Network motifs provide signatures that characterize metabolism and produce novel insights into the evolutionary history of the Eukaryotic cell

by

Erin Rachael Shellman

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2012

Doctoral Committee:

Doctor Charles Burant, Co-Chair
Professor Santiago Schnell, Co-Chair
Professor Daniel Forger
Professor Robert Kennedy
Professor Xiaoxia Lin
Professor Leslie Satin

For my Mother, who told me that I could be whoever I wanted,

and meant it.

ACKNOWLEDGEMENTS

Thank you to the people, academic and otherwise, who made it possible for me to complete my PhD studies and research. Specifically, Conner Sandefur served as a mentor, therapist and friend throughout graduate school, and was constantly supportive during my career search. Likewise, many thanks to Ethan Obie Romero-Severson, Grant Meadors, Ian Spicknall, Meghan O'Grady Milbrath and Bryan Mayer (Collectively known as *The Edit Club*) for the Oxford commas, tense corrections, and many, many beers. I am also indebted to Tanya Salyers, Brian Beck, Dave Ziegler, and Chris Hesse for the career advice.

My research was made possible with the financial, academic and mentoring support I received from my two advisors Chuck Burant and Santiago Schnell. Thank you both for hanging in there through the ups and downs of my thesis projects. Yu Chen and Nina Lin provided 60 metabolic network reconstructions that were essential to my third chapter, and for that I am grateful. Thanks are also in order for those who contributed to the creation of this handy LaTeX template. You guys saved me a lot of stress.

Above all, I absolutely could not have completed my PhD without the love and devotion of my partner, Bryan Mayer and my mother, Caryl Carver (Mamakita). Countless nights of debugging and trouble-shooting would have been insurmountable trials if not for Bryan's patience. Likewise, my mother was always a source of emotional support and encouragement. Thank you all for your kindness and patience during my thesis work.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

LIST OF ABBREVIATIONS

CI Confidence Interval

FANMOD A tool for fast network motif detection

FFP Feature Frequency Profiles

RECON1 Human Metabolic Network Reconstruction

SBML Systems Biology Mark-up Language

SP Significance Profile

ABSTRACT

Network Motifs Provide Signatures that Characterize Metabolism

by

Erin Rachael Shellman

Chairs: Doctor Charles Burant and Professor Santiago Schnell

A motif is a small, repeated pattern that is over-represented in a network compared to its abundance in a collection of random graphs. Motifs are of chief interest in network theory and systems biology because their over-expression may determine the topological properties that give rise to dynamic behaviors in biological systems. Motifs also provide novel functional evidence that can help unravel mechanisms of molecular evolution. In this work, we analyze metabolic network motifs, where metabolites are represented by nodes and biochemical associations are represented by edges. We find that metabolic network motifs can be characterized by their enzyme class associations and therefore, their biochemical functionality. Further, we demonstrate that cellular organelles display motif distributions that can be distinct and likely reflect the organelle's distinct metabolic role in the cell. We follow this analysis by assessing the relationship between motif participation and the property of tolerance to random component failure in the *E. coli* metabolic network. We find that the metabolic network displays higher levels of failure tolerance than seen in Erdős-Rényi random graphs, and that some motifs have unique structural properties in metabolism. Finally, we apply the methodology of motif mining and analysis to assess specific hypotheses of Eukaryotic organelle evolution. Specifically, we present novel evidence suggesting that an α-proteobacterium may not have been the ancestor of modern mitochondria. We independently validate this result using phylogenetic analysis and find that mitochondrial genomes tend to fall within the same clades as δ- and ϵ-proteobacteria. Based on this validation we propose a new hypothesis that modern mitochondria are not derived from α-proteobacteria, but are instead derived from a member of the δ- or ϵ-proteobacterial families.

CHAPTER I

Introduction

Reductionism has been the predominant paradigm for accumulating scientific knowledge for centuries and is an effective framework to conduct scientific inquiries. The methodology of reducing complex systems to the interactions of their parts has been successfully applied to the biological sciences and has resulted in the discovery of individual components that, when taken together, yield massive, functioning systems [6]. Today it is commonplace to take biological measurements using microarray technology that can yield tens of thousands of data points simultaneously. As a result, it has been estimated that the amount of data that would be collected for any given molecular pathway in 2011 would equal the amount of data collected on that pathway throughout history [20]. With this deluge of data has come the realization that the majority of interesting biological problems cannot be answered by interrogating just one gene or one molecule, but instead must be approached by analyzing the functionality of large systems of interacting genes and molecules. The methods of reductionism must be redefined to meet the needs of 21^{st} century science, and in the biological sciences these new methods are often found in the field of systems biology.

1.1 Systems Biology

"...It is about putting together rather than taking apart, integration rather than reduction..."

–Denis Noble [94]

Systems biology is a framework for utilizing genome-level data to conduct predictive, hypothesis-driven science. It offers an experimental approach that is distinct from that of classical biology.

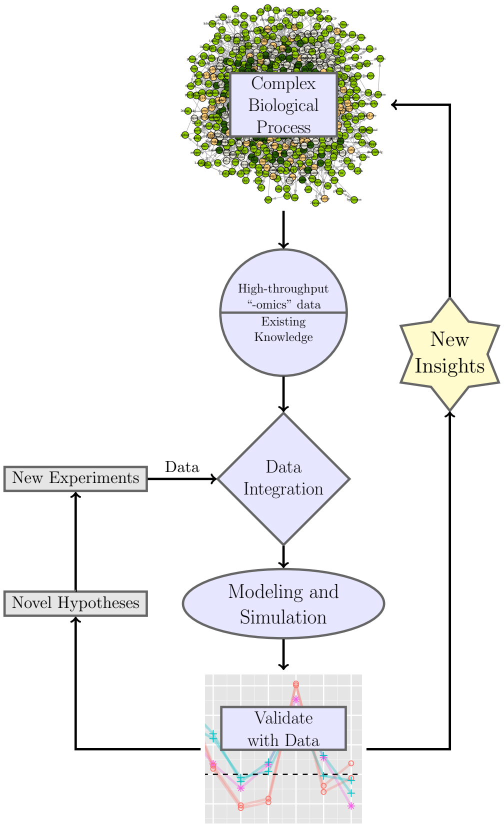


Figure 1.1: General work-flow of research in system biology. Starting with a complex biological system, existing knowledge is integrated and then modeled. Simulation results are then validated against experimental results and either recapitulate them and provide new insights, or suggest novel hypotheses and new experiments.

Rather than reducing complex systems to their simpler constituents, systems methods are top-down and can generate novel hypotheses and fill gaps in knowledge.

Rather than trying to understand a single component and working up to contextualize that component in a complex system, a systems methodology starts with a complex system and work down towards the goal of predictive modeling (Figure 1.1). Starting with a system, the next step is to integrate the network with existing experimental data, for example high-throughput data or observations from the literature. Following data integration, the systems biologist would then model and simulate the system. In the modeling phase it is important to establish clear outcomes that can be compared to data from experiments in order to validate the system model. This comparison identifies areas where the model fails to replicate experimental results and also areas where the model makes new, novel predictions. If the model is inadequate for predicting the desired outcomes it is necessary to hypothesize why, conduct additional computational and non-computational experiments, and gather data necessary to further improve and refine the model. The methodology can be repeated indefinitely until the model is able to adequately capture and predict all the behaviors of the system under investigation.

Systems methods rely on data acquired by the reductionist paradigm, and are thus not meant to replace reductionist methodologies. Instead systems biology is a framework within which to contextualize and interpret new findings at the level of whole systems.

1.1.1 An Application of the Systems Biology Work-flow

The iterative process of prediction and refinement can eventually expand to incorporate the entire system. Not surprisingly, one of the chief objectives of systems biology is the construction and simulation of a complete cell [59]. This goal was recently attained for the human pathogen *Mycoplasma genitalium* [50] and provides a stellar example of an application of the systems biological work-flow. To begin, Karr *et al.* identified the processes they were interested in modeling, all known cellular processes of *M. genitalium.* No single modeling framework is applicable to all cellular processes, so the authors divided the models of cellular processes into 28 modules that were modeled separately and combined at each time step. In the data integration stage, they manually curated parameters from over 900 publications resulting in over 1,900 parameter values. The parameters were validated by reproducing the results of knock-out and knock-down experiments from the literature. Then, using independent datasets they found that the model predicted accurate fluxes through glycolysis and the pentose phosphate pathway. The model was able to incorporate the function of each of the 525 genes in the *M. genitalium* genome, describe the complete life cycle at the level of discrete molecules and predict measurable cellular behaviors and phenotypes.

In addition to replicating known values, one of the key goals of systems biology is to make novel predictions that can lead to new hypotheses and experiments. One novel prediction made by Karr *et al.* was the rate of protein-protein collisions within the cell, which is currently unmeasurable [50]. They further predicted that most protein-protein collisions are initiated by either RNA- or DNA-polymerase, which causes displacement of single-stranded binding proteins or structural maintenance of chromosome proteins. This is a novel result of the model which can be used as starting point for describing new hypotheses and designing new experimental protocols. If not for the methodology of systems biology, it is possible that these measurements and predictions would have taken many years to be measured, or may never have surfaced.

1.1.2 Metabolic Network Reconstruction

Metabolic network reconstructions are used to organize and contextualize high-throughput molecular data. Reconstructions are networks of chemical reactions that are often laboriously hand-curated from textbooks and literature then experimentally and computationally validated. For example, the Human Metabolic Network Reconstruction is a collection of metabolic reaction mechanisms and metabolic enzymes compiled from many sources including Gene Ontology, Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, and EntrezGene [28]. Data from these sources are matched by overlapping identifiers and integrated into a complete model of metabolic reactions.

The number of genome-level metabolic network reconstructions has sharply increased from the first in 1999 [31] to well over 50 hand-curated reconstructions and hundreds of *in silico*-generated reconstructions. There are five general categories of applications of metabolic network reconstructions: contextualizing high-throughput data, metabolic engineering, hypothesis-driven discovery, studying ecological relationships between organisms, and network property discovery [77]. The work in subsequent chapters will focus on network property discovery, ecological relationships between organisms and hypothesis-driven discovery.

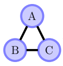
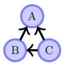
1.1.3 Dynamic and Static System-level Models

Dynamic models, like that of *M. genitalium*, are essential contributions to the field of systems biology, but it is also important to understand the "static" component of biochemical networks [59, 70], that is, network structure and composition. Static features do not capture cellular processes over time, but instead characterize the capacity and capabilities of the network. Many topological and architectural characteristics of biological networks are shared with other complex networks, like the World Wide Web, language and social networks [6, 70], and methodologies developed for non-biological systems can often be applied to understand biological networks.

1.2 Network Theory

Networks are mathematical abstractions that describe the relationships between discrete entities. A network can be represented as a graph containing nodes (vertices) connected by edges. In this work, the words "network" and "graph" will be used interchangeably. In biological networks we often discuss the relationship between local- and global-properties of networks. Global properties are those that apply to the entire graph (*e.g.* diameter, clustering coefficient and degree distribution), while local properties often describe features of individual nodes (*e.g.*, shortest path, degree and centrality). The Human Metabolic Network Reconstruction (Human Metabolic Network Reconstruction (RECON1)) provides a clarifying example of global and local properties (Figure 1.2). Although the network is dense and highly connected, there is a hierarchical structure that is visually apparent. Nodes cluster into their respective organelles, and within those organelle clusters, certain metabolites cluster into subgraphs with neighboring metabolites. One of the key goals of network biology is to understand how the small, local clusters give rise to the emergent structure of the complete graph.

1.2.1 Review of Relevant Graph Terminology

Graphs can be either directed or undirected. *Undirected graphs* illustrate associative relationships but cannot convey sequential information (*e.g.*, ). *Directed graphs* contain information about the direction and sequence of information flow through the network (*e.g.* ). Both types of graphs have their applications. For instance, in social networks it might be not obvious which way influence flows through a group of friends, so using an undirected graph is appropriate. In a metabolic network we typically know the direction of reaction pathways, so directed graphs are appropriate.

Many interesting features of networks can be derived using a series of metrics such as degree, closeness and betweenness centrality, and the clustering coefficient. The *degree* of a particular node in a network is the total number of edges connected to the node. For directed graphs, degree can be further reduced into in- and out-degrees.

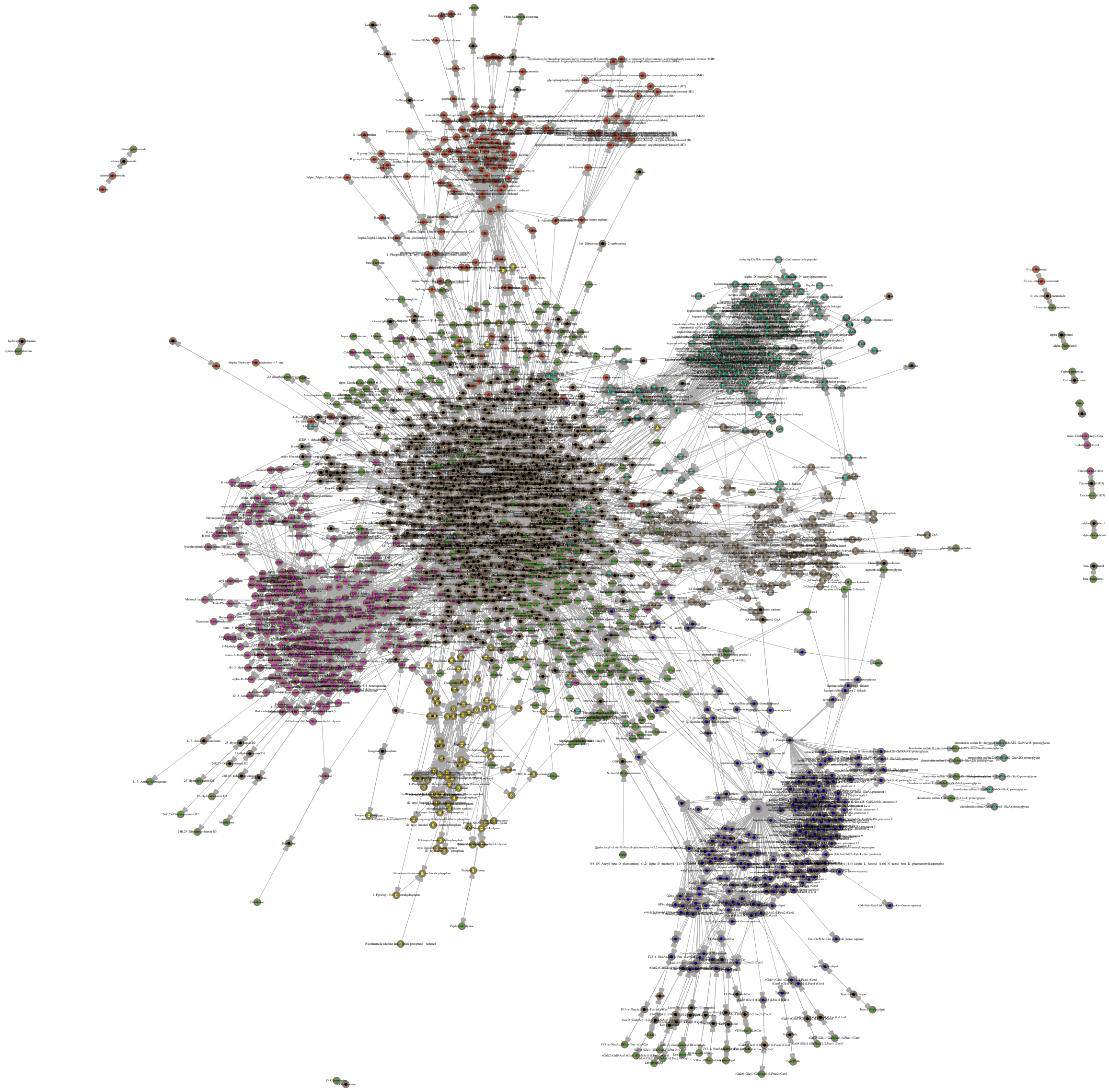
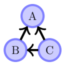


Figure 1.2: Graphical representation of the Human Metabolic Network Reconstruction [28]. Each node represents a metabolite and each edge indicates directional association of metabolites to one another. Nodes are colored by the organelle in which they reside.

The *in-degree* of a node is the total number of edges going into the node, and the *out-degree* is the number of edges emanating from the node. In undirected graphs the in-degree equals the out-degree. In the graph 

, the out-degree of node C is 2, while the in-degree is 0. The in-degree of node A on the other hand is 2, while the out-degree is 0.

Another important graphical metric is centrality. *Closeness centrality* is the average distance from one node to all other nodes in the graph (also called the average shortest path). The closeness centrality metric assigns large centrality to nodes with the smallest path distance to the other nodes in the graph [6, 8].

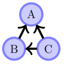
The previous two metrics, degree and closeness, provide good measurements of the highly connected nodes in a graph, but they overlook intermediate nodes that may be fundamental in connecting separate modules of the graph. The *betweenness centrality* is the number of shortest paths between all nodes to all other nodes that pass through a given node. In essence this metric captures the number of times a particular node is passed through when traversing from one node to another, or how often a particular node is between all other nodes [8].

The *clustering coefficient* is a measure of the tendency of a collection of nodes to cluster into highly connected groups. Many non-random networks show high degrees of nodal clustering when compared to random graphs. For instance, the metabolic networks of 43 organisms had clustering coefficients that exceeded those of random scale-free graphs by an order of magnitude [83].

Network properties such as average degree and average centrality describe the global properties of engineered networks well, but the meaning and implications of these measurements in biological networks must be demonstrated. Basler *et al.* evaluated properties of metabolic networks and found that many organizational network properties, such as average path length and clustering coefficient, emerge as the result of positive evolutionary pressure [9]. This finding suggests that mathematical network characteristics contain biological meaning and can provide insights into how evolution has shaped the organization of and function of metabolism.

1.2.2 Biological Networks Display Modularity

Biological networks typically follow a scale-free degree distribution, which means that the probabilty of observing a node with k edges follows a power-law (*i.e.*, $P(K) \sim k^{-\gamma}$) [9, 70, 83]. A scale-free degree distribution implies the presence of nodes with degrees that greatly exceed the average degree in the network, called *hubs*. These hubs often serve as intermediaries that connect groups of nodes into *modules*, which are separable clusters that can primarily function independently [39, 51]. For example, modular features in RECON1 are visually apparent by the manner in which metabolites cluster into organelles and then into smaller groups within those organelles (Figure 1.2).

Like many biological networks, metabolic networks display modularity [83]. Michoel *et al.* developed a network motif aggregation statistic to quantitatively test whether modules in protein-protein interaction, post-translational modification and transcriptional regulatory networks are the result of the aggregation of network motifs [68]. They found that the Feed-Forward Loop structure () significantly aggregates into modules. The aggregation of motifs into modules was likewise observed by Kashtan and Alon who tested the hypothesis that modules emerge spontaneously due to changes in the environment [51]. Using an evolutionary algorithm they found that networks spontaneously evolve modules when environments are added or removed *in silico*. Evolved modular networks also showed higher fitness levels and enrichment of several network motifs, again including the feed-forward loop motif. Finally, Basler *et al.* found the clustering coefficient in metabolic networks to be under positive selective pressure and not purely the result of thermodynamic constraints [9].

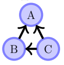
It is not clear why modularity occurs in biological networks, or by what mechanisms modules evolve. Many have suggested that they evolve through duplication or the benefits that they may confer such as robustness or stability [57]. It appears to be the case that modules are beneficial because they allow for rapid adaptation to changes in environments by making minimal changes to the module. For example, in the process of chemotaxis many modules work independently including nutrient sensing, cellular orientation and metabolism [51]. If the environment changes and a new energy source emerges, only a few elements of each module need to be adjusted to quickly respond and nothing needs to be built from scratch.

1.2.3 Network Motifs

A network motif (or simply "motif") is a small, repeated pattern or subgraph that is over-represented (enriched) in a network in comparison to its abundance in a random graph [69, 70]. Under-represented (suppressed) motifs are often referred to as "anti-motifs," however in this work we will refer to enriched and suppressed subgraphs simply as "motifs." As previously described, many network motifs were identified as aggregating in modules within biological networks.

1.2.4 Network Motifs May Imply Biological Function

Motifs are of chief interest in network theory and systems biology because significantly enriched motifs may determine the dynamic properties of whole systems [71]. Dynamic behaviors that have been linked to specific motifs include bistability and ultrasensitivity [89], failure tolerance [71] and network stability [27, 62, 81]. In addition, motifs provide a reduced, simplified framework in which to describe global functionality without losing resolution [2, 89, 90, 101]. For instance, Vázquez *et al.* showed in the transcription and metabolic networks of *E. coli* and *S. cerevisiae* that motif abundances could be predicted from two global parameters describing modular and scale-free topology, demonstrating that global and local graph properties are mutually defined [98].

Milo *et al.* showed that certain types of networks have unique motif distributions [70]. Specifically, transcription and signal transduction networks had distinctive distributions when compared to non-biological networks like the World Wide Web and social networks. One of the defining features of biological networks in comparison to engineered networks is the enrichment of the Feed-Forward Loop motifs (). The feed-forward loop structure is ubiquitous in biological networks and is associated with a breadth of functions including decreased response time of gene expression after a stimulus [71], pulsatility [70], and reliable information processing [55].

Above all, motifs are of interest because they provide a source of novel data that provide insights into molecular evolution [6]. For instance, one could hypothesize that motifs conferring advantages to the organism would be preferentially enriched, whereas those that are potentially harmful would suppressed [82]. Conant and Wagner showed that gene circuits have evolved independently and repeatedly in the transcription networks of *E. coli* and *S. cerevisiae* resulting in the convergent evolution of two motifs, the 4-node bi-fan and the 3-node feed-forward loop [22]. This finding provides strong evidence that the feed-forward loop and bi-fan motifs are ideal structures for their roles in transcription networks. This point is also made by Klemm and Bornd-holdt who defined a measure of dynamic reliability of information processing and showed that the enrichment of the feed-forward motif is correlated with the ability of that structure to produce reliable signals [55]. A similar finding was made by Prill *et al.* that motif enrichment and robustness to small perturbations were positively correlated so that those motifs that were most stable were also the most abundant [81].

When taken together, these studies highlight the crucial role motifs play in biological networks ranging from signal transduction to metabolism. Despite their important roles, motif structure has not been extensively studied in metabolic networks. In this work, we aim to describe the distribution of motif abundances in a wide range of organisms, begin to characterize their role in the network property of failure-tolerance and apply motif analysis to specific biological hypotheses. Our overarching hypothesis is that metabolic motifs capture relevant functional information that can be used to compare the metabolism of different species and organelles. That is, motifs provide a reduced, compact framework that can be used as a proxy for metabolic functionality, and thus be employed to make inferences about species-level differences in metabolism.

1.3 Specific Aims

Aim 1: Identify, characterize and compare the metabolic motifs present in 21 distinct organisms and seven organelles. In this aim we compile the metabolic networks of 21 organisms from their network reconstructions and mine them for motifs of node-size three.

Hypotheses: We hypothesize that each organelle will have a unique significance profile which is reflective of its distinct function in the cell. Further we hypothesize that motifs can be characterized by their enzymatic associations.

Aim 2: Assess the metabolic network of *E. coli* for the property of failure tolerance, and relate it to the relative abundances of particular network motifs.

Hypothesis: We predict that the motifs that displayed enrichment in the cytosol of *E. coli* as identified in Aim 1 (motifs 2, 3, 7 through 13), will have the property of increased failure tolerance.

Aim 3: Apply the methodology of motif mining and analysis to test specific hypotheses of organelle evolution.

Hypotheses: Based on the finding from Aim 1 that certain motifs can be mapped to specific chemical and biological functions, we hypothesize that organelles most closely related to their ancestral species will display similar significance profiles. Further, we hypothesize that organelles derived from endosymbiosis will display distinct patterns of enrichment compared with organelles derived from membraneous infolding.

CHAPTER II

Network motifs provide signatures that characterize metabolism

2.1 Introduction

Life can be studied at many different strata ranging from the molecular level to the ecosystem. Regardless of the stratum, a fundamental characteristic of life is a high degree of order which is divided into hierarchical levels of organization and function [83]. Because metabolism is a fundamental process shared among all living things, it directly influences every stratum of biological function. Molecules are activated by metabolites, and ecosystems are forged to satisfy metabolic requirements. Understanding the emergent, organizational properties of metabolism is one way to unravel molecular evolution [62], and is thus a crucial goal in the field of systems biology [6].

As a consequence of advances in the field of molecular biology, particularly in sequencing technology, it is now common to assemble genome-level metabolic networks by integrating known biochemical pathways with genomic annotation [40]. These large biochemical networks are often referred to as metabolic network reconstructions [77]. Many organizing principles of biological networks have been described as a result of the availability of large biological networks, and commonalities among metabolic, signaling and transcription networks have emerged. For instance, biological networks share global properties like scale-free degree distributions [9, 83] and modularity [39, 51]. They also share local properties such as patterns of network motif enrichment that are unlike those of engineered networks [70]. A network motif (or just "motif") is a repeated pattern or subgraph that is over- or under-represented in a network compared to its expected abundance in a collection of random graphs [33, 69]. Motifs are of chief interest in systems biology because their patterns of enrichment may determine the dynamical properties of whole networks [71]. Dynamic behaviors that have been linked to specific motifs include bistability and ultrasensitivity[89], failure tolerance [71] and network stability [27, 62, 81]. In addition, motifs provide a reduced, simplified framework in which to describe global functionality without losing resolution [2, 89, 90, 101]. Vázquez *et al.* [98] demonstrated that local graphical parameters could be used to predict global properties, and likewise that global properties could predict local network features in the transcription in metabolic networks of *E. coli* and *S. cerevisiae.* This result demonstrates the possibility of making global, organism-level inferences by characterizing local properties with motifs.

Since global properties of metabolic networks can be described using motif distributions, it is possible to make inferences about molecular evolution and comparative metabolic functionality. One method to measure functionality of motifs is with the collection of enzymes associated with that motif. Most biological reactions require enzymes for catalysis, and as a result the amount and type of enzymes associated with a particular organism partially characterize the organism's range of function.

In this work we characterized all 3-node motifs using Enzyme Commission numbers (EC) to show that in metabolism motif abundance is directly related to chemical and biological function. Further, we present a comparative analysis of the distributions of 3-node motifs in the metabolic pathways of 21 species[1, 12, 17, 21, 25, 28, 29, 35–37, 44, 53, 58, 64, 84, 86, 91, 95, 96, 102] by compartmentalizing the metabolism in the cellular organelles: cytosol, endoplasmic reticulum (ER), Golgi, mitochondrion, nucleus and peroxisome. Fully compartmentalized metabolic networks enabled us to test whether the motif distribution is unique for each structure in the hierarchical organization of the cell. We found that each organelle has a unique metabolic signature which is indicative of its role in the cell. Finally, we illustrated that motifs are able to capture biological differences between species.

2.1.1 Hypotheses

We hypothesize that network motifs contain biochemical meaning and can be uniquely characterized by the types of reactions and pathways in which they participate. Further, we hypothesize that metabolic processes are largely organelle-specific and that each of the organelles under investigation will exhibit unique motif distributions, reflecting their distinct roles in the cell.

2.2 Methods

2.2.1 Selection of Metabolic Network Reconstructions

All data in this work were from previously published metabolic network reconstructions (Table 2.2.1). 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.

Table 2.1: List of organisms analyzed in the present work, and relevant network characteristics.

Species	Kingdom	Nodes	Edges	Compartment	
A. thaliana	Plantae	1501	3411	Cytosol	[25]
		50	122	Mitochondrion	
		57	112	Peroxisome	
C. reinhardtii	Protista	660	2165	Cytosol	[17]
		25	58	Golgi	
		260	652	Mitochondrion	
		48	56	Nucleus	
C. thermocellum	Bacteria	516	1604	Cytosol	[84]
D. ethenogenes	Bacteria	501	1498	Cytosol	[1]
E. coli	Bacteria	908	2863	Cytosol	[36]
H. pylori	Bacteria	400	1194	Cytosol	[96]
H. salinarum	Archaea	526	1269	Cytosol	[44]
H. sapiens	Animalia	779	2181	Cytosol	[28]
		184	402	ER	
		234	591	Golgi	
		189	351	Lysosome	
		352	905	Mitochondrion	
		85	173	Nucleus	
		135	335	Peroxisome	
G. sulfurreducens	Bacteria	466	908	Cytosol	[64]
M. acetivorans	Archaea	697	1832	Cytosol	[58]
M. barkeri	Archaea	542	1602	Cytosol	[37]
M. musculus	Animalia	842	2399	Cytosol	[91]
		182	400	ER	
		262	643	Golgi	
		205	383	Lysosome	
		385	1019	Mitochondrion	
		85	176	Nucleus	
		140	342	Peroxisome	
M. tuberculosis	Bacteria	486	1417	Cytosol	[35]
P. pastoris	Fungi	571	1774	Cytosol	[21]
		19	22	ER	
		16	20	Golgi	
		225	576	Mitochondrion	
		36	62	Nucleus	
		74	161	Peroxisome	
S. aureus	Bacteria	549	1657	Cytosol	[12]
S. cerevisiae	Fungi	528	1657	Cytosol	[29]
		15	18	ER	
		11	17	Golgi	

		214	531	Mitochondrion	
		30	45	Nucleus	
		73	186	Peroxisome	
S. typhimurium	Bacteria	852	3102	Cytosol	[95]
T. maritima	Bacteria	727	2478	Cytosol	[102]
V. vulnificus	Bacteria	831	2494	Cytosol	[53]
Z. mays	Plantae	1418	2463	Cytosol	[86]
		60	78	Mitochondrion	
		50	51	Peroxisome	

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 [11]. Neither COBRA nor Matlab were used for analysis, but these criteria insured that the reconstructions were curated using similar protocols and 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, a dummy variable indicating the reversibility of each reaction and the subsystem to which the reaction belonged (*e.g.* "Folate Biosynthesis," "TCA Cycle," "Salvage Pathway of ATP").

2.2.2 Graph Construction

With the stoichiometric matrices from each of the 21 metabolic network reconstructions, we generated a list of reaction equations. Reversibility of reactions was considered so that all reverse reactions were included in the motif mining procedure. The list of equations was used to generate a FANMOD input file according to FANMOD specifications [100].

2.2.3 Identifying Enriched or Suppressed Motifs

A tool for fast network motif detection (FANMOD) was employed to identify motifs in metabolic networks [99]. FANMOD enumerates all subgraphs of a specified size in a network rather than estimating frequencies. For computational and analytical tractability this work focused on 3-node motifs (Table 2.2.5). 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. The choice of 1,000 random graphs was based on the typical comparison size in literature, but results did not change with 500 nor 5,000 random graphs. 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 normalization step is necessary because z-scores tend to increase with network size, resulting in biased comparisons when network sizes vary [70]. The z-score is computed with equation 2.1:

$$Z_i = N_{met_i} - \hat{\mu}(N_{random_i})/\hat{\sigma}(N_{random_i}) \tag{2.1}$$

where N_{met_i} is the number of occurrences of motif i in a metabolic network and N_{random_i} is the number of occurrences of motif i in a random network. The resultant z-scores are normalized and yield the Significance Profile (SP) (motif distribution):

$$SP_i = Z_i/\sqrt{\sum Z_i^2} \tag{2.2}$$

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 $p \leqslant 0.05$.

Motif mining is a computationally costly task which, in general, cannot be performed on a personal computer. Motif mining was conducted on a high performance computing cluster at the University of Michigan Center for Advanced Computing.

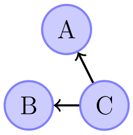
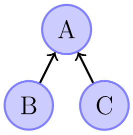
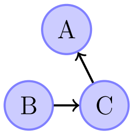
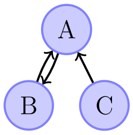
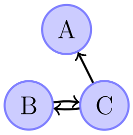
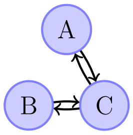
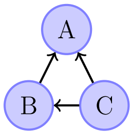
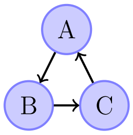
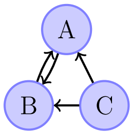
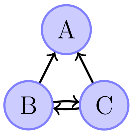
2.2.4 Choosing a Random Background

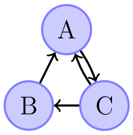
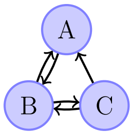
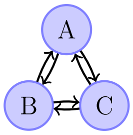
The results of any motif mining procedure are sensitive to the choice of random background used for generating the random graphs used for comparison. In this work, we generated random graphs of equal size and connectivity using the method of edge switching along with the "global constant" randomization model [100]. Global constant randomization holds the total number of bidirectional edges constant, while any particular node may gain or lose a bidirectional edge. A small comparison of the three randomization models ('local,' 'global' and 'no regard') was done, and the results did not change appreciably.

2.2.5 Substrate Graphs

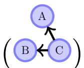
All motifs were represented as substrate graphs. Substrate graphs represent associativity of nodes, rather than mechanistic relationships like those of a bipartite graph. Each graph type has its advantages and disadvantages. For instance, when using bipartite graphs of size three, it is possible to generate motifs that contain no biological meaning. For example a bipartite motif might contain two nodes that represent reactions and one that represents a metabolite, which is not a valid chemical mechanism. Similarly, because substrate graphs are associative, we cannot know the chemical mechanism from the motif structure.

Table 2.2: Structure, FANMOD identifier and label for all 3-node motifs.

	Motif Structure	Fanmod ID	Name
1		6_000000110	V-Out
2		36_000100100	V-In
3		12_000001100	3-Chain
4		164_010100100	Mutual In
5		14_000001110	Mutual Out
6		78_001001110	Mutual V
7		38_000100110	Feed-forward Loop
8		140_010001100	3-Loop
9		166_010100110	Regulated Mutual
10		46_000101110	Regulating Mutual

11		102_001100110	Mutual and 3-Chain
12		174_010101110	Semi-Clique
13		238_011101110	Clique

2.2.6 Metabolic Characterization of Motifs

As previously mentioned, one cannot immediately infer reaction mechanisms from substrate graphs. To accomplish this, we enumerated every possible mechanism capable of yielding each of the 3-node motifs. To illustrate, take the first motif () which has two possible mechanisms: It could be either $C \to A$ and $C \to B$ or $C \to A + B$. In addition to enumerating both of the possible mechanisms, it is also necessary to enumerate all the combinations of reversibility. That is, it is possible that the correct mechanism for the first motif is $C \to A$ and $C \to B$, but that the $C \to B$ reaction is actually the reverse direction of a different reaction.

In order to characterize each motif, we used the stoichiometric matrices from the *E. coli*, *H. sapiens*, *M. barkeri* and *S. cerevisiae* metabolic network reconstructions. Stoichiometric matrices contain integers that denote whether a metabolite is produced, consumed or not a participant in a particular reaction. Negative integers denote consumption, positive integers denote production, and zeros denote absence. We generated a second stoichiometric matrix that contained the reverse mechanisms for all reversible reactions. We searched for motif mechanisms using a series of conditional tests in R. For example to find the reaction $C \to A$, we used:

```
which(Stoich[paste(motif1$nodeA[i]), ] > 0 &
Stoich[paste(motif1$nodeB[i]), ] == 0 &
Stoich[paste(motif1$nodeC[i]), ] < 0)
```

The *which* function returns the stoichiometric matrix column indices for reactions where node A is being produced ($A > 0$), node B does not participate ($B == 0$) and node C is being consumed ($C < 0$). Similar conditionals were used for all other combinations of reversibility.

2.3 Results

Prior to analysis we mined for motifs in the metabolic networks of 21 species (see Methods). Motifs are numbered as previously presented in the literature [70] and are roughly in order of increasing edge density. For example, motif 1 (V-out) has only two, non-reversible edges, while motif 13 (Clique) has six edges (or three fully reversible edges). Motif names briefly describe the biochemical relationship between the three nodes, and motif names and numbers will be used interchangeably.

2.3.1 Motifs can be uniquely characterized by their enzyme functionality

The first two digits of EC numbers in the networks yielded a total of 47 enzyme classes. For each enzyme class, we calculated the proportion of reactions associated with each motif and found that each of the 13 motifs was associated with a distinct catalog of enzymes (Figure 2.1).

The number and type of enzymes associated with the V-Out, V-in, 3-Chain and Feed-forward Loop motifs (motifs 1-3 and 7) was wide-reaching. In motifs one through three, 43 of the 47 total enzyme classes had non-zero proportions. This result implies that these motifs have a breadth of function, perhaps serving as intermediates between other motifs. The Feed-forward Loop motif (motif 7) had EC proportions similar to those of motifs 1-3, but was distinguished by increased proportions of EC 2.4

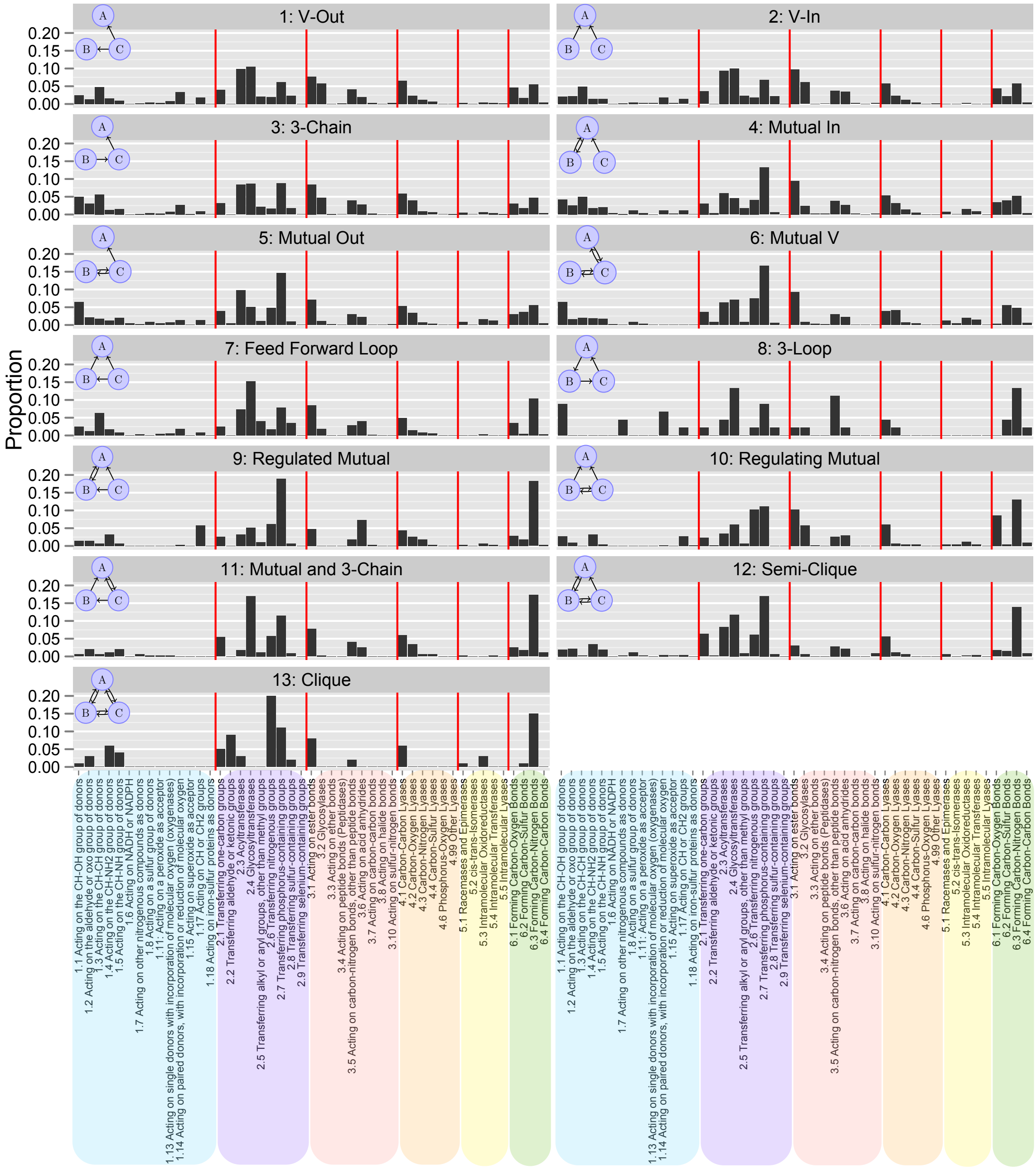


Figure 2.1: Proportions of two-digit Enzyme Commission numbers associated with each motif. EC data were obtained from the metabolic network reconstructions of *E. coli*, *H. sapiens*, *M. barkeri* and *S. cerevisiae*. Red vertical lines separate the six classes of EC number: oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases.

(Glycosyltransferases) and EC 6.3 (Forming carbon-nitrogen bonds).

Mutual In (motif 4), Mutual Out (motif 5) and Mutual V (motif 6) had EC distributions similar to one another. Enzymes that were key in characterizing these motifs include 1.1 (Acting on CH-OH group of donors), 1.6 (Acting on NADH or NADPH), 1.8 (Acting on a sulfur group of donors) and 5.4 (Intramolecular transferases). Although many of these enzymes are modestly represented, their presence remains a relevant characteristic. For example, EC 1.8 constitutes 1% of all enzymes associated with motifs 4-6, but was rarely found in all other motifs.

The EC distributions of Regulated Mutual (motif 9) and Regulating Mutual (motif 10) vary in key enzymes with respect to the other 11 motifs, but are similar to one another. A distinguishing characteristic of Regulated Mutual is the high level of EC 2.7 (Transferring phosphorus-containing groups) which comprises 20% of its total collection of enzymes, and nearly double that of motif 10. This is also true for ECs 1.17 (Acting on CH or CH2 group) and 3.6 (Acting on acid anhydrides) which are doubled in the Regulated Mutual motif versus the Regulating Mutual motif. The enzyme that distinguishes motif 10 from motif 9 and all others is EC 6.1 (Forming carbon-oxygen bonds) which comprises 9% of the enzymes associated with motif 10 and is twice to ten times the amount seen in all other motifs.

Mutual and 3-Chain (motif 11) and Semi-clique (motif 12) were similar in their enzyme proportions. Enzymes that distinguish these two from all other motifs are high proportions of glycotransferases (EC 2.4) and enzymes catalyzing carbon-oxygen bonds (EC 6.2). They differ primarily in the amounts of EC 2.3 (acyltransferase) which is four times greater in motif 12 compared to motif 11.

Finally, the motifs 3-Loop (motif 8) and Clique (motif 13) are particularly interesting because their EC distributions are sparser than the other motifs, suggesting a narrower range of function. The 3-Loop motif had non-zero proportions in just 19 of the 47 enzyme classes and Clique had non-zero proportions in only 17 of the 47.

These two motifs also displayed enzymes distributions that were unlike all other motifs. The 3-Loop (motif 8) shows proportions of ECs 3.5 (Acting on carbon-nitrogen bonds, other than peptide bonds) and 1.7 (Acting on other nitrogenous compounds as donors) that are at least twice the amount in all other motifs. Motif 13 is lacking any glycosyltransferases (EC 2.4) which are ubiquitous in every other motif.

2.3.2 Motif function recapitulates motif structure

Using the proportions of enzyme classes associated with each motif (Figure 2.1), we calculated a pairwise distance metric to quantify the level of similarity between the motifs. In agreement with the previous section, we found that motifs with similar structural features have similar proportions of enzyme classes in metabolic networks (Figure 2.2). The feed-forward structures (motifs 1, 2, 3 and 7) fall within their own cluster with motifs 1 and 2 showing more similarity with each other and less similarity with motifs 3 and 7. Motifs 4, 5 and 6 cluster together, but the motifs that share the structural property of one reversible edge and one non-reversible edge (4 and 5) cluster more closely to each other than to motif 6. This finding shows that the addition of one edge to a motif can distinguish its enzymatic associations (this is seen also in the clustering of motifs 3 and 7).

The findings depicted in figure 2.2 allowed us to conclude that motifs have chemical signatures that can be quantified with the EC numbers corresponding to the reactions in which they participate. Further, the similarity of the EC distributions is related to the structural features of the motifs such that motifs that are structurally similar also share similarities in their distributions of enzymes.

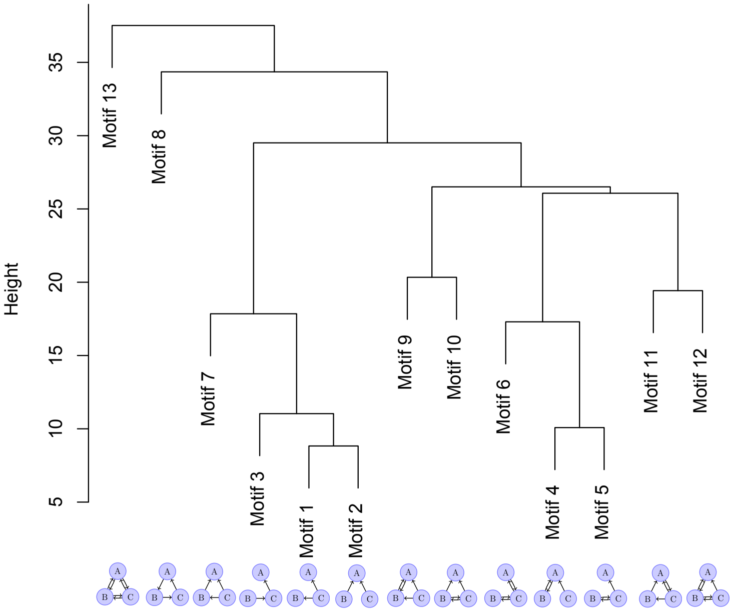


Figure 2.2: Dendrogram showing the distance between the proportions of EC numbers each of the 13 motifs.

2.3.3 Each cellular organelle displays a unique significance profile indicative of its unique role in the cell.

We used fully compartmentalized metabolic networks from 21 species to describe average motif distributions for each organelle. Each of the six organelles displayed a unique pattern of motif enrichment and suppression (Figure 2.3). Confidence intervals not containing the null, $z = 0$, are statistically significantly enriched or suppressed.

In the cytosolic compartment, 11 of the 13 motifs achieved statistical significance with the exception of Mutual Out (motif 5) and Mutual V (motif 6). The tightness of the CIs indicates relatively small variance between organisms and suggests that the local structure in the cytosol is well conserved across all kingdoms of life in our sample.

The ER had only two motifs that reached statistical significance, Regulated Mutual (motif 9) and Regulating Mutual (motif 10), both of which were suppressed. This is due primarily to inter-species variation in motif enrichment as seen from the

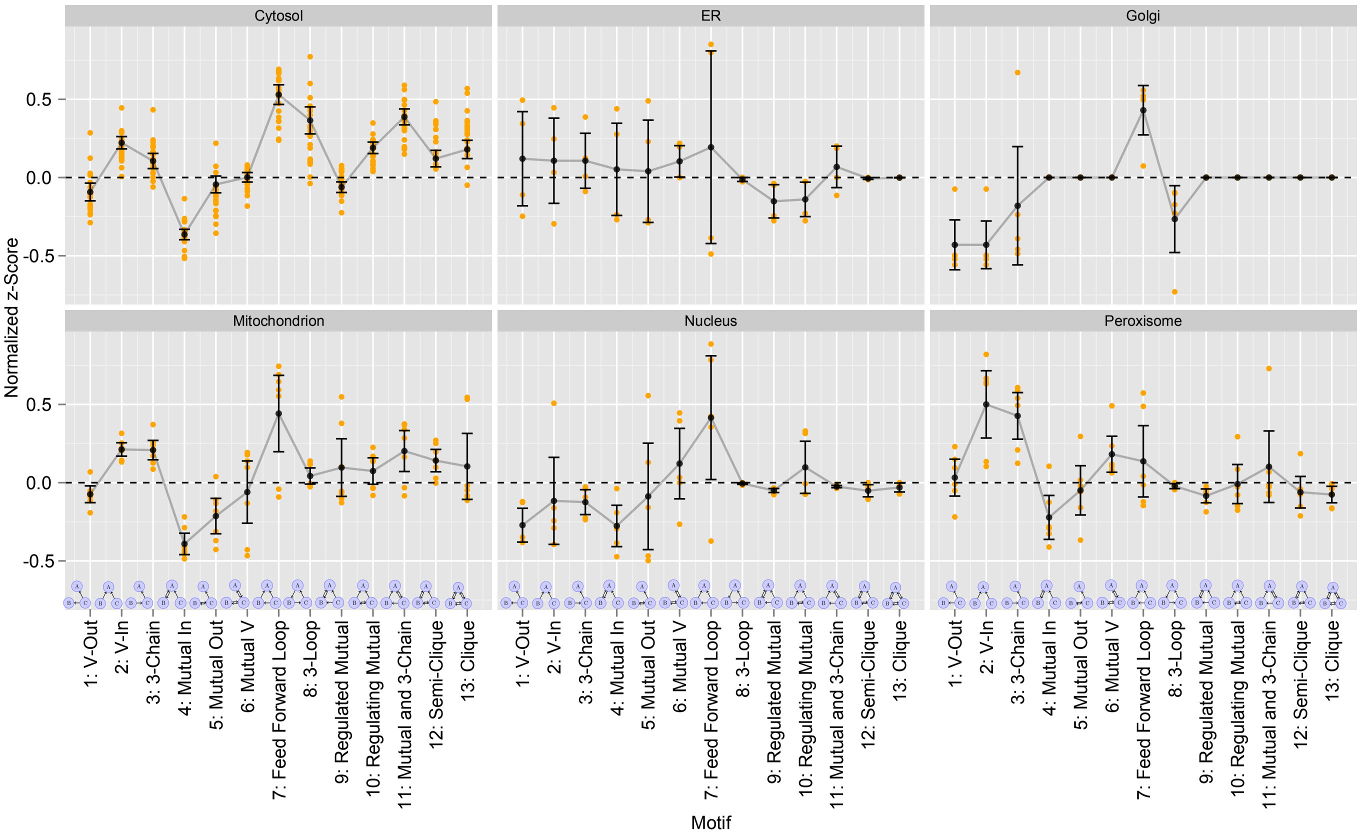


Figure 2.3: Average significance profiles of 3-node motifs by organelle. Black points (connected for visualization of relative abundances) indicate the average normalized z-score for each motif. Orange dots show the scatter of the normalized z-scores corresponding to each of the 21 organisms. Standard errors of the means were estimated from 1000 bootstrap samples of the normalized z-score and used to calculate 95% confidence intervals. Confidence intervals not containing the null, z = 0, are statistically significantly enriched or suppressed compared to the expected abundance in 1,000 random graphs.

scattered points in figure 2.3.

The Golgi showed enrichment in only one motif, Feed-forward Loop (motif 7), and suppression or absence in all others. This profile suggests that, unlike the cytosol, the Golgi performs a narrow set of metabolic functions, for example protein glycosylation, and therefore one type of motif is sufficient.

The nuclear motif distributions also displayed significant enrichment of the Feed-forward Loop motif (motif 7) and had high levels of inter-species variation (as seen from the points in figure 2.3).

An intriguing finding is the similarity of the cytosol, mitochondrion and peroxisome motif distributions. The profiles are remarkably similar with motifs 1 to 7 displaying the same pattern of enrichment and suppression (though not the same pattern of statistical significance) among all three organelles.

It is notable that the V-In (motif 2) and 3-Chain (motif 3) motifs are enriched in cytosol, mitochondrion and peroxisome but suppressed or non-significant in the ER, Golgi and nucleus. Recall, that these motifs were associated with a wide range of enzyme classes and had non-zero proportions for nearly all 47 enzyme classes. Because the cytosol, mitochondria and peroxisomes contain a more varied and complex set of metabolic reactions and roles, it is reasonable that we see this pattern of enrichment.

2.3.4 Within-species motif enrichment can be used as an indicator of metabolic functionality.

The mitochondrial motif distribution provides an interesting example in which to evaluate inter-species variation in motif enrichment. The mitochondrial sample is reduced to include only seven species because prokaryotes do not contain mitochondria. There is very little variation in mitochondrial motif distributions between the two species in Animalia, *H. sapiens* and *M. musculus* (Figure 2.4). Likewise, the two Fungi, *S. cerevisiae* and *P. pastoris*, show identical distributions to one another, and to those of Animalia. The two plants, *A. thaliana* and *Z. mays*, have motif distributions unlike those of any of the other kingdoms, showing enrichment of both Regulated Mutual (motif 9) and Clique (motif 13) even while those motifs are primarily suppressed in the other kingdoms. Similarly, Mutual V (motif 6), Feed-forward Loop (motif 7), and Mutual and 3-Chain (motif 11) are suppressed in plants but primarily enriched in other kingdoms. The motif distribution of the protist *C. reinhardtii* is somewhat of a hybrid of the plant and the animal distributions, possibly reflecting commonalities with plants due to the photosynthetic elements of their metabolism.

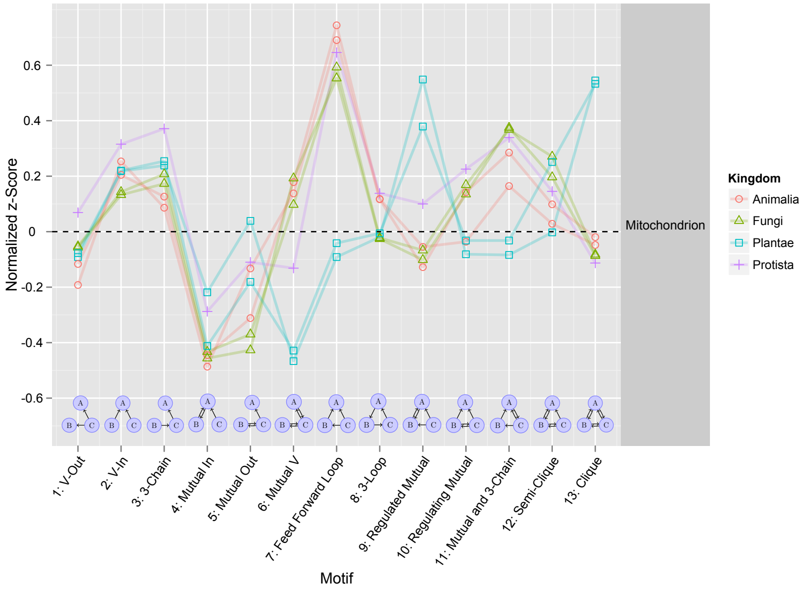


Figure 2.4: Mitochondrial motif distribution. Each line corresponds to an organism, and the line color denotes the kingdom of life to which the organism belongs.

We should expect some variation in motif distributions between plants and other organisms because the evolutionary history of plant mitochondria differs markedly from that of Bacteria, Fungi and Animalia [52].

The Clique motif (motif 13) is enriched in plant mitochondria, but suppressed in animals, fungi and protista. In the previous section, the Clique motif was found to be characterized by the transferral of aldehyde or ketonic groups (EC 2.2) and

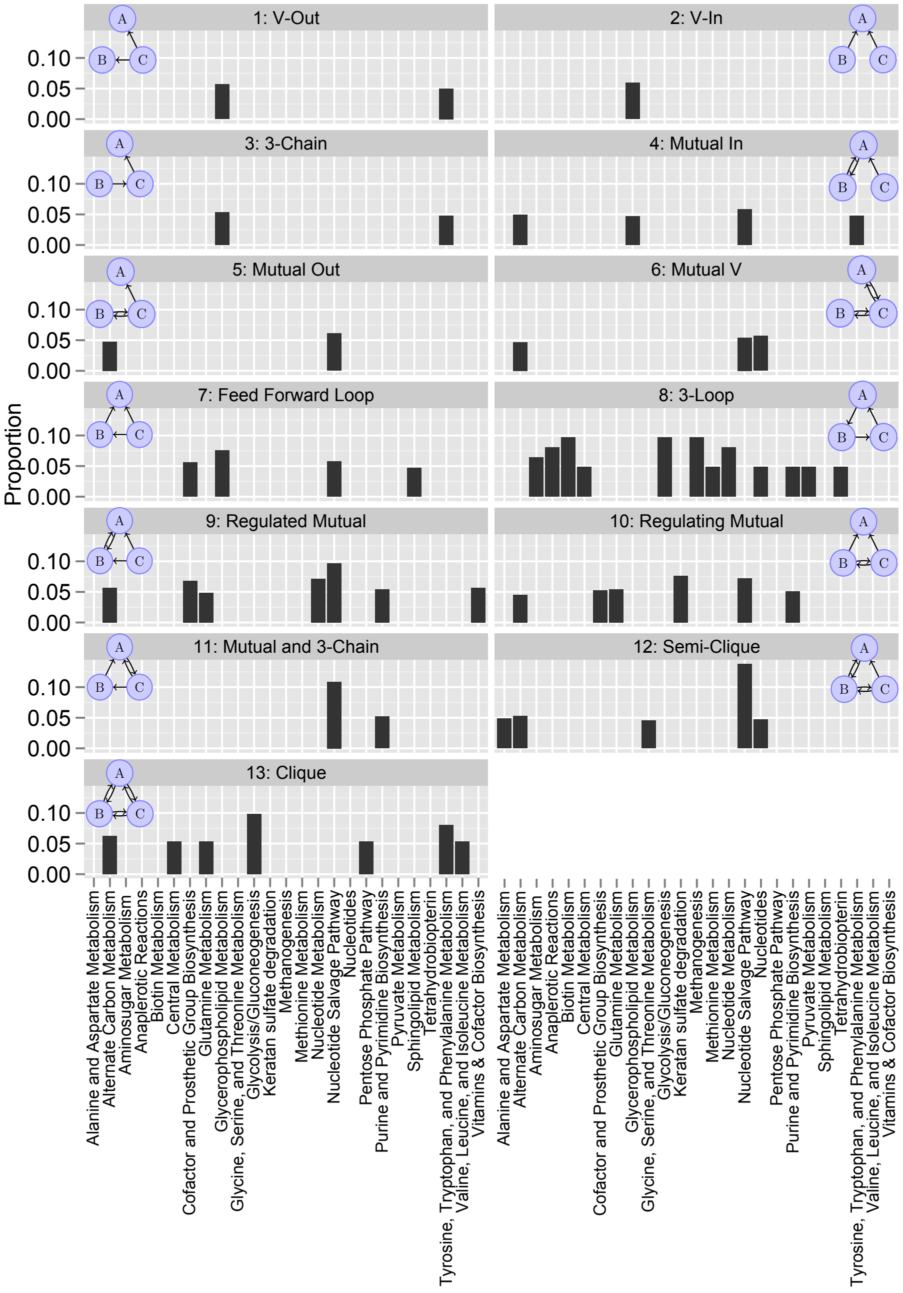


Figure 2.5: Barplots showing proportions of pathway participation for each motif. A minimum threshold of 5% was used for clarity.

intramolecular oxidoreductases (EC 5.3). Oxidoreductases are a class of enzymes that catalyze the transfer of electrons from one molecule to another, and they are common in the pathways of glycolysis and gluconeogenesis. In most organisms, the pathways of glycolysis and gluconeogenesis occur in the cytoplasm, however in plants these pathways are contained within the mitochondria [42]. Approximately 10% of the reactions of the Clique motif are considered part of glycolysis/gluconeogenesis, and this 10% constitutes the largest proportion for that motif (Figure 2.5). The remaining pathways of the Clique motif take place primarily outside of the mitochondria, which helps account for the suppression of the Clique in all other kingdoms.

The Mutual V (motif 6) motif is suppressed in plants while enriched in animals and fungi. Similarly, the Regulated Mutual (motif 9) motif is enriched in plants while suppressed in animals and fungi. Interestingly *C. reinhardtii*, a photosynthetic algae, follows the same pattern of suppression and enrichment as plants, suggesting that these two motifs may vary in photosynthetic organisms. Both motifs are associated with biochemical reactions in alternate carbon metabolism pathways (Figure 2.5), which vary between plants and animals due to the presence of chloroplasts in photosynthetic organisms. Chloroplasts create a cellular environment that is rich in carbohydrates, such as sucrose, fructose and glucose, and also rich in oxygen [76].

One of the many functions of a mitochondrion is fatty acid oxidation, which occurs less in plants than in other organisms [88]. The Feed-forward Loop motif (motif 7) is associated with fatty acid oxidation through the EC numbers 1.3, which refers to various types of oxidases and hydrogenases used in the beginning steps of fatty acid oxidation. Also, EC 1.1.1.35 and 1.1.1.211 which are dehydrogenases and 2.3.1.16 a acyltransferase involved in the conversion of coenzymes to acetyl-CoA. Because fatty acid oxidation is relatively rare in plants, we could expect for there to be less enrichment of the Feed-forward Loop motif in plants, which was the case here.

2.4 Discussion

In this work, we have characterized motifs in terms of their enzyme associativities, and we have estimated motif abundances in the metabolic networks of 21 organisms and 6 organelles. However, evaluating the properties of metabolic networks is only as useful as the reconstructions are valid. Many reconstructions are built using previous versions as starting points and thus perpetuate errors and biases that may have been present in previous incarnations of the networks. Due to high degrees of relatedness, we expect that those biases are consistent across most reconstructions because of their high degree of relatedness.

There is also a disconnect between the ever-growing number of fully sequenced genomes and the number of validated, usable network reconstructions to accompany these genomes. Currently, network reconstruction is massively time-consuming and largely done via manual curation. As a result of the time-intensive process of creating metabolic network reconstructions, our sample contained relatively few eukaryotes. Despite this, we expect that while the ensemble of enzymes and pathways associated with each motif will likely change as more reconstructions become available, motifs will still be function-specific.

Notwithstanding the previously mentioned limitations, the findings presented here improve on previous work [33, 97] on metabolic network motifs in two key ways. First, our analyses were restricted to include only manually-curated metabolic network reconstructions. We conducted a small analysis comparing the motif distributions of automated versus manually generated reconstructions and found that automated reconstructions systematically underestimate the number of reversible reactions in metabolic networks (unpublished data). Underestimation of reversibility results in underestimation of motifs with reversible edges (motifs 9-13) and overestimation of simpler motifs (motifs 1-3). Second, as a consequence of the high-quality, compartmentalized reconstructions we were able to present motif distributions for six distinct

Eukaryotic organelles which, to our knowledge, is a novel contribution.

2.4.1 Motifs with a breadth of associated enzyme classes could be intermediaries

In section 2.3.1 we saw that the feed-forward structural motifs (motifs 1-3 and 7) displayed wide-ranging enzymatic associations. We proposed that these motifs might be intermediaries connecting motifs of greater complexity (in terms of edge connectivity) into modules. In networks, modules are semi-autonomous units that can function primarily independently. It has been demonstrated in previous work [26, 68, 98] that motifs aggregate into functional modules in metabolic networks. Kashtan *et al.* showed that network modularity and motif aggregation evolve spontaneously in *in silico* networks exposed to changes in environment, and that the 3-Chain (motif 3) and Feed-forward Loop motifs (motif 7) in particular aggregate in modules [51]. We found that in metabolic networks, 3-Chain and Feed-forward Loop motifs have a breadth of enzyme associativity, perhaps because they aggregate within many metabolic modules. This is also supported by the motif enrichment levels seen in section 2.3.3. Besides the ER and peroxisome, all organelles showed enrichment of the Feed-forward Loop motif (motif 7). In the cytosol, mitochondrion and peroxiome enrichment of the 3-Chain motif (motif 3) was observed. This suggests that high abundances of motif 3 and 7 may contribute to the network modularity and perhaps the benefits conferred by that feature such as stability and robustness [57].

In contrast with motifs 1, 2, 3 and 7, motifs 8 and 13 displayed the narrowest range of enzymatic associativity with non-zero proportions in only 36-40% of all enzyme classes (Figure 2.1). Interestingly, these cyclic motifs were only significantly enriched in the cytosol and no other organelle. Cyclic motifs like 8 and 13 have been shown to have dynamically unstable properties in biological networks (transcription, signal transduction and neuronal signaling) [81] and to be unreliable in the context of information processing [55]. This could explain the lack of enrichment of these motifs in networks where metabolites are used as signaling molecules.

2.5 Conclusions

In this work we have shown that in metabolic networks motifs can be uniquely characterized by their enzymatic associations and therefore, their biochemical functionality. Further, we found that similarities in enzyme class proportions are explained by similarity in the structural features of the motifs. We also showed that cellular organelles display motif distributions that are distinct from one another and likely reflect their distinct metabolic roles in the cell.

Enzyme Commission numbers allowed us to uncover motif specificity at the chemical level, and pathway data allowed us to supplement the chemical information within a biological context. We were able to make inferences about higher-level biological function based solely on the structure of metabolic networks as described through motif distributions. This analysis demonstrates that network properties contain functional information that can be used to describe differences in metabolism between organisms.

The work presented here constitutes the first brush towards the goal of understanding metabolic network features across many forms of life. In the following chapter, we present an exploratory analysis if the *E. coli* metabolic network and assess the property of failure tolerance.

CHAPTER III

The metabolic network of *E. coli* displays distinct properties compared to Erdős-Rényi random networks

3.1 Introduction

In chapter II we characterized each 3-node motif using enzyme class associations and demonstrated that each organelle had a unique distribution of motif abundances. In this chapter, we present an exploratory analysis to begin to address why motifs contain evolutionary information, and what beneficial properties motifs may confer to an organism. Our goal in this work is two-fold. First we are interested in characterizing the network properties of the *E. coli* metabolic network. Second, we are interested in assessing the effect of motif participation on the robustness of the network to random component failure.

3.1.1 Motifs display many dynamic properties

The primary motivation for identifying motifs in biological networks is that their presence may provide insight into the organizational properties and evolutionary processes that gave rise to them [6, 81]. Significantly enriched motifs in biological networks imply positive selection pressure in favor of that motif. Likewise a suppressed motif could be detrimental to the survival of the organism and therefore be under negative selection pressure. These hypotheses are partially supported by the observation that topological features of biological networks differ markedly from those of random or synthetic networks [66]. While evolutionary reasoning is intuitive, characterizing the behaviors and beneficial features of motifs has proven a challenging task and remains an area of active research. The difficulty of this question is exacerbated by the need to select appropriate outcome measures with which to assess the behaviors and benefits of motifs. For instance, Ma *et al.* evaluated the network property of *adaptation* (or *stability*), which is a system's ability to respond to changes in inputs and then return to its pre-perturbed output level, even when the change persists [63]. Using a joint sensitivity and precision criterion, the authors found that incoherent feed-forward loops, loops in which one input is an activator and one is a suppressor, displayed adaptation more robustly than negative feedback loops, loops in which all inputs are suppressors. This result suggested that the configuration of the incoherent feed-forward loop was more sensitive to changes in input and therefore behaved with greater precision than other feed-forward loop configurations. Further, Prill *et al.* found that motif abundance was correlated with motif stability in metabolic, signaling and neuronal networks [81]. Prill found that the most stable motifs were the V-Out, V-In, 3-Chain and Feed-Forward Loop (motifs 1, 2, 3 and 7) motifs and the least stable were the Mutual V, Mutual and 3-Chain, 3-Loop, Semi-clique, and Clique (motifs 6, 11, 8, 12, and 13) motifs.

The ability of motifs to reliably transmit information has also been assessed. Klemm *et al.* found that motifs with feed-forward structures (motifs 1, 2, 3 and 7) always displayed reliable dynamics, while the least reliable motif is the 3-loop (motif 8) [55]. Like Prill *et al.* [81], Klemm *et al.* additionally found that motif abundance was correlated with the level of reliability so that those motifs that were most reliable were also most expressed.

Studies like those summarized above have been met with criticism. Ingram *et al.* argued that studies like those of Ma and Prill are too limited in scope and that the dynamic behaviors of motifs are actually broad and complex [48]. Further, Doyle and Csete argue that it is unclear that stability is even a property of the individual elements in biological networks and may instead be an emergent global property of networks [27]. If stability and robustness are primarily global properties, it is not obvious why these measures would be of interest towards the goal of characterizing motifs.

Despite these critiques, many have found that metabolic networks show a high degree of relatedness to one another but display motif distributions that are unlike those of other biological networks [33, 97]. This observation suggests that the design principles of metabolic networks may not be comparable to those of other biological networks like signal transduction or genetic networks. Van Nes *et al.* demonstrated that when the stability analysis of Prill [81] was applied to metabolic networks, structurally stable motifs were not enriched as previously reported [97]. The uniqueness of metabolic networks compared with other biological networks necessitates an exploratory analysis of basic network properties before progress in the characterization of metabolic motifs can be made.

3.1.2 Static properties of network motifs have not be adequately explored

Each of the above mentioned studies characterized dynamic behaviors of particular motifs. In general, dynamic systems are difficult to study because they require a great deal of data to inform parameter values in addition to knowledge about network structure. In order to characterize stability, it is necessary to supply rate parameters, initial conditions, and various levels of perturbation. In metabolism many of these parameters are simply not known and must be assumed, posing a major limitation to the generalizability of the results [48]. An alternative method that has not been well-studied is to characterize the static graphical features that are conferred by motifs.

When employed, static methods of analysis have uncovered many interesting features of biological networks. For example, Michoel *et al.* developed a network motif aggregation statistic to quantitatively measure whether modules in protein-protein interaction, post-translational modification and transcriptional regulatory networks were the result of the aggregation of network motifs [68]. They found that the feed-forward loop structure significantly aggregated into modules. Further, Basler *et al.* found the clustering coefficient in metabolic networks to be under positive selective pressure and not purely the result of thermodynamic constraints [9].

In this study, we expand upon the static analysis of Mirzasoleiman and Jalili [71] and assess local and global network properties following random component failure in the metabolic network of *E. coli.* In previous work, Mirzasoleiman and Jalili demonstrated that destruction of edges in protein and neuronal networks resulted in alterations to the motif distributions in those networks. We expand on their methodology and measure whether participation in network motifs causes the metabolic network of *E. coli* to be more resistant to random edges failures. Further, to determine whether the metabolic network responded to component failure in a manner distinct from random networks, we compared it with 1,000 Erdős-Rényi random graphs of equal node and edge size.

3.1.3 Hypotheses

We hypothesized that motifs that displayed enrichment in the metabolic network of *E. coli* as identified in Aim 1 will have the property of increased failure tolerance (green box in figure 3.1). Likewise, those motifs that were suppressed will display decreased failure tolerance (red box in figure 3.1). Specifically, we hypothesize that motifs 2, 3, and 7 - 13 will show increased failure tolerance, while motifs 1, 4, 5, and 6 will show decreased failure tolerance.

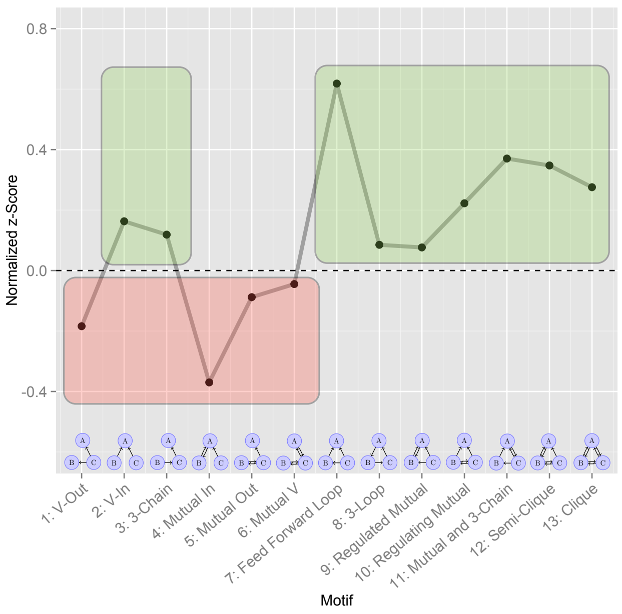


Figure 3.1: Motif profile of the *E. coli* metabolic network. Green boxes enclose motifs that are enriched in comparison with 1,000 random graphs, and red boxes enclose the motifs that are suppressed in comparison to 1,000 random graphs.

3.2 Methods

We used the *E. coli* metabolic network reconstruction version iAF1260 [36] as the network to analyze because of its relative simplicity in terms of cellular compartments and because of *E. coli*'s prevalence as a model organism in experimental research. The methods associated with the transformation of the stoichiometric matrix, motif mining with FANMOD, and determination of enrichment/suppression of network motifs can be found in the methods section in chapter II.

3.2.1 Network Parameters

3.2.1.1 Global Network Parameters

Global network parameters are those that describe features of the entire network rather than individual nodes or edges. We measured five global network parameters: Size, global transitivity, diameter, average path length and the power-law fit. *Size* is the total number of nodes in the network before and after edge destruction. *Global transitivity*, also called the *global clustering coefficient*, is the average ratio of fully connected triads (triangles) to connected triads [7], and corresponds to the level of connection between three nodes. The clustering coefficient ranges between 0 and 1, where 0 denotes no clustering and 1 denotes full clustering.

The *diameter* of the graph is the maximum shortest path (geodesic). To find the diameter, we first find the shortest paths from all nodes to all other nodes in the network and the largest of these is defined as the diameter (Figure 3.2 A). A similar measure is the *average path length*, which is measured by computing the shortest paths between all pairs of nodes and averaging them.

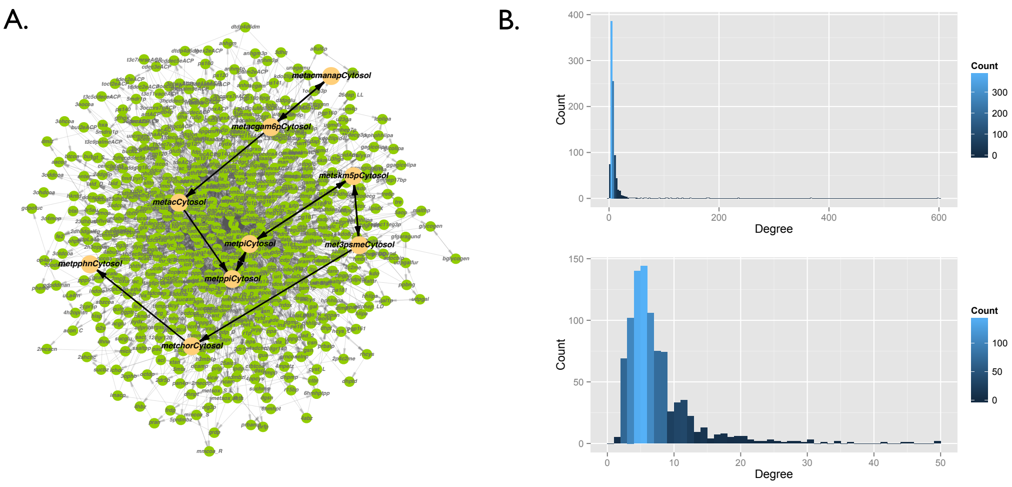


Figure 3.2: The diameter and degree distribution of the *E. coli* metabolic network. (A.) The diameter is denoted by yellow nodes connected with enlarged arrows. (B.) The complete degree distribution is displayed on the upper half of panel B. The color scale on the right indicates the density of points within the region. The majority of nodes have a degree near the average, while relatively few nodes have extremely large degrees. The bottom half of panel B contains more detail of the degree distribution below the extreme values.

Finally, the *power-law fit* is an estimate of the parameter γ which describes the exponential decay of the degree distribution:

$$P(K) \sim ck^{-\gamma} \tag{3.1}$$

Biological networks typically have scale-free degree distributions in which have most nodes have small degree, while few nodes have extremely high degree. The *E. coli* metabolic network has a scale-free distribution (Figure 3.2 B) with an average degree of 10 and a maximum degree of 598.

3.2.1.2 Local Network Parameters

Local network parameters are measurements that characterize the structure of a network at the level of the individual node or edge. We measured nine local network properties to assess failure tolerance in the context of motif participation. Of primary interest were the shortest out-paths and shortest in-paths which are measures of distances between nodes. The *shortest out-path* is the average shortest path length (or geodesic) from a reference node to all other nodes in the graph. Likewise, the *shortest in-path* measures the average shortest paths into a reference node from all other nodes in the graph.

Additionally we calculated the degree, closeness and betweenness centralities of each node. *Degree centrality* is the total number of edges connected to a given node. This measure was further divided into *in-* and *out-degrees*, which are the number of edges going into a node and the number of edges emanating from a node, respectively. Degree can be thought of as a measure of a node's popularity in a network [75].

Closeness centrality is the inverse of the mean length of the shortest paths between all nodes in a network [41]. The more central a node, the smaller its average distance to other nodes, and thus the larger its closeness centrality. Closeness centrality is a measure of how quickly information can spread from a given node to others in the network [75].

Betweenness centrality is a measure of the number of shortest paths from all nodes in a graph to all other nodes that pass through a particular node of interest [14]. Betweenness centrality captures the influence a particular node has on the network's ability to spread information [75]. In the context of biochemical reaction networks, metabolites with high betweenness centrality are intermediates that connect potentially disjoint reaction pathways.

Finally, we measured two features that quantify the amount of local clustering in the *E. coli* metabolic network: the local clustering coefficient, and Burt's constraint. The *local clustering coefficient* is defined similarly to the global clustering coefficient. It is the ratio of complete triangles to connected triads. *Burt's constraint* is a measure that captures the extent to which a node connects modules that are not otherwise linked [15]. A constraint value near zero indicates that a particular node is a bridge. The combination of low constraint and high betweenness centrality indicate bridging.

3.2.2 Generation of Random Graphs

One thousand Erdős-Rényi (ER) random graphs were generated for comparison with the *E. coli* metabolic network using the library "iGraph" in R. To insure that the random graphs were comparable, we simulated them to be identical in the number of nodes and edges.

3.2.3 Edge Destruction

To assess the feature of failure tolerance, we randomly destroyed the edges of the *E. coli* network and measured the effect on many graphical parameters. The biological analogue of edge destruction is the random removal of a set of metabolic reactions. We destroyed 1 to 10% of the edges and found that results did not change appreciably within this range (see appendix B). Destroying above 10% of the edges resulted in deterioration of the graph structure making the network structure unrealistic and inference difficult. Since there was no appreciable difference in the results in the 1 to 10% range, we restrict our analysis to the 10% edge destructions.

Destruction was conducted by randomly sampling the edges without replacement and deleting the sampled edges from the graph 10,000 times. Graphical parameters were then averaged over these 10,000 runs.

To determine statistical significance of the differences between graphical parameters before and after edge destruction, we calculated the mean and standard error of each measurement with 1,000 bootstrap re-samples. Using the bootstrap estimated standard errors, we calculated a 95% confidence interval for each network parameter.

All analyses in this work were conducted in R (version 2.15.1) with the library "iGraph."

3.3 Results

3.3.1 Global properties significantly vary between the *E. coli* metabolic network and the ER random networks

The global clustering coefficent, diameter, mean path length and power law fit were statistically significantly different between the *E. coli* metabolic network and ER random networks in the case of 0% edge destruction (Table 3.1). As expected, the *E. coli* metabolic network displayed higher levels of clustering compared (0.044) to random (0.01), but lower levels compared to the values reported in the literature, which typically range between 0.21 [46] and 0.40 [49]. The cause of this departure is apparent in the distribution of the clustering coefficients (Figure 3.3). The *E. coli* metabolic network has a large representation of metabolites with coefficients of zero. These metabolites correspond to endpoint metabolites that reside on the perimeter of the network (Figure B.5). Ignoring nodes with zero values, the mean clustering

Table 3.1: Comparison of Erdős-Rényi random graphs with the metabolic network of *E. coli* in the baseline condition and following 10% edge destruction

	0% Edge Destruction				**10% Edge Destruction**			
	ER Random Network		*E. coli* Metabolic Network		ER Random Network		*E. coli* Metabolic Network	
Parameter	Mean	95% CI	Mean	95% CI	Mean	95% CI	Mean	95% CI
Network Size	935		935		934.86	(934.83, 934.89)	933.594	(933.571, 933.617)
Clustering Coefficient	0.01	(0.010, 0.011)	0.044		0.009	(0.009, 0.01)	0.041	(0.0418, 0.0419)
Diameter	9.30	(9.26, 9.33)	8		10.21	(10.15, 10.26)	9.17	(9.15, 9.18)
Mean Path Length	4.47	(4.471, 4.473)	3.15		4.8	(4.752, 4.755)	3.29	(3.28, 3.30)
Power Law Fit	1.45	(1.44, 1.46)	1.39		1.40	(1.39, 1.41)	1.425	(1.42, 1.43)

coefficient in the *E. coli* metabolic network is 0.35, a value consistent with previous findings.

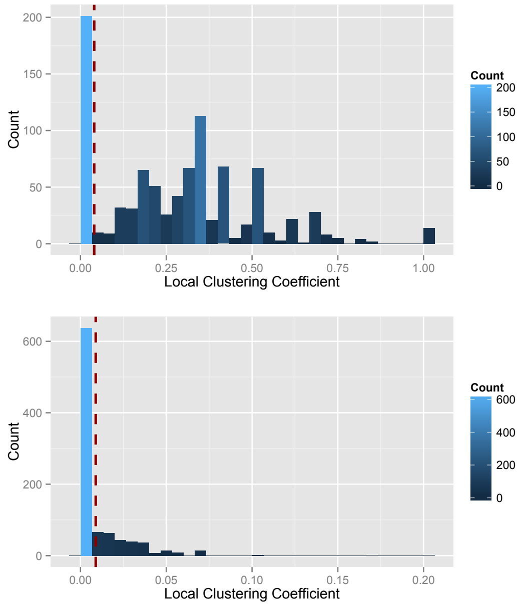


Figure 3.3: Comparison of the clustering coefficients of the *E. coli* metabolic network (top) and ER random networks (bottom). Red dashed lines indicate the mean values, and the blue intensities indicate the density of points contained in the bins.

The network diameter and mean path length were significantly higher in the random network than in the *E. coli* metabolic network. This result is consistent with the finding that ER random networks do not cluster and as a result have low edge density. The lack of edge density increases the length of the paths between nodes in random networks.

The degree distributions between the two network types varied dramatically (Figure 3.4). The *E. coli* metabolic network displayed a scale-free degree distribution typical of biological networks. Scale-free biological networks are characterized by the presence of nodes with degrees that greatly exceed the average degree, often acting as hubs connecting separate modules [39, 51]. Conversely, the degree distribution of the ER random networks displayed a bell shape curve and no nodes with extreme

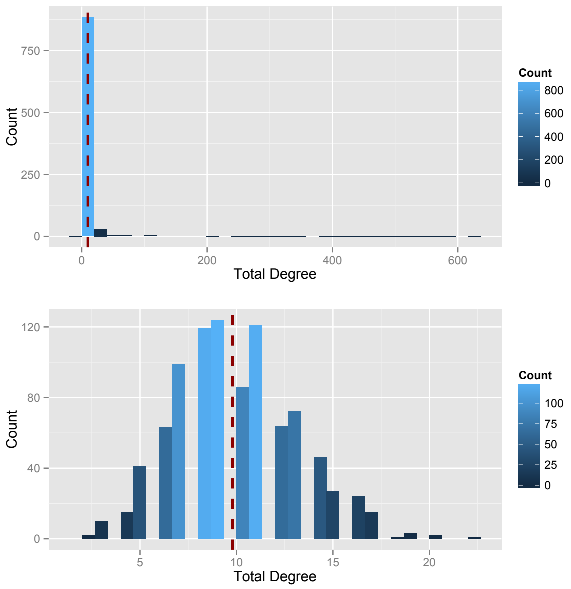


Figure 3.4: Comparison of the degree distributions of the *E. coli* metabolic network (top) and ER random networks (top). Both networks have an average degree of 9.80.

degree.

3.3.2 The *E. coli* metabolic network and ER random networks reveal similar levels of failure tolerance in global network parameters

In general, the two network types displayed similar levels of failure tolerance (Table 3.1). Both networks showed significant reductions in size following 10% edge destruction. Further, the clustering coefficient was reduced by 9% in the random networks, but only 7% in the metabolic network.

Conversely, the average diameter was increased by 15% in the metabolic network, but only 10% in the random networks. This is likely due to the inherent modularity of the *E. coli* metabolic network that implies a collection of hub nodes that serve as bridges. If a path to or from a hub node is destroyed, the ability to reach all other nodes in a graph could be severely reduced and increase the diameter.

These results suggest that while the topology of the *E. coli* metabolic network, as measured through the clustering coefficient and power law fit, was more tolerant to failure than the random networks, the diameter and path lengths and therefore ability of the network to disseminate metabolic information, were less tolerant to failure than the random networks.

3.3.3 Perturbation of Local Features

In the previous sections we established that the *E. coli* metabolic network and the ER random networks have fundamentally different structures. They vary in their levels of clustering, degree distributions and diameters. We are now interested in determining whether participation in particular network motifs confers the benefit of tolerance to component failure independently of network structure. To accomplish this, we evaluated local, node-level network properties in the context of motif participation before and after edge destruction.

3.3.4 The *E. coli* metabolic network displays higher levels of failure tolerance in local network parameters

To assess whether a graphical parameter displayed failure tolerance, it is necessary to evaluate the statistical significance of the mean value of the parameter before and after edge destruction. Recall that we are interested in assessing whether enriched motifs (motifs 2, 3 and 7 through 13) display higher levels of failure tolerance compared to suppressed motifs (motifs 1, and 4 through 6).

All motifs except the 3-Loop (motif 8) and Clique (motif 13) motifs displayed statistically significantly higher means in both their shortest out- and shortest in-paths (Figure B.3) following edge destruction. This result was also found in the ER random networks with every motif significantly higher following edge destruction except for the Semi-clique motif (motif 12).

There was no relationship between motif participation and the change in shortest

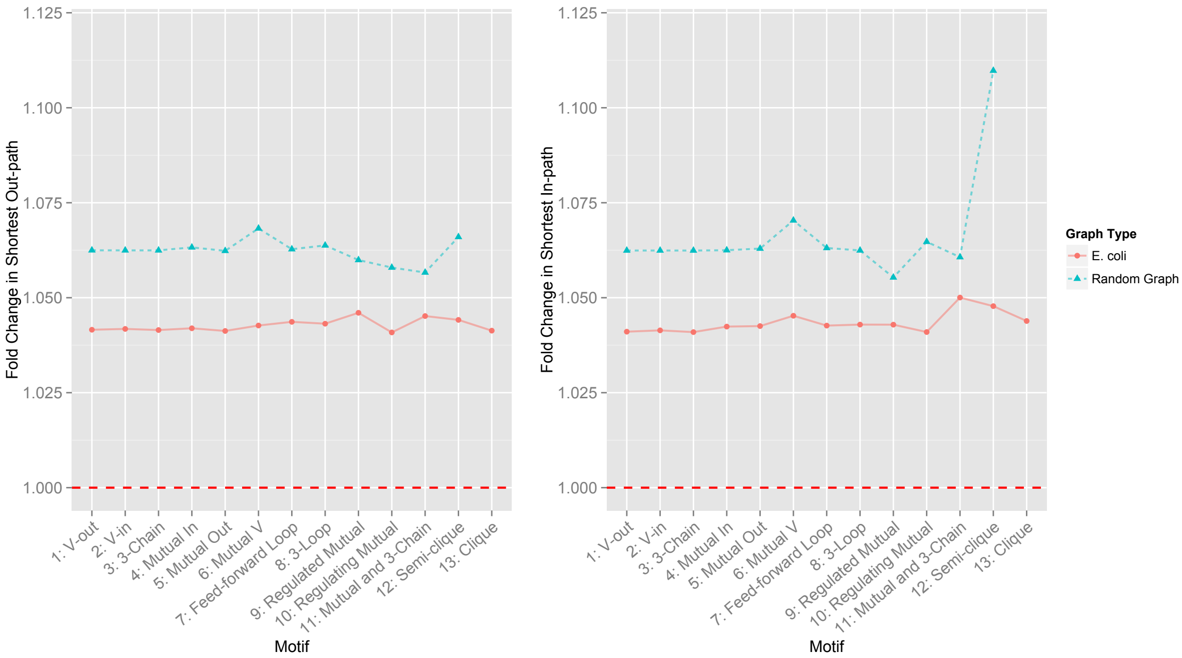


Figure 3.5: Fold changes in shortest out- and in-paths by motif. The red dashed line indicates the expected fold change if there is no change following 10% edge destruction.

paths following edge destruction (Figure 3.5). However, the *E. coli* metabolic network displayed significantly smaller increases in path length compared to the random networks. On average, the metabolic network saw increases in path length of 4%, while the random networks increased by 6%.

There was a positive relationship between degree and motif participation (Figure B.4) in the *E. coli* metabolic network. In the metabolic network, the average degree for the first five motifs was approximately ten. For motifs 6 through 13, the average degree was 23. The motifs with the highest degree were 3-Loop (motif 8) and Clique (motif 13) with degrees of 36 and 40, respectively. This finding demonstrates that motifs with greater edge density do not necessarily have greater degree. For example, motif 8 has three edges and a degree of 36 while motifs 4 and 5 have three edges but degrees of 10 and 11, respectively. Further, the relationship between motifs and degree was not seen in the random networks, suggesting that degree is an

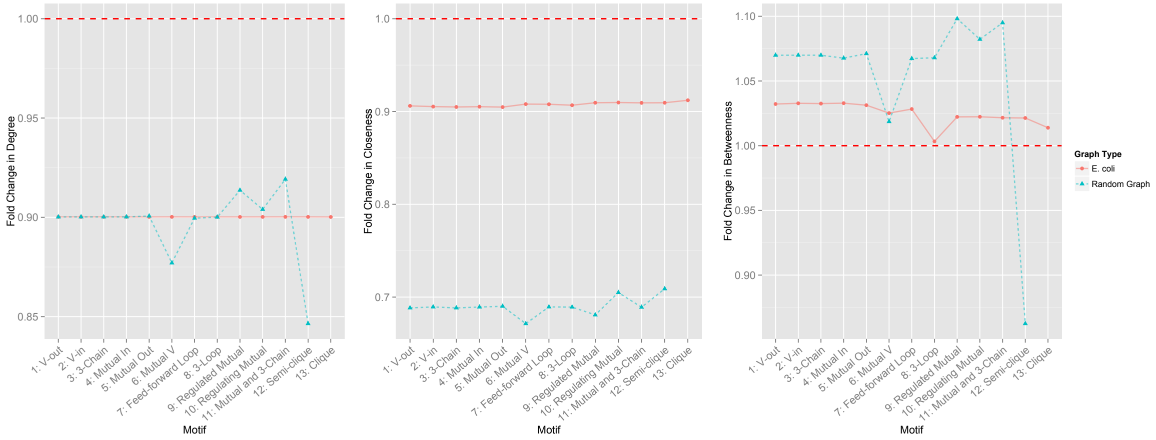


Figure 3.6: Fold changes in three centrality measures by motif. The red dashed line indicates the expected fold change if there is no change following 10% edge destruction.

indication not of motif complexity, but of network structure.

In both the *E. coli* metabolic network and the ER random networks, degree was reduced by approximately 10% following edge destruction (Figure 3.6). The effect of edge destruction on both types of networks was additive. That is, the change was constant across all motifs. The metabolic network and random networks displayed equal reductions in degree on average.

Closeness centrality decreased by 9% in the metabolic network and approximately 30% in the random networks following edge destruction (Figure 3.6). This indicates higher levels of failure tolerance, in the context of closeness, in the metabolic network (Figure B.4). Again, there was no relationship between motif participation and the change in closeness.

The fold change of betweenness centrality displayed a relationship with motif participation. Specifically in the metabolic network, the 3-Loop (motif 8) and Clique (motif 13) motifs were more tolerant to failure than the other 11 motifs. The Mutual V (motif 6) and Semi-clique (motif 12) motifs showed increased failure tolerance in the random networks. In general, the metabolic network was more tolerant to destruction than the random networks.

Finally, we discovered several interesting relationships by evaluating two related network parameters, the clustering coefficient and Burt's constraint. The clustering coefficient was largest in the Feed-forward Loop (motif 7), and Mutual and 3-Chain (motif 11) motifs (Figure 3.7). This result is consistent with previous findings that the Feed-forward Loop motif aggregates in biological networks creating dense modules [68]. The Mutual and 3-Chain (motif 11) motif is similar in structure to the Feed-forward loop but with an additional reversible edge. We also found a positive trend of increasing clustering coefficients with increased edge density in both the metabolic and random networks (Figure 3.7). In the random networks, motifs 1 through 6 had an average clustering coefficient of 0.01 and 0.03 for motifs 7 though 12. In general, this relationship was repeated in the metabolic network except in the case of motifs 8 and 13 which had relatively small clustering coefficients.

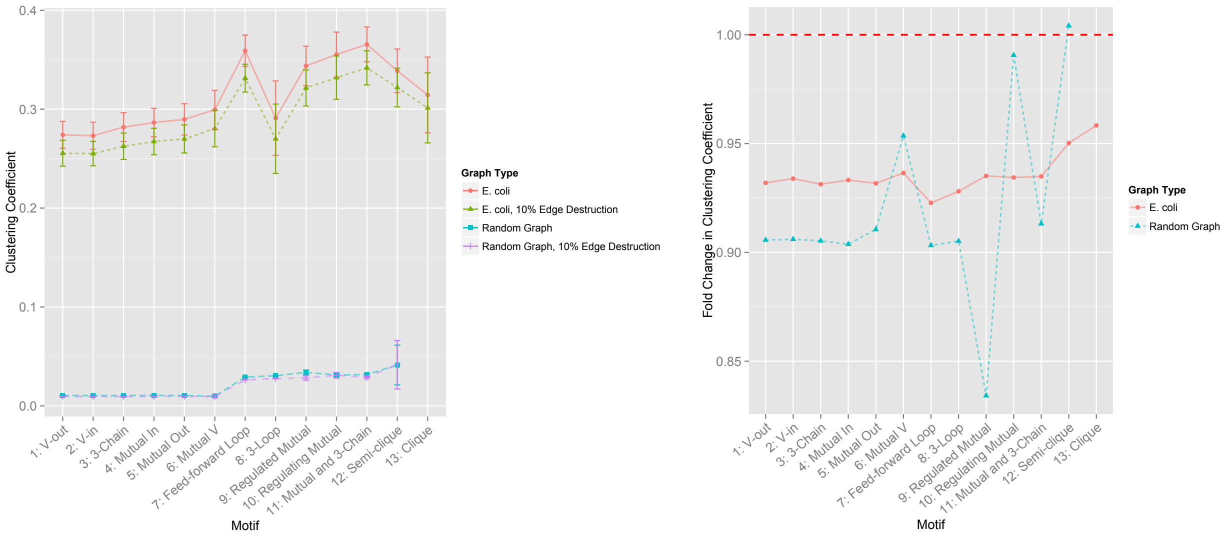


Figure 3.7: Fold changes in clustering coefficient by motif. The red dashed line indicates the expected fold change if there is no change following 10% edge destruction.

We observed a modest relationship between participation in the Semi-clique (motif 12) and Clique (motif 13) motifs and increased failure tolerance of the clustering coefficient. In the metabolic network, participation in the Semi-clique motif resulted

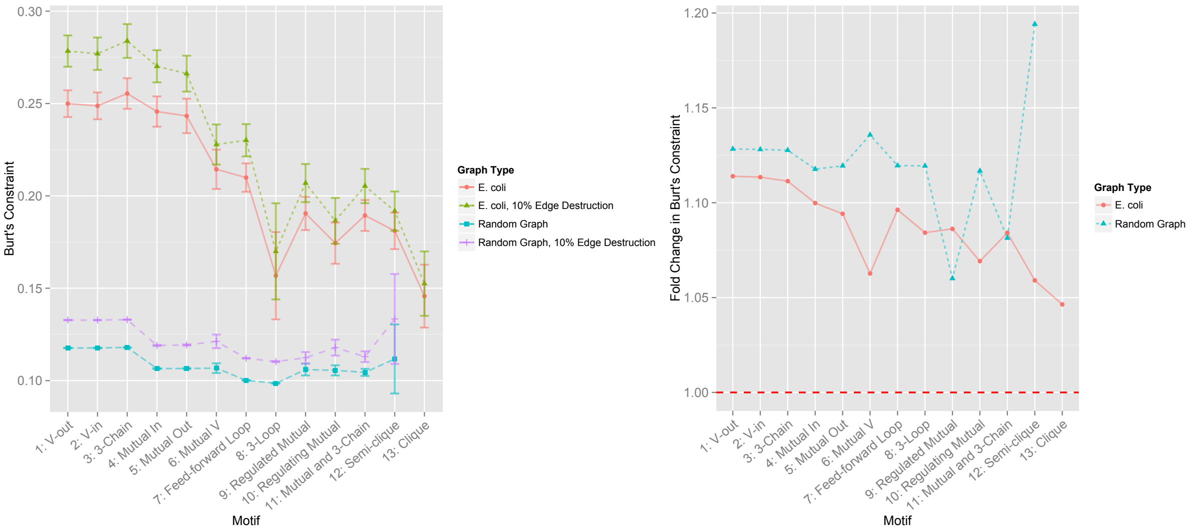


Figure 3.8: Fold changes in Burt's constraint by motif. The red dashed line indicates the expected fold change if there is no change following 10% edge destruction.

in a reduction in the clustering coefficient of only 5% compared with an average of 7% in the other motifs. Likewise, in the random network, participation of motif 12 resulted in an increase of 0.04% compared to an average decrease of 9% in the other 11 motifs. Again, the metabolic network was more tolerant to failure in general, and showed an average reduction in the clustering coefficient of 6% compared to an average reduction of 8% in the ER random networks.

Burt's constraint is a measure that indicates the extent to which a particular node is a bridge connecting clusters that would otherwise not be connected. A low constraint indicates that a node is a bridge. We found that the 3-Loop and Clique motifs (motifs 8 and 13) had the smallest constraints, suggesting that these motifs in particular have specific structural roles in the metabolic network (Figure 3.8). Recall that motifs 8 and 13 also had the largest betweenness centralities, demonstrating that they have more shortest paths passing through their nodes than any other motif structure. When taken together these results indicate that the metabolites participating in motifs 8 and 13 are structural metabolites connecting disparate metabolic

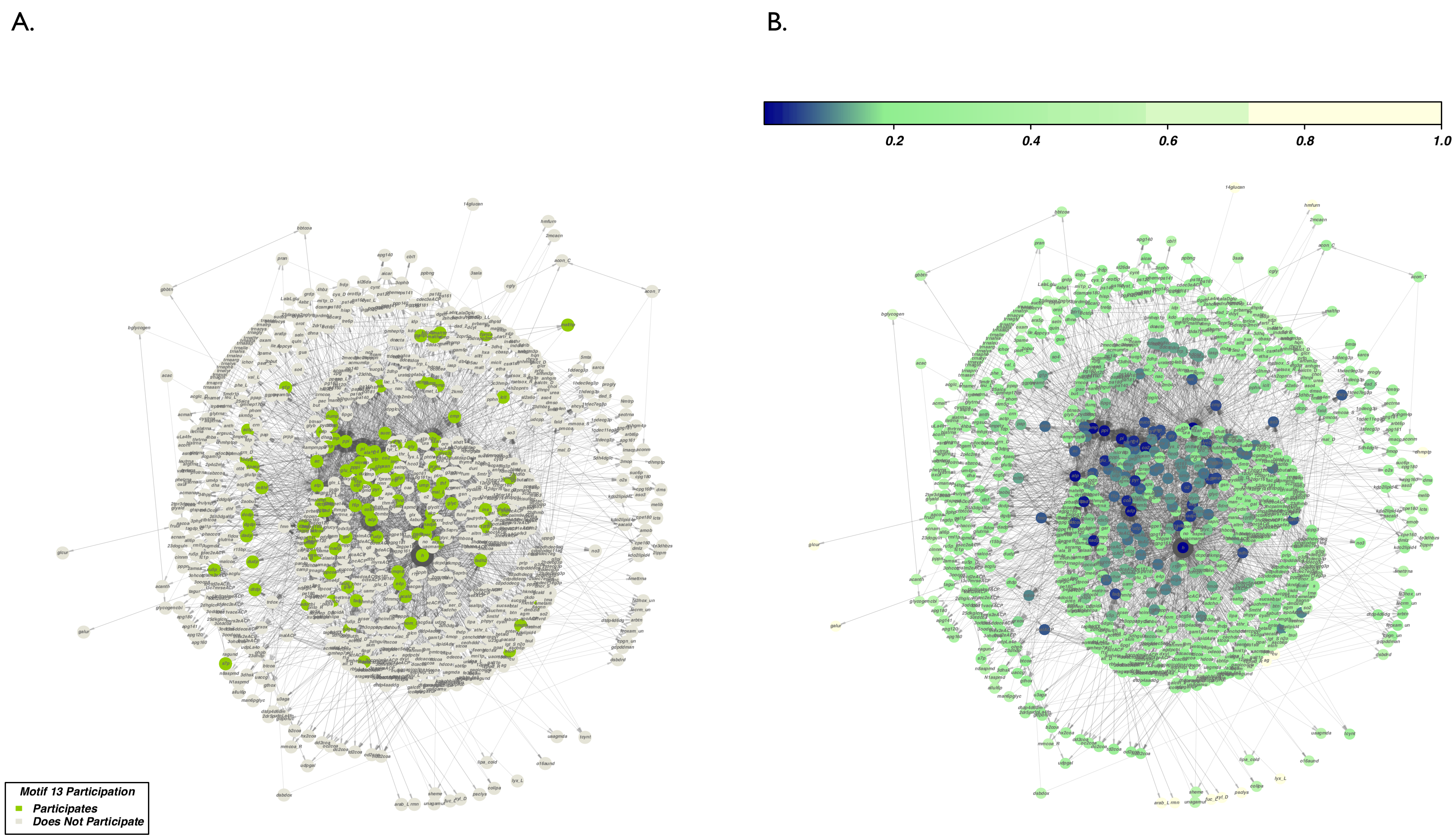


Figure 3.9: (A) Distribution of Clique motifs and (B) values of Burt's constraint in the *E. coli* metabolic network.

modules or pathways.

This structural property is visible from the distribution of the Clique (motif 13) motif in the metabolic network and the distribution of constraint values. Metabolites that participate in the Clique motif do not aggregate, but instead appear uniformly throughout the primary module of the network (Figure 3.9 A). The nodes that participate in the Clique motif also have the lowest constraint values (Figure 3.9 B).

In the *E. coli* metabolic network there was a downward relationship between the fold change in the constraint measurement and the motifs of greater edge density (Figure 3.8). The motifs with the lowest constraint values also tended to have the most tolerance to failure.

3.4 Discussion

In this study we showed that, contrary to our hypothesis, participation in particular motifs did not, in general, have an effect on failure tolerance in response to 10% edge destruction. However, we discovered that the metabolic network of *E. coli* displayed higher levels of failure tolerance overall compared to the ER random networks. Further, we discovered that certain motifs, specifically the 3-Loop (motif 8) and Clique (motif 13) motifs, have specific structural roles in the *E. coli* metabolic network.

As mentioned by Barabási and Oltvai, a major impediment to the characterization of the dynamic properties of motifs is that they never occur in isolation [6]. This limitation applies to the present static analysis as well. Our mean and standard error calculations were based on participation in a motif of interest, but ignored whether a metabolite simultaneously participated in other motifs. This is an unavoidable limitation in the study of real, biological networks due to their high levels of edge density and clustering. It is a challenge to find metabolites that only participate in one type of motif, and impossible to find metabolites that only participate in that motif once. In future work, additional simulation studies could be conducted to rigorously characterize the relationship between component failure and motif participation, but even then it would be unclear if the results would be generalizable to real, biological networks.

For simplicity our analysis was restricted to the metabolic network of *E. coli*, but this analysis could be expanded to include the 21 organisms and organelles investigated in the analysis of Aim 1. A broader analysis would allow us to investigate whether our findings are generalizable to all metabolism.

3.4.1 The Feed-forward Loop, 3-Loop and Clique motifs indicate structural network properties

Although we did not find evidence to support our hypothesis that enriched motifs would show higher levels of failure tolerance, we did uncover interesting structural features of the *E. coli* metabolic network.

First, we found that the clustering coefficient of the Feed-forward Loop motif (motif 7) was statistically significantly higher than all other motifs of equal edge density. This finding is consistent with previous work that found that the Feed-forward Loop aggregated into clusters [51, 68]. It is particularly striking when compared to the ER random networks where the Feed-forward loop again displayed a statistically significantly higher clustering coefficient compared to the first six motifs, but was no greater compared to the 3-Loop motif (motif 8) or motifs of greater edge density. The discrepancy between the metabolic network and random networks suggests that the Feed-forward Loop motif did not aggregate in the random networks but did aggregate in the metabolic network. It has been demonstrated that the transcriptional regulatory network of *E. coli* is enriched with Feed-forward Loop motifs that aggregate into modules and define the topological structure of the network [26]. Our findings suggest that the metabolic network of *E. coli* may share structural similarities with its transcriptional network, and that the structural features are responsible for its tolerance to failure.

Second, we found that the 3-Loop (motif 8) and Clique (motif 13) motifs bridged distinct clusters in the *E. coli* metabolic network. One explanatory example of this finding is the metabolite pyruvate. Pyruvate is a participant in both motifs 8 and 13 and is a known hub in metabolic networks. It is the end metabolite in glycolysis, the starting metabolite in gluconeogenesis, and can also be converted to alanine or to ethanol and is therefore a unifying metabolite. Beyond being a hub in the *E. coli* metabolic network, pyruvate has a small constraint value (0.04) indicating bridging behavior.

Previously, Prill *et al.* found that the 3-Loop and Clique motifs displayed the least stable dynamics and were the least abundant motifs in transcription, neuronal and signaling networks [81]. Consistent with that analysis, we found that the 3-Loop and Clique motifs are less abundant than the other 12 motifs, but are significantly enriched in the *E. coli* metabolic network compared to what could be expected at random. This indicates that these motifs are responsible for biological functionality, and in the case of metabolism that functionality is bridging between pathways. The uniqueness of the function of these two motifs suggests that a dynamic analysis that is appropriate for the other 11 motifs may be inappropriate for the 3-Loop and Clique motifs because they may behave in a fundamentally different manner. This point underscores the value of rigorously characterizing static network features before addressing the more challenging question of dynamic stability.

3.5 Conclusions

In this work we have shown that in the context of static network parameters, there is no relationship between motif participation and failure tolerance, but that the metabolic network of *E. coli* is more failure tolerant in general than Erdős-Rényi random networks. Further, we demonstrated that several motifs contribute structural features to the metabolic network of *E. coli*.

In the following chapter, we present an application of motif analysis to test hypotheses in organelle evolution and phylogeny.

CHAPTER IV

Metabolic network motifs provide novel evidence of the evolutionary origin of six Eukaryotic organelles

4.1 Introduction

In the previous chapters we demonstrated that metabolic motif distributions varied by enzyme class and organelle localization. Furthermore, we found that motifs could be uniquely characterized by the type of reactions in which they participate (chapter II). This finding suggests that, beyond network characteristics, motifs also contain biological information. We discovered the 3-Loop and Clique motifs are links that connect distinct metabolic modules (chapter III). In this chapter, we present an application of motif analysis to the question of the evolutionary origin of Eukaryotic organelles. We begin with a brief review of current theories, then present results and discussion.

4.1.1 The history of organelle evolution is convoluted by horizontal gene transfer

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* [85], a model of organelle evolution in which one microbe lives inside another giving rise to organelles found in modern Eukaryotes. This hypothesis has had support, however debate continues regarding the origin of nucleus [72], peroxisome [87], mitochondrion and even the host-cell that served as the venue for endosymbiotic events. 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 [4, 67]. 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 [4]. Similarly, the genome sequence of the red alga *Cyanidioschizon merolae* supports the hypothesis that plant plastids derived from a single endosymbiotic event [67].

Despite the successes of sequencing technology, genetic 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 [78]. There is a lack of consistency among alignment methods that extends beyond distantly related organisms and occurs even within well-studied organisms like mice [18]. Beyond these methodological concerns, horizontal gene transfer (HGT) makes interpretation of phylogenetic trees difficult because temporal relationships and ancestry are convoluted by the repeated exchange of genetic information [3, 43]. It has been estimated that as much as 90% of proteobacterial genes have been influenced by HGT at some point during their evolution making ancestry difficult to determine even within closely related lineages [56].

As noted by Christian de Duve [24], 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 [40, 77]. 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 [69, 70]. Eom *et al.* showed that the distributions of network motifs in 43 metabolic networks contained taxonomic meaning [33]. 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 (chapter II). These results are supported by the observation that many biological and engineered networks, for example yeast transcription networks and Internet linking structure, share global network properties, demonstrating that motif distributions contain key information about system-level organization [70].

In this study we present novel evidence for the origin of six Eukaryotic organelles by testing hypotheses of organelle origin using distributions of motif abundances. Specifically we are interested in comparing the distributions of Eukaryotic organelles with α-Proteobacteria and methanogenic Archaea because of their prominence in the literature [80]. An α-Proteobacterium is generally accepted to be the ancestor of modern mitochondria [4, 30, 34], while methanogenic Archaea are often hypothesized to be the source of the host cell or of the nucleus [65, 72]. Further, we propose a new methodology for constructing phylogenetic trees by incorporating metabolic signatures to pinpoint regions of genomically estimated phylogenies that may be spurious.

4.1.2 Hypotheses

Based on the finding from chapter II that motifs can be mapped to specific chemical and biological functions and therefore that motifs contain biological information, we hypothesize that organelles will display motif distributions (profiles) that are more similar to their ancestral microbes than to microbes that are non-ancestral. Further, we hypothesize that organelles derived from endosymbiosis will show a distinct profile compared to those organelles derived from membraneous infolding.

4.2 Methods

Many of the methods employed in this chapter are identical to those of chapter II. Specifically the construction of the graphs from metabolic network reconstructions, motif mining in FANMOD and the z-score methodology for determining enrichment of the motifs. The methods previously described will not be repeated here, and interested readers should refer to the methods section of chapter II.

4.2.1 Metabolic network reconstructions

In this application we are interested in testing specific mechanisms of organelle evolution, specifically those relating methanogenic Archaea and α-Proteobacteria. Because there are many more fully-sequenced genomes than there are genome-level metabolic network reconstructions, we employed the *in silico* method of Chen and Lin [19], Pipeline for Metabolic Nework Reconstruction (PEER), to generate a sufficient 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 (Table 4.2.1). To supplement the *in silico*-curated reconstructions, the manually-curated reconstructions used in chapter II were included and yielded a total of 11 α-, 10 β-, 10 δ-, 11 ϵ-, 13 γ-Proteobacteria and 12 methanogenic Archaeal metabolic networks.

Table 4.1: List of additional mircrobes included in Aim 3 and relevant network characteristics.

Species	Kingdom	Class	Nodes	Edges
A. caulinodans	Bacteria	α	1502	2520
A. avenae citrulli	Bacteria	β	1448	2394
A. baumannii	Bacteria	γ	1421	2309
A. dehalogenans	Bacteria	δ	1299	2196
A. butzleri	Bacteria	ϵ	981	1626
B. bacilliformis	Bacteria	α	933	1417
B. bacteriovorus	Bacteria	δ	1215	2039
B. japonicum	Bacteria	α	1765	3026
B. bronchiseptica	Bacteria	β	1481	2403
B. parapertussis	Bacteria	β	1448	2374
B. pertussis	Bacteria	β	1379	2185
B. petrii	Bacteria	β	1534	2554
B. aphidicola	Bacteria	γ	612	1108
B. cenocepacia	Bacteria	β	1824	3128
B. cepacia	Bacteria	β	1891	3227
B. mallei	Bacteria	β	1615	2730
B. multivorans	Bacteria	β	1687	2792
B. pseudomallei	Bacteria	β	1699	2883
C. concisus	Bacteria	ϵ	894	1488
C. curvus	Bacteria	ϵ	906	1488
C. fetus	Bacteria	ϵ	901	1492
C. hominis	Bacteria	ϵ	874	1438
C. jejuni	Bacteria	ϵ	850	1433
C. ruthia magnifica	Bacteria	γ	920	1550
C. crescentus	Bacteria	α	1522	2468
D. desulfuricans	Bacteria	δ	1024	1676
D. vulgaris	Bacteria	δ	786	1312
D. nodosus	Bacteria	γ	792	1336
E. ruminantium	Bacteria	α	690	1011
G. diazotrophicus	Bacteria	α	1436	2297
G. sulfurreducens	Bacteria	δ	1090	1815
H. acinonychis	Bacteria	ϵ	873	1434
H. hepaticus	Bacteria	ϵ	874	1464
M. loti	Bacteria	α	1750	3012
Methanobacterium Sp. AL-21	Archaea		807	1312
Methanobacterium Sp. SWAN1	Archaea		827	1338
M. smithii	Archaea		832	1374
M. maripaludis	Archaea		806	1340

M. vannielii	Archaea		784	1339
M. voltae	Archaea		734	1204
M. mazei	Archaea		921	1488
M. stadtmanae	Archaea		710	1114
M. hungatei	Archaea		830	1389
M. fervidus	Archaea		740	1209
M. xanthus	Bacteria	δ	1449	2469
Nitratiruptor	Bacteria	ϵ	931	1547
N. hamburgensis	Bacteria	α	1365	2310
P. carbinolicus	Bacteria	δ	1157	1946
P. propionicus	Bacteria	δ	1093	1820
P. aeruginosa	Bacteria	γ	1640	2772
R. felis	Bacteria	α	771	1041
S. cellulosum	Bacteria	δ	1605	2755
S. amazonensis	Bacteria	γ	1434	2442
S. putrefaciens	Bacteria	γ	1390	2354
S. meliloti	Bacteria	α	1658	2848
Sulfurovum	Bacteria	ϵ	1026	1723
S. aciditrophicus	Bacteria	δ	1024	1710
T. crunogena	Bacteria	γ	1067	1735
X. campestris	Bacteria	γ	1501	2522
Y. pestis	Bacteria	γ	1381	2369

The *in silico* reconstruction procedure has three primary steps: pre-reconstruction, automated curation and network revision. During pre-reconstruction, the pipeline identifies the collection of metabolic reactions known to exist in the organism with enzymes and proteins from genomic annotation. This first step produces a large pool of reactions that can be removed or modified in the following step.

The automated curation step uses growth conditions to modify and constrain the reactions proposed in the pre-reconstruction step. We curated nutrient data from the literature via Google Scholar using species names and the search phrases "culture conditions" and "isolation." Each culture media was recorded as well as growth temperatures, experimental conditions and pH. The automated reconstruction step then provides a series of alternative reconstructions by constraining the original reaction pool in light of the nutrient data.

Finally, network revision is performed to choose the "best" possible reconstruction for a given organism by minimizing a network revision penalty parameter. Each possible network revision, for example adding a reversible edge to a reaction mechanism, is assigned a penalty parameter. A linear solver then selects the combination of revisions that optimizes a particular criteria, such as biomass synthesis.

4.2.2 Determining statistically significant differences in enrichment

Similarly to chapter II, statistical significance was determined using 1,000 bootstrap estimates of the mean and standard error of the normalized z-scores for each organism and organelle. 95% CIs are indicated on each relevant figure and non-overlapping CIs indicate statistically significantly different enrichment.

4.2.3 Calculation of distance metrics

To quantify similarity of motif profiles we calculated a Canberra distance metric [61]. The Canberra metric is similar to a Euclidean or Manhattan distance, but is better suited for data that scatter around the origin as in the case here.

4.2.4 Construction of phylogenetic trees and calculation of leaf-wise distances

To avoid the pitfalls of whole-genome multiple alignment, genomic phylogenies were created using the alignment-free method of Feature Frequency Profiles (FFP) [93]. This method is appropriate when creating phylogenies from whole-genomes where the amount of homology is potentially low. Additionally, FFP requires fewer assumptions than multiple-alignment methods, which assume certain mutation rates, various amounts of HGT and the presence of homologous genes that may not exist [78].

FFP is a method similar to text comparison methods that work 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.

To employ FFP properly it is essential to select the appropriate l-mer resolution (length). This was done following the guidelines provided by the authors in the technical documentation. To find the lower limit, we count the number of features for each feature length l. The most abundant feature length, l_{Hmax}, is the lower limit of the optimal resolution. The maximum feature length, l_{CREmin}, can be found by estimating the relative entropy error between an $l-2$ Markov model and the observed frequency of a particular l-mer [92]. Special care is needed in this case because the size of mitochondrial genomes is much smaller than full nuclear 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. Trees created with l-mers within the optimal range are not expected to vary appreciably as demonstrated by bootstrap and jackknife resampling [93].

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) [38], 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 [13]. 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-dervied 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.

4.3 Results

Prior to analysis we mined for motifs in the metabolic networks of 72 species 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 [70] and are roughly in order of increasing edge density (Figure 4.1). 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% confidence intervals (Figure 4.1). Based on the previous finding that taxonomic relationships could be replicated using motif distributions [33], we expect that endosymbiotic organelles will display higher levels of similarity with their ancestor microbe compared to other, non-ancestral microbes.

4.3.1 Motif distributions provide novel evidence of the origin of Eukaryotic organelles.

The original host cell, within which serial endosymbiotic events are thought to have occurred, has been posited to be a methanogenic Archaen [65], a Bacterium or an early proto-Eukaryote [16, 79] 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 4.1). 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.

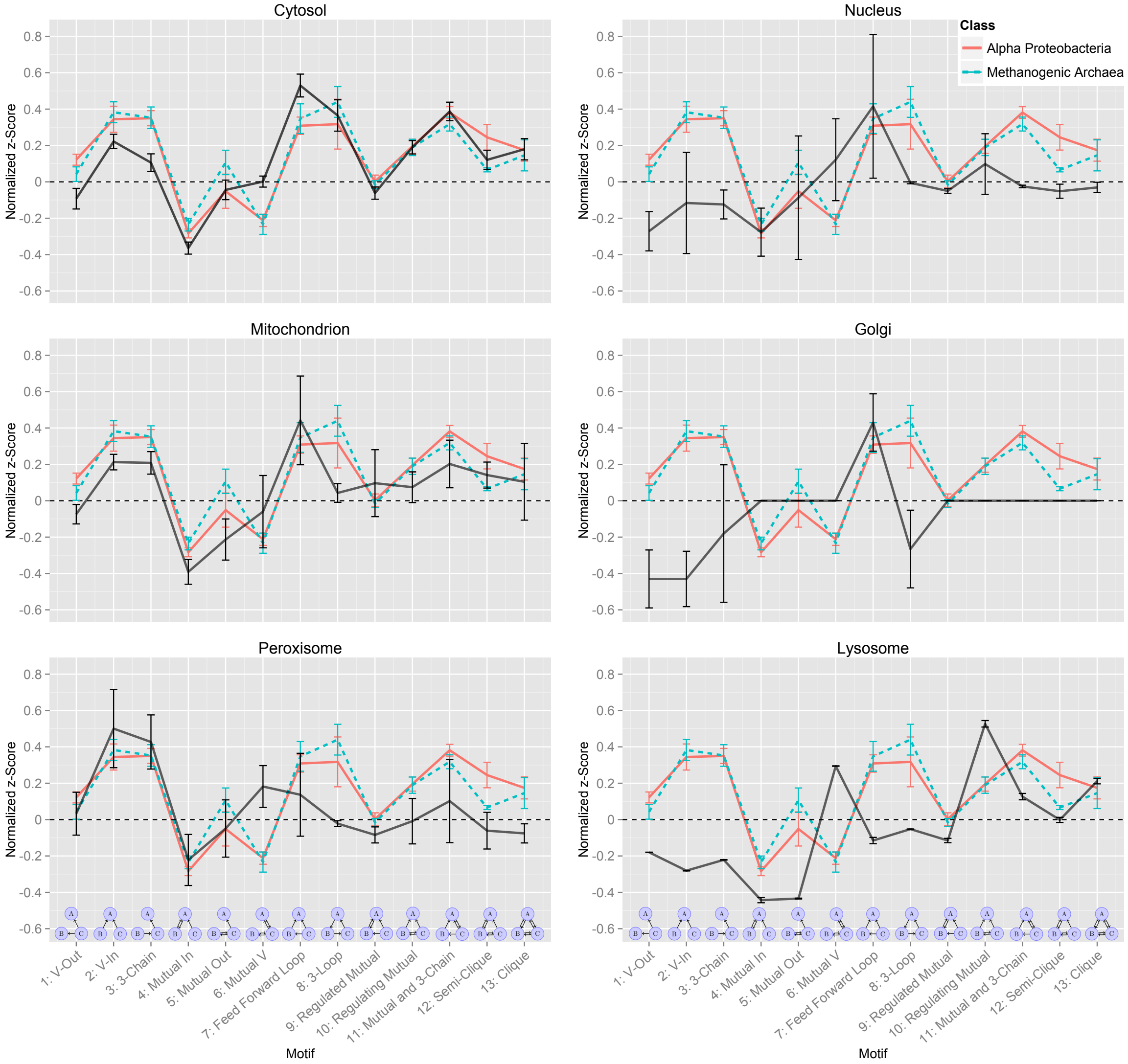


Figure 4.1: Motif distributions of six Eukaryotic organelles (solid black) compared with the profiles of α-Proteobacteria (solid red) and methanogenic Archaea (dashed blue). Non-overlapping confidence intervals indicate statistical significance at $p \leq 0.05$.

To test whether the host cell might instead have been derived from a microbe in a different family of proteobacteria, we compared the cytosolic motif profile with that of the β-, δ-, ϵ- and γ-Proteobacterial motif profiles (Figure 4.2). We found again that the proteobacterial motif profiles and cytosolic motif profile were relatively dissimilar, but that the cytosolic profile appeared most like that of a γ-Proteobacteria.

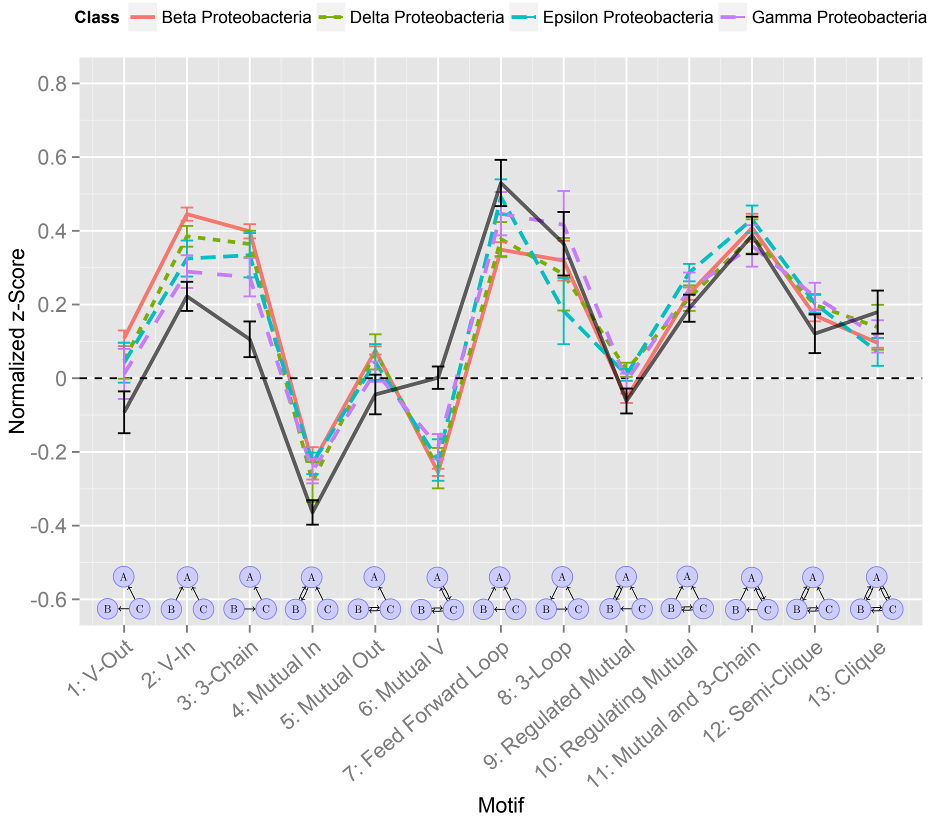


Figure 4.2: Comparison of the motif distributions of the Eukaryotic cytosol and four classes of proteobacteria, β, δ, ϵ and γ. The cytosolic profile most resembles a γ-proteobacterium and is only signifincantly different in four of the 13 motifs: 3-Chain, Mutual V, Regulated Mutual and Semi-clique.

The origin of the peroxisome is nebulous. Schlüter *et al.* argued that since the peroxisomal membrane is comprised entirely of Eukaryotic proteins, an endosymbiotic origin is unlikely [87]. 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 [23, 24]. 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 (Figure 4.1). Specifically, under-expression 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 suggest 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 (Figure 4.1). First, all three 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 supports the hypothesis that each of these organelles is derived from membranous infolding and not from endosymbiotic events. The consistency of these profiles, coupled with the fact that there is strong evidence that the Golgi is a product of membraneous infolding [73] suggests that both the nucleus and the lysosome are also membrane-derived organelles and not former endosymbionts.

Lastly, we find an apparent dissimilarity of the mitochondrial motif profile with that of the α-Proteobacteria (Figure 4.1). It is hypothesized that modern mitochondria are the product of an endosymbiotic event with a host cell and an early α-Proteobacterium [4, 30, 34]. However, the mitochondrial motif profile instead appears more like that of δ- or ϵ-Proteobacteria (Figure 4.3) rather than an α-Proteobacterium. Only four of the 13 motifs were significantly different compared to either δ- or ϵ-Proteobacteria, while six of 13 were significantly different compared to the α-Proteobacteria profile. This result is also apparent from the distances between the profiles (Figure 4.3, Panel A). The mitochondrial profile clusters with δ-

or ϵ-Proteobacteria while the α- or β- and γ-Proteobacteria cluster separately with each other again suggesting that mitochondria are metabolically akin to δ- or ϵ-Proteobacteria.

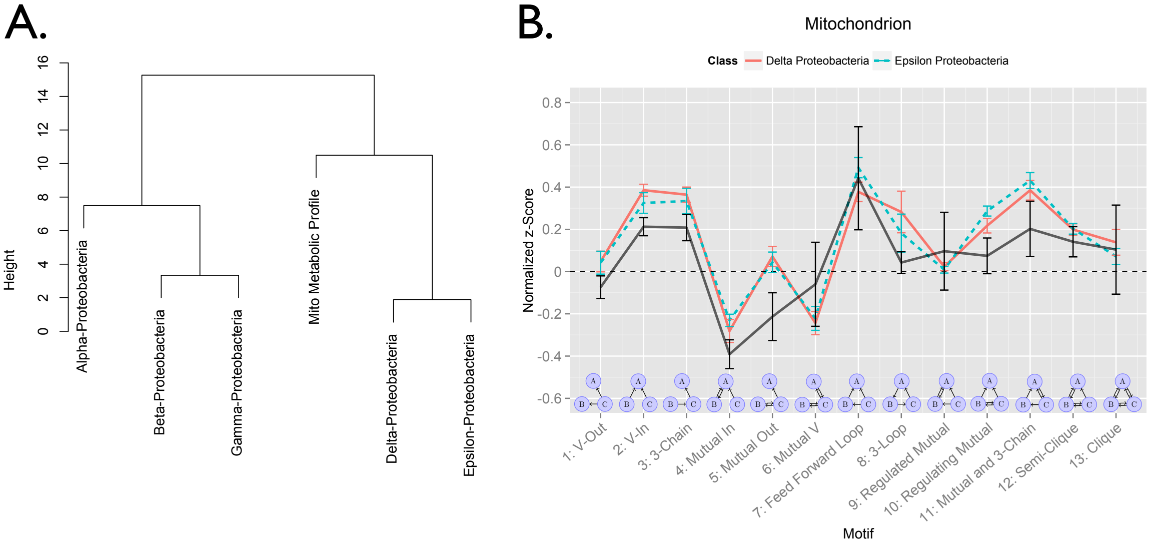


Figure 4.3: Comparison of Mitochondrial motif distribution with δ- and ϵ-Proteobacteria. Panel (A) contains a dendrogram of the distances between the motif profiles compared to the mitochondrial profile. The mitochondrial profile clusters most closely with δ- and ϵ-Proteobacteria profiles and not the α-Proteobacteria profile. Panel (B) contains the motif distributions of the mitochondrion (solid black), δ- (solid red) and ϵ-Proteobacteria (dashed blue).

4.3.2 Whole-genome phylogenies independently validate species similarities predicted with motif distributions

To validate 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 4.4). 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.

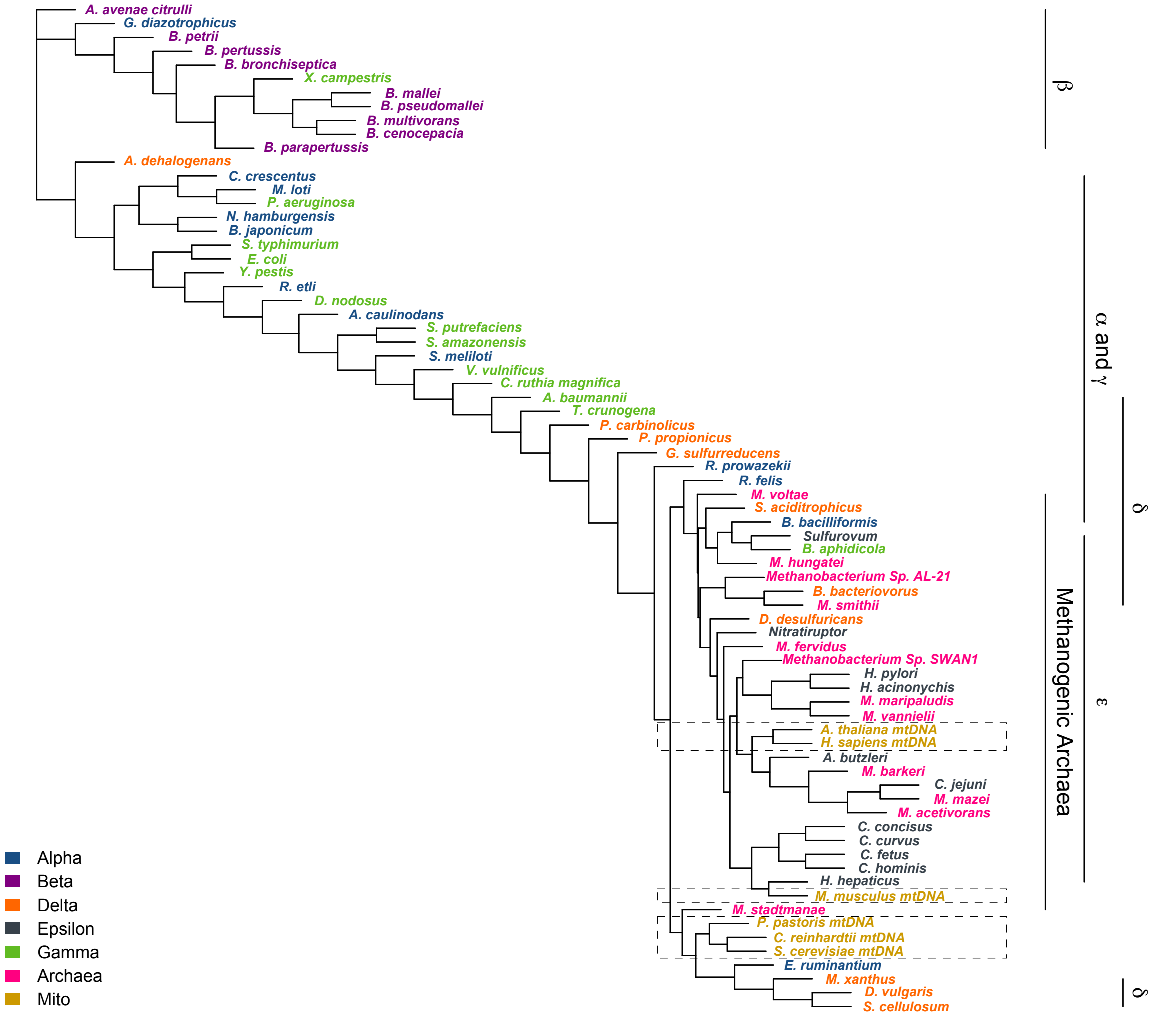


Figure 4.4: Genomic phylogeny created with feature frequency profiles. Each member of the proteobacterial family, Archaea and mitochondrial genome is colored and labeled on the right-hand side. Mitochondrial genomes are enclosed with dashed boxes for clarity. In general the mtDNA falls with the clades of the δ- and ϵ-Proteobacteria as well as the methanogenic Archaea.

4.3.3 Metabolic phylogenies provide an additional level of insight into ancetral relationships

Because of the difficulties inherent in the interpretation of phylogenetic relationships, we propose a second level of analysis to locate regions of trees where ancestral relationships might be spurious. To produce a clear example, we have reduced the species in this analysis from 67 to 18 to include only manually curated metabolic networks (Figure 4.5). 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.

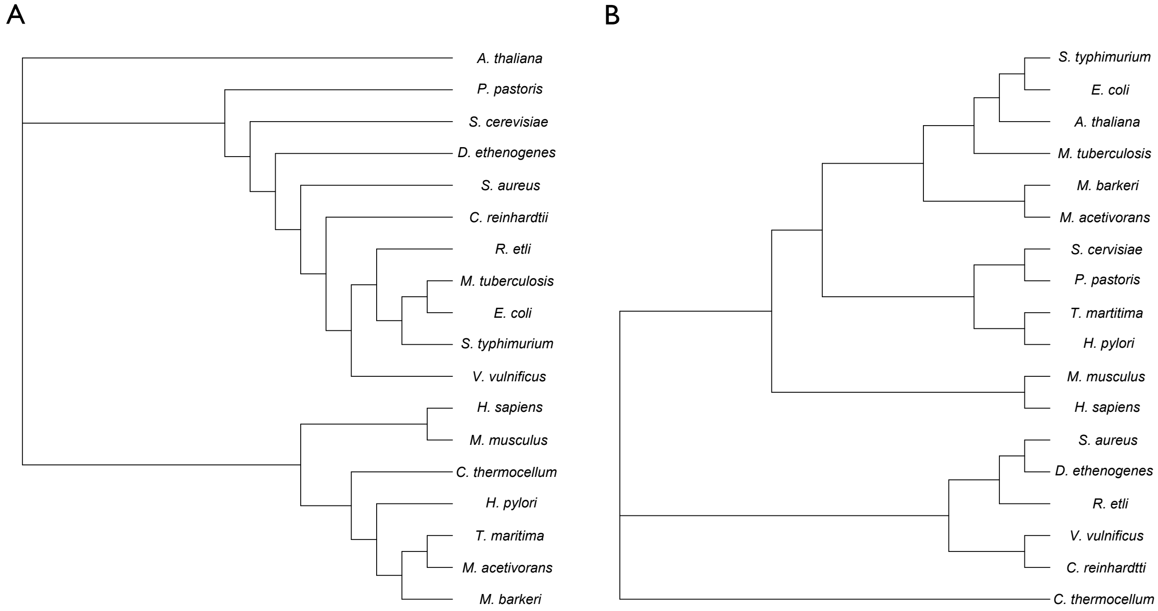


Figure 4.5: Phylogenies based on (A) whole genomes and (B) metabolic significance profiles.

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 [5]. For example, it is known that *T. maritima* shares approximately 24% of its genome with thermophilic bacteria [74], further demonstrating that HGT is largely a function of 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 [13]. The heatmap in Figure 4.6 shows the differences of the pairwise distances between the leafs of each of the phylogenies in Figure 4.5 A and B. The heatmap indicates the level of agreement in the leaf distances between the two phylogenies, that is, how similar the trees are to one another.

The genomic-derived phylogenetic relationship between two Archaea, *M. barkeri* and *M. acetivorans*, and the Bacteria, *T. maritima*, provides an example of a major limitation of phylogeny construction using genetics alone. *T. maritima* shares approximately 24% of its genome with Archaea [74, 80], 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. Using the metabolic phylogeny, *T. maritima* falls next to *H. pylori*, a fellow extremophile with which it likely shares stabilizing metabolic reaction pathways crucial for living in mutagenic conditions such as high temperatures or low pH. Conversely, *M. tuberculosis* has been shown to fall more closely to the Proteobacterial rather than Actinobacterial clade in previous work [54], a result that

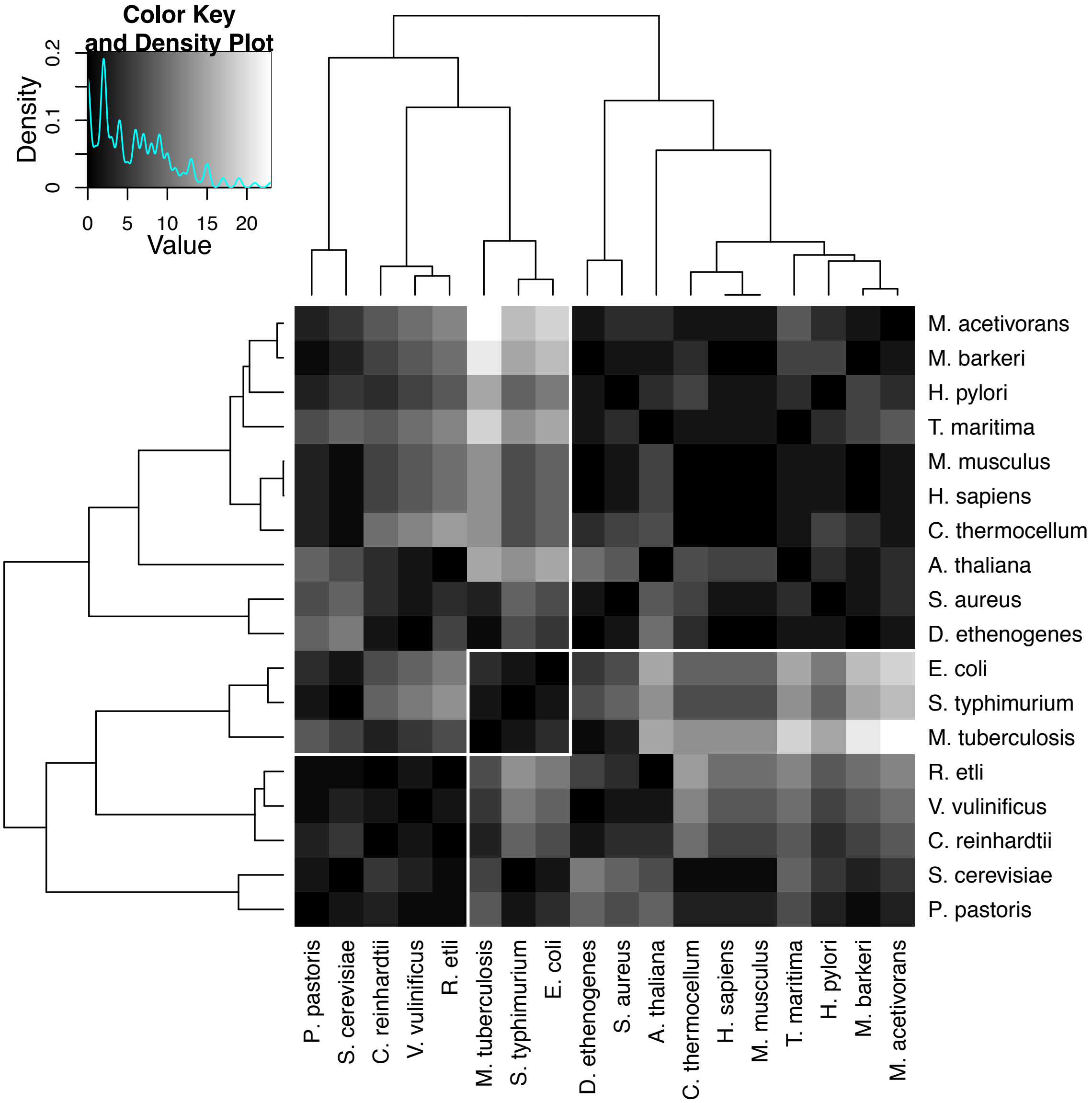


Figure 4.6: Heatmap of the differences between pairwise leaf distances. Differences near zero are shown in black and denote regions of relative agreement of the metabolic and genomic phylogenies. The areas of high agreement are enclosed in white boxes. Large differences are shown in white and denote regions of incongruence between the two trees.

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 than *M. tuberculosis* acquired many genes from early Proteobacteria via horizontal gene tranfer, or that Actinobacteria are actually more closely related to Proteobacteria than was initally 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 the integration of both phylogenies clarified a seemingly spurious relationship.

4.4 Discussion

In this work, we presented novel evidence of the origin of six Eukaryotic organelles by quantifying metabolic similarity with relative motif abundances. Further, we used these motif abundances to add an additional level of 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 organisms, 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. These organelles were never independently functioning organisms, and thus lack many of the reactions found in microbes. Instead they were evolved for isolating repetitive tasks, for instance recurrent glycosylation reactions in the Golgi and hydrolytic reactions in the lysosome.

The motif profile of the cytosolic compartment was unlike that of both α-Proteobacteria and methanogenic Archaea, which supports the hypothesis that the original host cell was a proto-Eukaryote. This proto-Eukaryote likely possessed many features of a modern Eukaryote, including a nucleus, endoplasmic reticulum, Golgi body and the cellular machinery necessary to engulf extracellular microbes.

It has been theorized that the nucleus may have been the first endosymbiont [47, 60]. For example, the syntrophic hypothesis for the origin of the nucleus describes a symbiotic relationship between a methanogenic Archaea and δ-Proteobacteria [72]. This mechanism involves a consortium of δ-Proteobacteria surrounding one methanogenic Archaea that eventually becomes engulfed to form the early nucleus. Others have postulated an ancestral Archaeon that acquired a Bacterial endosymbiont, although the cellular machinery required to acquire this Bacterial partner has never been demonstrated in any extant Archaea [24, 80] or Archaeal fossil evidence. A third hypothesized source of the Eukaryotic nucleus is from a vesicle created from in-folding of cellular membranes [60]. The in-folding mechanism is supported here due to the similarity of the nucleus motif profile to that of the Golgi and lysosome. Unlike the cytosol, mitochondria and peroxisome, the likely origin of the lysosome and Golgi is invagination and in-folding of plasma membranes. The formation of vesicles was evolutionarily advantageous because it allowed for internal digestion and isolation of complex reaction pathways [24].

Our results further support the hypothesis proposed by de Duve that peroxisomes were not a Eukaryotic product as has been suggested [87], but instead was an early endosymbiont which perhaps predated the mitochondrion. The metabolic functionality of peroxisomes and mitochondria overlap, which helps explain why the motif profile of the peroxisome looks more like that of the mitochondrion than the cytosol.

The postulation that mitochondria do not appear to be metabolically α-Proteobacterial in nature is undoubtedly controversial, however inspection of the literature reveals that there is little definitive evidence in support of the α-Proteobacterial endosymbiont model. For instance, two phylogenetic studies of yeast and other Eukaryotic mitochondrial proteins showed that, respectively, only 10 and 14% of mitochondrial proteins are attributable to α-Proteobacteria [30]. Additionally, phylogenetic analysis of several glycolytic enzymes in Eukaryota and Proteobacteria showed greater likeness in γ-Proteobacteria, or an ancestor of γ, than to α [32]. There is also temporal support for the notion that mitochondria are dervied from δ- or ϵ-Proteobacteria. The ϵ and δ group is thought to have evolved first [45], at roughly 2.85 billion years ago and the α at roughly 2.3 bya [10], giving the δ and ϵ groups more time to forge an endosymbiotic relationship. Finally, our phylogeny based on the FFP of whole genomes places mtDNA more closely with the δ and ϵ families, supporting the hypothesis that the early mitochondrion was not an α-proteobacteria.

Although the motif distributions alone are a single line of evidence corresponding to only one cellular system, we believe that our results justify further research into the relatively mysterious lineage of ancient δ- and ϵ-Proteobacteria as possible ancestors of modern mitochondria. We believe that these results provide robust evidence of organelle origin because they are based on known biological functionality and not on genetic information, which can be difficult to interpret and which does not always imply function. Further, through the complementary examples of *T. maritima* and *M. tuberculosis* we have shown that integrating a second level of cellular information with genomic-based phlyogenies can improve inferences of evolutionary relatedness and identify regions of trees where spurious connections may lie. Despite all this, it is true that the validity of the evidence presented here is absolutely dependent on the quality of the network reconstructions and the depth of knowledge of the reactions on which the reconstructions are built. A limitation of this work is the relative lack of representative organisms. There was only one plant and one protist available to represent their entire kingdoms, however this relative dearth can only shrink as more high-throughput data become available.

In closing, it is worth stating that the majority of the 18 reconstructions used in this research were assembled in different laboratories using various naming conventions and curation protocols. The consistency of the emergent properties of the organelles is remarkable and serves as a validation of the use of genome-level reconstructions for evaluating molecular evolution. Metabolic network reconstructions are revolutionizing the emerging field of comparative metabolomics in much the same way that high-throughtput sequencing technology revolutionized comparative genomics, and we have only just begun to uncover the role energy metabolism played in shaping the world as we know it.

4.5 Conclusions

In light of our novel metabolic data and consistent with previous work, we support the notion that a proto-Eukaryote containing a nucleus, which was not endoymbiotic in origin, predated the first endosymbionts. Further, due to the relatively divergent shape of the peroxisomal motif distribution, it seems likely that this organelle predated the mitochondrion.

CHAPTER V

Concluding Remarks

5.1 Thesis Summary

In chapter I we described the paradigm of systems biology as being the iterative process of modeling, validation, and refinement (Figure 1.1). We applied this framework throughout, with the ultimate goal of generating new insights about the evolutionary origin of the Eukaryotic cell by analyzing and comparing the abundance of metabolic network motifs between species and organelles.

In chapter II we analyzed 21 genome-level metabolic network reconstructions and demonstrated that motifs could be characterized by the ensemble of enzymes with which they were associated. Further, we found that motifs with structural similarities displayed similarity in enzyme association. We confirmed our hypothesis that each organelle would display a unique distribution of motif abundances, supporting the overarching hypothesis that network motifs can be used as a proxy for metabolic function. Finally, we used the enzymatic associations to highlight an example of how motif abundances could be applied to make inferences about biochemical functionality in mitochondria.

Building upon the findings of chapter II, in chapter III we conducted an exploratory analysis of the metabolic network of *E. coli.* We found that there was no relationship between motif participation and failure tolerance as measured by local network parameters such as centrality and shortest path-length in response to 10% edge destruction. However, we found that the *E. coli* metabolic network was more failure tolerant overall compared to Erdős-Rényi random networks. Further, we uncovered two structural properties of the *E. coli* metabolic network: clustering of the Feed-forward loop motif and bridging behavior in the 3-Loop and Clique motifs.

Finally, in chapter IV we applied the methodology of motif mining and analysis to test specific hypotheses of Eukaryotic organelle evolution. Specifically, we presented novel evidence suggesting that a δ- or ϵ-proteobacterium may have been the ancestor of modern mitochondria rather than an α-proteobacterium. We also concluded that, because the motif profile of the peroxisome was more like that of the cytosol and mitochondrion, the peroxisome was an endosymbiont and not derived from the cytosolic-ER matrix. We validated this finding using phylogenetic trees constructed from whole genomes and found that, consistent with the motif profiles, mitochondrial genomes tended to fall within the same clades as the δ- and ϵ-proteobacteria. This independent validation led us to the new hypothesis that modern mitochondria are not derived from α-proteobacteria, but are instead derived from a δ- or ϵ-proteobacteria.

The next step in the systems biological work-flow is to conduct new experiments to test the hypothesis that the ancestor of mitochondria is an δ- or ϵ-proteobacteria. This experimentation is outside of the scope of this work, but is an essential element that remains to be completed.

5.2 Future Work

In addition to experimental validation regarding the δ- and ϵ-proteobacteria, there are many opportunities to expand on the work presented in this thesis. Our goal was to thoroughly characterize three-node metabolic motifs, but these merely scratch the surface of potential analyses.

With FANMOD it is possible to mine for motifs of up to size eight, yielding thou-

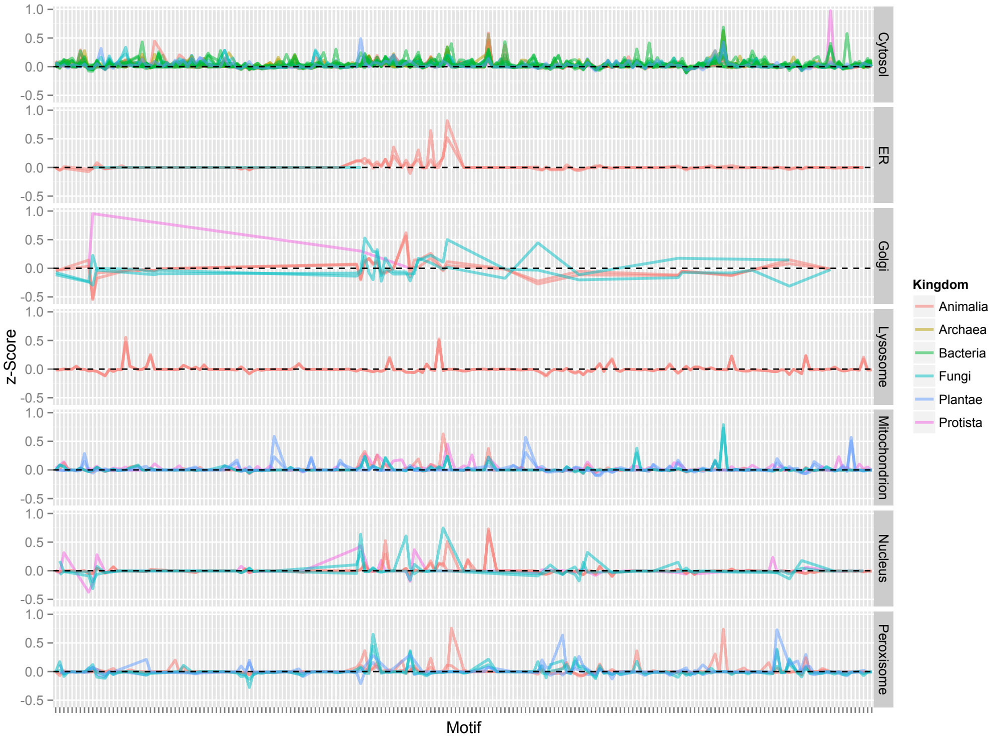


Figure 5.1: 4-node motif distributions in seven organelles.

sands of additional patterns to interrogate for biological meaning. Increasing the motif size by just one node creates a substantially more difficult analysis problem as can be seen from the 4-node motif distributions in the metabolic networks (Figure 5.1). There are 13 motifs of size three, 199 motifs of size four, 8,427 motifs of size five, and over 20,000 of size six. Increasing the node size substantially increases the number motifs to assess for statistical significance and results in many more patterns of enrichment. The dramatic increase in complexity resulting from increased motif size necessitates a more complicated analysis plan that can be built on the results of this thesis. Correction for multiple testing is a practical extension to the current analysis that would improve statistical inferences and help avoid false positives in enrichment. It would also be necessary to identify all motifs that are isoforms to insure that abundances were correctly calculated.

A second extension of this work is to investigate motifs with node and/or edge coloring. Motifs with edge coloring expand upon non-colored motifs by allowing motifs to contain structural *and* functional information. For instance, it is possible to color motif edges using stoichiometric values (Figure 5.2), which yields a more detailed biochemical relationship.

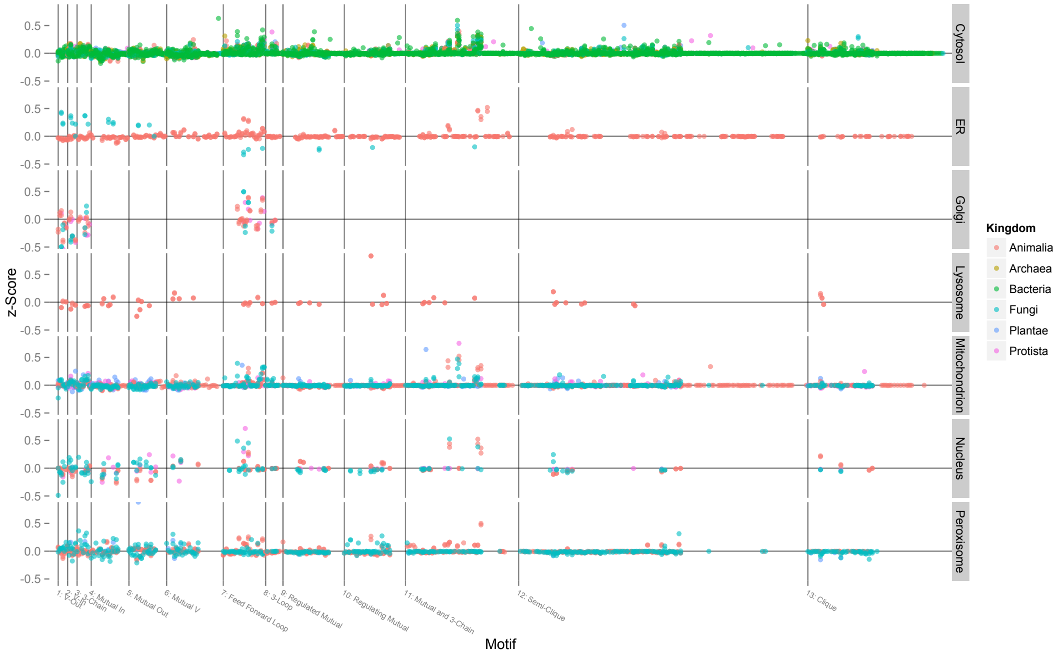


Figure 5.2: 3-node motifs with stoichiometric relationships encoded by edge coloring. Because these are 3-node motifs there are still only 13 *structural* motifs, but many more stoichiometrically *functional* motifs.

Qian *et al.* successfully applied the analysis of colored motifs to the *C. elegans* neural network by using node color to describe the neuron type [82]. They discovered that interneurons were over-represented within enriched motifs. This finding supported the hypothesis that the motifs transduce messages from sensor neurons towards muscles and thus control locomotion. Colored motifs combined two independent sources of information and allowed Qian to make insights that were not possible with either data source alone.

A parallel analysis could be done using metabolic networks and bipartite graphs. The work presented in this thesis dealt strictly with substrate graphs in which all nodes are metabolites, and all edges are reactive associations. The analysis could also be conducted using bipartite graphs in which nodes represent both metabolites and reactions. Bipartite graphs are more realistic in the sense that their structures denote actual biochemical reaction mechanisms, but they are more difficult to interpret. For instance, in generating random background for comparison of bipartite graphs it is possible to get a 3-node motif that contains two reaction nodes and one metabolite node. This motif structure does not correspond to a realistic biological situation, and convolutes interpretation of motif abundance.

All of the work presented here dealt with static network features, that is, features that do not change over time. While analyses of this nature are analytically more tractable than dynamic analyses, they do not provide a realistic representation of actual biological processes which change over time in response to inputs from their surroundings. There are many possible extensions of this work that could begin to describe the dynamic properties of metabolic networks. For example, it is likely that not all metabolic pathways in the network are continuously active. Thus, one analysis could integrate the metabolic network with measurements of gene expression levels. Through this analysis it would be possible to see how the structure of the genetically active portion of the network changes over time in response to a stimulus. The more layers of biological data we can integrate with the network, the greater the insights to be gained from further contextualizing the network structure in terms of its interactions with other cellular processes.

5.3 Implications

There are many important implications of this work. First, because motifs can be associated with metabolic enzymes, we were able to make inferences about possible higher-level biological function based solely on the structure of metabolic networks as described through motif distributions. This finding suggests that it is possible to assess metabolic similarity between two organisms by evaluating their relative abundances of 3-node motifs. That is, the motif distributions provided a simple, compact framework for assessing metabolic likeness that could be used to generate insights into complex biological relationships like microbial communities, parasitic behavior and symbiotic adaptation.

The characterization of 3-node motifs was limited by a lack of adequate species diversity of the metabolic network reconstructions, particularly in the kingdom of Eucarya. In order to fully understand the organizational properties of metabolism that may be shared among all life, it is essential that we increase attention to the development of methods for computationally constructing high-quality metabolic network reconstructions.

A goal in network research is the characterization of the roles of motifs and elucidation of the reasons that motif structures are selectively enriched. We found that two motifs, the 3-Loop and Clique motifs, that have been characterized as displaying unstable dynamics in the literature [55, 81] have the unique structural property of linkers between metabolic modules. The uniqueness of their role in metabolism implies that analyses of their dynamics in the cell should be similarly unique. A major implication of this work is that a one-size-fits-all approach to evaluating the dynamics of motifs in real, biological networks is inappropriate. This partially explains why dynamic studies like that of Prill *et al.* [81] have been unable to explain why motifs with unstable dynamics persist in biological networks at levels exceeding what could be expected by chance.

The results from our final aim have major implications in the field of evolutionary biology. They suggest that there is evidence to support alternative hypotheses in organelle evolution, specifically that the peroxisome was indeed an endosymbiont and that modern mitochondria are not derived from an α-proteobacterium. These findings suggest that new experimental research is necessary to unravel the history of all Eukaryotic organisms. Beyond the hypotheses generated, the use of motifs as a proxy of metabolic function is a novel method for combining functional information with traditional genetic methods. Using multiple levels of biological data improves our picture of reality and reduces biases in our conclusions that occur as the result of a limited cellular scope.

Using methods like mass spectrometry it is now possible to collect volumes of metabolomic data at resolutions that were unthinkable even five years ago. However, just like in the fledgling years of the genomic era, researchers in metabolism must step back and thoroughly elucidate the fundamental organizational and structural properties of metabolism. This work is meant to be a first brush towards that goal.

APPENDICES

APPENDIX A

Chapter 2 Supplement

A.1 Chapter 2 Supplemental Figures

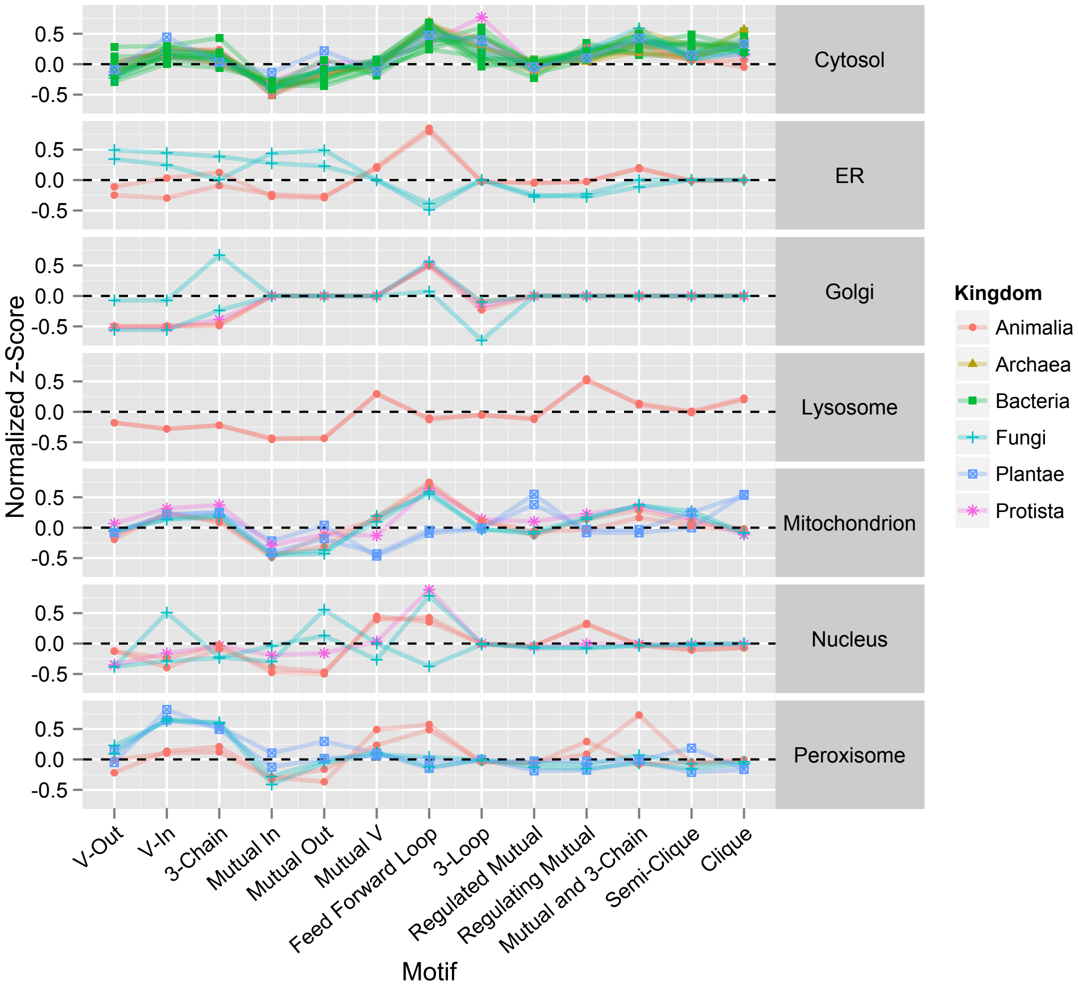


Figure A.1: Significance Profile of 3-node motifs by organelle. Line colors indicate the kingdom of life to which each organism belongs.

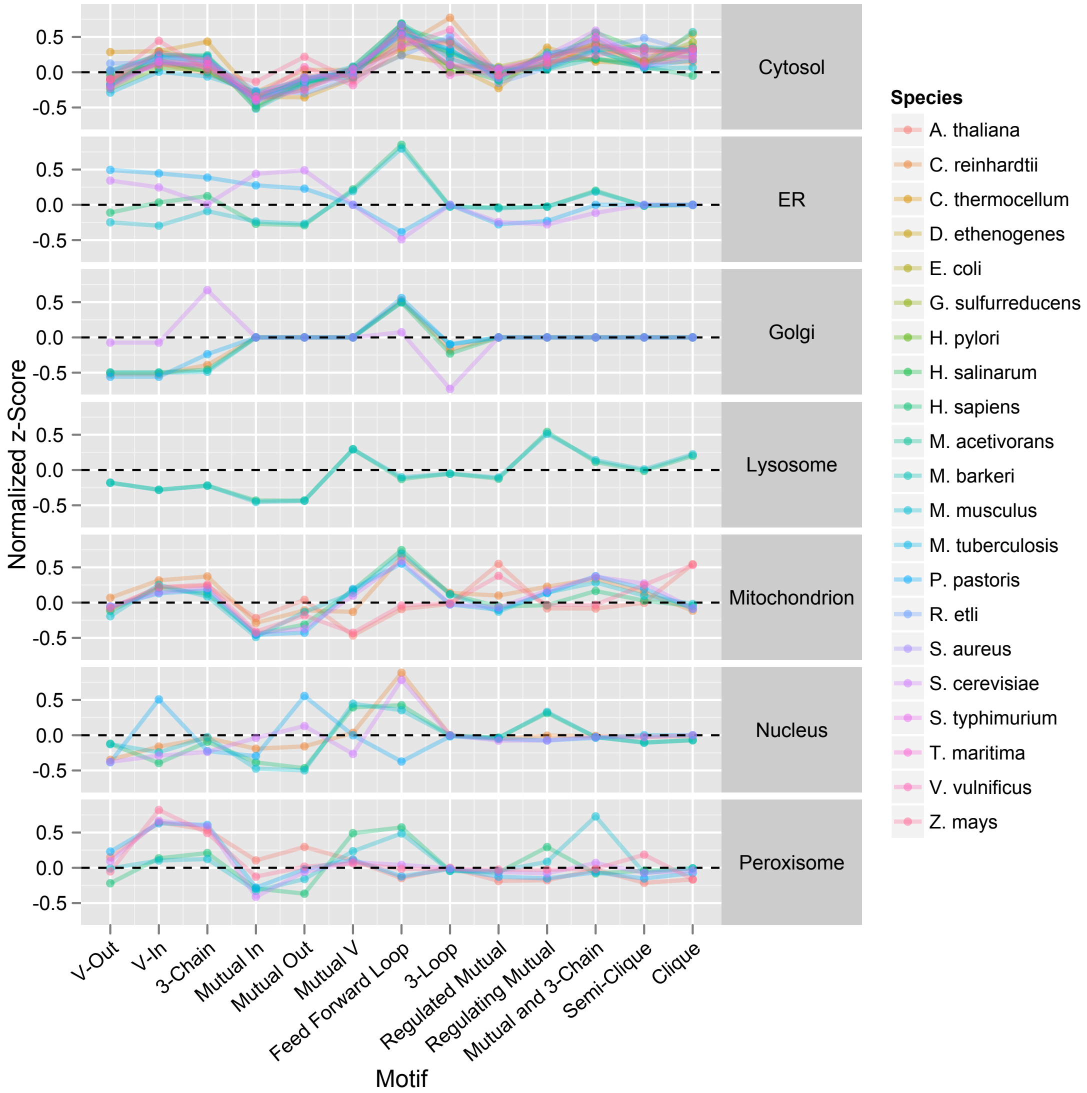


Figure A.2: blarp

APPENDIX B

Chapter 3 Supplement

B.1 Chapter 3 supplemental results

The graph size was relatively invariant to destruction of edges (Figure B.1). The intact *E. coli* graph had a total of 935 nodes and was only reduced by one node on average after 10% edge destruction.

The global transitivity decreased with each percent increase in edge destruction. At 10% edge destruction, the global transitivity was reduced by 5% from 0.044 to 0.042. The maximum reduction following edge destruction was approximately 83% suggesting that even with the destruction of 10% of the edges, the graph retains at least 83% of its clustering structure.

The diameters of the *E. coli* metabolic networks were relatively sensitive to edge destruction. The average fold change did not change until 5% edge destruction or higher, but the range of fold changes was wide. For example the average fold change at 4% edge destruction was the null value of 1, however some of the graphs had increases of diameter of as much as 50%. This was again seen at 10% edge destruction where the average change in diameter was a 13% increase, but ranged to as much as a 75%

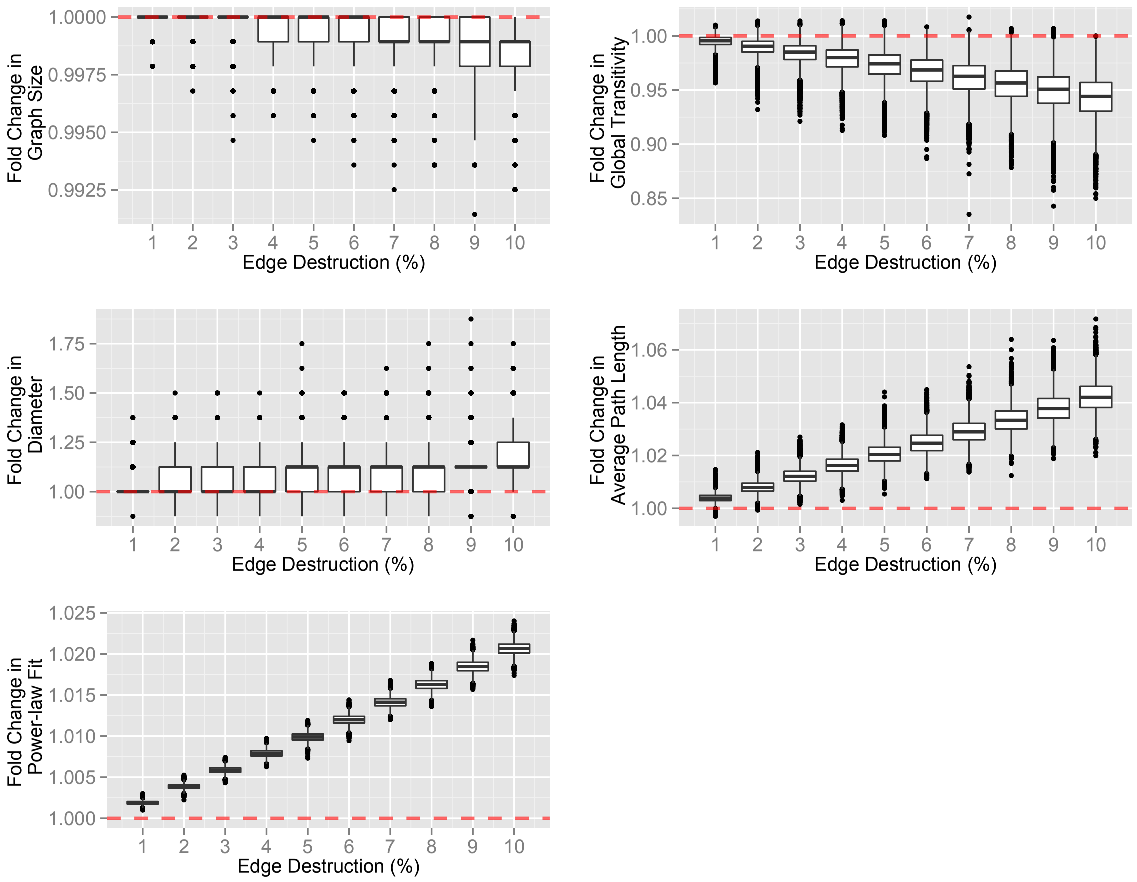


Figure B.1: Global graphical characteristics of the *E. coli* metabolic network reconstruction for varying levels of edge destruction.

increase. A 75% increase corresponds to a diameter of length 16 versus the original diameter of 9.

The average path length from a starting node to all other nodes in the network grew steadily on average with each percent increase in edge destruction. Despite the consistent increase the magnitude was actually quite small, with a maximum of approximately 7%. Like the graph size, this robustness to destruction is likely attributable to the fact that most nodes have at least 10 associated edges suggesting that they have alternate routes to all other nodes in the network.

Finally, the power law parameter γ increases with each percent increase in broken edges. Again the magnitude of the increases is very small with a maximum change of about 2.5%.

In addition to the global properties investigated above, we measured nine local, or node-level, properties (Figure B.2). The shortest-out and -in paths increased steadily with each percent increase in node destruction. There is a corresponding decrease in the closeness centrality with each percent increase in edge destruction. This is a logical result since closeness centrality is measured in terms of the shortest paths.

The mean betweenness centrality does not vary appreciably with increases in edge destruction, but the variance does.

It is interesting that for many local measures, for example degree, there appeared to be a threshold. The effect of breaking 1 or 2% of the edges was equal followed by a steady trend after 3% edge destruction.

B.1.1 Failure Tolerance

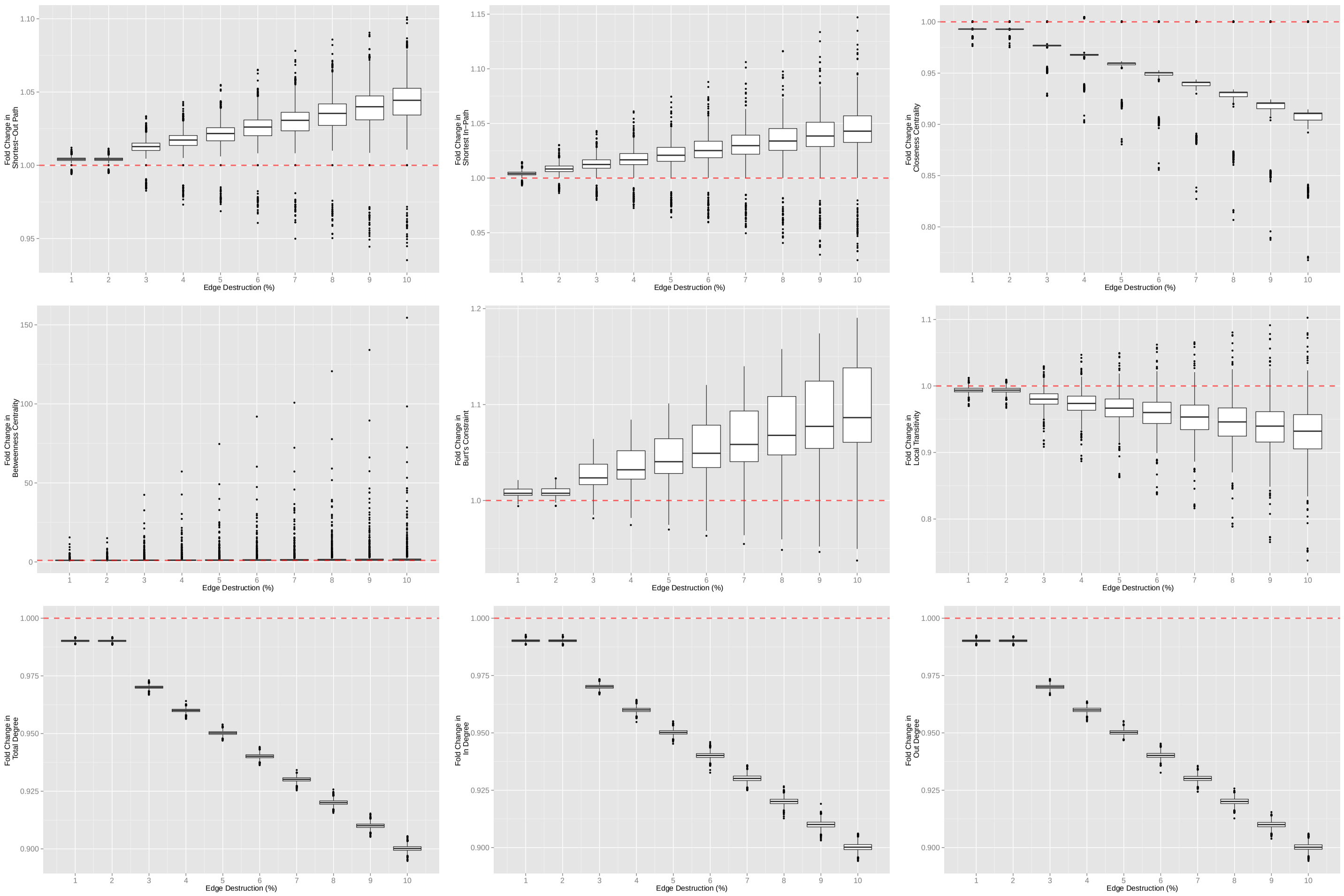


Figure B.2: Fold changes in local graphical characteristics of the *E. coli* metabolic network reconstruction for varying levels of edge destruction.

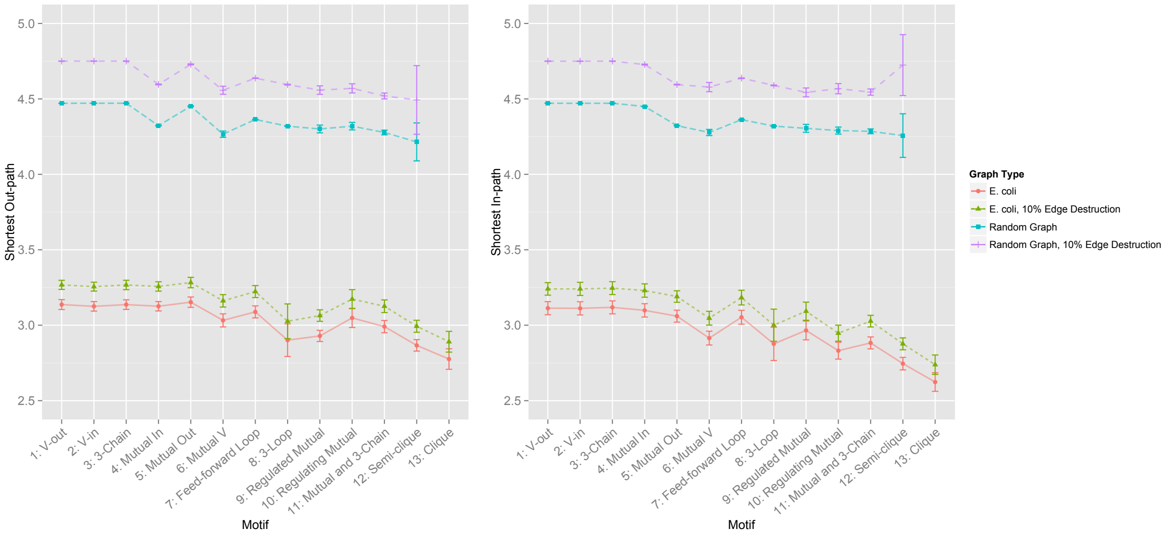


Figure B.3: Shortest Out- and In-paths by motif.

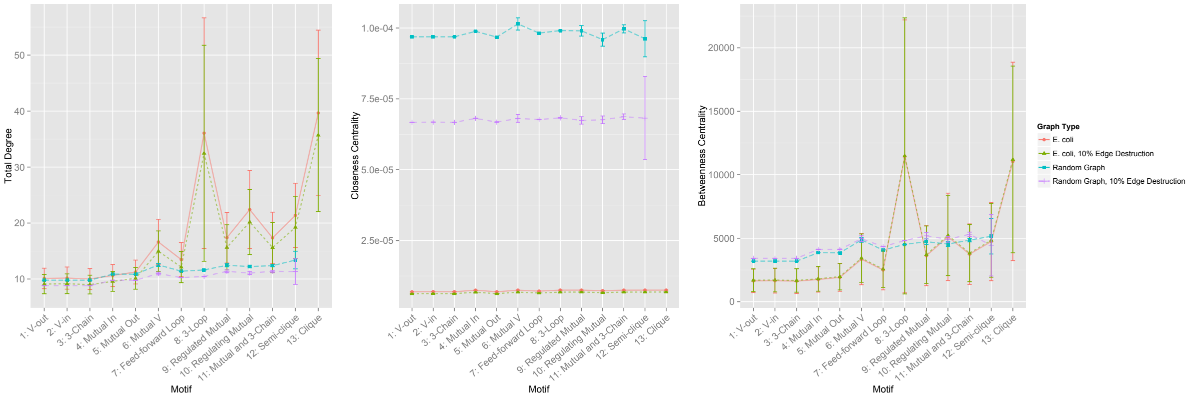


Figure B.4: Centrality measures by motif.

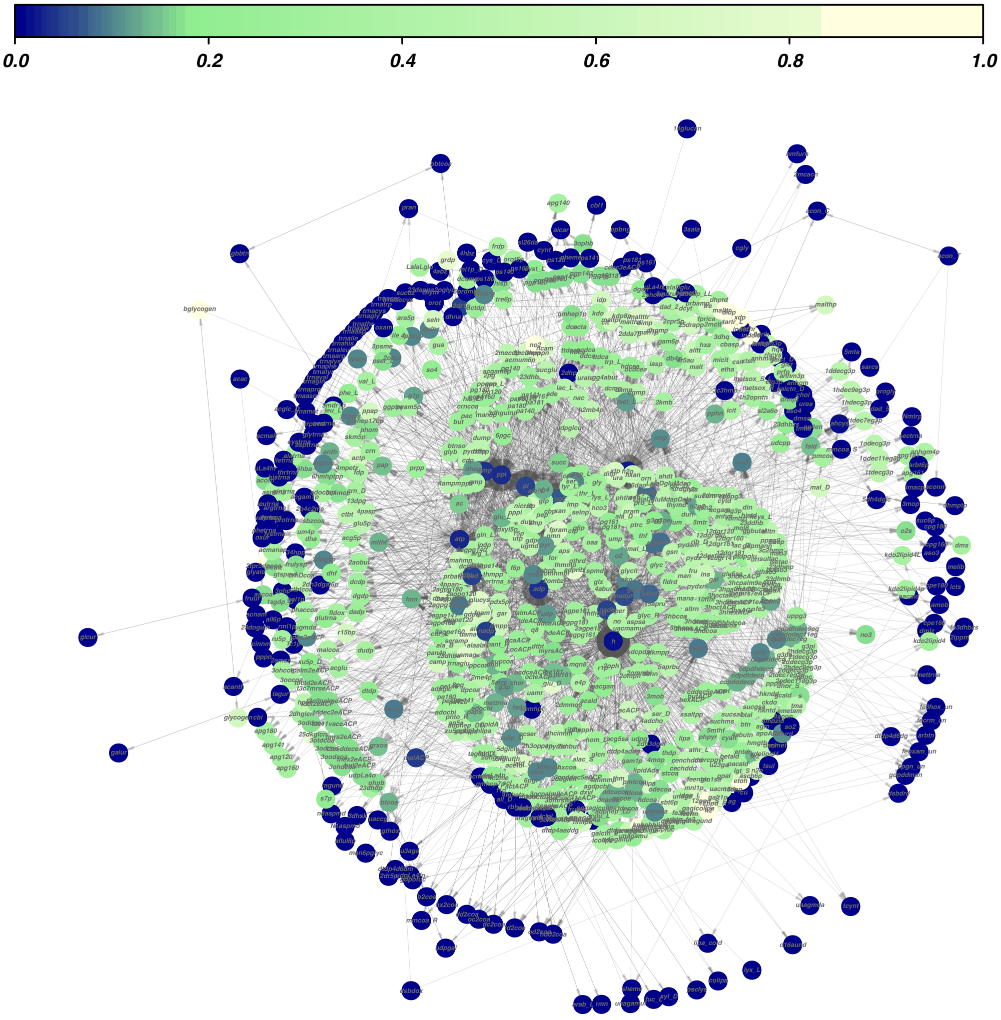


Figure B.5: The distribution of the clustering coefficient in the *E. coli* metabolic network.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] M Ahsanul Islam, Elizabeth A Edwards, and Radhakrishnan Mahadevan. Characterizing the metabolism of Dehalococcoides with a constraint-based model. *PLoS Comput Biol*, 6(8):1–16, 2010.

[2] Uri Alon. Network motifs: theory and experimental approaches. *Nat Rev Genet*, 8(6):450–461, 2007.

[3] Cheryl P Andam and J Peter Gogarten. Biased gene transfer in microbial evolution. *Nat Rev Micro*, 9(7):543–555, 2011.

[4] Siv G E Andersson, Alireza Zomorodipour, Jan O Andersson, Thomas Sicheritz-Ponten, U Cecilia M Alsmark, Raf M Podowski, A Kristina Naslund, Ann-Sofie Eriksson, Herbert H Winkler, and Charles G Kurland. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature*, 396(6707):133–140, 1998.

[5] L Aravind, Roman L Tatusov, Yuri I Wolf, D Roland Walker, and Eugene V Koonin. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends in Genetics*, 14(11):442–444, 1998.

[6] Albert-László Barabási and Zoltán N Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113, February 2004.

[7] A Barrat, M Barthélemy, R Pastor-Satorras, and A Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752, 2004.

[8] Alain Barrat, Marc Barthelemy, and Alessandro Vespignani. *Dynamical Processes on Complex Networks*. Cambridge, University Press, Cambridge, 3 edition, 2010.

[9] Georg Basler, Sergio Grimbs, Oliver Ebenhöh, Joachim Selbig, and Zoran Nikoloski. Evolutionary significance of metabolic network properties. *J R Soc Interface*, November 2011.

[10] Fabia U Battistuzzi and S Blair Hedges. A Major Clade of Prokaryotes with Ancient Adaptations to Life on Land. *Molecular Biology and Evolution*, 26(2):335–343, 2009.

[11] Scott A Becker, Adam M Feist, Monica L Mo, Gregory Hannum, Bernhard ØPalsson, and Markus J Herrgard. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat Protoc*, 2(3):727–738, 2007.

[12] Scott A Becker and Bernhard ØPalsson. Genome-scale reconstruction of the metabolic network in Staphylococcus aureus N315: an initial draft to the two-dimensional annotation. *BMC Microbiol*, 5:8, 2005.

[13] J Bluis and D G Shin. Nodal distance algorithm: calculating a phylogenetic tree comparison metric. *Bioinformatics and Bioengineering, 2003. Proceedings. Third IEEE Symposium on*, pages 87–94.

[14] Ulrik Brandes. A faster algorithm for betweenness centrality*. *The Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[15] Ronald S Burt. Structural Holes and Good Ideas. *American Journal of Sociology*, 110(2):349–399, 2004.

[16] T Cavalier-Smith. Archaebacteria and Archezoa. *Nature*, 339(6220):100–101, 1989.

[17] Roger L Chang, Lila Ghamsari, Ani Manichaikul, Erik F Y Hom, Santhanam Balaji, Weiqi Fu, Yun Shen, Tong Hao, Bernhard ØPalsson, Kourosh Salehi-Ashtiani, and Jason A Papin. Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism. *Mol Syst Biol*, 7:518, 2011.

[18] Xiaoyu Chen and Martin Tompa. Comparative assessment of methods for aligning multiple genome sequences. *Nat Biotech*, 28(6):567–572, 2010.

[19] Yu Chen and Xiaoxia Nina Lin. A Bioinformatic pipeline for automated genome-wide metabolic network reconstruction. May 2012.

[20] Han-Yu Chuang, Matan Hofree, and Trey Ideker. A Decade of Systems Biology. *Annu. Rev. Cell Dev. Biol.*, 26:721–744, July 2010.

[21] Bevan K S Chung, Suresh Selvarasu, Camattari Andrea, Jimyoung Ryu, Hyeokweon Lee, Jungoh Ahn, Hongweon Lee, and Dong-Yup Lee. Genome-scale metabolic reconstruction and in silico analysis of methylotrophic yeast Pichia pastoris for strain improvement. *Microbial Cell Factories*, 9(50):1–15, 2010.

[22] Gavin C Conant and Andreas Wagner. Convergent evolution of gene circuits. *Nat Genet*, 34(3):264–266, 2003.

[23] Christian de Duve. EVOLUTION OF THE PEROXISOME. *Annals of the New York Academy of Sciences*, 168(2):369–381, 1969.

[24] Christian de Duve. The origin of eukaryotes: a reappraisal. *Nat Rev Genet*, 8(5):395–403, 2007.

[25] Cristiana Gomes de Oliveira Dal'Molin, Lake-Ee Quek, Robin William Palfreyman, Stevens Michael Brumbley, and Lars Keld Nielsen. AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. *Plant Physiol*, 152(2):579–589, February 2010.

[26] Radu Dobrin, Qasim Beg, Albert-Laszlo Barabasi, and Zoltan Oltvai. Aggregation of topological motifs in the Escherichia coli transcriptional regulatory network. *BMC Bioinformatics*, 5(1):10, 2004.

[27] John Doyle and Marie Csete. Motifs, control, and stability. *PLoS Biol*, 3(11):e392, November 2005.

[28] Natalie C Duarte, Scott A Becker, Neema Jamshidi, Ines Thiele, Monica L Mo, Thuy D Vo, Rohith Srivas, and Bernhard ØPalsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *PNAS*, 104(6):1777–1782, February 2007.

[29] Natalie C Duarte, Markus J Herrgå rd, and Bernhard ØPalsson. Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res*, 14(7):1298–1309, July 2004.

[30] Sabrina D Dyall, Mark T Brown, and Patricia J Johnson. Ancient Invasions: From Endosymbionts to Organelles. *Science*, 304(5668):253–257, 2004.

[31] Jeremy S Edwards and Bernhard O Palsson. Systems Properties of the Haemophilus influenzaeRd Metabolic Genotype. *Journal of Biological Chemistry*, 274(25):17410–17416, 1999.

[32] Victor V Emelyanov. Mitochondrial connection to the origin of the eukaryotic cell. *European Journal of Biochemistry*, 270(8):1599–1618, 2003.

[33] Young-Ho Eom, Soojin Lee, and Hawoong Jeong. Exploring local structural organization of metabolic networks using subgraph patterns. *J Theor Biol*, 241(4):823–829, August 2006.

[34] Christian Esser, Nahal Ahmadinejad, Christian Wiegand, Carmen Rotte, Federico Sebastiani, Gabriel Gelius-Dietrich, Katrin Henze, Ernst Kretschmann, Erik Richly, Dario Leister, David Bryant, Michael A Steel, Peter J Lockhart, David Penny, and William Martin. A Genome Phylogeny for Mitochondria Among alpha-Proteobacteria and a Predominantly Eubacterial Ancestry of Yeast Nuclear Genes. *Molecular Biology and Evolution*, 21(9):1643–1660, 2004.

[35] Xin Fang, Anders Wallqvist, and Jaques Reifman. Development and analysis of an in vivo- compatible metabolic network of Mycobacterium tuberculosis. *BMC Systems Biology*, 4(160):1–24, 2010.

[36] Adam M Feist, Christopher S Henry, Jennifer L Reed, Markus Krummenacker, Andrew R Joyce, Peter D Karp, Linda J Broadbelt, Vassily Hatzimanikatis, and Bernhard ØPalsson. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol*, 3:121, 2007.

[37] Adam M Feist, Johannes C M Scholten, Bernhard ØPalsson, Fred J Brockman, and Trey Ideker. Modeling methanogenesis with a genome-scale metabolic reconstruction of Methanosarcina barkeri. *Mol Syst Biol*, 2:2006.0004, 2006.

[38] J Felsenstein. *PHYLIP (Phylogeny Inference Package) version 3.6.* University of Washington, Seattle., Department of Genome Sciences, 2005.

[39] Paul François and Vincent Hakim. Design of genetic networks with specified functions by evolution in silico. *Proceedings of the National Academy of Sciences of the United States of America*, 101(2):580–585, 2004.

[40] Christof Francke, Roland J Siezen, and Bas Teusink. Reconstructing the metabolic network of a bacterium from its genome. *Trends in Microbiology*, 13(11):550–558, 2005.

[41] Linton C Freeman. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239, 1978.

[42] Philippe Giegé, Joshua L Heazlewood, Ute Roessner-Tunali, A Harvey Millar, Alisdair R Fernie, Christopher J Leaver, and Lee J Sweetlove. Enzymes of Glycolysis Are Functionally Associated with the Mitochondrion in Arabidopsis Cells. *The Plant Cell Online*, 15(9):2140–2151, 2003.

[43] J Peter Gogarten and Jeffrey P Townsend. Horizontal gene transfer, genome innovation and evolution. *Nat Rev Micro*, 3(9):679–687, 2005.

[44] Orland Gonzalez, Susanne Gronau, Michaela Falb, Friedhelm Pfeiffer, Eduardo Mendoza, Ralf Zimmer, and Dieter Oesterhelt. Reconstruction{,} modeling and analysis of Halobacterium salinarum R-1 metabolism. *Mol. BioSyst.*, 4(2):148–159, 2008.

[45] Radhey S Gupta. The phylogeny of proteobacteria: relationships to other eubacterial phyla and eukaryotes. *FEMS Microbiology Reviews*, 24(4):367–402, 2000.

[46] Dapeng Hao, Cong Ren, and Chuanxing Li. Revisiting the variation of clustering coefficient of biological networks suggests new modular structure. *BMC Systems Biology*, 6(1):34, 2012.

[47] Tokumasa Horiike, Kazuo Hamada, Shigehiko Kanaya, and Takao Shinozawa. Origin of eukaryotic cell nuclei by symbiosis of Archaea in Bacteria is revealed by homology-hit analysis. *Nat Cell Biol*, 3(2):210–214, 2001.

[48] Piers J Ingram, Michael P H Stumpf, and Jaroslav Stark. Network motifs: structure does not determine function. *BMC Genomics*, 7:108, 2006.

[49] H Jeong, B Tombor, R Albert, Z N Oltvai, and A L. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.

[50] Jonathan R Karr, Jayodita C Sanghvi, Derek N Macklin, Miriam V Gutschow, Jared M Jacobs, Benjamin Bolival, Nacyra Assad-Garcia, John I Glass, and Markus W Covert. A Whole-Cell Computational Model Predicts Phenotype from Genotype. *Cell*, 150(2):389–401, 2012.

[51] Nadav Kashtan and Uri Alon. Spontaneous evolution of modularity and network motifs. *Proceedings of the National Academy of Sciences of the United States of America*, 102(39):13773–13778, 2005.

[52] Paul Kenrick and Peter R Crane. The origin and early evolution of plants on land. *Nature*, 389(6646):33–39, 1997.

[53] Hyun Uk Kim, Soo Young Kim, Haeyoung Jeong, Tae Yong Kim, Jae Jong Kim, Hyon E Choy, Kyu Yang Yi, Joon Haeng Rhee, and Sang Yup Lee. Integrative genome-scale metabolic analysis of Vibrio vulnificus for drug targeting and discovery. *Mol Syst Biol*, 7:1–15, 2011.

[54] Rhoda J Kinsella, David A Fitzpatrick, Christopher J Creevey, and James O McInerney. Fatty acid biosynthesis in Mycobacterium tuberculosis: Lateral gene transfer, adaptive evolution, and gene duplication. *Proceedings of the National Academy of Sciences*, 100(18):10320–10325, 2003.

[55] Konstantin Klemm and Stefan Bornholdt. Topology of biological networks and reliability of information processing. *Proceedings of the National Academy of Sciences of the United States of America*, 102(51):18414–18419, 2005.

[56] Thorsten Kloesges, Ovidiu Popa, William Martin, and Tal Dagan. Networks of Gene Sharing among 329 Proteobacterial Genomes Reveal Differences in Lateral Gene Transfer Frequency at Different Phylogenetic Depths. *Molecular Biology and Evolution*, 28(2):1057–1074, 2011.

[57] István A Kovács, Robin Palotai, Máté S Szalay, and Peter Csermely. Community Landscapes: An Integrative Approach to Determine Overlapping Network Module Hierarchy, Identify Key Nodes and Predict Network Dynamics. *PLoS ONE*, 5(9):e12528, 2010.

[58] Vinay Satish Kumar, James G Ferry, and Costas D Maranas. Metabolic reconstruction of the archaeon methanogen Methanosarcina Acetivorans. *BMC Systems Biology*, 5(28):1–10, 2011.

[59] V Lacroix, C G Fernandes, and M.-F. Sagot. Motif Search in Graphs: Application to Metabolic Networks. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 3(4):360–368, 2006.

[60] J A Lake and M C Rivera. Was the nucleus the first endosymbiont? *Proceedings of the National Academy of Sciences*, 91(8):2880–2881, 1994.

[61] G N Lance and W T Williams. A General Theory of Classificatory Sorting Strategies. *The Computer Journal*, 9(4):373–380, 1967.

[62] Fangting Li, Tao Long, Ying Lu, Qi Ouyang, and Chao Tang. The yeast cell-cycle network is robustly designed. *Proceedings of the National Academy of Sciences of the United States of America*, 101(14):4781–4786, 2004.

[63] Wenzhe Ma, Ala Trusina, Hana El-Samad, Wendell A Lim, and Chao Tang. Defining Network Topologies that Can Achieve Biochemical Adaptation. *Cell*, 138(4):760–773, 2009.

[64] R Mahadevan, D R Bond, J E Butler, A Esteve-Nuñez, M V Coppi, B O Palsson, C H Schilling, and D R Lovley. Characterization of Metabolism in the Fe(III)-Reducing Organism Geobacter sulfurreducens by Constraint-Based Modeling. *Applied and Environmental Microbiology*, 72(2):1558–1568, 2006.

[65] William Martin and Miklos Muller. The hydrogen hypothesis for the first eukaryote. *Nature*, 392(6671):37–41, 1998.

[66] Aurelien Mazurie, Samuel Bottani, and Massimo Vergassola. An evolutionary and functional assessment of regulatory network motifs. *Genome Biology*, 6(4):R35, 2005.

[67] Geoffrey I McFadden and Giel G van Dooren. Evolution: Red Algal Genome Affirms a Common Origin of All Plastids. *Current Biology*, 14(13):R514 – R516, 2004.

[68] Tom Michoel, Anagha Joshi, Bruno Nachtergaele, and Yves de Peer. Enrichment and aggregation of topological motifs are independent organizational principles of integrated interaction networks. *Mol. BioSyst.*, 7(10):2769–2778, 2011.

[69] R Milo, S Shen-Orr, S Itzkovitz, N Kashtan, D Chklovskii, and Alon U. Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298:824–827, 2002.

[70] Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. Superfamilies of Evolved and Designed Networks. *Science*, 303(5663):1538–1542, 2004.

[71] Baharan Mirzasoleiman and Mahdi Jalili. Failure Tolerance of Motif Structure in Biological Networks. *PLoS ONE*, 6(5):e20512, 2011.

[72] David Moreira and Purificación López-García. Symbiosis Between Methanogenic Archaea and δ-Proteobacteria as the Origin of Eukaryotes: The Syntrophic Hypothesis. *Journal of Molecular Evolution*, 47(5):517–530, 1998.

[73] Kevin Mowbrey and Joel B Dacks. Evolution and diversity of the Golgi body. *FEBS Letters*, 583(23):3738–3745, 2009.

[74] Karen E Nelson, Rebecca A Clayton, Steven R Gill, Michelle L Gwinn, Robert J Dodson, Daniel H Haft, Erin K Hickey, Jeremy D Peterson, William C Nelson, Karen A Ketchum, Lisa McDonald, Teresa R Utterback, Joel A Malek, Katja D Linher, Mina M Garrett, Ashley M Stewart, Matthew D Cotton, Matthew S Pratt, Cheryl A Phillips, Delwood Richardson, John Heidelberg, Granger G Sutton, Robert D Fleischmann, Jonathan A Eisen, Owen White, Steven L Salzberg, Hamilton O Smith, J Craig Venter, and Claire M Fraser. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of Thermotoga maritima. *Nature*, 399(6734):323–329, 1999.

[75] M E J Newman. A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39–54, 2005.

[76] Graham Noctor, Rosine De Paepe, and Christine H Foyer. Mitochondrial redox biology and homeostasis in plants. *Trends in Plant Science*, 12(3):125–134, 2007.

[77] Matthew A Oberhardt, Bernhard O Palsson, and Jason A Papin. Applications of genome-scale metabolic reconstructions. *Mol Syst Biol*, 5:1–15, 2009.

[78] Aloysius Phillips, Daniel Janies, and Ward Wheeler. Multiple Sequence Alignment in Phylogenetic Analysis. *Molecular Phylogenetics and Evolution*, 16(3):317–330, 2000.

[79] Anthony Poole and David Penny. Eukaryote evolution: Engulfed by speculation. *Nature*, 447(7147):913, 2007.

[80] Anthony M Poole and David Penny. Evaluating hypotheses for the origin of eukaryotes. *BioEssays*, 29(1):74–84, 2007.

[81] Robert J Prill, Pablo A Iglesias, and Andre Levchenko. Dynamic Properties of Network Motifs Contribute to Biological Network Organization. *PLoS Biol*, 3(11):e343, 2005.

[82] Jifeng Qian, Arend Hintze, and Christoph Adami. Colored Motifs Reveal Computational Building Blocks in the ¡italic¿C. elegans¡/italic¿ Brain. *PLoS ONE*, 6(3):e17013, 2011.

[83] E Ravasz, A L Somera, D A Mongru, Z N Oltvai, and A L Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, August 2002.

[84] Seth B Roberts, Christopher M Gowen, J Paul Brooks, and Stephen S Fong. Genome-scale metabolic analysis of Clostridium thermocellum for bioethanol production. *BMC Systems Biology*, 4(31):1–17, 2010.

[85] L Sagan. On the origin of mitosing cells. *J Theor Biol*, 14(3):255–274, March 1967.

[86] Rajib Saha, Patrick F Suthers, and Costas D Maranas. Zea mays irs1563: A comprehensive genome-scale metabolic reconstruction of maize metabolism. *PLoS ONE*, 6(7):e21784, 2011.

[87] Agatha Schlüter, Stéphane Fourcade, Raymond Ripp, Jean Louis Mandel, Olivier Poch, and Aurora Pujol. The evolutionary origin of peroxisomes: an ER-peroxisome connection. *Mol Biol Evol*, 23(4):838–845, April 2006.

[88] R R Sederoff. Molecular Mechanisms of Mitochondrial-Genome Evolution in Higher Plants. *The American Naturalist*, 130:S30—-S45, 1987.

[89] Najaf A Shah and Casim A Sarkar. Robust Network Topologies for Generating Switch-Like Cellular Responses. *PLoS Comput Biol*, 7(6):e1002085, 2011.

[90] Shai S Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet*, 31(1):64–68, 2002.

[91] Martin I Sigurdsson, Neema Jamshidi, Eirikur Steingrimsson, Ines Thiele, and Bernhard ØPalsson. A detailed genome-wide reconstruction of mouse metabolism based on human Recon 1. *BMC Systems Biology*, 4(140):1–13, 2010.

[92] Gregory E Sims. *FFP 3.18 - Feature Frequency Profile Phylogenetics Package*. Lawrence Berkeley National Lab, 3.18 edition, February 2012.

[93] Gregory E Sims, Se-Ran Jun, Guohong A Wu, and Sung-Hou Kim. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences*, 106(8):2677–2682, 2009.

[94] Janos Szabad. Putting together rather than taking apart. *EMBO Rep*, 11(12):904–906, 2010.

[95] Ines Thiele, Daniel Hyduke, Benjamin Steeb, Guy Fankam, Douglas Allen, Susanna Bazzani, Pep Charusanti, Feng-Chi Chen, Ronan Fleming, Chao Hsiung, Sigrid De Keersmaecker, Yu-Chieh Liao, Kathleen Marchal, Monica Mo, Emre Ozdemir, Anu Raghunathan, Jennifer Reed, Sook-Il Shin, Sara Sigurbjornsdottir, Jonas Steinmann, Suresh Sudarsan, Neil Swainston, Inge Thijs, Karsten Zengler, Bernhard Palsson, Joshua Adkins, and Dirk Bumann. A community effort towards a knowledge-base and mathematical model of the human pathogen Salmonella Typhimurium LT2. *BMC Systems Biology*, 5(1):8, 2011.

[96] Ines Thiele, Thuy D Vo, Nathan D Price, and Bernhard ØPalsson. Expanded metabolic reconstruction of Helicobacter pylori (iIT341 GSM/GPR): an in silico genome-scale characterization of single- and double-deletion mutants. *J Bacteriol*, 187(16):5818–5830, August 2005.

[97] P van Nes, D Bellomo, M J T Reinders, and D de Ridder. Stability from Structure: Metabolic Networks Are Unlike Other Biological Networks. *EURASIP Journal on Bioinformatics and Systems Biology*, 2009(1):1687–4153, 2009.

[98] A Vázquez, R Dobrin, D Sergi, J.-P. Eckmann, Z N Oltvai, and A.-L. Barabási. The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *Proceedings of the National Academy of Sciences*, 101(52):17940–17945, 2004.

[99] Sebastian Wernicke and Florian Rasche. FANMOD: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152–1153, 2006.

[100] Sebastian Wernicke and Florian Rasche. Fast Network Motif Detection: Manual. *Bioinformatics*, 22(9):1152–1153, June 2006.

[101] Denise M Wolf and Adam P Arkin. Motifs, modules and games in bacteria. *Current Opinion in Microbiology*, 6(2):125–134, 2003.

[102] Ying Zhang, Ines Thiele, Dana Weekes, Zhanwen Li, Lukasz Jaroszewski, Krzysztof Ginalski, Ashley M Deacon, John Wooley, Scott A Lesley, Ian A Wilson, Bernhard Palsson, Andrei Osterman, and Adam Godzik. Three-Dimensional Structural View of the Central Metabolic Network of Thermotoga maritima. *Science*, 325(5947):1544–1549, 2009.