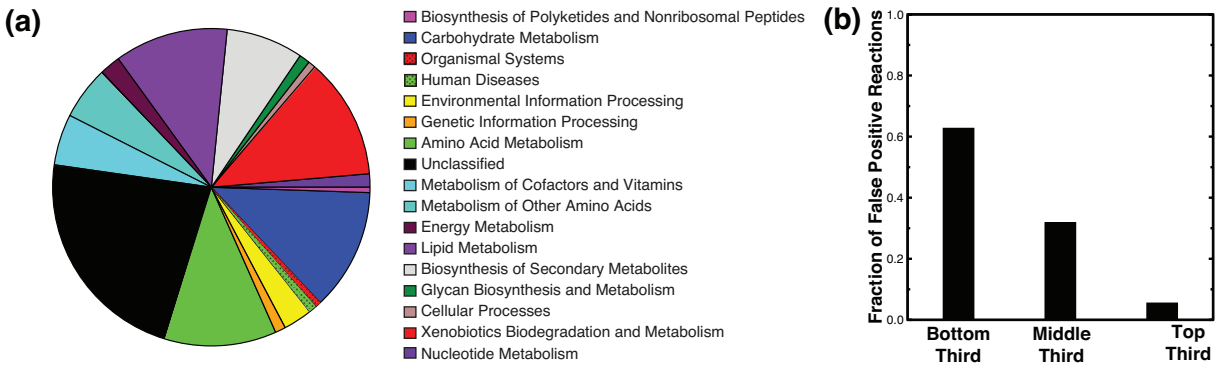

Figure 5. Examination of pathway annotations and conservation for false positive reactions identified in *E. coli*. We take the set of false positives from our predicted network (identified by comparison to known network reconstructions) and find that when we examine the pathway annotations the largest single group is "Unclassified" (**a**). Furthermore, when the reactions are binned by their conservation values we find that the bulk of the false positive assignments belong to the lowest third of conservation values (**b**). These two points suggest that the false positive set from our predictions is composed primarily of reactions that are relatively uncharacterized and poorly studied. This could mean that the actual amount of "true" false positives is smaller than currently suggested by comparison to other network reconstructions and has a substantial composition of present, but currently unknown reactions.

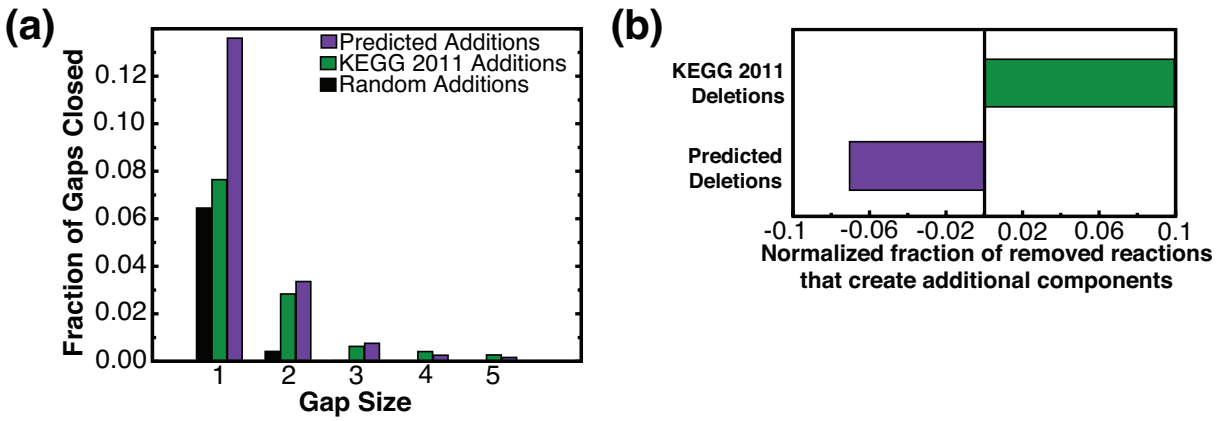

Figure 6. Predicted networks are more connected than ones constructed strictly from the database. a, Fraction of gaps between two network components that are closed when reactions are added to a metabolic network versus gap size. Both the predicted network additions and the KEGG 2011 additions close more gaps in the KEGG 2009 network than would be expected by random chance. However, there are significantly more additions using the predicted networks when a gap size of 1 is considered. **b**, Relative fraction of removed reactions that introduce an additional network component in comparison to random removals for predicted deletions and KEGG 2011 deletions. Reactions "removed" by the predictions introduce additional network components less frequently than random removals would. The opposite is true for the reactions removed in the KEGG 2011 network where the removals actually introduce additional components more frequently than random removals. The predictions significantly differ from the KEGG curation efforts in this behavior (predicted deletions $z_{score} = -0.97$, KEGG 2011 deletions $z_{score} = 1.23$).

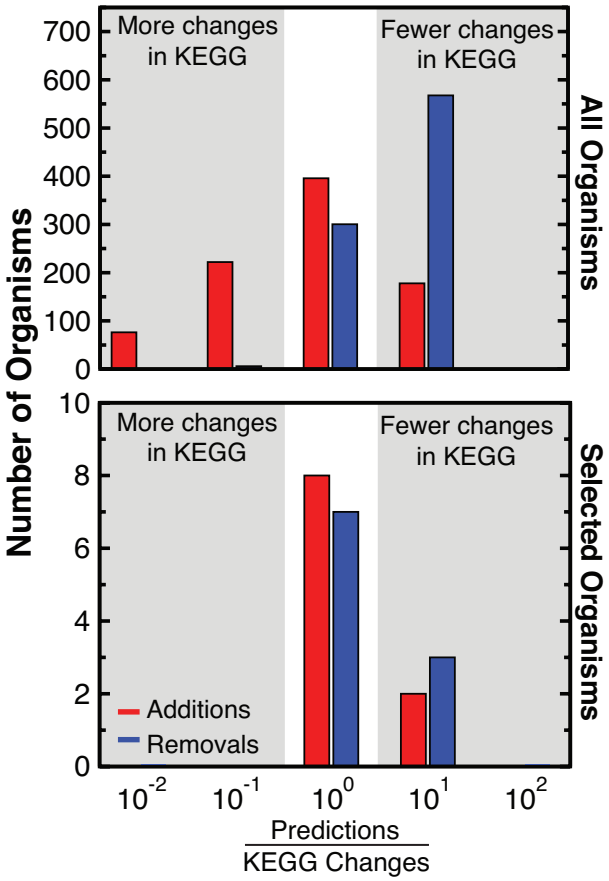

Figure 7. Comparison of changes in the KEGG database over two years against the "changes" predicted. When all organisms are considered there is a distinct bias for more reactions to be added to organisms in comparison to the number of reactions predicted to be added. When we consider only the set of 10 well studied organisms considered in Fig. 1 there is largely agreement between our predictions and the actual changes that occur in KEGG for both additions and removals.

Tables

Domain	Clade	Number of Organisms
Archaea (54)	Crenarchaeota	19
	Euryarchaeota	34
	Nanoarchaeota	1
Bacteria (750)	Acidobacteria	2
	Actinobacteria	59
	Alpha Proteobacteria	96
	Bacillales	59
	Bacteroides	12
	Beta Proteobacteria	60
	Chlamydia	12
	Clostridia	35
	Cyanobacteria	36
	Deinococcus Thermus	5
	Delta Proteobacteria	21
	Epsilon Proteobacteria	23
	Fusobacteria	1
	Gamma Proteobacteria	196
	Green Nonsulfur Bacteria	9
	Green Sulfur Bacteria	9
	Hyperthermophilic Bacteria	11
	Lactobacillales	61
	Magnetococcus	1
	Mollicutes	21
	Planctomyces	1
	Spirochete	17
	Termite Group	1
	Verrucomicrobia	2
Eukaryotes (70)	Animals	19
	Fungi	27
	Plants	3
	Protists	21

Table 1. Number of organisms considered by taxonomic clade

Consensus Networks	KEGG 2009 (K09)	KEGG 2011) (K11)	Ma Zeng 2003 (M03)	Zeng 2011 (Z11)	iAF1260 (iAF)	Global Network $f_\times = 0.14$ (GN)
K09, Z11, iAF		0.88	0.81			0.70
K09, M03, iAF		0.85		0.69		0.71
K09, Z11, GN		0.92	0.81		0.78	
K09, M03, GN		0.87		0.68	0.79	
K11, Z11, iAF	0.90		0.81			0.71
K11, M03, iAF	0.86			0.68		0.71
K11, Z11, GN	0.93		0.80		0.78	
K11, M03, GN	0.87			0.67	0.79	
M03, iAF, GN	0.81	0.80		0.66		
Z11, iAF, GN	0.83	0.82	0.82			

Table 2. Database network accuracy in comparison to a set of consensus networks. We show the accuracy measurements for database networks against ten constructions of the consensus network. For each consensus network construction we only show the results for the databases that were not used to build the consensus network.