BIOINFORMATICS DISCOVERY NOTE

Vol. 21 no. 8 2005, pages 1311–1315
doi:10.1093/bioinformatics/bti167

Systems biology

Topology of small-world networks of protein–protein complex structures

Antonio del Sol*, Hirotomo Fujihashi and Paul O'Meara

Bioinformatics Research Project, Research and Development Division, Fujirebio Inc., 51 Komiya-cho, Hachioji-shi, Tokyo 192-0031, Japan

Received on September 6, 2004; revised on November 11, 2004; accepted on November 17, 2004

Advance Access publication January 19, 2005

ABSTRACT

The majority of real examples of small-world networks exhibit a power law distribution of edges among the nodes, therefore not fitting into the wiring model proposed by Watts and Strogatz. However, protein structures can be modeled as small-world networks, with a distribution of the number of links decaying exponentially as in the case of this wiring model. We approach the protein–protein interaction mechanism by viewing it as a particular rewiring occurring in the system of two small-world networks represented by the monomers, where a re-arrangement of links takes place upon dimerization leaving the small-world character in the dimer network. Due to this rewiring, the most central residues at the complex interfaces tend to form clusters, which are not homogenously distributed. We show that these highly central residues are strongly correlated with the presence of hot spots of binding free energy.

Contact: ao-mesa@fujirebio.co.jp

Supplementary information: http://www.fujirebio.co.jp/support/index.php (under construction).

INTRODUCTION

Networks have become a powerful and useful tool for modeling and understanding the evolution of different complex systems (Kuramoto, 1984; Strogatz and Steward, 1993; Braiman *et al*., 1995; Gerhardt *et al*., 1990; Nowak and May, 1992). Although the connection topology is frequently assumed to be completely random or completely regular (Watts and Strogatz, 1998; Bollabas, 1985), in many cases both of these models seem to give a simplistic representation of real complex systems. Indeed, many real networks lie somewhere between the extremes of order and randomness with respect to their topological characteristics. This is the case of the so-called small-world network, where any pair of vertices can be connected through just a few links. The topology of these kinds of networks are characterized by large values of the clustering coefficient (as for regular graphs), defined as the average over all vertices of the fraction of the number of connected pairs of neighbors for each vertex, and small values of the characteristic path length (as for random graphs), defined as the average minimal distance between all pairs of vertices in the graph.

The representation of protein structures as small-world networks has recently become an interesting approach to study a variety of problems associated to protein function and structure, such as the identification of key residues involved in the protein folding mechanism (Vendruscolo *et al*., 2002) and the correlation between the topological properties of protein conformations and their kinetic ability to fold (Dokholyan *et al*., 2002; Greene and Higman, 2003), or the identification of functional sites in protein structures (Shemesh *et al*., 2004) among other examples.

An interesting application of small-world networks would be the representation of protein–protein complexes as such networks, in order to elucidate different structural characteristics associated with the presence of residues that contribute the most to the binding free energy (hot spots), which are unevenly distributed at the binding interface (Bogan and Thorn, 1998). Although different approaches involving sequence and structural information or energetic calculations have been proposed to study and predict hot spots of binding free energy (Kortemme and Baker, 2002; Sheinerman and Honig, 2002; Verkhivker *et al*., 2002; Ma *et al*., 2003; Brinda *et al*., 2002), the small-world representation of protein–protein complexes could give another complementary view on this problem.

Here, we show that the protein–protein interaction mechanism can be viewed as a specific rewiring occurring in the system of two small-world networks represented by the monomers, where a rearrangement of links takes place upon dimerization leaving the small-world character in the dimer network. Due to this specific rewiring, a rearrangement of residue centrality occurs, leading to the appearance of a significant percentage of central residues at the protein–protein interface. The analysis of 18 protein complexes with experimentally annotated hot spots of binding free energy shows that the most central residues at the protein–protein interface, responsible for the small-world character, are strongly correlated with the presence of hot spots.

SYSTEMS AND METHODS

Datasets

A dataset of 42 dimer complexes, which each contained at least one monomeric structure was obtained by searching the protein data bank (PDB) (http://www.rcsb.org/pdb/) (Berman *et al*., 2000) and the structural classification of proteins (SCOP) database (http://www.scop.berkeley.edu/) (Murzin *et al*., 1995). The non-complexed structures were chosen if they had an identical sequence to their bound form with no insertions and deletions. If any of the complexes contained more than two structures in the unbound form the most recently solved structures were used. As a result, a dataset of 58 monomers was compiled.

*To whom correspondence should be addressed.

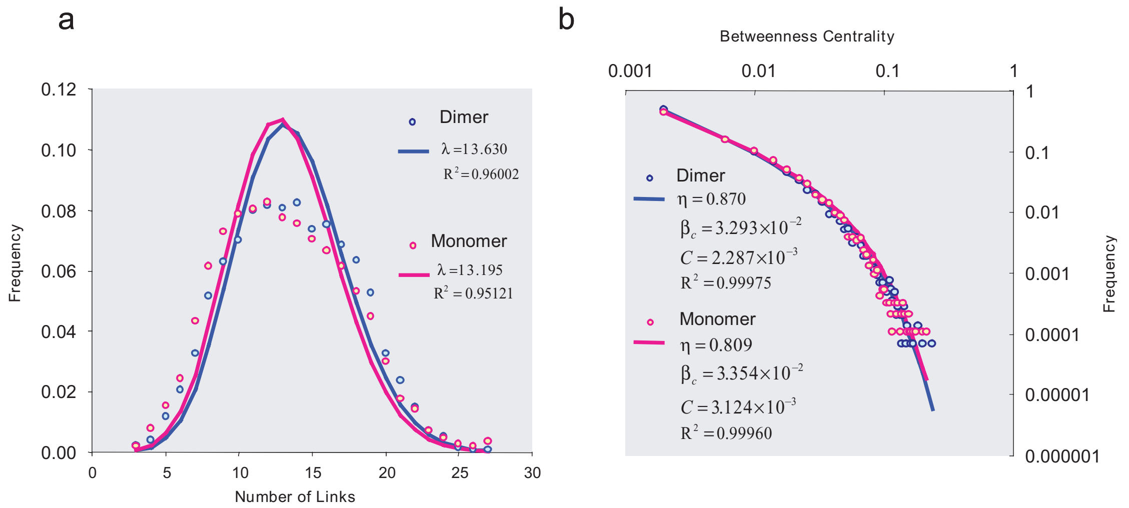

Fig. 1. Frequency distributions of the residue number of links and betweenness centrality averaged over both sets of monomers and dimers. **(a)** Bell-shaped Poisson frequency distribution of the residue number of links averaged in both the monomers (shown with the pink dots) and dimers (shown with the blue dots). The discrete Poisson fit $P(x) = \lambda^x e^{-\lambda}/x!$ is illustrated with the pink and blue lines for the monomers and dimers respectively. The average residue number of links λ and the correlation coefficients squared R^2 are shown in the graph. **(b)** Frequency distribution of betweenness centrality averaged over both sets of monomers (shown with the pink dots) and dimers (shown with the blue dots). The frequency distributions follow a power law with an exponential cut-off $P(\beta) = C\beta^{-\eta} \exp(-\beta/\beta_c)$ which is illustrated in the graph with the pink and blue lines for the monomers and dimers respectively. The data has been graphed using a logarithmic scale with the power law-scaling exponent η, exponential cut-off β_c, constant C and the correlation coefficients squared R^2 for both datasets shown in the graph. There was no statistically significant difference between the monomer and dimer frequency distributions in both (A) and (B).

A set of 18 protein complexes with experimental information on hot spot residues was obtained by searching the Alanine Scanning Energetics database (ASEdb) (http://www.140.247.111.161/hotspot/index.php) (Thorn and Bogan, 2001). Experimentally measured hot spots of binding free energy were defined as residues with a change in binding free energy greater than or equal to 1.0 Kcal/mol. Some additional data were used from previous studies in phenylalanine substitutions (Mainfroid *et al.*, 1996).

The conservation of residues in the protein complexes was analyzed based on multiple sequence alignments generated by ClustalW (Thompson *et al.*, 1994), using homologous protein sequences obtained from the Swissprot database (http://www.us.expasy.org/sprot/) (Boeckmann *et al.*, 2003).

The accessible surface areas (ASAs) of the protein complexes were determined using the DSSP program (Kabsch and Sander, 1983). Experimental enrichment of hot spot information was obtained from the literature (Bogan and Thorn, 1998).

The protein graphs

The protein structures are modeled as networks with amino acid residues being the vertices and all atom contacts between them the edges. Atom contacts are defined when the distance between at least one atom of residue i is at a distance $\leq$5.0 Å from an atom of residue j (Greene and Higman, 2003).

The characteristic path length L is defined as the average minimal distance between all pairs of vertices in the graph, calculated by:

$$L = \frac{1}{N_{\mathrm{p}}} \sum_{j>i} l_{ij},$$

where N_{p} represents the number of pairs of vertices of the graph, and l_{ij} is the minimal path between vertices i and j (Vendruscolo *et al.*, 2002).

The clustering coefficient C is defined as the average over all vertices of the fraction of the number of connected pairs of neighbors for each vertex, calculated by:

$$C = \frac{1}{N_{\mathrm{v}}} \sum_{i} \frac{n_i}{N_i(N_i - 1)/2},$$

where N_{v} is the number of vertices, N_i is the number of neighbors of the vertex i, and n_i is the actual number of edges between the neighbors of i (Vendruscolo *et al.*, 2002).

Statistical analysis

The frequency distributions of the residue number of links and betweenness centrality averaged over both sets of monomers and dimers were plotted and analyzed using Systat statistical software packages. The Kolmogorov–Smirnov test was used to test the statistically significant difference between the monomer and dimer frequency distributions.

Our analysis was carried out on a PC Linux cluster with 40 nodes (dual 3.02 GHz Xeon), and on a Windows PC (3.0 GHz Pentium IV).

DISCUSSION

We start by modeling protein structures as networks (see Systems and Methods). We base our analysis on a representative set of 42 biologically diverse protein complexes (with one or both of their unbound structures available), and find, in agreement with previous studies (Vendruscolo *et al.*, 2002; Dokholyan *et al.*, 2002; Greene and Higman, 2003), that both the dimer and monomer structures exhibit small-world character in accordance with their values of clustering coefficients and characteristic path lengths, in comparison with random and regular graphs with the same number of vertices and average number of neighbors (see Supplementary material). Figure 1a illustrates the frequency distribution of the residue number of links N averaged in both sets of monomers and dimers, indicating that both distributions are Poisson-like, where $P(x) = \lambda^x e^{-\lambda}/x!$ (with the average residue number of links λ), with no statistically significant difference between them. The concept of betweenness centrality used in sociology (Freeman, 1977), defined for each vertex k as the number of pairs of vertices with the shortest path among them passing through k normalized by the total number of pairs of vertices, is a good

Table 1. Statistically significant high betweenness (*z-score* ≥3.0) residues obtained from the 18 complexes analyzed, and their correlation to hot spots of binding free energy

Protein complex	PDB code and chain identifier	Statistically significant high betweenness residues (*z-score* ≥ 3)	Clusters (ratio > 0.8)
Hormone/receptor	1a22AB	18A,178A,365B	[**18A**] [175A,**178A**,**365B**,369B]
Enzyme/inhibitor	1a4yAB	33A,63A,150A,27B,31B,41B,89B,93B	[**33A**,**63A**,**31B**] [**150A**,**27B**,**93B**] [263A,**93B**] [318A,375A,**89B**] [434A,**41B**] [**27B**,**31B**]
Enzyme/inhibitor	1brsAD	27A,73A,38D,39D	[**27A**,**73A**,**38D**,**39D**]
Immune system protein	1bxiAB	30A,55A	[**30A**,33A,34A,37A] [50A,51A,54A,**55A**,56A]
Enzyme/inhibitor	1cbwCD	15D,17D	[**15D**,**17D**]
Immune system protein, receptor	1cdcBA	29A,31A,32A,33A,16B,31B,32B	[**29A**,**31A**,81A,29B,**31B**] [**31A**,**32A**,**33A**,38A,81A] [**29A**,**31B**,**32B**,38B] [**16B**,**32B**]
Enzyme/inhibitor	1dfjEI	146I	[**146I**,202I]
Antibody/antigen	1fccAC	28C	[27C,**28C**,31C,43C]
Antibody/antigen	1fvcAB	36A,38A,89A,37B,39B	[**36A**,**89A**,91A,105B] [**38A**,**37B**,**39B**,95B]
Antibody/antigen	1gc1CG	29C,43C,46C	[**29C**,81C] [**29C**,85C] [**43C**,44C,59C] [44C,**46C**]
Cytokine	1il8AB	25A,25B	[**25A**,27A,**25B**,27B]
Toxin/receptor	1jckCD	26D,60D,176D	[55C,20D,23D,**176D**] [23D,**26D**,90D,210D] [**26D**,**60D**,90D,210D]
Hydrolase	1pp2RL	31L,31R	[5L,9L,**31R**] [**31L**,5R,9R]
Isomerase	1ypiAB	12A,64A,77A,82A,98A,12B,46B,77B,98B	[**12A**,**64A**,**77A**,**98A**,**77B**,**98B**] [**82A**,**12B**] [**82A**,**46B**]
Enzyme/inhibitor	2ptcEI	15I,17I,19I	[**15I**,**17I**] [**17I**,**19I**]
Antibody/antigen	3hfmHY	58H	[**58H**]
Hormone/receptor	3hhrAB	21A,178A,164B,165B	[**21A**,172A] [64A,42B,43B,44B,**164B**,169B] [64A,43B,44B,103B,**164B**,169B] [175A,176A,**178A**,104B,169B] [**178A**,**164B**,**165B**,169B]
Cytokine	3inkCD	43D,45D,68D	[42D,**43D**] [42D,**68D**][**43D**,**45D**]

The types of protein complexes (column 1) with their corresponding PDB code and chain identifiers (column 2) are shown in the table along with their respective statistically significant high betweenness residues (column 3). The clusters including statistically significant high betweenness residues and experimentally annotated hot spots are also illustrated for each complex (column 4). The clustering ratio in each case was assumed to be ≥0.8, and it is defined as *ratio* $= N_e/[N_v(N_v-1)/2]$, where N_e is the number of edges among residues in the cluster, and N_v is the number of residues in the cluster. In columns 3 and 4, the green colored residues represent experimentally annotated hot spots and the blue colored residues represent statistically significant high betweenness residues, for which no experimental information on binding free energy is available. In each of the clusters, residues occurring in both columns 3 and 4 are shown in bold.

indicator of the centrality of the vertex in the network. The frequency distribution of the residue betweenness centrality β averaged in both sets of monomers and dimers follows a power law with an exponential cut-off $P(\beta) = C\beta^{-\eta}\exp(-\beta/\beta_c)$, with the corresponding values for the power law scaling exponent η and the exponential cut-off β_c approximately the same in the monomer and the dimer structures, and no statistically significant difference between the betweenness centrality distributions in the two cases (Fig. 1b). Unlike the frequency distribution of the residue number of links, the betweenness centrality frequency distribution is quite inhomogeneous, showing that a high number of residues have a small value of the betweenness centrality while only a few residues have a large value. This protein representation is in agreement with the wiring model proposed by Watts and Strogatz (1998), where an important role is played by the short cuts, responsible for the small values of the characteristic path length, while the clustering coefficient values remain high.

We study the protein–protein interaction mechanism using this representation of protein structures as small-world networks in order to elucidate some of the important topological changes occurring upon dimerization and the existence of topological determinants possibly related to key residues in the complex stability.

The process of dimerization between monomers can be viewed as a particular rewiring (rather than preferential attachment) in the system of the two monomers (each corresponding to a small-world network) due to the conformational changes, with the removal and addition of links occurring in each monomer, the formation of new links between the monomers, but on the other hand, leaving the frequency distributions of the residue number of links and betweenness centrality with no statistically significant difference between both sets of monomers and dimers (see Fig. 2 in Supplementary material). Interestingly, due to this rewiring process, new central residues (with statistically significant high values of central betweenness *z-score* ≥ 3.0) which are not homogenously distributed appear mainly at the protein–protein interfaces, while other previously central residues in the monomeric structures lose their centrality in the dimer structure. Conversely, there are a number of central residues in the monomer structures, which remain central in the complex (see Fig. 3 in Supplementary material).

Perhaps the most interesting result of this work is the strong correlation between the statistically significant central residues at protein–protein interfaces (topological determinants) with the most contributing residues to the binding free energy in protein–protein interactions. Experimental results based on Alanine scanning mutagenesis (Thorn and Bogan, 2001) and phenylalanine substitution (Mainfroid *et al*., 1996) of protein–protein interfaces has shown that the free energy contribution of individual amino acids in protein–protein binding is not uniformly distributed at the binding site; instead there are hot spots of binding free energy ($\Delta\Delta G \geq$ 1.0 Kcal/mol) comprised of a small subset of residues at the complex interface (Bogan and Thorn, 1998). Our analysis based on a set of 18 protein complexes with experimental information on hot spot residues and covering different biological examples of protein–protein interactions shows that the statistically significant high betweenness residues (z-*score* ≥ 3.0) occurring at the protein–protein interfaces are not uniformly distributed, but instead cluster together, surrounded by regions of residues with relatively low values of betweenness centrality, resembling that of the aforementioned free energy of binding distribution. More detailed analysis reveals a clear tendency of the statistically significant high betweeness residues to be located in hot spot regions, with the experimentally annotated hot spots exhibiting statistically significant high betweenness values in the majority of the cases. Table 1 shows that in the 18 complexes analyzed, 81% of these central residues form clusters with an experimentally annotated hot spot at the cluster center with 22 of these statistically significant high betweenness residues been actual hot spots (see Fig. 4 in Supplementary material).

The remaining 19% of our predicted residues occur mainly in those examples of protein complexes with little experimental information on hot spot residues, such as the enzyme/Inhibitor complex 2ptcEI, which contains only one experimentally annotated hot spot of binding free energy. On the other hand, these residues tend to be clustered together, are highly correlated with the experimental data on hot spot enrichment, and are generally conserved in sequence alignment or non-exposed to the solvent in the dimer structure, indicating that many of them are candidates of hot spots.

Despite the complexity involved in real physical interaction networks occurring in the protein structures, our simple network representation of the latter provides some insight into this complicated picture. Indeed, by using only one network topology characteristic (betweenness centrality) we are able to identify hot spot regions at protein–protein interfaces, taking into account the global topology of the complex whilst keeping its simplicity, which in combination with the reduced computational requirements are clear advantages of our method over previous physical models proposed to identify hot spots of binding free energy (Kortemme and Baker, 2002; Sheinerman and Honig, 2002). On the other hand, the graph-spectral method proposed by Brinda *et al*., including some additional information, such as residue solvent accessibility and sequence conservation, shows that the betweenness centrality turns out to be a better and simpler predictor of hot spot regions. There is a possibility that the correspondence between energy hot spots and structurally conserved residues remarked upon by Ma *et al*., could be related to the tendency of energy hot spots to remain central in the interacting network.

Finally, we should mention that a graph theoretical representation method similar to ours has been proposed by Shemesh *et al*. for identifying functional sites in protein structures. These authors reported that the most central residues in protein structure networks are found in functional sites (catalytic or ligand binding sites). Although their measure of centrality differs from our definition of betweenness centrality, it would be interesting to explore the possibility of using the information of residue centrality in the monomeric structures in order to improve the current methods of protein dockings. Some initial results in this direction have been addressed in our recent work (del Sol and O'Meara, 2004), where we show that some central residues in the monomeric structures remain central after dimerization and that possible information on hot spots of binding free energy could be obtained from the unbound structures. We are planning to continue this study in the future.

ACKNOWLEDGEMENTS

We would like to acknowledge interesting discussions in issues related to small-world view of protein structures with Dr Alfonso Valencia (CNB), and thank Professor Ruth Nussinov (NCI, Tel Aviv University) for helpful discussions on protein–protein interactions.

REFERENCES

Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Bogan,A.A. and Thorn,K.S. (1998) Anatomy of hot spots in protein interfaces. *J. Mol. Biol.*, **280**, 1–9.

Bollabas,B. (1985) *Random Graphs*. Academic Press, London.

Braiman,Y., Lindner,J.F. and Ditto,W.L. (1995) Taming spatiotemporal chaos with disorder. *Nature*, **378**, 465–467.

Brinda,K.V., Kannan,N. and Vishveshwara,S. (2002) *Protein Eng.*, **15**, 265–277.

del Sol,A. and O'Meara,P. (2004) Small-world network approach to identify key residues in protein–protein interaction. *Proteins*, **58**, 672–682.

Dokholyan,N.V., Li,L., Ding,F. and Shakhnovich,E.I. (2002) Topological determinants of protein folding. *Proc. Natl Acad. Sci. USA*, **99**, 8637–8641.

Freeman,L.C. (1977) A set of measures of centrality based on betweenness. *Sociometry*, **40**, 35–43.

Gerhardt,M., Schuster,H. and Tyson,J.J. (1990). A cellular automaton model of excitable media including curvature and dispersion. *Science*, **247**, 1563–1566.

Greene,L.H. and Higman,V.A. (2003) Uncovering network systems within protein structures. *J. Mol. Biol.*, **334**, 781–791.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22,** 2577–2637.

Kortemme,T. and Baker,D. (2002) A simple model for binding free energy hot spots in protein–protein complexes. *Proc. Natl Acad. Sci. USA*, **99**, 14116–14121.

Kuramoto,Y. (1984) Chemical oscillation. In *Waves and Turbulence*. Springer, Berlin.

Ma,B., Elkayam,T., Wolfson,H. and Nussinov,R. (2003) Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc. Natl Acad. Sci. USA*, **100**, 5772–5777.

Mainfroid,V., Mande,S.C., Hol,W.G., Martial,J.A. and Goraj,K. (1996) Stabilization of human triosephosphate isomerase by improvement of the stability of individual alpha-helices in dimeric as well as monomeric forms of the protein. *Biochemistry*, **35**, 4110–4117.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Nowak,M.A. and May,R.M. (1992) Evolutionary games and spatial chaos. *Nature*, **359**, 826–829.

Sheinerman,F.B. and Honig,B. (2002) On the role of electrostatic interactions in the design of protein–protein interfaces. *J. Mol. Biol.*, **318**, 161–177.

Shemesh,A., Amitai,G., Sitbon,E., Shklar,M., Netanely,D., Venger,I. and Pietrokovski,S. (2004) Structural analysis of residue interaction graphs. *The First Structural Bioinformatics Meeting*, ISMB/ECCB2004. pp. 22–23.

Strogatz,S.H. and Steward,I. (1993) Coupled oscillators and biological synchronization. *Sci. Am.*, **269**, 102–109.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Thorn,K.S. and Bogan,A.A. (2001) ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, **17**, 284–285.

Vendruscolo,M., Dokholyan,N.V., Paci,E. and Karplus,M. (2002) Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev. E*, **65**, 061910-1–061910-4.

Verkhivker,G.M., Bouzida,D., Gehlhaar,D.K., Rejto,P.A., Freer,S.T. and Rose,P.W. (2002) Monte carlo simulations of the peptide recognition at the consensus binding site of the constant fragment of human immunoglobulin G: the energy landscape analysis of a hot spot at the intermolecular interface. *Proteins*, **48**, 539–557.

Watts,D.J. and Strogatz,S.H. (1998) Collective dynamics of small-world networks. *Nature* (London), **393**, 440–442.