doi:10.1016/j.jmb.2003.08.061 J. Mol. Biol. (2003) 334, 781–791

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Uncovering Network Systems Within Protein Structures

Lesley H. Greene* and Victoria A. Higman

Oxford Centre for Molecular Sciences and Department of Chemistry, Central Chemistry Laboratory, University of Oxford, South Parks Road Oxford OX1 3QH, UK

*Corresponding author

Traditionally, proteins have been viewed as a construct based on elements of secondary structure and their arrangement in three-dimensional space. In a departure from this perspective we show that protein structures can be modelled as network systems that exhibit small-world, single-scale, and to some degree, scale-free properties. The phenomenological network concept of degrees of separation is applied to three-dimensional protein structure networks and reveals how amino acid residues can be connected to each other within six degrees of separation. This work also illuminates the unique features of protein networks in comparison to other networks currently studied. Recognising that proteins are networks provides a means of rationalising the robustness in the overall three-dimensional fold of a protein against random mutations and suggests an alternative avenue to investigate the determinants of protein structure, function and folding.

Keywords: networks; scale-free; small-world; protein structure; protein folding

The study of networks has recently become a blossoming area in science across many disciplines.[1–6] Any system of interconnected things, such as links between websites, business and social relationships between people and the connectivity of generators, transformers and sub-stations through power-lines, can be seen as a network. Studies of numerous and diverse networks such as the world-wide web, social networks, and electrical power grids have provided significant advances in the understanding of the topology, growth and dynamics of these systems.[1,2,7,8] Two key features underlying the connectivity of many real-world networks are small-world character[9] and scale-free behaviour.[7] The principles derived from the elucidation of these two network forms have been most recently applied to the biological realm revealing that molecular interaction networks involved in cellular, metabolic and transcriptional regulatory processes show small-world and scale-free behaviour.[2,10–14] Efforts have also just begun to apply network concepts to transition-states in protein folding,[15,16] protein-fold occurrence[17] and distribution[18] and to investigate the connectivity between protein folds.[19,20] We have applied network principles to native protein structures to further the boundaries of our understanding of the underlying determinants of protein folds. We present an alternative perspective by proposing that protein structures are in their most basic form networks.

Supplementary data associated with this article can be found at doi: 10.1016/j.jmb.2003.08.061

Abbreviation used: Ig, immunoglobulin.

E-mail address of the corresponding author: lesley.greene@bioch.ox.ac.uk

Results and Discussion

Conceptualising proteins as networks

Traditionally, the three-dimensional folds of proteins have been perceived as a construct based on elements of secondary structure and fold arrangement.[21–24] An alternative way in which to conceptualise and model protein structures is to consider the contacts between atoms in amino acid residues as a network of interactions irrespective of secondary structure and fold type. There is a natural distinction of contacts into two types: long-range and short-range interactions. Long-range interactions occur between residues that are distant from each other in the primary structure but are close in the tertiary structure. These interactions are important for defining the overall topology. Short-range interactions occur between

0022-2836/$ - see front matter © 2003 Published by Elsevier Ltd.

residues that are local to each other in both the primary and tertiary structure. These are found in secondary structure elements and are important in defining their structure.

For most networks what is termed a node and a link is fairly straight forwardly established. With proteins, however, we are faced with several ways in which to model the network of interactions. When looking at protein transition states, Dokholyan *et al.*[16] and Vendruscolo *et al.*[15] set the C^{α} atoms to be the nodes, and established a link between two nodes, if the atoms were within 8.5 Å of each other. In chemical terms, however, this is a simplification of the interactions within a protein. Side-chains play the pivotal role in both forming and fixing a protein structure, and any information on their orientation is lost in an analysis based solely on C^{α} atoms. We, therefore, use native structures where we are able to consider each amino acid to be a node, and a link to be established between two nodes, for any two atoms from two amino acid residues that are within 5 Å of each other. This produces a network with multiple links between nodes (Figure 1) in contrast to most other networks analysed in the literature, which generally only consider single links between nodes. We use multiple links as they more closely resemble the actual features of protein structures and give an indication of the strength of the interaction between residues. Our protein networks were constructed in two different ways, once using all, i.e. short and long-range, interactions, and once only using long-range interactions. We find that between 5% and 25% of the residues within each protein do not have any long-range contacts and are, therefore, not part of the long-range network.

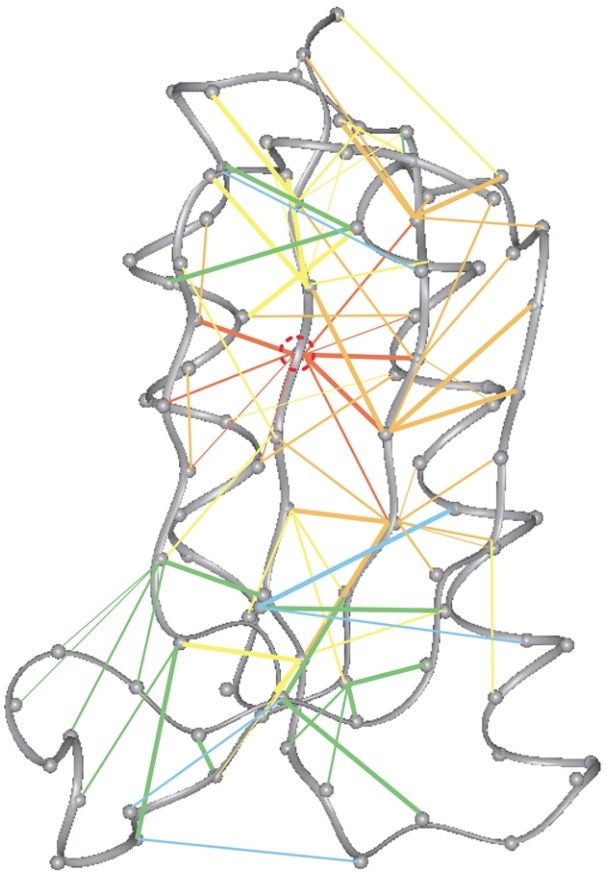

Figure 1. Example of an internal long-range interaction protein network. A ribbon diagram represents the polypeptide backbone of the mixed α/β protein, ribosomal S6 (pdb accession code; 1ris). α-Carbons for the 97 amino acid residues are shown as spheres. Illustrated are the contacts from valine 6 (denoted by a dotted circle) in the core of the protein and going out to five degrees of separation. Colour-coding by degrees is as follows: first (red), second (orange), third (yellow), fourth (green) and fifth (blue). The calculated number of links are weighted to take into account multiple links between the same nodes. Schematically we show the variation in the number of links between two nodes by variation in line thickness. Single links are depicted by thin coloured lines; there are two additional degrees of line thickness to indicate where there are two or more links between nodes.

Protein structures exhibit small-world behaviour

Our first goal was to determine if the pattern of links resembled small-world or random networks. A small-world network is one which has a relatively short characteristic path length, L, and a high clustering coefficient, C.[9] L is given by the number of links in the shortest path between two vertices averaged over all pairs of vertices. C is defined as set out by Watts & Strogatz:[9] if a vertex v has k_v neighbours, then the maximum number of links between these neighbours is $[k_v(k_v-1)]/2$. C_v gives the fraction of these possible links that actually exist and C is then defined as the average C_v over all vertices v. C is a measure of local clustering, which means that if two vertices X and Y are both connected to a third, Z, then for large C there is a high probability that X and Y are also directly linked to one another. The L and C parameters are used to define a small-world network by comparing it to a random network (one in which all edges have been randomly rewired) of the same size: $C_{\text{small-world}} \gg C_{\text{random}}$ and $L_{\text{small-world}} \gtrsim L_{\text{random}}$.[9]

We focused initially on the networks that consider protein structures as a system of long-range interactions only and analysed the clustering coefficients and characteristic path lengths. Intuitively we would expect a system where we have eliminated short-range interactions to have a low clustering coefficient and not to exhibit small-world behaviour. This is confirmed by the results shown in Figure 2. If we apply the formulae $L_{\text{random}} \sim \ln N/\ln k$ and $C_{\text{random}} \sim k/N$[15] for a random network with the same number of nodes, N, and average number of links, k, we find, as expected, that the long-range interaction network essentially does not differ from the random network.[9] When we considered the pattern of

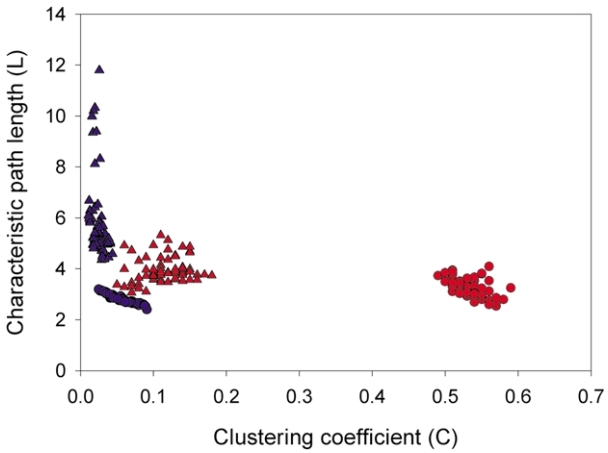

Figure 2. Small-world analysis of representative protein structures: clustering coefficient (C) *versus* characteristic path length (L). Shown in red are networks of the 65 representative proteins listed in Methods. Shown in blue are random networks with the same number of nodes and average connectivity. Triangles depict the long-range interaction networks, circles the long and short-range interaction networks.

links created by both long and short-range interactions together the results clearly indicate that proteins exhibit small-world properties in comparison to random networks (Figure 2). The results from this study concur with the small-world analysis conducted by Vendruscolo *et al*. using the simplified C^{α} model.[15] However, they do not find, as we have, distinct differences in network behaviour when looking at long *versus* long and short-range interactions.

We further investigated small-world behaviour in these systems by analysing protein structure networks under a range of conditions (data not shown). For the long-range network we find that C increases compared to C_{random} as we change the contact distance from 4 Å → 5 Å → 6 Å, however, not sufficiently to qualify for small-world behaviour. L decreases compared to L_{random} as the contact distance increases from 4 Å → 5 Å → 6 Å, which is to be expected, as the number of nodes remains constant but the number of links increases. A study was also conducted by changing the long-range residue number cutoff. C increases when going to shorter long-range cutoffs (from 10 to 4) as expected when the number of local interactions is increased. Although, at the cutoff of four C has doubled compared to the cutoff of ten, it is still only approximately 0.2 which is still too small to qualify for small-world behaviour. As the long-range cutoff increases (from 10 to 18), C tends towards C_{random}. L remains roughly equal, though L_{random} increases slightly as the cutoff is raised from 4 to 18. For the short and long-range network C remains the same as the contact distance increases from 4 Å → 5 Å → 6 Å. L decreases as the contact distance increases from 4 Å → 5 Å → 6 Å because as above the number of nodes stays constant but the number of links increases and so the average path length between nodes is lowered. L remains greater than L_{random} so small-world behaviour is seen at all three contact distance cutoffs.

Analysis of the connectivity distribution

Our second goal was to determine if the distribution pattern resembled scale-free networks. These characteristically have many nodes with few links and a few nodes with many links which when plotted display a skewed distribution. This is in contrast to a random network in which all nodes have about the same number of links (which therefore gives the network a characteristic scale) and the distribution of links approximates a bell-curve indicative of a Poisson distribution. Scale-free networks also follow a power-law distribution where the fraction of nodes having k edges, $p(k)$, decays as a power law $p(k) \sim k^{-\gamma}$ (where γ is generally between 2 and 3).[7,25] This scale-free characteristic, formalised by a power-law expression, also links networks with numerous and seemingly disparate phenomena such as earthquakes, fluctuations in stock market prices and fractals.[5,7]

First, we analysed the distribution pattern for both the long and short-range interactions. These results revealed that the pattern of links tends towards a bell-shaped Poisson distribution and is not scale-free (Figure 3A). These results also show that for all nine protein folds the mode number of links for a given node is approximately 100. Next we analysed the distribution pattern for exclusively long-range interactions and showed that for all nine folds there are a small number of nodes with many links and a large number of nodes with only a small number of links (Figure 3B). This distribution pattern is consistent with scale-free behaviour and initially suggested that the underlying long-range interaction network may be scale-free.

To explore further a possible relationship between the long-range interaction network and scale-free network systems three types of analysis were conducted. First, the distribution of long-range contacts in the 65 representatives of the nine protein folds were plotted on a double logarithmic scale to determine if one quantity can be expressed as a power of another quantity following, $N(k) = k^{-\gamma}$ and $\log N(k) = -\gamma \log k$.[26] Here, $N(k)$ is the distribution of the number of nodes in a given protein, k is the number of links and γ is the degree exponent. A straight-line fit of data on a log–log plot is a standard way in which to measure the distribution and identify the power-law dependence by calculating the slope.[26] One interpretation of this study of long-range links is that it approximates a power-law distribution as the log of $N(k)$ plotted against the log of k scales with a power-law for large k as evidenced by the linear arrangement, but only in the tail of the distribution (Figure 4A).[7,27] The scaling exponent for protein

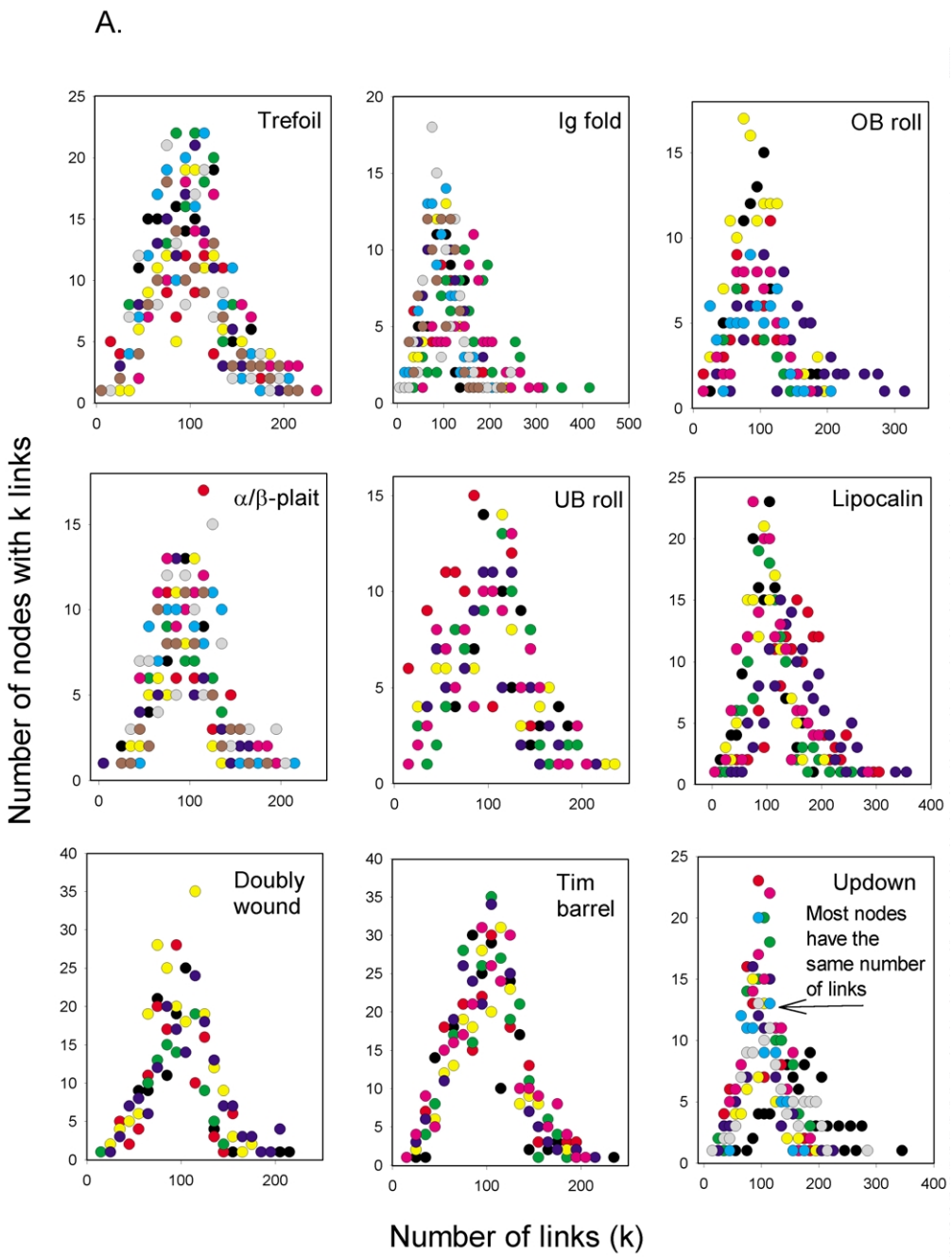

Figure 3 (*legend opposite*)

networks is around 3.7 and is unrelated to protein size or fold-arrangement. Because of the modest size of the protein network the scaling region is less prominent than for some larger networks. While for these larger networks γ has been reported as being in the range of 2–3, γ here is still in the range of other smaller networks where $\gamma = 4.0$.[7]

To investigate further the underlying long-range network connectivity we conducted a second study which takes into account that power-law distributions generally exhibit a characteristic exponential cut-off due to the finite size of the system.[28,29] Networks which are scale-free with exponential cutoffs can be found for example in scientific collaboration networks, protein–protein interaction networks and movie-actor networks.[25,28,30] We might expect this type of behaviour in protein structure networks because the limitations on the number of contacts an amino acid can form with others due to restrictions on space within the three-dimensional fold and the finite size of the atoms composing the amino acid residues should inevitably generate a cut-off in the degree distribution. We applied the following equation, $k^{(-\gamma)} \exp(-k/k_c)$, where γ is the

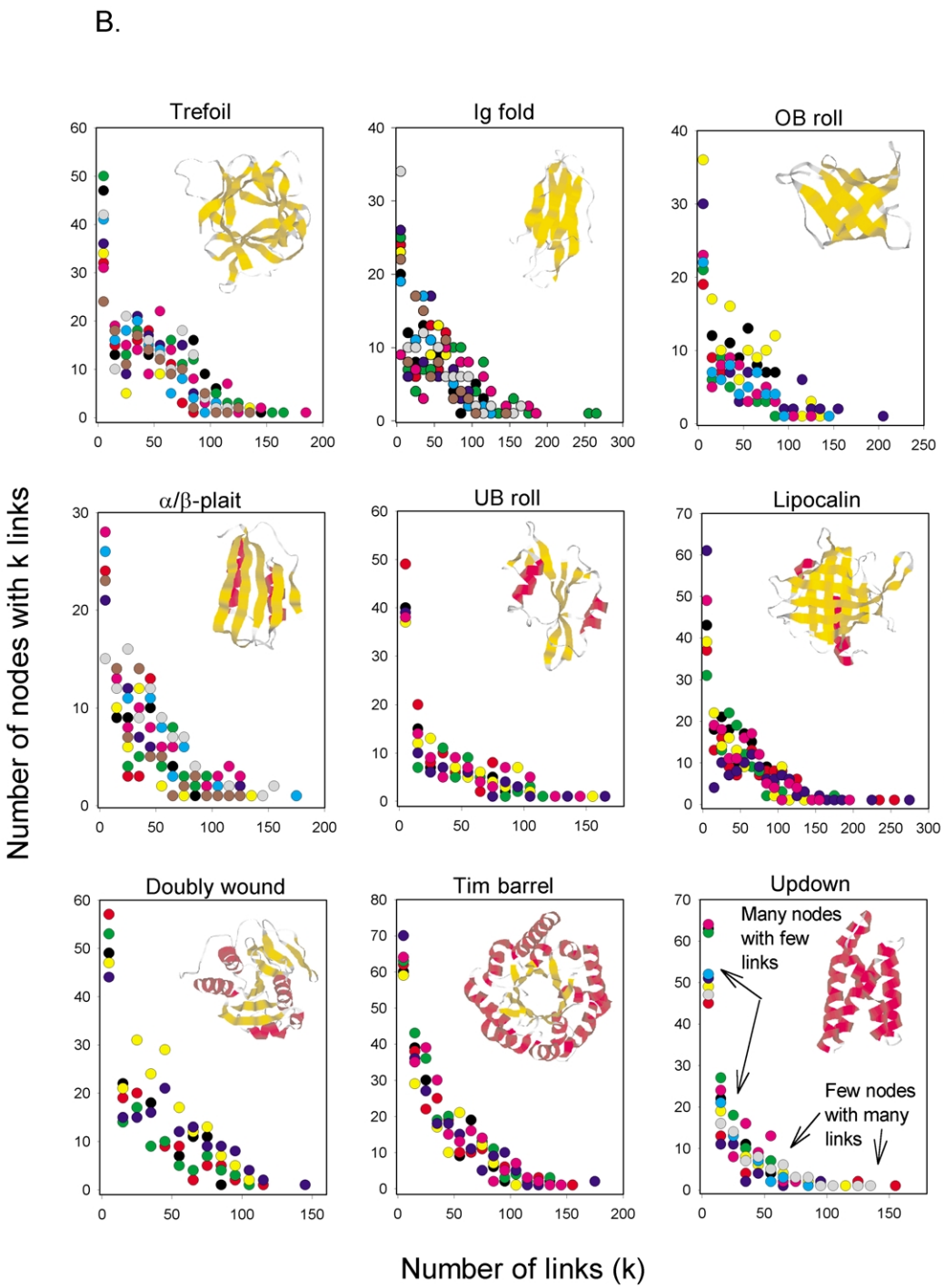

Figure 3. Analysis of long and short-range interactions in representative protein structures. Each protein analysed within a fold-type is shown in different colours on the scatter plot. Data are clustered in groups of ten. A, Distribution of long and short-range interactions in nine representative protein folds. B, Distribution of long-range interactions in nine representative protein folds. Included are ribbon drawings of the nine representative protein folds. Fold type and Pdb accession codes are as follows: trefoil (1i1b); Ig fold (1agd:b); OB roll (1mjc); α/β-plait fold (2acy); UB roll (1sha:a); lipocalin (1rbp); doubly wound (3eca:a); tim barrel (1ho4:a); updown helical bundle (1fap:b). Drawings are created with Rasmol, ver. 2.7.1. All data in this Figure are plotted and analysed with using Sigma Plot, ver. 4.1 (SPSS).

power-law scaling exponent and k_c is the exponential cutoff.[28] We believe this is a more appropriate way to analyse our data in comparison to the method illustrated in Figure 4A. The results in Figure 4B indicate that there is a power-law distribution for small k (as opposed to large k identified in Figure 4A), but the exponential cutoff is quite severe within this system under these conditions.[29]

This result led us to conduct a third type of analysis in which the data were graphed onto a linear-log plot. In this type of analysis data points which fall on a straight line indicate that

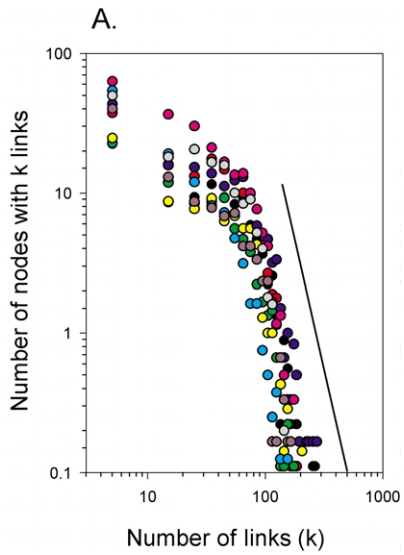

Figure 4 (*legend on p.788*)

the underlying form of distribution is exponential, which is another variation of real-world networks.[25] The results of this study further confirm that the distribution of contacts in a protein network is dominated by an exponential (Figure 4C) term. This result is not totally unexpected. Networks with exponentially decaying tails, also termed single-scale networks, can be found in other systems such as forms of yeast networks, some food webs and the power grid of southern California.[25,31,32] We also explored variations on the degree distributions using different protein structure parameters (data not shown). At 4 Å and 6 Å contact cutoffs in two distinct fold types studied (Updown and Tim barrels) the distributions are qualitatively the same on a semi-log plot as with 5 Å. The distributions are also consistent between the ten residue cutoff length and cutoff lengths of 6, 14 and 18 residues. At a cutoff distance of four residues the distribution begins to deviate and tend slightly towards a Poisson distribution. We attribute this to the beginning of secondary structure effects.

Degrees of separation in protein networks

The above results suggest that the behaviour in terms of networks is the same irrespective of protein-fold type, size and secondary structure composition. Where secondary structure, however, does become important is when considering the small-world phenomenon of degrees of separation.[1,9,33] This is a measure of how many links it takes to get from one node within a network to another. If on average any two nodes within a system can be connected through, for example, six links, we would say that the average degree of separation for that network is six.[33,34] We applied this small-world concept to the long-range network in proteins and found that the average degree of separation between any two amino acid residues is smaller than six (Figures 1 and 5). When analysing the results in terms of secondary structure, we find that all-β proteins have a lower average degree of separation than mixed α/β and all-α proteins of comparable size and that the average degree of separation increases at a slower rate as network size increases. This can be rationalised by considering the intrinsic nature of interactions involving each type of secondary structure. Within helical structures most inter-residue interactions are within a helix and only a few are between helical interfaces. It will, therefore, require a number of links to traverse residues within an all-helical fold. The β-strands on the other hand generally form sheets and this architecture requires fewer links to traverse residues that are sequentially farther apart. The concept of degrees of separation may have relevance to the mechanism of protein folding. We envisage it as a bridge between time and shape. The links dictate the final structure and we perceive them forming in degrees within relevant biological time. The network could originate from a select number of residues, which is akin to a nucleus,[35,36] and becomes structured on a faster folding time-scale than the majority of the amino acid residues.

Conclusion

The results in this report show that proteins can be successfully modelled as networks. There is a richness of variation in the forms of real-world networks. They can be small-world, random or regular.[9] The degree distributions can range from Poisson distributions to scale-free, single-scale with fast decaying exponential tails, broad-scale which show a power-law regime with an exponential

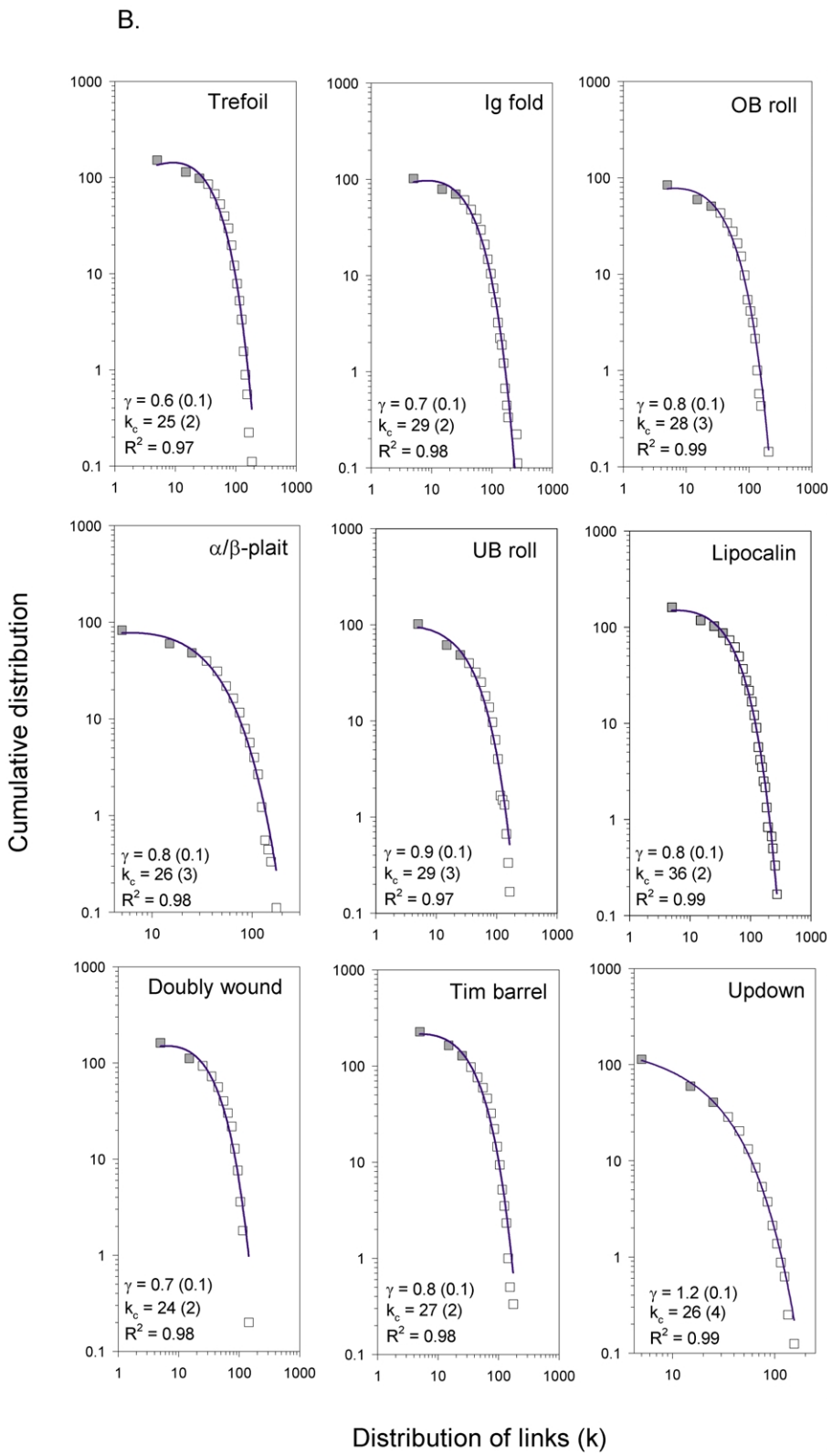

Figure 4 (*legend on p.788*)

cut-off and Gaussian distributions.[25,31] When considering protein structure networks composed of both long and short-range interactions the network is clearly small-world with a Poisson type of degree distribution. The long-range interaction network interestingly is not small-world and the distribution of contacts while having an underlying scale-free behaviour is dominated by an exponential term indicative of a single-scale system. The results of the small-world and connectivity

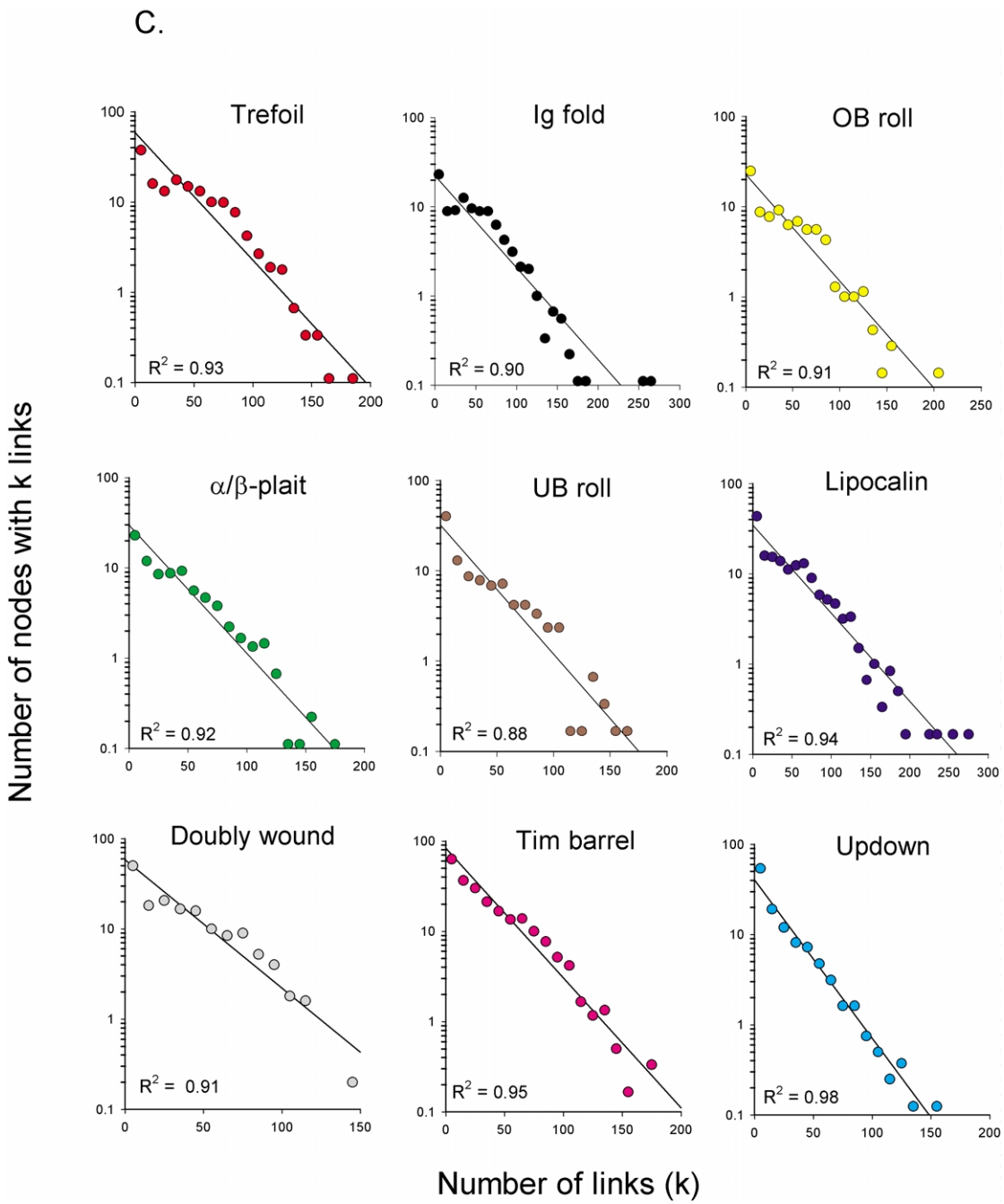

Figure 4. Analysis of the distribution of long-range links plotted on log–log and semi-log scales. A, Each curve represents the connectivity averaged for the multiple representative members from the nine fold types: trefoils (red), Ig fold (black), OB roll (yellow), α/β-plait fold (green), UB roll (dark red), tim barrels (pink), lipocalins (blue), doubly wound (grey) and updown bundles (cyan). The data have been logarithmically binned, with a bin ratio of 10. The tail of the distribution ($k > 45$) was fit by linear regression and the scaling exponent γ, for each fold-type ranges between 3.2 and 4.3. The average γ for the ninefold types is 3.7($\pm$0.3) with the fit denoted by the black line using the equation, $y = 10^9 x^{-3.7}$. B, The nine different fold types were analysed to determine the precise γ and exponential cutoff (kc) in the degree distribution of the log-transformed data. Data are plotted as cumulative distribution to reduce the noise in these small network systems and analysed following the functional form, $k^{-(\gamma-1)} \exp(-k/k_c)$.[43] The fit by non-linear regression is shown as a blue line. Grey squares denote the region which is fit by the power-law form; white squares denote the region fit by the exponential term. The observed values for the fit of the data are shown in the individual graphs with standard errors in parentheses. The P values for both the power-law and exponential terms are less than 10^{-3}, indicating a reasonably good fit. C, The distribution for each fold type is graphed on a semi-log plot with the slope of the line determined by linear regression. All data in this Figure were plotted and analysed using Sigma Plot, ver. 8 (SPSS).

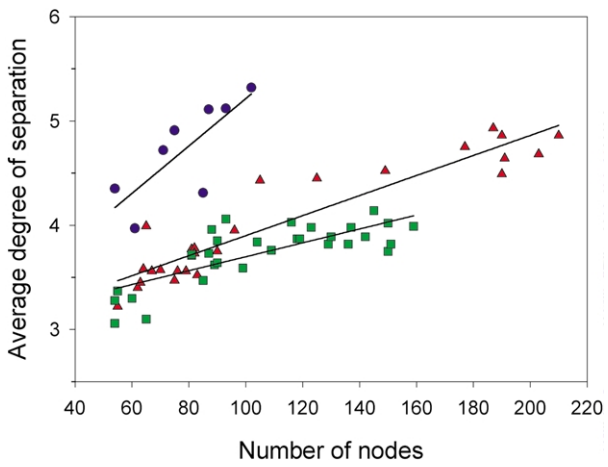

Figure 5. Average degree of separation plotted *versus* number of nodes in long-range interaction networks of the 65 representative proteins listed in Methods. The proteins are grouped according to secondary structure content as follows: α-helical proteins (blue circles), mixed α/β proteins (red triangles), β-sheet proteins (green squares).

distribution studies are consistent in all 65 proteins irrespective of topology, secondary-structure composition, domain size or origin of species. We also show in this work that the network of long-range interactions, which is key to the protein's topology, can be formed within six degrees of separation. However, this latter study unlike the others does have a direct correlation to secondary-structure composition.

Network systems can be analysed for robustness which equates to their degree of tolerance against errors.[37] The scale-free connectivity distribution identified in other network systems has been shown to confer robustness to the network against random attacks but not directed attacks.[37] Dunne *et al*. find that networks which do not exhibit small-world or scale-free behaviour but show a skewed distribution like the networks studied here display a similar pattern of robustness.[38] Initial studies on the protein structure networks presented here reveal that the long-range interaction network resembles these systems in their robustness to random *versus* directed attacks: targeted removal of the most highly connected amino acid residues has a substantially greater affect on the integrity of the network structure in contrast to a random removal of amino acid residues comprising the network (see Supplementary Material). In this way, modelling proteins as networks opens up an alternative way to analyse naturally occurring mutations *in vivo* and directed mutations *in vitro*. It will be interesting to see comprehensive theoretical and experimental studies that correlate mutations to connectivity and the effects on structure, function, thermodynamic stability and folding kinetics.

Proteins are particularly interesting network systems because they generally have multiple links between the same two nodes. This form of connectivity has gone largely unstudied in previous network systems, where the models have considered only one link between nodes. Additionally, we propose that protein domains offer a promising system in which to investigate fundamental principles governing networks more generally, such as the determinants of network topology, evolution, dynamics and formation. Protein domains provide unique challenges which include the relationship between links and thermodynamics as well as kinetics, the spatial constraints of individual folds and primary structure considerations as these networks are formed by sequentially linked nodes. In addition, proteins can be self-contained or interconnected systems of multiple proteins and are amenable to rigorous study with numerous high-resolution biophysical techniques and molecular-dynamics simulations.

Having shown the presence of network properties within protein structures, recasting our view of proteins towards networks promises to become a fruitful avenue in which to address the relationship between sequence and structure. This encompasses the mechanism of protein folding, biological function and *de novo* design. The most difficult problem in this area, that of predicting a protein structure from sequence information alone, may be facilitated by applying network models[7] such as fitness, growth and preferential attachment and developing new models towards this aim. We propose that a more evident scale-free correlation in protein structure networks may be found in less well-structured protein states, such as transition state structures or very early folding intermediates where there is a very sharp delineation between only a few residues in the nucleus which have a large number of native long-range contacts in comparison to the large majority of residues which have very few to no native long-range contacts. Our work shows that, in effect, in the native state there are too many residues which have an intermediate number of contacts for scale-free behaviour to be clearly delineated. So with these residues being less well connected in an intermediate transitional-like structure the underlying scale-free distribution may be more pronounced. In addition, scale-free behaviour is associated with models of growth and the networks studied here are the end product of a dynamic process. This again suggests that real scale-free behaviour may be more evident in protein structures that are still part of the folding process. Further work along these lines is required to test rigorously this proposal.

Towards understanding the evolution of protein folds, these findings in present day proteins are suggestive of the idea that the primordial proteins shared these network characteristics and they may have been fundamental to their initial selection and conservation.

Methods

Protein structures

Our studies focused on nine highly populated fold types representing the four protein classes: all-α, all-β, mixed α/β and mixed α + β.[39] Representative members of each fold-type were selected to ensure diversity in function and sequence. The selected proteins also sample the range of life within all three major branches (archae bacteria, eubacteria and eukaryotes). Pdb codes grouped by fold-type are as follows: trefoil (1ava:c, 1ce7:b, 1eyl:a, 1hwn:b, 1ilb, 1jly:b, 2aai:b, 1wba, 2fgf); Ig (1agd:b, 1bih:a, 1cd8, 1epf:d, 1hyx:l, 1ten, 1tlk, 3kbp:c, 1hng:a); OB roll (1b8a:a, 1cuk, 1eif, 1eov:a, 1gvp, 1hro, 1mjc); α/β-plait (1b7f, 1cvj:b, 1dar, 1feo, 1psd:a, 1ris, 1urn:c, 2acy, 1aye); UB roll (1ayc:a, 1bf5:a, 1cj1:a, 1d4t:a, 1fbv:a, 1sha:a); lipocalin (1bbp:a, 1beb:a, 1bj7, 1gm6:a, 1obp:a, 1rbp); doubly wound (1bmt:a, 1dio:a, 1pdo, 3eca:a, 3rab:a); tim barrels (1eun:a, 1g4t:a, 1ho4:a, 1nsj, 1rpx:a, 7tim:a); updown (1e86:a, 1fap:b, 1h7i:a, 1jr8:a, 1j8m:f, 1vlt:a, 256b:a, 2hmz:a). These criteria ensure that this study can be generalized to essentially all proteins with only one caveat; we have not taken into account in this report the subset of protein folds that require cofactors to form and stabilise the native-state because this biases the interactions between the amino acid residues. These include for example the globins and may form the basis for future studies.

Computational analysis

In the present work we analyse three-dimensional structures with all heavy atoms. The cut-off for contacts between atoms is 5 Å, which approximates the upper limit for attractive London–van-der-Waals forces,[40] and the contacts were determined using the program Contact.[41] Interactions are considered long-range if they occur between amino acid residues that are ten or more residues apart in the primary sequence. Short-range interactions are those occurring between amino acid residues that are less than ten residues apart. This is a conservative estimate based on the determination that the average length of an α-helix and a β-strand are 11 and 6 residues, respectively.[42] Further analysis of small-world and scale-free properties was carried out using programs specifically written in Fortran77 and run on a Sun workstation for this purpose.

Acknowledgements

We are indebted to Janet Moloney for inspiring conversations regarding power laws, networks and protein structures. We thank Jane Richardson for very helpful suggestions on how to better analyse protein structures as networks. We are grateful to Christina Redfield & Jonathan Jones for valuable discussions and critically reading this manuscript. L.H.G. acknowledges the National Science Foundation for an International Research Fellows Award. V.A.H. acknowledges awards from St. Peter's College (Oxford), the BBSRC and OCMS. This is a contribution from the Oxford Centre for Molecular Sciences which is supported by the BBSRC, MRC and EPSRC.

References

1. Strogatz, S. H. (2001). Exploring complex networks. *Nature*, **410**, 268–276.
2. Wolf, Y. I., Karev, G. & Koonin, E. V. (2002). Scale-free networks in biology: new insights into the fundamentals of evolution? *BioEssays*, **24**, 105–109.
3. Watts, D. J. (1999). *Small Worlds: The Dynamics of Networks between Order and Randomness*, Princeton University Press, Princeton, New Jersey.
4. Kauffman, S. A. (2000). *Investigations*, Oxford University Press, New York.
5. Barabasi, A.-L. (2002). *Linked: The Science of Networks*, Persus Publishing, Cambridge.
6. Buchanan, M. (2002). *Small World: Uncovering Nature's Hidden Networks*, Weidenfeld Nicolson, London.
7. Barabasi, A.-L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**, 509–512.
8. Albert, R. & Barabasi, A.-L. (2000). Topology of evolving networks: local events and universality. *Phys. Rev. Letters*, **85**, 5234–5237.
9. Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *Nature*, **393**, 440–442.
10. Maslov, S. & Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science*, **296**, 910–913.
11. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabasi, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.
12. Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. (2002). Network motifs in the transcriptional regulation network of *Eschericia coli*. *Nature Genet.* **31**, 64–68.
13. 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.
14. Luscombe, N. M., Qian, J., Zhang, Z., Johnson, T. & Gerstein, M. (2002). The dominance of the population by a selected few: power-law behavior applies to a wide variety of genomic properties. *Genome Biol.* **3**, 0040.1–0040.7.
15. Vendruscolo, M., Dokholyan, N. V., Paci, E. & Karplus, M. (2002). Small-world view of the amino acid residues that play a key role in protein folding. *Phys. Rev. E*, **65** 061910-1–061910-4.
16. Dokholyan, N. V., Li, L., Ding, F. & Shakhnovich, E. I. (2002). Topological determinants of protein folding. *Proc. Natl Acad. Sci. USA*, **99**, 8637–8641.
17. Qian, J., Luscombe, N. M. & Gerstein, M. (2001). Protein family and fold occurrence in genomes: power-law behavior and evolutionary model. *J. Mol. Biol.* **313**, 673–681.
18. Koonin, E. V., Wolf, Y. I. & Karev, G. P. (2002). The structure of the protein universe and genome evolution. *Nature*, **420**, 218–223.
19. Wutchy, S. (2001). Scale-free behavior in protein domain networks. *Mol. Biol. Evol.* **18**, 1694–1702.
20. Dokholyan, N. V., Shakhnovich, B. & Shakhnovich, E. I. (2002). Expanding protein universe and its

origin from the biological Big Bang. *Proc. Natl Acad. Sci. USA*, **99**, 14132–14136.
21. Levitt, M. & Chothia, C. (1976). Structural patterns in globular proteins. *Nature*, **261**, 552–558.
22. Orengo, C. A., Michie, A. D., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
23. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
24. Lesk, A. M. (2001). *Introduction to Protein Architecture*, Oxford University Press, Oxford, pp. 59–125.
25. Amaral, L. A. N., Scala, A., Barthelemy, M. & Stanley, H. E. (2000). Classes of small-world networks. *Proc. Natl Acad. Sci. USA*, **97**, 11149–11152.
26. Bak, P. (1997). *How Nature Works: The Science of Self-Organized Criticality*, Oxford University Press, Oxford pp. 1–32.
27. Doye, J. P. K. (2002). Network topology of a potential energy landscape: a static scale-free network. *Phys. Rev. Letters*, **88** 238701-1–238701-4.
28. Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proc. Natl Acad. Sci. USA*, **98**, 404–409.
29. Watts, D. J. (2003). *Six Degrees: The Science of a Connected Age*, William Heinemann, Random House, UK pp. 101–129.
30. Jeong, H., Mason, S. P., Barabasi, A.-L. & Oltvai, Z. N. (2001). Lethality and centrality in protein-networks. *Nature*, **411**, 41.
31. Dunne, J. A., Williams, R. J. & Martinez, N. D. (2002). Food-web structure and network theory: the role of connectance and size. *Proc. Natl Acad. Sci. USA*, **99**, 12917–12922.
32. Guelzim, N., Bottani, S., Bourgine, P. & Kepes, F. (2002). Topological and causal structure of the yeast transcriptional regulatory network. *Nature Genet.* **31**, 60–63.
33. Milgram, S. (1967). The small-world problem. *Psychol. Today*, **2**, 60–67.
34. Guare, J. (1990). *Six Degrees of Separation: A Play*, Vintage, New York.
35. Fersht, A. R. (1995). Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. *Proc. Natl Acad. Sci. USA*, **92**, 10869–10873.
36. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1994). Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry*, **33**, 10026–10036.
37. Albert, R., Jeong, H. & Barabassi, A.-L. (2000). Error and attack tolerance in complex networks. *Nature*, **406**, 378–382.
38. Dunne, J. A., Williams, R. J. & Martinez, N. D. (2002). Network structure and biodiversity loss in food webs: robustness increases with connectance. *Ecology Letters*, **5**, 558–567.
39. Holm, L. & Sander, C. (1998). Dictionary of recurrent domains in protein structures. *Proteins: Struct. Funct. Genet.* **33**, 88–96.
40. Tinoco, I. Jr, Sauer, K. & Wang, J. C. (1995). *Physical Chemistry: Principles and Applications in Biological Sciences*, 3rd edit., Prentice-Hall, New Jersey pp. 456–544.
41. Collaborative Computational Project, Number 4 (1994). The CCP4 suite: programs for protein crystallography. *Acta Crystallog. sect. D*, **50**, 760–763.
42. Schultz, G. E. & Schirmer, R. H. (1979). *Principles of Protein Structure*, Springer, New York pp. 66–107.
43. Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Rev.* **45**, 167–256.

Edited by J. Thornton

(Received 25 March 2003; received in revised form 14 August 2003; accepted 15 August 2003)

Supplementary Material for this paper comprising one Figure is available on Science Direct