THE ROYAL SOCIETY

doi 10.1098/rspb.2001.1711

The small world inside large metabolic networks

Andreas Wagner[1] **and David A. Fell**[2*]

[1]*Department of Biology, University of New Mexico, 167A Castetter Hall, Albuquerque, NM 87131-1091, USA*
[2]*School of Biological and Molecular Sciences, Oxford Brookes University, Headington, Oxford OX3 0BP, UK*

The metabolic network of the catabolic, energy and biosynthetic metabolism of *Escherichia coli* is a paradigmatic case for the large genetic and metabolic networks that functional genomics efforts are beginning to elucidate. To analyse the structure of previously unknown networks involving hundreds or thousands of components by simple visual inspection is impossible, and quantitative approaches are needed to analyse them. We have undertaken a graph theoretical analysis of the *E. coli* metabolic network and find that this network is a small-world graph, a type of graph distinct from both regular and random networks and 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 *Caenorhabditis elegans*. Moreover, the connectivity of the metabolites follows a power law, another unusual but by no means rare statistical distribution. This provides an objective criterion for the centrality of the tricarboxylic acid cycle to metabolism. The small-world architecture may serve to minimize transition times between metabolic states, and contains evidence about the evolutionary history of metabolism.

Keywords: metabolism; graph theory; small world; evolution; origin of life

1. INTRODUCTION

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 will arise which has, as yet, been little explored by molecular biologists: how to adequately represent and analyse the structure of such large networks. While it is trivial to understand the structure of isolated metabolic pathways, transcriptional cascades, or signalling pathways constructed from a small number of gene products, networks consisting of anywhere from hundreds to tens of thousands of components are less easily described. Theory suitable for analysing large networks exists only for perfectly ordered or certain completely random networks. Similarly, theoretical exploration of network properties (Kauffman 1967; Glass & Hill 1988; Chiva & Tarroux 1995; Wagner 1996) also depends on the creation of model networks that are truly qualitatively and quantitatively representative of the biological ones.

Here, we analyse the structure of a large metabolic network, that of *Escherichia coli* intermediary metabolism for energy generation and small building block synthesis. One type of question to ask for a given metabolic network concerns the availability and yield of transformation routes from nutrients to end-products. This is traditionally answered by consideration of the presence or absence of all the necessary steps of the classical biochemical pathways, but it can be argued that interpretation of metabolic networks in terms of these classical pathways fails to reveal the full potential of the network (Fell & Small 1986; Schuster *et al*. 1999, 2000; Edwards & Palsson 2000). Hence methods have been developed to enumerate the full repertoire of potential pathways, such as elementary modes analysis (Schuster *et al*. 1999, 2000). Another type of question that can be asked concerns the identity of the key intermediary metabolites that must be generated by catabolism for use in anabolism, and that therefore define the centre of metabolism dividing catabolism from anabolism. There is only partial agreement over the identity of these metabolites (e.g. Ingraham *et al*. 1983; Holmes 1986) and no objective criteria for their choice.

Thus the aim of this study was to characterize the structure of this particular metabolic network, to determine whether it can be objectively said to have a centre, and if so, to determine the identity of the central metabolites.

2. THEORY AND METHODS

Mathematically, the behaviour of a metabolic network can be captured as a system of ordinary differential equations in the metabolite concentrations. A compact expression of this equation system is obtained by use of the stoichiometry matrix, $\mathbf{N}$, whose elements n_{ij} represent the number of molecules of metabolite i formed (or, if negative, consumed) in a reaction step j. Given a vector of metabolite concentrations $\mathbf{S}$ and a vector of reaction rates $\mathbf{v}$, the equation is (e.g. Heinrich & Schuster 1996)

$$\frac{\mathrm{d}\mathbf{S}}{\mathrm{d}t} = \mathbf{N} \cdot \mathbf{v}. \tag{2.1}$$

Even numerical solutions of this equation are impractical for whole metabolic networks because the reaction rates are complicated (and often unknown) functions of the metabolite concentrations. For metabolic steady states, however (where $\mathrm{d}\mathbf{S}/\mathrm{d}t = \mathbf{0}$), the stoichiometric matrix $\mathbf{N}$ imposes a set of linear constraints on feasible solutions. Because the stoichiometric matrix contains the full information about the structure of the network, these have been termed structural constraints (Reder 1988) in the field of metabolic control analysis. Linear programming, null space analysis, convex analysis and elementary modes analysis (Heinrich & Schuster 1996; Schilling *et al*.

*Author for correspondence (daf@brookes.ac.uk).

Received 24 January 2001 *Accepted* 12 April 2001

1999) have all been applied to metabolism and are essentially different explorations of the structural constraints.

Here, we choose a different graph theoretic representation derived from the stoichiometric equations. Because most of the reactions of metabolism are multi-molecular, some form of hypergraph (Graham *et al*. 1995) would be needed to retain the full information content of the stoichiometry matrix. We instead decided to use a simple graph representation of metabolism, because hypergraphs are much less intuitive constructs than graphs and the tools our analysis needs have not yet been developed for them. One might argue that a directed graph (Graham *et al*. 1995) is a better choice, i.e. a graph where each edge has a direction, because of the existence of irreversible reactions. Again, we deliberately avoided directed graphs because perturbations can travel backwards through an enzyme-catalysed irreversible step, even in the absence of reverse net flow of matter. For example, in control analysis, it is the flux and concentration control coefficients of an enzyme (Fell 1997) and not the reversibility of the reaction that show whether it is possible for a change in activity of an enzyme to propagate effects into the part of the network 'upstream' of the reaction (Hofmeyr 1989; Sen 1991). A directed substrate graph would not capture this behaviour.

Based on publicly available information (Neidhardt 1996; Selkov *et al*. 1996; Pramanik & Keasling 1997; Bairoch 1999; Karp *et al*. 1999) 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 (Neidhardt 1996), (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), oxidative phosphorylation (6), amino acid and polyamine biosynthesis (95), nucleotide and nucleoside biosynthesis (72), folate synthesis and 1-carbon metabolism (16), glycerol 3-phosphate and membrane lipids (17), riboflavin (9), coenzyme A (11), NAD(P) (7), porphyrins, haem 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 the author for correspondence upon request. From these reaction equations, the stoichiometry matrix was automatically generated from the reaction list using the software package INDIGO (Fell & Sauro 1990). From this matrix, the substrate and reaction graph were derived omitting the metabolites CO_2, NH_3, SO_4, thioredoxin, organic phosphate and pyrophosphate. Upon removal of one or more metabolites, other vertices in the graph may become isolated; any such vertices were removed before analysis.

We considered 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 joined by an edge $e = (S_1, S_2) \in E_S$, the edge set of this graph, when they occur (either as substrates or products) in the same chemical reaction (figure 1*b*). The reaction graph, $G_R = (V_R, E_R)$, has a vertex set V_R consisting of all chemical reactions in the network. Two reactions are joined by 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 figure 1*c*).

For both graphs (Graham *et al*. 1995), the degree k of a vertex is the number of other vertices to which it is adjacent. Two vertices v_0, v_i are connected if there exists a path, i.e. a sequence of adjacent vertices $v_0, v_1, \ldots v_{i-1}, v_i$ from v_0 to v_i. We will be concerned only with connected graphs, i.e. graphs where all vertex pairs are connected, since the law of mass conservation and the fact that the carbon of all biomass is ultimately derived from CO_2 imply that metabolic networks must be 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 path length between two vertices, averaged over all pairs of vertices. Another important quantity (Watts & Strogatz 1998) 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 maximum possible m is $[k_v(k_v - 1)]/2$, in which case all the vertices are connected to each other, and we define $C(v) := 2m/[k_v(k_v - 1)]$. $C(v)$ measures the 'cliquishness' of the neighborhood of v, i.e. what fraction of the vertices adjacent to v are also adjacent to each other. By extension, the clustering coefficient C of the graph is defined as the average of $C(v)$ over all v.

The properties of the metabolic graph can be compared to the benchmark case of a random graph with the same number of vertices n and mean degree $\bar{k}$ by exploiting the available statistical theory of random graphs (Bollobás 1985). Importantly, random connectivity and a close variant k-regular random connectivity (Graham *et al*. 1995), have frequently been the assumptions of choice during more than three decades of modelling genetic networks (Kauffman 1967; Glass & Hill 1988; Chiva & Tarroux 1995; Wagner 1996). It is thus useful to see how the actual structure of a cell 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 $\bar{k}$, ($\bar{k} \ll n$), the probability p of two vertices being connected is given by $p = \bar{k}/(n - 1)$. Such graphs show (i) a binomial distribution of vertex degree k, (ii) a very small clustering coefficient $C = (\bar{k} - 1)/n$, 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 (Watts 1997). Thus, among all connected

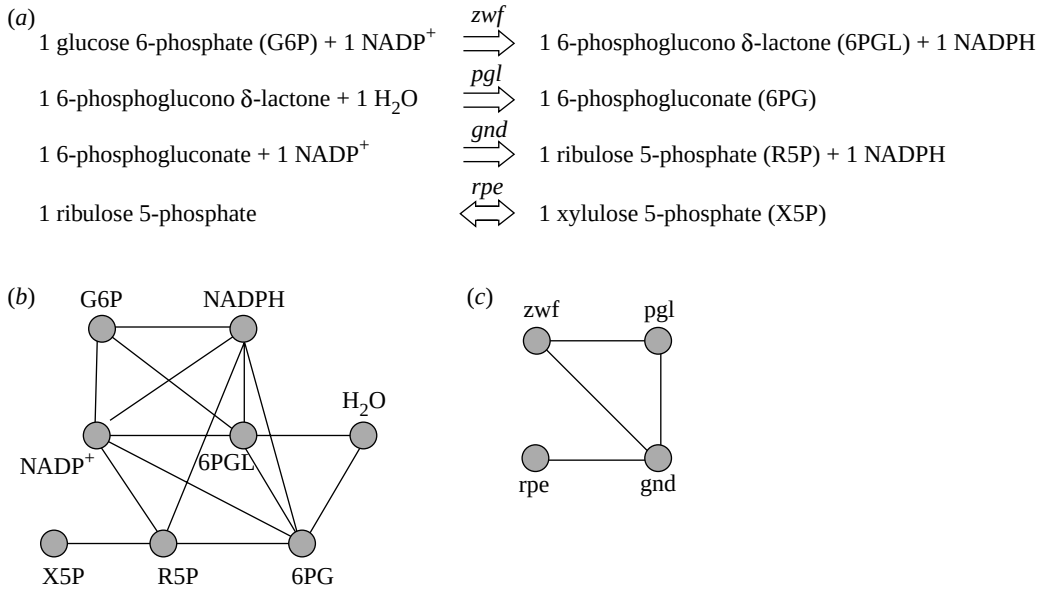

Figure 1. Graphical representation of metabolic networks. (*a*) Four stoichiometric equations taken from the pentose-phosphate pathway of *E. coli*. Names in parentheses are acronyms for compounds used in (*b*). Acronyms above arrows are the identifiers for the genes encoding the respective reactions (enzymes) (*zwf*, glucose-6-phosphate dehydrogenase [EC 1.1.1.49]; *pgl*, 6-phosphogluconolactonase [EC 3.1.1.31]; *gnd*, 6-phosphogluconate dehydrogenase [EC 1.1.1.43]; *rpe*, ribulose-phosphate 3-epimerase [EC 5.1.3.1]). (*b*) Substrate graph derived from stoichiometric equations. (*c*) Reaction graph derived from stoichiometric equations.

graphs with the same number of vertices and edges, random graphs are among the most rapidly traversed.

Graph analysis software was written in C++ using the LEDA library of data types (Mehlhorn & Naeher 1999).

3. RESULTS

Because of the ubiquity of the metabolites adenosine triphosphate (ATP), adenosine diphosphate (ADP) and nicotinamide adenine dinucleotide (NAD), as well as its phosphorylated and reduced forms (Stryer 1995), 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 *E. coli*. The variation in connectivity of both types of graph greatly exceeds that of random graphs. Like networks found in neurobiology and ecology (Cohen & Briand 1984; Murre & Sturdy 1995), metabolic graphs are sparse, i.e. the average degree ($\bar{k}$) of each vertex (metabolite or reaction) is small, 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 the degree of the metabolic graphs, however, is up to 20-fold greater than that of the corresponding random graph with $p = \bar{k}/(n-1)$, implying that some vertices in metabolic graphs have many more, and others many fewer, neighbours 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 allows a statistical definition of 'key metabolites' or 'key reactions', particularly highly connected vertices in metabolite graphs. For example,

Table 1. Elementary statistics of the substrate and reaction graphs.

(Shown are: the number of nodes, n; the mean degree, $\bar{k}$, and the standard deviation in degree, σ_k. For reference, standard deviation in degree is also shown for 100 numerically generated random graphs with the same n and $\bar{k}$ as those of the metabolic graphs. Two versions of each metabolic graph were analysed, one in which the metabolites ATP, ADP, NAD, NADP, NADH and NADPH were eliminated, and another one in which ATP etc. were included.)

graph	n	$\bar{k}$	σ_k	σ (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

for the substrate graph, one might define a key metabolite as one whose vertex degree k_m exceeds the average $\bar{k}$ by three standard deviations:

$$k_m > \bar{k} + 3\sigma = \bar{k} + 3\sqrt{\frac{k(n-1-k)}{n-1}}. \qquad (3.1)$$

Applying this to the substrate graph with $\bar{k} = 4.76$ (table 1) identifies 13 key metabolites with $k_m > 11.25$, of which the five most highly connected are glutamate, coenzyme A, 2-oxoglutarate, pyruvate and glutamine (table 2; left

Table 2. Thirteen key metabolites of *E. coli* metabolism.

(These are 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 path length (right column). NAD, ATP and their derivatives would be the most highly connected metabolites, but are not shown in the table.)

rank by degree	connectivity	rank by mean path length	importance number
glutamate	51	glutamate	2.46
pyruvate	29	pyruvate	2.59
CoA	29	CoA	2.69
2-oxoglutarate	27	glutamine	2.77
glutamine	22	acetyl CoA	2.86
aspartate	20	oxoisovalerate	2.88
acetyl CoA	17	aspartate	2.91
phosphoribosyl*PP*	16	2-oxoglutarate	2.99
tetrahydrofolate	15	phosphoribosyl*PP*	3.10
succinate	14	anthranilate	3.10
3-phosphoglycerate	13	chorismate	3.13
serine	13	valine	3.14
oxoisovalerate	12	3-phosphoglycerate	3.15

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 & Palsson (1993) followed Ingraham *et al*. (1983) 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, a triose phosphate, 3-phosphoglycerate, phosphoenolpyruvate, pyruvate, oxaloacetate, 2-oxoglutarate, acetyl CoA and succinyl CoA. Holmes (1986) chose a smaller subset of eight key precursors from which all cell biomass could be produced: glucose 6-phosphate, a triose phosphate, 3-phosphoglycerate, phosphoenolpyruvate, pyruvate, oxaloacetate, 2-oxoglutarate, and acetyl CoA. The most highly represented pathway in table 2 is the tricarboxylic acid cycle, especially if the amino acids derived directly from it by transamination are counted.

The high variance in connectivity warrants a closer look at the distribution of metabolite degrees for substrate graphs (figure 2) and reaction graphs (figure 3). Each figure shows a histogram of degree versus frequency, together with a rank distribution of vertices (metabolites or reactions), where the vertex with the highest degree was assigned rank unity. 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 is proportional to $k^{-\tau}$. Although displaying frequency data as a log–log binned histogram (figure 2*a*) is the most common way of visualizing a power law, much statistical information is lost by binning. However, 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 (figure 2*b*), although less confidence can be placed in its estimated value of the exponent τ because of the small network size. 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 record (Gopikrishnan *et al*. 1998; Keitt & Stanley 1998; Albert *et al*. 1999; Newman & Eble 1999). Their broad tail reflects a relative overabundance of the rare large events, objects, or highly connected metabolites. Although it is held by some that power laws reflect deep commonalities among many processes in nature (Bak 1990), power laws might result from pooling log–normal distributions, which are commonly found in nature (B.-L. Li, personal communication). The distribution of vertex degrees in the reaction graph does not follow a simple power law (figure 3). The rank versus degree plot (figure 3*b*) shows that it defies a straightforward classification, and appears to be governed by at least two qualitatively different regimes.

Metabolic graphs are, in fact, small-world graphs, like the architecture of the *Caenorhabditis elegans* nervous system, the power grid of the western United States, the structure of some sociological networks (Watts & Strogatz 1998), and the World Wide Web (Albert *et al*. 1999). The small-world graph was formally characterized by Watts (Watts 1997; Watts & Strogatz 1998) and is best illustrated by friendship networks in sociology, where small-worldness is known as 'six degrees of separation'. This followed original empirical work in sociology (Milgram 1967) that has since been confirmed for some completely mapped sociological networks (Watts & Strogatz 1998). The formal definition of a small-world graph is that it is sparse but much more highly clustered than an equally sparse random graph ($C \gg C_{\mathrm{random}}$), and its characteristic path length L is close to the theoretical minimum shown by a random graph ($L \approx L_{\mathrm{random}}$). The reason a graph can have small L despite being highly clustered is that a few nodes connecting distant clusters are sufficient to lower L (Watts & Strogatz 1998). It follows that 'small-worldness' is a global graph property that cannot be found by studying local graph properties.

Figure 4*a* demonstrates that the *E. coli* metabolic network is indeed more highly clustered (17 times more)

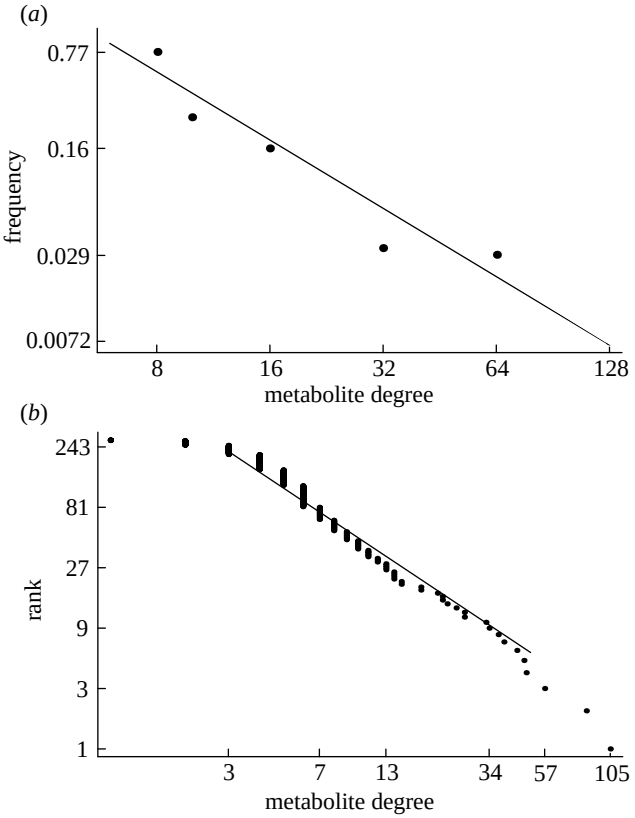

Figure 2. The power law distribution of metabolite connectivity in the substrate graph. (*a*) Log–log histogram of the relative frequency of metabolites with a given degree k. Vertices were binned into five intervals according to degree ($1 \leqslant k < 8, \ldots 64 \leqslant k < 128$) where values on the abscissa indicate the upper boundaries. Coefficient of determination $r^2 \approx 0.93$, $\tau = -1.59 \pm 0.21$. (*b*) 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, $\tau = -1.3 \pm 0.02$. Assuming that the degree of a metabolite can be described by a random variable D, plotting data as in (*a*) estimates 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 $P(D > k) \propto k^{-\tau}$.

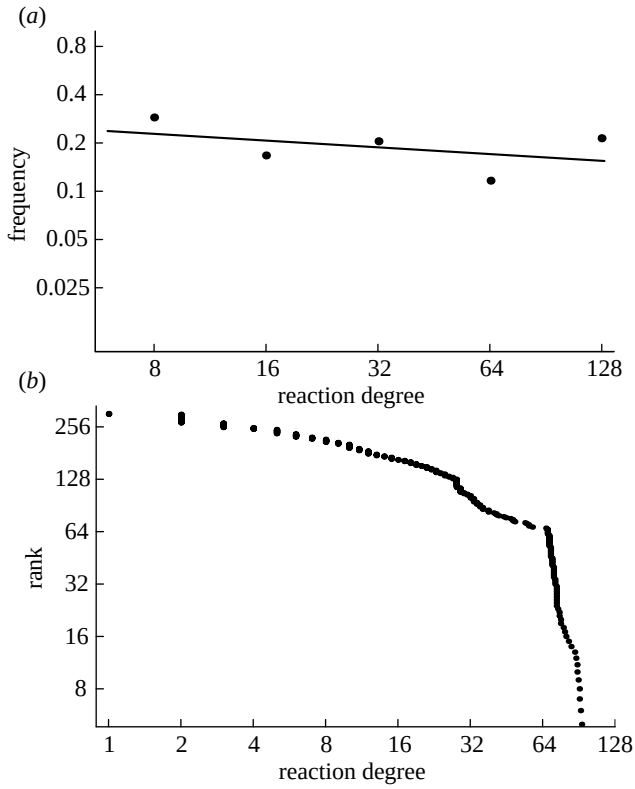

Figure 3. Degree distribution in the reaction graph. Plotted are the degree of nodes in the reaction graph versus binned frequency in (*a*) and rank in (*b*), as in figure 2: (*a*) already indicates that the degree distribution does not follow a power law $\tau = -0.13 \pm 0.16$, and (*b*) shows further that no simple cumulative probability function would appropriately approximate the rank distribution shown.

than random graphs. However, its characteristic path length (within 5% or less than 0.1 steps, of that of an equally sparse random graph) is very small (figure 4*b*). The high clustering coefficient of the substrate graph can be shown to be the result of local interactions within metabolic pathways, the 'cliques' in this network. To illustrate this, we analysed separately the substrate graphs of 10 of the longest individual pathways in our metabolic network. The analysed pathways comprise 203 substrates and include glycolysis, the tricarboxylic acid cycle, biosyntheses of riboflavin, folate, histidine, branched-chain amino acids, aromatic amino acids, threonine and lysine, arginine, putrescine and spermidine, porphyrin and haem, and coenzyme A. Their mean clustering coefficient is $C = 0.44$ ($\sigma = 0.14$, $n = 10$), not significantly different from that of $C = 0.48$ measured for the whole network. When considered as separate pathways, the coefficient of variation s in vertex degree (mean vertex degree averaged over 10 pathways, $\bar{k} = 3.2$) is found to be $s = 0.52$, 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, for which $s = 0.39$. 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 10 separate pathways, which is $L = 3.08$ ($s = 0.62$), and thus not much smaller than the $L = 3.88$ observed for the whole network.

4. DISCUSSION

Like most graph theoretical models, our model of metabolic networks omits most quantitative information, and is suited only to analyse network topography. Jeong *et al.* (2000) have simultaneously developed a graph analysis of metabolism with a more complicated graph that includes enzymes and enzyme–substrate complexes as intermediates along with the substrates. Furthermore, they represented reversible reactions twice, once for each

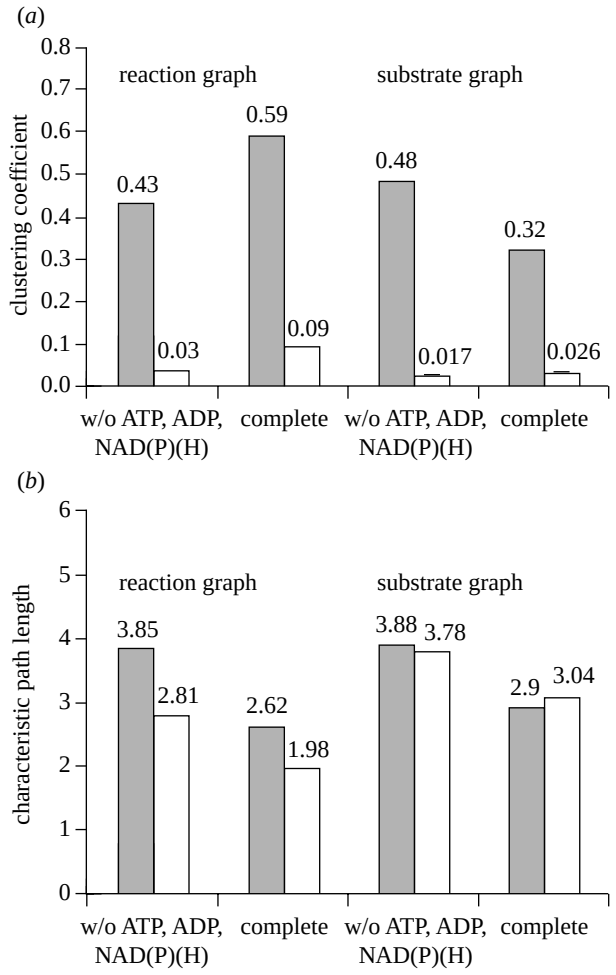

Figure 4. Metabolic network graphs are small-world graphs. (*a*) Clustering coefficients C and (*b*) characteristic path length L for the reaction graphs, substrate graphs (both shaded), and random graphs (empty bars). In (*b*), the similarity is put in context by the maximal L for the same connectivity ($L_{\max} \approx n/2(k+1)$; Watts 1997). Using table 1, $L_{\max}$ is 15.14, 5.38, 24.2 and 16.9 from left to right for the metabolic graphs shown. Values shown for random graphs are mean and standard deviations (error bars) over 100 numerically generated connected random graphs with the same probability of connection between vertices.

direction. Whilst this preserves more information, it also inflates the size of the graph. They too observed power-law scaling in the connectivity, but were more concerned about robustness of the network properties to deletion of nodes, which they claimed was equivalent to mutation. In fact, in their representation, a deleterious mutation in an enzyme would correspond to the removal of a set of edges, or all nodes corresponding to substrate complexes of that enzyme, not a substrate node as they investigated.

What might be the functional or phylogenetic significance of our observed pattern of 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 the observed structure is determined by chemical constraints alone. We cannot strictly exclude this possibility, but some evidence suggests this could not be the only constraint. First, the biosyntheses of various compounds, such as lysine and isopentenyl diphosphate, occur by different routes in different organisms (Rohmer *et al*. 1993). Recent analysis of the tricarboxylic acid 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 (Meléndez-Hevia *et al*. 1996). Moreover, considerable variation exists in the presence or absence of particular reactions in the tricarboxylic acid cycle in 19 prokaryotes with completely sequenced genomes (Huynen *et al*. 1999). Strikingly, in a majority of these species, the tricarboxylic acid cycle appears incomplete or absent. 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.

Could the observed network structure be an indicator of the evolutionary history of metabolism? Barabási & Albert (1999) have recently proposed a mathematical model that generates large graphs from small graphs by adding nodes and edges. If links to new nodes are made preferentially from 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 ones 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. Ribonucleotide 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 (Benner *et al*. 1989). Glycolysis and the tricarboxylic acid cycle are perhaps the most ancient metabolic pathways, and various of their intermediates (2-oxoglutarate, 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 first used (Benner *et al*. 1989; Taylor & Coates 1989; Morowitz 1992; Kuhn & Waser 1994; Waddell & Bruce 1995; Lahav 1999). The potential relation between evolutionary history and connectivity of metabolites corroborates a postulate put forward by Morowitz (1992), namely that intermediary metabolism recapitulates the evolution of biochemistry. Our highly connected metabolites pyruvate, 2-oxoglutarate, acetyl CoA and oxaloacetate are identified by Morowitz (1999) 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.

What 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 between metabolic states has been recognized and discussed by other authors (Easterby 1986; Schuster & Heinrich 1987; Cascante *et al*. 1995). Any response to a perturbation and transition to a new metabolic state requires that information about the perturbation has spread within the network. Watts & Strogatz (1998) studied how fast perturbations spread through small-world networks. 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. Thus small-worldness may allow a metabolism to react rapidly to perturbations.

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.

Financial support by the Santa Fe Institute for a visit by D.A.F. that enabled the completion of this work is gratefully acknowledged.

REFERENCES

Albert, R., Jeong, H. & Barabási, A. L. 1999 Internet: diameter of the world-wide web. *Nature* **401**, 130–131.

Bairoch, A. 1999 The ENZYME data bank in 1999. *Nucleic Acids Res.* **27**, 310–311.

Bak, P. 1990 Self-organized criticality. *Physica A* **163**, 403–409.

Barabási, A. L. & Albert, R. 1999 Emergence of scaling in random networks. *Science* **286**, 509–512.

Benner, S. A., Ellington, A. D. & Tauer, A. 1989 Modern metabolism as a palimpsest of the RNA world. *Proc. Natl Acad. Sci. USA* **86**, 7054–7058.

Bollobás, B. 1985 *Random graphs*. London: Academic Press.

Cascante, M., Meléndez-Hevia, E., Kholodenko, B. N., Sicilia, J. & Kacser, H. 1995 Control analysis of transit-time for free and enzyme-bound metabolites—physiological and evolutionary significance of metabolic response-times. *Biochem. J.* **308**, 895–899.

Chiva, E. & Tarroux, P. 1995 Evolution of biological regulation networks under complex environmental constraints. *Biol. Cybernet.* **73**, 323–333.

Cohen, J. E. & Briand, F. 1984 Trophic links of community food webs. *Proc. Natl Acad. Sci. USA* **81**, 4105–4109.

Easterby, J. S. 1986 The effect of feedback on pathway transient response. *Biochem. J.* **233**, 871–875.

Edwards, J. S. & Palsson, B. O. 2000 The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics and capabilities. *Proc. Natl Acad. Sci. USA* **97**, 5528–5533.

Fell, D. A. 1997 *Understanding the control of metabolism*. London: Portland Press.

Fell, D. A. & Sauro, H. M. 1990 Metabolic control analysis by computer: progress and prospects. *Biomed. Biochim. Acta* **49**, 811–816. See (http://members.tripod.co.uk/sauro/biotech.htm).

Fell, D. A. & Small, J. R. 1986 Fat synthesis in adipose tissue: an examination of stoichiometric constraints. *Biochem. J.* **238**, 781–786.

Glass, L. & Hill, C. 1988 Ordered and disordered dynamics in random networks. *Europhys. Lett.* **41**, 599–604.

Gopikrishnan, P., Meyer, M., Amaral, L. A. N. & Stanley, H. E. 1998 Inverse cubic law for the distribution of stock-price variations. *Eur. Phys. J.* B **3**, 139–140.

Graham, R. L., Groetschel, M. & Lovasz, L. (eds) 1995 *Handbook of combinatorics*. Cambridge, MA: MIT Press.

Heinrich, R. & Schuster, S. 1996 *The regulation of cellular systems*. New York: Chapman & Hall.

Hofmeyr, J.-H. S. 1989 Control pattern analysis of metabolic pathways: flux and concentration control in linear pathways. *Eur. J. Biochem.* **186**, 343–354.

Holmes, W. H. 1986 The central metabolic pathways of *Escherichia coli*: relationship between flux and control at a branch point, efficiency of conversion to biomass, and excretion of acetate. *Curr. Top. Cell. Reguln* **28**, 69–105.

Huynen, M. A., Dandekar, T. & Bork, P. 1999 Variation and evolution of the citric acid cycle: a genomic perspective. *Trends Microbiol.* **7**, 281–291.

Ingraham, J. L., Maaløe, O. E. & Neidhardt, F. C. 1983 *Growth of the bacterial cell*. Sunderland, MA: Sinauer Associates.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A. L. 2000 The large-scale organization of metabolic networks. *Nature* **407**, 651–654.

Karp, P., Riley, M., Paley, S., Pellegrini-Toole, A. & Krummenacker, M. 1999 Eco Cyc: Encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.* **27**, 55–58.

Kauffman, S. A. 1967 Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* **22**, 437–467.

Keitt, T. H. & Stanley, H. E. 1998 Dynamics of North-American breeding bird populations. *Nature* **393**, 257–260.

Kuhn, H. & Waser, J. 1994 On the origin of the genetic code. *FEBS Lett.* **352**, 259–264.

Lahav, N. 1999 *Biogenesis*. New York: Oxford University Press.

Mehlhorn, K. & Naeher, S. 1999 *The LEDA platform of combinatorial computing*. Cambridge University Press.

Meléndez-Hevia, E., Waddell, T. G. & Cascante, M. 1996 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. *J. Mol. Evol.* **43**, 293–303.

Milgram, S. 1967 The small-world problem. *Psychol. Today* **2**, 60–67.

Morowitz, H. J. 1992 *Beginnings of cellular life: metabolism recapitulates biogenesis*. New Haven: Yale University Press.

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

Murre, J. M. & Sturdy, D. P. F. 1995 The connectivity of the brain: multilevel quantitative analysis. *Biol. Cybernet.* **73**, 529–545.

Neidhardt, F. C. 1996 *Escherichia coli and Salmonella*: *molecular and cellular biology*. Washington, DC: ASM Press.

Newman, M. E. J. & Eble, G. J. 1999 Power spectra of extinction in the fossil record. *Proc. R. Soc. Lond.* B **266**, 1267–1270.

Pramanik, J. & Keasling, J. D. 1997 Stoichiometric model of *Escherichia coli* metabolism: incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. *Biotechnol. Bioeng.* **56**, 398–421.

Reder, C. 1988 Metabolic control theory: a structural approach. *J. Theor. Biol.* **135**, 175–201.

Rohmer, M. M. K., Simonin, P., Sutter, B. & Sahm, H. 1993 Isopentenyl diphosphate synthesis in bacteria does not proceed via the acetate/mevalonate pathway used by mammals. *Biochem. J.* **295**, 517–524.

Schilling, C. H., Schuster, S., Palsson, B. O. & Heinrich, R. 1999 Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol. Prog.* **15**, 296–303.

Schuster, S., Dandekar, T. & Fell, D. A. 1999 Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.* **17**, 53–60.

Schuster, S., Fell, D. A. & Dandekar, T. 2000 A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.* **18**, 326–332.

Schuster, S. & Heinrich, R. 1987 Time hierarchy in enzymatic-reaction chains resulting from optimality principles. *J. Theor. Biol.* **129**, 189–209.

Selkov, E. (and 11 others) 1996 The metabolic pathway collection from EMP—the enzymes and metabolic pathways database. *Nucleic Acids Res.* **24**, 26–28.

Sen, A. K. 1991 Quantitative analysis of metabolic regulation: a graph-theoretic approach using spanning trees. *Biochem. J.* **275**, 253–258.

Stryer, L. 1995 *Biochemistry*. New York: Freeman.

Taylor, F. J. R. & Coates, D. 1989 The code within the codons. *Biosystems* **22**, 177–187.

Varma, A. & Palsson, B. O. 1993 Metabolic capabilities of *Escherichia coli*. 1. Synthesis of biosynthetic precursors and cofactors. *J. Theor. Biol.* **165**, 477–502.

Waddell, T. G. & Bruce, G. K. 1995 A new theory on the origin and evolution of the citric acid cycle. *Microbiología Sem.* **11**, 243–250.

Wagner, A. 1996 Does evolutionary plasticity evolve? *Evolution* **50**, 1008–1023.

Watts, D. J. 1997 The structure and dynamics of small-world systems. PhD thesis, Cornell University, Ithaca, NY.

Watts, D. J. & Strogatz, S. H. 1998 Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442.

As this paper exceeds the maximum length normally permitted, the authors have agreed to contribute to production costs.