J. R. Soc. Lond. Interface (0000)
DOI: 10.1098/rsif0000.0000

Spontaneous emergence of modularity in cellular networks

Ricard V. Solé[†] and Sergi Valverde

Complex Systems Lab, ICREA-UPF, Dr Aiguader 88, 08003 Barcelona, Spain
Santa Fe Institute, 1399 Hyde Park Road, NM 87501, USA

Modularity is known to be one of the most relevant characteristics of biological systems and appears to be present at multiple scales. Given its adaptive potential, it is often assumed to be the target of selective pressures. Under such interpretation, selection would be actively favouring the formation of modular structures, which would specialize in different functions. Here we show that, within the context of cellular networks, no such a selection pressure is needed to obtain modularity. Instead, the intrinsic dynamics of network growth by duplication and diversification is able to generate it for free and explain the statistical features exhibited by small subgraphs. The implications for the evolution and evolvability of both biological and technological systems are discussed.

Keywords: ***Complex networks, Modularity, evolvability, tinkering, network biology***

1. INTRODUCTION

Biological and technological systems both exhibit a common pattern of modular organization. A modular system is formed by quasi-independent parts that are tightly integrated within themselves but also exhibit a certain degree of interdependency among them (Schlosser and Wagner, 2004). Modularity is considered a prerequisite for the adaptation of complex organisms and their evolvability (Gerhardt and Kirschner, 1997; Raff, 1996; Calabretta et al., 2000).

Modularity is particularly obvious in cellular networks (Ravasz et al., 2002), where it can be detected at the topological level. These networks include the webs of interactions among proteins, genes, enzymes and metabolites or signaling molecules. It has been argued that modularity is likely to have been selectively favoured by evolution. In that case, explaining its origins would require a functional view of biological networks (Hartwell et al., 1999).

Within the context of network theory (Dorogovtsev and Mendes, 2003; Bornholdt and Schuster, 2003; Boccaletti et al., 2006; Koonin et al. 2006) the given system is represented as a graph $\Omega = (V, E)$ composed by a set of N nodes (say proteins) $V = \{v_i\}$ and a set of links $e_{ij} \in E$ indicating if a connection exists between nodes v_i and v_j. An example is shown in figure 1: the human protein interaction network. Here we can see a few proteins having a large number of links (the hubs) surrounded by many proteins having just a few connections. This type of heterogeneous networks is very common and is characterized by a probability distribution $P(k)$ of having a node with k links which falls off as a power law with a cut-off, i. e.

$$P(k) \sim (k + k_0)^{-\gamma} e^{-k/k_c} \qquad (1)$$

[†]Author for correspondence (ricard.sole@upf.edu).

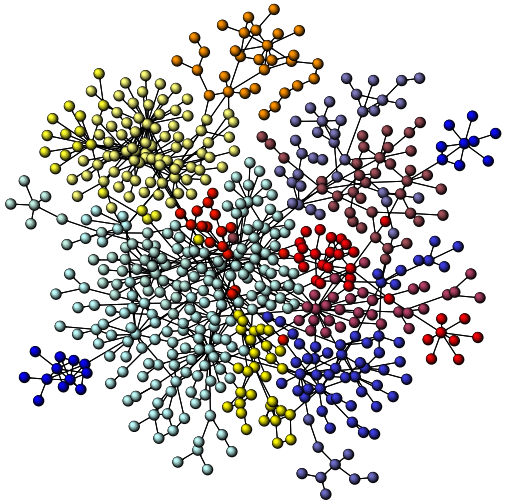

Figure 1. A modular network is here illustrated by means of the human proteome (data obtained from the DIP database: http://dip-doe-mbi.ucla.edu). Nodes are proteins and links indicate their physical (protein-protein) interaction. A standard algorithm for identifying topological modules has been used (see text) and effectively detects several well-defined groups of tightly related proteins. Modules appear indicated by different colours.

Here k_0 is a constant and $2 < \gamma < 3$ denotes the scaling exponent (typically close to $\gamma \sim 2.5$). The cut-off k_c is a characteristic degree indicating the presence of a maximum number of links. The hubs tend to have important roles (Albert et al., 2000) particularly when looking at regulatory elements such as transcription factors (Rodriguez-Caso et al., 2005) where the most connected nodes are often proto-oncogenes or tumor

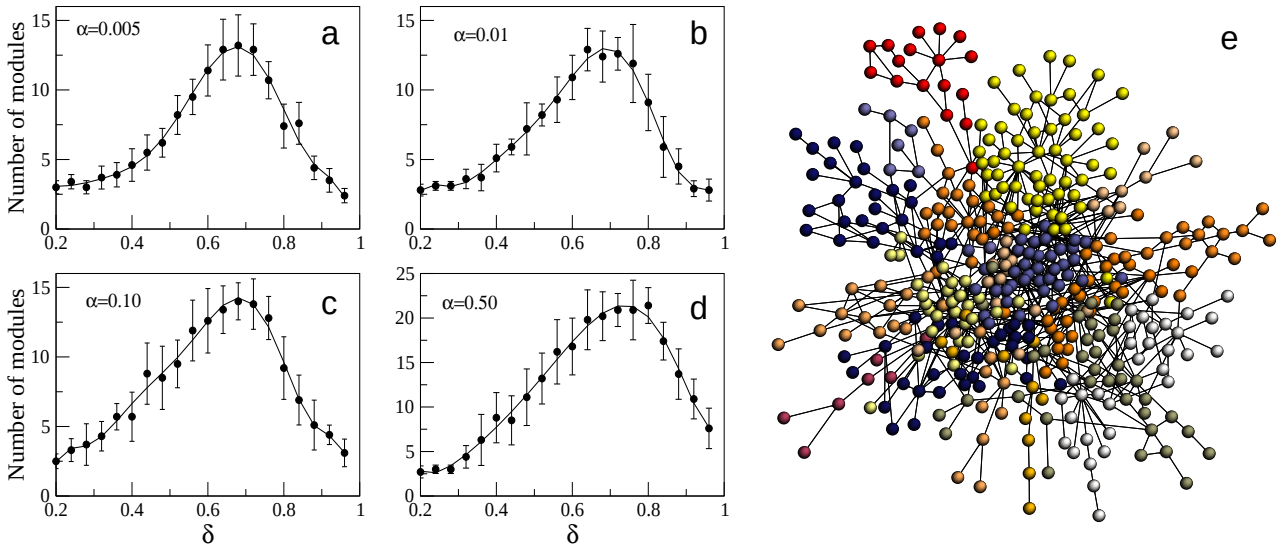

Figure 2. Modularity in tinkered model networks. Here we show (a-d) the average number of modules (μ) of networks generated using the DD model. Here μ is computed on the largest connected component for each δ and averaged over 50 replicas. We used four different α values (indicated inside each plot) and a maximal network size $N = 1000$. A well defined maximum is observed at $\delta^* \approx 0.7$ in all cases. The network shown in (e) is an example of the modular graphs obtained at δ^*.

supressor genes and their failure typically involves some proliferative disorder.

At its smallest scale, modules are defined by means of subgraphs involving three or four elements (Wolf and Arkin, 2003). These subgraphs have received considerable attention in relation with the so-called network motifs (Milo et al., 2002, 2004). Roughly speaking, motifs are patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks. The analysis of their statistical distribution reveals that each class of natural and artificial network seems to display a common patterns of motif abundances. The statistical pattern is thus interpreted as functionally meaningful. Under this view, motif abundances -as well as modularity- would be a consequence of selective forces. Is that the case? Recently, a model for the evolution of modularity and network motifs has been suggested, based on a genetic programming approach (Lipson et al., 2002; Kashtan and Alon, 2005). The model evolves electronic circuits under an environment that changes itself in a modular manner. However, the view of network substructures as resulting from pure selection or optimization has been questioned in a number of studies (Solé et al., 2002a; Banzhaf and Kuo 2004, Guimerà et al., 2004; Mazurie et al., 2005; Rice et al., 2005; Valverde and Solé, 2005; Solé and Valverde, 2006; Kuo et al., 2006) suggesting that the abundance of motifs does not necessarily reflect functional advantages.

In this paper we show that an alternative explanation for modularity exists, associated to the inevitable constraints imposed by the rules driving the growth of cellular networks. Specifically, since biological entities typically evolve by tinkering (Jacob, 1977, Solé et al., 2002a) widely reusing, combining and reconecting available parts, some patterns of network organization will be essentially inevitable. As a consequence, a both subgraph patterns and modular features would be largely a byproduct of the network generative rules.

2. GROWING NETWORKS BY DUPLICATION

The approach taken here is inspired by a physics view of biological complexity (Albert and Barabási, 2002) namely searching for generic mechanisms responsible for global patterns. The question being addressed here is how much of the modular organization of complex networks might result from just the rules of network growth by tinkering. The view here is thus topological, with no direct link to functional traits. In this context, we will restrict ourselves to a graph theoretic description of protein-protein interactions, as previously followed by several authors (Wuchty, 2001; Solé et al., 2002b, Vazquez et al., 2003, Pastor-Satorras et al., 2003; Goh et al., 2005; Colizza et al., 2005; Ispolatov et al., 2005; Foster et al., 2006). These models involve some type of duplication-divergence (DD) growth dynamics. This approach considers single-gene duplication events as the leading mechanism of genome growth. This is of course an approximation to the real complexities associated to genome growth dynamics. Although single gene duplication is considered to be the driving force behind the evolution of complex organisms (Wagner, 2001) several scales of duplication need to be considered, including whole-genome duplications (Maere et al., 2005).

After gene duplication has taken place, rapid divergence occurs and many redundant genes become silenced (i. e. become pseudogenes). Changes in wiring are associated to the emergence of novelty and new functionalities (Patthy, 1999). In our work presented here we only consider such simple approximation based on single-gene events. We will use one of the simplest DD models of protein network evolution (Vazquez et al., 2003) which involves the following set of rules, to be applied a given number of times, until N nodes are

present. Asuming that we have a graph of size n, we iterate the following rules:

1. Duplication: choose a node $v_i \in V$ at random and duplicate it, thus generating a new node v_{n+1}.
2. Link deletion: the new node shares a set of neighboring nodes $\{v_j\}$ with its predecessor. For each common pair of common links, i., e. $e_{i,j}$ and $e_{n+1,j}$ we choose one of them and delete it with probability δ. This rule thus removes (probabilistically) redundant relations among proteins.
3. Link addition: a link is added among nodes v_i and v_{n+1} with probability α. This is a small number and allows new functionalitiues to emerge by linking the twin proteins.

This model has been solved analytically and it has been shown that it exhibits a phase transition at a given deletion rate. This can be shown by constructing a dynamical equation for the average degree K_n of the simulated protein network after n nodes have been introduced. It can be shown that K_n evolves following the discrete system (Vazquez et al., 2003):

$$K_{n+1} = \frac{nK_n + 2\alpha + (1-2\delta)K_n}{n+1} \tag{2}$$

where we can see that the number of proteins n is also a time scale. Using the continuous approximation $K_{n+1} - K_n \sim dK_n/dn$ and assuming that n is large, we have a differential equation

$$\frac{dK_n}{dn} = \frac{2\alpha}{n} + \frac{1-2\delta}{n}K_n \tag{3}$$

By solving it, we obtain the time evolution of the average degree:

$$K_n = \frac{2\alpha}{2\delta - 1} + \left(K_1 - \frac{2\alpha}{2\delta - 1}\right) n^{1-2\delta} \tag{4}$$

It is easy to check that a steady degree K^* is achieved for $\delta > \delta_c = 1/2$, namely:

$$K^* = \lim_{n\to\infty} K_n = \frac{2\alpha}{2\delta - 1} \tag{5}$$

whereas for $\delta < \delta_c$ link removal is too slow and the average connectivity is high. There is thus a critical deletion rate $\delta_c = 1/2$ separating a strongly connected proteome from a sparse one (which would also include many small subgraphs). Since real protein maps are known to be rather sparse (with average connectivities around $\langle k \rangle \sim 3-5$) we should expect to find appropriate removal rates at values $\delta > \delta_c$. Actually, it has been shown (Vazquez et al. 2003) that for $\alpha = 0.1$ and $\delta = 0.7$ the model is consistent with several properties observed in the Yeast proteome (such as scale free topology, small world behavior, graph correlations and robustness against node deletion). By using appropriate measures, we will show that both modules and non-random distributions of subgraphs are an expected byproduct of network growth by duplication.

3. MODULARITY FROM TINKERING

3.1. Modules

We will first analyse the emergence of modular patterns in the previously described model. In order to provide a quantitative measure, we will use a specific algorithm[‡] of community detection (Clauset et al., 2006; see also Newman, 2006). The method considers a decomposition of the graph Ω (figure 2) into a set Γ_μ of subgraphs $C_i \in \Gamma_\mu$ defining a partition $\mathcal{C}$. Obviously, many possible C_r-partitions are possible. Using the adjacency matrix of the graph, $A = (a_{ij})$, and assuming a given partition, the fraction of edges that fall within subsets of $\mathcal{C}$ will be given by:

$$f(\mathcal{C}) = \frac{\sum_{i,j} a_{ij}\delta(C_i, C_j)}{\sum_{i,j} a_{ij}} \tag{6}$$

where $\delta(a,b) = 1$ if $a = b$ and zero otherwise. Using $m = \sum_{ij} a_{ij}/2$ we can also write:

$$f(\mathcal{C}) = \frac{1}{2m}\sum_{i,j} a_{ij}\delta(C_i, C_j) \tag{7}$$

In order to define an appropriate modularity index, the previous measure needs to be compared with the expectation from a randomly wired graph with identical number of nodes and links. Let us indicate as k_i the degree of v_i, which is obtained from the adjacency matrix as: $k_i = \sum_j a_{ij}$.

The expected probability of having a link connecting two arbitrary nodes v_i and v_j will be simply $k_i k_j/2m$ and thus we can defined modularity Q in terms of the average difference between the observed and the expected value of f, namely

$$\mathcal{Q} = \frac{1}{2m}\sum_{i,j}\left[a_{ij} - \frac{k_i k_j}{2m}\right]\delta(C_i, C_j) \tag{8}$$

which is properly normalized between zero (random network) and one (a single module is present).

The modularity of a network will be defined as the maximum $Q = \max\{\mathcal{Q}\}$ as evaluated by the search algorithm (Clauset et al., 2006). The size of the best partition μ defines the (potential) number of modules. Here we apply this measure to the largest connected component of the network generated by the algorithm described in section 2. For each δ, and a fixed α value, we generate 50 simulated networks, each one starting from a small graph of four fully-connected elements and ending once a graph with $N = 10^3$ nodes is obtained. Four different values of α have been used. The results are shown in figure 2, where a one-hump curve $\mu(\delta)$ is obtained in all four cases. A maximum is reached for $\delta^* \approx 0.7$ which is actually the deletion rate that gave best fit statistics compared with yeast proteome data (Vazquez et al., 2003).

The origins of the maximum can be understood in terms of two basic, conflicting components associated to the growth rules. At low δ values, the system is

[‡]All these approaches are heuristics. It has been shown that modularity maximization is a NP complete problem (Brandes et al., 2006) and thus there is in general no optimal partition

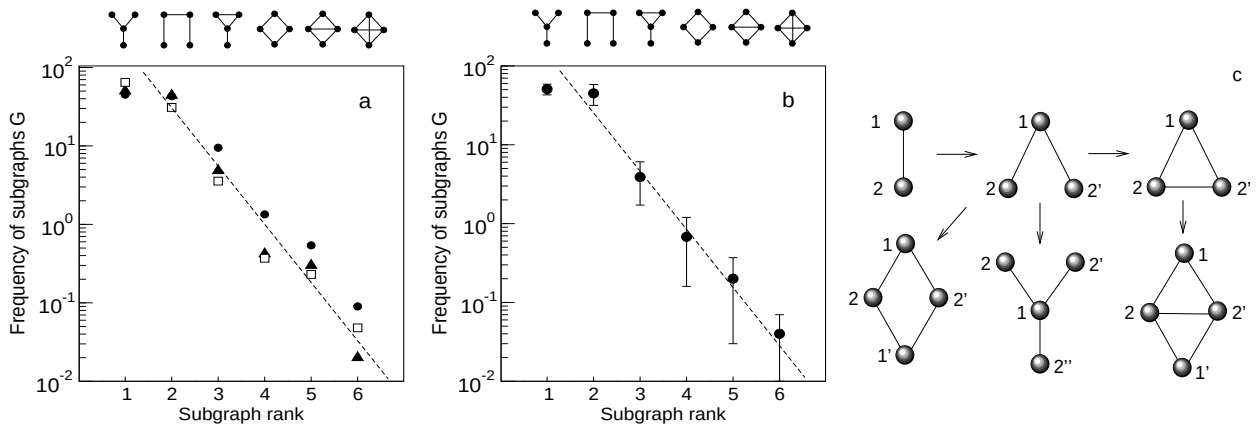

Figure 3. (a) Subgraph census for three different protein networks: human interactome (HI, filled triangles), yeast proteome (YP, filled circles) and the subset of human transcription factors (HTF, open squares). The exponential fit for $r \geq 2$ gives: $\eta_{HI} = 1.81 \pm 0.20$, $\eta_{YP} = 1.79 \pm 0.21$ and $\eta_{HTF} = 1.58 \pm 0.21$. In (b) the corresponding graph census is shown from runs of the DD model, averaged over 50 replicas. The exponent here is $\eta_{DD} = 1.71 \pm 0.20$. The data for the protein networks was gathered from the DIP database: http://dip-doe-mbi.ucla.edu and from (Rodriguez-Caso et al., 2005). In (c) some examples of the transitions between different subgrahs (using DD rules) are shown.

highly connected, being all nodes members of the giant component and having a large degree. As a consequence, we should not expect to observe a large number of modules. On the other hand, as δ increases, the network becomes more and more sparse and for $\delta > \delta_c$ it starts to get fragmented. The largest component in this domain will be formed by highly heterogeneous groups of loosely linked subgraphs. Close to the transition δ_c many elements belong to the largest connected component, but the number of modules is not large since they share many links. With increasing δ it is more likely to find groups of connected nodes that still share few links. But further increasing δ implies breaking of Ω into many small subgraphs which will display a low modularity.

3.2. Subgraph census

A second level of analysis can be performed by considering the frequency of subgraphs of a given size, also known as subgraph census (see Wasserman and Faust, 1994 and references therein). Here we have studied the census of $n = 4$ subgraphs, since it provides a reasonable number of different structures. More importantly, most examples whose functional relevance have been described in detail fall within this class (Milo et al., 2002, 2004; Valverde and Solé, 2005).

The results are shown in figure 3, where we can compare the observed patterns of subgraph abundances in real (a) and simulated (b) networks, respectively. For the real datasets, we have used the human interactome, the subset of transcription factors and the yeast proteome. The plots display the percent of subgraphs found in each network against the subgraph rank r. After a plateau, the frequency of subgraphs rapidly decays as an exponential function, i. e.

$$N(r) \sim e^{-\eta r} \tag{9}$$

with $\eta \approx 1.8$. Such pattern is also found in the distribution of subgraphs obtained from the DD model. Using the δ^* value that gives the maximum number of modules we also obtain an exponential decay, with $\eta_{DD} = 1.71 \pm 0.20$, consistently with the real datasets.

The common pattern shared by both real and simulated graphs is consistent with a rule-driven mechanism of network evolution. In this context, it is interesting to see that the different subgraphs are easily connected through DD events (figure 3c).

4. DISCUSSION

What drives the emergence of modularity in evolution? Is it a function-driven mechanism or instead the byproduct of more fundamental, dynamical rules, as it has been suggested in other contexts? (Kauffman, 1993). Although it is true that modularity is an essential feature of biological structures, our analysis suggests that it might be a byproduct of the multiplicative nature of duplication-rewiring mechanisms. Such a tinkering process (Jacob, 1976; Solé et al., 2002) inevitably leads to fluctuations in network structure due to its multiplicative nature: the rich gets richer and the graph will be organized around hubs. Moreover, the local amplification of subgraph abundances obtained from the DD process is also responsible for the decay observed in $N(r)$.

The presence of an optimal level of modularity at a given δ value is an important result of our study with potential implications to evolution by selection, at least at some levels. Selection might have been present at the level of link deletion. By removing the right amount of links, a large connected graph can be obtained, which will be both heterogeneous (and thus robust against random node deletion) and modular. In this context, although our results do not rule out the role of selection and functional adaptation in explaining modularity, they suggest that strong constraints are imposed by the rules of network growth. Thus topological patterns (including heterogeneity and modular organization) would be an emergent property of evolutionary tinkering. Evolvability might have strongly benefited from such features, since heterogeneity and

modularity immediately favour robustness and specialization, respectively. Further work should explore how these results can be extended to a more detailed level of description of network evolution.

Acknowledgments

We thank the members of the Complex Systems Lab for useful discussions. This work has been supported by a grant MCyT FIS2004-05422, by the European Union within the 6th Framework Program under contracts FP6-001907 (DELIS) the James S. McDonnell Foundation and by Santa Fe Institute.

REFERENCES

Albert, R., Jeong, H. & Barabási, A.-L. 2000. Error and attack tolerance of complex networks. *Nature* **406**, 378-382.

Albert, R. & Barabási, A.-L. 2002. Statistical mechanics of complex networks *Rev. Mod. Phys.* **74**, 47-97.

Banzhaf, W. & Kuo, P.D. 2004. Network motifs in natural and artificial transcriptional regulatory networks. *J. Biol. Phys. Chem.*, **4**,85-92.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. and Hwang, D. U. 2006. Complex networks: structure and dynamics. *Phys. Rep* **424**, 175-308.

Bornholdt, S. and Schuster, H. G. (2003) *Handbook of graphs and networks*. Wiley, Berlin.

Brandes, U., Delling, D., Gaertler, M., Goerke, R., Hoefer, M., Nikoloski, Z. & Wagner, D. 2006. Maximizing Modularity is hard. arXiv.org physics/0608255.

Calabretta, R., Nolfi, S., Parisi, D. & Wagner, G. P. 2000. Duplication of modules facilitates the evolution of functional specialization. *Artificial Life* **6**, 69-84.

Colizza, A. Flammini, A. Maritan, A. Vespignani, 2005. Characterization and modeling of. protein-protein interaction networks. *Physica A* **352**, 1-27.

Clauset, A., Newman, M. E. J. and Moore, C. (2004) *Phys. Rev.* E**70** 066111.

Dobrin, R., Beg, Q., Barabsi, A.-L. &Oltvai, Z. 2004. Aggregation of topological motifs in the Escherichia coli transcriptional regulatory network. *BMC Bioinformatics* **5**, 10.

Dorogovtsev, S. N. and Mendes, J. F. F. (2003) *Evolution of Networks: From Biological Nets to the Internet and WWW*, Oxford U. Press, New York.

Foster, D. V., Kauffman, S. A. and Socolar, E. S. 2006. Network growth models and genetic regulatory networks. *Phys. Rev. E***73**, 031912.

Gerhart, J. & Kirschner, M. 1997. *Cells, Embryos, and Evolution*. Blackwell. MA.

Goh, K.-I., Kahng, B. & Kim, D. 2005. Evolution of the Protein Interaction Network of Budding Yeast: Role of the Protein Family Compatibility Constraint. *J. Korean Phys. Soc.* **46**, 551-555.

Guimerà, R., Sales-Pardo, M. & Amaral, L. 2002. Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E***66**, 056120.

Hartwell, L.H., Hopfield, J.J., Leibler, S. & Murray, A.W. 1999 From molecular to modular cell Biology. *Nature* **402** suppl., C47-C52.

Ispolatov, I., Krapivsky, P. L. & Yuryev, A. 2005. Duplication-divergence model of protein interaction network. *Phys. Rev. E* **71**, 061911.

Jacob, F. 1977. Evolution as tinkering. Science **196**, 1161-1166.

Kashtan, N. & Alon, U. 2005. Spontaneous evolution of modularity and network motifs *Proc. Natl. Acad. Sci. USA* **102**, 13773-13778.

Kuo, P.D., Banzhaf, W. and Leier, A. 2006. Network topology and the evolution of dynamics in an artificial regulatory network model created by whole genome duplication and divergence. *Biosystems* **85**, 177-200.

Kauffman, S. A. 1993. *The origins of order*. Oxford U. Press, New York.

Kim, J., Krapivsky, P. L., Kahng, B. & Redner, s. 2002. Infinite-Order Percolation and Giant Fluctuations in a Protein Interaction Network, *Phys. Rev. E* **66**, 055101.

Klemm, K. & Bornholdt, S. 2005. Topology of biological networks and reliability of information processing. *Proc. Natl. Acad. Sci. USA* **102**, 18414-18419.

Koonin, E. V., Wolf, Y. I. & Karev, G. P (eds.) 2006. *Power Laws, Scale-Free Networks and Genome Biology*. Springer, New York.

Lipson, H., Pollack, J. B. & Suh, N. P. 2002. On the origin of modular variation. *Evolution* **56**, 1549-1556.

Maere, S. et al. 2005. Modeling gene and genome duplications in eukaryotes. *Proc. Natl. Acad. Sci. USA* **102**, 5454-5459.

Mazurie, A., Bottani, S. & Vergassola, M. 2005. An evolutionary and functional assessment of regulatory network motifs *Genome Biol* **6**, R35.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. 2002 Network motifs: Simple building blocks of complex networks. *Science* **298**, 824-827.

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M. & Alon, U. 2004. Superfamilies of designed and evolved networks. *Science* **303**, 1538-1542.

Newman, M. E. J. 2006. Modularity and community structure in networks. Proc. Natl. Acad. Sci. USA **103**, 8577-8582.

Ohno, S. 1970. *Evolution by gene duplication*. Springer, Berlin.

Pastor-Satorras, R., Smith, E. & Solé, R. V. 2003. Evolving protein interaction networks from gene duplication. *J. Theor. Biol.* **222**, 199-210.

Raff, R. A. 1996. *The Shape of Life*. Chicago U. Press. Chicago.

Ravasz, E., Somera, S.L., Mongru, D.A., Oltvai, Z.N. & Barabási, A.-L. 2002. Hierarchical organization of modularity in metabolic networks. *Science,* **297**, 1551-1555.

Reichardt, J. and Bornholdt, S. 2006. Statistical mechanics of community detection. *Phys. Rev E* **74**, 01640.

Rice, J. J., Kershenbaum, A. and Stolovitzky, G. 2005. Lasting impressions: motifs in protein-protein maps may provide footprints of evolutionary events. *Proc. Natl. Acad. Sci. USA* **102**, 3173-3174.

Rodriguez-Caso, C., Medina, M. A. & Solé, R. V. 2005. Topology, tinkering and evolution of the human transcription factor network *FEBS Journal* **272**, 6423-6434.

Schlosser, G. & Wagner, G. P. 2004 *Modularity in development and evolution*. Chicago U. Press.

Solé, R.V., Ferrer i Cancho, R., Montoya, J. M. & Valverde, S. 2002a. Selection, Tinkering, and Emergence in Complex Networks. *Complexity* **8**, 20-33.

Solé, R.V., Pastor-Satorras, R., Smith, E. & Kepler, T., S. 2002b. A model of large-scale proteome evolution. *Adv. Complex Syst.* **5**, 43-54.

Solé, R.V. & Valverde, S. 2006. Are Network Motifs The Spandrels of Cellular Complexity?. *Trends Ecol Evol.* **21**, 419-22.

Valverde, S. & Solé, R.V., 2005, Network Motifs in Computational Networks: A Case Study in Software Architecture, *Phys. Rev. E* **72**, 026107.

Vazquez, A., Flammini, A., Maritan, A. & Vespignani, A. 2003. Modeling of protein interaction networks, *ComplexUs* **1**, 38-44.

Wagner, A. 2001. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. and Evol.* **18**, 1283-1292.

Wagner, A. 2003. Does natural selection mold molecular networks? Science STKE, 41-43.

Wasserman, S. & Faust, K. 1994. *Social network analysis*. Cambridge U. Press, UK.

Wolf, D. M. & Arkin, A. P. 2003. Motifs, modules and games in bacteria. *Curr Opin Microbiol* **6**, 125-134.