BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 13 2007, pages 1616–1622
doi:10.1093/bioinformatics/btm150

Systems biology

A network-based method for target selection in metabolic networks

R. Guimerà*, M. Sales-Pardo and L.A.N. Amaral

Northwestern Institute on Complex Systems (NICO) and Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL 60208, USA

Received on January 11, 2007; revised on April 5, 2007; accepted on April 13, 2007

Advance Access publication April 26, 2007

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: The lack of new antimicrobials, combined with increasing microbial resistance to old ones, poses a serious threat to public health. With hundreds of genomes sequenced, systems biology promises to help in solving this problem by uncovering new drug targets.

Results: Here, we propose an approach that is based on the mapping of the interactions between biochemical agents, such as proteins and metabolites, onto complex networks. We report that nodes and links in complex biochemical networks can be grouped into a small number of classes, based on their role in connecting different functional modules. Specifically, for metabolic networks, in which nodes represent metabolites and links represent enzymes, we demonstrate that some enzyme classes are more likely to be essential, some are more likely to be species-specific and some are likely to be both essential and specific. Our network-based enzyme classification scheme is thus a promising tool for the identification of drug targets.

Contact: rguimera@northwestern.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Several groups have stressed the threat that the lack of development of new antimicrobial drugs poses to public health (Bax *et al*., 2000; Norrby *et al*., 2005; Spellberg *et al*., 2004). With the ever-increasing amount of biological data available, the promise of systems biology is to help in alleviating this problem by uncovering new drug targets (Nikolsky *et al*., 2005). A promising approach in this regard is to map the interactions between biochemical agents—such as proteins and metabolites—onto complex networks (Amaral and Ottino, 2004; Newman, 2003), and then study the properties of these networks to gain insight into key biological processes.

In spite of some efforts to uncover new drug targets using this approach (Jeong *et al*., 2001; Klamt and Gilles, 2004; Palumbo *et al*., 2005; Rahman and Schomburg, 2006), one may argue that the results have, so far, been modest. A potentially important reason for this is that the structure of complex biochemical networks is typically characterized in terms of global properties (Jeong *et al*., 2000, 2001; Ma and Zeng, 2003; Tanaka, 2005; Wagner and Fell, 2001). In particular, a lot of attention has been paid to the degree distribution of the nodes in these networks, i.e. the distribution of number of protein interactions per protein in the proteome (Jeong *et al*., 2001) and the distribution of the number of other metabolites into which a certain metabolite can be transformed through metabolic reactions in the metabolome (Jeong *et al*., 2000; Ma and Zeng, 2003; Ravasz *et al*., 2002; Tanaka, 2005; Wagner and Fell, 2001). The caveat is that global quantities are appropriate only when one of two very strict conditions is fulfilled: (i) the network lacks a modular structure (Guimerà and Amaral, 2005b; Han *et al*., 2004; Hartwell *et al*., 1999; Holme *et al*., 2003; Ravasz *et al*., 2002), or (ii) the network has a modular structure but (a) all functional modules were formed according to the same mechanisms, (b) all functional modules have similar properties and (c) the interface between functional modules is statistically similar to the bulk of the modules, except for the density of links.

To our knowledge, no real biochemical network fulfills either of the two conditions above, which implies that global properties are unlikely to provide insight into the mechanisms responsible for the formation and evolution of these networks (Guimerà *et al*., 2007), or to facilitate the discovery promising therapeutic targets. Alternative approaches that take into consideration the modular structure (Danon *et al*., 2005; Girvan and Newman, 2002; Guimerà and Amaral, 2005a; Newman and Girvan, 2004) of real-world complex networks are, thus, necessary. One such approach is the *cartographic* approach (Guimerà and Amaral, 2005a, b), which enables one to group nodes into a small number of roles according to their pattern of intra- and inter-module connections.

Recently, we demonstrated that the role of a node conveys significant information about the importance of the node, and about the evolutionary pressures on it (Guimerà and Amaral, 2005a, b). Here, we report that in metabolic networks—in which nodes represent metabolites and links represent enzymes—some link types, i.e. some enzyme classes, are more likely to be essential, some are more likely to be species-specific and some are likely to be both essential and specific. Our network-based enzyme classification scheme is thus a promising tool for the identification of drug targets.

*To whom correspondence should be addressed.

Intriguingly, we also find that some crucial enzymes are not backed up by alternative pathways.

2 METHODS

2.1 Description of the datasets

To build the 18 metabolic networks (Table 1 and Figs 1 and 3) from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Goto *et al*., 1998), we use the data compiled in the LIGAND section (Kanehisa and Goto, 2000) of the database (as of March–April 2006). To remove carrier metabolites (such as water or ATP), we determine the *main reactant pairs* in each reaction (Kanehisa *et al*., 2006). Reactant pairs are pairs of metabolites that have atoms or atom groups in common on both sides of a chemical reaction formula. In contrast with *cofactor pairs* and *leave pairs*, main reactant pairs correspond to the most relevant biochemical transformations in a reaction. Main substrates and products are listed in the *reaction_main.lst* file. We connect all the main substrates to all the main products in a reaction, but not substrates to substrates or products to products.

For each organism, we only take into account reactions that are catalyzed by an enzyme that the organism is able to synthesize, and reactions that are explicitly classified as spontaneous. We obtain the enzymes necessary for each reaction from the *reaction* file, and the enzymes synthesized by each organism from the organism databases in KEGG.

For our analysis of *Escherichia coli* and *Helicobacter pylori* (Figs 2 and 4), we use the set of reactions compiled by B. Ø. Palsson's Systems Biology Research Group, at UCSD (Reed *et al*., 2003; Thiele *et al*., 2005). To obtain the modules and the roles of each metabolite in the flux balance analysis (FBA) reconstruction of these metabolic network, we need to remove carrier metabolites from the metabolic network. To this end, we automatically match each reaction in the FBA database to a reaction in the KEGG database, and use the LIGAND section of the KEGG (as of November 2005) to identify the main reactant pairs in the reaction. We find such a match for ~80% of the reactions listed in the FBA databases. For the remaining reactions, we remove carrier metabolites manually.

2.2 Module identification

For a given partition of the nodes of a network into modules, the modularity M of this partition is (Guimerà *et al*., 2004; Newman and Girvan, 2004)

$$M \equiv \sum_{s=1}^{N_M}\left[\frac{l_s}{L}-\left(\frac{d_s}{2L}\right)^2\right], \quad (1)$$

where N_M is the number of modules, L is the number of links in the network, l_s is the number of links between nodes in module s, and d_s is the sum of the connectivities (degrees) of the nodes in module s. The modularity of a partition is high if the number of within-module links is much larger than expected from chance alone.

The objective of a module identification algorithm is thus to find the partition with the largest modularity. We use simulated annealing to find the partition with the largest modularity (Guimerà and Amaral, 2005a, b; Guimerà *et al*., 2004). Danon *et al*. (2005) have recently shown that this method is the most accurate method in the literature to date.

2.3 Role definition

Nodes with similar roles are expected to have similar relative within-module connectivity (Guimerà and Amaral, 2005a, b). If κ_i is the number of links of node i to other nodes in its module s_i, $\overline{\kappa}_{s_i}$ is the average of κ over all the nodes in s_i and $\sigma_{\kappa_{s_i}}$ is the standard deviation of κ in s_i, then we define the within-module degree z-score as

$$z_i = \frac{\kappa_i - \overline{\kappa}_{s_i}}{\sigma_{\kappa_{s_i}}}. \quad (2)$$

Table 1.

Species	Nodes	Links	M	95% confidence interval M_{rand}
A.fulgidus	303	366	0.813	0.736–0.756
A.pernix	300	387	0.797	0.699–0.723
M.jannaschii	223	277	0.813	0.714–0.726
P.aerophilum	335	421	0.811	0.727–0.735
P.furiosus	302	384	0.813	0.706–0.734
S.solfataricus	367	455	0.813	0.724–0.748
B.subtilis	649	863	0.815	0.718–0.730
E.coli	739	1009	0.810	0.705–0.717
F.nucleatum	378	473	0.816	0.726–0.742
H.pylory	360	438	0.837	0.734–0.758
M.leprae	451	578	0.814	0.722–0.742
T.elongatus	448	546	0.830	0.743–0.767
A.thaliana	607	792	0.825	0.722–0.734
C.elegans	431	569	0.818	0.706–0.722
H.sapiens	792	1056	0.842	0.721–0.733
P.falciparum	280	363	0.815	0.696–0.720
S.cerevisiae	570	776	0.814	0.702–0.714
S.pombe	503	664	0.827	0.715–0.727

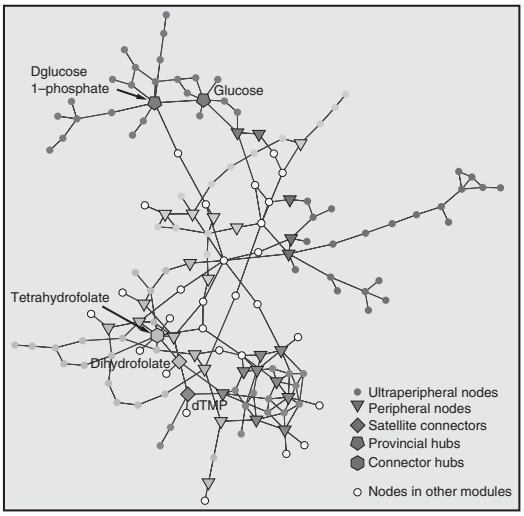

Fig. 1. Modularity and roles in the studied metabolic networks. In Table 1, we show the number of nodes and links in each network, the modularity M of the best partition obtained using simulated annealing, and the 95% confidence interval for the modularity M_{rand} of the randomizations of the network (Guimerà *et al*., 2004) (see Methods Section). All networks display a modularity that is significantly larger than that of their corresponding randomizations, which demonstrates that all the networks are truly modular. In the diagram, we show a portion of the FBA reconstruction of the metabolic network of *E.coli* (Reed *et al*., 2003). We depict nodes from five typical modules (colored) as well as the neighbors of these nodes that do not belong to any of the five modules (white). The different shapes of the colored nodes correspond to different roles.

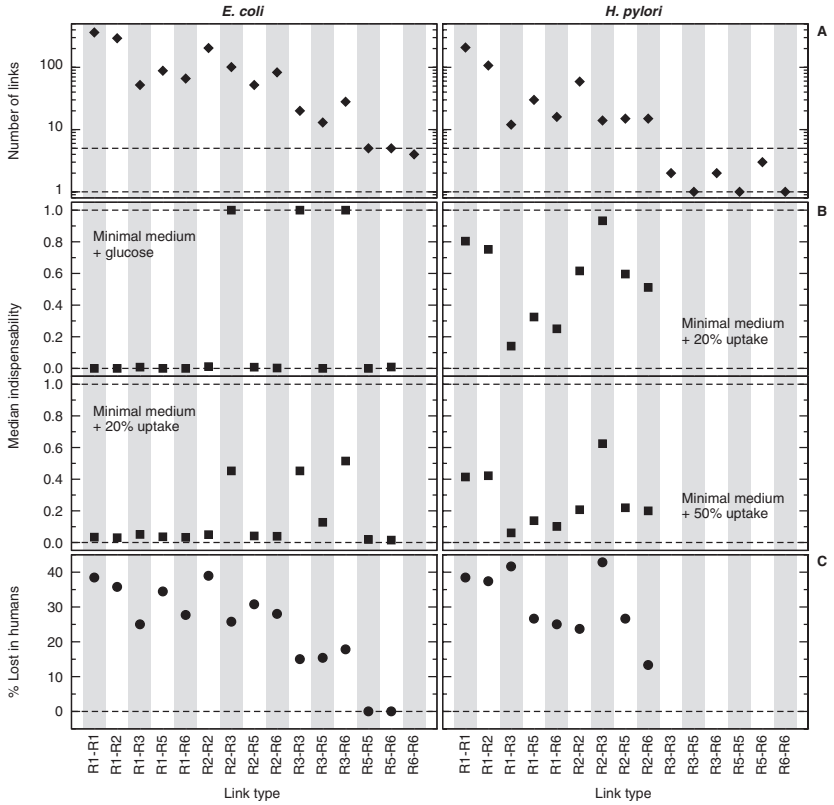

Fig. 2. Link type indispensability and conservation for *E.coli* (left) and *H.pylori* (right). (**A**) Distribution of link types. The higher the number of links of a certain type, the more reliable the estimation of the corresponding indispensability. Link types represented by fewer than five links (dashed line) are not considered significant and are therefore disregarded in the following analysis. (**B**) Median indispensability as a function of link type for different media (see Methods Section for the definition of the media; 'Minimum medium +20% uptake' indicates to a random medium with a fraction $f=0.20$ of all possible uptake fluxes added to the minimum uptake fluxes). Note that, since we consider 100 different partitions of each metabolic network, the same link can contribute to more than one link type. Regardless of the medium considered (see Methods Section), most link types in *E.coli* have a median close to zero, which means that about half of the links of those types can be removed without significantly altering the growth rate. In contrast, links of type R2–R3, R3–R3 and R3–R6, have a significantly higher median indispensability (I $\sim$1.0 for the glucose medium and I $\sim$0.45 for richer media with $f=0.20$). This means that, when removed, at least half of the links of these types have a markedly negative effect on the growth rate. Results for *H.pylori* are noisier due to the smaller total number of reactions and to the higher average essentiality. Remarkably, though, links of type R2–R3 continue to be the most essential in all media considered. (**C**) Percentage of links in *E.coli* and *H.pylori* that are *lost in humans*. We call a link lost in humans if none of the proteins catalyzing the link have a significantly similar match in humans [we consider two proteins significantly similar when the expectation value returned by BLAST (Altschul *et al.*, 1997) is smaller than 10^{-2}]. The lowest ratios occur for links involving global hub metabolites (R6) and, as hypothesized, links of type R2–R3 do not have a significantly lower than average loss rate.

The within-module degree z-score measures how 'well-connected' node i is to other nodes in the module.

Different roles can also arise because of the connections of a node to modules other than its own (Guimerà and Amaral, 2005a, b). We define the participation coefficient P_i of node i as

$$P_i = 1 - \sum_{s=1}^{N_M} \left(\frac{\kappa_{is}}{k_i} \right)^2 \qquad (3)$$

where κ_{is} is the number of links of node i to nodes in module s, and k_i is the total number of links of node i. The participation coefficient of a node is therefore close to one if its links are uniformly distributed among all the modules and zero if all its links are within its own module.

We classify as *non-hubs* those nodes that have low within-module degree ($z < 2.5$). Depending on the amount of connections they have on other modules, non-hubs are further subdivided into (Guimerà and Amaral, 2005a, b): (R1) *ultra-peripheral nodes*, i.e. nodes with all their links within their own module ($P \leq 0.05$); (R2) *peripheral nodes*, i.e. nodes with most links within their module ($0.05 < P \leq 0.62$); (R3) *satellite connectors*, i.e. nodes with a high fraction of their links to other modules ($0.62 < P \leq 0.80$) and (R4) *kinless nodes*, i.e. nodes with links homogeneously distributed among all modules ($P > 0.80$). We classify as *hubs* those nodes that have high within-module degree ($z \geq 2.5$). Similar to non-hubs, hubs are divided according to their participation coefficient into: (R5) *provincial hubs*, i.e. hubs with the vast majority of links within their module ($P \leq 0.30$); (R6) *connector hubs*, i.e. hubs with many links to most of the other modules ($0.30 < P \leq 0.75$) and (R7) *global hubs*, i.e. hubs with links homogeneously distributed among all modules ($P > 0.75$).

Note that, since we use a stochastic module identification algorithm, small differences do exist between different partitions of the network. Nevertheless, results are highly consistent (Guimerà and Amaral, 2005a, b), so that by obtaining 100 quasi-optimal partitions, we can estimate the likelihood that a link is of a certain type.

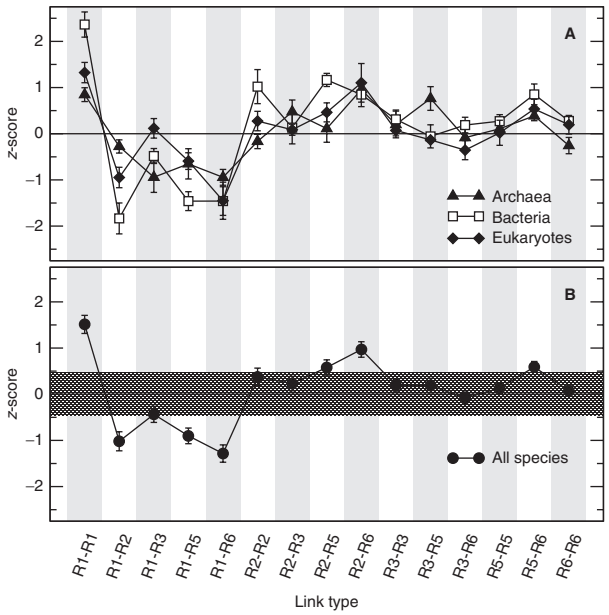

Fig. 3. Role-to-role connectivity profiles. For each network, we calculate the number r_{ij} of links between nodes belonging to roles i and j, and compare this number to the number R_{ij} of such links in a properly randomized network (see Methods Section). R_{ij} is normally distributed, so we use the z-score, $z_{ij} = (r_{ij} - \langle R_{ij}\rangle)/(\sqrt{\langle R_{ij}^2\rangle - \langle R_{ij}\rangle^2})$, to obtain a profile of over- and under-representation of link types. The brackets $\langle \ldots \rangle$ denote an average over 100 randomizations of the network (see Methods Section). (**A**) Average z-score for the abundance of each link type for archaea, bacteria and eukaryotes. (**B**) Average z-score for the abundance of each link type for all the species considered. The shaded region in panel (B) represents the 95% confidence interval for the random network expectation: points outside this region indicate statistically significant over- or under-represented link types.

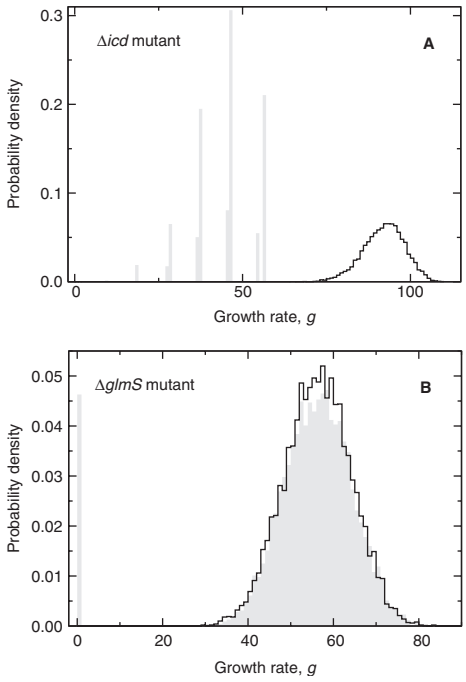

Fig. 4. *In silico* analysis of gene indispensability. We plot the distribution of growth rates for mutant (gray) and wild type (black line) *E.coli*, obtained with FBA from an ensemble of 10 000 random media (see Methods Section). (**A**) Mutant Δ*icd* in very rich random media ($f = 0.80$). The *icd* gene codes for the enzyme isocitrate dehydrogenase, which reversibly transforms isocitrate into 2-oxoglutarate (this link has a high probability of being of type R2–R3) and is listed as non-essential both in the PEC database and by Gerdes *et al.* (2003). Although FBA confirms that the mutant is always viable ($g > 0$), it also shows that its growth rate is significantly smaller than the growth rate of the wild type, which would render the mutant inviable in most selective environments. (**B**) Mutant Δ*glmS* in rich random media ($f = 0.50$). The *glmS* gene codes for the enzyme that mediates the transformation of L-glutamine into glucosamine-6P (which is of type R2–R3 with a high probability), and is reported to be non-essential in the PEC database and essential by Gerdes *et al.* (2003). FBA shows that the mutant has a growth rate very similar to the growth rate of the wild type in 95.4% of the media, but is inviable ($g = 0$) in 4.6% of the media.

2.4 Network randomization and statistical models

We use two different network randomization schemes. To assess the significance of the modular structure of a network (Fig. 1), we randomize all the links in the network while preserving the degree of each node. To uniformly sample all possible networks, we use the Markov-chain Monte Carlo switching algorithm (Maslov and Sneppen, 2002; Itzkovitz *et al.*, 2004). In this algorithm, one repeatedly selects pairs of links and swaps one of the ends of the links.

For the analysis of the over- and under-representation of links between pairs of roles, it is crucial to preserve not only the degree of each node, but also the modular structure of the network and the role of each node. Therefore, we restrict the Markov-chain Monte Carlo switching algorithm to pairs of links that connect nodes in the same pair of modules. In other words, we apply the Markov-chain Monte Carlo switching algorithm independently to links whose ends are in modules 1 and 1, 1 and 2 and so forth for all pairs of modules. This method guarantees that, with the same partition as the original network, the modularity of the randomized network is the same as that of the original network (since the number of links between each pair of modules is unchanged) and that the role of each node is also preserved.

2.5 Flux balance analysis and link indispensability

Flux balance analysis (FBA) (Edwards and Palsson, 2000a) enables us to computationally estimate metabolic fluxes in an organism. FBA is based on the idea that metabolic fluxes producing and consuming a certain metabolite must, in the steady state, balance each other. Additionally, some amounts of certain metabolites can be obtained from or excreted to the extracellular medium. These considerations impose a set of linear constraints on the metabolic fluxes. One can further assume that, among all possible flux solutions satisfying all the constraints, the one realized in nature is the one that maximizes the growth rate of the organism. With these assumptions, the problem of finding the flux distribution becomes a linear programming problem, which can be solved using standard numerical techniques (Winston and Venkataramanan, 2002).

For our analysis, we use the set of reactions compiled for *E.coli* (Reed *et al.*, 2003) and *H.pylori* (Thiele *et al.*, 2005). In our simulations

of gene deletions, we use ensembles of random media that are built as follows. A minimal medium is defined for each species (Reed *et al.*, 2003; Thiele *et al.*, 2005) (the *E.coli* minimal medium contains oxygen to simulate aerobic conditions). To the minimal medium, we add a randomly selected fraction f of all possible uptake fluxes, which are set to a maximum of 10 mmol/g h. Each random medium is evaluated 200 times, each time with a different selection of the uptake fluxes. Additionally, for *E.coli* we simulate a glucose medium, which consists of the minimal medium plus a glucose uptake rate of up to 10 mmol/g h.

Following the work of Edwards and Palsson, (2000b), we define the indispensability of a link as the ratio between the growth rate g_Δ of the mutant and the growth rate g of the wild type organism. In particular, the indispensability I is defined as

$$I = 1 - \frac{g_\Delta}{g}. \quad (4)$$

3 RESULTS

3.1 Functional modules and metabolite roles in metabolic networks

In a metabolic network, each node represents a metabolite, and two metabolites are connected if there is a biochemical reaction that transforms one into the other, so that, in general, links represent reactions and the enzymes catalyzing them. However, a single link may correspond to several parallel reactions and/or enzymes (if distinct reactions transform one metabolite into the other and/or if different enzymes are involved in such transformation). Moreover, a single reaction may correspond to different links (if there is more than one reactant or product), while a few reactions occur spontaneously and therefore have no associated enzymes.

We start by quantifying the modular structure of the metabolic network for 18 different organisms obtained from the KEGG database (Goto *et al.*, 1998; Kanehisa and Goto, 2000) (see Methods Section). We use simulated annealing to find the partition of the network into modules that maximizes the modularity (Guimerà and Amaral, 2005b; Guimerà *et al.*, 2004) (see Methods Section), and assess the significance of the modular structure of each network by comparing it to a large number of randomizations of the same network (Guimerà *et al.*, 2004) (Fig. 1). We find that all networks have a significant modular structure.

Next, we determine the role of each node (see Methods Section). In our cartography, we classify nodes into seven roles according to their pattern of inter- and intra-module connections (Guimerà and Amaral, 2005a, b) (Fig. 1): (R1) ultra-peripheral nodes, (R2) peripheral nodes, (R3) satellite connectors, (R4) kinless nodes, (R5) provincial hubs, (R6) connector hubs and (R7) global hubs. Similarly, links can be classified in several types according to the roles of the nodes they connect. For example, R3–R5 links connect satellite connectors to provincial hubs. For simplicity, and because roles R4 and R7 rarely occur in real complex networks (Guimerà and Amaral, 2005a, b), we focus on roles R1, R2, R3, R5 and R6, and on the links connecting nodes with these roles.

We surmise that links of different types in the network correspond to enzymes with different functions and importance in the metabolism. Additionally, because metabolites with certain roles are significantly conserved across species (especially satellite connectors and connector hubs) while others are not (Guimerà and Amaral, 2005b), links of certain types are more likely to be species-specific.

3.2 Reactions with the highest median indispensability involve satellite connector metabolites

These considerations suggest a novel approach to identify promising drug targets. Indeed, by considering links of certain types, we should be able to identify enzymes that are indispensable for an organism but not needed by other organisms. Given that nodes with roles R3 and R6 are highly conserved across organisms, we hypothesize that links of types R3–R3, R3–R6 and R6–R6 will be, in general, essential to an organism. These links are, however, likely to be non-specific. In contrast, links connecting highly conserved metabolites (R3 or R6) to less conserved metabolites (e.g. R2) may still be essential because of the conserved end, and specific because of the non-conserved end. These link types thus appear to be natural candidates for drug targets.

To investigate this possibility, we analyze in depth the metabolic network of two bacteria: *E.coli* and *H.pylori* (the human pathogen responsible for gastritis and peptic ulcer disease). We select these two organisms because one can use FBA to estimate the fluxes of each of the reactions in their metabolism (Edwards and Palsson, 2000a; Reed *et al.*, 2003; Thiele *et al.*, 2005). Importantly, FBA also enables us to computationally test the effect of gene deletions that result in a specific reaction not taking place (Edwards and Palsson, 2000a; Thiele *et al.*, 2005).

We analyze systematically the effect of removing links from the metabolic networks of *E.coli* and *H.pylori*, i.e. we simulate the removal of all enzymes catalyzing the reactions connecting a certain pair of metabolites (Fig. 2). We quantify the effect of a link deletion by the *indispensability* I, which measures the relative difference between the growth rate of the mutant and that of the wild type (see Methods Section): a link with $I = 0$ can be removed without affecting the growth rate, whereas a link with $I = 1$ is indispensable for the survival of the organism.

We find that for all types of links there are instances of high indispensability. Our results indicate, however, that certain link types have a significantly higher likelihood of being indispensable than others (Fig. 2B), which confirms that our node and link classification scheme captures important biological information. In particular, we find that the links with the highest indispensability in both *E.coli* and *H.pylori* involve satellite connector metabolites (R3). Satellite connectors are metabolites that, despite participating in a relatively small number of biochemical reactions, bridge several different modules. The finding that links involving satellite connectors have high median indispensability may thus account for the high conservation rate of satellite connectors reported earlier (Guimerà and Amaral, 2005b): Since many reactions involving satellite connectors are indispensable, it is unlikely that a mutation causing the loss of the ability to process these metabolites results in a viable organism.

3.3 High indispensability link types are not over-represented in the metabolism

Biological systems are considered mutationally robust if they are able to function even after genetic mutations occur. Mutational robustness through redundancy and the so-called distributed robustness are thought to provide an evolutionary advantage to organisms, which would explain the pervasiveness of robustness in biological systems (Wagner, 2005). Within this context, however, our findings on link indispensability are somewhat puzzling: some links in the metabolism are highly fragile and the these links often involve satellite connector metabolites, which seems to contradict the hypothesis that fragile links are random evolutionary 'accidents'.

A relevant question is, thus, whether: (i) evolution tries to provide backups for links in link types that are likely to be fragile, however it cannot backup all of these links so that those link types are still more fragile than one would expect, or (ii) fragile link types are fragile and (or because) evolution does not back them up. To address this question quantitatively, we consider the FBA reconstructions for *E.coli* and *H.pylori* as well as the KEGG reconstructions for these and other bacteria, archaea and eukaryotes. For each network, we obtain the role-to-role connectivity profiles by computing the number r_{ij} of links between nodes belonging to roles i and j, and comparing this number to the number R_{ij} of such links that are expected to appear purely by chance (Fig. 3, see Methods Section and Supplementary Fig. S1).

We find that the role-to-role connectivity profiles are consistent across domains, which indicates that the profiles are a result of the fundamental tasks that all metabolic networks need to carry out, and not of the particularities of each metabolic network. Remarkably, we also find that the most indispensable link types (R2–R3, R3–R3, and R3–R6) are not significantly or consistently over-represented in the 18 organisms we consider (Fig. 3). This result seems to indicate that fragile link types are fragile and (or because) evolution does not back them up.

Although we do not have a definite theory to understand this finding, we put forward two plausible explanations, both of them closely connected to the modular structure of metabolic networks. One possibility is that satellite connectors have chemical properties that make them unique in their ability to bridge two or more modules whose metabolites have, otherwise, little in common. In this case, it would be impossible for organisms to develop alternative pathways. A second hypothesis is that some 'fragile' links between modules provide a net evolutionary advantage, even at the price of making the metabolism less robust. This advantage may arise from the fact that fragile links between modules (i.e. those that do not have backups) enable some degree of module independence and can act as system-level regulation points: a small change in the fluxes through these links can reshape the *global* distribution of the metabolic fluxes, and can switch on and off whole modules (for similar ideas in neuroscience, (Hilgetag and Kaiser, 2004; Sporns *et al.*, 2000)).

3.4 Enzymes connecting peripheral metabolites and satellite connectors are promising drug targets

Besides posing interesting fundamental questions about the evolution of metabolism, our analysis also opens the door for novel approaches to exploit the information hidden in metabolic networks. We have established that, at least in *E.coli* and *H.pylori*, biochemical reactions involving satellite-connector metabolites are particularly vulnerable. These reactions could, therefore, be of great importance to metabolic engineering and be natural targets for drugs.

Enzymes connecting peripheral metabolites and satellite connectors (R2–R3) seem the most promising as drug targets, given that peripheral metabolites (R2) are not highly conserved across species (Guimerà and Amaral, 2005b). This means that it should be possible to find vulnerable and species-specific links of type R2–R3 in pathogens. To investigate this possibility, we study the conservation in humans of the enzymes that correspond to different types of links in *E.coli* and *H.pylori* (Fig. 2C). We find that, as hypothesized, the most conserved enzymes correspond to links involving connector hubs (R6), and that enzymes corresponding to R2–R3 links are not significantly more conserved than enzymes of types not involving R6 nodes.

4 CONCLUSION

As we have shown, FBA does predict that enzymes of types R2–R3, R3–R3 and R3–R6 have greater median indispensability than other enzyme types. The true test of our predictions, however, will have to come from new experimental studies. New experiments are necessary because two important aspects reduce the usefulness of existing system-level databases on gene essentiality (Gerdes *et al.*, 2003; Thiele *et al.*, 2005) (Fig. 4).

The first limitation of existing databases is that essentiality is assumed to be a binary variable, i.e. either a gene is considered to be essential or it is considered to be non-essential. This simplification overlooks the fact that the absence of certain genes in a mutant may severely reduce its growth rate, which would render the mutant inviable in selective environments (Fig. 4A). Concretely, a host is a selective environment because of the immune response triggered by the presence of the pathogen. Another selective environment is one in which the mutant competes for limited resources with the wild type or with other organisms.

The second limitation of current databases is that they store experimental results on essentiality obtained for rich media, whereas many genes will be essential only when certain compounds are not available (Fig. 4B). Papp *et al.* (2004) have predicted that, for *Saccharomyces cerevisiae*, over 50% of the genes that are non-essential in rich media will be essential in more stringent conditions.

Our results therefore suggest that a more thorough description of essentiality is necessary, and our network-based method to identify and classify enzymes may be a useful guide for further and more exhaustive experiments, specially on enzymes corresponding to links of type R2–R3, which are likely to have fundamental and practical importance.

In the absence of such comprehensive experimental studies, we have nonetheless been able to verify that some enzymes that correspond to R2–R3 links in *E.coli* have been previously suggested or are being investigated as potential drug targets. The enzymes encoded by genes *glmS* (EC 2.6.1.16) *pfs*

(EC 3.2.2.16 3.2.2.9) and *ptsI* (EC 2.7.3.9) catalyze transformations that our method assigns a high probability (>33%) of being of type R2–R3 in *E.coli*, and are research targets for bacterial infections according to the Therapeutic Target Database (TTD) (Chen *et al*., 2002). It is important to note that these enzymes also catalyze transformations of types other than R2–R3. This turns out to be a quite general situation for the enzymatic targets listed in the TTD; most target enzymes listed in the database correspond to multiple links in the *E.coli* metabolic network, often including several R1–R1 and R1–R2 links and sometimes including links that involve connector hubs (R6) and, to a lesser extent, provincial hubs (R5). This suggests that the 'effectiveness' of current enzymatic targets might be more related to removing many links in the metabolic network, with the associated lack of specificity and potential side effects, than to their ability to actually interfere with critical species-specific links. We believe that our results will help clarifying these issues, providing justification to why certain enzymes are better targets than others, and will open the door to the guided discovery of new targets.

ACKNOWLEDGEMENTS

We thank V. Hatzimanikatis, W.M. Miller, A. Mongragón, D.B. Stouffer and L.K. Wing for useful comments and suggestions. We thank J. Reed, I. Thiele and the Systems Biology Research Group at UCSD for their help and for making their FBA reconstructions publicly available. R.G. and M.S.-P. gratefully acknowledge support from the Fulbright Program. L.A.N.A. gratefully acknowledges support from a NIH/NIGMS K-25 award, from the J.S. McDonnell Foundation, and from the W.M. Keck Foundation.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Amaral,L.A.N. and Ottino,J. (2004) Complex networks: Augmenting the framework for the study of complex systems. *Eur. Phys. J. B*, **38**, 147–162.

Bax,R. *et al*. (2000) The millennium bugs – the need for and development of new antibacterials. *Int. J. Antimicrob. Agents*, **16**, 51–59.

Chen,X. *et al*. (2002) TTD: Therapeutic Target Database. *Nucleic Acids Res*., **30**, 412–415.

Danon,L. *et al*. (2005) Comparing community structure identification. *J. Stat. Mech.: Theor. Exp*., P09008.

Edwards,J.S. and Palsson,B.Ø. (2000a) The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl Acad. Sci. USA*, **97**, 5528–5533.

Edwards,J.S. and Palsson,B.Ø. (2000b) Robustness analysis of the Escherichia coli metabolic network. *Biotechnol. Prog*., **16**, 927–939.

Gerdes,S.Y. *et al*. (2003) Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. *J. Bacteriol*., **185**, 5673–5684.

Girvan,M. and Newman,M.E.J. (2002) Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA*, **99**, 7821–7826.

Goto,S. *et al*. (1998) LIGAND: chemical database for enzyme reactions. *Bioinformatics*, **14**, 591–599.

Guimerà,R. and Amaral,L.A.N. (2005a) Cartography of complex networks: modules and universal roles. *J. Stat. Mech.: Theor. Exp*., P02001.

Guimerà,R. and Amaral,L.A.N. (2005b) Functional cartography of complex metabolic networks. *Nature*, **433**, 895–900.

Guimerà,R. *et al*. (2004) Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E*, **70**, art. no. 025101.

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

Han,J.-D.J. *et al*. (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**, 88–93.

Hartwell,L.H. *et al*. (1999) From molecular to modular biology. *Nature*, **402**, C47–C52.

Hilgetag,C.C. and Kaiser,M. (2004) Clustered organization of cortical connectivity. *Neuroinformatics*, **2**, 353–360.

Holme,P. *et al*. (2003) Subnetwork hierarchies of biochemical pathways. *Bioinformatics*, **19**, 532–538.

Itzkovitz,S. *et al*. (2004) Reply to ''comment on 'subgraphs in random networks' ''. *Phys. Rev. E*, **70**, art. no. 058102.

Jeong,H. *et al*. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–654.

Jeong,H. *et al*. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

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

Kanehisa,M. *et al*. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*., **34**, D354–D357.

Klamt,S. and Gilles,E.D. (2004) Minimal cut sets in biochemical reaction networks. *Bioinformatics*, **20**, 226–234.

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

Maslov,S. and Sneppen,K. (2002) Specificity and stability in topology of protein networks. *Science*, **296**, 910–913.

Newman,M.E.J. (2003) The structure and function of complex networks. *SIAM Review*, **45**, 167–256.

Newman,M.E.J. and Girvan,M. (2004) Finding and evaluating community structure in networks. *Phys. Rev. E*, **69**, art. no. 026113.

Nikolsky,Y. *et al*. (2005) Biological networks and analysis of experimental data in drug discovery. *Drug. Discov. Today*, **10**, 653–662.

Norrby,S.R. *et al*. (2005) Lack of development of new antimicrobial drugs: a potential serious threat to public health. *Lancet Infect. Dis*., **5**, 115–119.

Palumbo,M.C. *et al*. (2005) Functional essentiality from topology features in metabolic networks: a case study in yeast. *FEBS Lett*., **579**, 4642–4646.

Papp,B. *et al*. (2004) Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature*, **429**, 661–664.

Rahman,S.A. and Schomburg,D. (2006) Observing local and global properties of metabolic pathways: 'load points' and 'choke points' in the metabolic networks. *Bioinformatics*, **22**, 1767–1774.

Ravasz,E. *et al*. (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.

Reed,J.L. *et al*. (2003) An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR) *Genome Biol*., **4**, R54.

Spellberg,B. *et al*. (2004) Trends in antimicrobial drug development: implications for the future. *Clin. Infect. Dis*., **38**, 1279–1286.

Sporns,O. *et al*. (2000) Connectivity and complexity: the relationship between neuroanatomy and brain dynamics. *Neural Netw*., **13**, 909–922.

Tanaka,R.(2005) Scale-rich metabolic networks. *Phys. Rev. Lett*., **94**, art. no. 168101.

Thiele,I. *et al*. (2005) Expanded metabolic reconstruction of Helicobacter pylori (iIT341 GSM/GPR): an in silico genome-scale characterization of single- and double-deletion mutants. *J. Bacteriol*., **187**, 5818–5830.

Wagner,A. (2005) Distributed robustness versus redundancy as causes of mutational robustness. *Bioessays*, **27**, 176–188.

Wagner,A. and Fell,D.A. (2001) The small world inside large metabolic networks. *Proc. Roy. Soc. B*, **268**, 1803–1810.

Winston,W.L. and Venkataramanan,M. (2002) *Introduction to Mathematical Programming: Applications and Algorithms*. 4th edn. Duxbury Press, Belmont, CA, USA.