Proc. Idea-Finding Symposium
Frankfurt Institute for Advanced Studies (2003) 000–000

Idea-Finding Symposium
Frankfurt Institute for Advanced Studies
Frankfurt, Germany
April 15–17, 2003

Analyzing Molecular Networks

K. Sneppen,[1] S. Maslov[2,a] and K.A. Eriksen[1]

[1] Nordita, Blegdamsvej 17, 2100 Copenhagen Ø, Denmark

[2] Department of Physics, Brookhaven National Laboratory
Upton, New York 11973, USA

Abstract. A general scheme for analyzing patterns in complex networks is presented. A given network is compared to a distribution of randomized networks, which are constructed such that all previously recognized features of the network is preserved. In particular we find patterns that distinguish the Internet from previously analyzed molecular networks: Highly connected nodes are at the center of the Internet, but at the periphery of molecular regulatory/signaling networks.

1. Introduction

Networks have emerged as a unifying theme in complex systems research, because such systems consist of many mutually interacting components. These components are not identical as say electrons in condensed matter physics. Instead each of them has a unique identity separating it from others. The very basic question one may ask about a complex system is which other components a given component interacts with? System-wide this information can be visualized as a graph whose nodes correspond to individual components and edges to their mutual interactions. Such a network can be thought of as a backbone of the complex system along which propagate various signals and perturbations.

Living organisms provide us with a quintessential paradigm for a complex system. Therefore, it should not be surprising that in biology networks appear on many different levels: from genetic regulation and signal transduction in individual cells, to neural system of animals, and finally to food webs in ecosystems. However, complex networks are not limited to living systems: in fact they lie at the foundation of an increasing number of artificial systems. The most prominent example of this is the Internet and the World Wide Web being correspondingly the "hardware" and the "software" of the network of communications between computers.

An interesting common feature of many complex networks is an extremely broad, often scale-free, distribution of connectivities (defined as the number of immediate neighbors) of their nodes [1, 2]. While the majority of nodes in such networks are each connected to just a handful of neighbors, there exist a few hub nodes that have a disproportionately large number of interaction partners. The histogram of connectivities is an example of a low-

ISBN 963 000 000 0

level topological property of a network. While it answers the question about how many neighbors a given node has, it gives no information about the identity of those neighbors. It is clear that most of non-trivial properties of networks lie in the exact way their nodes are connected to each other. However, such connectivity patterns are rather difficult to quantify and measure. By just looking at many large complex networks one gets the impression that they are wired in a rather haphazard way. One may wonder which topological properties of a given network are indeed random, and which arose due to evolution and/or fundamental design principles and limitations? Such non-random features can then be used to identify the network and better understand the underlying complex system.

2. A Null Model for Network Analysis

In this section we propose a universal recipe for how such information can be extracted. To this end we first construct a proper null randomized model of a given network. An example of such a network is shown in Fig. 1, which indeed have a broad degree distribution as seen in Fig. 2. As was pointed out in [3], broad distributions of connectivities in most real complex networks indicate that the connectivity is an important individual characteristic of a node and as such it should be preserved in any meaningful randomization process. In addition to connectivities one may choose to preserve some other low-level topological properties of the network. Any higher level topological property, such as e.g. the pattern of

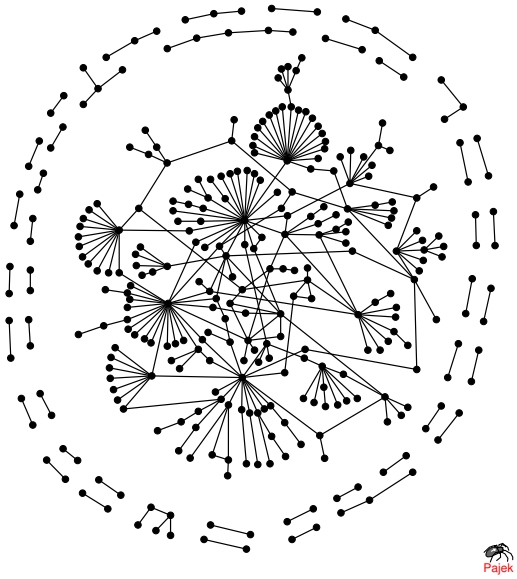

Fig. 1. Protein interaction network in the nucleus of a (Saccromyces cerevisia) yeast cell (from Maslov et al. (2002)). Each node is a protein, each link a binding interaction detected in two hybrid experiment of Ito et al.

correlations between connectivities of neighboring nodes, the number of loops of a certain type, the number and sizes of components, the diameter of the network, spectral properties of its adjacency matrix, can then be measured in the real complex network and separately in an ensemble of its randomized counterparts. Dealing with the whole ensemble allows one to put error bars on any quantity measured in the randomized network. One then concentrates only on those topological properties of the complex network that significantly deviate from the null model, and, therefore, are likely to reflect its basic design principles and/or evolutionary history.

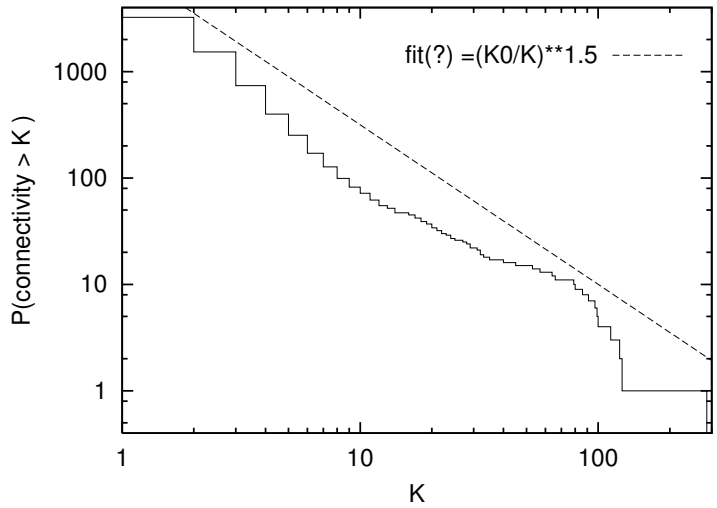

Fig. 2. Overall distribution of connectivities in the two hybrid measurement of Ito et al. (2001). We show the cumulative distributions $P(>K)$, as these allows for better judgment of potential power law fit, as shown with line. Notice that if $P(>K) \propto 1/K^{1.5}$, then $N(K) = dP/dK \propto 1/K^{2.5}$.

The *local rewiring algorithm* that randomizes a network yet strictly conserves connectivities of its nodes [4, 5] consists of repeated application of the elementary rewiring step shown and explained in detail in Fig. 3. It is easy to see that the number of neighbors of every node in the network remains unchanged after an elementary step of this randomization procedure. The directed network version of this algorithm separately conserves the number of upstream and downstream neighbors (in- and out-degrees) of every node.

Once an ensemble of randomized versions of a given complex network is generated, the abundance of any topological pattern is compared between the real network and characteristic members of this ensemble. This comparison can be quantified using two natural parameters: 1) the ratio

$$R(j) = \frac{N(j)}{\langle N_r(j) \rangle}, \tag{1}$$

where $N(j)$ is the number of times the pattern j is observed in the real network, and $\overline{N_r(j)}$

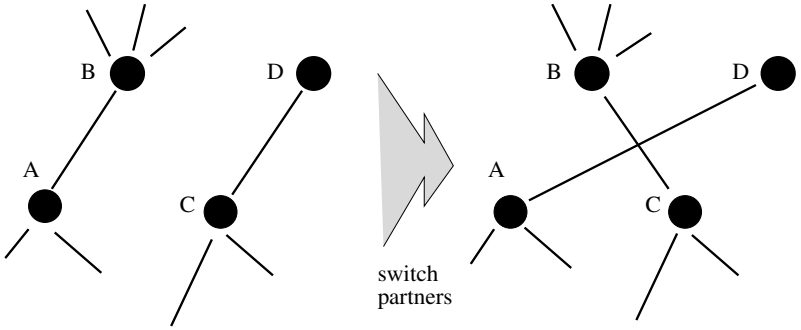

Fig. 3. One elementary step of the local rewiring algorithm. A pair of edges A–B and C–D is randomly selected. They are then rewired in such a way that A becomes connected to D, and C to B, provided that none of these edges already exist in the network, in which case the rewiring step is aborted, and a new pair of edges is selected. The last restriction prevents the appearance of multiple edges connecting the same pair of nodes.

is the average number of its occurrences in the ensemble of its random counterparts; 2) the Z-score of the deviation defined as $Z(j) = [N(j) - \overline{N_r(j)}]/\Delta N_r(j)$, where $\Delta N_r(j)$ is the standard deviation of $N_r(j)$ in the randomized ensemble. This general idea was recently applied to protein networks in yeast [4] and *E. coli* [7], and subsequently applied

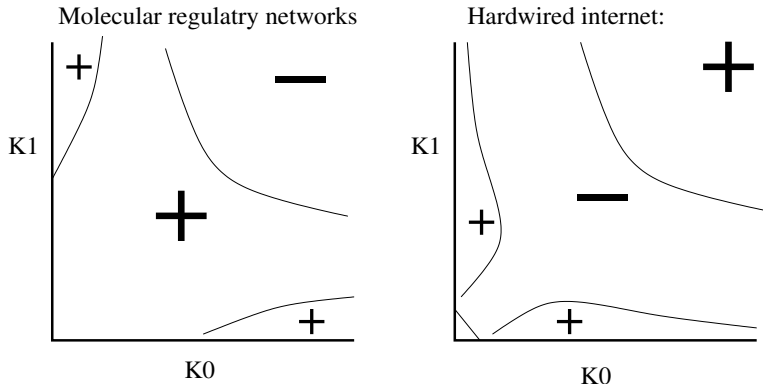

Fig. 4. Correlation profile for the hardwired Internet and for regulatory molecular networks in yeast. A + means that one see more of the corresponding connections in the real network than its randomized counterpart, a – implies a relative suppression of the connections. Notice that in protein networks, highly connected nodes tend to avoid connecting to each other. Also notice that medium connected proteins, that is proteins with 3–8 connections, preferentially associate directly to each other. These intermediately connected proteins may form the "computational core" of the protein networks.

to analysis of the hardwired Internet (the millennium snapshot of the Internet (data from January 2, 2000), when $N = 6474$ Autonomous Systems were linked by $E = 12572$ bi-directional edge), see Maslov et al. [14]. In Fig. 4 we show the qualitative difference between molecular networks analyzed in [4] and the Internet analyzed in [14]. One sees that these networks exhibit roughly opposite hierarchical features. We speculate that this reflects the limited specificity of individual proteins, in distinguishing between different exit channels. In contrast, the Internet is made of computers (autonomous systems), with huge internal specificity.

We stress that the correlation profile is by no means the only topological pattern one can investigate in a given complex network, with other examples being its spectral dimension [10], the betweenness of its edges and nodes [11, 8], feedback circuits, feed-forward loops, and other small network motifs [7]. The interplay between such higher order structures and how they are influenced by the correlation profile of the network is discussed in [14].

3. Evolution of Networks

Networks are not static. They change in time, and this evolution can in some cases be followed. For molecular networks changes in may occur by gene duplication. In fact, gene duplication followed by functional divergence of associated proteins is a major force shaping molecular networks in living organisms [16]. Thus about $1/3$ of the proteins in most organism have gene duplicates within the same organism. A pair of proteins which are generated by such a gene duplication is called paralogs. Recent availability of system-wide data for yeast S. Cerevisiae [17–20] have allowed us [15] to access the effects of gene duplication on robustness and plasticity of molecular networks.

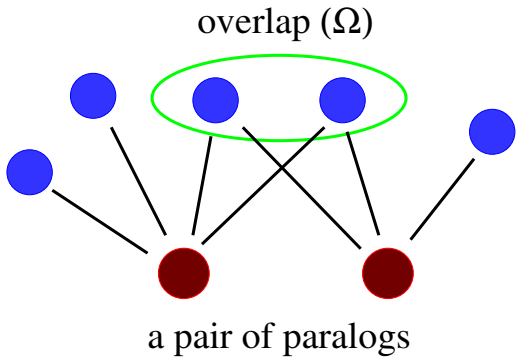

Fig. 5. Illustration of the concept of overlap in a molecular network. For a pair of paralogs the overlap Ω is defined as the number of common neighbors they have in the network. In the case of transcription network the regulatory overlap $\Omega_{\rm reg}$ counts transcription factors regulating both paralogs, while for the physical interaction network the interaction overlap $\Omega_{\rm int}$ counts their common binding partners. The pair of paralogs used in this illustration has the overlap $\Omega = 2$ out of the total of 5 distinct neighbors of the pair. That corresponds to a normalized overlap of $2/5 = 0.40$.

To this end we have measured (see Fig. 5): 1) The similarity of positions of duplicated genes in the transcription regulatory network [17] given by the number of transcription regulators they have in common (their upstream regulatory overlap); 2) The similarity of the set of binding partners [18,19] of their protein products (their downstream overlay), and 3) their ability to substitute for each other in knock-out experiments [20]. These measures reflect, correspondingly, the upstream and downstream properties of molecular networks around the duplicated genes.

In any case a time development of the properties 1–3) is estimated by averaging over a huge number of protein pairs which all have similar divergence in sequence from each other. When the proteins are newly duplicated they are nearly identical, when they are maybe 80% identical they have duplicated for maybe 100 million years ago, and when they are even weaker related to each other the time since divergence is even longer. In any case the identification of time with relative similarity is very uncertain, and in the analysis we therefore only use similarity as a measure in itself, and thus compare divergence of functions around a protein pair, with divergence of its intrinsic sequences.

We found [15] that the upstream transcriptional regulation of duplicated genes diverges fast, losing on average 4% of their common transcription factors for every 1% divergence of their amino acid sequences. In contrast, the set of physical interaction partners of their protein products changes much slower. The relative stability of downstream functions of duplicated genes, is further corroborated by their ability to substitute for each other in gene knockout experiments. We believe that the combination of the upstream plasticity and the downstream robustness is a general feature determining the evolvability of molecular networks.

Analysis of these types of network evolution data are also found in [21–23].

4. Conclusions

In summary we have proposed a general algorithm to detect characteristic topological features in a given complex network. In particular, we introduced the concept of the *correlation profile*, which allowed us to quantify differences between different complex networks even when their connectivity distributions are similar to each other. Applied to the Internet, this profile identifies hierarchical features of its structure, and helps to account for the level of clustering in this network.

Finally we briefly outlined the possibility for analyzing evolution of molecular networks, by utilizing genome as a pale-ontological record, and focusing in particular on the about $1/3$ of the genes in any organism that have duplicates within the same organism. This have lead us to the general conclusion that molecular networks use downstream robustness to facilitate fast upstream evolvability of network rewiring. This allow living organisms to use old proteins in new situations, and thereby develop new traits and possibly new biological species.

Acknowledgement

Work at Brookhaven National Laboratory was carried out under Contract No. DE-AC02-98CH10886, Division of Material Science, U.S. Department of Energy.

References

1. A.-L. Barabasi and R. Albert, *Science* **286** (1999) 509.
2. M. Faloutsos, P. Faloutsos and C. Faloutsos, *Comp. Commun. Rev.* **29** (1999) 251.
3. M.E.J. Newman, S.H. Strogatz and D.J. Watts, Random graphs with arbitrary degree distributions and their applications, *Phys. Rev. E* **64** (2001) 026118 1–17; M.E.J. Newman, to appear in *Handbook of Graphs and Networks*, S. Bornholdt and H.G. Schuster, eds. [cond-mat/0202208].
4. S. Maslov and K. Sneppen, *Science* **296** (2002) 910.
5. These algorithms first appeared in the context of random matrices in: D. Gale, *Pacific J. Math.* **7** (1957) 1073–1082; H.J. Ryser, in *Recent Advances in Matrix Theory*, Univ. of Wisconsin Press, Madison, 1964, pp. 103–124. More recently they were used in the graph-theoretical context: R. Kannan, P. Tetali and S. Vempala, *Random Structures and Algorithms* **14** (1999) 293–308.
6. This algorithm also first appeared in the mathematical literature: E.A. Bender and E.R. Canfeld, *Journal of Combinatorial Theory* **A24** (1978) 296.
7. S.S. Shen-Orr, R. Milo, S. Mangan and U. Alon, *Nature Genetics* **31** (2002) 64.
8. R. Pastor-Satorras, A. Vazquez and A. Vespignani, *Phys. Rev. Lett.* **87** (2001) 258701 1–4.
9. Website maintained by the NLANR Measurement and Network Analysis Group at http://moat.nlanr.net/ .
10. S. Bilke and C. Peterson, *Phys. Rev. E* **64** (2001) 036106.
11. M. Girvan and M.E.J. Newman, cond-mat/0112110 (2001).
12. D. Watts and S. Strogatz, *Nature* **293** (1998) 400.
13. N. Metropolis et al., *J. Chem. Phys.* **21** (1953) 1087.
14. S. Maslov, K. Sneppen and A. Zaliznyak, cond-mat/0205379 (2002), submitted to PNAS.
15. S. Maslov, K. Sneppen and K.A. Eriksen, submitted to *Genome Research* (2003).
16. S. Ohno, *Evolution by gene duplication*, Springer-Verlag, Berlin, 1970.
17. T.I. Lee et al., *Science* **298** (2002) 799–804.
18. P. Uetz et al., *Nature* **403** (2000) 623–627.
19. T. Ito et al., *Proc. Natl. Acad. Sci. USA* **98** (2002) 4569–4574.
20. G. Giaever et al., *Nature* **418** (2002) 387–391.
21. Z. Gu, D. Nicolae, H.H.-S. Lu and W.-H. Li, Rapid divergence in expression between duplicate genes inferred from microarray data, *Trends in Genetics* **18** (2002) 609–613.
22. A. Wagner, The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes, *Mol. Biol. Evol.* **18** (2001) 1283–1292.
23. Z. Gu et al., Role of duplicate genes in genetic robustness against null mutations, *Nature* **421** (2003) 63–66.