Computational Biology and Chemistry 53 (2014) 242–250

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Metabolic network motifs can provide novel insights into evolution: The evolutionary origin of Eukaryotic organelles as a case study

Erin R. Shellman [a], Yu Chen [b], Xiaoxia Lin [b], Charles F. Burant [c,d,*], Santiago Schnell [d,a,*]

[a] Department of Computational Medicine & Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, USA
[b] Department of Chemical Engineering, University of Michigan School of Engineering, Ann Arbor, MI, USA
[c] Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, MI, USA
[d] Department of Molecular & Integrative Physiology, University of Michigan Medical School, Ann Arbor, MI, USA

ARTICLE INFO

Article history:
Received 1 April 2014
Received in revised form 15 September 2014
Accepted 15 September 2014
Available online 21 September 2014

Keywords:
Metabolism
Eukaryotic cell
Evolution
Phylogenies
Network motifs
Enzyme classification

ABSTRACT

Phylogenetic trees are typically constructed using genetic and genomic data, and provide robust evolutionary relationships of species from the genomic point of view. We present an application of network motif mining and analysis of metabolic pathways that when used in combination with phylogenetic trees can provide a more complete picture of evolution. By using distributions of three-node motifs as a proxy for metabolic similarity, we analyze the ancestral origin of Eukaryotic organelles from the metabolic point of view to illustrate the application of our motif mining and analysis network approach. Our analysis suggests that the hypothesis of an early proto-Eukaryote could be valid. It also suggests that a δ- or ε-Proteobacteria may have been the endosymbiotic partner that gave rise to modern mitochondria. Our evolutionary analysis needs to be extended by building metabolic network reconstructions of species from the phylum Crenarchaeota, which is considered to be a possible archaeal ancestor of the eukaryotic cell. In this paper, we also propose a methodology for constructing phylogenetic trees that incorporates metabolic network signatures to identify regions of genomically-estimated phylogenies that may be spurious. We find that results generated from our approach are consistent with a parallel phylogenetic analysis using the method of feature frequency profiles.

1. Introduction

The sequence of events in the early history of the Eukaryotic cell remain as mysterious today as they were in 1967 when Lynn Margulis described the serial endosymbiosis hypothesis (Sagan, 1967), a model of organelle evolution in which one microbe lives inside another. Debate continues regarding the origin of nuclei (Moreira and Lopez-Garcia, 1998), peroxisomes (Schlüter et al., 2006), mitochondria and even the host-cell that served as the venue for endosymbiotic events. Recent work argues that the eukaryotic ancestor has its origins in the phylum Crenarchaeota (Williams et al., 2013). Considerable progress has been made in the field of sequencing technology that has enabled geneticists and evolutionary biologists to interrogate the genomes of bacteria and mitochondria and discover commonalities between them (Andersson et al., 1998; McFadden and van Dooren, 2004). For example, a comparison of the α-proteobacterium *Rickettsia prowazekii* and *Saccharomyces cerevisiae* indicated that *R. prowazekii* was the likely ancestor of modern mitochondria based on the similarity of ribosomal RNA sequences (Andersson et al., 1998). Similarly, the genome sequence of the red alga *Cyanidioschizon merolae* supports the hypothesis that plant plastids were derived from a single endosymbiotic event (McFadden and van Dooren, 2004).

Despite the successes of sequencing technology, genomic methods are not without limitations and controversies. For instance, alignment methods make many assumptions regarding substitution rates and, more fundamentally, that homologous genes even exist between divergent species (Phillips et al., 2000). Consistency among the results of various alignment methods can be lacking both among distantly related organisms and well-studied organisms like mice (Chen and Tompa, 2010). Beyond these methodological concerns, horizontal gene transfer makes interpretation of phylogenetic trees difficult because temporal relationships and ancestry are convoluted by the repeated exchange of genetic information (Andam and Gogarten, 2011; Gogarten and Townsend, 2005). However, phylogenetic trees still

* Corresponding authors at: Department of Molecular & Integrative Physiology, University of Michigan Medical School, Ann Arbor, MI, USA. Tel.: +1 734 615 8733; fax: +1 734 615 6553.
E-mail addresses: burantc@med.umich.edu (C.F. Burant), schnells@umich.edu (S. Schnell).

http://dx.doi.org/10.1016/j.compbiolchem.2014.09.006
1476-9271/

provide a robust approach for the analysis of the evolutionary relationship between organisms (Lapierre et al., 2014).

As noted by de Duve (2007), modern methods for elucidating the evolutionary history of the Eukaryotic cell tend to focus strictly on genetic arguments and ignore other key cellular properties. This need not be the case. Today it is common to assemble genome-level metabolic networks by integrating known biochemical pathways with genomic annotation to yield networks describing functional properties of organisms (Francke et al., 2005; Oberhardt et al., 2009). Characterizing these biochemical networks using concepts borrowed from graph theory, such as network motifs, has proven to be a fruitful method to understand organism-level functional features. Network motifs are small, repeating patterns or subgraphs that are over- or under-represented in comparison to their abundance in a random graph (Milo et al., 2002; Milo et al., 2004). Eom et al. (2006) showed that the distributions of network motifs in 43 metabolic networks contained taxonomic meaning. That is, known taxonomic families could be reproduced using relative motif abundances from metabolic networks. Additionally, in past work we showed that metabolic network motifs could be characterized by their enzyme associations, suggesting that in metabolism, motif abundance is related to enzymatic functionality (Shellman et al., 2013).

In this study, we illustrate the use of metabolic network motifs to explore the evolutionary origin of six Eukaryotic organelles from the metabolic point of view. Our analysis characterizes the origin of cell organelles by comparing distributions of network motif abundances. Due to the limited source of metabolic network reconstructions, we limit our network motif analysis by comparing the distributions of Eukaryotic organelles with α-Proteobacteria and methanogenic Archaea because of their prominence in the literature (Poole and Penny, 2007b). An α-Proteobacterium is considered to be the ancestor of modern mitochondria (Andersson et al., 1998; Dyall et al., 2004; Esser et al., 2004), while methanogenic Archaea are often hypothesized to be the source of the host cell or of the nucleus (Martin and Muller, 1998; Moreira and Lopez-Garcia, 1998). Our results could be improved by building metabolic network reconstructions of species from the phylum Crenarchaeota, which is considered to be a possible archaeal ancestor of the eukaryotic cell (Williams et al., 2013). Further, we propose a new methodology for interpreting phylogenetic trees by incorporating metabolic signatures to pinpoint regions of genomically estimated phylogenies that may be spurious.

2. Materials and methods

In this work, we used 20 previously published metabolic network reconstructions plus 60 newly generated Proteobacteria and Archaeal reconstructions (more details below). The networks include representative species from six kingdoms of life and six distinct organelles. Organelles were analyzed as separate networks so that the motif distributions of individual organelles could be described. The network reconstructions were minimally processed, but several highly connected cofactors (ATP, ADP, AMP, NAD, NADH, NADP, NADPH, NH_3, CoA, H_2O and H^+) were removed from each network for clarity. Reactions associated with transports across membranes were also removed because they are not of metabolic interest, and cannot be said to belong to only one organelle.

Criteria for inclusion in this study was that the reconstruction (1) must be curated in the Systems Biology Mark-up Language (SBML) and (2) readable by the COBRA toolbox in Matlab (Becker et al., 2007). Neither COBRA nor Matlab were used for analysis, but these criteria insured that the reconstructions were adequately formatted and vetted for typographical errors. Once each reconstruction was read into Matlab, we exported relevant data as plain text files for the motif mining procedure. Specifically, we extracted the stoichiometric matrix, the reaction and metabolite names, the reversibility information of each reaction and the subsystem to which the reaction belonged (e.g., "Folate Biosynthesis," "TCA Cycle," "Salvage Pathway of ATP").

2.1. Identification of network motifs

For computational and analytical tractability, this work focused on network motifs of node-size three. All motifs were represented as substrate graphs. Substrate graphs represent associativity of nodes, rather than mechanistic relationships from elementary reactions like those of a bipartite graph. To infer reaction mechanisms from substrate graphs, we enumerated every possible enzyme catalyzed mechanism in the metabolic reaction capable of yielding each of the 13 possible motifs of node-size three following Shellman et al. (2013). For a list of the 13 possible motifs of node-size three, we refer the readers to Milo et al. (2002). For example, the network motif V-Out ( ) has two potential mechanisms: either substrate C transformed into product A and substrate C transformed into product B ($C \rightarrow A$ and $C \rightarrow B$) or substrate C transformed into products A and B ($C \rightarrow A + B$). The enzyme class EC 2.7 (transferring phosphorus-containing groups) is commonly found in metabolic mechanisms catalyzing transformations with the motif V-Out (Shellman et al., 2013). In the graphical representation, we included reversible reactions when they were present in the metabolic stoichiometric matrix. We generated network graph reconstruction as fast network motif detection (FANMOD) input files, following the FANMOD specifications. FANMOD is a software tool for fast network motif detection in graph representations of networks (Wernicke and Rasche, 2006).

2.2. Identifying enriched or suppressed motifs

FANMOD was employed to identify motifs in metabolic networks (Wernicke and Rasche, 2006). We estimated enrichment of particular motifs by comparing them to 1000 random networks of equal node and edge size. If a motif appeared more often in the metabolic network than in the random networks it was considered an enriched motif. Following motif enumeration, we calculated normalized z-scores to compare the number of motifs identified in each metabolic network with the average number in the random graphs. The z-score is computed as:

$$Z_i = N_{\mathrm{met}_i} - \frac{\mathrm{mean}(N_{\mathrm{random}_i})}{\mathrm{std}(N_{\mathrm{random}_i})}$$

where N_{met_i} is the number of occurrences of motif i in the metabolic network and N_{random_i} is the number of occurrences of motif i in a random network. The mean and standard deviation (std) of N_{random_i} were calculated using 1000 random graphs generated by FANMOD. The resultant z-scores are normalized and yield the significance profile or motif distribution:

$$\mathrm{SP}_i = \frac{Z_i}{\sqrt{\Sigma Z_i^2}}$$

Normalized z-scores range from −1 to 1 and any motif with a z-score > 0 is considered enriched. Likewise, any motif with a z-score < 0 is considered suppressed. Motifs with z-scores equal to 0 appear in the network as often as could be expected at random. To assess whether motifs were statistically significantly enriched or suppressed, we calculated the mean and standard error of the normalized z-scores for each motif using 1000 bootstrap samples and constructed 95% confidence interval (CI)s. CIs not containing