Evolution of Hierarchy in Bacterial Metabolic Networks

Aaron Goodman, Dylan Rhodes, Laura Groenendaal

December 9, 2014

Abstract

How does hierarchy evolve? In this paper, we present a model to study evolution of flow hierarchy using bacterial metabolic networks. We find that hierarchical flows in metabolic networks are conserved across closely related species of bacteria. Additionally, an evolution-based, generative model is implemented and parameterized to mimic bacteria of different environments. Although the model ultimately does not effectively simulate real-world metabolic networks, intriguing results from this analysis point towards possibilities for further study.

1 Introduction

We typically think of hierarchy as the antithesis of self-organization. Hierarchical systems, organized from above, are often contrasted with systems organized by the spontaneous coordination of their component parts. However, recent work by Luo and Magee argues that hierarchy itself can be an emergent property of self-organized systems, since they found a high degree of hierarchy in such natural systems as supply chains, software systems, food webs, and neuronal circuits.

Our group is studying the evolution of hierarchy in the metabolic networks of well known, single-celled bacteria. The evolution of networks is difficult to study over a long time span because many computational and record keeping tools have only been around for the past couple decades. However, a rich source of data is present in the biological sciences, where we can study an array of organisms and make data-driven inferences about their evolutionary adaptations over time. Metabolic networks in particular have been increasingly studied over the last decade as gene sequencing methods have improved. In addition to their inherent interest as naturally occurring, complex structures, metabolic networks seem to share several basic mathematical properties with other types of real world graphs, such as power-law degree distribution and high scores for modularity [6]. Therefore, insights into their generative process are relevant to broader questions about how hierarchical networks evolve.

2 Prior Work

2.1 Evolution of Metabolic Networks

Early explorations of biological networks, including protein-protein interactions,transcription factor networks, and metabolic pathways have exposed universally recurring patterns. In particular, researchers have found that most biological networks take on power law degree distributions and have high clustering coefficients relative to randomly generated graphs of similar size. High average clustering coefficient is indicative of modular structure, since it implies that a network is composed of small, dense subgraphs. While modularity is intuitively relatable to the function of metabolic networks on the grounds of mostly independent biochemical pathways, it seems to conflict with a scale free degree distribution. Indeed, researchers have struggled to reconcile these properties in generative models, often turning to contrived examples which fail to mimic the structure and generative process of their biological counterparts or which match only a subset of the desired characteristics.

Minimal work has been done on the evolution of metabolic network, with the exception of Kreimer et. al which studied the metabolic networks of 138 bacterial species and their inferred ancestors in order to track the evolution of network modularity [2].

2.2 Hierarchy Metrics

Although hierarchy has often been neglected as a tool for network analysis, much progress has been made in recent years towards developing better metrics for characterizing hierarchical networks. Muchnik et al. showed in their 2007 paper that hierarchy generation algorithms elicit different motifs of networks when they measure different aspects of their modular structures. As one example, two algorithms run on the neural network of *C. Elegans* produced strikingly different, yet mutually informative results. A partitioning algorithm based on local hierarchies generated a partition which mimicked the flow of information from sensory to motor neurons, while a centrality based algorithm elicited a hierarchy resembling the spatial organization of the organism's neurons [5]. The latter is an instance of containment hierarchy, which assumes that nodes with close equivalence measures share lower common ancestors. The former, as a measurement for flow hierarchy, is more relevant for our own analysis of metabolic networks.

Recent work has proposed metrics for flow hierarchy in order to glean greater insights into the structure of real world networks. Flow hierarchy follows the insight that hierarchy emerges when the influence of certain nodes on others begins to differ. Most directed networks are structured as flow hierarchies, where nodes can be influenced by nodes at lower levels, and are not strictly nested according to level complexity. In Luo and Magee's definition of flow hierarchy, a containment ordering criterion of common ancestry does not apply, since it is the direction of flows of network resources that determines orders of levels [3]. By defining flow hierarchy around unidirectionality, they treat hierarchy as an architectural property that detects the extent to which local flows conform to the holistic direction of the graph.

The metric Luo and Magee choose to assess flow hierarchy, however, cannot handle impure forms of hierarchy. Suppose there were two acyclic networks, one where lower-level links flow into all of the the layers above them, and another where links at each level only flow into another. Luo and Magee‘s hierarchy degree would be identical in both cases. The Luo metric can thus yield trivial results when the network expands: as new nodes are randomly added, it is likely that the hierarchy score will decrease, because the probability of a cycle emerging increases. By defining hierarchy as a simple measure for asymmetry of flows, their metric offers insufficient specificity over their relative influence. The GRC measure introduced by [**?**], on the other hand, captures the heterogeneity of influence in an network, addressing some of these problems by sacrificing the elegance of Luo‘s measure.

3 Data and Methods

3.1 Metrics

We implemented two different hierarchy metrics to analyze the structure of our real world and generated networks. In order to evaluate their correspondence with our collected data, we also included in our analysis the Clauset-Newman-Moore modularity score as well as other topological properties (average path length, degree assortativity, average degree), and a measure of phylogenetic similarity derived from a tree of life dataset.

3.1.1 Luo Hierarchy

Our primary metric is the hierarchy score proposed by Luo and Magee, which calculates the proportion of edges retaining their local direction in the network, or more intuitively, the proportion of edges not included in any cycle. Their chosen algorithm constructs edge adjacency and network adjacency matrices, raises the edge adjacency matrix to the power p to find the edge distance matrix, and then evaluates its diagonal to determine whether a given edge is included in any cycle of the network.

$$h = 1 - \frac{\sum_{i=1}^{E} c_i}{E} \tag{1}$$

Where E is the number of edges in the network and c_i is an indicator variable for whether the edge i in a cycle. Graphs violating asymmetry in flows will yield a hierarchy metric that tends towards 0, while highly acyclic graphs will move closer to 1. The Luo hierarchy metric can be computed efficiently by identifying all of the strongly-connected components in the graph, and then identifying edges that connect nodes within the same SCC as part of a cycle, so $c_i = 1$ and $c_i = 0$ for all other edges.

3.1.2 Global Reaching Centrality

We have also implemented a second metric for flow hierarchy to confirm the results we obtained using Luo's metric. Mones, Vicsek and Vicsek introduced a hierarchy measure defined as the global reaching centrality (GRC) of a network, obtained by generalizing the concept of m reach centrality proposed by Borgatti's 2006 paper [**?**]. GRC is a heterogenous distribution of local reaching centrality, the maximum number of nodes reachable from each given node in the network. More formally, local reaching centrality, $C_R(i)$, is defined as the proportion of nodes in the network that can be reached through the outgoing edges of node i, or the number of nodes with a finite positive directed distance from node i divided by N-1. The global reaching centrality (GRC) is then defined as the difference between the maximum and the average value of these generalized local reach centralities over the network V:

$$GRC = \frac{\sum_{i \in V} |C_R^{max} - C_R(i)|}{N - 1} \tag{2}$$

3.2 Reconstructing the Networks

The bacterial metabolic networks shared with us by Kreimer *et al.* lacked essential data about the directionality of the reactions [2]. Their networks were unsuitable for our analysis because they protrayed enzyme reactions as uniformly reversible, which undermined the results of our graph partitioning algorithms and flow hierarchy metrics. We redressed this problem by reconstructing the networks, according to the method outlined in [4]. Using the REST API of the Kyoto Ecyclopedia of Genes and Genomes (KEGG) database, we augmented the data. We reconstructed the networks as directed networks by identifying which enzymes were present in each bacterial species, found the reactions that involved these enzymes, and created a nearly bipartite graph with nodes for metabolites and enzymes and edges between the enzymes and their metabolites. A few metabolites exist in equilibrium between two forms, and these were noted as edges between the two metabolites. Not only did these improved networks prove more informative for our analysis of flow hierarchy, they are also more true to the structure of our metabolic networks, for they include metabolites as well as enzymes.

3.3 Network Visualizations

Networks were visualized using Gephi and a ForceAtlas layout. Prior to vizualization, entries in the graph that referred to classes of enzymes and metabolites that were connected globally to on average more than 1% of the graph were removed. This eliminated much of the visual clutter, to allow for a clearer visualization of the network.

3.4 Phylogenetic Analysis

In order to understand how hierarchy evolved, we looked at how hierarchy scores varied along the phylogenetic tree. To perform this analysis, we used the tree of life dataset of Ciccarelli *et al.* [1]. This tree had 204 species, of which 124 we were able to find corresponding bacteria in our network datasets.

We calculated the pairwise tree distance between all of the represented bacteria in order to quantify their ancestral similarity. We then compared their ancestral similarity to the absolute difference of the number of

nodes, number of edges, mean node degree, diameter, SCC fraction, average clustering coefficient, Clauset modularity, and Luo and GRC hierarchies. The differences in these scores was correlated to the ancestral tree distances using Spearman's rank correlation.

3.5 Generative Models

3.5.1 Configuration Model

As a baseline for comparison with our evolutionary model, we implemented an algorithm to construct configuration models for directed graphs and ran it several times over the metabolic networks of each bacterium. The configuration model's synthetic output retains the number of nodes, number of edges, in-degree distribution, and out-degree distribution of each node. The algorithm first disconnects each edge, leaving each node with some number of in and out edge stubs disconnected from the rest of the graph. The algorithm then rewires these stubs together uniformly at random. Duplicate edges and self-loops are both disregarded if generated and replaced with a different combination of end points which still maintains the graph's original properties.

3.5.2 Evolutionary Model

The ultimate goal of our project was to implement a generative model with properties which mimicked those of real bacteria living under different environments. Thus, the model was designed to expose parameters with intuitive connections to real evolutionary processes with the aim of describing lifestyle conditions through alternative parameterizations. In the vein of other evolutionary models, we chose a time step based approach in which an initial graph is augmented over discrete steps by processes chosen probabilistically. Our final model contained four possible processes for each time step. Therefore, it could be parameterized with the number of time steps and a distribution over each of the four processes.

Name of Process	Description of Effect
Edge Deletion	Choose an edge uniformly at random and remove it from the graph.
Subgraph Duplication	Choose a node n uniformly at random and draw a size s from an exponential distribution with mean 1. For the neighborhood around n of size s, duplicate each node and edge and attach this subgraph to the parents of n in the original network.
Node Insertion	Choose if the new node should be a metabolite or enzyme with probability 1/2. Generate a distribution over the other class of nodes in which half of the probability mass is distributed uniformly and half is distributed in proportion to each node's in-degree. Select a node from this distribution and construct an edge from it to the new node.
Edge Insertion	Select a metabolite or enzyme uniformly at random and add an edge from it to a node drawn uniformly at random from the other class of nodes

Table 1: Processes included in evolutionary model

Each process was carefully designed to emulate the effect of a class of common evolutionary mutations. Edge deletion mirrors a point mutation which restricts the ability for an enzyme to catalyze a particular reaction. Subgraph duplication corresponds to partial gene duplication, a major driver of evolutionary diversity, in which parts of functional pathways are replicated in the genome, leaving individual copies free to mutate without effecting the viability of the organism. Node and edge insertions mimic the mutation of enzymes which create variants to catalyze new reactions or produce new metabolites. Notably, node insertions follow a partial preferential attachment model based on in-degree which emulates the propensity for organisms to develop pathways which make use of overabundant metabolites.

4 Results

4.1 Graph Overviews

A variety of hierarchical and non-hierarchical features can be seen in the metabolic networks. Figure 4.1 shows the metabolic network of *E. coli*. The graph is made up of a center of the graph is made up of tightly connected modules, some long pathway chains, and some networks that involve combining several metabolites in a multistep process. One particular example of this, is the synthesis of Malonyl-Coenzyme A which is used in many fatty acid synthesis pathwasy. This is in the upper right corner of the figure, and can be seen in more detail in Figure 2

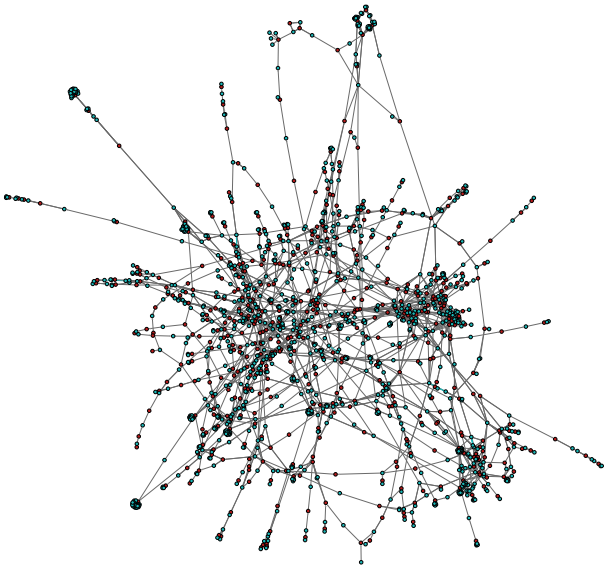

Figure 1: Visualization of *E. coli* metabolic network. Enzymes are red and metabolites are blue

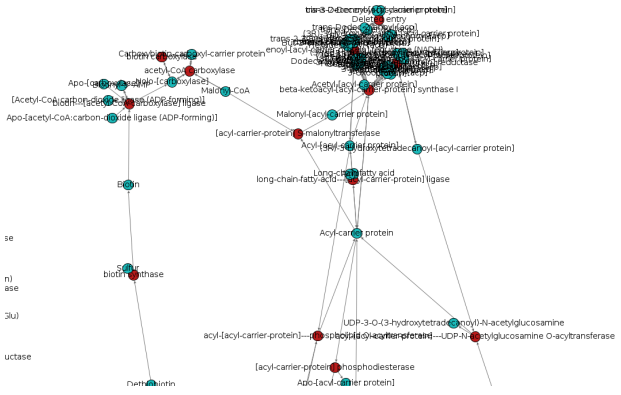

Figure 2: Closeup of hierarchy in fatty acid synthesis pathways. Enzymes are red and metabolites are blue

4.2 Descriptive Statistics

Model	Clust. Coef.	Luo	Global Reachability	Modularity	Effective Diameter
Real Network	0.004	0.128	0.130	0.634	6.76
Erdos-Renyi	0.003	0.171	0.083	0.428	9.35
Configuration Model	0.043	0.139	0.149	0.416	5.74

Table 2: Network statistics for generative models based on Escherichia coli HS

We also generated an ensemble of common generative models parameterized to resemble the size of each metabolic network in order to gain some insight into the set of properties which related to the metabolic network structure. In table 5, the results of this analysis are presented for a particular bacterium in our dataset, Escherichia coli HS.

4.3 Phylogenetic Analysis

The Luo hierarchy score is similar between closely related species. The evolutionary distance and Luo hierarchy can be seen in Figure 3.

Using the pairwise distance metrics discussed in Section 3.4, we found that all of the metrics were significantly correlated with evolutionary distance $p < 10^{-100}$. These results are shown in Table 4.3. The most strongly correlated metric was the Luo hierarchy, with $\rho = 0.42$. However we also note that the mean node degree difference is also highly correlated with ancestral distance. Thus, further investigation is needed to normalize the Luo hierarchy to the degree distribution.

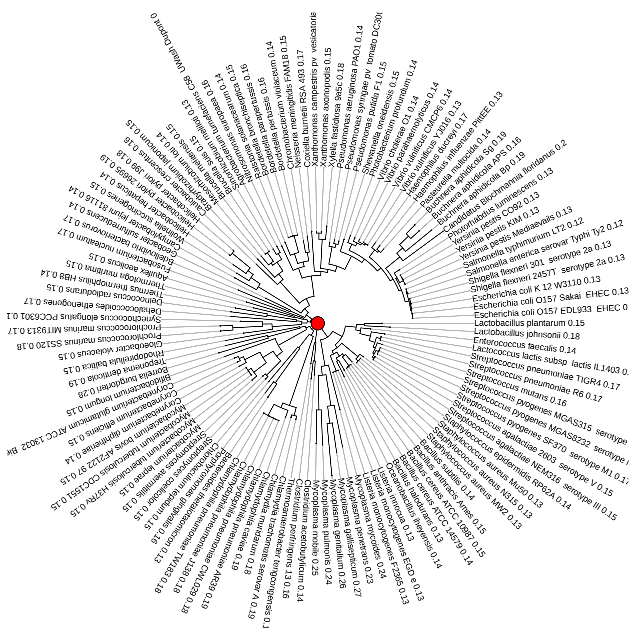

Figure 3: Luo Hierarchy score (at the end of the leaf labels) is conserved across closely related species.

Statistic	Spearman Correlation
Number of Nodes	0.35
Number of Edges	0.34
Average Clustering Coefficient	0.35
Luo Hierarchy	0.42
GRC Hierarchy	0.11
Clauset Modularity	0.30
SCC Fraction	0.34
Number of SCCs	0.12
Full Diameter	0.20
Effective Diameter	0.41
Mean Node Degree	0.41

Table 3: Correlation between pairwise evolutionary tree distances and graph metrics.

Habitat	Node Count	Edge Count	Luo Hierarchy	Global Reaching Centrality	Modularity
Aquatic	1668	4153	0.159	0.157	0.648
Host-associated	1339	3319	0.171	0.154	0.655
Multiple	1711	4406	0.149	0.149	0.645
Specialized	1512	3775	0.155	0.153	0.645
Terrestrial	1860	4865	0.146	0.151	0.639

Table 4: Aggregate network statistics for bacteria of differing lifestyles

4.4 Generative Model Results

In order to parameterize the evolutionary model effectively, we performed analysis on the characteristics of the bacteria in our dataset which lived within different habitats. Some of the representative statistics are depicted in the table 4. The results drove us to focus on host-associated bacteria and those which live in multiple environments. This choice was motivated by several factors, data driven and otherwise. First of all, host-associated and multiple environment bacteria were the two classes best represented in our dataset at 141 and 88 members apiece. Second, these two classes exhibited fairly divergent Luo hierarchy metric scores. The host associated type contained more than half of the bacteria in the dataset with very high ($>$ 0.2) Luo hierarchy score, while more than three quarters of the bacteria in the multiple type fell below the overall average hierarchy score. Third, these two classes of bacteria are the most intuitively differentiable in terms of lifestyle, since host-associated bacteria are well suited for a single, very specific environment while bacteria in the multiple class thrive in a variety of locales. This important distinction is also reflected in the average number of nodes of members of the two classes: multiple habitat bacterial genomes contain a greater number of enzymes and metabolites since they must be able to synthesize chemicals from a variety of precursors, whereas host-associated organisms can operate successfully with fewer enzymes in their more static habitat.

The choice of these two classes of bacteria motivated the parameterization of our evolutionary model in separate ways. We created a faux host-associated organism by allocating a greater probability mass to the chances of deleting edges and preferentially attaching nodes, since host-associated organisms tend to prefer smaller genomes and have fewer distinct metabolic pathways of greater complexity than their counterparts in multiple environments, since they must use a smaller range of precursors to produce their metabolites. The synthetic organism suited for multiple environments had a large probability mass assigned to subgraph duplication and random edge insertion, since bacteria which live in multiple environments must have many well connected pathways which can synthesize the chemicals they require from varied precursors. We also parameterized a baseline synthetic organism with equal probability mass assigned to the possible evolutionary

processes at each time step.

To test our hypotheses, we generated one hundred synthetic metabolic networks for each of the three parameterizations and calculated aggregate summary statistics over the groups. We hoped to observe higher Luo hierarchy metrics for the host-associated organisms than for the multiple environment class. We also expected ballpark similarity with the results from the real world dataset. The results can be seen in table 5.

Model Type	Node Count	Edge Count	Luo Hierarchy	Global Reaching Centrality	Modularity
Multiple	1093	2032	0.630	0.689	0.659
Host-associated	432	421	0.975	0.280	0.812
Baseline	631	601	0.984	0.287	0.838

Table 5: Aggregate statistics for synthetic bacteria of differing lifestyles

As you can see, the results of our experiment were not especially good. The synthetic bacteria networks failed to resemble those of the real world bacteria in most qualities. However, there are still some useful insights which can be gained from these results. First of all, subgraph duplication and random edge insertion appear vital for synthetic generative models to achieve Luo hierarchy metrics in a reasonable range. The synthetic bacteria generated for multiple environment suitability were by far the most realistic in node degree, Luo hierarchy, and modularity scores. However, they were notably less similar to the real world networks in terms of general reaching centrality than the other two parameterizations of the synthetic model. This result is fairly intuitive; the other two classes were less well-connected than the multiple environment bacteria, because they had a greater probability mass allocated to random edge deletion. Although the baseline and host-associated models are very poor proxies for real metabolic networks, they do suggest that connectivity must be kept in check to approximate true graphs. Finally, it is clear that preferential node attachment is not a realistic evolutionary step, since it inflates the Luo hierarchy score to artificially high values.

5 Conclusion

Although our model was not an effective proxy for the metabolic network data, the process of generating it yielded promising results. We found that preferential attachment was not an adequate model for the evolution of hierarchy in metabolic networks.

These preliminary insights should be pursued further. First, the evolution of hierarchy has implications for industry, since metabolic network structure is of critical importance to developing novel biotechnology compounds. Most new drugs are derived from naturally occuring small molecules, but the molecules can be difficult to extract, and the organisms that create these molecules can rarely be cultivated in the lab. Thus biologists use transgenic approaches to insert the metabolic pathways into bacteria that can be cultured easily. However simply putting the relevant genes into the bacteria is not enough and our analysis of evolution of hierarchy will contribute to a systems view of metabolic networks.More critically, greater understanding of hierarchy has implications for fields ranging from the economic to the political. For example, in organization of firms it is often thought that smaller firms have a flatter, less hierarchical structre, and as the firms grow and evolve they become more hierarchical. Our results above suggest that the preferential attachment mechanism, or 'rich get richer' phenomenon, may not be primary in either natural nor human-made networks.

References

[1] F. D. Ciccarelli, T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork. Toward automatic reconstruction of a highly resolved tree of life. *Science (New York, N.Y.)*, 311(5765):1283–7, Mar. 2006.

[2] A. Kreimer, E. Borenstein, U. Gophna, and E. Ruppin. The evolution of modularity in bacterial metabolic networks. *PNAS*, 105(19):6976–81, May 2008.

[3] J. Luo and C. Magee. Detecting evolving patterns of self-organizing networks by flow hierarchy measurement. *Complexity*, 2011.

[4] H. Ma and A. Zeng. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19(2):270–277, 2003.

[5] L. Muchnik, R. Itzhack, S. Solomon, and Y. Louzoun. Self-emergence of knowledge trees: Extraction of the Wikipedia hierarchies. *Physical Review E*, 76(1):016106, July 2007.

[6] E. Ravasz and A.-L. Barabási. Hierarchical organization in complex networks. *Physical Review E*, 67(2):026112, Feb. 2003.

Team Member Contributions

Aaron wrote the code to integrate the data from Kreimer *et al.* and KEGG, reconstruct the networks, calculate the Luo hierarchy, created the phylogenetic tree, wrote code to compute distance matrices and find correlations, wrote code to create the dendogram with hierarchy scores, and created the network visualizations.

Dylan wrote the script to fetch enzyme and metabolite names using the KEGG API, implemented the evolutionary model and the configuration model, performed data analysis on the habitat-segmented bacteria, wrote code to calculate an ensemble of network statistics, and wrote a graph visualization in MATLAB which took advantage of Louvain community detection and made it into the milestone report but not the final one.

Laura implemented the Global Reaching Centrality metrics, wrote a script to simultaneously add label and color information to GraphViz dot files for visualization, ran ridge regressions on hierarchy scores against habitat classifications, implemented the first pass synthetic evolutionary model for the milestone report, and ran a battery of network statistics on alternative evolutionary models.