Algebraic Comparison of Metabolic Networks, Phylogenetic Inference, and Metabolic Innovation

Christian V. Forst [a,*], Christoph Flamm [c], Ivo L. Hofacker [c], and Peter F. Stadler [b,c,d]

[a] *Bioscience Division, Los Alamos National Laboratory, Mailstop M888, Los Alamos, NM 87545, USA*

[b] *Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*

[c] *Department of Theoretical Chemistry University of Vienna, Währingerstraße 17, A-1090 Wien, Austria*

[d] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*

[*] Corresponding author:
Christian V. Forst, Bioscience Division, Los Alamos National Laboratory, Mailstop M888, Los Alamos, NM 87545, USA

Tel.: +1 (505) 665-5268, FAX: +1 (505) 665-3024, E-Mail: `chris @ lanl.gov`

Abstract

Metabolic networks are naturally represented as directed hypergraphs in such a way that metabolites are nodes and enzyme-catalyzed reactions form (hyper)edges. The familiar operations from set algebra (union, intersection, and difference) form a natural basis for both the pairwise comparison of networks and identification of distinct metabolic features of a set of algorithms. We report here on an implementation of this approach and its application to the procaryotes. We demonstrate that metabolic networks contain valuable phylogenetic information by comparing phylogenies obtained from network comparisons with 16S RNA phylogenies. We then used the same software to study metabolic innovations in two sets of organisms, free living microbes and *Pyrococci*, as well as obligate intracellular pathogens.

Key words: Set algebra, metabolic networks, phylogeny, *Pyrococci*, intracellular pathogens

Manuscript *24 June 2005*

1 Introduction

The metabolic networks of a wide variety of organisms, in particular procaryotes, have been reconstructed by means of a combination of genomic annotations with biochemical and physiological data, see e.g. (Becker & Palsson, 2005). These networks are compiled in databases, in particular in the `KEGG` resource (Kanehisa *et al.*, 2004).

Large scale bacterial phylogenies that are based on single genes are notoriously plagued by gene transfer, gene duplication, gene deletion, and functional replacement of genes. The same holds for various approaches towards utilizing gene content for phylogenetic purposes, discussed e.g. by Fitz-Gibbon & House (1999); Ma & Zeng (2004); Snel *et al.* (1999, 2002); Wolf *et al.* (2001); Yang *et al.* (2005). A recent article by Hong *et al.* (2004) addressed this issue by considering the presence or absence of 64 individual subpathways that were identified based on the COG division (Tatusov *et al.*, 1997) of the National Center for Biotechnology Information. A related approach, based on comparisons of individual pathways was discussed by discussed Dandekar *et al.* (1999); Forst & Schulten (1999, 2001) and Heymans & Singh (2003). The pathways necessary for such approaches can be derived from a given metabolic network either "by hand" or using automated procedures such as metabolic flux analysis, see e.g. (Schilling & Palsson, 1998; Schilling *et al.*, 2000; Schuster *et al.*, 2000; Xiong *et al.*, 2004; Gagneur & Klamt, 2004).

Instead of attempting to first reconstruct individual pathways, we take here a more global view by grounding our analysis in the direct comparison of the metabolic networks. While the application of generic graph distances or similarity measures, see e.g. (Bunke & Shearer, 1997) is certainly appealing, they cannot be used in a straightforward manner for metabolic networks. The reason is that chemical reaction networks do not have a simple representation as graphs, at least not when metabolites are represented as nodes and reactions as edges. Instead, a metabolic network is naturally described by a directed hypergraph (Zeigarnik, 2000), or, equivalently, by a directed bipartite graph, in which metabolites and reactions (or, equivalently, the enzymes that catalyze the reactions), are represented by two different types of vertices. More global comparison of metabolic networks, in terms of various network indices and networks motifs, can be found in Zhu & Qin (2005).

This contribution is organized as follows: In the next section we summarize an algebraic approach to comparisons and manipulations of chemical reaction networks that is motivated by set theory. We briefly describe the `C` library that implements this approach. We then demonstrate that the symmetric difference of two metabolic networks can be used to derive a distance measure that is suitable for reconstructing phylogenetic relationship from metabolic

network data. More interestingly, however, the same approach can be used directly to extract those subnetworks of the metabolism that are innovations in the particular subtree of the phylogeny. We illustrate our approach using pathogenic procaryotes as an example.

2 The Algebra of Directed Hypergraphs

A metabolic network is defined by its metabolites and the system of reactions that inter-convert them. We denote the set of metabolites by X. A chemical reaction can be described as a pair of multisets (E^-, E^+), where $E^- \subseteq X$ is the set of educts in the reaction and $E^+ \subseteq X$ is the set of reaction products. Slightly more generally, we can replace the multisets by an ordinary sets and instead define the multiplicities of product and educt metabolites by means of the stoichiometric coefficients $n^+_{x,E}$ and $n^-_{x,E}$ of the products and educts, respectively. A metabolic network is thus a pair $(X, \mathcal{E})$ where $\mathcal{E}$ is a set of reactions. Such a structure is known as a directed hypergraph $\mathfrak{M}(X, \mathcal{E})$, see e.g. (Zeigarnik, 2000). The *stoichiometric matrix* $\mathbf{S}$ of the network has the entries

$$\mathbf{S}_{xE} = n^+_{x,E} - n^-_{x,E} \tag{1}$$

For completeness, we remark that the set $E^c = E^+ \cap E^-$ are the catalysts of the reaction E. Furthermore, a reaction is autocatalytic if $n^+_{x,E} - n^-_{x,E} \neq 0$ for some $x \in E^c$. By abuse of notation we write $E = E^+ \cup E^-$ for the set of metabolites involved in the reaction E. Furthermore, we write $\mathrm{supp}\mathcal{E} = \bigcup\{E | E \in \mathcal{E}\}$ for the set metabolites that actually take part in the reactions. We call a network $\mathfrak{M}(X, \mathcal{E})$ *clean* if $X = \mathrm{supp}\mathcal{E}$ and define the *clean up operator* as $\lfloor \mathfrak{M} \rfloor = (\mathrm{supp}\mathcal{E}, \mathcal{E})$. Furthermore, for a given set $\mathcal{E}$ of reactions and set A metabolites we define

$$\mathcal{E}[A] = \left\{ E \in \mathcal{E} | (E^+ \cup E^-) \subseteq A \right\} \tag{2}$$

The restriction of a network $\mathfrak{M}(X, E)$ to a set A of metabolites is defined as the clean network

$$\mathfrak{M}[A] = \lfloor\, (A, \mathcal{E}[A]) \,\rfloor \ . \tag{3}$$

For short we write $\mathfrak{M}[\mathcal{E}] = \mathfrak{M}[\mathrm{supp}\mathcal{E}]$ for the restriction with respect to a set of reactions. The number of reactions in a network $\mathfrak{M}$ will be denoted by $\|\mathfrak{M}\|$.

In order to compare networks in a systematic way, we need to be able to determine the differences and the commonalities of two networks. In the following

let $\mathfrak{M}'(X', \mathcal{E}')$ and $\mathfrak{M}''(X'', \mathcal{E}'')$ be two networks. Of course we have $\mathfrak{M}' = \mathfrak{M}''$ iff $X' = X''$ and $\mathcal{E}' = \mathcal{E}''$. The empty network will be denotes by $\varnothing$.

Union. The union $\mathfrak{M} = \mathfrak{M}' \cup \mathfrak{M}''$ is defined as the network $(X' \cup X'', \mathcal{E}' \cup \mathcal{E}'')$. Note that $\mathfrak{M}$ is clean if both $\mathfrak{M}'$ and $\mathfrak{M}''$ are clean.

Intersection. The intersection $\mathfrak{M} = \mathfrak{M}' \cap \mathfrak{M}''$ is defined as the clean network

$$\mathfrak{M} = \lfloor (X' \cap X'', \mathcal{E}' \cap \mathcal{E}'') \rfloor \tag{4}$$

Note that $(\mathcal{E}' \cap \mathcal{E}'')[X' \cap X''] = \mathcal{E}' \cap \mathcal{E}''$.

Difference. The *difference* $\mathfrak{M} = \mathfrak{M}' \setminus \mathfrak{M}''$ is defined as the clean network

$$\mathfrak{M} = \lfloor (\mathrm{supp}(\mathcal{E}' \setminus \mathcal{E}''), \mathcal{E}' \setminus \mathcal{E}'') \rfloor \tag{5}$$

The difference Network contains all reactions occurring in $\mathfrak{M}'$ but not $\mathfrak{M}''$, and all metabolites occurring in the remaining reactions.

The *strict difference* $\mathfrak{M} = \mathfrak{M}' \backslash\backslash \mathfrak{M}''$ is the clean network

$$\mathfrak{M} = \lfloor (X' \setminus X'', (\mathcal{E}' \setminus \mathcal{E}'')[X' \setminus X'']) \rfloor \tag{6}$$

The new network contains only those metabolites occurring in $\mathfrak{M}'$ but not $\mathfrak{M}''$, and only those reactions from $\mathfrak{M}'$ that can be performed with the remaining metabolites. Thus, we have $\|\mathfrak{M}' \backslash\backslash \mathfrak{M}''\| \leq \|\mathfrak{M}' \setminus \mathfrak{M}''\|$.

Symmetric difference. The *symmetric difference* $\mathfrak{M} = \mathfrak{M}' \triangle \mathfrak{M}''$ is defined as the clean network $\mathfrak{M} = \lfloor (\mathfrak{M}' \cup \mathfrak{M}'') \setminus (\mathfrak{M}' \cap \mathfrak{M}'') \rfloor$.

Strict symmetric difference. The *strict symmetric difference* $\mathfrak{M} = \mathfrak{M}' \diamondsuit \mathfrak{M}''$ is $\mathfrak{M} = \lfloor (\mathfrak{M}' \cup \mathfrak{M}'') \backslash\backslash (\mathfrak{M}' \cap \mathfrak{M}'') \rfloor$.

The Vienna Reaction Network Library `Vienna-RNL` implements these basic set-theoretic operations on chemical reaction networks[1]. It provides basic ANSI C data structures for chemical reactions and their networks, IO routines for reading and writing and various formats, as well as set operations such as the union, intersection, or difference of two chemical reaction networks. It is intended for the use in conjunction with the user's own C programs or PERL scripts.

An extension of the IO-Routines for reading and writing SBML (Hucka *et al.*, 2004), an XML based dialect for the standardized representation of systems

[1] `http://www.tbi.univie.ac.at/software/Vienna-RNL/`

biology models, is currently being implemented. The capability of reading and writing SBML will make the functionality of the Vienna Reaction Network Library accessible to about 80 other software systems [2] which support SBML.

3 Phylogenies from Networks

Datasets were retrieved from the `KEGG` database on metabolic networks (Kanehisa *et al.*, 2004), which holds genomic and network data of about 20 Archaea, 200 Bacteria and 20 Eucarya, where in particular the data of many Eukaryotes are incomplete. In a preparatory step we decomposed the individual `KEGG`-pathways into their chemical reactions and combined these to a complete network for each organism.

The simplest approach to inferring phylogenetic relationships form metabolic networks is to use a distance measure d on the set of reaction networks. We use here

$$d(\mathfrak{M}', \mathfrak{M}'') = \frac{\|\mathfrak{M}' \triangle \mathfrak{M}''\|}{\|\mathfrak{M}'\| + \|\mathfrak{M}''\| - \|\mathfrak{M}' \cap \mathfrak{M}''\|} = \frac{\|\mathfrak{M}' \triangle \mathfrak{M}''\|}{\|\mathfrak{M}' \cup \mathfrak{M}''\|} \tag{7}$$

Alternatively, the strong symmetric difference $\mathfrak{M}' \diamondsuit \mathfrak{M}''$ could be used to define a difference measure. Furthermore, other normalizations of the difference measure could be used. We have observed, however, that equ.(7) performs best with respect to reproducing trusted 16S RNA phylogenies. Distance-based network phylogenies are computed using the `Fitch` algorithm (Fitch & Margoliash, 1967) implemented in the `phylip` package (Felsenstein, 1996) as well as using the splits-decomposition algorithm from the `SplitsTree` package (Huson, 1998).

An example comprising a selection of bacterial and archaeal metabolic networks is shown in Fig. 1. The phylogeny inferred from the metabolic networks conforms almost perfectly with the neighbor-joining tree computed from the 16S rRNA sequences of the same organisms. The rRNA sequences were aligned using `clustalx`. The minor discrepancies are due to poorly resolved nodes as can be seen in the split-decomposition network below.

The use of distance measures reduces the available information on the network structure already in the first step. We therefore complement distance based phylogenetic analysis with parsimony methods using reaction content: For a given set of organisms, we calculated the union of all networks as maximal network. For each organism we then constructed a reaction profile reflecting

[2] `http://www.sbml.org/`

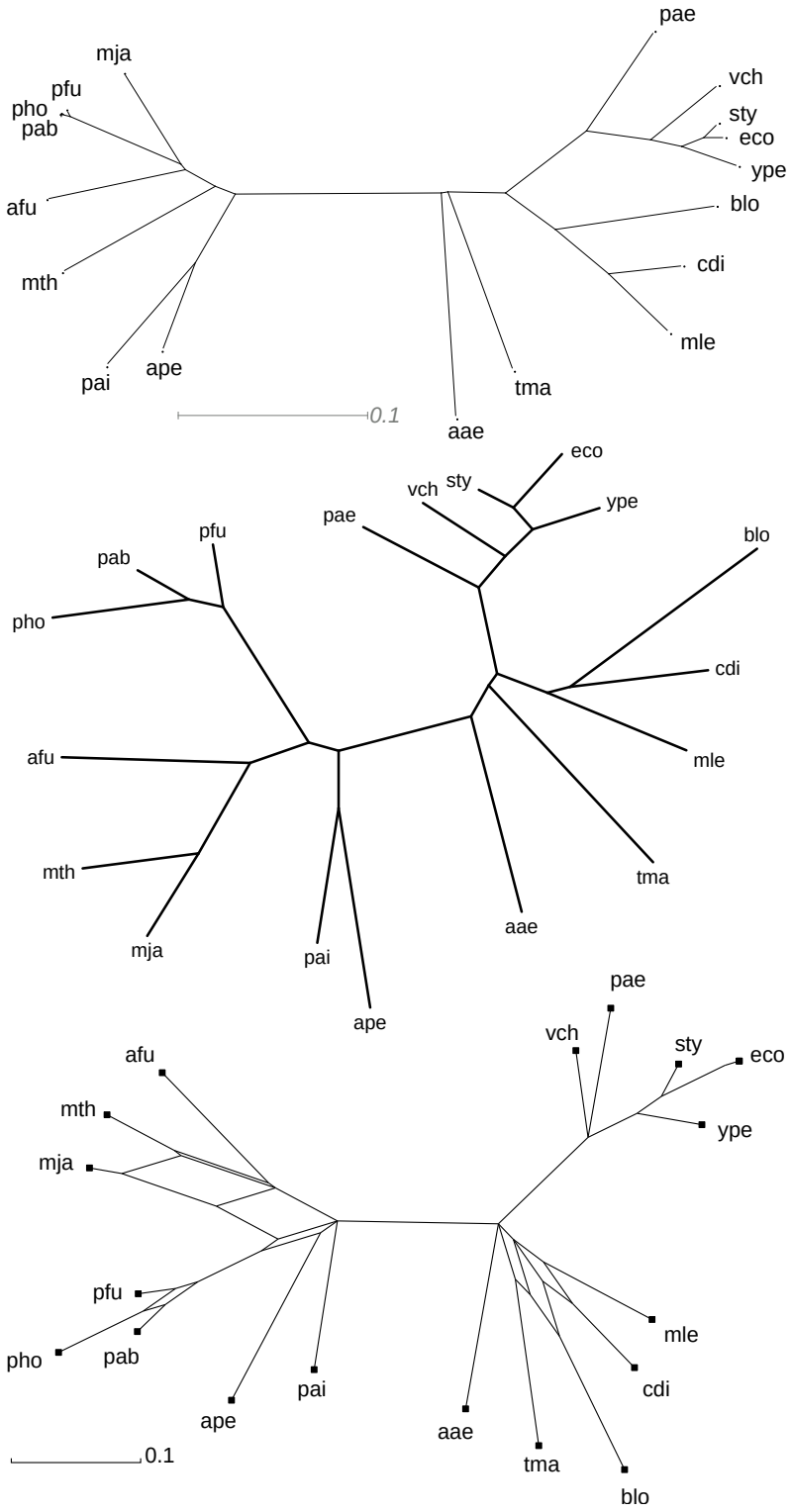

Fig. 1. Unrooted phylogenies. (top) Neighbor-joining tree of 16S rRNA sequences. (center) Phylogenetic tree calculated from metabolic network data using the Fitch algorithm for distance matrices. (bottom) Phylogenetic tree calculated from metabolic network data using Splits decomposition with the Fitch-Margoliash power 2 fit for distance matrices. Species abbreviations are collected in Table 1 in the Appendix.

presence or absence of a particular reaction in the metabolic network of the respective organism. This approach thus is reduced to reconstructing phylogenies from character-tables that represent the presence/absence of particular reactions in the reaction network. It should be noted that this is similar, but not quite the same, as using the presence or absence of orthologous enzymes (see e.g. (Fitz-Gibbon & House, 1999; Ma & Zeng, 2004; Snel *et al.*, 1999, 2002; Wolf *et al.*, 2001; Yang *et al.*, 2005)). The main difference is that the network based approach tolerates horizontal gene transfer and functional replacements (Hong *et al.*, 2004).

4 Metabolic Innovations

The algebraic approach to metabolic network evolution can also be used in a straightforward way to trace the history of metabolic innovations. To this end, consider a (trusted) unrooted phylogenetic tree T in which each leaf of T is labeled with the metabolic network $\mathfrak{M}_k$ of the corresponding taxon k. Each edge e of T defines a split, i.e., a bipartition $\sigma_e = \{U_e, \bar{U}_e\}$ of the set of taxa. Here we regard splits as directed. Note that mathematically we can define innovations at each split in both directions. One of the two subsets U or $\bar{U}$, however, contains the ancestral state, hence only one direction makes biological sense: this is the one where the ancestral state (root of the tree) is located in the sub-set $\bar{U}$. This knowledge has to be provided externally.

Consider an (arbitrary) directed split $\sigma = (U, \bar{U})$ on the given set of taxa, i.e., a pair of sets of taxa $(U, \bar{U})$ such that $U \neq \emptyset$, $\bar{U} \neq \emptyset$, and $U \cap \bar{U} = \emptyset$. We define the differential metabolic network

$$\mathfrak{D}(\sigma) = \left(\bigcup_{k \in U} \mathfrak{M}_k \right) \setminus \left(\bigcup_{k \in \bar{U}} \mathfrak{M}_k \right) \tag{8}$$

The network $\mathfrak{D}(\sigma)$ describes the *metabolic innovations* in U relative to the "background" $\bar{U}$.

As discussed in the previous section, network phylogenies are rather sensitive with respect to life-style and environmental constraints. The organisms whose metabolic networks have been utilized to compute the 16s rRNA tree shown in Fig. 1 are capable to freely live in the environment with a reasonably large capacity for adaptation.

As a first example we analyzed the unique metabolic network from the *Pyrococcus* genus. Figure 2 shows the network phylogeny from Fig. 1 with the *Pyrococcus spp.* clade highlighted. The resulting differential network indicates reactions present in *Pyrococcus spp.* but absent in all other organisms

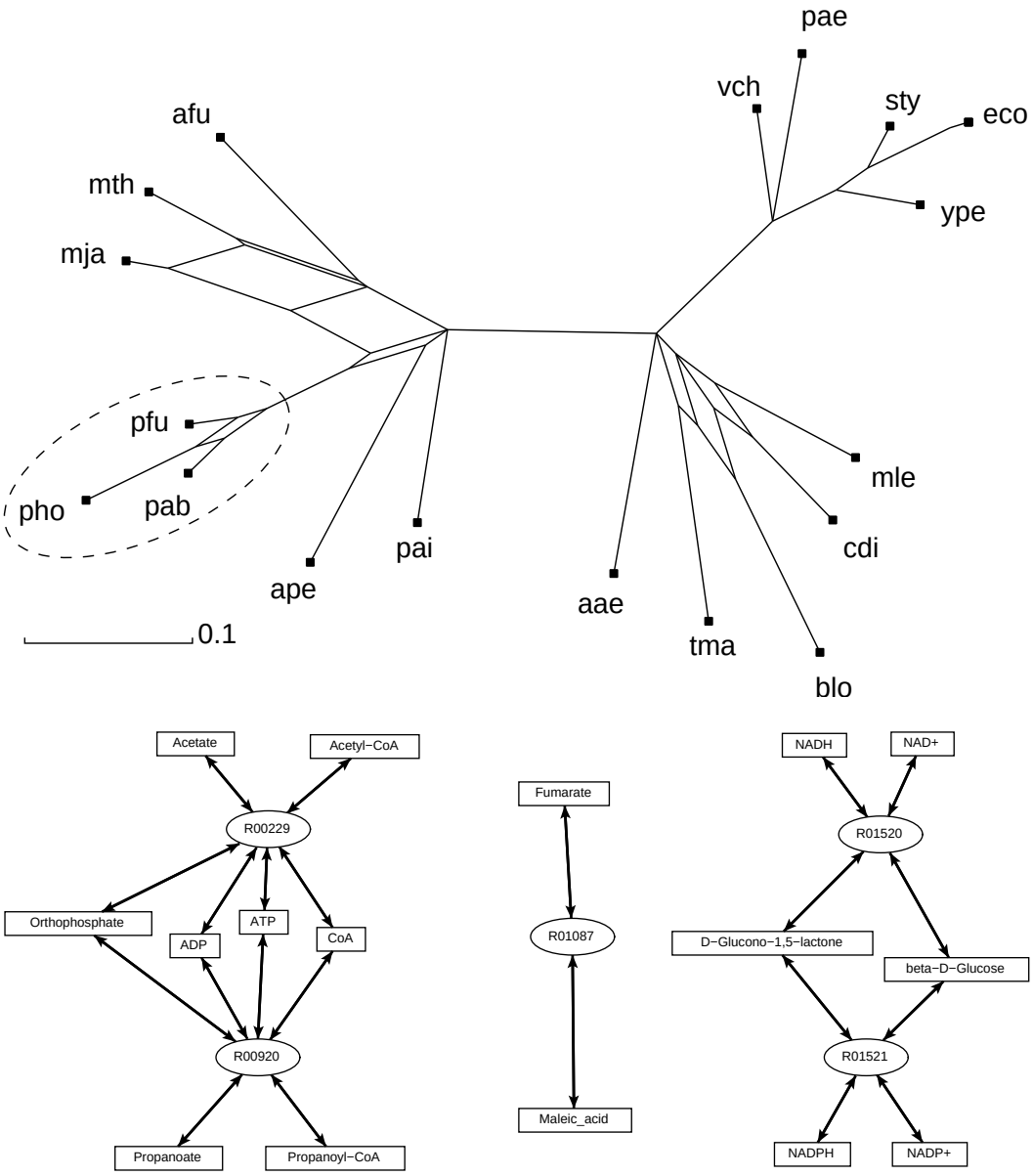

Fig. 2. (top) The *Pyrococcus spp.* clade has been selected (dashed oval) for differential network analysis. (bottom) Differential metabolic network. Numbers in the ovals refer to reaction ids in the KEGG database.

of the phylogeny (Figure 2). For example, reaction $R01087$ is catalyzed by Maleate cis-trans-isomerase which is utilized in maleate assimilating and high-temperature bacteria. A second sub-network involving both ADP-forming acetate and propanoate CoA ligases is potentially used in the organisms to convert between acetate and propanoate and their corresponding CoA forms.

As a second example we analyzed our set of reference organism (Figure 1) with obligatory intracellular pathogens. Figure 3 shows the phylogeny with the selected pathogens (dashed oval). Interestingly, *Mollicutes*, such as *My-*

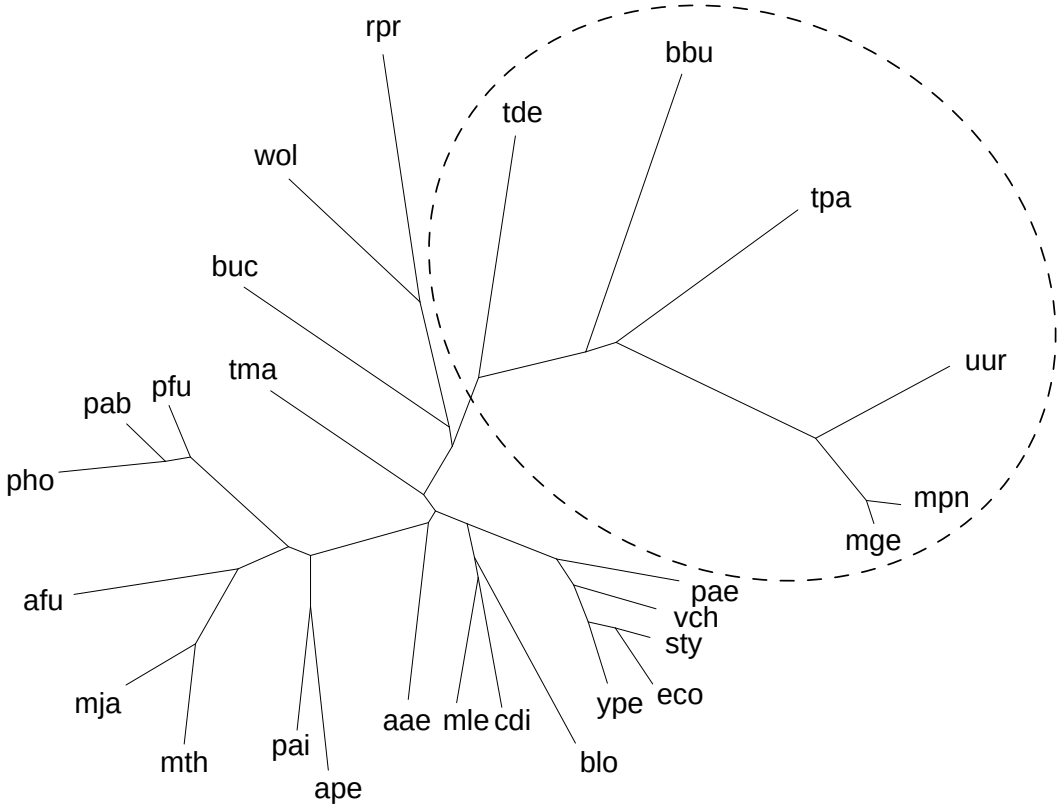

Fig. 3. Unrooted network phylogeny using PHYLIP with the Fitch-Margoliash algorithm. A set of obligatory intracellular pathogens has been selected (dashed oval) for differential network analysis (see text).

coplasmae and *Spirochaetes*, such as *Treponema* are grouped together. They all possess a minimal gene-set, and thus a highly optimized and host-dependent metabolic network. Surprisingly, this set of organisms has specific reactions that are absent in the remaining organisms of the phylogeny. Figure 4 shows the corresponding differential network which consist of five sub-network. The two largest network involves sugar-conversions and parts of glycolysis. Smaller networks correspond to formylation of tetrahydrofolate as well as cholin and carnitine pathways.

5 Discussion

The Vienna Reaction Network Library introduced above treats chemical reaction networks, and metabolic networks in particular, as directed hypergraphs. A framework borrowed from set algebra provides natural definitions of unions, intersections, and differences that can be used to compare the metabolic networks of difference organisms. We have demonstrated that metabolic networks convey phylogenetic information and can indeed be used to infer phylogenetic

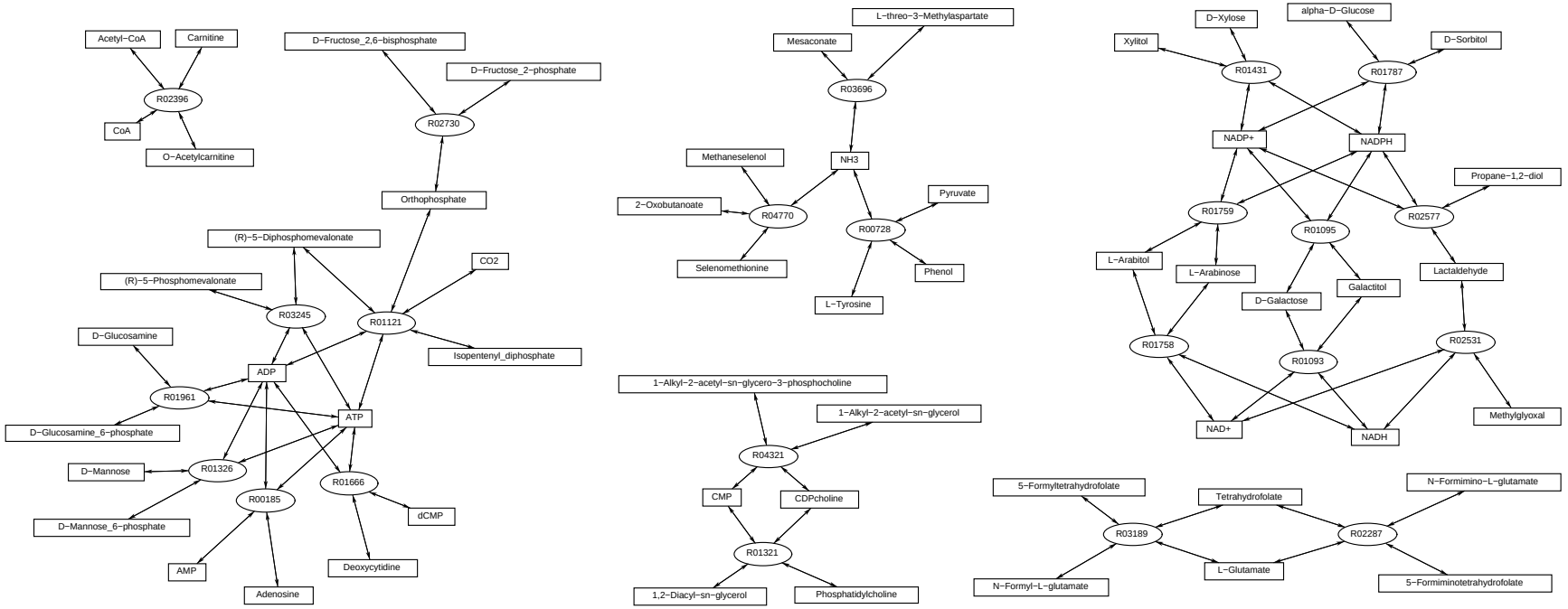

Fig. 4. Differential network corresponding to split shown in Figure 3. These reactions are specializations of the intracellular parasites.

relationships of free-living organisms in a way that is similar to gene-content based approaches. In contrast to the latter, however, metabolic network based phylogenies are less sensitive to the effects of horizontal gene transfer and functional replacement.

Differences of metabolic networks among subtrees of a trusted phylogeny, or more generally, along any split of interest in a set of organisms can be computed directly, making it easy to study metabolic innovations in particular clades. A first application of our network phylogeny analysis involved three members of the *Pyrococcus spp.* clade. The metabolic reactions resulting from the split between the *Pyrococci* and the remaining organisms involve the maleate cis-trans-isomerase reaction, ADP-forming acetate and propanoate CoA ligase reactions as well as beta-D-Glucose:NAD(P)+ 1-oxoreductase.

Our second example considers a class of intra-cellular pathogens that includes *Mycoplasmae*, *Ureaplasmae*, and *Spirochete*. Their restricted repertoire of metabolic reactions reflect the specialized life-style. Many metabolic pathways are not required in such a rich environment and have been lost in the course of evolution. On the other hand, constructing a network phylogeny including these microbes, we observe metabolic reactions assembling an unconnected network that is present in this set of intracellular pathogens and absent in remaining organism. Such reactions include phosphorilization and conversions of sugars and derivates, deaminating lyase reactions, and reactions involving carnitine, choline and tetrahydrofolate.

At present, metabolic network data are compiled by a multitude of methods, and at least in part are constructed by genomic similarity with other organisms. Strictly speaking, therefore, we cannot view metabolic network data such as those complied in the KEGG database as independent from genomic data. With the recent advances of experimental techniques in metabolomics (see e.g. (Brown *et al.*, 2005; Griffin, 2004; Sumner *et al.*, 2003)), however, the situation is rapidly improving.

Our comparative approach to metabolic network analysis, which focuses on individual reactions rather than on aggregate feature such as pathways, simplifies the identification of metabolic innovations and, in particular, facilitates the recognition of organisms as potential biological threat agents based on their metabolic repertoire. Furthermore, the ability to easily identify differences in metabolic capacity between pathogens should be useful towards a refined classification of pathogenicity based on metabolic capabilities.

In this contribution we have restricted ourselves to unweighted networks. Distance measures between networks, however, could be refined by attaching weights to both vertices and (hyper-)edges without requiring significant algorithmic changes. These could reflect, e.g., how essential a reaction or a

metabolite is for each organism. With the increasing amount and accuracy of available data it might also be feasible to devise a stochastic model of the evolution of metabolic networks, which could then be turned into a scoring scheme for a generalized version of (local) graph alignment along the lines of Berg & Lässig (2004).

Acknowledgments

This work was supported in part by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Project No. P15893, by the German *DFG* Bioinformatics Initiative BIZ-6/1-2, by the Austrian *Gen-AU bioinformatics integration network* sponsored by BM-BWK and BM-WA, and by the Laboratory Directed Research and Development program of the Los Alamos National Laboratory, Project No. 20040184ER.

References

Becker, S. A. & Palsson, B. O. (2005). Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus N315*: an initial draft to the two-dimensional annotation. *BMC Microbiology* **5**, 8 [epub].

Berg, J. & Lässig, M. (2004). Local graph alignment and motif search in biological networks. *Proc. Natl. Acad. Sci. USA* **101**, 14689–14694.

Brown, S. C., Kruppa, G. & Dasseux, J. L. (2005). Metabolomics applications of FT-ICR mass spectrometry. *Mass Spectrom. Rev.* **24**, 223–231.

Bunke, H. & Shearer, K. (1997). On a relation between graph edit distance and maximum common subgraph. *Pattern Rec. Let.* **18**, 689–694.

Dandekar, T., Schuster, S., Snel, B., Huynen, M. A. & Bork, P. (1999). Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochem. J.* **343**, 115–124.

Felsenstein, J. (1996). Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.* **266**, 418–427.

Fitch, W. & Margoliash, E. (1967). Construction of phylogenetic trees. *Science* **155**, 279–284.

Fitz-Gibbon, S. T. & House, C. H. (1999). Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res.* **27**, 4218–4222.

Forst, C. V. & Schulten, K. (1999). Evolution of metabolism: A new method for the comparison of metabolic pathways using genomic information. *J. Comp. Biol.* **6**, 343–360.

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

Gagneur, J. & Klamt, S. (2004). Computation of elementary modes: a unifying framework and the new binary approach. *BMC Bioinformatics* **5**, 175 [epub].

Griffin, J. L. (2004). Metabolic profiles to define the genome: can we hear the phenotypes? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **359**, 857–871.

Heymans, M. & Singh, A. K. (2003). Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics* **19**, Suppl. 1, i138–i346.

Hong, S. H., Kim, T. Y. & Lee, S. Y. (2004). Phylogenetic analysis based on genome-scale metabolic pathway reaction content. *Appl. Microbiol. Biotechnol.* **65**, 203–210.

Hucka, M., Finney, A., Bornstein, B. J., Keating, S. M., Shapiro, B. E., Matthews, J., Kovitz, B. L., J., S. M., Funahashi, A., Doyle, J. C. & Kitano, H. (2004). Evolving a lingua franca and associated software infrastructure for computational systems biology: the Systems Biology Markup Language (SBML) project. *Syst. Biol.* **1**(1), 41–52. Doi: 10.1049/sb:20045008.

Huson, D. H. (1998). `SplitsTree`: analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68–73.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucl. Acids Res.* **32**, D277–D280.

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

Schilling, C. H., Letscher, D. & Palsson, B. Ø. (2000). Theory for the systematic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.* **203**, 229–248.

Schilling, C. H. & Palsson, B. Ø. (1998). The underlying pathway structure of biochemical reaction networks. *Proc. Natl. Acad. Sci. USA* **95**, 4193–4198.

Schuster, S., Fell, D. A. & Dandekar, T. (2000). A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnol.* **18**, 326–332.

Snel, B., Bork, P. & Huynen, M. A. (1999). Genome phylogeny based on gene content. *Nat. Genet.* **21**, 108–110.

Snel, B., Bork, P. & Huynen, M. A. (2002). Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.* **12**, 17–25.

Sumner, L. W., Mendes, P. & Dixon, R. A. (2003). Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry* **62**, 817–836.

Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997). A genomic

perspective on protein families. *Science* **278**, 631–637.

Wolf, Y. I., Rogozin, I. B., Grishin, N. V., Tatusov, R. L. & Koonin, E. V. (2001). Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.* **1**, 8 [epub].

Xiong, M., Zhao, J. & Xiong, H. (2004). Network-based regulatory pathways analysis. *Bioinformatics* **20**, 2056–2066.

Yang, S., Doolittle, R. F. & Bourne, P. E. (2005). Phylogeny determined by protein domain content. *Proc. Natl. Acad. Sci. USA* **102**, 373–378.

Zeigarnik, A. V. (2000). On hypercycles and hypercircuits in hypergraphs. In: *Discrete Mathematical Chemistry* (Hansen, P., Fowler, P. W. & Zheng, M., eds.), vol. 51 of *DIMACS series in discrete mathematics and theoretical computer science*. Providence, RI: American Mathematical Society.

Zhu, D. & Qin, Z. S. (2005). Structural comparison of metabolic networks in selected single cell organisms. *BMC Bioinformatics* **6**, 8 [epub]. Doi:10.1186/1471-2105-6-8.

Table 1. Metabolic networks used in this study.

Domain			Species	KEGG Id	Genomic Sequence
Bacteria	Proteobacteria	Gamma	Escherichia coli K-12 MG1655	eco	U00096
			Buchnera aphidicola	buc	BA_000003
			Salmonella typhi CT18	sty	NC_003198
			Yersinia pestis CO92	ype	NC_003143
			Vibrio cholorae	vch	NC_002505
			Pseudomonas aeruginosa	pae	NC_002516
		Rickettsiales	Rickettsia prowazekii	rpr	NC_000963
			Wolbachia endosymbiont	wol	NC_002978
	Firmicutes	Mollicutes	Mycoplasmae genitalium	mge	L43967
			Mycoplasmae pneumoniae	mpn	NC_000912
			Ureaplasmae urealyticum	uur	NC_002162
	Spirochaetes		Borrelia burgdorferi	bbu	AE000783
			Treponema pallidum	tpa	NC_000919
			Treponema denticola	tde	NC_002967
	Actinobacteria		Mycobacterium leprae	mle	NC_002677
			Bifidobacterium longum	blo	NC_004307
			Corynebacterium diphtheriae	cdi	NC_002935
	Hyperthermophilic bacteria		Aquifex aeolicus	aae	AE000657
			Thermotoga maritima	tma	AE000512
Archaea	Euryarchaeota		Methanococcus jannaschii	mja	NC_000909
			Methanobacterium thermoautotrophicum	mth	NC_000916
			Archaeoglobus fulgidus	afu	NC_000917
			Pyrococcus horikoshii	pho	BA000001
			Pyrococcus abyssi	pab	NC_000865
			Pyrococcus furiosus	pfu	NC_003413
	Crenarchaeota		Aeropyrum pernix	ape	BA000002
			Pyrobaculum aerophilum	pai	NC_003364