Graphical analysis of biocomplex networks and transport phenomena

K.-I. Goh, B. Kahng*, and D. Kim

School of Physics, Seoul National University NS50, Seoul 151-747, Korea

Many biocomplex networks such as the protein interaction networks and the metabolic networks exhibit an emerging pattern that the distribution of the number of connections of a protein or substrate follows a power law. As the network theory is developed recently, several quantities describing network structure such as modularity and degree-degree correlation have been introduced. Here we investigate and compare the structural properties of the yeast protein networks for different datasets with those quantities. Moreover, we introduce a new quantity, called the load, characterizing the amount of signal passing through a vertex. It is shown that the load distribution also follows a power law and its characteristics is related to the structure of the core part of the biocomplex networks.

1. Introduction

Recently biocomplex systems have drawn considerable attentions since their emergent behaviors, arising from diverse interactions and adaptations are more than the sum of individual components [1,2]. Such complex systems may be described in terms of graphs, consisting of vertices and edges, where vertices and edges represent substrate or proteins, and their mutual reactions or interactions in metabolic networks or protein interaction networks, respectively [3-6]. In the last century, biologists mainly focused their interests on the identification of individual molecules and their functions in relation to macroscopic biological phenomena. However, it is recently believed that thousands of genes and their products such as proteins, RNA and small molecules, function in a complete and concerted way [7]. Thus it is natural to invoke the graph theory which helps us to visualize how molecules in a given organism function together in concerted ways.

The cellular components such as genes, proteins, and other molecules, connected by all physiologically relevant interactions, form a full weblike molecular architecture in a cell [8]. In such an architecture, genes are known to play a structural role, determining the scope and passing the information in a hereditary manner to subsequent generations. The functional role of gene is expressed through protein. At the biological level, proteins rarely act alone; rather they interact with other proteins to perform particular cellular functions. Thus protein-protein interactions play pivotal roles in various aspects of the structural and functional organization of the cell and their complete description is indispensible to thorough understanding

* Corresponding author. E-mail: kahng@phya.snu.ac.kr

of the cell. Proteins can be viewed as vertices of a protein-protein interaction network in which two proteins are connected if they can physically attach to each other, forming a complex network called the protein interaction network (PIN). Recently, high-throughput data-collection methods such as protein chips or semi-automated yeast two-hybrid screens have been introduced, that help to determine which proteins interact with each other in large scale. In particular, organisms with sequenced genomes such as the yeast *Saccharomyces cerevisiae* provide important test beds for analyzing such a PIN [9].

In this manuscript, we investigate the structural property of the PIN in graph theoretic aspect and also the transport phenomena on such complex networks. We first introduce several quantities describing network structure in Section 2. The structural properties of the *S. cerevisiae* PIN are analyzed specifically in Section 3. In Section 4, we consider a transport problem on complex networks. The final Section is devoted to the conclusions and discussions.

2. The degree distribution, the degree correlation function, and the clustering coefficient

Retrospectively, the graphical approach was initiated by Erdős and Rényi (ER) [10] in 1960, who were the first to study the statistical aspect of random graphs using the probabilistic method. Thus, modeling random networks has a long history, and has been particularly active as a branch of combinatorial graph theory. In graph theory, one of interesting quantities is the degree, defined as the number of edges connecting to a given vertex. The degree distribution of the ER network follows a Poisson distribution. Recently, however, there were findings that the degree distribution of the PIN follows a power law,

$$P_D(k) \sim k^{-\gamma}, \qquad (1)$$

where k means degree and γ is the degree exponent. The network displaying a power-law degree distribution is called scale-free (SF) network. Besides the PIN, SF networks [11] are ubiquitous in real-world networks such as the world-wide web (WWW) [12-14], the Internet [15-17], the citation network [18] and the author collaboration network of scientific papers [19-20], and the metabolic networks in biological organisms [21]. The SF behavior of the degree distribution can be generalized into the Pareto form,

$$P_D(k) \sim (k+k_0)^{-\gamma}, \qquad (2)$$

with a constant k_0.

In fact, the degree distribution of the yeast PIN fits better to this Pareto form, which will be discussed later.

It is known that the degrees of the two vertices located at the ends of an edge are correlated to each other. As the first step, such degree-degree correlation can be quantified in terms of the average of the degrees over neighbors of proteins with degree k as a function of k, denoted by $\langle k_{\mathrm{nn}}\rangle(k)$. In most biological networks, the function $\langle k_{\mathrm{nn}}\rangle(k)$ exhibits a decreasing behavior with increasing k. The decaying behavior is expressed roughly by another power law as

$$\langle k_{\mathrm{nn}}\rangle(k) \sim k^{-\nu}. \qquad (3)$$

On the other hand, the degree-degree correlation can also be described in terms of the assortativity coefficient introduced by Newman, which is defined as

$$r = \frac{\langle k_1 k_2\rangle - \langle (k_1 + k_2)/2\rangle^2}{\langle (k_1^2 + k_2^2)/2\rangle - \langle (k_1 + k_2)/2\rangle^2}, \qquad (4)$$

where k_1 and k_2 are the degree of two end vertices, respectively, if an edge, and $\langle\ldots\rangle$ denotes the average over all edges. It is nothing but the Pearson correlation coefficient for the degrees of two end vertices over all edges, normalized so that $-1 \le r \le 1$. r is negative when the function $\langle k_{\mathrm{nn}}\rangle(k)$ exhibits decreasing behavior like the case of the PIN. In fact, the assortativity coefficient was introduced to characterize social networks, which have positive values of r in general. Thus vertices with higher degree tend to connect to those with lesser (more) degrees in PIN (social networks).

Many real world biocomplex networks have modular structures within them. Such modular structures are characterized in terms of the clustering coefficient. Let C_i be the local clustering coefficient of a vertex i, defined as $C_i=2e_i/k_i(k_i\text{-}1)$, where e_i is the number of edges present among the neighbors of vertex i, out of its maximum possible number $k_i(k_i\text{-}1)/2$. The clustering coefficient of a network, C, is the average of C_i over all vertices. $C(k)$ means the mean clustering coefficient over the vertices with degree k. When a network is modular and hierarchical, the clustering function follows a power law,

$$C(k) \sim k^{-\beta}, \qquad (5)$$

for large k, and C is independent of system size N [22,23].

3. Graph theoretic analysis of the yeast protein interaction network

There are a number of existing databases [24-26] or large-scale data sets [7,9,27-30] that store the information on the protein interactions in the yeast. As all biological data are subject to some errors and incompleteness, which database to use is not a trivial problem. Without having a unified one only, we have tried to access as many data as we can, including those from the four major large-scale datasets, (i) the large-scale yeast two-hybrid data by Uetz et al. [9,28] and (ii) by Ito et al. [27], as well as the curated databases, (iii) the Munich Information Center for Protein Sequences (MIPS) [24] and (iv) the database of the interacting proteins (DIP) [25] as of March 2003. We also collected data from following additional sources: (a) Two-hybrid data by Tong et al. [29], (b) Mass spectrometry protein complexes analysis data (filtered one) by Ho et al. [30]. After trimming the synonyms and other redundant entries manually, the resulting network consists of 16174 interactions (excluding self-interactions) between 5002 vertices in terms of distinct open reading frames. We denote this data as “integrated” one. Topological features of the resulting integrated network are summarized in Table I and Fig. 1, which contain the comparison with topological features from individual databases. We measure various quantities describing the structural properties of the PIN based on our dataset as follows:

(i) *Giant cluster*---Among 5002 proteins, as many as 4927 (98%) forms a giant cluster.

(ii) *Mean degree*---The mean degree $\langle k \rangle$, i.e., the average number of interaction partners per protein, is $\langle k \rangle \approx 6.44$ excluding self-interactions, which is larger than previous estimates, $\langle k \rangle \approx 2$~3 based on [9,27,31].

(iii) *Degree distribution*---It has been reported that $P_D(k)$ follows a power law, Eq. (1), with $\gamma \approx 2.4$~2.7 [32] or a power law with exponential cutoff in the form of $P_D(k) \sim (k+k_0)^{-\gamma}\exp(-k/k_c)$ with $\gamma \approx 2.45$, $k_0=1$, and $k_c \approx 20$ [31]. Based on our dataset, we found, however, that the connectivity distribution fits better to the generalized Pareto function, Eq. (2) with $\gamma \approx 3.5$ and $k_0 \approx 8.4$. That is, the PIN is scale-free. Note that the exponent $\gamma \approx 3.5$ is rather larger than previous measured values, $\gamma \approx 2.4$~2.7.

(iv) *Assortativity*---The assortativity coefficient r [33] is negative as r=-0.137, i.e., the PIN is dissortatively mixed, meaning that proteins with a small number of interaction partner are likely to connect to those with a large number of interaction partner, and *vice versa*, compared with its random counterpart whose r value is typically null.

(v) *Average of neighbor's degree*---The function $\langle k_{\mathrm{nn}} \rangle(k)$ exhibits a decreasing behavior with increasing k, a common behavior to dissortatively mixed networks. The decaying behavior is expressed roughly by another power law, Eq. (3), with $\nu \approx 0.2$~0.3, where the value ν is smaller than a previous estimated value 0.5~0.6 [34] based on the dataset by Ito et al.

(vi) *Clustering*---The clustering coefficient, C, is obtained to be C=0.131, larger than the values based on the data by Uetz et al. and by Ito et al.

(vii) *Hierarchical modularity*--- The average clustering function $C(k)$ is likely to be constant for small k, while it decreases with increasing k for large k. Such a behavior is comparable to the ones measured from other databases as shown in Fig. 1.

Putting all these together, the yeast protein interaction network is scale-free, dissortatively mixed, highly clustered, and organized in a highly modular manner. The topological characteristics from our dataset and its comparison to other ones are summarized in Table I and in Fig. 1. Such structural properties are universal for different species, so that they could be used as a test bed to find incomplete protein interactions.

4. Classification of scale-free networks

While the emergence of the scale-free behavior in complex networks is intriguing and has a number of important consequences in its own right, there may exist other hidden orders in the scale-free networks. In this section, we introduce a candidate for this, the load distribution, and show that we can classify a range of real-world and model-generated scale-free networks into two distinct classes. We argue that such classification is rooted from the distinct topological features of the *shortest pathways* in the network.

4.1. Load distribution

Let us suppose that a signal is sent from a vertex *i* to *j* ($i \to j$), along the shortest pathway between them [35]. In the information network such as the Internet, data packet is normally transmitted along the shortest pathways, however, for biological networks, it is not, even though the shortest pathways are the major flux canal. Nevertheless, here we consider the signal transport along the shortest pathways for simplicity. If there exist more than one shortest pathways, the signal would encounter one or more branching points. In this case, the signal is presumed to take one of them with equal probability, and the signal is effectively divided evenly over the branches at each branching point as it travels. Then the load $\ell_k^{i\to j}$ at a vertex *k* is defined as the amount of signals passing through that vertex *k*. Note that $\ell_k^{i\to j} = 0$ for vertices which do not fall on the shortest pathway (*i*→*j*). Also note that the contribution from the pathway (*i*→*j*), $\ell_k^{i\to j}$, may be different from that of (*j*→*i*), $\ell_k^{j\to i}$, even for undirected networks. Then we define the load ℓ_k of a vertex *k* as the accumulated sum of $\ell_k^{i\to j}$ over all pairs of senders and receivers: $\ell_k = \sum_{i,j} \ell_k^{i\to j}$. Here, we do not take into account the time delay of signal transfer at each vertex or edge, so that all signals are delivered in a unit time, regardless of the distance between any two vertices. So the load is a static variable for a given number of vertices *N*. The definition of the load is illustrated in Fig. 2. Since the packets are conserved, the total load contributed by one pair is simply related to the shortest pathway length d_{ij} between them, by $\sum_k \ell_k^{i\to j} = d_{ij} + 1$. Thus we have the sum rule for ℓ_k:

$$\sum_k \ell_k = \sum_{i,j} (d_{ij} + 1) \equiv N(N-1)(D+1) \sim N^2 D, \qquad (6)$$

where *D* is called the diameter. The quantity we defined as load is closely related to the one used in sociology called "betweenness centrality" (BC) which quantifies how much power is centralized to a person in social networks [36,37].

We focus our interest on the manner how ℓ_k are distributed. Once a SF network is generated artificially or adopted from the real world, we select an ordered pair of vertices (*i*, *j*) on the network, and identify the shortest pathway(s) between them and measure the load on each vertex along the shortest pathway using the modified version of the breath-first search algorithm introduced by Newman [37] and independently by

Brandes [38].

We have measured load ℓ_k of each vertex *k* for SF networks with various γ. It is found numerically that the load distribution $P_L(\ell)$ follows the power law [35],

$$P_L(\ell) \sim \ell^{-\delta}. \tag{7}$$

When the indices of the vertices are ordered according to the rank of the load, we have $\ell_1 \geq \cdots \geq \ell_N$. Then, the power-law behavior of the load distribution implies that

$$\frac{\ell_i}{\sum_j \ell_j} \sim \frac{1}{N^{1-\alpha}} \frac{1}{i^\alpha}, \tag{8}$$

with

$$\delta = 1 + 1/\alpha. \tag{9}$$

The relation, Eq. (8), is valid in the region [39], $\ell_{\min} < \ell < \ell_{\max}$, where

$$\ell_{\min} \sim \ell_{\max}/N^\alpha \sim \begin{cases} ND & \text{if } \alpha < 1 \\ ND/\ln N & \text{if } \alpha = 1 \\ N^{2-\alpha}D & \text{if } \alpha > 1. \end{cases} \tag{10}$$

Based on numerical measurements of load exponents for a variety of SF networks, we find that the load exponent is likely to be robust, independent of the details of network structure such as the degree exponent γ as long as γ is in the range 2<γ<3 and other details such as the mean degree, the directionality of edge, and so on [35]. Thus we may categorize the SF networks according to the load distributions of them. We found two classes, say, class I and II [40]. For the class I, the load exponent is δ ≈ 2.2(1) and for the class II, it is δ ≈ 2.0(1). We conjecture the load exponent for the class II to be exactly δ = 2 since it can be derived analytically for simple models. We will show that such different universal behaviors in the load distribution originate from different generic topological features of networks.

4.2. Real-world and artificial networks investigated

A few network examples that we find to belong to the class I with δ ≈ 2.2(1) include:

(i) The protein interaction network of the yeast *S. cerevisiae* compiled by Jeong et al. [31] (PIN1), where vertices represent proteins and the two proteins are connected if they interact.

(ii) The core of protein interaction network of the yeast *S. cerevisiae* obtained by Ito et al. (PIN2) [27].

(iii) The metabolic networks for 5 species of eukaryotes and 32 species of bacteria in Ref. [21], where vertices represent substrates and they are connected if a reaction occurs between two substrates via enzymes. The reaction normally occurs in one direction, so that the network is directed.

(iv) The Barabási-Albert (BA) model [41] when the number of incident edges of an incoming vertex $m \geq 2$.

(v) The stochastic model for the protein interaction networks introduced by Solé et al. [42].

For both (i) and (v), the degree distribution is likely to follow a generalized power-law with a cut-off. Despite this abnormal behavior in the degree distribution for finite system, the load distribution follows a pure power law with the exponent $\delta \approx 2.2(1)$. The representative load distributions for real world networks (ii), and (iii) are shown in Fig. 3(a).

The networks that we find to belong to the class II with $\delta = 2.0$ include:

(vi) The Internet at the autonomous systems (AS) level as of October, 2001 [43].

(vii) The metabolic networks for 6 species of archaea in Ref. [21].

(viii) The WWW within www.nd.edu domain [12].

(ix) The BA model with m=1 [41].

(x) The deterministic model by Jung et al. [44].

In particular, the networks (ix) and (x) are of tree structure, where the edge load distribution can be solved analytically. The load distributions for real-world networks (vi) and (viii) are shown in Fig. 3(b).

4.3. Topology of the shortest pathways

To understand the generic topological features of the networks in each class, we particularly focus on the topology of the shortest pathways between two vertices separated by a distance d. We define the *mass-distance relation* $M(d)$ as the mean number of vertices on the shortest pathways between a given pair of vertices, averaged over all pairs separated by the same distance d. If the shortest pathway topology is simple and resembles a fractal with the fractal dimension D_F, $M(d)$ would behave like $\sim d^{DF}$ for large d, while if is tree-like, one would expect $M(d) \sim d$. We find that the mass-distance relation behaves differently for each class; For the class I, $M(d)$ behaves nonlinearly (Figs. 4a-b), while for the class II, it is roughly linear (Figs. 4c-d).

For the networks belonging to the class I such as the PIN2 (iii) and the metabolic network for eukaryotes (iv), $M(d)$ exhibits a non-monotonic behavior (Figs. 4a-b), *viz.*, it exhibits a hump at $d_h \approx 10$ for (iii) or $d_h \approx 14$ for (iv). To understand why such a hump arises, we visualize the topology of the shortest pathways between a pair of vertices, taken from the metabolic network of a eukaryote organism, *Emericella nidulans* (*EN*), as a prototypical example for the class I. Fig. 5(a) shows such a graph with linear size 26 edges (d=26), where an edge between a substrate and an enzyme is taken as the unit of length. From Fig. 5(a), one can see that there exists a blob structure inside which vertices are multiply connected, while vertices outside are singly connected. The characteristic of the class I is that the blob is localized in a small region. To give a visual image of the existence of the localized blob, we show the global snapshot of the shortest pathways in the *EN* metabolic network in Fig. 6.

For the class II, the mass depends on distance linearly, $M(d)$~Ad for large d (Fig. 4c-d). Despite the linear dependence, the shortest pathway topology for the case of A>1 is more complicated than that of the simple tree structure where A≅1. Therefore, the SF networks in the class II are subdivided into two types, called the class IIa and IIb, respectively. For the class IIa, A>1 and the topology of the shortest pathways includes multiply connected vertices (Figs. 5(b) and (c)), while for the class IIb, A≅1 and the shortest pathway is almost singly connected (Fig. 5(d)). Examples in real world networks in the class IIa are the Internet at the AS level (A~4.5) and the metabolic network for archaea (A~2.0), while that in the class IIb is the WWW (A~1.0).

The WWW is an example belonging to the class IIb. For this network, the mass-distance relation exhibits $M(d)$~1.0d, suggesting that the topology of the shortest pathway is almost singly connected, which is confirmed in Fig. 5(d). When a SF network is of tree structure, one can solve the distribution of load running through each edge analytically, and obtain the load exponent to be δ=2.

4.4. Application to the metabolic networks

In biological perspectives, the power of the shortest pathway analysis and the resulting classification is exemplified by the success in the categorizing the domains of life. In Fig. 7, we show the mass-distance relations of the metabolic networks of all 43 species that we considered, grouped by the domains. Evidently, $M(d)$ for archaea behave differently from that for bacteria and eukaryotes. The eukaryotes have the class I-type metabolic networks and the archaea have the class II-type ones. The existence of the blob in eukaryotes and lack thereof in archaea implies the formation of such architecture might be driven by evolutionary pressure. One advantage of having the class I-type topology is that it is more resilient to the targeted attack on highly connected vertices [40]. It would be interesting to extend such idea to a more realistic situation for the metabolic stability.

5. Conclusion and discussion

We have studied the structural properties of the yeast protein interaction networks and the transport phenomena along the shortest pathways on biocomplex networks from the graph theoretic viewpoint. Thanks to recent development of data collection and graph analysis methods, the structural properties of the yeast protein interaction networks have been unveiled rapidly. Here we analyzed the degree distribution, the degree-degree correlation, and the clustering coefficient of the yeast interaction networks for several different datasets available [9,24,25,27,31] and also for an integrated data we constructed. The yeast PIN is found to be strongly dissortative and highly modular. We believe that such analysis could be helpful for understanding the evolution of the protein interaction networks and finding protein interactions yet undiscovered. Moreover, we investigate the transport problem along the shortest pathways on biocomplex networks such as metabolic networks. We found that the load distribution follows a power law, and its exponent is robust, insensitive to detailed structural properties. We could classify real-world networks into two classes based on this property and also on the topological features of the shortest pathways. In particular, we find the metabolic networks for archaea belongs to the different class from that for bacteria and eukaryotes. The shortest pathway structure is simple for archaea. While further theoretical understandings are needed in relation to the robustness of the load distribution, at the moment, it would be interesting to notice that the load distribution is closely related to the structure of the core part of biocomplex networks.

Acknowlegement

This work is supported by the KOSEF Grant No. R14-2002-059-01000-0 in the ABRL program.

References

1. K. Ziemelis and L. Allen, Complex systems, Nature **410**, 241 (2001) and following review articles on complex systems.
2. R. Gallagher and T. Appenzeller, Complex systems, Science **284**, 87 (1999) and following viewpoint articles on complex systems.
3. S. H. Strogatz, Exploring complex networks, Nature **410**, 268--276 (2001).
4. R. Albert and A.-L. Barabási, Statistical mechanics of complex networks, Rev. Mod. Phys. **74**, 47 (2002).
5. S. N. Dorogovtsev and J. F. F. Mendes, *Evolution of Networks* (Oxford University Press, Oxford, 2003).
6. M. E. J. Newman, The structure and function of complex networks, SIAM Rev. **45**, 167 (2003).
7. A. C. Gavin et al., Functional organization of the yeast proteome by systematic analysis of protein complexes, Nature **415**, 141--147 (2002).
8. E. M. Marcotte et al., A combined algorithm for genome-wide prediction of protein function, Nature **402**, 83--86 (1999).
9. P. Uetz, et al., A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*, Nature **403**, 623--627 (2000).
10. P. Erdős and A. Rényi, On the evolution of random graph, Publ. Math. Inst. Hung. Acad. Sci. Ser. A **5**, 17 (1960).
11. A.-L. Barabási, R. Albert, and H. Jeong, Mean-field theory of scale-free networks, Physica A **272**, 173 (1999).
12. R. Albert, H. Jeong, and A.-L. Barabási, Diameter of the World Wide Web, Nature **401**, 130--131

(1999).

13. B.A. Huberman and L.A. Adamic, Growth dynamics of the World Wide Web, Nature **401**, 131 (1999).
14. A. Broder et al., Graph structure of the World Wide Web, Computer Networks **33**, 309 (2000).
15. M. Faloutsos, P. Faloutsos, and C. Faloutsos, On power-law relationship in the Internet topology, Comput. Commun. Rev. **29**, 251 (1999).
16. R. Pastor-Satorras, A. Vázquez, and A. Vespignani, Dynamical and correlation properties of the Internet, Phys. Rev. Lett. **87**, 258701 (2001).
17. K.-I. Goh, B. Kahng, and D. Kim, Fluctuation-driven dynamics of the Internet topology, Phys. Rev. Lett. **88**, 108701 (2002).
18. S. Redner, How popular is your paper? Eur. Phys. J. B **4**, 131 (1998).
19. M. E. J. Newman, The structure of scientific collaboration, Proc. Natl. Acad. Sci. USA **98**, 404 (2001).
20. A.-L. Barabási, H. Jeong, R. Ravasz, Z. Neda, T. Vicsek, and A. Schubert, On the topology of the scientific collaboration networks, Physica A **311**, 590-614 (2002).
21. H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, Large-scale organization of metabolic networks, Nature **407**, 651 (2000).
22. E. Ravasz et al. Hierachical organization of modularity in metabolic networks, Science **297**, 1551--1555 (2002).
23. E. Ravasz and A.-L. Barabási, Hierachical organization in complex networks, Phys. Rev. E **67**, 026112 (2003).
24. H. W. Mews et al. MIPS: analysis and annotation of proteins from whole genomes, Nucl. Acids Res. **32**, D41--D44 (2004).
25. L. Salwinski et al. The Database of Interacting Proteins: 2004 update, Nucl. Acids Res. **32**, D449--D451 (2004).
26. G. D. Bader, D. Betel, and C. W. Hogue, BIND: the Biomolecular Interaction Network Database, Nucl. Acids Res. **31**, 248--250 (2003).
27. T. Ito et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome, Proc. Natl. Acad. Sci. U.S.A. **98**, 4569--4574 (2001).
28. B. Schwikowski, P. Uetz, and S. Fields, A network of protein-protein interactions in yeast, Nat. Biotechnol. **18**, 1257--1261 (2000).
29. A. H. Tong et al. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules, Science **295**, 321--324 (2002).
30. Y. Ho et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry, Nature **415**, 180--183 (2002).
31. H. Jeong et al. Lethality and centrality in protein networks, Nature **411**, 41--42 (2001).
32. A. Wagner, How the global structure of protein interaction networks evolves, Proc. R. Soc. Lond. B

270, 457--466 (2003).

33. M. E. J. Newman, Assortative mixing in networks, Phys. Rev. Lett. **89**, 208701 (2002).
34. S. Maslov and K. Sneppen, Specificity and stability in topology of protein networks, Science **296**, 910--913 (2002).
35. K.-I. Goh, B. Kahng, and D. Kim, Universal behavior of load distribution in scale-free networks, Phys. Rev. Lett. **87**, 278701 (2001).
36. L. C. Freeman, A set of measure of centrality based on betweenness, Sociometry **40**, 35 (1977).
37. M. E. J. Newman, Scientific collaboration networks II: Shortest paths, weighted networks, and centrality, Phys. Rev. E **64**, 016132 (2001).
38. U. Brandes, A faster algorithm for betweenness centrality, J. Math. Sociol. **25**, 163 (2001).
39. K.-I. Goh, B. Kahng, and D. Kim, Packet transport and load distribution in scale-free network models, Physica A **318**, 72 (2003).
40. K.-I. Goh, E. Oh, H. Jeong, B. Kahng and D. Kim, Classification of scale-free networks, Proc. Natl. Acad. Sci. USA **99**, 12583 (2002).
41. A.-L. Barabási and R. Albert, Emergence of scaling in random networks, Science **286**, 509 (1999).
42. R. Solé, R. Pastor-Satorras, E. Smith, and T. Kepler, A model of large-scale proteome evolution, Adv. Complex. Syst. **5**, 43 (2002).
43. D. Meyer, *University of Oregon Route Views Archive Project* (http://archive.routeviews.org) (2001).
44. S. Jung, S. Kim, and B. Kahng, Geometric fractal growth model for scale-free networks, Phys. Rev. E **65**, 056101 (2002).

Table I. **Topological characteristics of the Yeast PIN for various datasets:** N is the number of proteins with at least one interacting partner, $\langle k \rangle$ the mean degree, r the assortativity coefficient, C the clustering coefficient, N_1 the size of the giant cluster, and N_2 the size of the second giant cluster. Self-interactions are eliminated throughout the analysis.

	Uetz	ITO	JEONG	MIPS	DIP	Integrated
N	1331	3279	1846	1991	4713	5002
$\langle k \rangle$	2.10	2.68	2.39	2.66	6.30	6.44
r	-0.145	-0.176	-0.162	0.055	-0.136	-0.137
C	0.071	0.037	0.153	0.271	0.122	0.131
N_1	924	2839	1458	1439	4626	4927
N_2	8	6	7	11	3	3

Figures

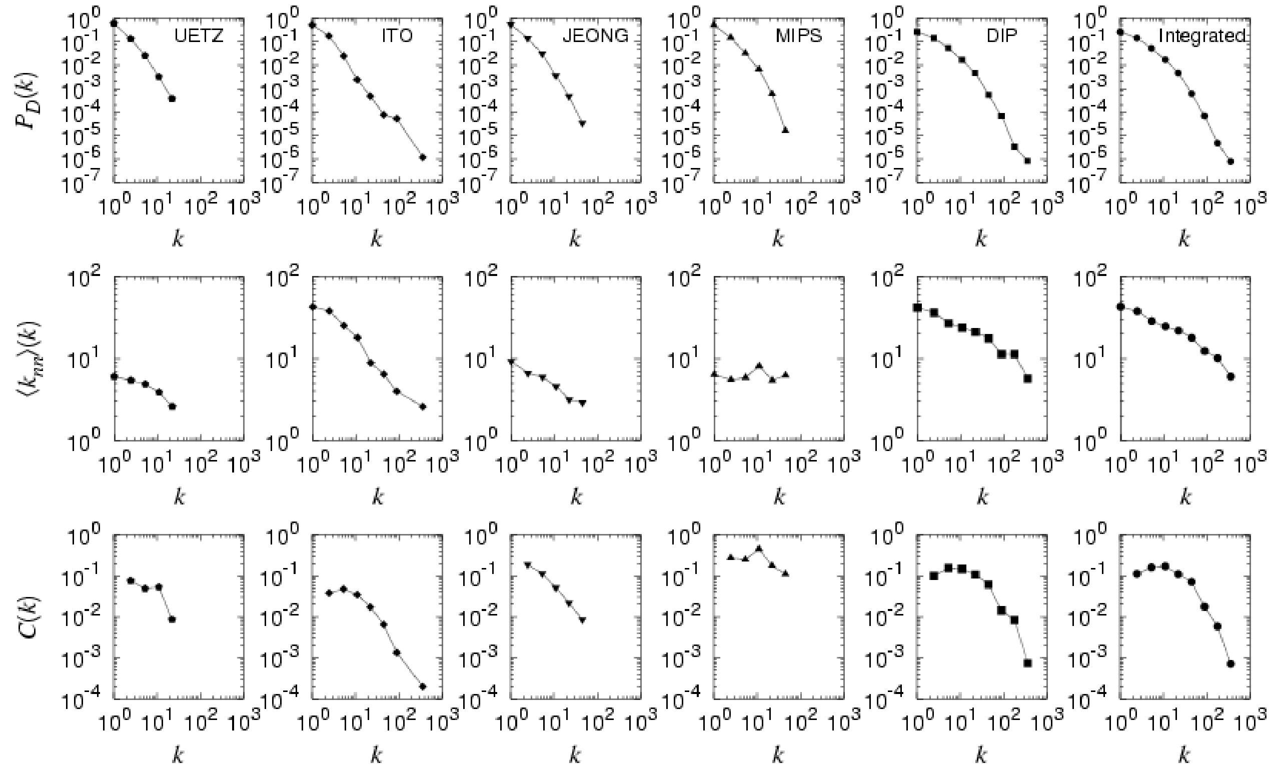


Figure 1. **Topological characteristics of the Yeast PIN for various datasets**: That of Uetz et al. [9], Ito et al. [27], Jeong et al. [31], MIPS [24], DIP [25], and the integrated one. Shown are the degree distribution $P_D(k)$, the average of the neighbor degree $\langle k_{\mathrm{nn}}\rangle(k)$, and the local clustering function $C(k)$. All data points are logarithmically binned. The ranges of abscissae and ordinates are fixed for easy comparison.

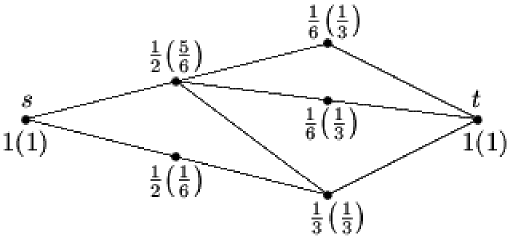

$\frac{1}{6}\left(\frac{1}{3}\right)$
$\frac{1}{2}\left(\frac{5}{6}\right)$
s
t
$1(1)$
$\frac{1}{6}\left(\frac{1}{3}\right)$
$1(1)$
$\frac{1}{2}\left(\frac{1}{6}\right)$
$\frac{1}{3}\left(\frac{1}{3}\right)$

Figure 2. **Illustration of the definition of load:** The load at each vertex due to a unit packet transfer from the vertex i to the vertex j. In this diagram, only the vertices along the shortest paths between (i, j) are shown. The quantity in parenthesis is the load due to the one from j to i.

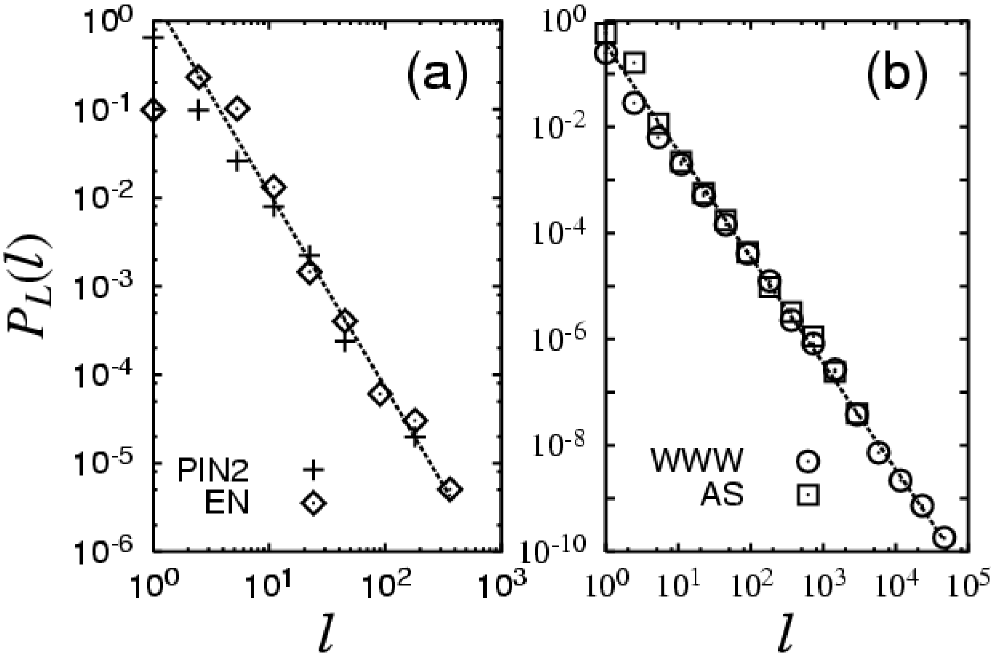


Figure 3. **Load distributions for the two classes:** (a) The PIN of the yeast (ii) and the metabolic network of a eukaryote *Emericella nidulans* (iii), belonging to the class I. (b) WWW within www.nd.edu domain (xi) and the Internet ASes (xiii), which belong to the class II.

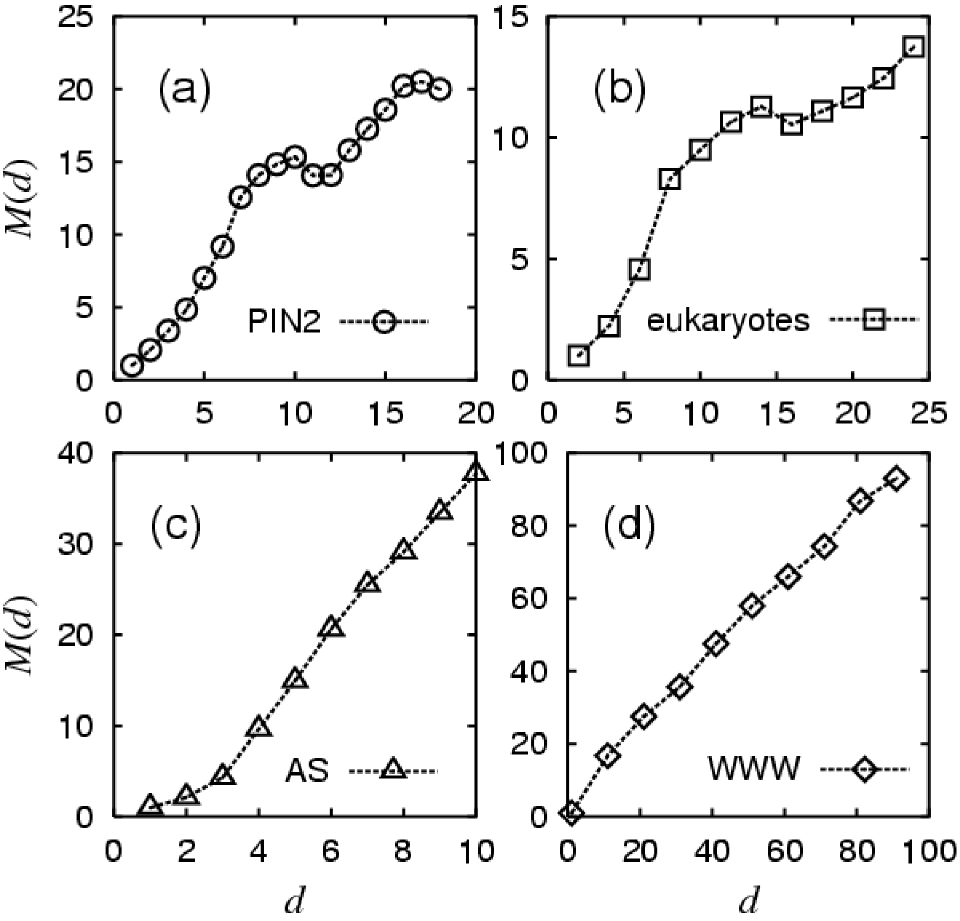
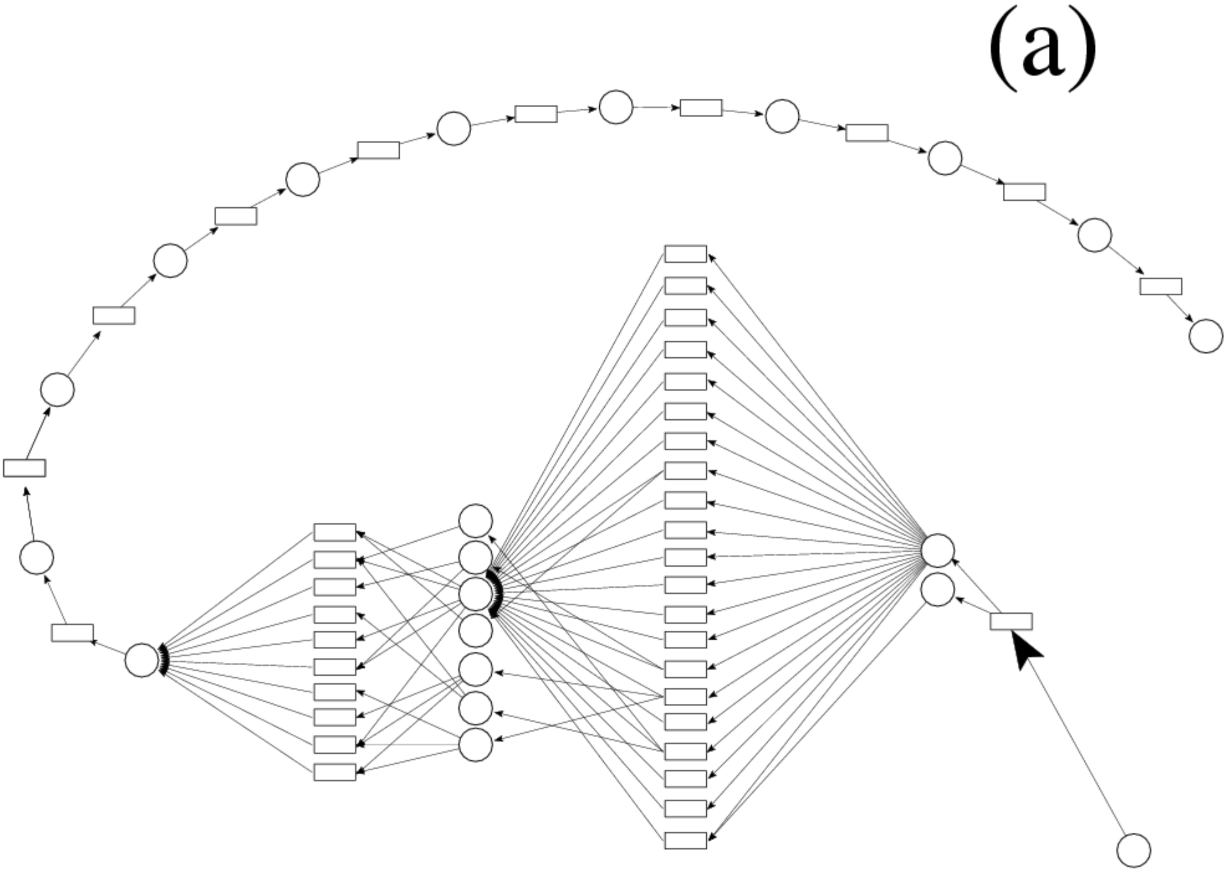
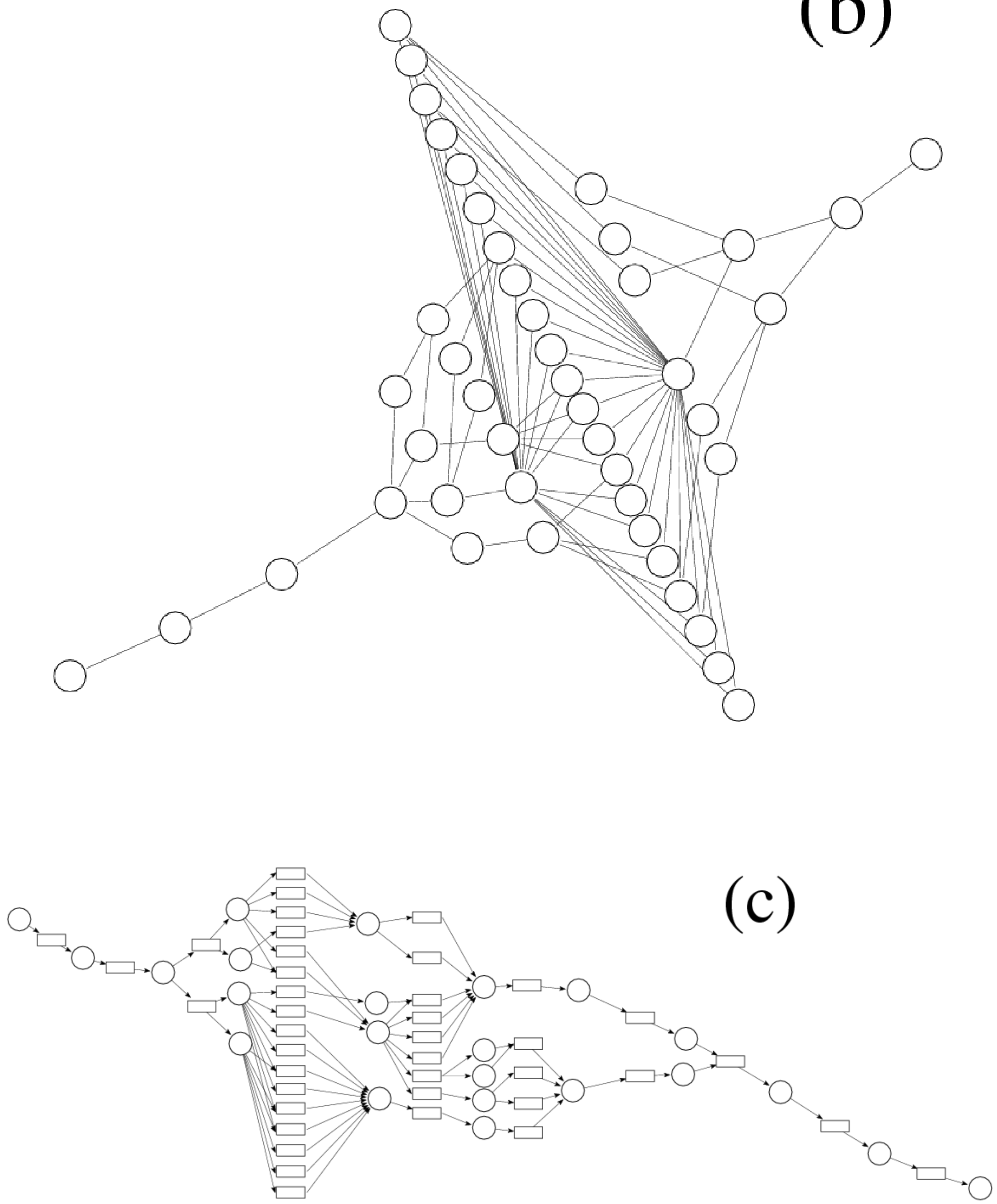


Figure 4. **Mass-distance relation for prototypical SF networks:** The yeast PIN (a), the metabolic networks of eukaryotes (b), the Internet at the AS level (c), and the WWW within nd.edu domain (d).


(a)


(b)
(c)

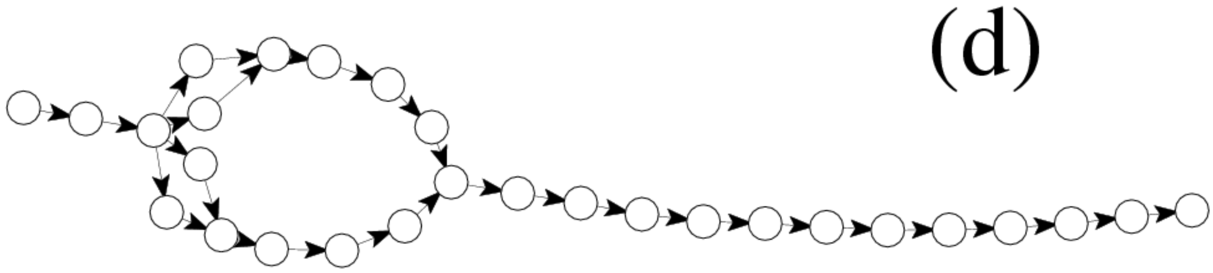


Figure 5. **Topology of the shortest pathways:** (a) The metabolic network of a eukaryote *E. nidulans* of length 26. (b) The Internet at AS level of length 10. (c) The metabolic network of an archae *Methanococcus jannaschii* of length 20. (d) WWW of www.nd.edu with length 20. In (a) and (c), circles denote substrates and rectangles denote intermediate states.

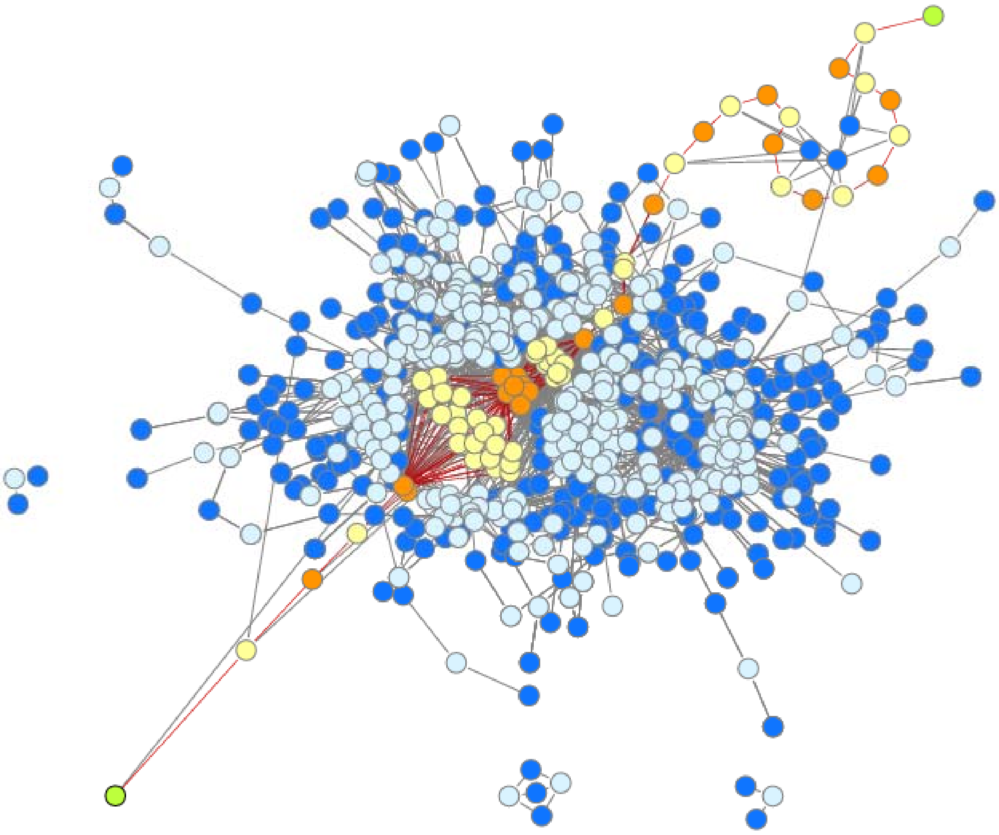


Figure 6. **Global snapshot of the metabolic network of *E. nidulans*.** The metabolites are shown in blue and the enzymes in light blue. Highlighted in orange (metabolites) and yellow (enzymes) are the shortest pathways of longest length, *d*=26, whose starting and end points are indicated in green.

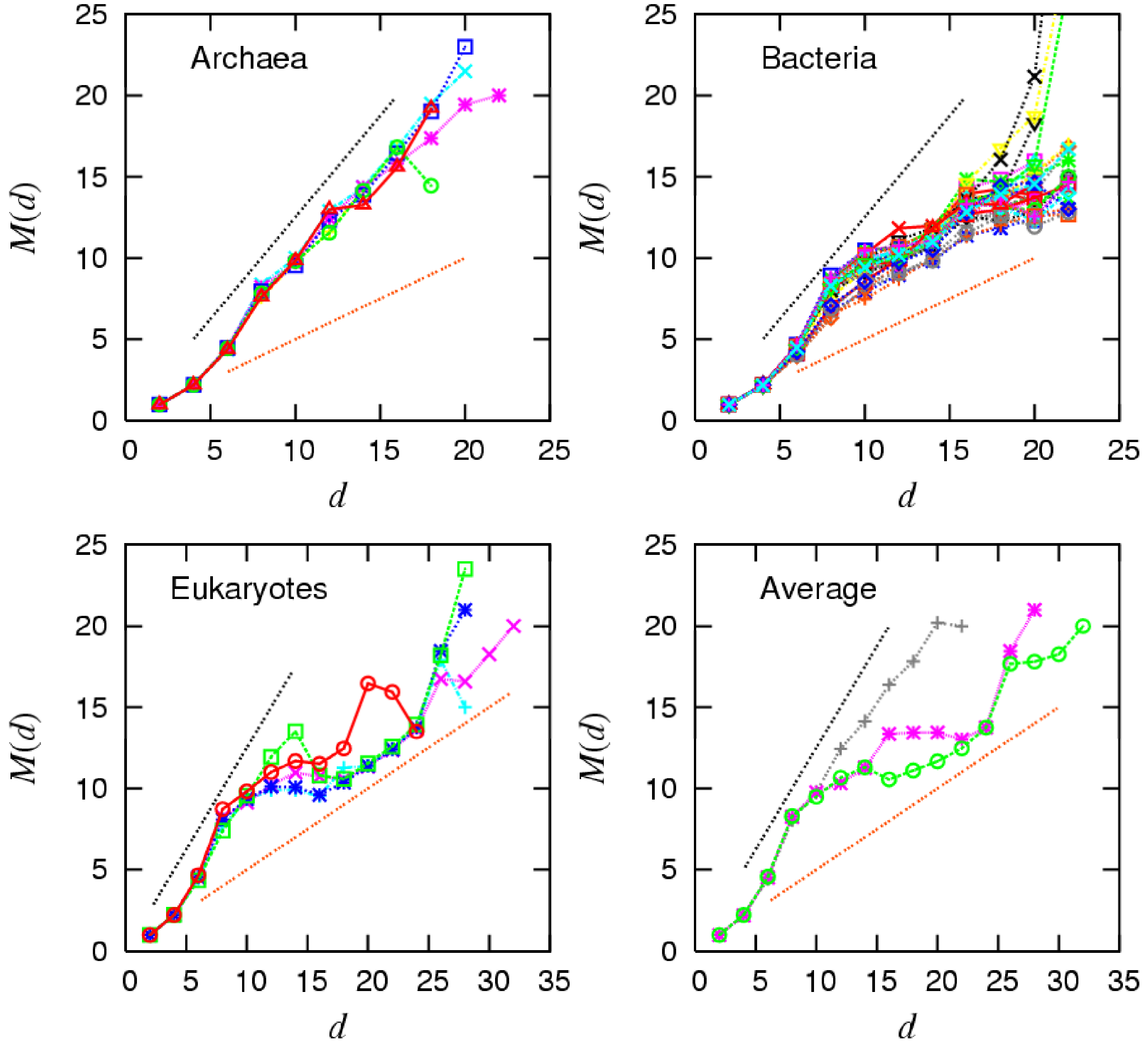


Figure 7. **Mass-distance relations for the metabolic networks of the three domains of life:** 6 archaea, 32 bacteria, and 5 eukaryotes, respectively, are plotted. In the bottom-right panel, $M(d)$ averaged over all species in each domain are compared. + stands for the data for archaea, ✳ for bacteria, and ○ for eukayotes. The straight lines have slopes 1.25 (black) and 0.5 (orange), respectively, drawn for the eye. Note that since we count only the metabolites in $M(d)$, $M(d) = 0.5d$ for singly-connected shortest pathways.