Modeling the topology of protein interaction networks

Christian M. Schneider,¹,‡ Lucilla de Arcangelis,¹,² and Hans J. Herrmann¹,³

¹Computational Physics, IfB, ETH Zurich, Schafmattstrasse 6, CH-8093 Zurich, Switzerland
²Department of Information Engineering and CNISM, Second University of Naples, I-81031 Aversa (CE), Italy
³Departamento de Física, Universidade Federal do Ceará, 60451-970 Fortaleza, Ceará, Brazil

(Dated: August 1, 2011)

A major issue in biology is the understanding of the interactions between proteins. These interactions can be described by a network, where the proteins are modeled by nodes and the interactions by edges. The origin of these protein networks is not well understood yet. Here we present a two-step model, which generates clusters with the same topological properties as networks for protein-protein interactions, namely, the same degree distribution, cluster size distribution, clustering coefficient and shortest path length. The biological and model networks are not scale free but exhibit small world features. The model allows the fitting of different biological systems by tuning a single parameter.

PACS numbers: 64.60.aq, 89.75.Fb, 87.15.km, 87.23.Kg

I. INTRODUCTION

The study of complex networks in biology promises new fruitful insights about the functionality of genes and proteins [1][2]. Since the interactions between proteins determine their functionality, the properties and the origin of the interaction networks have attracted much attention [3][4]. They consist of protein complexes, which are connected in a large, constantly evolving, cluster [5]. The analysis of hundreds of protein complexes has established that some of the relevant structural features are the contact area, the shape of the interfaces, the complementarity of surface shapes, and the interaction-mediating forces. Although not all interactions have been discovered yet, numerous studies have been performed and many data sets are available [6][7]. One important outcome of these studies is that most protein networks show a wide range of variability in the number of nodes and edges and the average connectivity degree (Table I). They appear not to be scale-free, namely, the distribution of connectivity degrees is not a power law although it stretches over a significant number of orders of magnitude. Moreover, they do not consist of one single cluster but in addition to a large component many small clusters of interactions are also detected. These results suggest that the specific features of biological networks express different underlying mechanisms than do other networks, like social interaction networks or the internet [8][9]. In fact, it has been speculated that gene duplication is the dominant evolutionary force in shaping biological networks [10][11]. Conversely, non-biological networks are typically driven by additive growth processes [12] such as, for instance, preferential attachment [20], but many other mechanisms like rewiring [21], aging [22], or fitness [23] have been investigated. However, none of these models can reproduce the full topology of protein networks like, for instance, the emergence of isolated clusters found in real biological networks (Fig. 1).

Here we propose a different model, which reproduces many topology properties of protein interaction networks. We do not consider the details of the biochemical mechanisms at the basis of each interaction, nor classify proteins in classes as in other approaches [3][4][24]. Conversely, we follow a simple probabilistic approach.

II. THE MODEL

The procedure starts with a fully connected network of $N$ sites and $M = N(N-1)/2$ edges. The number of nodes is equal to the number of nodes of the biological network considered, $N = N_{bio}$. The evolution is performed according to the following steps:

(i) Choose at random a node $i$.
(ii) Choose at random an edge $e_{ij}$ and remove it with a

TABLE I. List of organisms from STRING 8.2 data set [26] investigated here. Columns report the number of nodes $N$, the number of edges $M$, the average degree $\langle k \rangle$, and the value of the model parameter $\alpha$ used here. Edges between pairs of proteins represent an 80% reliability of protein interaction. NF belongs to Actinobacteria, BJ to Alphaproteobacteria, and all other bacteria belong to the Gammaproteobacteria class.

| Organism                  | $N$  | $M$  | $\langle k \rangle$ | $\alpha$ |
|---------------------------|------|------|----------------------|---------|
| Nocardia farcinica (NF)   | 3582 | 12045| 6.7                  | 1.50    |
| Bradyrhizobium japonicum (BJ) | 4883 | 19261| 7.9                  | 1.00    |
| Aeromonas hydrophila (AH) | 2708 | 9050 | 6.7                  | 0.75    |
| Citrobacter koseri (CK)   | 3373 | 8212 | 4.9                  | 0.50    |
| Escherichia coli (EC)     | 3204 | 13091| 8.2                  | 0.75    |
| Pseudomonas aeruginosa (PA) | 3794 | 14252| 7.5                  | 0.75    |
| Serratia proteamaculans (SP) | 3373 | 8187 | 4.9                  | 0.75    |
| Vibrio cholerae (VC)      | 2512 | 8612 | 6.9                  | 1.00    |
| Saccharomyces Cerevisiae (SC) | 4771 | 54607| 22.9                 | 1.75    |
| Homo Sapiens (HS)         | 11102| 136930| 24.7                | 1.75    |

* schnechr@ethz.ch
probability $p_{i,j}$ related to the degree $k_j$ of the neighbor $j$ of node $i$:

$$p_{i,j} = \frac{p_j}{N_i} \quad \text{with} \quad p_j = \begin{cases} k_j^{-\alpha} & k_j > 1 \\ 0 & \text{otherwise} \end{cases}$$

(1)

and with $N_i$ the normalization $N_i = \sum_{j=1}^{k_i} p_j$. $\alpha > 0$ is the only free parameter of the model and controls the relative robustness of edges belonging to highly connected nodes with respect to edges of sites with low $k$. This rule implies that “the poor get poorer. The case $\alpha = 0$ implies that all sites have the same probability to lose edges and the process reduces to a random depletion.

(iii) Repeat this procedure for another node $i$ until the number of edges $M$ in the network equals the number of nodes $N$.

(iv) Choose at random two nodes $i$ and $j$. Add an edge between these nodes with probability

$$p_{i,j} = \frac{[N_c(i,j)]^2}{(k_i k_j)},$$

(2)

where $N_c(i,j)$ is the number of neighbors that nodes $i$ and $j$ have in common. This step supposes that, if two given nodes are able to interact with the same nodes, they have a high probability to interact with each other.

(v) Repeat this procedure for another random pair of nodes $i$ and $j$ until the number of edges $M$ in the network equals the number of edges of the modeled biological network $M_{bio}$.

These rules are based on the assumption that the evolution is controlled by two basic mechanisms: (i) preferential depletion: the lower the node degree, the lower the probability to maintain interactions [25]; (ii) similarity: the more common neighbors two nodes share, the higher is the probability to have an interaction.

The first mechanism is important for the emergence of isolated clusters and a maximal degree, while the second one is necessary to generate networks with a high clustering coefficient and assortativity. It is interesting to notice that the implementation of the depletion mechanism alone generates scale-free networks and does not reproduce the topology of protein-protein interaction networks [25].

### III. RESULTS

The biological networks are obtained from the STRING 8.2 data set [26], where a combined score of 80% is used to decide whether two proteins interact. We tested our algorithm on the ten different biological networks listed in Table I. For each organism we determine a value of the parameter $\alpha$ which provides a good fit (Table I) for the degree distribution. All results for model networks are averages over 100 independent runs for bacteria and 10 runs for the other two networks. In Fig. 1 we show an example for a biological network and the corresponding model network, with the same number of nodes and edges and $\alpha = 0.75$. Both networks have one large...
cluster with dangling ends, shown in the center of both graphs. Moreover, both networks have a large number of small clusters, placed on the border of each network. For both networks highly connected nodes are placed in the largest cluster, whereas small clusters are made of low-degree nodes. Since the topology is not a quantified differentiation property to decide whether two networks are similar, we calculate some fundamental properties characterizing the connectivity and the structure of the two networks. The model has by construction the same numbers of nodes $N$ and of edges $M$ as the biological one and therefore the average degrees per node $\langle k \rangle$ are exactly the same. To provide more information on the connectivity level of the two networks, we measure first the degree distribution. In Fig. 3 we show the degree distribution of different biological networks and their numerical counterparts. The biological networks are not scale-free and the numerical data reproduce the data very well by tuning the parameter $\alpha$. We observe that the value of the exponent $\alpha$ controls the maximum degree and the exponential cutoff of the distribution. For $\alpha = 0$ the exponential cutoff is at $k = 1$ and therefore the degree distribution a pure exponential. By increasing $\alpha$, the range of the initial regime increases and the exponential cutoff moves toward larger $k$ values. To tune the parameter, we compare the tail of the degree distribution for different $\alpha$ values and choose the one which fits best. In the procedure the smallest allowed degree is $k = 1$; the model then generates one large network and many small clusters, as in biological systems. We characterize this complex structure by evaluating the cluster size distribution. The cluster size is defined in terms of both the number of nodes, $S_N$, and the number of edges, $S_M$, belonging to the cluster. Figure 3 shows the cluster size distributions for different biological and numerical networks. Both distributions exhibit a regime consistent with a power law with an exponent $\sim -4.4$, for the size in terms of sites, and an exponent $\sim -2.7$ for the size in terms of edges. The faster decay found for the first distribution suggests that the structure is highly clustered, as will be confirmed later. Furthermore, in most cases the size of the largest connected cluster is comparable (Table II). Interestingly, numerical data for $f(S_M)$ also reproduce the fluctuations at small sizes observed in biological data. These are not the effect of statistical noise, but measure the relative weight of the population of clusters with few edges, whose patterns can be simply identified. The level of connectivity in the system is measured by the average clustering coefficient $C_{\text{max}}$ and the shortest path length $l_{\text{max}}$. The error bars are $1\%$, $2\%$, $4\%$ and $2\%$, respectively.

![FIG. 3. (Color online) Frequency $f(S)$ of finding a cluster with a given number of nodes $S_N$ for CK (circles), EC (triangles), and VC (stars) and with a given number of edges $S_M$ (inset) for AH (circles), BJ (triangles), and VC (stars), and their corresponding model networks (lines). Top and bottom data sets are shifted vertically by one decade, upward and downward.](image)

| System  | $S_M^{\text{max}}$ | $S_N^{\text{max}}$ | $C_{\text{max}}$ | $l_{\text{max}}$ | $S_M^{\text{max}}$ | $C_{\text{max}}$ | $l_{\text{max}}$ |
|---------|--------------------|--------------------|-----------------|-----------------|--------------------|-----------------|-----------------|
| AH      | 1785               | 7986               | 0.5             | 5.5             | 1996               | 8300            | 0.65            | 5.08            |
| BJ      | 2807               | 16453              | 0.5             | 6.2             | 3551               | 18043           | 0.61            | 4.85            |
| CK      | 2032               | 6609               | 0.5             | 8.1             | 2471               | 7398            | 0.59            | 5.81            |
| EC      | 2677               | 12620              | 0.5             | 6.2             | 2354               | 12250           | 0.71            | 4.92            |
| NF      | 1683               | 9435               | 0.5             | 6.8             | 2569               | 11257           | 0.49            | 4.33            |
| PA      | 2613               | 13024              | 0.5             | 7.3             | 2778               | 13263           | 0.68            | 5.09            |
| SP      | 1778               | 5911               | 0.5             | 6.4             | 2484               | 7468            | 0.54            | 5.33            |
| VC      | 1717               | 7726               | 0.5             | 5.6             | 1842               | 8028            | 0.60            | 4.74            |
| SC      | 4711               | 54570              | 0.4             | 3.7             | 3351               | 53012           | 0.81            | 3.89            |
| HS      | 10890              | 136799             | 0.4             | 3.9             | 7864               | 133576          | 0.66            | 4.15            |

![FIG. 4. (Color online) Visualization of small-world properties of biological networks. Average shortest path length $l_k$ of sites of degree $k$ versus $k$ for AH (circles), EC (triangles), and VC (stars) and clustering coefficient $C_k$ (inset) of sites of degree $k$ versus $k$ for AH (circles), BJ (triangles), and EC (stars), and their corresponding numerical networks (lines). Top and bottom data sets for $l_k$ are shifted vertically by factors of 2 and 0.5.](image)
the topological properties exhibit similar behavior. It is also possible to infer the $\alpha$ value by analysis of only 75% of the entire protein data set. From the statistical point of view our model seems to be a good candidate for modeling the topology of protein interaction networks. However, the ingredients we implement are not well established for protein interaction networks, although they are present in other biological systems. Stem cells are an example of depletion. When a stem cell specializes and becomes a particular cell (a red blood cell, a muscle cell, or even a neuron) it loses the ability to interact with cells from other types [28]. Moreover, the similarity concept can be interpreted as the establishment of interacting protein families [29][30].

IV. DISCUSSION

In conclusion, we present a statistical model, which reproduces surprisingly well many topological properties of protein interaction networks. The model is based on a twofold mechanism for evolution, namely, preferential depletion and similarity. By fitting a single parameter, we are able to generate networks that reproduce protein interaction networks for different bacteria as well as Saccharomyces cerevisae and Homo sapiens. We wish to stress that not only do the largest clusters exhibit the same connectivity properties but also the small-cluster distributions show very good agreement between biological and model data. The clustering coefficient and the average path length suggest that highly connected nodes are placed in the largest cluster and preferentially connected to nodes with high degree. The systematic analysis of the network structure for a number of biological systems indicates that protein interaction networks are not scale-free but rather exhibit small-world properties. Further research should be performed to better understand the origin of this dual mechanism in protein interaction networks.

We acknowledge financial support from the ETH Competence Center “Coping with Crises in Complex Socio-Economic Systems (CCSS) through ETH Research Grant No. CH1-01-08-2 and FUNCAP.

[1] L.H. Hartwell et al., Nature 402, C42 (1999).
[2] H. Jeong et al., Nature 407, 651 (2000).
[3] P. Uetz et al., Nature 403, 623 (2000).
[4] A.C. Gavin et al., Nature 415, 141 (2002).
[5] A.L. Barabási and Z.N. Oltvai, Nature Rev. Genet. 5, 101 (2004).
[6] T. Ideker, T. Galitski and L. Hood, Annu. Rev. Genomics Hum. Genet. 2, 343 (2001).
[7] L. Hood, Mech. Ageing Dev. 124, 9 (2003).
[8] Editorial, Nat. Cell Biol. 8, 1179 (2006).
[9] B. Schwikowski, P. Uetz and S. Fields, Nat. Biotechnology, 18, 1257 (2000).
[10] R. Friedman and A. Hughes, Genome Res. 11, 373 (2002).
[11] P.T. Spellman et al., Mol. Biol. Cell 9, 3273 (1998).
[12] J. DeRisi, V. Iyer and P.Brown, Science 278, 680 (1997).
[13] Z. Gu et al., Mol. Biol. Evol. 19, 256 (2002).
[14] http://interactome.dfci.harvard.edu/
[15] A.H.Y. Tang et al., Science 303, 808 (2004).
[16] L. Hakes, D.L. Robertson and S.G. Oliver, BMC Genomics 6, 131 (2005).
[17] R. Albert, H. Jeong and A.L. Barabási, Nature (London) 401, 130 (1999).
[18] R. Albert and A.L. Barabási, Rev. Mod. Phys. 74, 47 (2002).
[19] A. Bhan, D. Galas, and D.G. Dewey, Bioinformatics 18, 1486 (2002).
[20] A.L. Barabási and R. Albert, Science 286, 509 (1999).
[21] P.L. Krapivsky and S. Redner, Phys. Rev. E 63, 066123 (2001).
[22] L.A.N. Amaral et al., Proc. Natl. Acad. Sci. U.S.A. 97, 11 149 (2000).
[23] G. Bianconi and A.L. Barabási, Europhys. Lett 54, 436 (2001).
[24] T. Ito et al., Proc. Natl. Acad. Sci. U.S.A. 98, 4569 (2001).
[25] C.M. Schneider, L. de Arcangelis and H.J. Herrmann, Europhys. Lett. 5, 16005 (2011).
[26] http://string-db.org
[27] D.J. Watts and S.H. Strogatz, Nature (London)393, 440 (1998).
[28] F.H. Gage, Science 287, 1433 (2000).
[29] J.Park, M. Lappe and S.A. Teichmann J. Mol. Biol. 307, 929 (2001).
[30] S.A. Teichmann, J. Mol. Biol. 324, 399 (2002).
[31] V. Batagelj and A. Mrvar, Connections 21(2), 47 (1998).