Published in final edited form as:

Comput Biol Chem. 2014 December ; 53PB: 242–250. doi:10.1016/j.compbiolchem.2014.09.006.

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,*}, and Santiago Schnell^{d,a,*}

^a Department of Computational Medicine & Bioinformatics, University of Michigan Medical School, Ann Arbor, Michigan, USA

^b Department of Chemical Engineering, University of Michigan School of Engineering, Ann Arbor, Michigan, USA

^c Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, Michigan, USA

^d Department of Molecular & Integrative Physiology, University of Michigan Medical School, Ann Arbor, Michigan, USA

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.

Keywords

metabolism; eukaryotic cell; evolution; phylogenies; network motifs; enzyme classification

*Corresponding authors. burantc@med.umich.edu. schnells@umich.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

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

As noted by de Duve (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.

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").

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.

(Shellman et al., 2013). For a list of the 13 possible motifs of node-size three, we refer the readers to Milo et al. (Milo et al., 2002). For example, the network motif V-Out (A) 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).

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 1,000 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:

$$Z_t = N_{\text{met}_t} - \frac{\text{mean}(N_{\text{random}_t})}{\text{std}(N_{\text{random}_t})}$$

where N_{met_t} is the number of occurrences of motif t in the metabolic network and N_{random_t} is the number of occurrences of motif t in a random network. The mean and standard deviation (std) of N_{random_t} were calculated using 1,000 random graphs generated by FANMOD. The resultant z-scores are normalized and yield the significance profile or motif distribution:

$$SP_t = Z_t / \sqrt{\sum Z_t^2}$$

Normalized z-scores range from -1 to 1 and any motif with a z-score greater than 0 is considered enriched. Likewise, any motif with a z-score less than 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 1,000 bootstrap samples and constructed 95% confidence interval (CI)s. CIs not containing the null, $z = 0$, were statistically significantly enriched or suppressed at $p = 0.05$.

Motif mining was conducted on the high performance computing cluster Flux (Intel Nehalem/I7 64-bit cluster with over 4016 Nehalem cores with an average of 4GB RAM/

core) at the University of Michigan Center for Advanced Computing. All data were analyzed in R (v 2.15.0) and visualized with the package “ggplot2.” Motif visualizations were created in LaTeX with the package “TikZ.”

Metabolic network reconstructions

In this work, we were interested in illustrating the use of metabolic network motif mining to study Eukaryote organelle evolution. Because there are many more fully-sequenced genomes than there are genome-level metabolic network reconstructions, we employed a previously developed bioinformatics pipeline (Chen, 2012) to generate a sample of proteobacteria and Archaeal reconstructions for analysis. Briefly, the reconstructions were generated by selecting ten representative species from each of α -, β -, δ -, ϵ -, γ -Proteobacteria, as well as ten species of methanogenic Archaea. In total, there were 11 α -, 10 β -, 10 δ -, 11 ϵ -, 13 γ -Proteobacteria and 12 methanogenic Archaeal metabolic networks. As we mentioned before, an ideal evolutionary analysis would also require metabolic network reconstructions of species from the phylum Crenarchaeota. However, network reconstructions are labor intensive to construct and validate.

Construction of phylogenetic trees and calculation of leaf-wise distances

To avoid the complications of whole-genome multiple alignment, genomic phylogenies were created using the alignment-free method of Feature Frequency Profiles (FFP) (Sims et al., 2009). This method is appropriate when creating phylogenies from whole-genomes where the amount of homology is potentially low.

FFP is a method similar to text comparison methods that are based on the premise that like-texts (or genomes) use a like-vocabulary. Since genomes do not contain “words,” FFP uses the differences in l-mer frequencies to estimate inter-species distances. The frequency of all l-mers of a particular length are assembled into a profile. The profiles are compared to the Jensen-Shannon Divergence measure to estimate the dissimilarity between genomes. The minimum l-mer length was found to be 7 for mtDNA and 11 for nuclear genomes. The upper limits were 16 and 26 for mtDNA and nuclear DNA respectively. An l-mer length of 14 was used to create the phylogenies in this work because it constitutes a value near the middle of the range.

Whole genomes from each of the 72 species were downloaded from the NCBI genome database. Consensus trees were based on 1,000 bootstrap samples, and the consensus phylogeny was generated using the “Consense” function in Phylip (v. 3.69), an open-source suite of programs for inferring and investigating phylogenies.

The metabolic phylogeny was created in R (v. 2.14.2) using the “ape” and “ade4” packages. First, Canberra distances were calculated to measure the distance between the motif profiles of each of the organisms. These distances were then used to construct a tree in Newick format.

The nodal distance metric (or leaf-wise distance) measures the pairwise distances between one leaf on a phylogenetic tree to all other leafs (Bluis and Shin, 2003). It was calculated by counting the pairwise path length from one leaf to all other leafs in the tree. This calculation

was performed for each tree, the metabolically-derived tree and the genomically-derived tree and then the difference in the distances was measured. Differences near zero indicate agreement of the relative position of a leaf in both trees, while large differences indicate disagreement of the relative position of the leaf in both trees.

Results

Prior to analysis we mined for motifs in the metabolic networks of 80 species (see, list of species in **SUPPLEMENTARY MATERIAL, TABLE 1**) and calculated normalized z-scores to compare relative motif abundances across organisms and organelles (see **Methods**). Motifs are numbered as previously presented in the literature (Milo et al., 2004) and are roughly in order of increasing edge density (**FIGURE 1**). To evaluate the statistical significance of motif enrichment, we calculated means and standard errors of the normalized z-scores using bootstrap re-sampling and constructed 95% CIs (**FIGURE 1**). Based on the previous finding that taxonomic relationships could be replicated using motif distributions (Eom et al., 2006), we expected that endosymbiotic organelles will display higher levels of similarity with their ancestor microbe compared to other, non-ancestral microbes.

Motif distributions can be used to investigate theories of organelle evolution

The original host cell, within which serial endosymbiotic events could have occurred, has been hypothesized to be a methanogenic Archaeon (Martin and Muller, 1998). This cell could be a Bacterium or an early proto-Eukaryote (Cavalier-Smith, 1989; Poole and Penny, 2007a) with many of the features of a modern Eukaryote (nucleus and membraneous structures such as the ER). We found the average motif profile of the Eukaryotic cytosol to be dissimilar compared to the profiles of both α -Proteobacteria and methanogenic Archaea (**FIGURE 1**). Seven of thirteen motifs were significantly different from Archaea, and six of thirteen were different from α -Proteobacteria. The relative dissimilarity of the cytosolic motif profiles with the motif profiles of the α -Proteobacteria and methanogenic Archaea suggests that neither are good fits as an ancestor of the host cell. The analysis of metabolic networks of other bacterial and possible archaeal ancestors could provide more clues into the origins of the cytoplasm.

The origin of the peroxisome is nebulous. Schlüter *et al.* (Schlüter et al., 2006) argued that since the peroxisomal membrane is comprised entirely of Eukaryotic proteins, an endosymbiotic origin is unlikely. Conversely, there is morphological and chemical evidence supporting the hypothesis that peroxisomes are derived from an endosymbiotic event that may have occurred before the evolution of the mitochondrion (de Duve, 1969; de Duve, 2007). We found that the motif profile of the peroxisome shared many features with that of the cytosol and the mitochondrion, suggesting that it is metabolically akin to a former free-living microbe. Specifically, under-representation of the V-Out (motif 1), Mutual In (motif 4), Mutual Out (motif 5), and Regulated Mutual (motif 9) motifs and enrichment of the V-In and 3-Chain motifs (motifs 2 and 3). Like the cytosol, the peroxisomal profile does not share many features with the profiles of the α -Proteobacteria nor the methanogenic Archaea, which suggests that these organisms are not, metabolically speaking, good candidates for the origin of the peroxisome.

The nucleus, Golgi and lysosome share two general properties in their motif profiles. First, all displayed suppression in the first five motifs. This is in contrast to the pattern of enrichment in the profiles of cytosol, mitochondrion and peroxisome which displayed enrichment of the second and third motifs. Second, rather than having several enriched motifs as in the cytosolic profile, the lysosome has four and the nucleus and Golgi each have only one. This result suggests that each of these organelles could be derived from membranous in-folding and not from endosymbiotic events. The similarity of these profiles, coupled with the fact that there is evidence that the Golgi is a product of membranous in-folding (Mowbrey and Dacks, 2009), suggests that both the nucleus and the lysosome are also membrane-derived organelles, not former endosymbionts.

Lastly, we find an apparent dissimilarity of the mitochondrial motif profile with that of the α -Proteobacteria. It has been hypothesized that modern mitochondria are the product of an endosymbiotic event with a host cell and an early α -Proteobacterium (Andersson et al., 1998; Dyall et al., 2004; Esser et al., 2004). However, the mitochondrial motif profile instead appears more like that of δ - or ϵ -Proteobacteria (FIGURE 2) rather than an α -Proteobacterium within the limited number of network reconstructions we studied.

Whole-genome phylogenies are consistent with species similarities predicted by metabolic network motif distributions

To illustrate the finding that the mitochondrial motif profile is more similar to that of a δ - or ϵ -Proteobacterial motif profile, we employed the method of feature frequency profiles (FFP) to create a phylogenetic tree from the genomes of all organisms in our dataset (FIGURE 3). The dissimilarity between mitochondria and α -Proteobacteria is also observed at the genomic level. The mtDNA of six Eukaryotes (colored in gold) falls more closely within the clades of δ - and ϵ -Proteobacteria than to α -Proteobacteria.

Metabolic phylogenies could provide an additional level of insight into ancestral relationships

Because there are sometimes inherent difficulties interpreting phylogenetic relationships, we propose a combination of genomic and metabolic profile analysis to locate regions of trees where ancestral relationships might be spurious. In our analysis, we reduced the species from 72 to 18 to include only manually curated metabolic networks (FIGURE 4). Manual curation insures that the metabolic information is of the highest quality and is the most reliable. The first panel (A) contains a phylogeny based on whole genomes and the second (B) contains a phylogeny based on comparative motif distributions in the cytosolic compartment.

The genomic phylogeny constitutes three primary clades, the first containing the only plant *A. thaliana*, the second containing both fungi, *S. cerevisiae* and *P. pastoris*, as well as the protist *C. reinhardtii* and the majority of bacteria, and the third containing *H. sapiens*, *M. musculus*, three extremophilic bacteria and both methanogenic Archaea, *M. acetivorans* and *M. barkeri*. The extremophilic bacteria *C. thermocellum*, *H. pylori* and *T. maritima* likely share more adaptive genomic properties with Archaea than with bacteria since Archaea tend to favor harsh, hot environments (Aravind et al., 1998). For example, *T. maritima* shares

approximately 24% of its genome with thermophilic Archaea (Nelson et al., 1999), suggesting that horizontal gene transfer is associated with environmental commonality rather than ancestral relatedness.

The phylogeny based on motif abundances shows a distinct topology that includes many more tightly paired clades. The kingdoms of life show high levels of congruence among one another (for instance, *H. sapiens* and *M. musculus* cluster together), suggesting that the motif distributions are capturing relevant metabolic similarity.

To assess the level of taxonomic congruence between the phylogenies, we calculated pairwise node-distance matrices by counting the path lengths separating each leaf from every other leaf in the tree (Bluis and Shin, 2003). The heatmap in **FIGURE 5** shows the differences of the pairwise distances between the leaves of each of the phylogenies in **FIGURE 4A** and **B**. The heatmap shows a good level of agreement in the leaf distances between the two phylogenies, that is, how similar the trees are to one another.

The genome-derived phylogenetic relationship between two Archaea, *M. barkeri* and *M. acetivorans*, and the bacterium, *T. maritima*, illustrate how phylogeny construction using genetics alone could have some limitations. *T. maritima* shares approximately 24% of its genome with Archaea (Nelson et al., 1999; Poole and Penny, 2007b), so it is no surprise that it appears genetically more similar to Archaea than to its Bacterial counterparts, however this genetic connection can lead to spurious evolutionary conclusions when viewed in isolation. *M. tuberculosis* has been shown to fall more closely to the Proteobacterial rather than Actinobacterial clade in previous work (Kinsella et al., 2003), a result that is reproduced here with *M. tuberculosis* sharing a clade with *E. coli*, *S. typhimurium* and *V. vulnificus*, all γ -Proteobacteria, as well as *R. etli*, an α -Proteobacterium. The unusual tree topology from the perspective of *M. tuberculosis* suggests one of two possibilities: either that *M. tuberculosis* acquired many genes from early Proteobacteria via horizontal gene transfer, or that Actinobacteria are actually more closely related to Proteobacteria than was previously thought. The metabolic phylogeny supports the latter claim, because *M. tuberculosis* again falls within a clade shared by *E. coli* and *S. typhimurium*. In this example, we illustrate how the integration of both phylogenies can clarify a seemingly spurious relationship. A more systematic analysis, including species outside the groups studied, will provide a more complete picture of the evolutionary origin of Eukaryotic organelles.

Discussion

In this work, we introduced 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. To illustrate our analysis, we studied the origin of six Eukaryotic organelles by quantifying metabolic similarity with relative motif abundances. Further, we used motif abundances to provide an additional level of functional information in the construction of phylogenetic trees.

The heterogeneity of motif distributions observed in the cytosol, mitochondrion and peroxisome corresponds to the heterogeneity of function of these organelles. For instance,

the cytosol is the venue for primary metabolism, but it is also the channel through which many signaling molecules must pass. This heterogeneity can be attributed to the fact that at one point, each of these organelles were independent microbes and relied on a collection of metabolic pathways which were imported into the host cell. Conversely, the homogeneity of motifs in the nucleus, Golgi and lysosome suggests lower levels of metabolic complexity and that the range of tasks done within them is relatively narrow. This suggests that these organelles were likely never independently functioning organisms, and thus lack many of the reactions found in microbes. Instead they could have evolved to isolate repetitive tasks, for instance recurrent glycosylation reactions in the Golgi and hydrolytic reactions in the lysosome.

Our limited analysis of the evolutionary origin of Eukaryotic organelles using metabolic network motifs illustrates that cytosolic compartment is unlike that of both α -Proteobacteria and methanogenic Archaea. This suggests that the original host cell was a proto-Eukaryote. In the **SUPPLEMENTARY MATERIAL**, we compared the cytosolic profile with that of β -, δ -, ϵ - and γ -proteobacteria. The **SUPPLEMENTARY FIGURE 1** shows that the cytosolic profile is unlike that of those bacteria. This suggests that the proto-Eukaryote could thus neither be a bacterium nor an archaean. In addition, we found that the mitochondria motif profile is different to that of the α -Proteobacterial. A phylogenetic analysis of several glycolytic enzymes in Eukaryota and Proteobacteria showed greater likeness in γ -Proteobacteria, or an ancestor of γ -Proteobacteria, than to α -Proteobacteria (Emelyanov, 2003). Our phylogeny based on the FFP of whole genomes places mtDNA more closely with the δ - and ϵ -Proteobacteria families. We discuss other insights about the evolutionary origin of Eukaryotic organelles in the **SUPPLEMENTARY MATERIAL**, which are based on our network motif analysis. It is important to emphasize that a more systematic analysis, which includes species outside the groups studied, will provide a more complete picture of the evolutionary origin of Eukaryotic organelles.

In this work, metabolic network motif distributions serve as a single line of evidence to investigate evolution, which correspond to only one cellular system. Metabolic network motif distributions can be used to investigate evolutionary processes, because they are based on biochemical functionality and not strictly on genetic information, which can be difficult to interpret and does not always imply function. Further, we illustrate that integrating genomic and metabolic information can improve inferences of evolutionary relatedness and identify regions of trees where spurious evolutionary connections may lie through the complementary use of genomic-based phylogenies.

The power of using metabolic network motif distributions depends on the quality of the network reconstructions and the depth of knowledge of the reactions on which the reconstructions are built. It also depends on the methods used to identify statistical significant subgraphs as network motifs. In the literature, the power and limitations of different network motifs identification approaches is an active area of research [see, Sandefur et al. (2013) for a recent review]. We employed the normalized z-score, which identifies subgraphs as overrepresented or underrepresented motifs if they have z-score values greater or less than zero, respectively (Milo et al., 2004). This is a mainstream approach to identifying network motifs. Normalized z-score permits a statistical comparison

of a particular motif across different networks. It requires the generation of a set of biophysico-chemical realistic randomized networks for comparison. It also requires testing the assumption of normality in the networks under investigation. If metabolic network motif distributions are going to be used to investigate evolutionary relationships, it is fundamental to increase the source of metabolic network reconstructions available in the literature. As more high-throughput data become available, it will be possible to develop more metabolic network reconstructions.

Metabolic network reconstructions are revolutionizing the emerging field of comparative metabolomics in much the same way that high-throughput sequencing technology revolutionized comparative genomics. By investigating evolution using metabolic networks, we will begin to uncover the role metabolism has played in shaping the world as we know it.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Allison Ho (University of Michigan) for reading the revised version of this manuscript. This work was partially supported by the National Institute of Diabetes and Digestive and Kidney Diseases (R01 DK079084), the Michigan Nutrition Obesity Research Center (P30 DK089503), the Michigan Regional Comprehensive Metabolomics Resource Core (U24 DK097153) the Robert C. and Veronica Atkins Foundation, the A. Alfred Taubman Research Institute, the James S. McDonnell Foundation under the 21st Century Science Initiative, Studying Complex Systems Program (Grant No. 220020223). In addition, ERS acknowledges support from the Systems and Integrative Biology Training Grant (T32 GM008322).

References

- Andam CP, Gogarten JP. Biased gene transfer in microbial evolution. *Nature Reviews Microbiology*. 2011; 9:543–555. [PubMed: 21666709]
- Andersson SGE, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UCM, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*. 1998; 396:133–140. [PubMed: 9823893]
- Aravind L, Tatusov RL, Wolf YI, Walker DR, Koonin EV. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends in Genetics*. 1998; 14:442–444. [PubMed: 9825671]
- Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO, Herrgard MJ. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nature Protocols*. 2007; 2:727–38. [PubMed: 17406635]
- Bluis J, Shin DG. Nodal distance algorithm: Calculating a phylogenetic tree comparison metric. Third IEEE Symposium on Bioinformatics and Bioengineering - Bibe 2003, Proceedings. 2003:87–94.
- Cavalier-Smith T. Molecular Phylogeny - Archaeobacteria and Archezoa. *Nature*. 1989; 339:100–101.
- Chen XY, Tompa M. Comparative assessment of methods for aligning multiple genome sequences. *Nature Biotechnology*. 2010; 28:567–U53.
- Chen Y. Metabolic network reconstruction nad modeling of microbial communities 2012 PhD thesis, University of Michigan. <http://hdl.handle.net/2027.42/96092>
- de Duve C. Evolution of the peroxisome. *Annals of the New York Academy of Sciences*. 1969; 168:369–81. [PubMed: 5270945]
- de Duve C. Essay - The origin of eukaryotes: a reappraisal. *Nature Reviews Genetics*. 2007; 8:395–403.

- Dyall SD, Brown MT, Johnson PJ. Ancient invasions: From endosymbionts to organelles. *Science*. 2004; 304:253–257. [PubMed: 15073369]
- Emelyanov VV. Mitochondrial connection to the origin of the eukaryotic cell. *European Journal of Biochemistry*. 2003; 270:1599–1618. [PubMed: 12694174]
- Eom YH, Soojin L, Jeong H. Exploring local structural organization of metabolic networks using subgraph patterns. *Journal of Theoretical Biology*. 2006; 241:823–829. [PubMed: 16504210]
- Esser C, Ahmadinejad N, Wiegand C, Rotte C, Sebastiani F, Gelius-Dietrich G, Henze K, Kretschmann E, Richly E, Leister D, Bryant D, Steel MA, Lockhart PJ, Penny D, Martin W. A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Molecular Biology and Evolution*. 2004; 21:1643–1660. [PubMed: 15155797]
- Francke C, Siezen RJ, Teusink B. Reconstructing the metabolic network of a bacterium from its genome. *Trends in Microbiology*. 2005; 13:550–558. [PubMed: 16169729]
- Gogarten JP, Townsend JP. Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology*. 2005; 3:679–687. [PubMed: 16138096]
- Kinsella RJ, Fitzpatrick DA, Creevey CJ, McInerney JO. Fatty acid biosynthesis in *Mycobacterium tuberculosis*: Lateral gene transfer, adaptive evolution, and gene duplication. *Proceedings of the National Academy of Sciences of the United States of America*. 2003; 100:10320–10325. [PubMed: 12917487]
- Lapierre P, Lasek-Nesselquist E, Gogarten JP. The impact of HGT on phylogenomic reconstruction methods. *Brief Bioinform*. 2014; 15:79–90. [PubMed: 22908214]
- Martin W, Muller M. The hydrogen hypothesis for the first eukaryote. *Nature*. 1998; 392:37–41. [PubMed: 9510246]
- McFadden GI, van Dooren GG. Evolution: Red algal genome affirms a common origin of all plastids. *Current Biology*. 2004; 14:R514–R516. [PubMed: 15242632]
- Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: Simple building blocks of complex networks. *Science*. 2002; 298:824–827. [PubMed: 12399590]
- Milo R, Itzkovitz S, Kashtan N, Levitt R, Shen-Orr S, Ayzenshtat I, Sheffer M, Alon U. Superfamilies of evolved and designed networks. *Science*. 2004; 303:1538–1542. [PubMed: 15001784]
- Moreira D, Lopez-Garcia P. Symbiosis between methanogenic archaea and delta-proteobacteria as the origin of eukaryotes: The syntrophic hypothesis. *Journal of Molecular Evolution*. 1998; 47:517–530. [PubMed: 9797402]
- Mowbray K, Dacks JB. Evolution and diversity of the Golgi body. *FEBS Letters*. 2009; 583:3738–3745. [PubMed: 19837068]
- Nelson KE, Clayton RA, Gill SR, Gwinn ML, Dodson RJ, Haft DH, Hickey EK, Peterson LD, Nelson WC, Ketchum KA, McDonald L, Utterback TR, Malek JA, Linher KD, Garrett MM, Stewart AM, Cotton MD, Pratt MS, Phillips CA, Richardson D, Heidelberg J, Sutton GG, Fleischmann RD, Eisen JA, White O, Salzberg SL, Smith HO, Venter JC, Fraser CM. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature*. 1999; 399:323–329. [PubMed: 10360571]
- Oberhardt MA, Palsson BO, Papin JA. Applications of genome-scale metabolic reconstructions. *Molecular Systems Biology*. 2009; 5:320. [PubMed: 19888215]
- Phillips A, Janies D, Wheeler W. Multiple sequence alignment in phylogenetic analysis. *Molecular Phylogenetics and Evolution*. 2000; 16:317–330. [PubMed: 10991785]
- Poole A, Penny D. Engulfed by speculation. *Nature*. 2007a; 447:913–913. [PubMed: 17581566]
- Poole AM, Penny D. Evaluating hypotheses for the origin of eukaryotes. *BioEssays*. 2007b; 29:74–84. [PubMed: 17187354]
- Sagan L. On origin of mitosing cells. *Journal of Theoretical Biology*. 1967; 14:225–274.
- Sandefur CI, Mincheva M, Schnell S. Network representations and methods for the analysis of chemical and biochemical pathways. *Mol Biosyst*. 2013; 9:2189–200. [PubMed: 23857078]
- Schlüter A, Fourcade S, Ripp R, Mandel JL, Poch O, Pujol A. The evolutionary origin of peroxisomes: An ER-peroxisome connection. *Molecular Biology and Evolution*. 2006; 23:838–845. [PubMed: 16452116]

- Shellman ER, Burant CF, Schnell S. Network motifs provide signatures that characterize metabolism. *Molecular BioSystems*. 2013; 9:352–360. [PubMed: 23287894]
- Sims GE, Jun SR, Wua GA, Kim SH. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences of the United States of America*. 2009; 106:2677–2682. [PubMed: 19188606]
- Wernicke S, Rasche F. FANMOD: a tool for fast network motif detection. *Bioinformatics*. 2006; 22:1152–1153. [PubMed: 16455747]
- Williams TA, Foster PG, Cox CJ, Embley TM. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature*. 2013; 504:231–6. [PubMed: 24336283]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Highlights

- We present a novel application of metabolic network motif that can provide insights into evolution
- We illustrate our novel application to investigate ancestral origin of Eukaryotic organelles
- Our preliminary analysis supports the hypothesis of an early proto-Eukaryote
- Our analysis also suggests that a δ - or ϵ -Poteobacteria may have been the endosymbiotic partner that gave rise to modern mitochondria
- We propose a methodology for constructing phylogenetic trees that incorporates metabolic network signatures

Figure 1.

Motif distributions of six Eukaryotic organelles (solid black) compared with the profiles of α -Proteobacteria (solid red) and methanogenic Archaea (dashed blue). Each of the 13, 3-node motifs is indicated with a name along the x -axis, as well as a small depiction of the structure of that motif. Normalized z -scores are along the y -axis and a z -score equal to 0 is indicated with a dashed line. z -scores above 0 indicate motif enrichment and those below indicate motif suppression. Non-overlapping confidence intervals indicate statistical significance at $p = 0.05$.

Figure 2.

Comparison of Mitochondrial motif distribution with δ - and ϵ -Proteobacteria. Panel (A) contains a dendrogram of the distances between the motif profiles compared to the mitochondrial profile. The mitochondrial profile clusters most closely with δ - and ϵ -Proteobacteria profiles and not the α -Proteobacteria profile. Panel (B) contains the motif distributions of the mitochondrion (solid black), δ - (solid red) and ϵ -Proteobacteria (dashed blue).

Figure 3. Genomic phylogeny created with feature frequency profiles. Each member of the proteobacterial family, Archaea and mitochondrial genome is colored and labeled on the right-hand side. Mitochondrial genomes are enclosed with dashed boxes for clarity. In general the mtDNA falls with the clades of the δ - and ϵ -Proteobacteria as well as the methanogenic Archaea.

Figure 4. Phylogenies based on (A) whole genomes and (B) metabolic significance profiles.

Figure 5.

Heatmap of the differences between pairwise leaf distances. Differences near zero are shown in black and denote regions of relative agreement of the metabolic and genomic phylogenies. The areas of high agreement are enclosed in white boxes. Large differences are shown in white and denote regions of incongruence between the two trees.