BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 22 2008, pages 2579–2585
doi:10.1093/bioinformatics/btn503

Phylogenetics

Phylogenetic distances are encoded in networks of interacting pathways

Aurélien Mazurie[1,2,*], Danail Bonchev[2], Benno Schwikowski[1] and Gregory A. Buck[2]

[1]Systems Biology Group, Institut Pasteur, 25 rue du Docteur Roux, 75015 Paris, France and [2]Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284-2030, USA

Received on May 19, 2008; revised on September 3, 2008; accepted on September 19, 2008

Advance Access publication September 26, 2008

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Although metabolic reactions are unquestionably shaped by evolutionary processes, the degree to which the overall structure and complexity of their interconnections are linked to the phylogeny of species has not been evaluated in depth. Here, we apply an original metabolome representation, termed Network of Interacting Pathways or NIP, with a combination of graph theoretical and machine learning strategies, to address this question. NIPs compress the information of the metabolic network exhibited by a species into much smaller networks of overlapping metabolic pathways, where nodes are pathways and links are the metabolites they exchange.

Results: Our analysis shows that a small set of descriptors of the structure and complexity of the NIPs combined into regression models reproduce very accurately reference phylogenetic distances derived from 16S rRNA sequences (10-fold cross-validation correlation coefficient higher than 0.9). Our method also showed better scores than previous work on metabolism-based phylogenetic reconstructions, as assessed by branch distances score, topological similarity and second cousins score. Thus, our metabolome representation as network of overlapping metabolic pathways captures sufficient information about the underlying evolutionary events leading to the formation of metabolic networks and species phylogeny. It is important to note that precise knowledge of all of the reactions in these pathways is not required for these reconstructions. These observations underscore the potential for the use of abstract, modular representations of metabolic reactions as tools in studying the evolution of species.

Contact: aurelien.mazurie@pasteur.fr

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

1 INTRODUCTION

Phylogenetic relationships between species are traditionally inferred from genomic data, based on observed mutations in the sequence of orthologous genes found in all studied species—a typical example being the SSU rRNA (16S rDNA) gene sequence (Olsen *et al.*, 1994). Results obtained are potentially biased, however, by the highly variable rates of evolution observed across species (Huynen and Bork, 1998). Moreover, identification of orthologs and paralogs in the genomes is complicated by gene duplication and loss, horizontal gene transfer and functional replacement events, resulting in misannotations.

Recently, higher level functional components have been considered as replacements for or complements of gene-based phylogenies. The annotations of the metabolic reactions are the most promising source of information due to the abstraction of the cellular functions they provide and their availability in numerous species (Kanehisa *et al.*, 2006). One approach to exploit this information is to calculate a distance between species based on the enzymatic genes found in their genome, or on the network of reactions they define by exchanging metabolites, or both; links between these two aspects have been demonstrated (Liu *et al.*, 2007). Examples include phylogenies inferred from the presence or absence of enzymes in the genomes, either alone (Ma and Zeng, 2004) or in combination with the metabolic network structure (Forst *et al.*, 2006; Oh *et al.*, 2006; Zhang *et al.*, 2006), from the similarity of enzyme sequences or functional annotation in combination with the comparison of their direct neighbors in the reactions network (Clemente *et al.*, 2007; Forst and Schulten, 2001; Heymans and Singh, 2003), from the presence or absence of pathways across species (Liao *et al.*, 2002) and from the completeness of pathways across species (Hong *et al.*, 2004).

In these studies, metabolic reactions are represented as directed or undirected graphs. Nodes either represent metabolites that are linked by the enzymes that process them, or enzymes linked by metabolites they exchange. However, in addition to the large amount of information required across all species for meaningful comparisons, these representations are potential sources of bias, whose impact has not been evaluated in phylogeny reconstruction. This presents several issues. First, incorporation of the so-called ubiquitous metabolites, e.g. water, connects functionally distant metabolites without real mechanistic biological meaning, producing an unrealistically small degree of separation of nodes (Ma and Zeng, 2003). The criteria by which metabolites should be included or excluded in this context are unclear. Second, the structure of these networks is highly sensitive to annotation errors, as, especially in newly sequenced genomes, the presence of orthologous enzymes in species is initially assessed by sequence similarity. In addition to the risk of false positives or negatives, the exact set of reactions in which the putative enzyme is involved may not match those in the reference species from which the annotations are transferred. This is even more critical when the transferred annotation is a generic

*To whom correspondence should be addressed.

enzyme name, such as an EC code, which, due to its abstract nature, can be associated to several distinct physical reactions.

Here, we describe a new representation in which metabolic reactions are represented as an undirected, weighted network of interacting pathways (NIPs). Nodes in NIPs are metabolic pathways, i.e. non-exclusive and consensual sets of metabolic reactions as defined by the reference source KEGG (Kanehisa *et al.*, 2006). Edges link overlapping pathways sharing at least one metabolite, i.e. at least one enzyme in each of the two pathways uses this metabolite as a substrate or product. This representation is designed to be less sensitive to common biases from annotation errors and other sources, since false positives and negatives are less likely to occur at the level of a pathway than for an enzyme. Still, NIPs depend on the definition of metabolic pathways proposed by reference databases. The wide use of KEGG as a reference pathway source shows, however, that these definitions are in practice employed as standard representations of metabolism by the biochemistry community and are unlikely to be greatly modified in the future. Other algorithmic-based representations of metabolic networks based on genome-scale data (e.g. gene expression, topology of the reaction network) have recently been proposed in the literature—see Aittokallio and Schwikowski (2006) for a review. The relationship between these novel representations and the phylogeny of species is currently under investigation.

We anticipated that this higher hierarchical level of organization of metabolic networks would reveal patterns of their evolution by being more focused on the notion of modularity, an emergent property of networks that has been studied extensively (Hartwell *et al.*, 1999; Papin *et al.*, 2004; Ravasz *et al.*, 2002; Spirin *et al.*, 2006) but which cannot be easily extracted from the genome sequence alone. This new representation is expected to better capture phylogenetic relationships among species than previous approaches, by focusing less on the components (enzymes and metabolites) of metabolic pathways and more on how they interact in a modular manner.

2 METHODS

The general approach used to measure the correlation between phylogenetic distances and structure of metabolic networks is summarized in Figures 1 and 2 and below.

2.1 Extraction of metabolic networks

Metabolic reactions were retrieved from two public sources, the December 2006 release of the KEGG database (Kanehisa *et al.*, 2006), and the November 2006 release of the Ma dataset (Ma and Zeng, 2003). The latter source is a manually curated version of the former for 107 species out of the 289 available from KEGG. We reconstructed two networks, a network of interacting pathways (NIP) and a network of interacting metabolites (NIM), for each of the species. NIPs were built by linking overlapping metabolic pathways sharing at least one metabolite (Fig. 1). For comparison, NIMs were also built by linking metabolites converted in a reaction occurring in at least one metabolic pathway. Edges in these two undirected graphs are weighted, either by the number of metabolites shared (in NIPs) or by the number of pathways in which metabolites are converted (in NIMs), respectively. The weight of a node is the sum of weights of its incident edges. Note that NIPs contain no information about the underlying metabolic reactions or the enzymes that catalyze them, and only keep information about which metabolic pathways are present in the species and how they overlap.

To account for the potential bias represented by ubiquitous metabolites, two variations of the NIPs and NIMs datasets were considered. The first, termed 'filtered', excludes all metabolites considered ubiquitous by the authors or the respective source. The second, termed 'unfiltered', only excludes water. To compare results obtained with the KEGG and Ma sources, the same 107 species were considered in both. A description of the metabolic pathways used for the construction of the NIPs is provided in Supplementary Table 1.

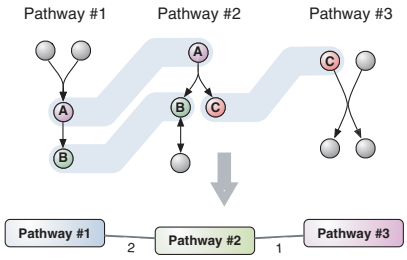

Fig. 1. Extraction of NIPs from metabolic networks. A list of all metabolites processed is compiled for each pathway known to exist in a given species; example is given here of fictive metabolites A, B and C processed in three metabolic pathways. Pathways that use or produce metabolites found in other pathways are linked together (shaded lines). Links are weighted by the number of metabolites exchanged.

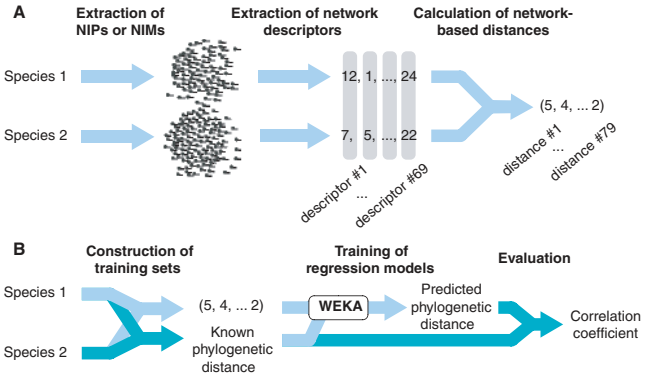

Fig. 2. General approach. (**A**) For any given species, metabolic networks are extracted and descriptors of their structure and complexity are calculated. Network-based distances between all pairs of species are derived from the NIP and NIM descriptors. In this example, the three descriptor values are numerical and the resulting distances are the arithmetic difference. (**B**) Correlation between network-based distances and 16S-based phylogenetic distance is measured by training regression models. The models heuristically search for combinations of network-based distances best predicting the phylogenetic distance. Performance is expressed as the Pearson's correlation coefficient between known and predicted phylogenetic distances.

2.2 Reference phylogenetic distances

The phylogenetic distance matrix used as a reference was derived from a multiple alignment of the gene sequences for the small subunit of the ribosomal RNA of each of the 107 species by employing a DNA sequence evolution model. The sequences were retrieved from the European ribosomal RNA database (Wuyts *et al.*, 2004) and the GenBank database (Benson *et al.*, 2006), and aligned using CLUSTALW (Chenna *et al.*, 2003). The DNA evolution model used, GTR+I+G, was the one best fitting the alignment data, as determined by MODELTEST (Posada and Crandall, 1998) using hierarchical likelihood ratio tests involving 56 different models available in PAUP* (Swofford, 2003). We excluded 9 of the 107 species due to uncertain identifier matching in the database. The 98 remaining species were grouped into 80 taxa

to include strains of the same species, resulting in 12 Archaea, 60 Bacteria and 8 Eukarya representing 15%, 75% and 10% of the total, respectively. The list of these 80 taxa with their main taxonomic ranks (domain, kingdom and class) and the KEGG identifiers of the associated species and strains are presented in Supplementary Table 2. Phylogenetic trees were inferred from the resulting distance matrices using the neighbor-joining algorithm implemented by the NEIGHBOR program of the PHYLIP toolbox (Felsenstein, 1989).

2.3 Description of metabolic networks

Networks can be characterized both qualitatively and quantitatively using graph theory (Harary, 1969) and information theory (Weaver and Shannon, 1949), by applying a variety of topological, compositional and information-theoretic descriptors (Bonchev and Buck, 2005); i.e. quantities that are uniquely associated with specific aspects of network structure and complexity. Four categories of descriptors—degree, centrality, distance and cliques-related—were considered, with a total of 35 unique descriptors, some of them devised specifically for this study. Weighted and unweighted flavors of descriptors were considered for those descriptors related to node and edge count, and three different versions of their information content were used for those descriptors related to values distributions. An expanded set of 69 descriptors (35 unique plus 34 derivatives) was thus constructed (Supplementary Table 3 and associated references). Compositional descriptors (i.e. list of nodes) reporting only parts of metabolic networks (largest cliques, nodes at center) were selected to be highly sensitive to the whole network structure, thus lowering the risk of collision (similar values even when the network is significantly different). The values of these descriptors were calculated for each NIP and NIM using the NETWORKX library.[1]

2.4 Network-based distances between taxa

Based on the above expanded set of 69 network descriptors, we computed a pairwise distance vector between each pair of the 80 taxa (Fig. 2A). The distance between the values of each descriptor was calculated according to its type. For numeric descriptors, this distance was the absolute value of the difference. When the descriptor was a vector of numeric values (e.g. node degree distribution) we used three different distance functions; the sum of the absolute values of the difference between each element, the Manhattan and the Euclidean distance. When the descriptor was a set (e.g. a list of network nodes), we used the Jaccard distance—the ratio between the cardinality of the intersection and the cardinality of the union of the two sets. When taxa were represented by several strains or individuals, the distance between each of their descriptor values was taken as the mean of the pairwise distances calculated between the strains. The use of several distance calculations for some descriptors (see Supplementary Table 3) resulted in a distance vector of 79 distance values for each pair of taxa. A dataset was constructed as described for the Archaea, Bacteria, Eukarya and for the 80 taxa together.

2.5 Correlation estimation

The correlation between network-based distances and reference phylogenetic distances of taxa was assessed by training regression models to predict the latter from any combination of the former (Fig. 2B). Training sets were constructed to report, for each pair of taxa and for each metabolic network dataset, the two types of distances. These training sets are available as Supplementary Table 4. Supervised learning algorithms implemented in the WEKA toolbox (Witten and Frank, 1999 and Supplementary Table 5) were applied on the training sets to reproduce, i.e. predict, the phylogenetic distance from any combination of network distances. A Pearson's coefficient of the 10-fold cross-validation and that of the whole training set (referred to as q^2 and R^2, respectively) was calculated by comparing known and predicted phylogenetic distances. For a given training set, the correlation between network-based and phylogenetic distances was then taken as the highest q^2 obtained among all regression models. A high score would mean that phylogenetic distances are fully encoded in, i.e. they can be calculated from, the structure and organization of metabolic networks. To detect any overfitting, 10 randomized versions of each training set were also evaluated, in which reference phylogenetic distances were shuffled using the Fisher–Yates algorithm (Fisher and Yates, 1938).

Finally, we identified the smallest subset of network descriptors that still performs as well as the complete set. This was done using feature selection algorithms (Guyon and Elisseeff, 2003; Hall and Holmes, 2003) and a heuristic evaluation of subsets of descriptors on the regression models identified earlier as the best ones. A tool, METACLASSIFY, was developed to automate the training of the regression models and to retrieve the results.[2]

3 RESULTS AND DISCUSSION

3.1 Networks of interacting pathways

NIPs were constructed to represent the metabolism of species, as outlined in the Section 2 and Figure 1, from two metabolic reaction datasets: KEGG and Ma for the same 107 species, with ubiquitous metabolites either removed (filtered dataset) or kept (unfiltered).

A NIP contains 37% to 97% of all known metabolic pathways of the 107 species; an example is shown in Figure 3. Use of NIPs instead of the entire network of metabolic reactions (NIMs) represents an 8- to 11-fold compression of the network size, from an average of 507 metabolites down to an average of 63 pathways. NIPs are also more compact, with an average node degree of 9.0 ± 3.6 to 48.0 ± 15.6 (filtered and unfiltered version, respectively) to compare with values of 2.4 ± 0.1 to 5.1 ± 0.3 for NIMs. As shown below, this substantial compression nevertheless conserves all information needed to accurately reconstruct phylogeny of species.

3.2 Prediction of the phylogenetic distance

We assessed the correlation between metabolic network-based distances and phylogenetic distances by training regression models, for all pairs of species considered (Fig. 2 and Section 2). These models were trained to predict phylogenetic distance from any combination of network-based distances. The correlation coefficients between predicted and reference phylogenetic distances calculated from the 16S rRNA sequences, evaluated using 10-fold cross-validation (q^2) and on the whole training set (R^2) are given in Table 1. Their analysis led to the following observations.

First, the accuracy of the phylogenetic distance prediction from our set of 79 descriptors of metabolic network structure and complexity is high, for both NIPs and NIMs (q^2 of 0.92 ± 0.02 and 0.93 ± 0.04, respectively). This observation demonstrates the utility of metabolic network organization for phylogeny reconstruction, and compares very favorably with similar work (see below). The average relative error in phylogenetic distance prediction is highest for small distances (~18% for distances below 0.2), and decrease exponentially for larger distances (from ~4% to ~0.75% for distances above 0.2; data not shown).

Second, both types of metabolic network representations perform equally well, although NIPs are better than NIMs at reconstructing phylogeny of Eukarya (q^2 of 0.79 ± 0.12 and 0.70 ± 0.3, respectively). We show here that the amount of information required to build NIMs (i.e. the full set of metabolic

[1] https://networkx.lanl.gov/

[2] http://oenone.net/tools/

Fig. 3. Example of NIP. NIPs extracted from the filtered KEGG metabolic dataset for *Saccharomyces cerevisiae*. Node shade is proportional to the number of metabolic pathways overlapping with the represented one. Edge shade is proportional to the number of metabolites exchanged.

Table 1. Accuracy of the inferred phylogenetic distances

Domain	Network	Filtering	Source	q^2	R^2	Regression model
All	NIM	Filtered	KEGG	0.9482	0.9985	Functions.GaussianProcesses
All	NIM	Filtered	Ma *et al.*	0.8858	0.9983	Functions.GaussianProcesses
All	NIM	Unfiltered	KEGG	0.9683	0.9998	Functions.GaussianProcesses
All	NIM	Unfiltered	Ma *et al.*	0.9166	0.9993	Functions.GaussianProcesses
All	NIP	Filtered	KEGG	0.8859	0.9993	Functions.GaussianProcesses
All	NIP	Filtered	Ma *et al.*	0.9351	0.9996	Functions.GaussianProcesses
All	NIP	Unfiltered	KEGG	0.9169	0.9994	Functions.GaussianProcesses
All	NIP	Unfiltered	Ma *et al.*	0.9356	0.9997	Functions.GaussianProcesses
Archaea	NIM	Filtered	KEGG	0.4879	0.7713	Functions.LinearRegression
Archaea	NIM	Filtered	Ma *et al.*	0.4956	0.8403	Functions.SMOreg
Archaea	NIM	Unfiltered	KEGG	0.6405	0.9979	Functions.MultilayerPerceptron
Archaea	NIM	Unfiltered	Ma *et al.*	0.8336	0.9228	Functions.LinearRegression
Archaea	NIP	Filtered	KEGG	0.3579	0.9999	Functions.MultilayerPerceptron
Archaea	NIP	Filtered	Ma *et al.*	0.6968	0.915	Functions.LinearRegression
Archaea	NIP	Unfiltered	KEGG	0.6803	0.7034	Functions.SimpleLinearRegression
Archaea	NIP	Unfiltered	Ma *et al.*	0.7054	0.9987	Functions.MultilayerPerceptron
Bacteria	NIM	Filtered	KEGG	0.8267	0.9964	Functions.GaussianProcesses
Bacteria	NIM	Filtered	Ma *et al.*	0.7778	0.9964	Functions.GaussianProcesses
Bacteria	NIM	Unfiltered	KEGG	0.8497	0.9996	Functions.GaussianProcesses
Bacteria	NIM	Unfiltered	Ma *et al.*	0.8145	0.9991	Functions.GaussianProcesses
Bacteria	NIP	Filtered	KEGG	0.8539	0.9993	Functions.GaussianProcesses
Bacteria	NIP	Filtered	Ma *et al.*	0.8471	0.9986	Functions.GaussianProcesses
Bacteria	NIP	Unfiltered	KEGG	0.7867	0.9993	Functions.GaussianProcesses
Bacteria	NIP	Unfiltered	Ma *et al.*	0.8371	0.9993	Functions.GaussianProcesses
Eukarya	NIM	Filtered	KEGG	0.4776	0.9998	Functions.MultilayerPerceptron
Eukarya	NIM	Filtered	Ma *et al.*	0.9511	0.9898	Trees.REPTree
Eukarya	NIM	Unfiltered	KEGG	0.3981	0.8009	Functions.IsotonicRegression
Eukarya	NIM	Unfiltered	Ma *et al.*	0.966	0.9848	Lazy.LWL
Eukarya	NIP	Filtered	KEGG	0.7916	1	Functions.LinearRegression
Eukarya	NIP	Filtered	Ma *et al.*	0.768	0.997	Functions.SMOreg
Eukarya	NIP	Unfiltered	KEGG	0.6572	0.8125	Trees.DecisionStump
Eukarya	NIP	Unfiltered	Ma *et al.*	0.9525	1	Functions.LinearRegression

Correlation coefficients between reference 16S phylogenetic distances and distances predicted from descriptors of NIPs and NIMs, for the filtered and unfiltered versions of the metabolic pathways datasets are tested. The description of the regression models used is given in Supplementary Table 5. The correlation coefficient determined from the 10-fold cross-validation (q^2) and from the whole dataset (R^2) is given for each model. Values are given for all 80 taxa and for each domain individually.

reactions) is not necessary to perform good reconstructions, and can advantageously be replaced by NIPs (i.e. knowledge of which pathways are present and which metabolites they exchange). This observation is particularly important in the context of missing or erroneous genome annotations, which are a particular problem with newly sequenced genomes.

Third, unfiltered datasets perform better than filtered datasets. The additional structural information provided by ubiquitous metabolites slightly improves reconstructions of phylogenies. This effect is observed with equal strength in NIMs (q^2 of 0.94 ± 0.04 and 0.92 ± 0.04 for unfiltered and filtered datasets, respectively) and in NIPs (0.93 ± 0.01 and 0.91 ± 0.03, respectively).

When considering the species domains of Archaea, Bacteria and Eukarya independently, the performances are still good—an average q^2 of 0.61 ± 0.15, 0.82 ± 0.03 and 0.74 ± 0.21, respectively. However, high differences between the q^2 and R^2 in the Archaea and Eukarya indicate some overfitting that may be due to the small size of these domains (15% and 10% of the datasets, respectively).

No such overfitting could be detected when reconstructing phylogeny of all species, as shown by the small difference between q^2 and R^2 and by the low scores obtained with randomized training sets in which known 16S phylogenetic distances were shuffled (see Section 2). The highest q^2 achieved by regression models in these randomized sets was 0.07 and 0.08 for NIPs and NIMs, respectively. These results demonstrate that our approach is robust against overfitting: regression models do not report artifactual relationship between metabolic network structure and the phylogeny of species after being trained on deliberately incorrect datasets where this relationship was effectively destroyed.

3.3 Prediction of the phylogenetic tree

The performance of the phylogeny reconstruction from metabolic network descriptors was also evaluated by comparing the trees inferred from the predicted phylogenetic distances with the reference

Fig. 4. Example of predicted tree. Example of phylogenetic tree predicted from descriptors of NIPs from the unfiltered version of the Ma dataset, for all 80 taxa considered in this study (right). For comparison, the tree resulting from the 16S sequences is also given (left). These two trees show minor differences (highlighted in bold) as indicated by the high similarity scores obtained on subsets of taxa (Table 2). Branch lengths are displayed as equal for the purpose of display.

16S tree. An example of tree obtained is shown in Figure 4, with discrepancies highlighted.

By using the same reference tree, subset of taxa and scores, we directly compared the performance of our approach which those of Heymans and Singh (2003), Forst *et al.* (2006), Zhang *et al.* (2006) and Clemente *et al.* (2007), where phylogeny reconstruction from metabolic data was also considered (Table 2). These studies were shown to outperform previous similar approaches from Forst and Schulten (2001) and Liao *et al.* (2002). For the same sets of 16 and 8 taxa used in Heymans and Singh (2003) and Clemente *et al.* (2007) respectively, our approach achieved better second cousins scores of 0.3 to 0.737 and 0.625 to 1, to compare with the scores of 0.27 and 0.571, respectively reported. For the same set of 27 taxa used in Forst *et al.* (2006), our approach achieved better branch distance scores of 0.005 to 0.021 (except for three out of our eight metabolic datasets), to compare with the score of 0.023 reported. Finally, for the same set of 47 taxa used in Zhang *et al.* (2006), our approach achieved better Penny and Hendy's topological similarity scores of 0.7 to 0.95, to compare with the score of 0.386 reported.

Table 2. Accuracy of the inferred phylogenetic trees

Network	Filtering	Source	16 taxa from Heymans *et al.* Cousins	27 taxa from Forst *et al.* BSD	SD	47 taxa from Zhang *et al.* PH	8 taxa from Clemente *et al.* Cousins
NIM	Filtered	KEGG	**0.489**	0.053	6	**0.700**	**1.000**
NIM	Filtered	Ma *et al.*	**0.300**	**0.020**	6	**0.775**	**1.000**
NIM	Unfiltered	KEGG	**0.579**	**0.005**	**2**	**0.950**	**1.000**
NIM	Unfiltered	Ma *et al.*	**0.737**	**0.011**	**2**	**0.900**	**1.000**
NIP	Filtered	KEGG	**0.340**	**0.016**	8	**0.900**	**1.000**
NIP	Filtered	Ma *et al.*	**0.550**	**0.021**	**2**	**0.925**	**0.625**
NIP	Unfiltered	KEGG	**0.400**	0.259	24	**0.850**	**1.000**
NIP	Unfiltered	Ma *et al.*	**0.319**	0.058	4	**0.950**	**1.000**
Compared authors' results			0.27	0.023	2	0.386	0.571

Distance between reference 16S tree and trees inferred from predicted phylogenetic distances. Results are given for the same subsets of 16, 27, 47 and 8 taxa considered in Heymans and Singh (2003), Forst *et al.* (2006), Zhang *et al.* (2006) and Clemente *et al.* (2007), respectively. Cousins: second cousin similarity; a value of 1 means identical trees (Shasha *et al.*, 2004). BSD and SD: branch score distance and symmetric difference; a value of 0 means identical trees (Felsenstein, 1989). PH: Penny and Hendy's topological similarity (Paradis *et al.*, 2004; Penny and Hendy, 1985); a value of 1 means identical trees. Given in bold are those scores where our method performs comparably or better than studies cited above.

3.4 Best predictors of the phylogenetic distance

Descriptors of NIP structure and complexity do not contribute equally to phylogeny reconstruction. For the filtered NIP datasets from KEGG and Ma, we were able to significantly reduce their number from 79 to 22 descriptors in both datasets (Supplementary Table 6, abridged in Table 3 into 16 and 14 non-redundant descriptors for the KEGG and Ma datasets, respectively), while performing nearly as well in predicting the phylogenetic distances among the taxa. Our study is the first to identify the precise aspects of metabolic network structure and complexity that best encode the phylogeny of species.

Table 3. Descriptors best predicting phylogenetic distance (abridged)

Descriptor	KEGG	Ma *et al.*
Vertex clustering coefficient	#1	#3
Average vertex distance degree	#4	#1
Vertex eccentricity	#2	#5
Largest cliques	#10	#4
Information on vertex degree magnitude distribution	#6	#9
Diameter	#8	#8
Radius	#3	#13
Information on distance distribution	#11	#11
Clique distribution		#2
Vertices at center	#5	
Number of vertices		#6
Vertex degree distribution		#7
Information on clique distribution	#7	
Total graph distance	#9	
Information on vertex degree distribution		#10
Number of connected components		#12
Information on clique size distribution	#12	
Number of cliques	#13	
Vertex degree	#14	
Average edge betweenness centrality		#14
Information on distance degree magnitude distribution	#15	
Average graph distance	#16	

Subsets of NIP descriptors performing together nearly as well as the full set of 79 at predicting the phylogenetic distance among species, ranked by decreasing contribution to the prediction. Derivatives of the same descriptor were replaced by a representative (see Supplementary Table 3 and Section 2). Unabridged subsets are available as Supplementary Table 6. The subsets identified for the filtered KEGG and Ma datasets led to a q^2 of 0.876 and 0.905, respectively.

Analysis of these lists shows an interesting combination of descriptors related to degree distribution, distance distribution, clique composition and clique-size distribution. Importance of degree and distance distribution in describing NIPs supports the hypothesis of a link between the scale-freeness (Barabasi and Albert, 1999) and small-worldness (Watts and Strogatz, 1998) of biological networks and the phylogeny of species. A surprising result of our analysis is the apparent significant role of NIP cliques, i.e. groups of completely interconnected pathways. Large cliques are found in NIPs (up to 20 pathways), while NIMs typically have small cliques (3 to 5 metabolites). Metabolism of species is organized around a core of highly overlapping pathways, the structure and composition of which are important to distinguish these species. In terms of the KEGG nomenclature, this core is dominated by carbohydrate and amino acid metabolic pathways that preferentially exchange either pyruvate or acetyl-CoA (Supplementary Table 7).

Finally, the considerable contribution of weighted-type descriptors emphasize the importance of quantification of pathway cross-talk. Descriptors considering the strength of the connections between pathways are more predictive of the phylogenetic distance than their non-weighted version (where the number of metabolites shared by pathways is ignored). This could explain the advantage of keeping ubiquitous metabolites, which add information about the amount of metabolites pathways exchange.

4 CONCLUSIONS

To address the relationship between metabolic and phylogenetic information, we developed and used an abstract representation of metabolic reactions called Network of Interacting Pathways or NIP, together with an extensive set of descriptors of the structure and complexity of networks. We demonstrated that networks of metabolic reactions, as well as their simplified pathway-based representation, contain enough information to accurately predict phylogenetic distances among species. The full knowledge of all metabolic reactions involved is not required, and can advantageously be replaced by the knowledge of which pathways are present and which pathways overlap. Ubiquitous metabolites, usually ignored, are shown to slightly improve the reconstructions.

The success of our approach reveals that the organization of metabolic networks reflects, i.e. encodes, the phylogeny of the corresponding species. Evolution not only leaves its footprint on gene and protein sequences, but also in the fine wiring of functional modules—here, metabolic pathways. However, as shown by the few discrepancies observed between the reference phylogeny and the phylogeny reconstructed from metabolic networks, not all of the mutations leading to or following speciation lead to modifications in the structure and complexity of metabolic networks.

Using machine learning approaches we have been able, for the first time, to identify the most important features of pathway organization that best encode the phylogeny of species: scale-freeness, small-worldness, high average clustering coefficient and the presence of a core of densely overlapping pathways. Our results suggest that the efficient functioning of the living cells depends very strongly on fine details of the cross-talk among functional modules, which might be considered as an organizational principle of complex networks. While most approaches to identify functional modules in metabolic networks are based on the hypothesis that metabolic reactions are significantly denser within modules than across modules (Guimera and Nunes Amaral, 2005; Holme *et al.*, 2003; Kreimer *et al.*, 2008), our results suggest that connections between modules are very dense themselves, and of subtle complexity.

Compacting up to 11-fold the information contained in metabolic networks, NIPs represent a higher hierarchical level of the metabolic system that appears to encode essential evolutionary information and permits highly accurate quantitative predictions. Among the possible applications of the NIP representation, we are evaluating its use as a standard to assess network modularization approaches, and to explore the major differences in the organization of metabolic networks between major taxonomic groups.

ACKNOWLEDGEMENTS

We are grateful to Dr C. Turbeville, Dr J. Alves and Dr M. Rivera (VCU, Richmond) for useful discussions concerning phylogeny and phylogenetic tree reconstruction. We also thank other members of the Buck laboratory and the Center for the Study of Biological Complexity for useful support and discussions. Finally, we thank the anonymous referees for their constructive comments, which contributed for the better presentation of our study.

Funding: National Institutes of Health (grants R01AI050196, R01AI055347 and U54AI057168 to G.A.B.,PI); European union (grant LSHG-CT-2006-037469 to B.S.,PI).

Conflict of Interest: none declared.

REFERENCES

Aittokallio,T. and Schwikowski,B. (2006) Graph-based methods for analysing networks in cell biology. *Brief. Bioinform.*, **7**, 243–255.

Barabasi,A. and Albert,R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

Benson,D.A. *et al*. (2006) GenBank. *Nucleic Acids Res.*, **34**(Database issue), D16–D20.

Bonchev,D. and Buck,G. (2005) Quantitative measures of network complexity. In Bonchev,D. and Rouvray,D. (eds), *Complexity in Chemistry, Biology, and Ecology*. Springer, New York, 191–235.

Chenna,R. *et al*. (2003) Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.

Clemente,J.C. *et al*. (2007) Phylogenetic reconstruction from non-genomic data. *Bioinformatics*, **23**, e110–e115.

Felsenstein,J. (1989) PHYLIP - phylogeny inference package (version 3.2). *Cladistics*, **5**, 164–166.

Fisher,R. and Yates,F. (1938) *Statistical Tables for Biological, Agricultural and Medical Research*. 3rd edn. Oliver and Boyd, London, pp. 26–27.

Forst,C.V. and Schulten,K. (2001) Phylogenetic analysis of metabolic pathways. *J. Mol. Evol.*, **52**, 471–489.

Forst,C.V. *et al*. (2006) Algebraic comparison of metabolic networks, phylogenetic inference, and metabolic innovation. *BMC Bioinformatics*, **7**, 67.

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

Guyon,I. and Elisseeff,A. (2003) An introduction to variable and feature selection. *J. Mach. Learn. Res.*, **3**, 1157–1182.

Hall,M. and Holmes,G. (2003) Benchmarking attribute selection techniques for discrete class data mining. *IEEE Trans. Knowl. Data Eng.*, **15**, 1437–1447.

Harary,F. (1969) *Graph Theory*. Addison-Wesley, Reading, MA.

Hartwell,L.H. *et al.* (1999) From molecular to modular cell biology. *Nature*, **402** (Suppl. 6761), C47–C52.

Heymans,M. and Singh,A.K. (2003) Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics*, **19** (Suppl. 1), i138–i146.

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

Hong,S.H. *et al*. (2004) Phylogenetic analysis based on genome-scale metabolic pathway reaction content. *Appl. Microbiol. Biotechnol.*, **65**, 203–210.

Huynen,M.A. and Bork,P. (1998) Measuring genome evolution. *Proc. Natl Acad. Sci. USA*, **95**, 5849–5856.

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

Kreimer,A. *et al*. (2008) The evolution of modularity in bacterial metabolic networks. *Proc. Natl Acad. Sci. USA*, **105**, 6976–6981.

Liao,L. *et al*. (2002) Genome comparisons based on profiles of metabolic pathways. In *Proceedings of the 6th International Conference on Knowledge-Based Intelligent Information and Engineering Systems*. Crema, Italy.

Liu,W. *et al*. (2007) A network perspective on the topological importance of enzymes and their phylogenetic conservation. *BMC Bioinformatics*, **8**, 121.

Ma,H.-W. 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.

Ma,H.-W. and Zeng,A.-P. (2004) Phylogenetic comparison of metabolic capacities of organisms at genome level. *Mol. Phylogenet. Evol.*, **31**, 204–213.

Oh,S. *et al*. (2006) Construction of phylogenetic trees by kernel-based comparative analysis of metabolic networks. *BMC Bioinformatics*, **7**, 284.

Olsen,G.J. *et al.* (1994) The winds of (evolutionary) change: breathing new life into microbiology. *J. Bacteriol.*, **176**, 1–6.

Papin,J.A. *et al.* (2004) Hierarchical thinking in network biology: the unbiased modularization of biochemical networks. *Trends Biochem. Sci.*, **29**, 641–647.

Paradis,E. *et al.* (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Penny,D. and Hendy,M. (1985) The use of tree comparison metrics. *Syst. Zool.*, **34**, 75–82.

Posada,D. and Crandall,K.A. (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.

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

Shasha,D. *et al.* (2004) Unordered tree mining with applications to phylogeny. In *ICDE '04: Proceedings of the 20th International Conference on Data Engineering*. IEEE Computer Society, Washington, DC, USA, p. 708.

Spirin,V. *et al.* (2006) A metabolic network in the evolutionary context: multiscale structure and modularity. *Proc. Natl Acad. Sci. USA*, **103**, 8774–8779.

Swofford,D. (2003) *PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods), version 4.0b 10*. Sinauer Associates Inc., Sunderland, MA.

Watts,D.J. and Strogatz, S.H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–442.

Weaver,W. and Shannon,C. (1949) *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Illinois. (republished in paperback 1963).

Witten,I.H. and Frank,E. (1999) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, San Francisco.

Wuyts,J. *et al.* (2004) The European ribosomal RNA database. *Nucleic Acids Res.*, **32**(Database issue), D101–D103.

Zhang,Y. *et al.* (2006) Phylophenetic properties of metabolic pathway topologies as revealed by global analysis. *BMC Bioinformatics*, **7**, 252.