Network Structure of Metabolic Pathways

D. J. Raine[1,*] **& V. Norris**[2]

[1]*Department of Physics and Astronomy, University of Leicester, LE1 7RH, UK*
[2]*Laboratoire des Processus Intégratifs Cellulaires, UPRES A CNRS 6037, Faculté des Sciences et Techniques de Rouen, 76821 Mont- Saint- Aignan, France*

The living cell is an autocatalytic network of metabolic pathways sustained far from equilibrium by the supply of matter and energy. At the abstract level we can regard a chart of cellular metabolic pathways as a network of undirected connexions between metabolites (the nodes of the network) each connected pair being related by an enzyme-catalyzed reaction. The clustering properties of the reaction networks can be obtained from maps of known metabolic pathways. For the number of nodes as a function of the number of connexions, a long-tailed distribution is obtained, which can be described by a power law. We also find evidence consistent with a power law in the relationship between regulatory proteins and genes. We investigate three models for the construction of metabolic networks, which we call the random connexion model, the random cluster model and the accumulation model. The last two of these give a long-tailed distribution of nodes. The random cluster and accumulation models also exhibit 'small-world' features, in agreement with the structure of real biological networks. We speculate on the possible implications for the evolution of metabolic networks.

Keywords: evolution, genes, metabolism, self-organized criticality, small world networks

1. INTRODUCTION

Abstract networks (or graphs) consist of nodes and links between them. The number of links per node, which we call the class of the node, may vary between nodes, but in a homogeneous network the statistical distribution of links will be the same from region to region across the network. Classically, the distinction amongst homogeneous networks has been made between fully ordered networks, in which the nodes are identical, and networks with nodes connected independently at random. More recently, it has been recognized that other non-random distributions of links are not only possible but occur naturally [1]. It is therefore important to investigate the structure of networks and to consider models that can reproduce observed distributions, since these will provide clues to the evolution and construction of such networks. Here we investigate the architecture of metabolic and genetic networks in biological cells.

*Author for correspondence.
+(44)116 252 2075 (voice), +(44)116 252 2070 (fax)
e-mail: jdr@leicester.ac.uk

In the following section we deal with the evidence from the literature and from our own investigations [2] that reveals the structure of metabolic networks. We find that they belong to a type of network in which the distribution of nodes as a function class follows a power law. These are the scale-free networks [3]. Such distributions are found in an increasing number of other examples [4] including the connectivity of the world wide web, the power grid of the US, and citations of scientific papers. In these cases there seems to be a clear explanation of the power law structure. Connexions in these networks are not laid down at random, but each grows by the addition of new nodes that are connected to a node of the existing network with a probability that depends on the pre-existing number of connexions. In other words, new nodes are connected preferentially to those nodes that are already most connected. For the examples cited, this assumption is qualitatively reasonable. For example, the most cited papers are the most likely to be read and hence the most likely to receive further citations. It is not so easy to see why metabolic networks should evolve in this manner. To see the problem, we consider metabolic networks as examples of autocatalytic networks [5] in which each reaction in the network (each link) is catalyzed by a molecule in the network. (Hence there are at least as many nodes that are also catalysts as there are links.) Consider now the

Received 5 June 2001

emergence of a new node (catalyst or metabolite). This might arise from an addition to the 'food set' (the external input of molecules to the network) or by some random mutation of a catalyst, or from the slow build up of the product of a weakly catalyzed reaction. It might be argued that the new metabolite is more likely to arise from one of the more highly connected nodes already present, but even so, a new catalyst would have a probability that was essentially random of catalyzing other new reactions in the network; it would not add preferentially to the highly connected nodes because the emergence of the new molecule is unrelated to its catalytic powers in other parts of the network.

These considerations raise the more general question therefore as to whether the structure of metabolic networks reflects the structure of a particular set of enzymes or can emerge in a random chemistry. The issue here is the level of self-organization in metabolic networks. Does the organization in these networks arise at the level of enzyme chemistry or at the level of network evolution, independent of the particular chemical details? Since a random growth model does not work (see below) and the scale-free model can be queried on the grounds of biological relevance it is of some interest to explore other models.

2. NUMERICAL EXPERIMENTS

We consider a metabolic network as a graph having metabolites as nodes linked by enzyme-catalyzed reactions. As an indicator of the connectivity of a network we form the distribution function of nodes with given numbers of links. A node with n links is assigned to class n. (So a metabolite of class 1 is an end-point.) If we take account of the directions of the reactions under physiological conditions then the graphs are directed with links either into or out of a given metabolite. In this case the classes refer to links in a given direction. Jeong et al. have shown that metabolic networks in 43 organisms, taken as directed reactions, have a self-similar but non-homogeneous distribution of nodes, following an approximate power law as a function of class [6]. The power laws have approximately the same slope for both educts and for products. Thus we expect an analysis which treats metabolic networks as undirected will produce similar results.

It is of interest to ask if this result is already visible in the well-known compendia of metabolic pathways. We took the Boehringer Mannheim chart [7], including all the reactions involving a given metabolite (not just those joined by explicit links in the visualization of the chart) but without regard to direction, to obtain the class of each metabolite. We also analyzed the Nicholson chart [8], but now counting the arrows into or out of each metabolite as they appear in the chart, with no account taken of links that are not explicitly indicated. We ignored connexions outside a metabolic chart itself and, since most molecules of class 1 take part in further reactions not included in the charts, we disregarded class 1 molecules in the analysis. Molecules such as ATP and water, which are ubiquitous but which do not appear on these charts as explicit nodes, were ignored as were the inputs of these molecules into reactions. We counted a total of 606 and 386 nodes in the Boehringer Mannheim and Nicholson charts respectively, restricting ourselves in the former case to general biochemical pathways plus those confined to unicellular organisms.

The result in both cases is an approximate power law distribution for the number of nodes $N(n)$ of class n:

$$N(n)=kn^{\alpha}. \qquad (1)$$

For the Boehringer Mannheim chart the least squares fit to (1) yields parameters k = 1906 and α = –2.7 (see figure 1). The comparison with the best fit

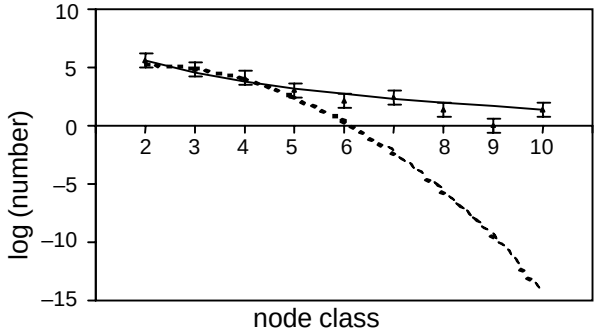

Figure 1. Best fit power law (solid line) and gaussian (dashed line) compared with the data (points), from the Boehringer Mannheim metabolic pathways chart [7]. Formal standard errors are shown. The best fit gaussian cannot reproduce both the core and tail.

gaussian distribution in figure 1 clearly shows the long tail in the data relative to a gaussian fit to the core. Attempts to fit the tail at the expense of the core give a worse formal chi-squared parameter. We conclude that the sample is not drawn from a gaussian distribution. For the Nicholson chart, the slope of the power law obtained from a log-log fit is –3.2 out to nodes of class 9, beyond which we have no data. That we obtain approximate agreement for two separate analyses gives us confidence for the view that the charts are representative and that the set of all metabolic pathways is approximately self-similar. Given that we are using a small sample from a heterogeneous subset of metabolic pathways the slope cannot be regarded as significantly different from that obtained by Jeong et al [6].

3. MODELS

We consider three models for the origin of structure in a metabolic network: random connexion, random cluster and accumulation. In the random connexion model we imagine the links between nodes to be made independently at random. Biologically, this would be the case if enzymes evolved with a random probability of catalyzing a potential reaction of the network [5]. We shall treat this as a special case of the random cluster model in which nodes are connected in groups. This might result from groups of physically associated enzymes being transferred as modules [9] or hyperstructures [10]. This model is similar, but not equivalent, to the food web models of Williams and Martinez [11]. A third model is based on the probability that a molecule acquires a new link (that is, becomes a substrate in a new reaction) is enhanced in accordance with the number of links that it already has. Such a probability might result if the number of links were proportional to the abundance of a molecule, since abundant molecules might be expected to offer more probable substrates for new reactions than scarce ones. It might also be the case that metabolites that have multiple links, and are therefore made in several ways, are more difficult to eliminate from an evolving metabolism and therefore have a higher than random survival probability. We call this the accumulation model by analogy with similar models in economics. It is similar but not equivalent to the scale-free model of Barabási and Albert [4] referred to above. In the accumulation model the network can grow by making additional links between existing nodes, as well as by adding new ones at each step in its construction.

To obtain the random connexion and random cluster distributions we use the following definition of the connectivity matrix a_{ij} of a graph. Let the nodes be labeled by $i = 1,2\ldots N$. Let the matrix element $a_{ij} = 1$ if node i is linked to node j, and let $a_{ij} = 0$ otherwise. We take $a_{ii} = 0$, so a node is not considered to be linked to itself, and $a_{ij} = a_{ji}$, so we are considering undirected graphs. Let p_c be a fixed integer between 1 and $(N - 1)/2$ and let p_l be fixed in [0,1]. Moving along each row of the matrix (a_{ij}) in turn let p and r be chosen independently from [0,1] with uniform random probability. For each row i, starting from column $j = i + 1$, choose $a_{ij} = 1$ for $j = i + 1,\ldots J = \min(N, i + [rp_c + 1])$ if $p > p_l$ and $a_{ij} = 0$ for this range of indices otherwise. The square brackets $[x]$ indicate the integer part of x. Now repeat, starting from $j = J + 1$ until $J = N$. Thus, clusters of nodes of all sizes (in the given range) are laid down at random, so a class of node of a given size can occur from all random combinations of smaller classes. The connexions from a given node are however correlated since a given link is likely to be one of a cluster. The model therefore differs from a classical random graph [12] except in the case $p_c = 1$ where the nodes are connected at random.

3.1 Random connexion model

The nodes are connected at random with $p_c = 1$. The result is the expected Poisson distribution of node classes [12] (figure 2). In this model, an autocatalytic network arises when the number of alternative final steps in the production of a molecule is so large that at least one will be catalyzed by another molecule of the network with probability near unity, as described by Kauffman [5]. We can think of this approximately as follows: each molecule of the network catalyzes a reaction in the network with some probability, and hence provides a random link in the graph. This is roughly equivalent to the case $p_c = 1$ above, which does not give a power law distribution of nodes. The equivalence is not exact because the Kauffman networks are anisotropic with larger molecules potentially linked by more pathways. Nevertheless, this difference is not crucial and Poisson statistics holds approximately in Kauffman's model, which limits the extent to which it can be taken to describe the evolution of real metabolic pathways.

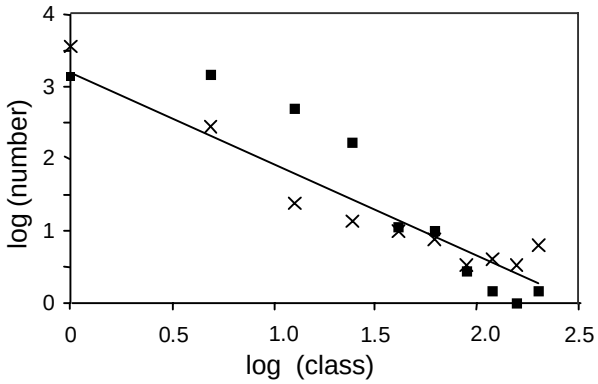

Figure 2. The distribution function for node classes from two models. Crosses: the average of 4 runs of the random cluster model for a network of 100 nodes. Squares: the average of 4 runs of the random connexion model for a similar network. Solid line: best fit straight line to the random cluster model. The values $p_c = 10$, $p_l = 1 - 2/N$, and $N = 100$ were chosen for ease of simulation.

3.2 The random cluster model

We have been able to generate a network of nodes with an approximate power law distribution of links in the random cluster model defined above. Figure 2 gives an example showing that for $p_c >> 1$ the result is an approximate power law distribution of node classes.

We can understand this roughly as follows. A cluster of size n at a node is linked randomly to n other nodes. Thus each time we insert a node in the tail of the class distribution (above the diagonal of the connexion matrix) we add a number of uncorrelated instances of lower class nodes (below the diagonal). The combination gives a long-tailed distribution that approximates to a power law. Note that as described here the model appears not to be statistically homogeneous. Since connexions can only be made from a given node to nodes with higher labels the later nodes are treated differently from the earlier ones. However, by filling in the whole matrix with random clusters, not just entries above the diagonal, and symmetrizing the resulting matrix we would treat all nodes on the same footing and we would obtain an equivalent distribution of node classes. Thus the apparent departure from homogeneity is not significant.

From the numerical simulation of the model we can only claim an approximate fit to a power law distribution; we cannot rule out an exponential tail. Barabási et al. [3] have looked at the conditions under which a power law distribution is possible and rule out models in which the number of nodes is fixed *a priori*. In our case the *maximum* number of nodes available to the growing network is fixed in order to simplify the programming, but the number of nodes actually connected in the largest connected set is not determined *a priori*. The fact that the network grows through clusters of connexions is an essential difference from a random connexion model.

3.3 Accumulation model

We label a set of N nodes sequentially and we build a network by adding a connexion (i, j), between nodes i and j ($i, j = 1 \ldots N$), with probability $p(1 + f(n))$, where p is fixed and less than $(1 + \max(f(n)))^{-1}$ and $f(n)$ is a function of the number of connexions n already present from node i. The function $f(n)$ is chosen such that the probability that a node acquires a new connexion increases with the number of connexions it already has. Provided this condition is satisfied the results are not qualitatively dependent on the form of f. We used both powers and exponentials. The results for $f(n) = \lambda n^2$, λ a constant, are shown in figure 3 and compared with a power law. It is clear that in this case the chosen parameters give rise to a tail in the distribution which can be described as an approximate power law.

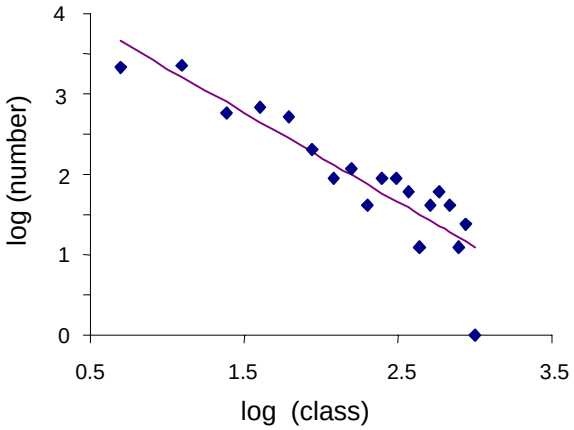

Figure 3. The accumulation model in which the probability of making the another connexion from a node with n present connexions is taken to be $0.01\times (1 + 0.0015\, n^2)$. The data was obtained for a network of 200 nodes.

4. SMALL WORLDS

Small worlds are networks that are linked in such a way that they exhibit a high degree of clustering (like ordered networks) but a relatively short average number of links between any two nodes (like random networks) [1]. Such networks are associated with a degree of robustness and efficiency [13]. To investigate the small-worlds connectivity of the random cluster model we translate the criteria of Watts and Strogatz into graph theoretic terms. Let $A = (a_{ij})$ be the matrix of connexions. The cliquishness of the network describes the average number of times any two nodes connected to a third node are themselves connected. This is equivalent to counting the number of closed triangles in the network, which is proportional to

$$c = tr\left(A^3\right) \tag{2}$$

We therefore define the relative clustering parameter $C = c/N$.

The other key parameter is the minimum number of steps connecting arbitrary pairs of points averaged over the graph. This is slightly harder to define in a computationally convenient way. We propose the following. Let the products $[A^n]$ be defined such that its $(i, j)^{\text{th}}$ element $[a^n{}_{ij}]$ is either 1 or 0 according to whether the corresponding elements of the product matrix A^n are positive or zero. The probability that an arbitrary pair of nodes is connected after l steps equals the fraction of the N $(N - 1)$ non-diagonal elements of $[A^n]$ that are non-zero. This gives

$$f = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}[a_{ij}{}^{l}]}{N(N-1)} \tag{3}$$

and suggests we define a path length parameter L as the value of l for which $f = 0.5$, say. In fact, in general, the product matrices break up into direct products, so f does not tend to 1 as $l \to \infty$. This is equivalent to saying that there are nodes that are not connected to anything, and these should really be removed before we analyze the network. We have therefore replaced N in equation (3) by the number of connected nodes (as determined for each numerical experiment). In practice, for the parameter values in our simulations, this makes only a small difference to the results.

We find the following. For a 100 node ordered network with each node linked to its four nearest neighbors, the cliquishness $C = 5.8$ and the length parameter $L = 14$. For the random network, C is small as we would expect , typically $C < 0.2$ and $L \sim 8$. As we increase the clustering, with p_c taking values between about 5 and 20 the clustering increases (C between about 0.2 and 2) but the length remains close to the random value ($L \sim 4$ - 6). This is indicative of small worlds behaviour and agrees qualitatively with the results of Jeong et al. [6]. We find a similar behaviour in the accumulation model with $C \sim 2$ and $L \sim 3$ for the parameters corresponding to figure 3.

5. OTHER CELLULAR NETWORKS

Clearly the enzymes, which can be thought of as pathways entering the chart from outside, as well as the genetic apparatus, are so far missing from our analysis of cellular networks. Some additional evidence for the structure of the complete network of molecular interaction in the cell can be obtained from the investigation of Thieffry et al. who carried out a similar analysis to ours for the genetic regulatory circuits in *Escherichia coli* [14]. The low number of extended regulatory circuits surprised these authors, but in fact their data shows some support for a power law distribution. One of their sets of data giving the numbers of proteins that regulate differently sized groups of genes is shown in figure 4. Similarly, Ramsden and Vohradský find power law behaviour in protein synthesis [15]. At present the data are not available for examining the interaction between regulation of gene expression, protein synthesis and metabolism, or for investigating the detailed evolution of catalytic networks, but the models reported here do provide some constraints on attempts at an integrative biology [16].

6. EVOLUTION OF RANDOM NETWORKS

It is tempting to consider the departure from randomness in the power law distribution of nodes,

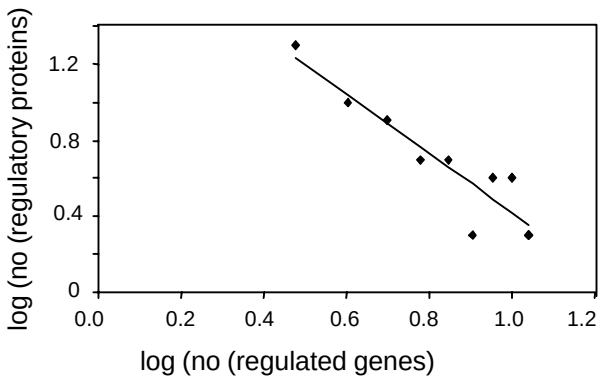

Figure 4. Regulated genes in *E. coli*.The number of proteins that regulate a given number of genes is plotted. The data is taken from Thieffry et al.[13] omitting proteins that regulate only one or two genes in order to show the power law tail. The best fit straight line has slope −1.44.

and small-world structure, in metabolic pathways, and possibly all cellular networks, as suggestive of self-organized criticality. To make this connexion explicit we need to relate the power law distribution to the dynamical process that generates it. For example, the classical exemplar of self-organized criticality in a model sand pile produces a power law distribution of avalanches in a response to an external supply of material. Could self-organized criticality in metabolic networks result from the dynamics of early cells? We can obtain some insight into these dynamics by considering the network produced by the fanning out of connexions from a single initial metabolite.

Let us label each metabolite by $s_i = +1, 0$. Initially we have $s_i = 0$ for all i. We choose an arbitrary node j and set $s_j = 1$ for this node. At the next step we set $s_i = 1$ for all nodes that are connected to node j. The vector (s_i) describes the state of the system at each stage and is obtained from the connexion matrix by $s_i \to [\Sigma_j a_{ij} s_j]$, where $[x]=1$ if $x > 0$ and $x = 0$ otherwise. This produces avalanches of all sizes of nodes with $s_i = 1$ (figure 5), characteristic of the critical state [17]. It is also of interest that the statistical distribution of node classes is governed by clusters of connexions over a range of sizes, which is characteristic of behaviour far from equilibrium.

At this stage the mapping onto a dynamical system has only an abstract significance. The crucial next step is to change the viewpoint to see the avalanches not as the turning on of nodes in an existing network, but as the way in which an autocatalytic network might have grown. (For this purpose we have to think of a metabolic network as part of an autocatalytic network.) Over some long timescale we envisage mutations or additions to the food set whereby a new enzyme or metabolite appears. This may result in the catalysis of one or more further

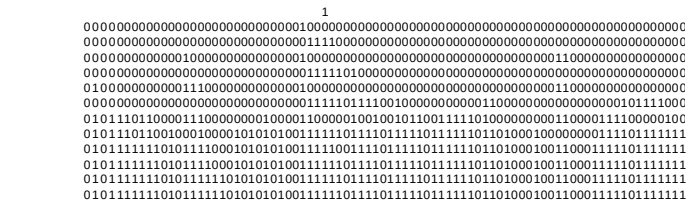

Figure 5. The development of avalanches from a single metabolite over part of the network. The top row shows the location of a randomly selected node, labeled 1. The label 0 for non-selected nodes used elsewhere in the table is suppressed in the first row for clarity. The second row shows the location of those nodes (in the part of the network displayed here) of those nodes that are connected to that in the top row in one representation of the random cluster model. Those not connected are labeled 0. Similarly the third row shows the nodes, labeled 1, connected to all those labeled 1 in the second row and those that are not connected, labeled 0. After 11 steps a steady pattern of connections has built up (at least in the part of the network displayed here) showing connected components of all sizes (numbers of contiguous 1s).

new reactions leading to a cascade of new enzymes and metabolites on a short timescale until the process comes to a halt. At each appearance of a new catalyst there will be an avalanche of some number of further new catalysts and metabolites. This presence of two timescales, one long, one short, is one ingredient of a self-organized critical system. The other aspect is some critical parameter that is self-tuning. Here we note the surprising fact that metabolic networks appear to have a constant diameter (average number of reaction steps between a random pair of nodes) independent of the size (number of nodes) of the network [6]. We speculate that networks may evolve in a self-similar manner to maintain the diameter at its critical value.

ACKNOWLEDGEMENTS

A preliminary version of this paper was presented at the International Conference on Complex Systems in Nashua, New Hampshire, May 21–26, 2000. We are grateful to Denis Thieffry for making his data available to us and for conversations with Dick D'Ari, Jeremy Ramsden, Lois Le Sceller, Janine Guespin, Tim Raine, Gordon Stewart and Dick Willingale. DJR acknowledges the hospitality of Camille Ripoll at the IFR 'Systemes Intégrés', Rouen.

REFERENCES

1. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* (*Lond.*) **393** (1999) 440–442.
2. Raine, D. J. and Norris, V. Self organisation in metabolic pathways. *Inter.Journal of Complex Systems* B Article, No.361
3. Barabási, A.-L., Albert, R. & Jeong, H. Mean-field theory for scale-free random networks. *Physica* A **272** (1999) 173–187.
4. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286** (1999) 509–512.
5. Kauffman, S. (1993) *The Origins of Order.* Oxford University Press, Oxford
6. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A.-L. The large-scale organization of metabolic networks. *Nature* (*Lond.*) **407** (2000) 651–654.
7. Michal, G. *Biochemical Pathways* (Boehringer Mannheim, Roche Diagnostics Corporation, Switzerland) (3rd edition, 1993)
8. Nicholson, D. E. *Metabolic Pathways* (16th edition) (Koch-Light Laboratories, Boston, Mass.) (1976)
9. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* (*Lond.*) **402** (1999) C47–52.
10. Norris, V., Alexandre, S., Bouligand, Y., Cellier, D., Demarty, M., Grehan, G., Gouesbet, G., Guespin, J., Insinna, E., Le Sceller, L., Maheu, B., Monnier, C., Grant, N., Onoda, T., Orange, N., Oshima, A., Picton, L., Polaert, H., Ripoll, C., Thellier, M., Valleton, J.-M., Verdus, M.-C., Vincent, J.-C., White, G. & Wiggins, P. Hypothesis: hyperstructures regulate bacterial structure and the cell cycle. *Biochimie* **81** (1999) 915–920.
11. Williams, R. J. & Martinez, N. D. Simple rules yield complex food webs. *Nature* (*Lond.*) **404**(2000)180–183.
12. Erdös, P. & Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5** (1960) 17–61.
13. Albert, R., Jeong, H. & Barabási, A. L. Error and attack tolerance of complex networks. *Nature* (*Lond.*) **406** (2000) 378–382.
14. Thieffry, D., Huerta, A. M., Perez-Rueda, E. & Collado-Vides, J. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. *BioEssays* **20** (1998) 433–440.
15. Ramsden, J. J. & Vohradský, J. Zipf-like behavior in procaryotic protein expression. *Phys. Rev.* E **58** (1998) 7777–7780.
16. Ripoll, C., Guespin-Michel, J., Norris, V. & Thellier, M. Defining integrative biology. *Complexity* **4** (1998) 19–20.
17. Bak, P. (1996) *How Nature Works: the Science of Self-Organized Criticality.* Copernicus, New York.