# Supporting Information:

## 1. Supporting methods:

### Construction of protein structure networks:

Each amino acid in the protein structure is represented as a node, and these nodes (amino acids) are connected by edges based on the strength of non-covalent interaction between the side chains of the two amino acid residues. The strength of interaction between two amino acid side chains is evaluated as a percentage given by (1, 2):

$I_{ij} = n_{ij} / sqrt(N_i * N_j)$

Where $I_{ij}$ = strength of interaction between residues i and j.

$n_{ij}$ = number of distinct side-chain atom pairs between i and j coming within a distance cut off of 4.5 Å units.

$N_i$ & $N_j$ = Normalization factors for residue types i and j. The normalization factors for the 20 different amino acid residues were evaluated from a non-redundant set of protein structures and are taken from the work of Kannan and Vishveshwara (1). This factor takes into account the differences in the sizes of the side chains of the different residue types and their propensity to make the maximum number of contacts with other amino acid residues in protein structures. The normalization factors for the 20 amino acid types are given in Table S1.

We then define a cutoff value $I_{min}$ and any residue pair with $I_{ij} > I_{min}$, is connected by an edge in the graph. Therefore, for a single protein, multiple protein structure graphs can be generated using different $I_{min}$s. This cutoff ($I_{min}$) is varied from 0% to 10% (very few nodes interact with a

value >10%), and as the interaction cutoff increases from 0% to 10%, the number of edges in the graphs decreases because, at higher cutoff, the number of nodes making the high level of interaction is less. Further, we ignore the connections between covalently linked backbone neighbors (i to i+1 and i to i-1 connections) so as to concentrate mainly on the long-distance non-covalent interactions stabilizing the protein tertiary structure. Additionally, the condition of symmetry is imposed such that if i is connected to j, then j is connected to i by definition. Thus, we quantify the interactions among the side chains of the residues and also construct amino acid-based protein structure graphs at varying strengths of interaction using this method. The protein structure networks have thus been constructed as a function of $I_{min}$ for a non-redundant set of 200 proteins and analyzed using the following parameters (2):

a) Degree distribution as a function of $I_{min}$: The degree of a node is the number of edges it makes in the graph. The distribution of this degree in a protein structure graph is a function of the $I_{min}$ at which it is constructed. The nature of the degree distribution is known to vary with $I_{min}$ in all globular proteins. We find that it varies from a bell-shaped normal distribution to a decay-type distribution with increasing $I_{min}$, as analyzed in a non-redundant dataset of 200 proteins (2).

b) Size of the largest cluster: A cluster is a set of connected amino acid residues in the graph. Clusters also vary as a function of the $I_{min}$. The size (in terms of the number of nodes) of the largest cluster in a graph is a well-known parameter used in network analysis and in percolation. We study this parameter as a function of $I_{min}$ and find that for all globular proteins with varying sizes, the size of the largest cluster follows a sigmoidal

profile with respect to the $I_{min}$ with a transition occurring within a narrow range of $I_{min}$ (2).

## 2. Supporting Text:

### a) The Erdõs-Rényi model:

The random graph originally defined by Erdõs and Rényi, which is the one of interest here, is a model consisting of 'N' nodes connected to one another by 'q' edges, wherein the probability of forming an edge between any two nodes is uniformly 'p'. Several properties of these graphs can be deduced through probabilistic arguments. For instance, the degree distribution, which is the distribution of the number of nodes having 'k' edges, is given by a binomial distribution. For large enough N, this distribution is given by the Poisson form:

y(k)=(pN)^k*exp(-pN)/k!

where y(k) is the probability of finding a node having 'k' edges. Another important property of the Erdõs-Rényi (ER) model is the existence of a percolation threshold characterized by a critical probability $p_c$ at which a giant cluster permeates the system. For the ER model, this critical probability is approximately 1/N; the giant cluster associated with this point scales as $N^{2/3}$. This random graph property can be viewed as an instance of critical phenomena where quantities of interest go to zero or infinity as simple power laws close to a transition point, and the exponents associated with these power-laws characterize the 'universality class' for the transition in question. (For further details, see ref. (3) and ref. (4))

### b) Random graph simulations:

We find that the protein structure graphs deviate from the ER model in significant ways. One of the main factors contributing to this aspect is the presence of the inherent backbone connections between the covalently linked neighboring amino acid residues in proteins. To further understand the effect of these covalent linkages, we carried out numerical simulations of a modified ER model, which accounts for the presence of a connected backbone. The details of the simulations are given in the methods section of the main paper. The degree distributions and largest cluster profiles of these modified ER random graphs are shown in Figs. S2 and S3, respectively. These random graphs show agreement with qualitative features of the ER model. However, as can be seen from Fig. S2, the deviations from the exact Poisson form are significant in these modified ER model. In fact, the peak height and position of the degree distribution curve for each size and probability shows a distinct shift compared to the corresponding best fit Poisson curve. The trend shown by this deviation is similar to that of the protein data only in some regions of parameter space. In particular, the decrease in peak height in the random graph case, when compared with the Poisson curve, is only seen for larger graphs (N = 400), while it is observed for all protein graphs. While this simulation by no means explains all the deviations exhibited by the proteins from ER behavior, it serves to show that the modified ER models with backbone constraints accounts for some of the properties shown by the protein network.

**c) Thermophile/mesophile analysis:**

To further understand the factors enhancing the stability of thermophilic proteins, we studied the number of hydrogen bonds, salt bridges and di-sulphide bridges in the thermophilic proteins and their mesophilic counterparts. The statistics is presented in Table S2.

**3. Supporting References:**

1. Kannan N, Vishveshwara S (1999) Identification of side-chain clusters in protein structures by a graph spectral method. J Mol Biol **292:** 441-464.

2. Brinda K V, Vishveshwara S (2005) A network representation of protein structures: implications to protein stability. Biophysical J **89:** 4159–4170.

3. R Albert R, Barabasi A-L (2002) Statistical mechanics of complex networks. Rev Mod Phys **74:** 47-97.

4. Stauffer D (1985) Introduction to Percolation Theory (Taylor & Francis Ltd, London).

**4. Supporting Figures and Legends:**

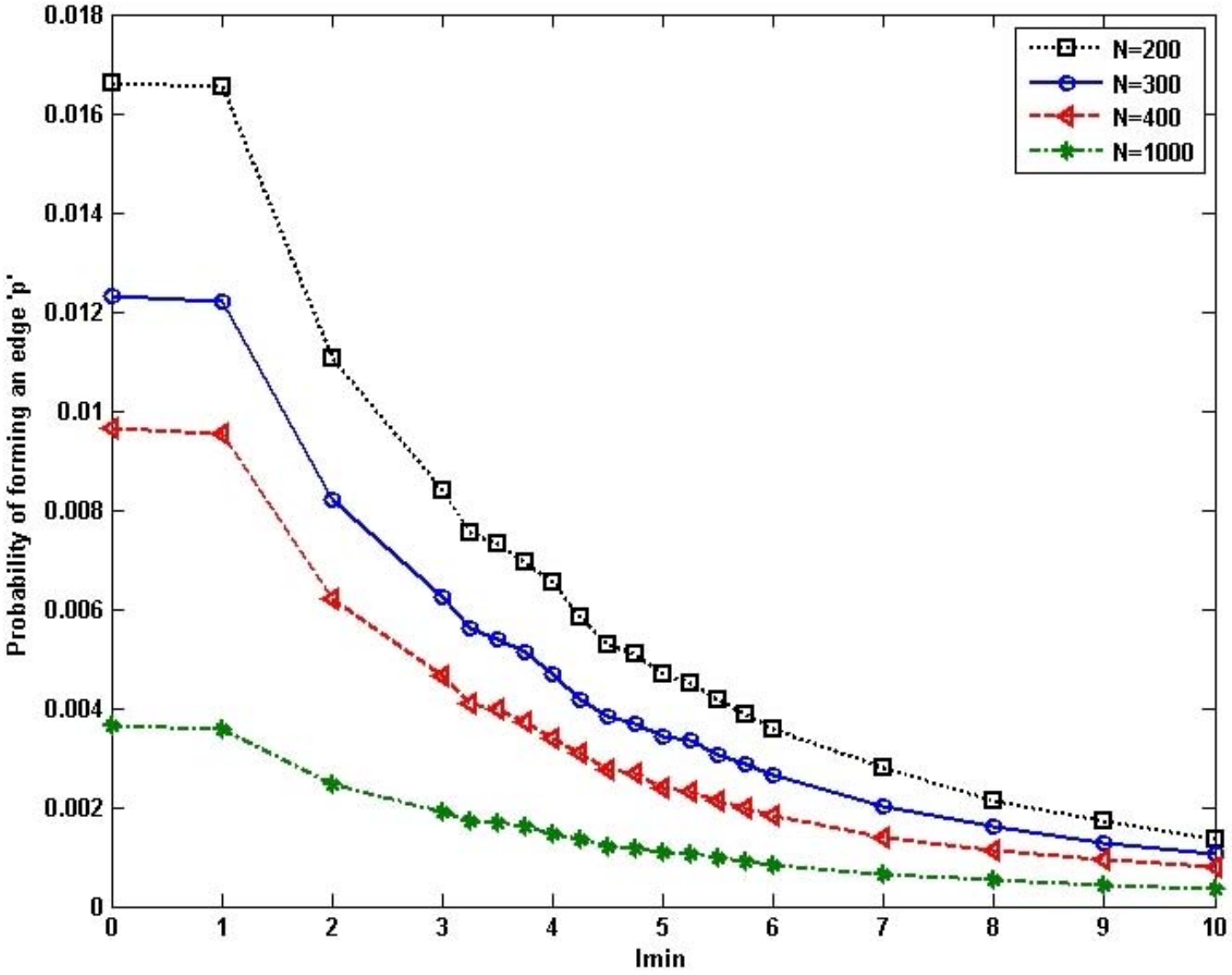


**Figure S1:** Correlation of $I_{min}$ to 'p' for proteins (averaged over 20 proteins) of different sizes (N), where 'p' is the probability of forming an edge and $I_{min}$ is the minimum interaction strength cut off chosen for edge formation. The plots for different protein sizes are shown in different colors. Note that the probability of forming an edge 'p' is dependent on the size of the protein, whereas the $I_{min}$ representation, by itself, is independent of the protein size.

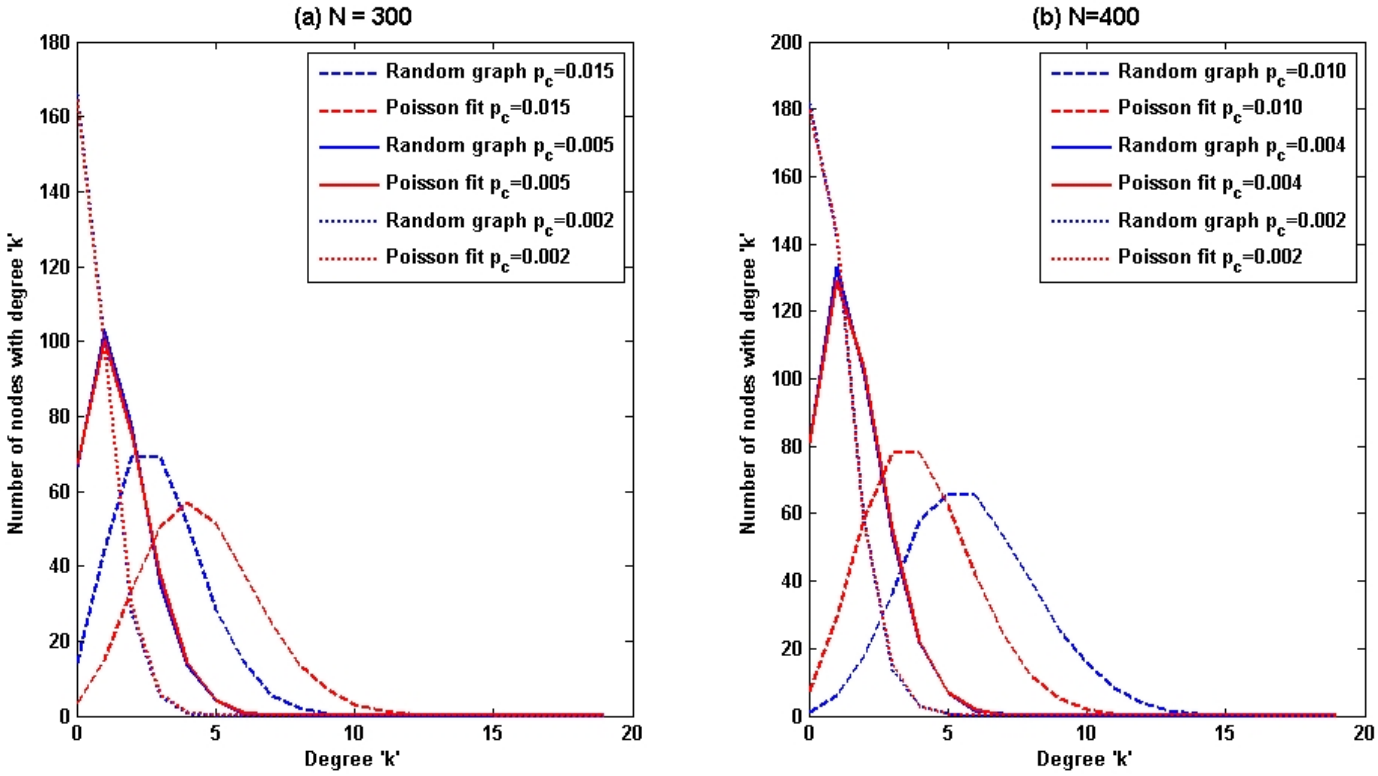


**Figure S2:** Degree distribution plots for random graphs of sizes (a) N=300; (b) N=400, constructed as mentioned in the supporting text. The degree distributions are shown in blue and the corresponding Poisson fits are shown in red. The plots for three different $p_c$ values, 0.015, 0.005 and 0.002 (for N=300) and 0.010, 0.004 and 0.002 (for N=400), are shown in dashed, line and dotted respectively.

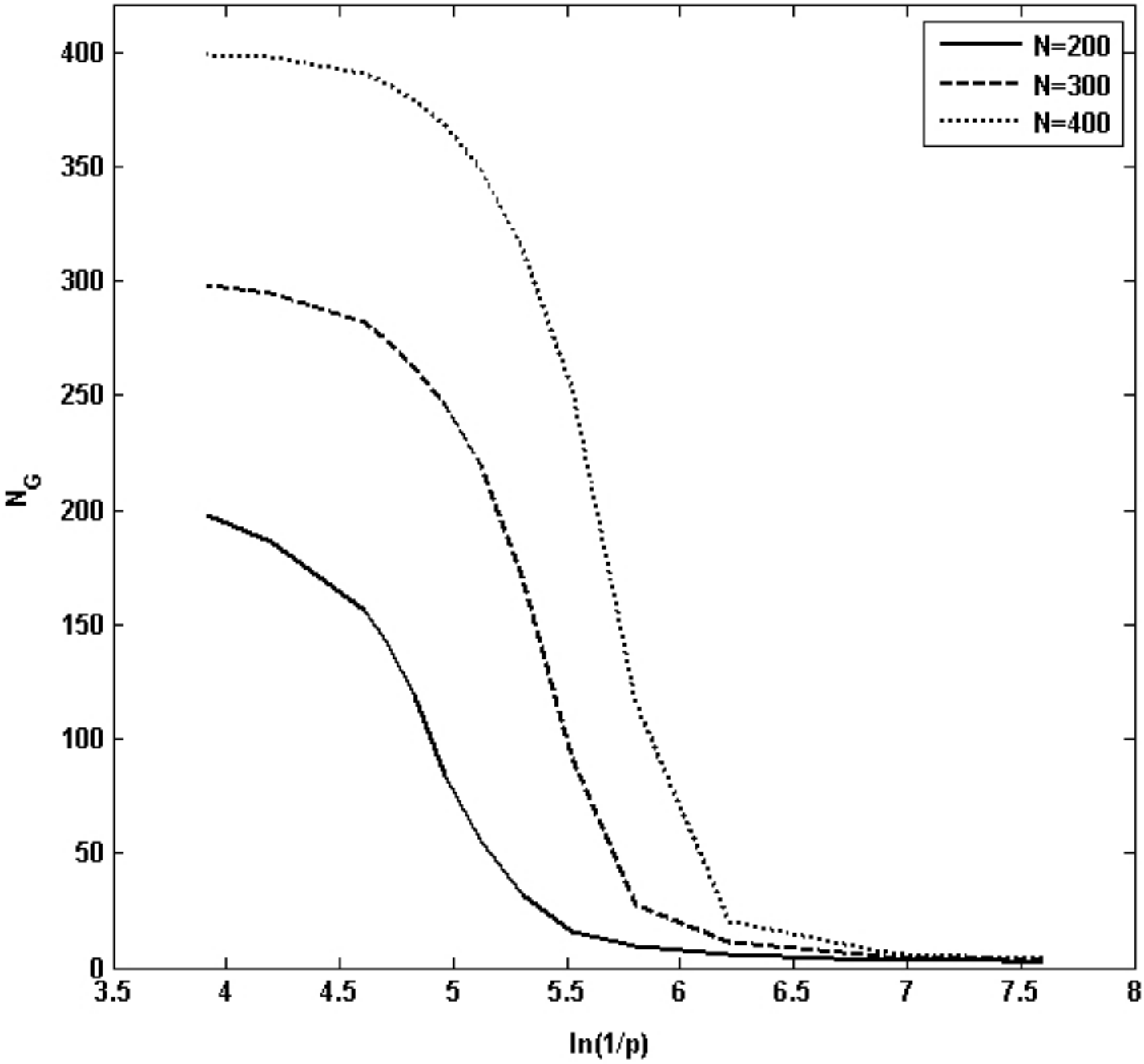


**Figure S3:** Plot of the size of the giant cluster ($N_G$) Vs ln(1/p) for random graphs of sizes 200, 300 and 400, shown as line, dashed and dotted respectively.

## 5. Supporting Tables:

### Table S1: Normalization values for the 20 amino acid residue types from reference 1

| Residue type | Normalization value |
|---|---|
| Ala | 55.7551 |
| Arg | 93.7891 |
| Asn | 73.4097 |
| Asp | 75.1507 |
| Cys | 54.9528 |
| Gln | 78.1301 |
| Glu | 78.8288 |
| Gly | 47.3129 |
| His | 83.7357 |
| Ile | 67.9452 |
| Leu | 72.2517 |
| Lys | 69.6096 |
| Met | 69.2569 |
| Phe | 93.3082 |
| Pro | 51.3310 |
| Ser | 61.3946 |
| Thr | 63.7075 |
| Trp | 106.703 |
| Tyr | 100.719 |
| Val | 62.3673 |

## Table S2: Thermophile-Mesophile Comparison of Interactions

| Families | | [a]Total Number of Disulphide Bonds | [b]Total Number of Salt Bridges | [c]Total number of Sidechain and Mainchain-Sidechain Hydrogen Bonds |
|---|---|---|---|---|
| **Protein Name** | **PDB ID (Thermophile, Mesophile)** | | | |
| Carboxypeptidase | 1obr<br>2ctc | 2<br>1 | 17<br>10 | 140<br>126 |
| Triose Phosphate Isomerase | 1btm<br>7tim | 0<br>0 | 11<br>13 | 69<br>81 |
| Reductase | 1ebd<br>1lvl | 0<br>1 | 14<br>10 | 102<br>98 |
| Signal Recognition Particle | 1ffh<br>1fts | 0<br>0 | 25<br>7 | 104<br>47 |
| Lactate Dehydrogenase | 1ldn<br>1ldm | 0<br>0 | 12<br>7 | 71<br>80 |
| Phosphoglycerate Kinase | 1php<br>3pgk | 0<br>0 | 19<br>2 | 118<br>30 |
| Neutral Protease | 1thl<br>1npc | 0<br>0 | 13<br>15 | 121<br>127 |
| Subtilisin | 1thm<br>1st3 | 0<br>0 | 9<br>6 | 115<br>102 |
| Glyceraldehyde-3-Posphate Dehydrogenase | 1vc2 (replaces 1gd1)<br>1gad | 0<br>0 | 11<br>16 | 120<br>124 |
| Xylanase | 1yna<br>1xyn | 1<br>0 | 4<br>1 | 81<br>49 |
| Adenylate Kinase | 1zip<br>1ak2 | 0<br>1 | 12<br>12 | 67<br>60 |
| Phosphofructo Kinase | 3pfk<br>2pfk | 0<br>0 | 10<br>8 | 89<br>92 |

[a] The disulphide bond information was obtained from the PDB file descriptions
[b] The salt bridge interactions are calculated using VMD with O-N distance <= 3.2 Å
[c] The Hydrogen bonds are calculated using Hbplus v3.15 with D-A distance <= 3.9 Å and DHA angle >= 120°

Note: Each row represents a specific family of proteins, where the first line represents the protein from a thermophile and the second line represents the protein from a mesophile.