The Small World Inside Large Metabolic Networks

Andreas Wagner
David Fell

SFI WORKING PAPER: 2000-07-041

SANTA FE INSTITUTE

The small world inside large metabolic networks

Andreas Wagner[1,2], David Fell[3]

[1] University of New Mexico, Department of Biology, 167A Castetter Hall
Albuquerque, NM 817131-1091
Tel. +1-505-277-2021; FAX +1-505-277-0304;
wagnera@unm.edu

[2] The Santa Fe Institute, 1399 Hyde Park Road,
Santa Fe, NM 87501

[3] School of Biological & Molecular Sciences,
Oxford Brookes University, Oxford OX3 0BP, UK.
Tel. +44 (0)1865 483247; FAX +44 (0)1865 484017;

Abstract

We analyze the structuture of a large metabolic network, that of the energy and biosynthesis metabolism of *Escherichia coli*. This network is a paradigmatic case for the large genetic and metabolic networks that functional genomics efforts are beginning to elucidate. To analyze the structure of networks involving hundreds or thousands of components by simple visual inspection is impossible, and a quantitative framework is needed to analyze them. We propose a graph theoretical description of the *E. coli* metabolic network, a description that we hope will prove useful for other genetic networks. We find that this network is a small world graph, a type of graph observed in a variety of seemingly unrelated areas, such as friendship networks in sociology, the structure of electrical power grids, and the nervous system of *C. elegans*. Moreover, its connectivity follows a power law, another unusual but by no means rare statistical distribution. This architecture may serve to minimize transition times between metabolic states, and also reflect the evolutionary history of metabolism.

The information necessary to characterize the genetic and metabolic networks driving all functions of a living cell is being put within our reach by various genome projects. With the availability of this information, however, a problem so far unknown to molecular biologists will arise: how to adequately represent and describe the structure of large genetic networks. While it is trivial to understand the structure of metabolic pathways, transcriptional cascades, or signalling pathways that consist of a small number of genes, networks consisting of anywhere from hundreds to tens of thousands of components are less easily described. If networks encountered in areas ranging from neurobiology to sociology are any guide, we can assume that the structure of genetic and metabolic networks will be a complex mix of ordered and random elements. However, a body of theory suitable to analyze large networks exists only for networks at either end of this spectrum, networks that are perfectly orderly or completely random. Thus, a quantitative framework suitable to describe the structure of large biological networks remains to be developed. A *description* of genetic and metabolic networks is, however, only a first step towards a second, deeper level of inquiry, which asks for an *explanation* of the structure of such networks. Needles to say, any explanation of a network's organization, whether it invokes historical accidents or functional optimality principles requires a prior understanding of what that organization is.

Here, we analyze the structure of a large metabolic network, that of the *Escherichia coli* intermediary metabolism for energy generation and small building block synthesis. To obtain insight into both structural and functional aspects of this network, we first propose and motivate a mathematical representation suitable for our purpose. We then introduce some quantities useful to describe the network's structure. Significantly, we find that this structure is similar to one observed for networks in several unrelated areas of science, such as sociology. Qualitatively new hypotheses on network function and evolution can be based upon the detection of such global organizational patterns. While this attempt is only a modest beginning, we hope that it points towards the kind of global properties of biological networks that will be found in years to come, and that similar approaches will lead to an understanding of the evolutionary forces shaping such networks.

Metabolic and genetic networks can be represented at multiple levels of mathematical resolution. The most fine-grained of these involves detailed modeling of each gene products' activity as a function of substrate, product, and effector concentration. This level of detail may be difficult to achieve even for the smallest networks, because individual enzymes may show quite complex behavior. A case in point is glutamine synthase, which has at least nine small-molecule effectors regulating its activity [1,2]. Modeling an entire metabolism on this level of detail is certainly hopeless. A second, coarser level of representation is metabolic control analysis, which requires less detailed information about enzyme kinetics. The hundreds of steps for which even this smaller amount of information would be required, together with needed information on metabolic fluxes and concentrations, makes this representation currently impractial for a network of this magnitude. A third and more manageable level of representation is that of stoichiometric equations describing network reactions (Fig. 1a), a representation for which the apparatus of metabolic flux balance analysis and linear programming can provide useful information on admissible steady state metabolic fluxes [3-6] Here, we choose a fourth, graph theoretic representation derived from stoichiometric equations. Our motivation to use a graph representation of metabolism instead of analyzing metabolic flux is twofold. First, in graph theory a mathematical toolbox is already in place that can be used to describe the structure of metabolic networks, and that is currently not available for stoichiometric equations. Second, a graph theoretic representation is a common denominator permitting a comparison of seemingly unrelated networks, which has recently led to the exposure of deep similarities among networks found in sociology, engineering, and neurobiology [7]. While undoubtedly crude, we note that a graph model still allows a qualitative analysis of network function, for example in terms of how disturbances propagate through a network [7].

Based on publicly available information [8-12] we 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 attempted to include only those that would occur under one particular condition: aerobic growth on minimal medium with glucose as sole carbon source and O_2 as electron acceptor. We deliberately omitted (i) reactions whose occurrence is reportedly strain-dependent [8], (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.

Our metabolic map comprises the following pathways: 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 (72), folate synthesis 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). The reaction list is available from DAF upon request.

From these reaction equations, a stoichiometric matrix [3] was automatically generated from the reaction list using the software package INDIGO [13] (http://members.tripod.co.uk/sauro/biotech.htm). From this matrix, the substrate and reaction graph were derived under omission of the following metabolites: CO_2, NH_3, SO_4, thioredoxin (oxidized and reduced form), organic phosphate (P_i) and pyrophosphate (PP_i). Graph analysis software was written in C++ using the LEDA library of data types [14].

We will consider two complementary representations of a metabolic network. The first of these is the 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 they occur (either as substrates or products) in the same chemical reaction (Fig. 1b). Second, consider the reaction graph $G_R=(V_R, E_R)$. Its vertex set V_R shall consist of all chemical reactions in the network. Two reactions are adjacent if there exists an edge, i.e., $(R_1, R_2) \in E_R$, the edge set of the reaction graph, if they share at least one chemical compound, either as substrate or as product (see Fig. 1c).

We will now briefly discuss why we have avoided two obvious alternatives to these representations. First, perhaps the most natural representation of a metabolic network is that of a hypergraph [15]. However, hypergraphs are much less intuitive constructs than graphs, and the tools our analysis needs have not yet been developed for them. Second, one might argue that the existence of irreversible chemical reactions would make a directed graph [15] a better choice, i.e., a graph where each edge has a direction. For instance, one might wish to connect a substrate S and a product P of an irreversible reaction as $S \rightarrow P$. However, we deliberately avoid directed graphs here for the following reasons. One of the uses of a graph representation of metabolic networks is to assess qualitatively how perturbations of either enzyme concentrations (e.g., via mutation) or substrate concentrations (e.g., via changes in consumption or availability) propagate along the network. A directed graph representation, however, would not accurately reflect how perturbations at individual vertices propagate. Consider the *substrate graph*, and a perturbation in the concentration of a compound that is the product of an irreversible reaction. Even for irreversible reactions, the concentration of a reaction product potentially affects the reaction rate by occupancy of the active site. Thus, a perturbation in a product concentration downstream of an irreversible reaction (S → P) can affect the reaction rate, and thus substrate concentrations "upstream" of that reaction. A directed substrate graph would not capture this behavior, because by definition P can not influence S in this representation. In a similar vein, in control analysis, flux and concentration control coefficients of an enzyme [16] and not the reversibility of the reaction, show how it is possible for a change in concentration of an enzyme to propagate into the part of the network "upstream" of the reaction [17,18]. A directed *reaction graph* would not capture this behavior. The two graph representations used here are complementary and obviously related. Consider three substrates S_1-S_3 in the substrate graph, connected by two reactions R_1, R_2. Then, R_1 and R_2 are nodes in the reaction graph connected by an edge corresponding to S_2. Conversely, consider three reactions R_1-R_3 in the reaction graph, connected by two substrates S_1, S_2. Then, S_1 and S_2 are vertices of the substrate graph connected by reaction R_2.

The terminology introduced now will apply equally to both types of graph [15]. The *degree k* of a vertex is the number of other vertices it is adjacent to. Two vertices v_0, v_i are *connected* if there exists a *path*, i.e., a sequence of adjacent vertices $v_0, v_1, ..., v_{i-1}, v_i$ from v_0 to v_i. We will be concerned only with connected graphs, i.e., graphs where all vertex pairs are connected. Notice that (i) the law of mass conservation, and (ii) the fact that the carbon of all biomass is ultimately derived from CO_2 imply that metabolic networks are connected. The path length *l* is defined as the number of edges in the shortest path between v_0 and v_i. The *characteristic path length L* of a graph is the pathlength between two vertices, averaged over all pairs of vertices. Another important quantity [19] is the *clustering coefficient C(v)* of a vertex *v*. Consider all k_v vertices adjacent to a vertex *v*, and count the number *m* of edges that exist among these k_v vertices (not including edges connecting them to *v*). The maximally possible *m* is $k_v(k_v-1)/2$, in which case all *m* vertices are connected to each other, and we define $C(v):=m/(k_v(k_v-1)/2)$. *C(v)* measures the "cliquishness" of the neighborhood of *v*, i.e., what fraction of the vertices adjacent to v are also adjacent to each other. In extension, the clustering coefficient *C* of the graph is defined as the average of *C(v)* over all *v*. It is best viewed as a measure of the graph's "cliquishness" (see also below).

In analogy to statistics, where hypotheses are tested by comparing a data set to some random distribution, we will find it useful to compare the properties of the metabolic graph to a reference graph, a random graph with the same number of vertices *n* and mean degree *k*. This is not to say that we believe that metabolic network graphs will be well approximated by random graphs, only that random graphs provide a useful benchmark to evaluate exactly how metabolic graphs are different from them. This is made feasible by the available statistical theory of random graphs [20]. Importantly, random connectivity and a close variant, *k*-regular random connectivity [15], have frequently been the assumptions of choice during more than three decades of modeling genetic networks [21-24]. It is thus useful to see how the actual structure of a cell biological network (albeit not a regulatory one) relates to one key assumption made in this tradition.

In connected sparse random graphs with *n* nodes and average degree *k* (k«n) , the probability *p* of two vertices being connected is given by $p=k/(n-1)$. Such graphs show (i) a binomial distribution of vertex degree *k*, (ii) a very small clustering coefficient $C=(k-1)/n$, close to the theoretically attainable minimum of zero for large *n*, and (iii) a characteristic path length that is also close to the theoretically attainable minimum, although no closed mathematical form exists [7]. Thus, among all connected graphs with the same number of vertices and edges, random graphs are among the most rapidly traversed.

Variation in connectivity of metabolic networks greatly exceeds that of random graphs. Because of the ubiquity of the metabolites adenosine triphosphate (ATP), adenosine diphosphate (ADP), nicotinamide adenine dinucleotide (NAD), as well as its phosphorylated and reduced forms [1], we explored two situations, one in which these metabolites are included, and another one in which they are omitted. Table 1 shows basic connectivity statistics for reaction and substrate graphs representing the central energy and biosynthetic metabolism of *Escherichia coli.* Similar to networks found in neurobiology or ecology [25,26], metabolic graphs are sparse (Table 1). That is to say that the average degree of each vertex (metabolite or reaction) is small compared to the maximally possible degree *k=n-1*. For our *E. coli* network, *k* is of order *log n*. In a random graph with *n* nodes and probability *p* of two nodes being connected, the degree of each vertex follows a binomial distribution with variance *(n-1)p(1-p)*. The variance in degree for the metabolic graphs, however, is up to 20-fold greater than that of the corresponding random graph with *p=k/(n-1)*. This implies that some vertices in metabolic graphs have many more, and others many fewer neighbors than vertices for a random graph. Given this enormous dispersion, *k*-regular random graphs would be particularly poor statistical models of metabolic networks.

Comparison to random graphs also lends itself to a statistical definition of "key-metabolites" or "key-reactions", particularly highly connected vertices in metabolite graphs . For example, for the substrate graph, one might define a key metabolite as one whose vertex degree k_m exceeds the average *k* by three standard deviations,

$$k_m > k + 3\sigma_{random} = k + 3\sqrt{\frac{k(n-1-k)}{n-1}}.$$

Applying this to the substrate graph with *k*=4.76 (Table 1), leads to k_m>11.25 and 13 key metabolites, of which the five most highly connected are glutamate, coenzyme A, α-ketoglutarate, pyruvate, and glutamine (Table 2; left column). This list overlaps with sets of key metabolic intermediates of *E.coli* used by other authors in metabolite balancing studies, where they represent the common biosynthetic source of all cell materials. For instance, Varma and Palsson [27] followed Ingraham *et al.* [28] in using a set of 12 biosynthetic precursors produced by the catabolism of all carbon sources: glucose 6-phosphate, fructose 6-phosphate, ribose 5-phosphate, erythrose 4-phosphate, triose phosphate, 3-phosphoglycerate, phosphoenolpyruvate, pyruvate, oxaloacetate, 2-oxoglutarate, acetyl CoA and succinyl CoA. Holmes [29] chose a smaller subset of 8 key precursors from which all cell biomass could be produced: glucose 6-phosphate, triose phosphate, 3-phosphoglycerate, phosphoenolpyruvate, pyruvate, oxaloacetate, 2-oxoglutarate, and acetyl CoA.

Substrate graphs have a power-law degree distribution. The high variance in connectivity warrants a closer look at the distribution of metabolite degrees ("connectivity"), which is shown in figures two and three for substrate graphs and reaction graphs, respectively. Each figure shows a histogram of degree vs. frequency, as well as a rank distribution of vertices (metabolites or reactions), where the vertex with the highest connectivity was assigned rank 1. Figure 2 reveals that the degree distribution of a substrate graph is consistent with a power-law, i.e., the probability *P(k)* of finding a vertex with degree *k*, $P(k) \propto k^{-\tau}$. While displaying frequency data as a log-log binned histogram is the most common way of visualizing a power law, much statistical information is lost by binning, resulting in little statistical confidence for a small graph such as this one (Fig. 2a). It is thus reassuring that the rank distribution which does not discard information and is essentially an estimate of the cumulative probability distribution of *k*, is also in good agreement with a power law (Fig. 2b) However, little confidence can be placed in the estimated value of the exponent τ (e.g., $\tau \approx 1.38$ from the rank distribution), because of the small network size. The large variance in degree discussed above is a consequence of the power law relation.

Power laws are "fat-tailed" probability distributions that have been detected in a variety of seemingly unrelated processes in nature and society, such as population size fluctuations in birds, price fluctuations in the stock market, the topography of the world wide web, or the magnitude of extinction events in the fossil [30-33]. Their fat tail reflects an overabundance of large events or objects, e.g., stock market crashes or highly connected metabolites. While it has been proposed that power laws reflect some deep commonalities among many processes in nature [34], alternative explanations resort to more mundane explanations. For instance, power laws may result from pooling log-normal distributions which are commonly found in nature (Li, pers. comm.).

The distribution of vertex degrees in the reaction graph does not follow a simple power law (Fig. 3). The rank vs.degree plot (Fig. 3b) shows that the it defies a straightforward classification, and appears to be governed by at least two qualitatively different regimes (Fig. 3b).

Metabolic graphs are small world graphs. What do the architecture of the *C. elegans* nervous system, the power grid of the western United States, the structure of some sociological network, and the world wide web have in common ? The surprising answer is that they are all small world graphs, a type of graph formally characterized by [7,19,30]. Small-world graphs are best illustrated with friendship networks in sociology, where small-worldness is known in folklore as "six degrees of separation". Friendship networks are sparse (each of $>2x10^8$ individuals in the United States is connected to at most 1000 "friends"), and highly clustered (one's friends tend to be friends of each other). This means that most of the few connections per individual are tied up in local interactions within "cliques" of individuals. Nevertheless, every individual in the U.S may be linked to every other individual by a short chain of acquaintances, as suggested by empirical work in sociology [35], a suggestion that has been confirmed for some completely mapped sociological networks [19]. A more formal definition of a small-world graph is that it is sparse and is much more highly clustered than an equally sparse random graph (C»C_{random}), but that its characteristic path length L is close to the theoretically possible minimum, that of a random graph ($L \approx L_{random}$). The reason why a graph can have small *L* despite being highly clustered is that few nodes connecting distant clusters may suffice to cause small *L* [7]. It follows that "small-worldness" is a global graph property that can not be found by studying local graph properties.

Figure 4a demonstrates that the *E.coli* metabolic network is much more highly clustered than random graphs. However, its characteristic path length is very small, and within one step of that of random graphs. Thus, it falls into the category of small world graphs, a feature that would not be obvious on the level of individual metabolic pathways. Substrate graphs illustrate this property particularly well. Their characteristic path length is within 5 percent (<0.1 steps) of that of an equally sparse random graph, although they are at least 17 times more clustered than random graphs. The high clustering coefficient is the result of local interations within metabolic pathways, the "cliques" in this network. To illustrate this, we analyzed separately the substrate graphs of ten of the longest individual pathways in our metabolic network, which were otherwise chosen arbitrarily. The analyzed pathways comprise 203 substrates and include glycolysis, the tricarboxylic acid cycle, biosyntheses of riboflavine, folate, histidine, branched chain amino acids, aromatic amino acids, threonine and lysine, arginin, putrescine and spemidine, porphyrene and heme, and coenzyme A. Their mean clustering coefficient, averaged over 10 pathways, calculates as $C=0.44$ ($\sigma=0.14,\ n=10$), not significantly different to that of $C=0.48$ measured for the whole network. We interpret this as an indication that the overall high clustering of the network is due to the individual metabolic pathways or modules. When considered as separate pathways, the coefficient of variation s in vertex degree (mean vertex degree averaged over ten pathways: k = 3.2) is found to be equal to $s=0.52$, which is much lower than that observed for the complete network ($s=1.01$; Table 1), and closer to that expected for a random graph with the same number of vertices (n=203) and average degree k $(s=0.39)$. This suggests that the highly connected metabolites linking the individual pathways into a connected network are responsible for the great variance in degree. Their high connectivity provides the "glue" of the network and is also responsible for the short pathlength. This is suggested by the mean characteristic path among each of the ten separate pathways, which is $L=3.08$ (σ=0.62), and thus not much smaller than the L=3.88 observed for the whole network.

Like most graph theoretical models, our model of metabolic networks omits most quantitative information, and is suited only to analyze network topography. However, it has two advantages to more fine-grained models. First, the required information (stoichiometric equations) is available. Second, by using a conceptual framework not restricted to chemistry, but important to many areas of science, it allows us to expose a deep structural similarity to seemingly unrelated networks. However, having identified a common design principle, one has to return to biology and think about its possible origins.

What might be the functional or phylogenetic significance of the observed patterns, a power law distribution of connectivity, and the small-world nature of the metabolic graph? It is of course possible that there is no such significance, because the laws of chemistry might constrain network structure so severely that only one design of a metabolic network can ensure that all basic cellular functions are fulfilled. In this case, the observed structure is determined by chemical constraints alone, and because the nature of these constraints is poorly understood, we could say little further. We can not strictly exclude this possibility, but some existing work speaks to the issue and suggests otherwise. First the biosyntheses of various compunds, such as lysine or isopentenyl diphosphate, occur by different routes in different organisms [36]. Recent analysis of the tricarboxylic acid (TCA) cycle from the viewpoint of chemical design showed that there are several chemically possible solutions to the tasks it performs, of which the solution realized in cells is the one that involves the fewest chemical transformations [37]. Moreover, considerable variation exists in the presence or absence of particular reactions in the TCA cycle in 19 prokaryotes with completely sequenced genomes [38]. Strikingly, in a majority of these species, the TCA cycle appears incomplete or absent [38]. If even key components of metabolism can show such variation, how much more variation must there be in more peripheral parts of a metabolic network? At the very least, these studies suggest that chemistry does allow flexibility in the design of a metabolic net. If this is the case, then the observed architecture may be a relic of evolutionary history, a product of evolutionary optimization, or a mixture of both. Which of these may be more important we may never know, but we will offer two speculations.

First, could the observed network structure be an indicator of the evolutionary history of metabolism? Barabási and collaborators [39] have recently proposed a mathematical model that generates large graphs from small graphs by adding nodes and edges. If new links between nodes are made preferentially between nodes that already have many links, then the resulting graphs are small-world graphs with power-law degree distributions. A key prediction is that vertices with many connections are vertices that have been added early in the history of the graph. Cast in terms of metabolism, if early in the evolution of life metabolic networks have increased in size by adding new metabolites, then the most highly connected metabolites should also be the phylogenetically oldest. Indeed, many of the most highly connected metabolites in Table 2 have a proposed early evolutionary origin. RNA cofactors such as coenzyme A, NAD, or GTP are among the most highly connected metabolites, and are thought to be among the remnants of an RNA world. 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 Table 2. Early proteins are thought to have used many fewer amino acids than extant proteins, and the highly connected amino acids glutamine, glutamate, aspartate, and serine are thought to be among those used earliest. [40-45]. The potential relation between evolutionary history and connectivity of metabolites corroborates a postulate put forth and defended forcefully by Morowitz [, 1992; 493], namely that intermediary metabolism recapitulates the evolution of biochemistry. Our highly connected metabolites pyruvate, α-ketoglutarate, acetyl CoA, oxaloacetate are identified by Morowitz [46] as belonging to the original core metabolism, and glutamate, glutamine and aspartate are the links from this core into the next earliest subset of compounds, the first amino acids.

Second, which aspect of metabolic function might a small-world network optimize? Metabolic networks need to react to perturbations, either perturbations in enzyme concentrations, or changes in metabolite concentrations. Because metabolic networks are connected, each component in the network may be affected by such perturbations, and thus the network as a whole must adapt to the changed conditions by assuming a different metabolic state. The importance of minimizing the transition time betwen metabolic states has been recognized and discussed by other authors [47,48]. Any response to a perturbation and transition to a new metabolic state requires that information about the perturbation has spread within the network. Watts and Strogatz [7] studied how fast perturbations spread through small-worldnetworks. Significantly, they found that the time required for spreading of a perturbation in a small-world network is close to the theoretically possible minimum for any graph with the same number of nodes and vertices. In other words, perturbations spread extremely fast, and small-worldness may thus allow a metabolism to react rapidly to them.

These hypotheses might not be tested easily. However, they serve to illustrate that a suitable mathematical framework can allow us to perceive global patterns of biological organization, patterns that are not visible on a local level, patterns that allow us to build qualitatively new kinds of hypotheses. Detecting order in the torrent of genomic data descending upon the life science community will certainly require such hypotheses.

Acknowledgments

Financial support by the Santa Fe Institute for a two-week visit of D.F. which led to completion of this manuscript is gratefully acknowledged.

References

1. Stryer, L. *Biochemistry* (Freeman, New York, 1995).
2. Rhee, S. G., Chock, P. B. & Stadtman, E. R. Regulation of Escherichia coli glutamine synthetase. *Advances in Enzymology* **62**, 37-92 (1989).
3. Heinrich, R. & Schuster, S. *The regulation of cellular systems* (Chapman and Hall, New York, 1996).
4. Schilling, C. H., Schuster, S., Palsson, B. O. & Heinrich, R. Metabolic pathway analysis: Basic concepts and scientific applications in the post-genomic era. *Biotechnology Progress* **15**, 296-303 (1999).
5. Schuster, S., Dandekar, T. & Fell, D. A. Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends in Biotechnology* **17**, 53-60 (1999).
6. Fell, D. A. & Small, J. R. Fat synthesis in adipose tissue: an examination of stoichiometric constraints. *Biochemical Journal* **238**, 781-786 (1986).
7. Watts, D. J. The structure and dynamics of small world networks. Ph.D. Thesis ;Cornell University. (1997).
8. Neidhardt, F. C. (ed.) *Escherichia coli and Salmonella* (ASM Press, Washington, D.C., 1996).
9. Karp, P., Riley, M., Paley, S., Pellegrini-Toole, A. & Krummenacker, M. EcoCyc: electronic encyclopedia of E.coli genes and metabolism. *Nucleic acids research* **27**, 55- (1999).
10. Pramanik, J. & Keasling, J. D. Stoichiometric model of Escherischia coli metabolism: incorporation of growth rate dependent biomass composition and mechanistic energy requirements. *Biotechnology and Bioengineering* **56**, 398-421 (1997).
11. Selkov, E. *et al.* The metabolic pathway collection from EMP: the enzymes and metabolic pathways database. *Nucleic Acids Research* **24**, 26-29 (1996).
12. Bairoch, A. The enzyme data bank in 1999. *Nucleic Acids Research* **27**, 310-311 (1999).
13. Fell, D. A. & Sauro, H. M. Metabolic control analysis by computer: progress and prospects. *Biomed. Biochim. Acta* **49**, 811-816 (1990).
14. Mehlhorn, K. & Naeher, S. *The LEDA platform of combinatorial computing* (Cambridge University Press, Cambridge, 1999).

15. Graham, R. L., Groetschel, M. & Lovasz, L. (eds.) *Handbook of combinatorics* (MIT press, Cambridge, 1995).

16. Fell, D. *Understanding the control of metabolism* (Portland Press, London, 1997).

17. Hofmeyr, J.-H. S. Control pattern analysis of metabolic pathways: flux and concentration control in linear pathways. *European Journal of Biochemistry* **275**, 253-258 (1991).

18. Sen, A. K. Quantitative analysis of metabolic regulation: a graph--theoretic approach using spanning trees. *Biochemical Journal* **275**, 253-258 (1991).

19. Watts, D. J. & Strogatz, S. H. Collective Dynamics of Small-World Networks. *Nature* **393**, 440-442 (1998).

20. Bollobás, B. *Random graphs* (Academic Press, London, 1985).

21. Kauffman, S. A. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of theoretical biology* **22**, 437-467 (1967).

22. Glass, L. & Hill, C. Ordered and Disordered Dynamics in Random Networks. *Europhysics Letters* **41**, 599-604 (1998).

23. Chiva, E. & Tarroux, P. Evolution of Biological Regulation Networks Under Complex Environmental Constraints. *Biological Cybernetics* **73**, 323-333 (1995).

24. Wagner, A. Does Evolutionary Plasticity Evolve? *Evolution* **50**, 1008-1023 (1996).

25. Murre, J. M. J. & Sturdy, D. P. F. The Connectivity of the Brain : Multilevel Quantitative-Analysis. *Biological Cybernetics* **73**, 529-545 (1995).

26. Cohen, J. E. & Briand, F. Trophic Links of Community Food Webs. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences* **81**, 4105-4109 (1984).

27. Varma, A. & Palsson, B. O. Metabolic capabilities of Escherichia coli. synthesis of biosynthetic precursors and cofactors. *Journal of theoretical biology.* **165**, 477-502 (1993).

28. Ingraham, J. L., Maaloe, O. & Niedhardt, F. C. *Growth of the bacterial cell.* (Sinauer, Sunderland, MA, 1983).

29. Holms, W. H. The central metabolic pathways of Escherischia coli: relationship between flux and control at a branch point , efficiency of conversion to biomass and excretion of acetate. *Current Topics Cell. Regul.* **28**, 69-105 (1986).

30. Albert, R., Jeong H., Barabasi, A.L. Internet: Diameter of the world-wide web. *Nature* **401**, 130-131 (1999).

31. Newman, M. E. J. & Eble, G. J. Power spectra of extinction in the fossil record. *Proceedings of the Royal Society of London Series B-Biological Sciences* **266**, 1267-1270 (1999).

32. Gopikrishnan, P., Meyer, M., Amaral, L. A. N. & Stanley, H. E. Inverse Cubic Law For the Distribution of Stock-Price Variations. *European Physical Journal B* **3**, 139-140 (1998).

33. Keitt, T. H. & Stanley, H. E. Dynamics of North-American Breeding Bird Populations. *Nature* **393**, 257-260 (1998).

34. Bak, P. Self-Organized Criticality. *Physica a* **163**, 403-409 (1990).

35. Milgram, S. The small world problem. *Psychology Today* **2**, 60-67 (1967).

36. Rohmer, M., M., K., Simonin, P., Sutter, B. & Sahm, H. Isopentenyl diphosphate synthesis in bacteria does not proceed via the
acetate/mevalonate pathway used by mammals:. *Biochemical Journal* **295**, 517-524 (1993).

37. Melendez-Hevia, E., Waddell, T. G. & Cascante, M. The Puzzle of the Krebs Citric-Acid Cycle : Assembling the Pieces of Chemically Feasible Reactions ; and Opportunism in the Design of Metabolic Pathways During Evolution. *Journal of Molecular Evolution* **43**, 293-303 (1996).

38. Huynen, M. A., Dandekar, T. & Bork, P. Variation and evolution of the citric acid cycle: a genomic perspective. *Trends in Microbiology* **7**, 281-291 (1999).

39. Barabasi, A.-L., Albert, R., Jeong, H. Mean-field theory for scale-free random networks. Los Alamos National Laboratory Preprint cond-mat/9907068 (1999).

40. Morowitz, H. J. *Beginnings of cellular life* (Yale University Press, New Haven, 1992).

41. Kuhn, H. & Waser, J. On the origin of the genetic code. *FEBS letters* **352**, 259-264 (1994).

42. Lahav, N. *Biogenesis* (Oxford University Press, New York, 1999).

43. Benner, S. A., Ellington, A. D. & Tauer, A. Modern metabolism as a palimpsest of the RNA world. *Proceedings of the National Academy of Sciences of the U.S.A.* **86**, 7054-7058 (1989).

44. Waddell, T. G. & Bruce, G. K. A new theory on the origin and evolution of the citric acid cycle. *Microbiologia sem* **11**, 243-250 (1995).

45. Taylor, B. L. & Coates, D. The code within the codons. *Biosystems* **22**, 177-187 (1989).

46. Morowitz, H. J. A theory of biochemical organization, metabolic pathways, and evolution. *Complexity* **4**, 39-53 (1999).

47. Easterby, J. S. The effect of feedback on pathway transient response. *Biochemical Journal* **233**, 871-875 (1986).

48. Cascante, M., Melendez--Hevia, E., Kholodenko, B. N., Sicilia, J. & Kacser, H. Control analysis of transit--time for free and enzyme--bound metabolites - physiological and evolutionary significance of metabolic response--times. *Biochemical Journal* **308**, 895-899 (1995).

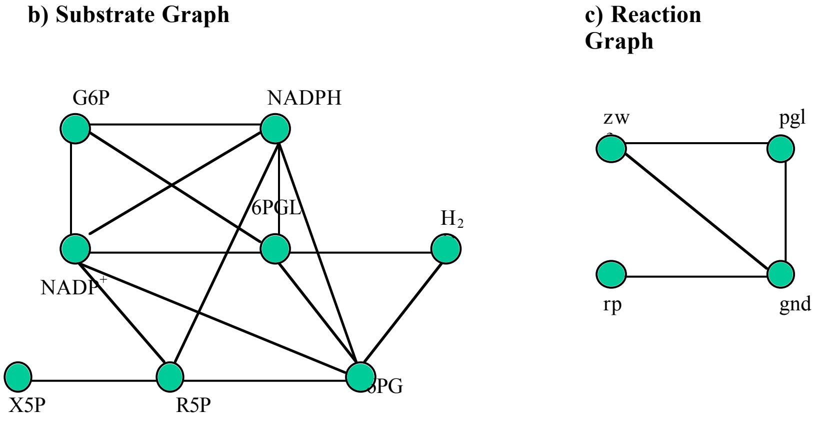
a) Stoichiometric Equations

1 Glucose 6-phosphate (G6P) + 1 NADP$^+$ $\overset{zwf}{\Rightarrow}$ 1 6-Phosphoglucono δ-lactone (6PGL) + 1 NADPH

1 6-Phosphoglucono δ-lactone + 1 H_2O $\overset{pgl}{\Rightarrow}$ 1 6-Phosphogluconate (6PG)

1 6-Phosphogluconate + 1 NADP$^+$ $\overset{gnd}{\Rightarrow}$ 1 Ribulose 5-phosphate (R5P) + 1 NADPH

1 Ribulose 5-phosphate $\overset{rpe}{\Leftrightarrow}$ 1 Xylulose 5-phosphate (X5P)



b) Substrate Graph

c) Reaction Graph

Fig. 1: Graph representation of metabolic networks. a) Four stoichiometric equations taken from the pentose-phosphate pathway of *E.coli* after [1,8]. Names in parentheses are acronyms for compounds used in b). Acronyms above arrows indicate genes encoding the respective reactions (enzymes) in *E. coli* (*zwf*: Glucose-6-phosphate dehydrogenase [EC 1.1.1.49]; *pgl*: 6-Phosphogluconolactonase [EC 3.1.1.31]; *gnd*: 6-Phosphogluconate dehydrogenase [EC 1.1.1.4]; *rpe*: Ribose-5-phosphate isomerase [EC 5.3.1.6]). **b)** Substrate graph derived from stoichiometric equations as describd in the text. **c)** Reaction graph derived from stoichiometric equations as described in the text.

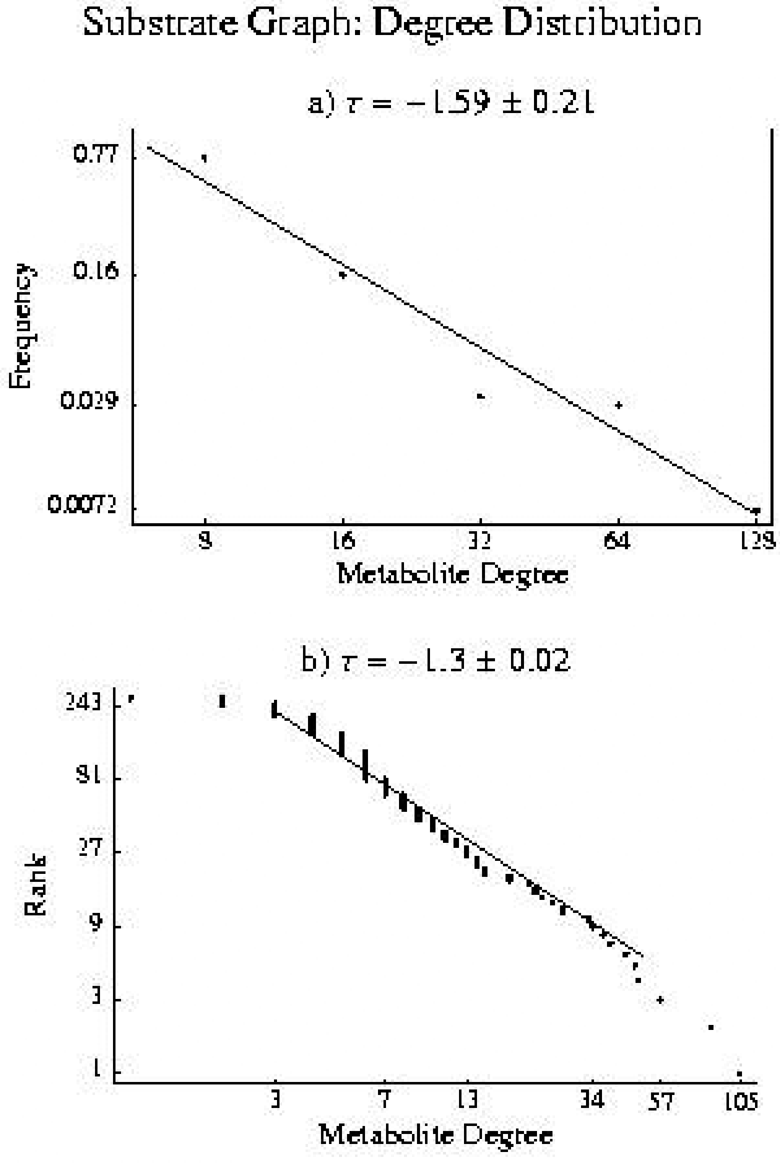


Fig. 2. The substrate graph shows a power law distribution of metabolite connectivity.

a) Log-log histogram of the relative frequency of metabolites with a given degree (number of connections) k. Vertices were binned into five intervals according to degree: $1 \leq k < 8$, $8 \leq k < 16$, $16 \leq k < 32$; 32; $32 \leq k < 64$; $64 \leq k < 128$). Values on the abscissa indicate the upper boundary of each interval. Coefficient of determination $r^2 \approx 0.93$. **b)** Metabolites were ranked according to the number of connections (degree) they have in the substrate graph. Shown is metabolite rank vs. degree on a log-log scale. Assuming that the degree of a metabolite can be described by a random variable D, plotting data as in a) can be used to estimate the probability function $P(\log D=k)$, whereas b) estimates the counter-cumulative probability function $P(\log D>k)$. Both a) and b) 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 τ because of the small network size.

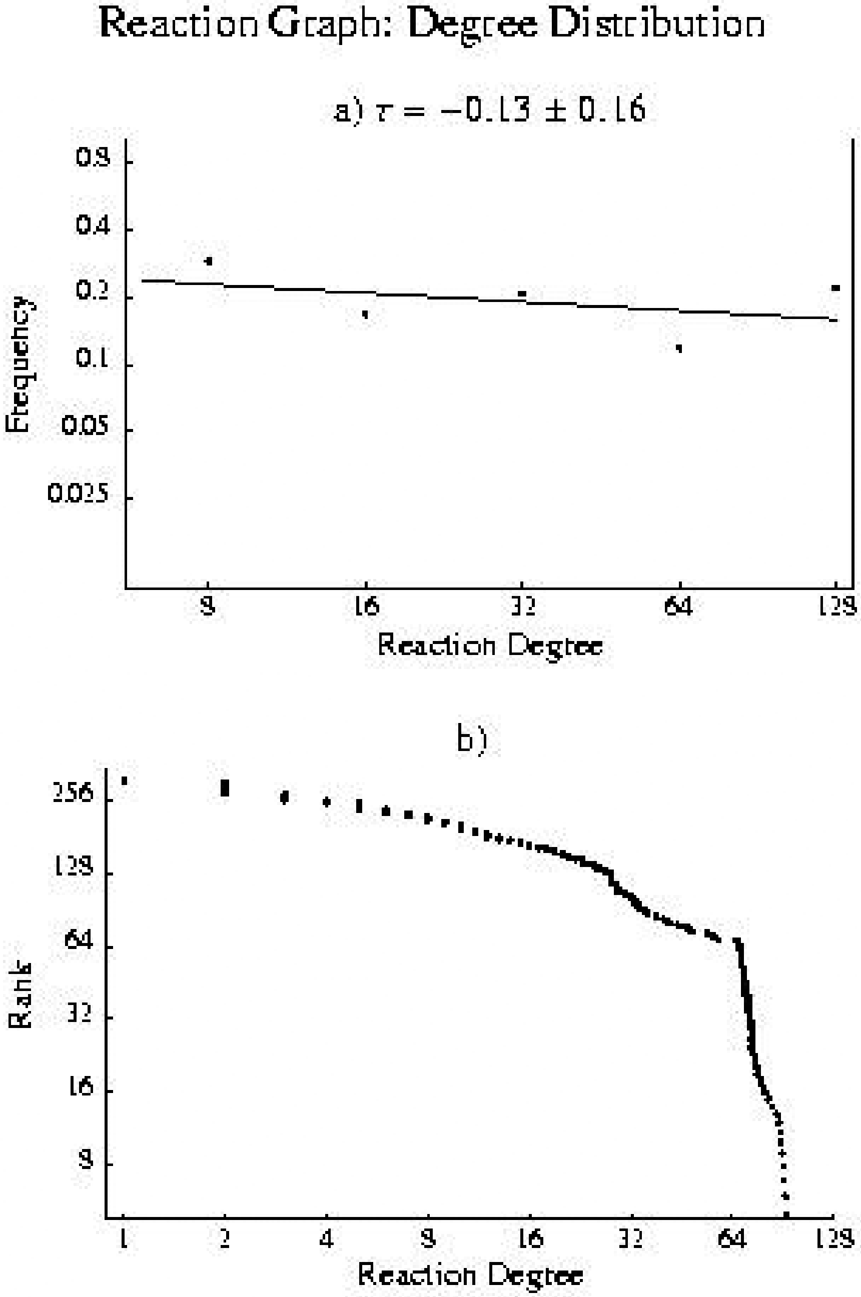
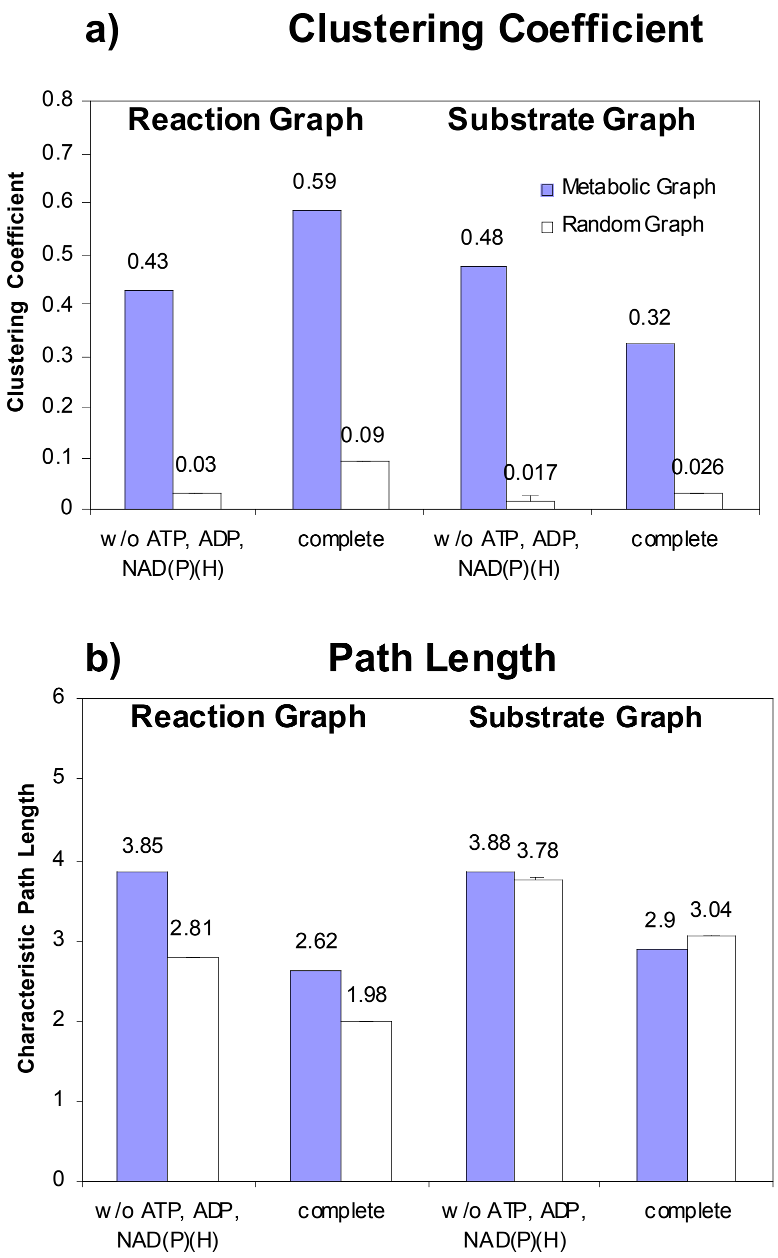


Fig. 3. Degree distribution in the reaction graph. Plotted are connectivity (degree) of nodes in the reaction graph vs. binned frequency in **a)** and rank in **b)**, completely analogous to Fig. 3 (see legend). a) already indicates that the degree distribution does not follow a power law, and b) shows further that no simple cumulative probability function would appropriately approximate the rank distribution shown. This is because there are at least two distinct regimes in the degree distribution, where the degree changes much slower among the most highly connected reactions among less highly connected reactions.


a) Clustering Coefficient

b) Path Length

Fig. 4. Metabolic network graphs are small-world graphs. Shown are characteristic path length L in **a)** and clustering coefficients C in **b)** for reaction graphs, substrate graphs, and random graphs. Notice the great difference between metabolic graphs and random graphs in clustering coefficient, and the similarity between metabolic graphs and random graphs in characteristic path length. This similarity becomes more striking if one considers how large L can become for graphs with the same connectivity (k) [7] shows that the maximally possible L is close to $n/2(k+1)$)]. Using Table 1, these values calculate as 15.14, 5.38, 24.2, and 16.9 from left to right for the metabolic graphs shown in the figure. The substrate graph conforms better to the small world model than the reaction graph. Values shown for random graphs are mean and standard deviations (error bars) over 100 numerically generated connected random graphs where the probability for two vertices to be connected was chosen as k/n (Table 1).

Table 1: Elementary Statistics of Substrate and Reaction Graphs. Shown are the number of nodes (*n*), the mean degree (*k*), and standard deviation in degree (σ_k) for reaction graph and substrate graph. For reference, standard deviation in degree is also shown for 100 numerically generated random graphs with the same *n* and *k* as those of the metabolic graphs. Two versions of each metabolic graph were analyzed, one in which the metabolites ATP, ADP, NAD, NADP, NADH, NADPH, CO_2, NH_3, SO_4, thioredoxin, phosphate and pyrophosphate were eliminated, and another one in which ATP, ADP, NAD, NADP, NADH, and NADPH were included. Upon removal of one or more metabolites, other vertices in the graph may become isolated. Such vertices were removed before analysis.

	n	k	σ_k	σ_k random graph	
Substrate Graph *w/o ATP, ADP, NAD(P)(H)*	275	4.76	4.79	2.12±0.08	
Substrate Graph	282	7.35	10.5	2.67±0.11	
Reaction Graph *w/o ATP, ADP, NAD(P)(H)*		311	9.27	9.59	3.01± 0.12
Reaction Graph	315	28.3	29.1	5.04 ± 0.21	

Table 2: Thirteen "key metabolites" defined as metabolites whose degree in the substrate graph lies at least three standard deviations beyond the mean metabolite degree. Also shown for comparison are the 13 metabolites with the shortest mean path length (also known as the "importance number"). These two indicators of a metabolite's centrality are correlated but not identical. Values in parentheses are metabolite degree (left column) and mean pathlength (right column). NAD, ATP and their derivatives would be the most highly connected metabolites, but are not shown in the table.

ranked by degree ("connectivity"}	ranked by mean path length ("importance number")
glutamate (51)	glutamate (2.46)
pyruvate (29)	pyruvate (2.59)
coenzyme A (29)	coenzyme A (2.69)
α-ketoglutarate (27)	glutamine (2.77)
glutamine (22)	acetyl CoA (2.86)
aspartate (20)	oxo-isovalerate (2.88)
acetyl-CoA (17)	aspartate (2.91)
phosphoribosyl pyrophosphate (16)	α-ketoglutarate (2.99)
tetrahydrofolate (15)	phosphoribosyl pyrophosphate (3.1)
succinate (14)	anthranilate (3.1)
3-phosphoglycerate (13)	chorismate (3.13)
serine (13)	valine (3.14)
oxo-isovalerate (12)	3-phosphoglycerate (3.15)