bioRxiv preprint doi: https://doi.org/10.1101/118299; this version posted May 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RESEARCH

Evolution of Hierarchy in Bacterial Metabolic Networks

Aaron Goodman and Marcus Feldman*

*Correspondence:
mfeldman@stanford.edu
Department of Biology, Stanford University, Stanford, CA, 94305, USA
Full list of author information is available at the end of the article

Abstract

Background: Through simulation studies and genomic data analysis, researchers have shown that modularity in metabolic networks is an important adaptive mechanism enabling survival in changing environments. Another global property of networks, flow hierarchy, is often used in optimizing information flow in designed networks. Hierarchy also arises in self-organized networks as an optimal way to aggregate flows while minimizing costs of connections.

Results: Using a comparative approach on 2,935 bacterial metabolic networks, we show that hierarchy evolves with modularity and is conserved to a high degree. Hierarchy in bacterial metabolic networks reflects a fundamental tradeoff between growth rate and biomass production, and reflects a bacteria's realized ecological strategy. Additionally, by inferring the ancestral metabolic networks, we find that hierarchy decreases with distance from the root of the tree, suggesting the important pressure of increased growth rate relative to efficiency in the face of competition.

Conclusions: Just as hierarchical character is an important structural property in efficiently engineered systems, we see that it evolves in self-organized bacterial metabolic networks, is reflective of the life-history strategy of the bacteria, and plays an important role in network organization and efficiency.

Keywords: modularity; hierarchy; metabolism; bacteria; reverse ecology

Background

In characterizing bacteria, we seek to understand both their internal processes and how they interact with other species and their environments. Techniques in cell and molecular biology have been very helpful in understanding the inner workings of bacteria, but do not address the ecological context of bacteria. Increasingly,

bioRxiv preprint doi: https://doi.org/10.1101/118299; this version posted May 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

metagenomic techniques are being used to simultaneously sequence all of the bacteria present in a given environment. However, these techniques can only tell us a limited amount of information about particular species, where they are found, their relative abundances, and co-occurrence patterns.

One way we can move beyond the correlational profiles of metagenomics to understand the underlying mechanisms is by studying a bacterium's metabolic network and its evolution. A bacterium's ability to reproduce depends on the efficiency of its metabolism, which we can study as a network of metabolites linked together by the enzymes that transform one metabolite into another [1]. These networks metabolic reflect the environmental pressures which guided bacterial evolution, and the structure of these networks varies across the bacterial kingdom. Thus, by studying its metabolic network we can begin to understand the ecological role the bacterium occupies. Such approaches have shown that metabolic networks can predict the minimal nutrients a bacterium needs to grow, as well as mutualistic relationships between bacteria, and their major ecological niche [2][3][4].

In studying these metabolic networks, we can draw from work done on networks in other fields, particularly the study of information accrual networks. In the process of synthesizing the complex molecules needed for survival, bacteria reduce the overall entropy in the cell. This reduction, accomplished by the synthesis pathways, can also be viewed as an increase of information, given the thermodynamic equivalence of entropy reduction and information accrual. Thus we can use the metabolic network graph to study the flow of information through the cell.

Information accrual networks are found in engineering, where they are used in the design of control systems, and in the social sciences to study the organization of the firm [5][6]. In these fields, networks are often constructed to have a hierarchical structure of information flow. In control systems, information is acquired at the lowest level of the hierarchy and transmitted to higher levels, where it is aggregated and passed upward; at the same time orders come from the top of the hierarchical network and are passed down to lower levels These networks are constructed with the topology imposed upon them. However, top-down network design is not necessary for the emergence of hierarchy. When measured in degree rather than in absolute terms, hierarchical characteristics are observed in many self-organized systems such as food webs, supply chains, and transcription factor networks [7].

bioRxiv preprint doi: https://doi.org/10.1101/118299; this version posted May 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The global reaching centrality (GRC) is a commonly used quantitative measure of flow hierarchy; it is defined as the average difference between the maximum local reaching centrality (i.e. fraction of nodes in the network accessible by each node of the network) and the local reaching centrality. Other, less widely used measures of flow hierarchy are an eigenvector centrality based method, the fraction of edges participating in cycles, or by decomposition into treeness, feedfowardness, and orderability [8][9] [7].

Previous study of the hierarchical nature of metabolics networks has focused on containment hierarchy, which characterizes the recursive property of networks consisting of modules containing other modules is the containment hierarchy, which is often measured by the cophenetic coefficient [14]. Metabolic networks have been shown to be composed of self-similar modules of metabolites and modules of modules and that these modules correspond to known pathways [15][16]. The hierarchical modular organization of metabolic networks has been hypothesized to increase evolvability of metabolism [17]. Simulations of Boolean logic networks have suggested that modularity evolves in changing environments, and it has been hypothesized that this would be reflected in bacterial metabolic networks, though this has not been borne out [10][11][12]. The difference in modularity of metabolic networks has been found to be moderately correlated with the phylogenetic divergence of the organisms. Based on imputed ancestral metabolic networks, there is a general trend of loss of modularity over evolutionary time due to the addition of peripheral pathways during niche specialization [13].

In this work, we focus on the heretofore neglected flow hierarchy. Flow and containment hierarchies describe different properties of the same network. The containment hierarchy of a neural network of *C. elegans* represents the spatial organization of the physical neurons and the flow hierarchy represents the flow of information from the sensory to motor neurons [18]. Flow hierarchy has been shown to evolve in networks that have costs to each connection [19].

Although flow hierarchy (hereafter refered to as hierarchy) has not been well studied in metabolic networks, it has been identified in a variety of self-organized networks, including food webs, neural networks, and the transcription factor network in *D. melanogaster*, where the degrees of hierarchy were significantly higher than would be expected in a random network with the same degree distribution [7].

bioRxiv preprint doi: https://doi.org/10.1101/118299; this version posted May 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Degree of hierarchy has also been used to make inferences about the forces guiding the development of networks [20]. Hierarchical characteristics in emergent social networks can predict the costs of maintaining information-sharing relationships, and in supply chains reflects the degree of market variability they are able to withstand [21].

In simulated evolution experiments of Boolean logic networks, the cost of maintaining links between nodes is the driving force in the emergence of hierarchy [19]. Similarly, in metabolic networks, maintaining catalytic abilities between metabolites incurs a cost either as a tradeoff between specificity and efficiency, or from production and replication of unused enzyme [20][22].

Studying the evolution of modularity, and of containment hierarchy, or the hierarchical composition of modules can tell us about the evolutionary contingencies underlying the construction of metabolic networks. However by studying the flow hierarchy we can learn about the ecological niche that the bacteria fill [17]. We show that there is more conservation of hierarchy than of modularity and that hierarchy evolves independently of modularity. We employ the reverse ecology principle to understand how the hierarchical character of a metabolic network reflects the life history strategy of a bacterium in relationship to the growth-yield tradeoff, and its environmental niche.

Results and Discussion

Networks

Networks were reconstructed from 2,935 bacteria species in the KEGG database. These networks were robust to misannotation of enzymes. In random perturbations of the metabolic network for *E. coli* with 10% of the reactions removed, 95% the networks had hierarchy scores within 12% of the true network, and with 10% of reactions reversed, within 6% of the true network.

Network sizes ranged from 76 to 1496 metabolites, with a mean of 848. The smallest was the obligate insect parasite *Nasuia deltocephalinicola* and the largest was the soil bacterium *Burkholderia lata*.

Hierarchy

Hierarchy scores for the metabolic networks were calculated using the GRC hierarchy score (see methods). The mean degree of hierarchy was 0.279, and ranged

bioRxiv preprint doi: https://doi.org/10.1101/118299; this version posted May 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

from 0.065, for the insect symbiote Candidatus *Nasuia deltocephalinicola*, to 0.385 for a *Blattabacterium* endosymbiont of *Nauphoeta cinerea*, an insect endosymbiote. The hierarchy score for *E. coli* strains was 0.269 (Figure 1). For comparison with a random network and real world networks, GRC hierarchy scores for an Erdős-Rényi random graph is 0.058, a scale free network 0.127, and a tree 0.997, an estuary food web 0.814, and the neuronal network of *C. elegans* [23].

Relationship to environment and growth rate

There is a fundamental ecological trade off between growth rate and yield, which is a result of the underlying efficiencies of the reactions. Bacteria that have a metabolism that produces the maximal growth rate per amount of carbon taken up will have suboptimal biomass production, and vice versa.

This tradeoff is representative of fundamentally divergent ecological strategies that bacteria use [24]. Furthermore, the tradeoffs between growth and yield are represented in the constraints on the metabolic network, such that high-yield strategies lead to more hierarchical networks. There is a tradeoff between enzyme specificity and efficiency, so when yield is favored there will be higher costs of maintaining edges in the network, which leads to hierarchy [25] [20]. Rapidly growing bacteria have more metabolic cycles which allow for metabolic flexibility at the cost of wasted energy, and these cycles decrease hierarchy [26]. The cost of maintaining unused enzymes in the genome is higher when efficiency is paramount [22].

Using a dataset of 111 bacteria with known growth rates, we see that the hierarachical character of the network correlates inversely with growth rate, Spearman $\rho = -0.31$, $p = 0.00065$, fig 2. Furthermore, there is evidence that carbon efficiency constraints on bacteria differs greatly by environment, and that the evolutionary dynamics of carbon usage niche specialization are stronger within populations [27][28]. When we control for the bacterial environment, we see a correlation of $\rho = -0.41$, which is significantly greater than 0 ($p = 0.0001$, and significantly greater than the correlation when not controlling for the environment $p = 0.003$).

Thus the hierarchical character of the metabolic networks reflects the growth rate of the organisms and their environmental niche. These constraints of edge weight and tradeoffs between hierachical and ahierachical networks in metabolism are similar to those made in social networks and supply chains [20] [21].

bioRxiv preprint doi: https://doi.org/10.1101/118299; this version posted May 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Relationship to other network properties

In addition to measuring hierarchy, we evaluated a number of other network statistics. We computed node count, edge count, modularity (as measured by Girvan-Newman betweenness centrality [29]), clustering coefficient, full diameter, effective diameter, number of strongly connected components, proportion of the nodes in the largest strongly connected component, and Luo Hierarchy score, an alternative metric of hierarchy that measures the proportion of edges that do not participate in any cycles. Edge and node count correlated most strongly with genetic distance. However, after these basic structural properties, the statistics that correlated most highly with genetic distance were the Girvan-Newman Modularity score and the GRC hierarchy score (Table 1). We also computed the partial correlation for each variable with genetic distance, controlling for the others, and found that the GRC metric had the highest partial correlation.

Hierarchy Over Time

The hierarchy of the KEGG bacteria and reconstructed ancestors seems to first increase, and then decrease with distance from the root of the tree (Figure 3). Interestingly, with the dataset of 2,935 from the latest KEGG database, the correlation of modularity and distance from the root of the tree found by Kreimer *et al.* [13] is actually reversed. Modularity appears to increase rather than decrease with distance from the root, (Figure 4). This correlation remains positive when restricting analysis to the species used by Kreimer *et al.*.

As bacteria specialize to niches in a given ecosystem, they take on different metabolic strategies, which are reflected in the hierarchical profile of the metabolic network. This difference in strategies is consistent with the rise and fall of hierarchy over the evolutionary trajectory. As microbes first adapt to new environments or habitats (niche *sensu* Grinnell) they must gain novel metabolic functions, which are added as pathways in the periphery in the network and which increases the hierarchical character [30]. As complex relationships develop within the habitats, and bacteria adapt to different resource use profiles and competitive strategies (niche *sensu* Elton), the hierarchical profile of the metabolic network diversifies. Thus, the decrease in hierarchy over evolutionary time is caused by more bacteria specializing in a rapid-growth strategy, but the increasing variance in hierarchy reflects the fact

bioRxiv preprint doi: https://doi.org/10.1101/118299; this version posted May 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

that not all bacteria adopt this strategy. In studying the adaptive strategies chosen by different bacteria, we may be able to make inferences about the bacteria and their environments, as well as the interplay between evolutionary and ecological dynamics.

Correlation of Modularity and Hierarchy

Hierarchy and modularity are global properties of metabolic networks. Both correlate with bacterial specialization, and both change with distance from the root of the phylogenetic tree. Using the method of phylogenetic independent contrasts to look for correlation independent of phylogenetic structure, we found a moderate inverse correlation between modularity and hierarchy (Pearson correlation $r = -0.18, p < 10^{-15}$), suggesting little evolutionary relationship between modularity and hierarchy [31]. Interestingly, simulated Boolean networks demonstrate a positive correlation between modularity and hierarchy [19].

Conclusion

Characterizing the hierarchical structure of metabolic networks is useful in understanding the constraints under which these networks evolve. Hierarchy correlates with phylogenetic divergence, as would be expected for a trait subject to natural selection. This correlation is similar to the correlation of phylogenetic distance and modularity, suggesting that the hierarchical organization of networks, like modular organization, is important for function. However, modularity should be viewed as complementary to, rather than supplanted by, hierarchy when analyzing the global organization of metabolic networks. Both structural properties are conserved across phylogenies and evolve together. A better understanding of the character of metabolic networks is valuable in the growing field of 'reverse ecology,' in which the observed networks can be used to make inferences on possible environments [2][32][33].

By algorithmically reconstructing the metabolic networks, we are able to perform a larger-scale analysis than has previously been reported. Although the reaction annotations in KEGG may be prone to errors or omissions, we find that the GRC hierarchy metric is robust to small amounts of reaction omissions or reversals. By expanding the scope of the analysis, we find that modularity is actually inversely

bioRxiv preprint doi: https://doi.org/10.1101/118299; this version posted May 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

correlated with distance from the root of the tree, contrary to what has been found in previous studies of a more limited set of bacteria.

From reconstructed ancestral metabolic networks, we are able to infer how hierarchy evolves in networks over time, and understand the interplay between evolutionary and ecological dynamics. Hierarchy shows an increase followed by a decrease across the phylogenetic tree, which is reflective of the adaptive process of bacteria, first to novel fundamental niches, and then to a realized niche. The net trend in decreasing hierarchy reflects a dominance of fast-growth, low-efficiency strategy.

Methods

Hierarchy Metric

Hierarchy scores were calculated using the global reaching centrality metric developed by Mones *et al.*, which is based on the local reaching centrality [23]. The local reaching centrality (LRC) of a node in a network is the fraction of the nodes of the network that can be reached starting at the focal node. More precisely, if the metabolic network is represented as a graph, $G = (V, E)$ it can be said that v reaches v' if there exists a series of edges $(v, v_i), (v_i, v_j)...(v_j, v') \in E$. Let $R(v)$ be the set of nodes $v' \in V$ where v' is reachable from v. Then the LRC of v is $\frac{|R(v)|-1}{|V|-1}$. The GRC is then $\frac{1}{|V|-1} \sum_{v \in V} \max_{v' \in V} \mathrm{LRC}(v') - LRC(v)$

Modularity Metric

The modularity metric was calculated using the SNAP package [34]. The modularity of a network is the optimal partitioning of the nodes into clusters to maximize $Q = \frac{1}{4m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} I_{ij} \right)$. Where m is the number of edges in the network, A_{ij} is the adjacency matrix, i.e. A_{ij} is 1 if there is an enzyme that converts metabolite i into metabolite j. k_i is the number of reactions that metabolite i participates in, and $I_{ij} = 1$ if i and j are in the same module, and -1 otherwise. Since finding the global optimal of Q is an NP-hard problem, we use the method developed by Girvan and Newman, which partitions the network by iteratively removing the edge with the highest betweenness centrality [29].

Robustness of Reconstruction

The KEGG database is large, with heavy manual curation; however, this does not mean that the data are always perfect. A reaction may be favorable in one direction

bioRxiv preprint doi: https://doi.org/10.1101/118299; this version posted May 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in a model organism in laboratory conditions, but might proceed in the opposite direction or become bidirectional in different environments or species. It is also possible that reactions are missing from the database, or that an enzyme placed in an orthology group based on the study of one species may catalyze a different reaction in other species. To evaluate robustness to errors in the KEGG database, we examined the network for the well-studied bacterium, *E. coli*. We performed 100 replicates dropping or reversing 10% of the reactions, evaluated the hierarchy scores of these networks, and calculated the spread of the central 95% of hierarchy scores.

Reconstruction of Genetic Distance

Following the methods often used in bacterial comparative genomics [35][3][36], for each of the 2,935 species, the 16s ribosomal sequence from KEGG was aligned to the Greenegenes database using PyNast, resulting in multiple sequence alignments for the 2,935 species [37][38]. The genetic distances between all pairs of bacterial species were computed using the Kimura distance metric [39].

Reconstruction of Networks

For each bacterial species, a network of metabolites was inferred based on the enzymes present in the genome, the reactions known to be catalyzed by the enzymes present or orthologous enzymes, and a database of reaction substrates and products. The KEGG database of the genomic content of the 2,935 bacterial genomes was used to identify which enzyme classes were present in each genome and which reactions were present [40]. The reaction information from KEGG was supplemented by a the bioreaction database from Stelzer *et al.* which excludes currency metabolites, improves on predictions of directionality of reactions, and, for reactions with multiple substrates and products, provides carbon tracking of which substrates are converted to which products [41]. Using this reaction information, networks were constructed with metabolites as nodes, and a directed edge was placed between metabolites if there was a reaction that converted one metabolite to another. If reactions were reversible, then bi-directional edges were added between the substrates and products.

bioRxiv preprint doi: https://doi.org/10.1101/118299; this version posted May 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Ancestral Networks

To construct the ancestral networks, a phylogenetic tree was reconstructed using RAxML 8.2.9 and the 16-state GTR nucleotide substitution model with gamma rate heterogeneity [42]. The branch-length weighted average bootstrap support of the partitions over 300 trees was 85.4. Using the maximum likelihood estimate of the best tree, at each interior node of the phylogenetic tree, a genome was constructed using the Fitch small parsimony algorithm. In cases where the presence or absence of a gene was equally parsimonious, the gene was randomly selected to be included. These ancestral genomes were then used to reconstruct ancestral networks, just as the networks were constructed on the leaves of the tree.

Niche strategies

Growth rate data for 113 bacterial species and environmental annotations for those bacteria, which for 68 species were gathered from NCBI, and manual curation following literature review was used for the remaining 45 [43][4]. Due to their low number, the two aquatic species in the data set were excluded from further analysis. Correlations were calculated as Spearman's ρ. To calculate correlation controlling for environment, ρ was calculated within each environment, and a species weighted-average across environments was computed. Due to several bacteria having the same growth rates, p-values were calculated using permutation tests rather than the Student's t-distribution approximation. Significance tests were performed with 100,000 permutations each. For the overall ρ, permutations were done across all bacteria. To test the strength of the habitat-controlled correlation, growth rates were permuted within habitat classes and the species-weighted ρ was computed for each permutation. To test the effect of controlling for the environment, habitat labels were permuted and the difference between the species weighted ρ and overall ρ was computed.

Consent to Publish

All authors have approved this manuscript for submission. This work has not been published or submitted elsewhere.

Competing interests

The authors declare that they have no competing interests.

Author Contributions

AG designed the experiment, carried out the analysis, and wrote the paper. MF provided critical guidence on the direction of the work, and revision of the manuscript.

bioRxiv preprint doi: https://doi.org/10.1101/118299; this version posted May 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Availability of the Data

The dataset supporting the conclusions of this article, specifically the reconstructed metabolic network (Supplementary File 1) and topological statistics (Supplementary File 2) of these networks, are included within the article (and its additional files).

Acknowledgments

We are grateful to B Callahan, E Borenstein and J Leskovec for helpful feedback with this work, and to editor Richard Goldstein and two anonymous reviewers for there insightful comments.

References

1. Ma, H., Zeng, A.-p.: Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. Bioinformatics **19**(2), 270–277 (2003)
2. Borenstein, E., Kupiec, M., Feldman, M.W., Ruppin, E.: Large-scale reconstruction and phylogenetic analysis of metabolic environments. Proceedings of the National Academy of Sciences of the United States of America **105**(38), 14482–7 (2008). doi:10.1073/pnas.0806162105
3. Levy, R., Borenstein, E.: Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. Proceedings of the National Academy of Sciences **110**(31), 12804–12809 (2013). doi:10.1073/pnas.1300926110
4. Freilich, S., Kreimer, A., Borenstein, E., Yosef, N., Sharan, R., Gophna, U., Ruppin, E.: Metabolic-network-driven analysis of bacterial ecological strategies. Genome biology **10**(6), 61 (2009). doi:10.1186/gb-2009-10-6-r61
5. Brooks, R.: A robust layered control system for a mobile robot. IEEE Journal on Robotics and Automation **2**(1), 14–23 (1986). doi:10.1109/JRA.1986.1087032. 1010.0034
6. Teece, D.J.: Firm organization, industrial structure, and technological innovation. Journal of Economic Behavior & Organization **31**(2), 193–224 (1996). doi:10.1016/S0167-2681(96)00895-5
7. Luo, J., Magee, C.: Detecting evolving patterns of selforganizing networks by flow hierarchy measurement. Complexity (2011)
8. Czégel, D., Palla, G.: Random walk hierarchy measure: What is more hierarchical, a chain, a tree or a star? Scientific Reports **5**(December), 17994 (2015). doi:10.1038/srep17994. arXiv:1508.07732v1
9. Corominas-Murtra, B., Goñi, J., Solé, R.V., Rodríguez-Caso, C.: On the origins of hierarchy in complex networks. Proceedings of the National Academy of Sciences of the United States of America **110**(33), 13316–21 (2013). doi:10.1073/pnas.1300832110. 1303.2503
10. Kashtan, N., Alon, U.: 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). doi:10.1073/pnas.0503610102
11. Parter, M., Kashtan, N., Alon, U.: Environmental variability and modularity of bacterial metabolic networks. BMC evolutionary biology **7**, 169 (2007). doi:10.1186/1471-2148-7-169
12. Takemoto, K.: Does Habitat Variability Really Promote Metabolic Network Modularity? PLoS ONE **8**(4), 2–10 (2013). doi:10.1371/journal.pone.0061348
13. Kreimer, A., Borenstein, E., Gophna, U., Ruppin, E.: The evolution of modularity in bacterial metabolic networks. Proceedings of the National Academy of Sciences of the United States of America **105**(19), 6976–81 (2008). doi:10.1073/pnas.0712149105
14. Lorenz, D.M., Jeng, A., Deem, M.W.: The emergence of modularity in biological systems. Physics of Life Reviews **8**(2), 129–160 (2011). doi:10.1016/j.plrev.2011.02.003. 1204.5999
15. Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N.: Hierarchical Organization of Modularity in Metabolic Networks **297**(August), 1551–1555 (2002)
16. Sales-Pardo, M., Guimerà, R., Moreira, A.A., Amaral, L.A.N.: Extracting the hierarchical organization of complex systems. Proceedings of the National Academy of Sciences of the United States of America **104**(39), 15224–15229 (2007). doi:10.1073/pnas.0703740104. 0705.1679
17. Braakman, R., Smith, E.: The compositional and evolutionary logic of metabolism. Physical biology **10**(1), 011001 (2013). doi:10.1088/1478-3975/10/1/011001. 1207.5532

bioRxiv preprint doi: https://doi.org/10.1101/118299; this version posted May 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

18. Muchnik, L., Itzhack, R., Solomon, S., Louzoun, Y.: Self-emergence of knowledge trees: Extraction of the Wikipedia hierarchies. Physical Review E **76**(1), 016106 (2007). doi:10.1103/PhysRevE.76.016106
19. Clune, J., Mouret, J.-B., Lipson, H.: The Evolutionary Origins of Hierarchy. PLOS Computational Biology **280**(1755), 20122863 (2016). doi:10.1098/rspb.2012.2863. 1207.2743v1
20. Nepusz, T., Vicsek, T.: Hierarchical self-organization of non-cooperating individuals. PLoS ONE **8**(12) (2013). doi:10.1371/journal.pone.0081449. 1308.0029
21. Hu, F., Zhao, S., Bing, T., Chang, Y.: Hierarchy in industrial structure: The cases of China and the USA. Physica A: Statistical Mechanics and its Applications **469**, 871–882 (2016). doi:10.1016/j.physa.2016.11.083
22. D'Souza, G., Waschina, S., Pande, S., Bohl, K., Kaleta, C., Kost, C.: Less is more: Selective advantages can explain the prevalent loss of biosynthetic genes in bacteria. Evolution **68**(9), 2559–2570 (2014). doi:10.1111/evo.12468
23. Mones, E., Vicsek, L., Vicsek, T.: Hierarchy measure for complex networks. PLoS ONE **7**(3), 1–10 (2012). doi:10.1371/journal.pone.0033799. 1202.0191
24. Lipson, D.A.: The complex relationship between microbial growth rate and yield and its implications for ecosystem processes. Frontiers in Microbiology **6**(JUN), 1–5 (2015). doi:10.3389/fmicb.2015.00615
25. Beardmore, R.E., Gudelj, I., Lipson, D.A., Hurst, L.D.: Metabolic trade-offs and the maintenance of the fittest and the flattest. Nature **472**(7343), 342–6 (2011). doi:10.1038/nature09905
26. Russell, J.B., Cook, G.M.: Energetics of bacterial growth: balance of anabolic and catabolic reactions. Microbiological reviews **59**(1), 48–62 (1995)
27. Sinsabaugh, R.L., Manzoni, S., Moorhead, D.L., Richter, A.: Carbon use efficiency of microbial communities: stoichiometry, methodology and modelling. Ecology Letters **16**(7), 930–939 (2013). doi:10.1111/ele.12113
28. Novak, M., Pfeiffer, T., Lenski, R.E., Sauer, U., Bonhoeffer, S.: Experimental tests for an evolutionary trade-off between growth rate and yield in E. coli. The American naturalist **168**(2), 242–51 (2006). doi:10.1086/506527
29. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America **99**(12), 7821–7826 (2002). doi:10.1073/pnas.122653799. 0112110
30. Da Silva, M.R., Ma, H., Zeng, A.P.: Centrality, network capacity, and modularity as parameters to analyze the core-periphery structure in metabolic networks. Proceedings of the IEEE **96**(8), 1411–1420 (2008). doi:10.1109/JPROC.2008.925418
31. Felsenstein, J.: Phylogenies and the Comparative Method. The American Naturalist **125**(1), 1–15 (1985). doi:10.1086/284325
32. Shapiro, B.J., Polz, M.F.: Ordering microbial diversity into ecologically and genetically cohesive units. Trends in Microbiology **22**(5), 235–247 (2014). doi:10.1016/j.tim.2014.02.006. NIHMS150003
33. Dekel, E., Mangan, S., Alon, U.: Environmental selection of the feed-forward loop circuit in gene-regulation networks. Phys. Biol. Phys. Biol **2**(2), 81–81 (2005). doi:10.1088/1478-3975/2/2/001
34. Leskovec, J., Sosič, R.: SNAP: A General-Purpose Network Analysis and Graph-Mining Library. ACM Transactions on Intelligent Systems and Technology **8**(1), 1–20 (2016). doi:10.1145/2898361
35. Chiu, H.C., Levy, R., Borenstein, E.: Emergent Biosynthetic Capacity in Simple Microbial Communities. PLoS Computational Biology **10**(7) (2014). doi:10.1371/journal.pcbi.1003695
36. Zaneveld, J., Turnbaugh, P.J., Lozupone, C., Ley, R.E., Hamady, M., Gordon, J.I., Knight, R.: Host-bacterial coevolution and the search for new drug targets. Current Opinion in Chemical Biology **12**(1), 109–114 (2008). doi:10.1016/j.cbpa.2008.01.015. NIHMS150003
37. DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., Andersen, G.L.: Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and Environmental Microbiology **72**(7), 5069–5072 (2006). doi:10.1128/AEM.03006-05
38. Caporaso, J.G., Bittinger, K., Bushman, F.D., Desantis, T.Z., Andersen, G.L., Knight, R.: PyNAST: A flexible tool for aligning sequences to a template alignment. Bioinformatics **26**(2), 266–267 (2010). doi:10.1093/bioinformatics/btp636
39. Kimura, M.: A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution **16**(2), 111–120 (1980). doi:10.1007/BF01731581. arXiv:1011.1669v3

bioRxiv preprint doi: https://doi.org/10.1101/118299; this version posted May 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

40. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M.: KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research **27**(1), 29–34 (1999). doi:10.1093/nar/27.1.29
41. Stelzer, M., Sun, J., Zeng, A.-P., Kamphans, T., Fekete, S.P.: An extended bioreaction database that significantly improves reconstruction and analysis of genome-scale metabolic networks. Journal of Chemical Information and Modeling **53**, 1689–1699 (2013). doi:10.1017/CBO9781107415324.004. arXiv:1011.1669v3
42. Stamatakis, A.: RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**(9), 1312–1313 (2014). doi:10.1093/bioinformatics/btu033. bioinformatics/btu033
43. Couturier, E., Rocha, E.P.C.: Replication-associated gene dosage effects shape the genomes of fast-growing bacteria but only for transcription and translation genes. Molecular Microbiology **59**(5), 1506–1518 (2006). doi:10.1111/j.1365-2958.2006.05046.x

bioRxiv preprint doi: https://doi.org/10.1101/118299; this version posted May 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

	Correlation	Partial Correlation
Node Count	0.26^{***}	0.03^{***}
Edge Count	0.27^{***}	-0.02^{*}
Modularity	0.28^{***}	0.05^{***}
GRC Hierarchy	0.28^{***}	0.17^{***}
Luo Hierarchy	0.22^{***}	-0.06^{***}
Largest SCC Fraction	0.30^{***}	0.12^{***}
Cluster Coefficient	0.13^{***}	0.03^{***}
Full Diameter	0.23^{***}	0.06^{***}
Effective Diameter	0.20^{***}	0.01
SCC Count	0.14^{***}	0.02^{**}
Mean Degree	0.24^{***}	-0.04^{***}

Table 1 Correlation of network statistics with phylogenetic distances, and partial correlation of network statistic with phylogenetic distance, controlling for the other variables. The correlation and partial correlation of GRC Hierarchy metric with genetic distance is higher than all other non-trivial metrics. ***: p value < 0.001, **: p value < 0.01, *: p value < 0.05.

bioRxiv preprint doi: https://doi.org/10.1101/118299; this version posted May 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

List of Figures

1 Histogram of GRC hierarchy scores of the 2,935 bacteria in the KEGG database. Mean degree of hierarchy is 0.279. Ranging from 0.065 for the insect symbiote Candidatus *Nasuia deltocephalinicola* to 0.385 for a *Blattabacterium* endosymbiont of *Nauphoeta cinerea*, an insect endosymbiont. The hierarchy score for *E. coli* strains is 0.269. 16

2 The relationship between hierarchical character and growth rate reflects fundamental tradeoffs between growth and yield, and is informative of the ecological niche the bacteria occupy. Overall growth rate is inversely correlated with hierarchy $\rho = -0.31$, $p = 0.00065$. When controlling for bacterial environment $\rho = -0.41, p = 0.0001$ The outlier in the facultative parasite pane is *Borrelia burgdorferi*, which is an obligate parasite that alternates between insect and vertebrate hosts, and thus is similar to the obligate parasites. The particular strain also lacks a number of enzymes in glycolysis pathway that are present in other *B. burgdorferi* strains which have hierarchy scores of 0.183 ± 0.002. 17

3 Hierarchy has a slight overall decreases with phylogenetic distance. Spearman's rank correlation, $\rho = -0.06, p < 10^{-6}$. Hierarchy appears to increase and then decrease further from the root of the tree. . . . 18

4 Modularity increases with phylogenetic distance. Spearman's rank correlation, $\rho = 0.31$, $p < 10^{-15}$. 19

bioRxiv preprint doi: https://doi.org/10.1101/118299; this version posted May 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

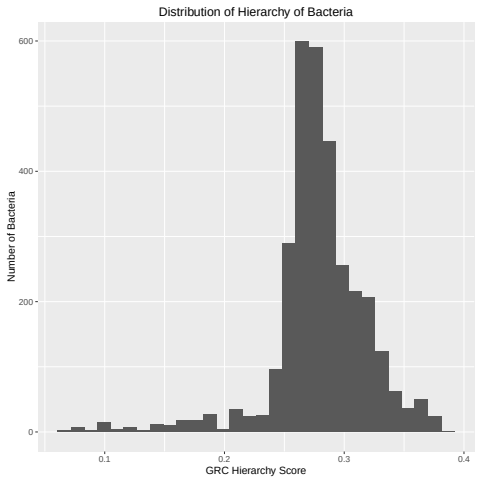

Figure 1 Histogram of GRC hierarchy scores of the 2,935 bacteria in the KEGG database. Mean degree of hierarchy is 0.279. Ranging from 0.065 for the insect symbiote Candidatus *Nasuia deltocephalinicola* to 0.385 for a *Blattabacterium* endosymbiont of *Nauphoeta cinerea*, an insect endosymbiont. The hierarchy score for *E. coli* strains is 0.269.

bioRxiv preprint doi: https://doi.org/10.1101/118299; this version posted May 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

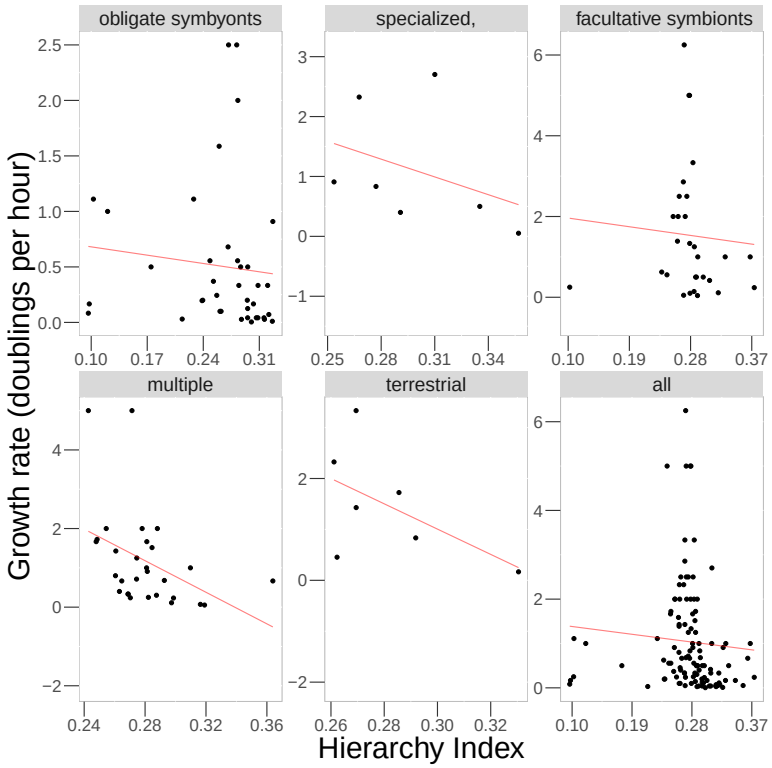

Figure 2 The relationship between hierarchical character and growth rate reflects fundamental tradeoffs between growth and yield, and is informative of the ecological niche the bacteria occupy. Overall growth rate is inversely correlated with hierarchy $\rho = -0.31$, $p = 0.00065$. When controlling for bacterial environment $\rho = -0.41, p = 0.0001$ The outlier in the facultative parasite pane is *Borrelia burgdorferi*, which is an obligate parasite that alternates between insect and vertebrate hosts, and thus is similar to the obligate parasites. The particular strain also lacks a number of enzymes in glycolysis pathway that are present in other *B. burgdorferi* strains which have hierarchy scores of 0.183 ± 0.002.

bioRxiv preprint doi: https://doi.org/10.1101/118299; this version posted May 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

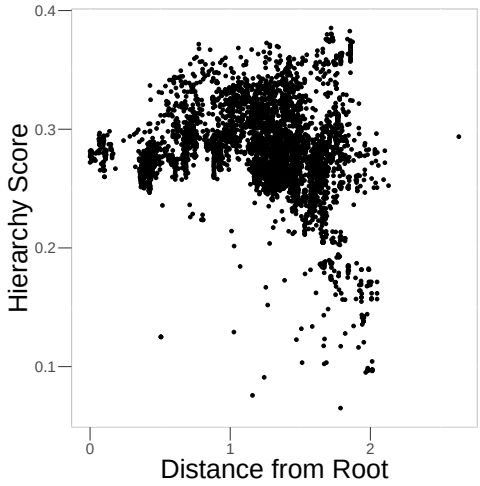

Figure 3 Hierarchy has a slight overall decreases with phylogenetic distance. Spearman's rank correlation, $\rho = -0.06, p < 10^{-6}$. Hierarchy appears to increase and then decrease further from the root of the tree.

bioRxiv preprint doi: https://doi.org/10.1101/118299; this version posted May 9, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

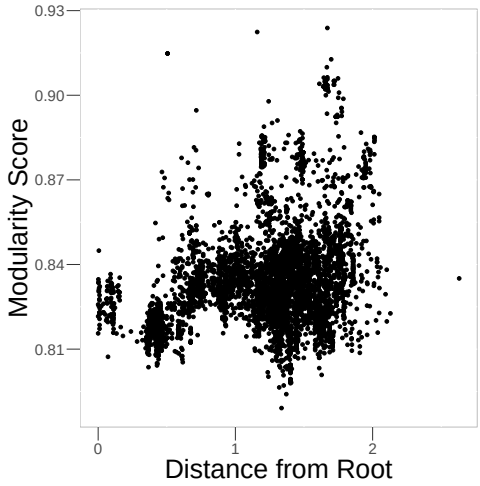

Figure 4 Modularity increases with phylogenetic distance. Spearman's rank correlation, $\rho = 0.31$, $p < 10^{-15}$.