The Large-scale Structure of Metabolic Networks: A Glimpse at Life's Origin?

Mapping an Organism's Network Structure

ANDREAS WAGNER

Metabolism comprises the network of interactions that provide energy and building blocks for cells and organisms, a network sustaining the living and allowing it to grow and reproduce. For well-studied organisms, especially microbes such as *Escherichia coli,* considerable information about metabolic reactions has been accumulated through decades of experimental work. This information, originally scattered through thousands of articles of original literature, has increasingly found its way into larger collections including encyclopedias [1] and online databases [2,3]. With easier availability of this information, it has become feasible to map the structure of large pieces of an organism's metabolic network.

REPRESENTING METABOLIC NETWORKS

David Fell and I [4,5] assembled a list of 317 stoichiometric equations involving 287 substrates that represent the central routes of energy metabolism and small-molecule building block synthesis in *E. coli*. Because there is considerable variation in the metabolic reactions realized under different environmental conditions, we included only reactions that would occur under one condition: aerobic growth on minimal medium with glucose as sole carbon source and O_2 as electron acceptor. We also deliberately omitted (i) reactions whose occurrence is reportedly strain dependent [1], (ii) biosyntheses of complex cofactors (e.g., adenosyl-cobalamine), which are not fully understood, and (iii) syntheses of most polymers (RNA, DNA, protein) because of their complex stoichiometry.

When faced with a complex assemblage of chemical reactions, the problem arises immediately of how to represent the resulting reaction network. Importantly, for most reactions only qualitative information is available—one may know the substrates and stoichiometry of a reaction but not much more. A mathematical representation that captures such qualitative information is that of a graph, for example, that of a substrate graph $G_S = (V_S, E_S)$. Its vertex set V_S consists of all chemical compounds (substrates) that occur in the network. Two substrates S_1, S_2 are *adjacent* if there exists an edge *e,* i.e., $e = (S_1, S_2) \in E_S$, the edge set of this graph, if the two substrates occur (either as substrates or products) in the same chemical reaction. Such a network representation has the advantage of being intuitive and simple. Other graph-like representations of metabolic networks are possible, including bipartite graphs and hypergraphs [6]. However, hypergraphs are much less intuitive constructs than graphs, and the many analysis tools available for graphs have not yet been developed to the same extent for other graph representations. One might argue that the existence of irreversible chemical reactions would suggest a directed graph

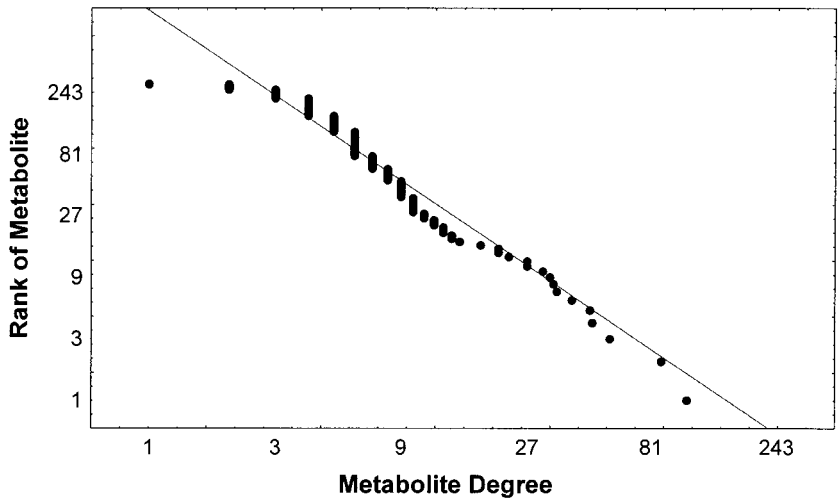
Andreas Wagner is an evolutionary biologist and an associate professor at the University of New Mexico (Department of Biology, 167A Castetter Hall, Albuquerque, NM 87131-1091; e-mail: wagnera@unm.edu). He is also a member of the external faculty at The Santa Fe Institute. He received his Ph.D. in 1995 from Yale University and became an Assistant Professor after two postdoctoral stints at the Institute for Advanced Study in Berlin, Germany (1995–1996) and SFI (1996–1998). He is interested in the evolution of genetic and metabolic networks, the evolution of robustness and evolvability in genetic systems as well as the evolution of organismal design features such as

© 2003 Wiley Periodicals, Inc.,

[6] representation, i.e., a graph where each edge has a direction. However, a directed graph representation would be inappropriate for an important application of graph representations of biological networks: to assess qualitatively how perturbations of either enzyme concentrations—via mutation—or substrate concentrations—via changes in consumption or availability—propagate along the network. The reason is that even for irreversible reactions, the concentration of a reaction product potentially affects the reaction rate by occupancy of the enzyme's active site. Reaction products can thus affect substrate concentrations "upstream" of irreversible reactions.

What is the structure of the *E. coli* substrate graph? It has much in common with graphs found in areas as different as computer science (the worldwide Web) and sociology (friendship and collaboration networks). It is a small world network [7], meaning that any two nodes (substrates) can be reached from each other through a path of very few edges, fewer than in other graphs of comparable size. Also, the distribution of the vertex degree d, the number of edges d connecting each substrate to other substrates is consistent with a power law, i.e., the probability $P(d)$ of finding a vertex with degree d is $P(d) \propto d^{-\tau}$ (Figure 1). (The exponent is less than two but can not be estimated very accurately due to small network size.) This degree distribution has also been found for reaction networks derived from different organisms [8]. It seems to be a universal characteristic of metabolic networks.

FIGURE 1



Connective distribution of *Escherichia coli* core metabolism. Metabolites were ranked according to the number of connections (degree) they have in the substrate graph. Shown is metabolite rank versus degree on a log-log scale. If D is a random variable describing metabolite degree, this rank plot estimates the counter-cumulative probability function $P(\log D > k)$. The data are consistent with a power-law distribution of D, i.e., $P(\log D > k) \propto e^{-k\tau}$ and thus $P(D > k) \propto k^{-\tau}$. However, little confidence can be placed in the estimated value of the exponent $\tau = 1.3$ because of the small network size. The following metabolic functions were included in the network whose degree distribution is presented: Glycolysis (12 reactions), pentose phosphate and Entner-Doudoroff pathways (10), glycogen metabolism (5), acetate production (2), glyoxalate and anaplerotic reactions (3), tricarboxylic acid cycle (10), oxydative phosphorylation (6), amino acid and polyamine biosynthesis (95), nucleotide and nucleoside biosynthesis and folate synthesis (72) and 1-carbon metabolism (16), glycerol 3-phosphate and membrane lipids (17), riboflavin (9), coenzyme A (11), NAD(P) (7), porphyrins, heme, and sirohaem (14), lipopolysaccharides and murein (14), pyrophosphate metabolism (1), transport reactions (2), glycerol 3-phosphate production (2), isoprenoid biosynthesis and quinone biosynthesis (13).

POWER LAWS AND ROBUSTNESS

Two complementary hypotheses figure prominently in explaining power-law degree distributions. First, Albert and collaborators [9] found that networks with power-law distributed degrees are robust to random perturbations in the following sense. Upon removal of randomly chosen nodes, the mean distance between network nodes that can still be reached from each other (via a path of edges) increases only very little. This distance is also known as the network diameter. In graphs with other degree distributions, network diameter can increase substantially upon node removal. Also, graphs with power-law degree distributions fragment less easily into large disconnected subnetworks upon random node removal. These findings have led Jeong and collaborators [8] to suggest that metabolic network graphs with power-law distributed degrees have such a degree distribution *because* this distribution provides robustness against perturbations.

It is difficult to assess the merit of this hypothesis for metabolic networks directly, for doing so would require comparing large metabolic networks of different structure. However, the ensemble of core metabolic reactions is very similar in most free-living organisms, and thus the global structure of metabolism is highly conserved. In addition, it is not easy to identify (i) the kinds of perturbations to which a metabolic networks would have adapted over billions of years and (ii) the reasons why short path lengths would provide an advantage to the organism. At most, one can venture an informed speculation. For metabolic networks, a possible advantage of small mean path

modularity. His lab addresses questions in this research area with a variety of approaches that range from functional genomics to mathematical population genetic modeling. A more detailed account including publications can be found at http://samba.unm.edu/-wagnera/.

© 2003 Wiley Periodicals, Inc.

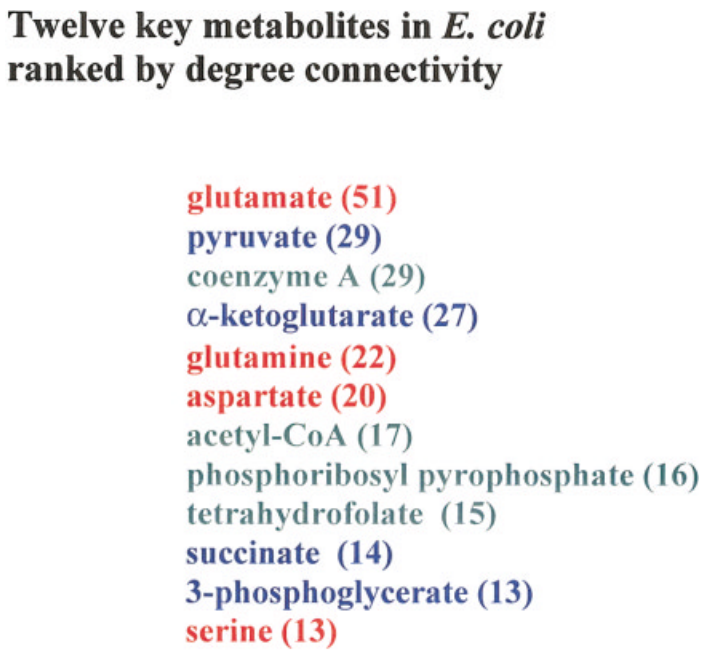
lengths stems from the importance of minimizing transition times between metabolic states in response to environmental changes [10–12]. Networks with robustly small average path lengths thus might adjust more rapidly to environmental change.

In contrast to this weak case for the selectionist explanation of the degree distribution, there may be a stronger case against it. One might ask whether power-law degree distributions might not be features of many or all large chemical reaction networks, whether part of an organism or not, and regardless of whether they perform any function that benefits from a robust network diameter. If so, then metabolic network degree distributions would join the club of other power-laws (such as Zipf's law of word frequency distribution in natural languages) whose existence does not owe credit to a benefit they provide.

Gleiss et al. [13] have assembled public information on a class of large chemical reaction networks that exist not only outside the living, but on spatial scales many orders of magnitude larger than organisms. These are the chemical reaction networks of planetary atmospheres, networks largely shaped by the photochemistry of their component substrates. The available data stems not only from earth's atmosphere, but also from other solar planets including Venus and Jupiter, planets with chemically vastly different atmospheres. These planets' atmospheres have been explored through remote spectroscopic sensing methods and through visits by planetary probes. The chemical reaction networks in these atmospheres, despite being vastly different in chemistry, have a degree distribution consistent with a power law [13]. This suggests that power-law distributions may be very general features of chemical reaction networks. The reasons why we observe them in cellular reaction networks may have nothing to do with the robustness they may provide.

The ensemble of core metabolic reactions is very similar in most free-living organisms, and thus the global structure of metabolism is highly conserved.

FIGURE 2



Highly connected metabolites in *Escherichia coli* are evolutionarily old. The list shows the 12 most highly connected metabolites in the *E. coli* core intermediary substrate network. The numbers in parentheses shows the degree (number of neighbors) of a metabolite in the substrate network as defined in the text. Green indicates the proposed remnants of a surface metabolism or an RNA world; red indicates the proposed early amino acids, and blue, the proposed early metabolites (in the tricarboxylic acid cycle or in glycolysis). The network was generated after the elimination of the compounds NAD, ATP, and their derivatives. These are even more highly connected than the compounds shown here. They are also evolutionarily ancient. See text for further details.

POWER LAWS AND DEEP TIME

Metabolic networks have a history. They have not been assembled in their present state at once. They have grown, perhaps over billion years, as organisms increased their metabolic and biosynthetic abilities. Having to take into account this history raises a question: How does a network arrive at a power-law degree distribution if it grows? The perhaps simplest mathematical model capable of growing power-law distributed networks involves only two simple rules [14]. First and unsurprisingly, it adds nodes to a graph. Second, it connects this node to previously existing nodes according to a second rule, where already highly connected nodes are more likely to receive a new connection than nodes of lesser connectivity. Over many node additions, a power-law degree distribution emerges. A great variety of variations to this model have been proposed (reviewed in {15]). They differ greatly in detail but retain in some way or another the rule that new connections preferably involve highly connected nodes. But more importantly, many of these models make a key prediction: Highly connected nodes are old nodes, nodes having been added very early in a network's history.

We may never know enough about the history of life and metabolism to distinguish between different ways in which metabolism might have grown. However, we can address this latter prediction, common to many different growth models. *Are highly connected metabolites old metabolites?* The answer will contain a speculative element, because the oldest metabolites are those that arose in the earliest days of the

© 2003 Wiley Periodicals, Inc.

living, close to life's origins. Also, that life forms as different as bacteria and humans have very similar metabolic structure suggests that the growth of metabolism has essentially been completed at the time the common ancestor of extant life emerged. The detailed structure of metabolism at this early time may remain in the dark forever. However, origin of life hypotheses make some clear predictions on the chemical compounds expected to have been part of early organisms. There are several of these hypotheses, and they are complementary in the respect most important here: They emphasize the origins of different aspects of life's chemistry. Some emphasize the origins of early genetic material (RNA). Others make postulates about the composition of the earliest proteins. Yet others ask about the earliest metabolites in energy metabolism. Each of them makes a statement about a different aspect of early life's chemistry.

Figure 2 shows the 12 most highly connected metabolites of the *E. coli* metabolic network graphs. Every single one of them has been part of early organisms according to at least one origin-of-life hypothesis. Colored in green are compounds such as coenzyme A, thought to have been a part of early RNA-based organisms [16]. The RNA moieties they contain are present in all organismal lineages. Some compounds in this group, such as tetrahydrofolate and coenzyme A, are thought to have played a role in precellular life that may have taken place on polykationic surfaces. Their merit in this regard is that they are elongate molecules with one anionic terminus. They are therefore able to flexibly tether other molecules to the substrate, thus localizing them while simultaneously increasing their potential to react with other compounds [17]. Colored in red in Figure 2 are amino acids that were most likely part of early proteins. This postulate is based on likely scenarios for the early evolution of the genetic code [18]. Shown in blue are compounds likely to be a part of the earliest energy and biosynthetic metabolism. Glycolysis and the TCA cycle are perhaps the most ancient metabolic pathways, and various of their intermediates (α-ketoglutarate, succinate, pyruvate, 3-phosphoglycerate) occur in Figure 2 [16,18–22]. The potential relation between evolutionary history and degree connectivity of metabolites corroborates a postulate put forth and defended forcefully by Morowitz [20], namely that intermediary metabolism recapitulates the evolution of biochemistry.

Thus, although the structure of metabolic networks may not be a reflection of their robustness, it may teach us about their history. Functional genomic experiments are unearthing the structure of many other genetic networks [23–25], some of which show a power-law degree distribution [5,26,27]. Perhaps their structure can also teach us important lessons about their ancient history.

Although the structure of metabolic networks may not be a reflection of their robustness, it may teach us about their history.

ACKNOWLEDGMENTS

I gratefully acknowledge financial support through National Institutes of Health grant GM63882.

REFERENCES

1. Neidhardt, F.C. *Escherichia coli* and Salmonella. ASM Press: Washington, DC, 1996.
2. Karp, P.; Riley, M.; Paley, S.; Pellegrini-Toole, A.; Krummenacker, M. EcoCyc: Electronic encyclopedia of *E. coli* genes and metabolism. Nucleic Acids Res 1999, 27, 55.
3. Ogata, H., Goto, S. et al. KEGG-Kyoto Encyclopedia of genes and genomes. Nucleic Acids Research 27, 29–34, 1999.
4. Fell, D.; Wagner, A. The small world of metabolism. Nat Biotechnol 2000, 18, 1121–1122.
5. Wagner, A.; Fell, D. The small world inside large metabolic networks. Proc Roy Soc London Ser B 2001, 280, 1803–1810.
6. Graham, R.L.; Groetschel, M.; Lovasz, L. Handbook of combinatorics. MIT Press: Cambridge, 1995.
7. Watts, D.J. Small Worlds. Princeton University Press: Princeton, NJ, 1999.
8. Jeong, H.; Tombor, B.; Albert, R.; Oltvai, Z.N., Barabasi, A.L. The large-scale organization of metabolic networks. Nature 2000, 407, 651–654.
9. Albert, R.; H. Jeong, Barabasi, A.L. Error and attack tolerance of complex networks. Nature 2000, 406, 378–382.
10. Easterby, J.S. The effect of feedback on pathway transient-response. Biochem J 1986, 233, 871–875.
11. Schuster, S.; Heinrich, R. Time hierarchy in enzymatic—reaction chains resulting from optimality principles. J Theor Biol 1987, 129, 189–209.
12. Cascante, M.; Melendezhevia, E.; Kholodenko, B.; Sicilia, J.; Kacser, H. Control analysis of transit-time for free and enzyme-bound metabolites: Physiological and evolutionary significance of metabolic response-times. Biochem J 1995, 308, 895–899.
13. Gleiss, P.M.; Stadler, P.F.; Wagner, A.; Fell, D.A. Small cycles in small worlds. Adv Complex Systems 2001, 4, 207–226.
14. Barabasi, A.L.; Albert, R.; Jeong, H. Mean-field theory for scale-free random networks. Physica A 1999, 272, 173–187.
15. Albert, R.; Barabasi, A.-L. Statistical mechanics of complex networks. 2002, 74(1), 47–97.
16. Benner, S.A.; Ellington, A.D.; Tauer, A. Modern metabolism as a palimpsest of the RNA world. Proc Natl Acad Sci USA 1989, 86, 7054–7058.
17. Wachtershauser, G. Before enzymes and templates: Theory of surface metabolism. Microbiol Rev 1988, 52, 452–484.
18. Kuhn, H.; Waser, J. On the origin of the genetic code. FEBS Lett 1994, 352, 259–264.
19. Taylor, B.L.; Coates, D. The code within the codons. Biosystems 1989, 22, 177–187.
20. Morowitz, H.J. Beginnings of cellular life. Yale University Press: New Haven, 1992.
21. Waddell, T.G.; Bruce, G.K. A new theory on the origin and evolution of the citric acid cycle. Microbiol Sem 1995, 11, 243–250.
22. Lahav, N. Biogenesis. Oxford University Press: New York, 1999.

© 2003 Wiley Periodicals, Inc.

23. Hughes, T.R.; Marton, M.J.; Jones, A.R.; Roberts, C.J.; Stoughton, R.; Armour, C.D.; Bennett, H.A.; Coffey, E.; Dai, H.Y.; He, Y.D.D.; Kidd, M.J.; King, A.M.; Meyer, M.R.; Slade, D.; Lum, P.Y.; Stepaniants, S.B.; Shoemaker, D.D.; Gachotte, D.; Chakraburtty, K.; Simon, J.; Bard, M.; Friend, S.H. Functional discovery via a compendium of expression profiles. Cell 2000, 102, 109–126.
24. Uetz, P.; Giot, L.; Cagney, G.; Mansfield, T.A.; Judson, R.S.; Knight, J.R.; Lockshon, D.; Narayan, V.; Srinivasan, M.; Pochart, P.; QureshiEmili, A.; Li, Y.; Godwin, B.; Conover, D.; Kalbfleisch, T.; Vijayadamodar, G.; Yang, M.J.; Johnston, M.; Fields, S.; Rothberg, J.M. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 2000, 403, 623–627.
25. Ito, T.; Chiba, T.; Ozawa, R.;Yoshida, M.; Hattori, M.; Sakaki, Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci USA 2001, 98, 4569–4574.
26. Jeong, H.; Mason, S.P.; Barabasi, A.-L.; Oltvai, Z.N. Lethality and centrality in protein networks. Nature 2001, 411, 41–42.
27. Wagner, A. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. Mol Biol and Evol 2001, 18, 1283–1292.

© 2003 Wiley Periodicals, Inc.