Growing Biochemical Networks: Identifying the Intrinsic Properties

Hugues Bersini, Tom Lenaerts and Francisco C. Santos

IRIDIA, CP 194/6, Université Libre de Bruxelles, Brussels, Belgium [**]

Abstract. How can a new incoming biological node measure the degree of nodes already present in a network and thus decide, on the basis of this counting, to preferentially connect with the more connected ones? Although such explicit comparison and choice is quite plausible in the case of man-made networks, like Internet, leading the network to a scale-free topology, it is much harder to conceive for biochemical networks. The computer simulations presented in this article try to respect simple and, as far as possible, basic biological characteristics such as the heterogeneity of biological nodes, the existence of natural hubs, the way nodes bind by mutual affinity, the significance of type-based network as compared with instance-based one and the consequent importance of the nodes concentration to the selection of the partners of the incoming nodes.

1 Introduction

Recent years have brought a resurgent interest for evolving networks showing interesting and far from random connectivity structure like a power-law or scale-free one (Barabási and Albert) (BA in the following) [1–7]. Although the most representative of these networks have been spotted in the human and social worlds (like the Internet or epidemic networks), the fact that such a connectivity structure allows the nodes to optimally connect (these networks exhibit the small-world and good robustness properties allowing a fast and reliable communication), has encouraged an increasing number of biological researchers to believe that this scale-free connectivity should hopefully be shared by biological networks. Based on rough experimental data, some cellular, genetic and chemical networks seem, as a matter of fact, to structure their connectivity in such a particular way [8–13].

In this paper, computer experiments of growing networks will be proposed as more elementary and tractable versions of a long tradition of simulations of immune networks and chemical reaction networks popular in Artificial Life. When adopting a more biological perspective, the BA preferential attachment [1] poses serious difficulties[1]. How could a new biological node, discovering and

[**] For further information contact bersini@ulb.ac.be

[1] The BA model for building scale-free graphs is made of two main steps: (i) at each time step a new node with m links is added to the network (*growth*) ; (ii) the probability p_i that a new vertex will be connected to a node i is $p_i = k_i / \sum ki$, k_i

observing *potential partners*, preferentially decide to connect with one of these on the basis of its connectivity? Although a human being can perform such a conscious choice while involved in the construction of any technological network (such as the Internet, the Web or public transportation) or entering a sexual network, namely to express a certain preference for some nodes to attach to, this same preferential choice appears quite unlikely in natural systems growing with no human intervention. In order to remain faithful to observed biological networks, some basic principles need to be incorporated: (i) every node has a different identity based on its physical properties defining its *type*; (ii) every node connects to a selected set of nodes based on mutual *attractiveness* (*affinity*) ; (iii) certain nodes have intrinsically more ways to connect than others i.e. they are *natural hubs* and (iv), since every node represents a type, biological networks are *type-based* network instead of *instance-based*.

Take for instance chemical [15] or metabolic networks [9]. Here molecules appear only once in the network and the connections refer to the reactions in which these molecules are involved. Every node corresponds to a molecular type and has a particular concentration, and this will play a key role in the attachment mechanism: a new node should connect to other nodes based on the distribution of concentrations since this determines the probability of interaction. Furthermore, it is well-known that some molecules are more reactive than others due to their structure. Consequently, some molecules are more attractive than others. Due to this, the connectivity of a node is not only an outcome of the growing history (like in [1]), but also a result of the nodes' intrinsic features. Some of these nodes have such a high attractiveness that they become hubs (in protein networks, p53 is famous for that [16]) prior to any connection with others; they were born like hubs. Given this interpretation of molecules, chemical reaction networks represent interactions between molecular types as opposed to molecular instances. The many technological and social networks observed recently to be scale-free are all instance-based. Only one single airport, one single Web site, one single computer server or one single sexual partner corresponds to the associated node in the respective network. In contrast, some of the few biochemical networks being plotted and discussed in the literature thanks to the existence of experimental data (such as chemical reaction network or proteome map) are type-based.

The next sections will introduce the basic ingredients of the successive computer simulations and discuss expected results obtained by three experiments: a first one with no natural hub and the two successive ones with high frequency and low frequency hubs. The purpose of these experiments is the study the im-

being the degree of node i (*preferential attachment*). The more partners a node has the more likely this same node will be the partner of a new node that enters the network. The application of this BA law during the growing of the network, instead of a pure random attachment law (where a new node would randomly connect with existing nodes), will give more chance to some nodes to acquire a larger connectivity, driving the distribution to a power-law decay for the number of nodes as a function of their number of partners (with an exponent -3) rather than an exponential decay produced by a random growing [14].

plications of these features on the modelling of complex (growing) networks. In this way, we hope to arrive at a biochemical plausible model of growing networks that can explain the observed degree distributions in biochemical data.

2 Basic ingredients of the computer models

Let' s describe the basic ingredients of our tentatively more biology-like computer simulations of network structural evolution. Every node of the network is a binary string of length N so that only 2^N nodes are possible when the network has been entirely filled. Reflecting the key-lock aspect of biological or chemical binding, a node n_i will connect with another one on the basis of their hamming distance (DH). So the simplest binding rule will be for a node n_i to connect to another one n_j if:

$$DH(n_i, n_j) > t \tag{1}$$

meaning an Hamming distance (DH) superior to a given threshold t.

Eventually, each node should potentially be able to connect with other nodes. Each node i has a certain concentration that, in a very first attempt, simply changes as an effect of the on-going recruitment of nodes. When a node, randomly chosen, is recruited into the network, either it exists already and thus its concentration is just incremented by 1 or it does not exist so far and is included in the network with a concentration initially fixed to 1.

As discussed in the introduction, one *instance node* is not the same thing as one *node type* and we are interested here in growing *type-based* networks. While one instance node will be able to connect with one possible partner and only one at the moment of its recruitment, various nodes of this same type (namely the same binary string) will be able to connect to different node types. Therefore the final connectivity topology will be drawn as a function of the types of node and not of the instance nodes for which we assume one and only one partner is allowed.

The simulation goes as follows:

a) In the beginning a small number of nodes with initial concentration equal to one are recruited in the system just to kick off the growing.

b) Afterwards, at each time step, a new instance node, generated randomly, will enter the network only if it connects with an existing node type. To test this, whenever a new random instance node is proposed, a given number of trials is done with existing nodes selected probabilistically as a function of their concentration, the more concentrated the more likely to be tested. So the probabilistic preferentially of attachment is here a function of the concentration of the current nodes instead of their connectivity.

c) The test is based on the defined affinity criterion (here it is the simplest one described in Equation (1)). If following this given number of trials, the test was never successful, a new random node is proposed.

d) Once the test is succeeded and if not already existing in the network, the new node gets in and creates a new node type with initial concentration one.

e) If the incoming node n_i, although able to connect to another node n_j, already exists as a node type in the network, it is not included as a new type in the network but instead its concentration is increased by one. This explains how the concentration of the types of node composing the network can change in time, reproducing a very elementary form of dynamics.

f) If the node n_j already contains the node n_i among its partners nothing changes besides the increase in concentration of n_i. While if n_i was not yet among the partners of n_j, it will be added as such. So, on account of the *type-based* nature of our network, the computer model allows at each simulation step either to add one node type to be connected with an existing one or just one new edge between two types of node already there.

g) The simulation stops after the recruitment of an arbitrary number of types of node. The graphical results of the simulation, in the form of the now classical log-log degree distribution, is presented below (Figure 1) for N=13, t=9 and the recruitment of 4000 nodes (approximately half the number of potential types).

The resulting network's degree distribution follows an exponential decay, which is not surprising as such distribution is theoretically obtained through a random growth rule [14]. In fact, no node is favoured during the attachment of any new node since the concentration of all nodes in the network approximately remains the same. The average degree is 2.72 very much below the 378 potential nodes any node could connect with. More problematic is the very weak value of the average clustering coefficient (coefficient discussed in [3]), 0.00067, showing that two partners of a same node have very low probability to bind (as a matter of fact, most of the nodes has zero value for their local clustering coefficient). Although in strong contrast with very high values observed for instance in neural and metabolic networks, this small number essentially reflects the way the attractiveness between two nodes is defined, by means of their hamming distance. This attractiveness test makes very improbable for a same lock to have two very different keys which could bind together. In the following, the presence of natural hubs will tend to recover from this unwelcome and challenging result.

3 Similar simulations but allowing the presence of natural hubs

One key observation done recently has been that the networks observed do not show the kind of homogeneous distribution that a random growing network would produce. These observations show a bigger heterogeneity in the connectivity of the nodes as compared to the randomly growing case. Moreover certain have a huge connectivity an are referred to as hubs. For most of the authors, the presence of this heterogeneity is just an outcome of the preferential attachment rule. In biology instead, types of nodes are far from identical, and some, just by

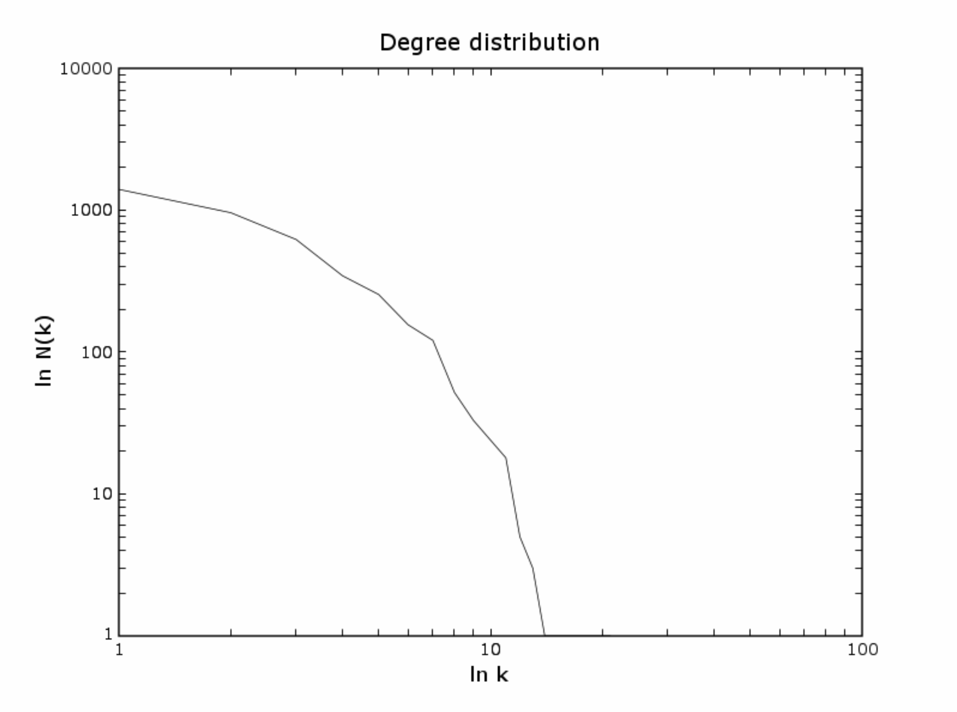

Fig. 1. The log-log degree distribution for a simulation based on the binding rule given in Equation (1). 4000 nodes are recruited in the network. The number of edges is 5457 giving an average degree of 2.72. The variance of the degree is 4.23. The clustering coefficient is 0.00067 and nearly all nodes have local clustering coefficient equal to 0. The plot takes an exponential shape typical of a random growing.

their internal shaping or constitution, are more naturally inclined to play the role of hub than others. We believe this to be the main reason for the presence of hubs detected in the real biological networks; hubs are not products but simple data of history.

We propose two refinements of the computer simulations introduced in Section 2. In the first one, the frequency of hubs will not be different than the frequency of any other node types while in the second one hubs will turn out to be much rarer (as a type not in necessarily in frequency). In the two following sections these refinements will be discussed.

3.1 Refinement 1: introducing attractiveness

In this simulation, see Figure 2, any node is a binary string of length N and possesses an additional attribute called its *attractiveness threshold* (AT_i) comprised in between [1,N-1]. The new rule of binding is that two nodes can bind if

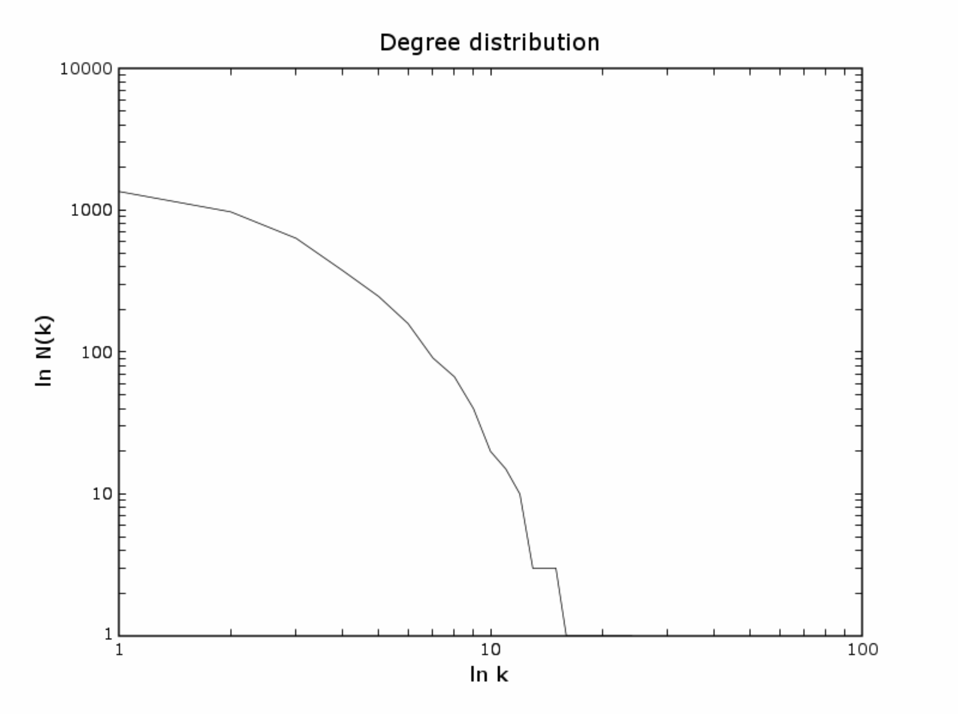

Fig. 2. The log-log degree distribution for a simulation based on the definition of nodes with *attractiveness threshold* between [0,9], node length = 10, the binding rule given in Equation (2) and high frequency hubs. 4000 nodes are recruited in the network. The number of edges is 5542 giving an average degree of 2.77. The variance of the degree is 4.51. The clustering coefficient is 0.00083. The plot takes an exponential shape typical of a random growing. The effect of natural hubs is negligible.

and only if:

$$DH(n_i, n_j) > min(AT_i, AT_j) \tag{2}$$

As a consequence, a node with a low attractiveness threshold has much more possibilities to connect than a node characterised by the same binary string but with a higher threshold. We performed a simulation with $N = 10$ [2]. A node with a small attractiveness threshold is a potential natural hub but, by the way it is defined, any type of hub is as likely as any other type. For instance, there will be as many strong hubs of length 10 (for which $AT = 1$) as any other type of length 10 and many nodes (i.e. 2^N) share a same attractiveness. Again, as shown in Figure 2, plotting the degree distribution following the recruitment

[2] The reason for this reduction in length being to have a same potential number of types as in the previous simulation ($2^{10} * 9$) despite the additional attribute differentiating types with same bit strings.

of 4000 random nodes, nothing really exciting is to be pointed out since the presence of these natural hubs does not really modify the distribution. Either hub or not, all types will tend to have a same concentration and so an equal chance to be selected. The curve is slightly shifted on the left with respect to the previous simulation (average degree = 2.77) and the presence of natural hubs slightly increases the average clustering coefficient to 0.00083. This is a situation that the last kind and most innovative simulation proposed in this paper will challenge. In this new set up, we want the hubs to be naturally much less frequent than the other types of node. The idea is that there are less biological ways to become a natural hub than an anonymous node. For a node to present one set of characteristics which makes it appealing for a certain category of potential nodes it is something, but possessing all characteristics to make it appealing for all categories of partners is a complete different story. On the whole, hubs should be harder and thus less frequent to appear in any environment, biological or technological.

3.2 Refinement 2: inverting dependence between hubness and type frequency

To achieve an inverse dependency between the *hubness* and the frequency, a node type can now have its length varying between 1 and N bits (so that at the end, the potential number of cell is equal to $2^{N+1} - 2$, we take $N = 12$ for the simulation). Nodes are defined so that the smaller one type is in size the more potential connections it can present. The new rule of binding is defined as follows. Let's call l_i the length of node n_i and make the new distance DHL between two nodes i and j of different length to be the Hamming distance between the l_i bits of n_i (here we will suppose l_i smaller than l_j) and l_i contiguous bits of n_j beginning from a position selected randomly between 1 and $(N - l_i)$. The two nodes will bind if:

$$DHL(n_i, n_j) = Min(l_i, l_j) \tag{3}$$

To some extent, the random choice of the position of the beginning bit to compare aims at reproducing the spatial orientation the two biological entities must adopt in order to bind or interact. Also, by avoiding a node to only connect to a complementary node with contiguous opposite bits departing always from the same constant position, the probability of forming triangular clusters is increased, an apparently important facet of biological networks. A node of length 1 (here *0* and *1*) becomes a very strong hub, able to connect to nearly every other nodes, while nodes of length N are just able to connect to a very limited amount of nodes. An interesting consequence of this new binding rule is that *a natural power law emerges simply by the definition of the nodes and the way they can connect.* It is indeed elementary to verify that for a given length l, each one of the 2^l nodes can approximately connect to $2^{(N-l+1)} + (l-2)$ other nodes so that the function relating the number of nodes with their potential number of partners adopts a power-law shape. In other words, we get *scale-free for free*

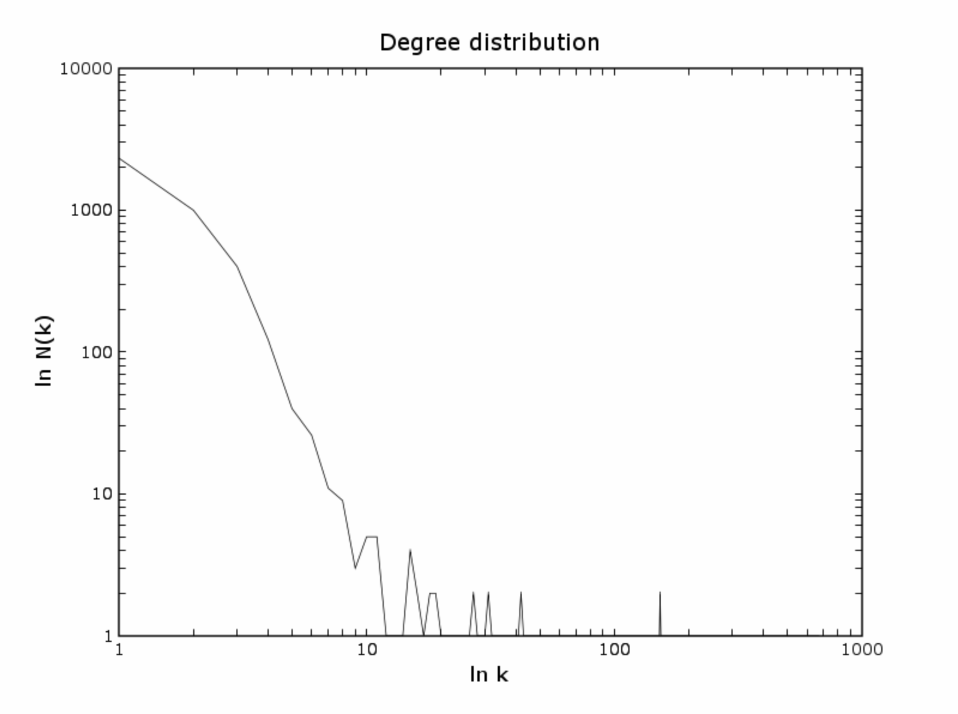

Fig. 3. The degree distribution for a simulation based on the definition of nodes with varying length between 1 and 12, the binding rule given in Equation (3), and in the presence of low frequency hubs, the smaller in length the stronger in attractiveness. 4000 nodes are recruited in the network. The number of edges is 5619 giving an average degree of 2.80. The variance of the degree is 353.07. The clustering coefficient is 0.040 with many nodes having a local clustering coefficient above 0.5. The plot takes an exponential shape typical of a random growing but now the effect of natural hubs is largely visible in the distribution tail, showing numerous picks at very high degree.

and the complete network with all types and all edges is inherently scale-free, independently of any growing mechanism. Results of the simulation (the log-log degree distribution) for $N = 12$ and again the random generation of 4000 nodes (all nodes whatever their length have equal chance to be generated) is plotted in Figure 3. Once again this distribution is shaped as an exponential decay coming out of the randomness of the recruitment, of the attachment rules and the somewhat homogeneity of the concentration, but a key difference appears in the tail of the curve: some peaks come into view, testifying for the presence of the natural hubs. Despite the maintenance of an exponential decay for small degrees, the effect of natural hubs, although less represented than poorly connected nodes, is very patent for higher degrees and makes the first, the second and the third simulations as well as their respective topologies very distinct in the final part of the distribution. Despite an average degree which is still small (2.81), strong hubs show up. For instance, the ten top hubs with their respective degree are: *1* (670), *0* (601), *11* (463), *100* (272),*00* (263), *01* (233), *110* (205), *10* (153), *000* (153), degrees much bigger than the average value. An exponential decay for low degree seems to be still compatible with the presence of hubs in the network at much higher degree (a very heavy and rugged tail), a presence that just mirrors their existence in reality, independently of any history of growing.

4 Discussion

Is preferential attachment a realistic method for growing biological networks? Based on biological observation, corresponding networks are more likely to grow in a less constrained and more random way, possibly producing the type of exponential distribution given by the various simulations presented in this paper. The question then becomes how these simulations correspond to the observations made in biological data. Since biological data is obtained using a relatively small amount of nodes, one can wonder whether the observed power-law distributions is correct. Besides hubs, high clustering and all the biological functional properties they could be responsible for are still compatible with an exponential decay at low degree, although their presence is often and (perhaps wrongly) written to be conditioned by a scale-free topology. When noticing the large influence of the concentration increase in the topology obtained by the simulations (mainly the third one), we need to restrain from definitive claims on network topology until at least a deeper attention is paid to the interplay between the concentration dynamics and the meta-dynamics, the original project of many authors adept of Alife and studying the evolution of biological networks 15 years ago [17–20].

A remarkable observation coming from our simulations with the varying nodes length is that in spite of a natural topology inherently scale-free, the simulation of the growing network somewhat counters this intrinsic topology by *insisting* in producing again an exponential distribution. This again illustrates the importance of the sampling done through the biological species in order to see how they do connect. A certain sampling could produce an exponential topology while another one, performed on the same biological system but selecting different species, would produce a scale-free version of it. This dependency on the sampling should be taken into account more carefully in, for instance, the study of protein-protein interaction map composition, which is undertaken by a lot of researchers, yet with very different outcomes [21, 22]. Moreover, a static sampling could produce a different topology than the one spontaneously adopted by the network when preserved in its natural environmental conditions.

References

1. Barabási,A.-L.; Albert, R. : Emergence of scaling in random networks. Science **286**, (1999). pp. 509-512.

2. Ferrer i Cancho, C., Janssen R., Solé, R.V. : The topology of technology graphs: small world pattern in electronic circuits. Phys. Rev.(2001). E **63**, 32767.
3. Newman, M.E.J. : The structure and function of complex networks. SIAM Review. **45** (2003). pp. 167-256
4. Pastor-Satorras, R., Vespignani, A. : Evolution and structure of the Internet: A statistical physics approach. Cambridge University Press (2004)
5. Dorogovtsev, D., Mendes, J.F.F. : Evolution of networks: From biological nets to the Internet and WWW. Oxford University Press (2003).
6. Solé, R.V., Pastor-Satorras, R., Smith, E., Kepler, T. : A model of large-scale proteome evolution. Adv. Complex Syst. **5**, (2002) pp. 43-54.
7. Strogatz, S. : Exploring Complex Networks. Nature **410** (2001) pp. 268-276.
8. Barabási, L-A., Oltvai, Z. N. : Network Biology: Understanding the cell's functional organization. Nature Reviews Genetics. **5**. (2004). pp. 101-113.
9. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., Barabási, A-L. : The large-scale organisation of metabolic networks. Nature **407**, (2000) 651–654.
10. Uetz, P. et al. : A comprehensive analysis of protein-protein interactions in Saccharomes cerevisiaie Nature **403** (2000) - pp. 623–627
11. Vazquez, A., Flamimi, A., Maritan, A. and Vespignani, A. : Modeling of Protein Interaction Networks In ComplexUs **1**, (2003) pp. 38–44.
12. Wagner, A., Fell, D.A. : The small world inside large metabolic networks. Proc. R. Soc. Lond. B. **268**. (2001). pp. 1803–1810.
13. Wagner, A. : How the global structure of protein interaction networks evolve. Proc. R. Soc. London B **270**. (2003) pp. 457–466.
14. Barabsi, A.-L., Albert, R., Jeong, H. : Mean-field theory for scale-free random networks Physica A 272, 173-187 (1999).
15. Temkin, O.N. and Zeigarnik, A.V. and Bonchev D. : Chemical reaction networks: a graph-theoretical approach, CRC Press (1996).
16. Vogelstein B., Lane D., Levine, A. J. : Surfing the p53 network Nature **408**, 307–310 (2000).
17. Bersini, H. : Immune Network and Adaptive Control. In Toward a Practice of Autonomous Systems. Proceedings of the first European Conference on Artificial Life, Varela and Bourgine, (1993) 217-225. MIT Press.
18. De Boer, R.J., Perelson, A.S. : Size and Connectivity as Emergent Properties of a Developing Immune Network Journal of Theor. Biology, **149** (1991) pp. 381–424
19. Detours, V., Bersini, H., Stewart, J., Varela, F.: Development of an Idiotypic Network in Shape Space, Journal of Theor. Biol. **170**, (1994)). pp. 401–404.
20. Varela, F., Coutinho, A. : Second Generation Immune Network, Immunology Today, **12** No 5 (1991) pp. 159-166.
21. Thomas, A., Cannings, R., Monk, N.A.M., Cannings, C. : On the structure of protein interaction networks Biochem. Soc. Trans. **31**, (1491–1496) (2003)
22. Michael P. H. Stumpf, Carsten Wiuf, Robert M. May Subnets of scale-free networks are not scale-free: Sampling properties of networks PNAS **102** 12 4221–4224 (2005)