12 Structural properties of metabolic networks: implications for evolution and modelling of metabolism.

D.A. Fell[1] and A. Wagner[2]
[1]School of Biological and Molecular Sciences, Oxford Brookes University, Headington, Oxford, OX3 0BP, U.K.
[2]Department of Biology, University of New Mexico, 167A Castetter Hall, Albuquerque, NM 87131-1091 U.S.A.

Introduction

Modelling and analysis of pathways is usually done with simple representative structures, such as linear sequences, short branched sequences or cycles. Yet the advent of complete genome sequences is focussing attention on the full phenotype of cells, and for metabolism this implies the total metabolic network expressed by the cell. For example, whether a knockout mutation will cause a serious metabolic defect can only be predicted in the context of the whole network, and the surprising fact is that essential enzymes prove to be a minority [1,2]. It is therefore reasonable to ask what are the structural characteristics of real metabolic networks, so that we can build realistic models of them.

In this context, the structure of a network means its connectivity, as encoded in a stoichiometry matrix derived from a list of the balanced equations of all the reactions that can occur in a cell under specified conditions. Thus kinetic considerations (including regulatory interactions) are excluded. However, defining the characteristics of a network structure is not straightforward. In general, a network could be regular, with every metabolite linked by m reactions to its immediate neighbours to give a lattice structure, but clearly, metabolism is not like that. At the other extreme, a network could be random, in terms of which metabolites are connected by a reaction. The general properties of random graphs can be predicted. The only clearly–defined alternative to regular and random networks is the ‘small world’ network [3], which combines the ‘local’ clustering of connections characteristic of regular networks with occasional ‘long range’ connections between clusters such as can be expected to occur in random networks.

By defining measures that distinguished between these three types of network, Watts & Strogatz [3] showed that several biological, technological and social networks were of the ‘small world’ type. Here we take the central metabolic network of *Escherischia coli*, as a relatively complete and well–studied representative of a typical metabolism, and determine its type according to these measures, since this has implications for the network’s dynamics and genesis.

Methods

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 [4–8] under aerobic growth with glucose as sole carbon source and O_2 as electron acceptor.

From these reaction equations, a stoichiometric matrix [9] was automatically generated from the reaction list using Sauro’s software package INDIGO (http://members.tripod.co.uk/sauro/biotech.htm). From this matrix a substrate graph was derived, with glucose and inorganic materials regarded as external.

The substrate graph $G_S = (V_S, E_S)$ is defined by a vertex set V_S consisting of all chemical compounds (metabolites) that occur in the network. Two metabolites are regarded as adjacent if they occur (either as substrates or products) in the same chemical reaction. Note that this graph indicates the existence of a possible direct influence of one metabolite on the concentration or rate of reaction of another. There is not a one to one correspondence between edges and reactions, nor is the graph directed so as to indicate substrate–product relationships.

In this graph, 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 $v_0, v_1, \dots v_{i-1}, v_i$ from v_0 to v_i. The metabolic network, because of mass conservation and because all the carbon of the biomass can be derived from a single source, must be a connected graph, i.e., all vertex pairs are connected. The path length l is defined as the number of edges in the shortest path between v_0 and v_i. The mean path length from a vertex is the average of the path lengths to all other vertices, and is also known as the importance number. The vertex with the lowest importance number is arguably the ‘centre’ of the graph, and is the justification for Erdos being the centre of the graph of mathematical collaborations on publications, and for Kevin Bacon being until recently the centre of a film star database. The characteristic path length L of a graph is the path length between two vertices, averaged over all pairs of vertices.

Another important quantity [3] is the clustering coefficient $C(v)$ of a vertex v, which 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.

The properties of the substrate graph were compared with those of a random graph with the same number of vertices n and mean degree k [10,11]. In connected sparse random graphs with n nodes and average degree k ($k \ll n$), the probability p of two vertices being connected is given by $p = k/(n-1)$. Expressions can be derived for the distribution of vertex degree, the clustering coefficient and characteristic path length of such graphs. Among all connected 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 [12].

Results

Because of the ubiquity of coenzymes such as ATP, ADP and NAD, etc, we report here on the results where they are omitted from the substrate graph of central energy and biosynthetic metabolism of *Escherichia coli*, leaving 275 metabolites. The substrate graph is sparse with the average degree of each metabolite only 4.76, small compared to the maximal possible degree $k = n - 1$. Variation in connectivity in the substrate graph is greater, with a standard deviation in degree (σ_k) of 4.79 compared with an average of 2.12 for corresponding random graphs. This implies that some vertices in substrate graphs have many more, and others many fewer neighbors than vertices for a random graph. Indeed, a histogram of degree *vs.* frequency, or a rank distribution of metabolites, where the metabolite with the highest connectivity was assigned rank 1 is consistent with a power-law (not shown). Given this large dispersion, k–regular random graphs (used in modelling neural and genetic networks) would be particularly poor statistical models of metabolic networks.

13 key metabolites of particularly high connectivity are listed in Table 12.1, of which the top five are glutamate, coenzyme A, 2–oxoglutarate, pyruvate, and glutamine. This list overlaps with sets of key metabolic intermediates of *E.coli* selected on other criteria by other authors in metabolite balancing studies, where they represent the common biosynthetic source of all cell materials. (e.g., Varma and Palsson [13] and Ingraham et al. [14] used a set of 12 biosynthetic precursors. Holmes [15] chose a smaller subset of 8 key precursors from which all cell biomass could be produced.)

The architecture of the *C. elegans* nervous system, the power grid of the western United States, the structure of some sociological networks and the world wide web, like the *E. coli* graph, are all small-world graphs, formally characterized by Watts [10] and Barabási and Albert [16]. Small-world graphs were first illustrated [17] with friendship networks in sociology (‘six degrees of separation’). Friendship networks are sparse (each of the 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). Even though most of the few connections per individual are tied up in

Table 12.1 Thirteen key metabolites of *E. coli* metabolism. Metabolites with connectivity significantly higher than the mean metabolite degree are shown. For comparison, the thirteen metabolites with the shortest mean path lengths (importance number) are shown.

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 pyroP	16	2-oxoglutarate	2.99
tetrahydrofolate	15	phosphoribosyl pyroP	3.10
succinate	14	anthranilate	3.10
3-phosphoglycerate	13	chorismate	3.13
serine	13	valine	3.14
oxoisovalerate	12	3-phosphoglycerate	3.15

local interactions within "cliques" of individuals, every individual in the U.S. may be linked to every other by a short chain of acquaintances.

The more formal definition of a small-world graph is that it is sparse but much more highly clustered than an equally sparse random graph ($C \gg C_{random}$), with a characteristic path length L that is close to the theoretically possible minimum (which is well approximated by a random graph [10]). The reason why a graph can have small L despite being highly clustered is that the few nodes connecting distant clusters suffice to lower L [10]. Hence ‘small-worldness’ is a global property not apparent from the local graph properties. The substrate graph illustrates this property particularly well. Its characteristic path length ($L = 3.88$) is only 3% (approximately 0.1 steps) above that of an equally sparse random graph, but it is 28 times more clustered ($C = 0.48$).

What might be the functional or phylogenetic significance of the observed pattern: the power law distribution of connectivity, and the small-world nature of the metabolic graph? It is possible that there is no such significance, because the laws of chemistry might almost completely constrain network structure. However, recent analysis of the tricarboxylic acid (TCA) cycle 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 [18]. This at least suggests that chemistry does allow flexibility in the design of a metabolic net, so the observed architecture may reflect both evolutionary history and evolutionary optimization.

Could the observed network structure be an indicator of the evolutionary history of metabolism? Barabási and Albert [16] have recently proposed a mathematical model that generates small-world graphs with power-law degree distributions: large graphs are made from small graphs by adding nodes and edges, with new links formed preferentially at nodes that already have many links. Consequently 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. Now many of the most highly connected metabolites in Table 12.1 have a proposed early evolutionary origin. Glycolysis and the TCA cycle are perhaps the most ancient metabolic pathways, and various of their intermediates (2–oxoglutarate, succinate, pyruvate, 3-phosphoglycerate) occur in Table 12.1. Early proteins are thought to have been made of many fewer amino acids than extant proteins, and the highly connected amino acids glutamine, glutamate, aspartate, and serine are thought to be those used earliest [19–24]. The potential relation between evolutionary history and connectivity of metabolites corroborates a postulate put forth and defended forcefully by Morowitz [19], namely that intermediary metabolism recapitulates the evolution of biochemistry. Our highly connected metabolites pyruvate, 2–oxoglutarate, acetyl CoA and oxaloacetate are identified by Morowitz [25] 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.

A small-world network might also optimize metabolic function. Metabolic networks are subject to perturbations, and every component in the network could be affected by such perturbations, because they are all connected. The importance of minimizing the transition time betwen metabolic states has been recognized and discussed by other authors [26,27]. Any response to a perturbation and transition to a new metabolic state requires that information about the perturbation has spread through the network. Watts and Strogatz [3] 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.

Whether or not there is special significance to the small-world character of metabolic networks, what is certain is that models of small idealised sections of metabolism can never be fully representative of the global properties of metabolism.

References

1. Thatcher, J. W., Shaw, J. M. and Dickinson, W. J. (1998) Marginal Fitness Contributions of Nonessential Genes in Yeast. *Proc. Natl. Acad. Sci. USA* **95**, 253–257.

2. Schilling, C. H. and Palsson, B. O. (1998) The Underlying Structure of Biochemical Reaction Networks. *Proc. Natl. Acad. Sci. USA* **95**, 4193–4198.

3. Watts, D. J. and Strogatz, S. H. (1998) Collective Dynamics of 'Small–World' Networks. *Nature (London)* **393**, 440–442.

4. Neidhardt, F. C. (1996) *Escherichia coli* and *Salmonella. Molecular and Cellular Biology.*, ASM Press, Washington DC.

5. Karp, P., Riley, M., Paley, S., Pellegrini-Toole, A. and Krummenacker, M. (1999) Eco Cyc: Encyclopedia of *Escherichia coli* Genes and Metabolism, *Nucleic Acids Res.* **27**, 55–.

6. Pramanik, J. and Keasling, J. D. (1997) Stoichiometirc Model of *Escherichia coli* Metabolism: Incorporation of Growth–Rate Dependent Biomass Composition and Mechanistic Energy Requirements. *Biotechnol. Bioeng.* **56**, 398–421.

7. Selkov, E., Basmanova, S., Gaasterland, T., Goryanin, I., Gretchkin, Y., Maltsev, N., Nenashev, V., Overbeek, R., Panyushkina, E., Pronevitch, L., Selkov, E. and Yunis, I. (1996) The Metabolic Pathway Collection from EMP — The Enzymes and Metabolic Pathways Database. *Nucleic Acids Res.* **24**, 26–28.

8. Bairoch, A. (1999) The ENZYME Data Bank in 1999. *Nucleic Acids Res.* **27**, 310–311.

9. Heinrich, R. and Schuster, S. (1996) *The Regulation of Cellular Systems*, Chapman and Hall, New York.

10. Watts, D. J. (1997) *The Structure and Dynamics of Small-World Systems*, Ph.D. thesis, Cornell University.

11. Bollobás, B. (1985) *Random Graphs*, Academic Press, London.

12. Mehlhorn, K. and Naeher, S. (1999) *The LEDA Platform of Combinatorial Computing*, Cambridge University Press, Cambridge.

13. Varma, A. and Palsson, B. O. (1993) Metabolic Capabilities Of *Escherichia coli.* 1. Synthesis of Biosynthetic Precursors and Cofactors. *J. Theoret. Biol.* **165**, 477–502.

14. Ingraham, J. L., Maalœ, O. E. and Neidhardt, F. C. (1983) *Growth of the Bacterial Cell.*, Sinauer Associates Inc., Sunderland, MA.

15. 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. Cellul. Regul.* **28**, 69–105.

16. Barabási, A. L. and Albert, R. (1999) Emergence of Scaling in Random Networks. *Science* **286**, 509–512.

17. Migram, S. (1967) The Small–World Problem. *Psychology Today* **2**, 60–67.

18. Meléndez-Hevia, E., Waddell, T. G. and Cascante, M. (1996) The Puzzle of the Krebs Citric Acic Cycle: Assembling the Pieces of Chemically Feasible Reactions and Opportunism in the Design of Metabolic Pathways During Evolution. *J. Mol. Evol.* **43**, 293–303.

19. Morowitz, H. J. (1992) *Beginnings of Cellular Life: Metabolism Recapitulates Biogenesis.*, Yale University Press, New Haven.

20. Kuhn, H. and Waser, J. (1994) On the Origin of the Genetic Code. *FEBS Letters* **352**, 259–264.

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

22. Benner, S. A., Ellington, A. D. and Tauer, A. (1989) Modern Metabolism as a Palimpsest of the RNA World. *Proc. Natl. Acad. Sci. USA* **86**, 7054–7058.

23. Waddell, T. G. and Bruce, G. K. (1995) A New Theory on the Origin and Evolution of the Citric Acid Cycle. *Microbiología Sem.* **11**, 243–250.

24. Taylor, F. J. R. and Coates, D. (1989) The Code Within the Codons. *Biosystems* **22**, 177–187.

25. Morowitz, H. J. (1999) A Theory of Biochemical Organization, Metabolic Pathways and Evolution. *Complexity* **4**, 39–53.

26. Easterby, J. S. (1986) The Effect of Feedback on Pathway Transient Response. *Biochem. J.* **233**, 871–875.

27. Cascante, M., Meléndez-Hevia, E., Kholodenko, B. N., Sicilia, J. and 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.