Chapter 3

Small-World Nature of Protein Contact Networks

3.1 Introduction

There have been several efforts to study protein structures as (graphs) networks. In these studies the effort has been to analyse globular proteins as systems composed of interacting parts. In recent years, with the elaboration of network properties in a variety of real networks, Vendruscolo et al. [6] showed that protein structures have small-world topology. Greene and Higman [7] studied the short-range and long-range interaction networks in protein structures of 65 proteins and showed that long-range interaction network is not small world and its degree distribution, while having an underlying scale-free behaviour, is dominated by an exponential term indicative of a single-scale system. Atilgan et al. [8] studied globular protein structures and analysed the network properties of the core and surface of the proteins. They established that, regardless of size, the cores have the same local packing arrangements. They also explained, with an example of binding of two proteins, how the small-world topology could be useful in efficient and effective dissipation of energy, generated upon binding.
3.2 Small-World topology of PCNs

The small-world nature of protein networks is a basic finding. The small-world nature of a network is reflected in two properties: high clustering compared to their random controls, and a logarithmic increase in the characteristic path length with increase in the size of the network.

![Figure 3.1: Distribution of sizes of proteins analysed.](image)

The function that a protein serves in the cell is decided by the structure of the protein. Proteins, owing to oft-repeated structural constructs, could be classified [65] (Structural Classification of Proteins, http:// scop.mrc-lmb.cam.ac.uk/scop/) based on their structural composition. We analysed 80 proteins (listed in Table Nos. 2.1, 2.2, 2.3, 2.4), 20 each from four major categories (α, β, α/β, α + β) of the SCOP structural classification. These are from diverse functional groups: hydrolase, transferase, protease, calcium binding, oxynedoreductase, antifungal, signalling, transport, toxin, coagulation factor etc. to name a few. The size of these proteins varied from 73 to 2359 amino acids. Fig. 3.1 shows the histogram of size of these proteins and their break-up across the structural classes.

We calculated the average clustering coefficient (C) and the characteristic path length (L) of the proteins. Fig. 3.2 (a) shows the L versus C plot. As seen in the figure, on the scale of 0 to 1, the proteins have a very high value of clustering coefficients. Apart from very high C, what is interesting
3.2 Small-World topology of PCNs

Figure 3.2: (a) $L-C$ plot of proteins from four structural classes. (b) Increase in the $L$ of proteins with logarithmic increase in size ($n_r$). The dotted line is a log-linear fit to the PCN data.

is that these 80 proteins are almost indistinguishable with this parameter. Thus while presenting a generic property (that of high clustering), of proteins similar to that of a large number of other complex networks, the small-world network result provides a grim picture in terms of our ability to correlate this specific network (geometric) parameter to the proteins’ structure and
3.2 Small-World topology of PCNs

For a network to be classified as a small-world network, apart from high clustering, its $L$ should increase only as a $\log(n_r)$. Such a logarithmic scaling of $L$ with $n_r$, makes it a small-world network, i.e. any node on the network could be reached from any other node in an exceptionally few number of steps. Fig. 3.2 (b) shows that the $L$ of these 80 PCNs scale logarithmically with the size of the network. These two properties thus ascertain the small-world nature of the PCNs across structural classes. Fig. 3.2 (b) also shows $L$ of random controls of PCNs (marked with an arrow).

Fig. 3.3 shows the summary plot of $L-C$ for all 80 proteins, with their Type-I random controls in the bottom-left, regular controls in the extreme-right, and PCNs in the middle. The inset of the figure shows the means and standard deviations of $L$ and $C$ of the corresponding data. As seen in the figure the $L$ of PCNs are of the same order of magnitude as those of their Type I random controls. PCNs of these proteins have very high clustering coefficients compared to their random controls (statistically significant, $p < 0.001$; Two-Sample Kolmogorov-Smirnov Test). The $L$ and $C$ computed here and in the rest of this chapter, for random and regular controls, were computed based on analytical formulae mentioned in Subsection 2.3.1 and
3.3 Globular and Fibrous Proteins

Most proteins are "globular" in their three-dimensional structure, into which the polypeptide chain folds into a compact shape. In contrast, "fibrous" proteins have relatively simple, elongated three-dimensional structure suitable for their biological function (see Fig. 3.4 (b)). The "small-world" nature of globular proteins was argued [8] to be required for enhancing the ease of dissipation of disturbances. If that were true, the fibrous proteins should depart from the small-world nature. We studied fibrous proteins and compared their network properties with globular proteins of comparable sizes. Table 3.3 shows the details of these proteins. As shown in the L–C plot in Fig. 3.4(a), fibrous proteins have larger $L$, although the $C$ are similar to those of globular proteins. Thus, in this respect, the fibrous proteins also show "small-world" properties. The average diameter for the fibrous proteins ($D = 15$) was found to be larger than that of the globular proteins ($D = 8.57$). This is expected because of the elongated structure of fibrous proteins. Despite this major difference in structure, the network properties of fibrous proteins and globular proteins are not very different. This indicates that the "small-world" property of proteins is generic and persists irrespective of structural differences.

<table>
<thead>
<tr>
<th>Sr.No.</th>
<th>PDB ID</th>
<th>$n_r$</th>
<th>$L$</th>
<th>$C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>1CGD</td>
<td>90</td>
<td>5.401</td>
<td>0.7463</td>
</tr>
<tr>
<td>F2</td>
<td>1CAG</td>
<td>88</td>
<td>5.274</td>
<td>0.6933</td>
</tr>
<tr>
<td>F3</td>
<td>1EI8</td>
<td>172</td>
<td>5.610</td>
<td>0.6045</td>
</tr>
<tr>
<td>F4</td>
<td>1QSU</td>
<td>89</td>
<td>5.337</td>
<td>0.6432</td>
</tr>
<tr>
<td>G5</td>
<td>1ABA</td>
<td>87</td>
<td>3.382</td>
<td>0.5942</td>
</tr>
<tr>
<td>G6</td>
<td>1AEG</td>
<td>86</td>
<td>4.066</td>
<td>0.5952</td>
</tr>
<tr>
<td>G6</td>
<td>1AYI</td>
<td>86</td>
<td>3.812</td>
<td>0.6025</td>
</tr>
<tr>
<td>G7</td>
<td>1CG6</td>
<td>88</td>
<td>3.740</td>
<td>0.6055</td>
</tr>
<tr>
<td>G8</td>
<td>1CEI</td>
<td>85</td>
<td>3.713</td>
<td>0.6024</td>
</tr>
<tr>
<td>G9</td>
<td>1CTJ</td>
<td>89</td>
<td>3.763</td>
<td>0.5968</td>
</tr>
<tr>
<td>G10</td>
<td>1DSL</td>
<td>88</td>
<td>3.404</td>
<td>0.5509</td>
</tr>
</tbody>
</table>

Table 3.1: List of four fibrous(F1–F4) and seven globular proteins(G5–G10) analysed.
Figure 3.4: (a) L–C plot of Fibrous and Globular proteins of comparable sizes. (b) Examples of three-dimensional structures of a fibrous and globular protein (not to the scale) with their PDB codes.

3.4 α and β Proteins

Figure 3.5: L–C plot for α and β proteins. Arrows indicate the means of C for α and β proteins.

As seen earlier (Figure 3.2(a)), both α and β class of proteins show small-world properties. Given that these are two distinct structural units one would want to know how that reflects on the global network parameters of α and
3.5 Degree Distributions of PCNs and LINs

The distribution of the degrees is an important property which characterises the network topology. The degree distribution of a random network is characterised by a Poisson distribution. The degree distribution of many real-world networks has been shown to be that of the scale-free type [92]. Many models have been proposed to explain the evolution of network and the degree distribution with which they are characterised at present.

We analysed the degree distribution of the 80 proteins mentioned above. Figure 3.6 shows the normalised degree distributions of $\alpha$, $\beta$, $\alpha + \beta$, and $\alpha/\beta$ protein networks. Figure 3.6 shows the scatter plot of normalised degree distributions ($P(k)$) of all 80 proteins of four different classes. Data points in each plot indicate $P(k)$ values for all the residues of 20 proteins of the respective class. Solid line is a Gaussian fit to the mean of $P(k)$ for each value of $k$.

As seen, shapes of these distributions are single humped, Gaussian-like [7]. Importantly, unlike in scale-free degree distributions the number of nodes with very high degree falls off rapidly. This is interesting as in scale-free networks high-degree nodes (hubs) are known to be the facilitators of communication across the network by providing shorter routes through them. Hence hubs would partially explain small-world nature. But, clearly, the distribution of contacts in proteins is dominated by an exponential term.

Figure 3.7 shows the 1-$\sigma$ standard deviation of the data of 20 PCNs for the normalised degree distribution of respective classes. Solid line is a Gaussian fit to $\langle P(k) \rangle$, the mean of $P(k)$ for each value of $k$. The Gaussian fit was obtained with

$$y(x) = \frac{A}{w\sqrt{\pi}/2} \exp\left(-\frac{2(x - xc)^2}{w^2}\right).$$
Figure 3.6: Scatter plot of degree distributions for (a) $\alpha$, (b) $\beta$, (c) $\alpha + \beta$, and (d) $\alpha/\beta$ proteins, 20 of each class. The solid line is the Gaussian fit through the means.

Table 3.2 gives parameter values of the goodness of fit. Here, $R^2$ is the coefficient of determination.

<table>
<thead>
<tr>
<th>Class</th>
<th>$x_c$</th>
<th>$w$</th>
<th>$A$</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>7.922</td>
<td>3.524</td>
<td>3.443</td>
<td>0.9126</td>
</tr>
<tr>
<td>$\beta$</td>
<td>8.175</td>
<td>5.429</td>
<td>6.555</td>
<td>0.9189</td>
</tr>
<tr>
<td>$\alpha + \beta$</td>
<td>7.506</td>
<td>4.216</td>
<td>5.535</td>
<td>0.9732</td>
</tr>
<tr>
<td>$\alpha/\beta$</td>
<td>7.961</td>
<td>5.192</td>
<td>6.146</td>
<td>0.9684</td>
</tr>
</tbody>
</table>

Table 3.2: Degree Distribution Curve Fitting. Parameters and goodness of fit.

Degree distribution of LINs is shown in Fig. 3.8. As seen the $P(k)$ of LINs show a single-scale decay with no typical node present in them.
3.6 Diameter of PCNs and LINs

The concept of diameter, strictly speaking, is applicable only to single-component graphs. Owing to the presence of the backbone connectivity, PCNs and its other versions are always single-component. Diameter is expected to scale with the number of nodes in the same way as the characteristic path length ($L$). Fig. 3.9 shows that $D$ does scale logarithmically with $n_r$. Diameter, since it is maximal of the distances between two nodes, the growth of $D$ with $n_r$ imposes upper limit on the rate of growth of $L$ with $n_r$.

3.7 $C-n_T$ Plot

Clustering coefficient is essentially the probability of formation of triangles in the network. In a random network the probability that a given node’s two first-neighbours themselves are connected is equal to that of any two randomly selected nodes are connected. Therefore, clustering coefficient ($C_{rand}$)
Figure 3.8: Scatter plot of degree distributions for LINs of (a) α, (b) β, (c) $\alpha + \beta$, and (d) $\alpha/\beta$ proteins, 20 of each class. Data points in each plot indicate $P(k)$ values for all the residues of 20 proteins of the LINs of the PCNs of respective class.

of a random graph is given by

$$C_{\text{rand}} = p = \frac{\langle k \rangle}{n_r}.$$  

(3.1)

Therefore, according to Eq. 3.1 when $C_{\text{rand}}/\langle k \rangle$ of random networks is plotted as a function of $n_r$ for varying sizes of the network, the data will show a linear nature with slope $-1$. The random controls of PCN show such behaviour as shown in Fig. 3.10 (The data pointed with an arrow).

Figure 3.10 also shows the change in $C$ with changing size of PCNs. Here the $C$ of PCNs do not change with the size of the network ($n_r$) which indicates that the PCNs, far from being random, show an indication of hierarchical structure [32] in them.
3.8 Discussion

Our results show that protein networks have "small-world" property regardless of their structural classification ($\alpha$, $\beta$, $\alpha+\beta$, and $\alpha/\beta$) and tertiary
structures (globular and fibrous proteins). Small world nature implies that PCNs have high degree of clustering [6–9] (compared to their random counterparts). Clustering, for protein structures, represents the extent/density of packing in the network. Thus higher order compaction, observed in proteins, is in agreement with what is expected from globular polymer chains in contrast to ‘randomly folded control polymers’.

Though small but definite differences exist between α and β classes, and fibrous and globular proteins. The size independence of the clustering coefficient in proteins indicates a departure from the random nature and an inherent modular organisation in the protein networks.

It is interesting to note that unlike other networks, PCNs while being small-world are not characterised by scale-free degree distribution. The absence of hubs in PCNs is understandable as there is a physical limit on the number of amino acids that can occupy the space within a certain distance around another amino acid. Such system-specific restrictions have been identified to be responsible for the emergence of different classes of networks with characteristic degree-distributions by Amaral et al. [93]. They observed that preferential attachment to vertices in many real scale-free networks [16] can be hindered by factors like ageing of the vertices (e.g. actors networks), cost of adding links to the vertices, or, the limited capacity of a vertex (e.g. airports network).