Protein residue networks from a local search perspective

Susan Khor
Department of Computer Science
Memorial University of Newfoundland, St John's NL Canada

slc.khor@gmail.com

Abstract

Proteins have been abstracted as a network of interacting amino acids and much attention has been paid to the small-world property of such networks, which we call protein residue networks (PRNs). Hitherto, a global search strategy such as Breath-First Search (BFS) is commonly used to measure the average path length of PRNs. We propose that a local search strategy is more appropriate because the inverse relationship between clustering and average path length in a local search better fits the notion that amino acids get closer to each other as a protein becomes more compact. This inverse relationship is also observed in data from a molecular dynamics (MD) simulation of a protein unfolding. To study local search on PRNs, we devised a greedy local search algorithm called EDS and compared the characteristics of BFS paths with EDS paths. While they are different in terms of variation in path length, search cost and link usage, they exhibit similarities in terms of hierarchy and centrality. We argue that the differences are preferable as they make EDS paths a better model of intra-protein communication. The similarities are also preferable as they imply the transferability of existing methods based on BFS centrality. Clustering coupled with strong transitivity helps to keep EDS paths short on PRNs by creating a store of potential short-cut edges. The ready availability of PRN edges that can act as short-cuts helps EDS avoid backtracking. The number of short-cuts scales linearly with protein size. Short-cut edges are enriched with short-range contacts, see higher usage (are more central), have stronger local clustering but weaker local community structure, and effect larger EDS path dilation. Throughout the paper, network statistics for PRNs from a MD simulation are reported to support our findings, and to observe how the network statistics change as a protein folds.

1. Introduction

The genetic system is a heritable mechanism for producing proteins which are the building-blocks of life. In general, proteins attain their structure necessary for functioning via the different possible physical and chemical interactions amongst their amino acid molecules within the constraints of their environment. One way of understanding protein structure is through their *contact maps*, which abstracts away the physical and chemical details and puts the spotlight on contacts between amino acid molecules of a protein. More formally, the contact map of a protein is the adjacency matrix **A** of a graph G representing the protein as a set of nodes V and a set of edges E. Typically, each node in V represents an amino acid, and an edge is placed between a node-pair if they satisfy certain conditions, e.g. if they are within an acceptable Euclidean distance from each other. We call such a graph (constructed in accordance with section 2.1) a *Protein Residue Network* (PRN). Other criteria have been used to construct the network of interacting amino acids within proteins. To avoid confusion, we will refer to the general class of networks

induced by protein contact maps as *residue interaction networks* (RINs), of which our PRNs are a specific form.

Pursuant to the introduction of a model of the small-world phenomenon in social networks [1], networks induced by protein contact maps were classified as small-world networks [2]. This study is performed on globular proteins, although both fibrous and membrane proteins are qualitatively small-world networks also [3, 4]. A *small-world network* (SWN) combines the order inherent in regular graphs, with the arbitrary connectivity of pure random graphs to supply the short-cut edges or long-range connections. The second half of this definition has undergone refinement. Briefly, for a local search algorithm to be able to find a short path from point *a* to point *b* in a SWN, the short-cut edges need not be long-range or even random, but multi-scaled [5, 6, 7]. More formally, a graph with N nodes and average degree K is identified as having SWN structure if: (i) its clustering coefficient C is significantly larger than the clustering coefficient of a comparable Erdos-Renyi (ER) random graph with $C_{\mathrm{ER}} \sim K / N$; and (ii) its characteristic path-length L approaches that of a comparable ER random graph with $L_{\mathrm{ER}} \sim \ln(N) / \ln(K)$. For constant K, the latter property implies that L increases logarithmically with N.

The clustering coefficient C reflects the probability that two unique nodes u and v which are directly connected to a third other node w, are themselves connected in the network forming a triangle. Typically, the clustering coefficient for a network with N nodes is $C = \frac{1}{N}\sum_{i}^{N} C_i$ where $C_i = \frac{2e_i}{k_i(k_i - 1)}$ is the clustering coefficient of a node i with degree k_i and e_i is the number of links amongst i's k_i direct neighbor nodes [1]. Links in an ER graph are independent of each other, so $C_{ER} = p = \frac{2M}{N(N-1)} = \frac{K}{(N-1)}$ where p is the probability of connecting two nodes in the ER graph, M is the number of links and $K = \frac{2M}{N}$. Typically, L is the average length of paths between all unique node-pairs in a network: $L = \frac{2}{N(N-1)}\sum_{i<j}^{N} \lambda(i, j)$ where $\lambda(i, j)$ is the length of a path (number of edges in a path) from node i to node j found through a global search such as breadth-first search (BFS). All networks considered in this paper are simple graphs.

The description of proteins as small-world networks is intuitive since the combination of order and randomness in a SWN parallels the coexistence of the highly ordered alpha helical and beta sheet secondary structures with random coils in protein molecules. Further, the short characteristic path length of a SWN appeals directly to the need for rapid communication between distantly located sites in a protein. Such efficient long-range intra-protein communication underpins allosteric interactions between cooperative binding sites which are crucial for proteins to be functional [10, 11]. Proteins may have more than one binding site, which may be (un)-occupied in concert. More commonly, activation of a site

regulates the binding receptivity of other sites on the same protein. The ability for such long-range interactions has also been observed in non-allosteric proteins, and is believed to be a fundamental phenomenon of all globular proteins [10, 12]. Indeed, the short characteristic path length of a SWN is a widely mentioned and frequently applied topological feature of RINs [2, 13, 14, 15, 16].

With a few exceptions [3, 17, 18], there is less discussion about the role clustering plays in proteins. This may be because clustering in RINs is seen as an inevitable consequence of protein packing in 3D space. However, the presence of clustering in RINs needs to be explained *and related to the characteristic path length* of RINs if we are to fully understand why proteins have SWN structure and to fully exploit a network model of proteins. We propose that the role of clustering in RINs can be better understood in the light of a *greedy local search*. In a greedy local search, information available to the algorithm to decide the next step in a path is confined to the neighbourhood within a small radius of the current node, and the algorithm moves to the neighbouring node that is closest to the target node. As such a greedy local search is *directed* and in our case, as in Kleinberg's [5, 6], by proximity to the target. Where necessary, L_G refers to the characteristic path length of a RIN found with a global search strategy, and L_W to one obtained with a local search strategy.

The majority of previous research on RINs has implicitly defined characteristic path length and other network statistics derived from path length such as betweenness centrality and closeness centrality, in terms of shortest paths found via a global search strategy such as BFS on unweighted RINs or Dijkstra's algorithm on weighted RINs [16, 18]. An exception is [19] which uses a Markov random walk on RINs. There are three issues with using global search on RINs (since PRNs are un-weighted, we will discuss global search in terms of BFS).

First, "[vibrational] energy flow in globular proteins resembles transport on a percolation cluster with channels through which energy flows easily and dead-end regions where energy flow stalls" [12]. Vibrational energy is due to stretching and twisting of molecular bonds, and is also generated to compensate for loss in translational and rotational degrees of freedom when two molecules associate [11]. The transport of such energy through specific pathways is believed to be a mechanism underlying allosteric interactions in proteins. Anisotropic energy flow has been experimentally observed to occur efficiently between two allosterically linked binding sites: FAI in subdomain IB and Sudlow site I in subdomain IIA, in the multi-domain BSA protein [20].

In a BFS on PRNs, there is very little variation in the lengths of paths, i.e. all nodes are easily reached from every other node (section 3.1). Thus BFS models a communication strategy where there is little specificity in inter-nodal communication. But protein sites are not created equal: certain sites are more actively involved in protein activity and are more evolutionarily conserved than others [21]. While there is a fair amount of redundant pathways, specific communication pathways within a protein molecule

have been traced, and some of these pathways exist to slow down communication as a way to absorb or localize the after effects of undesirable perturbations to maintain protein stability [14].

Second, a BFS progresses by radiating outward from a source node in all possible directions. Thus, BFS is directionless, and unlike energy flow in proteins which is anisotropic and sub-diffusive [12]. The lack of direction increases the volume of space explored by BFS, i.e. the number of sites or nodes visited during a search. While the average length of shortest paths found with BFS increases only logarithmically with network size (section 3.1), the cost of BFS in terms of the average number of unique nodes visited during the search increases linearly with network size (section 3.2). The linear search cost of BFS is known [8 p. 44]. In contrast, a greedy local search is more focused and less diffusive. Hence it is less costly and it turns out that for PRNs, it is possible for a Euclidean distance directed greedy local search with backtrack (our EDS algorithm is described in section 2.5) to produce paths with average length that increases logarithmically with network size, and at a cost that also increases logarithmically with network size. The small-world property of PRNs is preserved with local search, and as such PRNs are *navigable* small-world networks.

Third, the clustering coefficient of folded proteins is significantly larger than C_{ER}, and a high level of clustering forms a barrier to short inter-nodal path-lengths L_G. The direct relationship between C and L_G follows directly from $L_{ER} \sim \ln(N) / \ln(K)$. ER graphs with typical low p have little to no clustering and are therefore locally tree-like. The average degree K then approximates the branching factor of a search tree. By introducing cycles into the search tree, i.e. increasing clustering, some of the branches now loop back to a previous tree level. In other words, clustering reduces the effective branching factor, or effective K. Consequently, L_G becomes larger than L_{ER}.

However, if short inter-nodal distances are important for protein functionality, why should proteins take on a highly clustered conformation that prevents them from having shorter pathways? This contradiction does not arise with a local view of search where suitable clustering is an enabler, not an impediment, to shorter path lengths. The inverse relationship between C and L_W is more attuned to what happens when a protein folds or unfolds. Molecular dynamics (MD) simulation on the 2EZN protein (section 2.7) confirms the presence of a significant gap in C values between the set of PRNs for configurations at equilibrium, and the set of PRNs for non-equilibrium configurations. MD simulations of the 2EZN protein also reveal that C generally increases (decreases) and L generally decreases (increases) as the protein folds (unfolds) (Fig. 1).

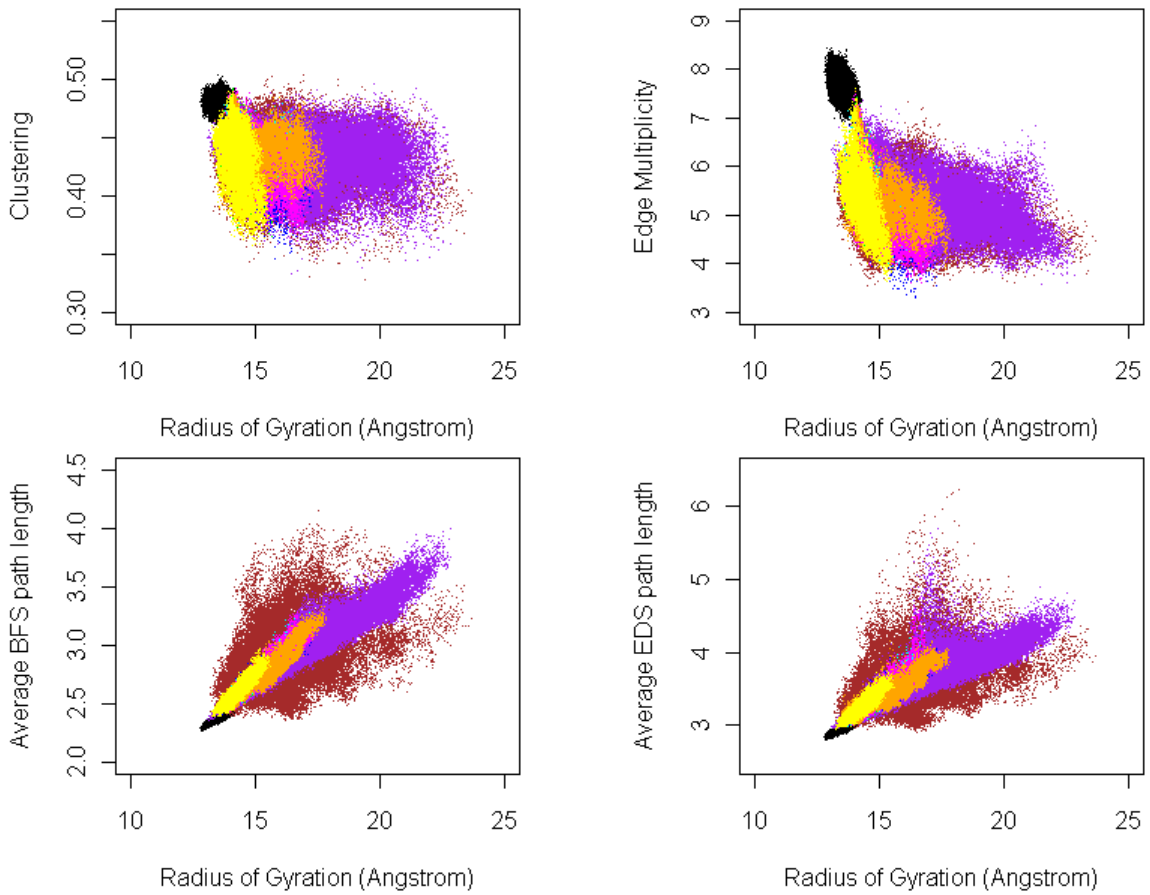


Fig. 1 Change in clustering, edge multiplicity and average path length as the 2EZN protein folds under MD simulation. Clustering and edge multiplicity (section 2.3) increases, while average path length decreases. The scatter plots show the respective network statistic for each snapshot in a run. Black points denote the native state statistics (section 2.7).

2. Materials and Method

Since we are proposing the use of local search to investigate PRNs, this study focuses quite heavily on the differences between global search implemented as BFS (Breath-First Search) and local search implemented as EDS (section 2.5). The differences are measured primarily in terms of path length (section 3.1) and search cost (section 3.2), but other differences stemming from path length differences, as well as some unexpected similarities are reported in section 3. To observe the impact of clustering on local search, a null model in the form of the MGEO networks, which are identical to PRNs in several respects including the 3D coordinates of nodes, is designed (section 2.2).

In all figures, the average (mean), standard deviation, median, first quartile and third quartile values for a variable are indicated by "avg", "sd", "mid", "Q1" and "Q3" respectively. A "BFS" prefix denotes that the statistic is measured with a set of paths found through Breadth-First Search. An "EDS" prefix denotes that the statistic is measured with a set of paths found with the EDS algorithm. Unless stated otherwise, significance tests are made with either R's t.test or Wilcox.test, in paired form where

applicable, and a p-value < 0.05 is required for significance. Eigenvalues and eigenvectors were calculated with GNU Octave 3.8.1 on Linux. In figures where the x-axis is labeled "Nodes", read as N (number of nodes).

2.1 PRN construction

Define a Protein Residue Networks (PRN) as a simple undirected graph $G = (V, E)$ with $|V| = N$ and $|E| = M$. The PRNs are constructed from the PDB coordinates files [22] with side-chain considerations following the method in [17]. Each amino acid (residue) is represented by a node and two nodes u and v are linked *iff* $| u - v | \geq 2$ and their interaction strength I_{uv} is ≥ 50%. $I_{uv} = \frac{n_{uv} \times 100}{\sqrt{N_u \times N_v}}$ where n_{uv} is the number of distinct atom-pairs (i, j) such that i is an atom of residue u, j is an atom of residue v and the Euclidean distance between atoms i and j, $ed(i, j)$ is ≤ 7.5 Å. All the atoms of a residue, including those of the backbone, are considered when calculating n_{uv} (this departs from [17] where only the side-chain atoms are used to calculate I_{uv}). N_u and N_v are normalization values by residue type. They are obtained from Table 1 in [23]. As in [17], peptide bonds are excluded, i.e. links are prohibited between nodes i and i+1 (the nodes of a PRN are labeled in the same order as they appear in the PDB file).

The threshold value-pair of 50% and 7.5 Å was chosen after some initial experiments to permit most intra-protein domain-based interactions identified in 3did [24] to be edges in most of the 2000 initial PRNs sampled at random from the list of proteins appearing in the 3did catalog (see Appendix A for overview of protein selection for network construction). A pair of PFAM domains are deemed able to interact with each other if they have at least five estimated contacts (hydrogen bonds, electrostatic or van der Waals interaction) between them [24]. The 3did links are the intra- or inter-chain residue-residue interactions between contacting PFAM domains of a protein as listed in the 3did catalog.

Each PRN satisfies the following criteria: (i) there are no missing atoms in the PDB coordinate file, (ii) all domain-based interactions cataloged in 3did are represented as links, (iii) the PRN is a single connected component, and (iv) there is a power-law relation between the number of nodes N and link density. The last condition reflects the sparseness of PRNs. The link density of PRNs is $\frac{2M}{(N-1)(N-2)} \sim KN^{-1}$ and average node degree K is constant for sparse graphs.

The PRNs (and MGEO networks from section 2.2) are reduced to the same set of 166 proteins after screening out outliers to permit pairwise comparison. Compared to the BFS paths, the EDS paths show much greater variation in length which section 3.1 argues is a more realistic model for proteins. However, the extremely large values skew path-length statistics and spills over to other path-length related statistics such as link usage and betweenness centrality. Therefore the networks were screened to limit the effect of

outliers by excluding networks whose average BFS or EDS path exceeds $ln(max(N)) = 7.482119$ where $max(N)$ is the number of nodes in the largest PRN in the set of 204 PRNs.

The edges or links of a PRN are partitioned into two sets according to their sequence distance. A link connecting nodes u and v is long-range if u and v are more than 10 residues apart on the protein sequence, i.e. $|u - v| > 10$ [25]. *Long-range* edges (*LE*) represent interactions between residues which are far apart on the protein sequence but close to each other in the tertiary structure. *Short-range* edges (*SE*) represent interactions between residues that are close to each other in the primary and the tertiary structures.

2.2 MGEO network construction

The MGEO networks are constructed in a similar manner as random geometric networks. Random geometric networks with random coordinates were proposed in [26] as null models for RINs. However, the MGEO networks use the PDB coordinates of the Cα atoms. To increase the number of single component MGEO networks, the nodes are first connected to form a "backbone" as follows: link u, the most recent node to join the backbone, to a node v which is not already in the backbone and is closest in Euclidean distance to v subject to $|u - v| \geq 2$. The backbone starts with a single node chosen uniformly at random. Due to the way the backbone is constructed, there is no guarantee that the backbone will connect all the nodes in a network. For example, in a network comprising five nodes, if the backbone is built with edge (1, 5) then (5, 3), it is impossible for nodes 2 or 4 to join the backbone. Next, the 3did edges are added. Finally, m edges are added so that a MGEO network will have identical link density as its PRN. $m = M - B - D$ where M and D are respectively the number of links and 3did edges of a PRN, and B is the number of links used to construct the backbone. There may be overlap between backbone links and 3did edges.

The m edges comprise the shortest (Euclidean distance between the Cα atoms) links such that $|u - v| \geq 2$ and that satisfy the *skip* condition, which defines how many shortest links to ignore between two link inclusion events. The skip condition is introduced to control clustering. Skipping a larger number of shortest links results in a MGEO network with lower C. Thus, the MGEO4 networks, where only every fifth shortest link is included, are significantly less clustered than the MGEO2 networks which include only every third shortest link. The contact maps of two PRNs and their MGEO4 networks are shown in Fig. 2.

2.3 Structural properties of PRNs and MGEO networks

The average degree for PRNs is constant with N (Fig. 3a), and PRNs have Gaussian degree distribution which can be explained by the excluded volume argument [9]. We confirm that the MGEO networks

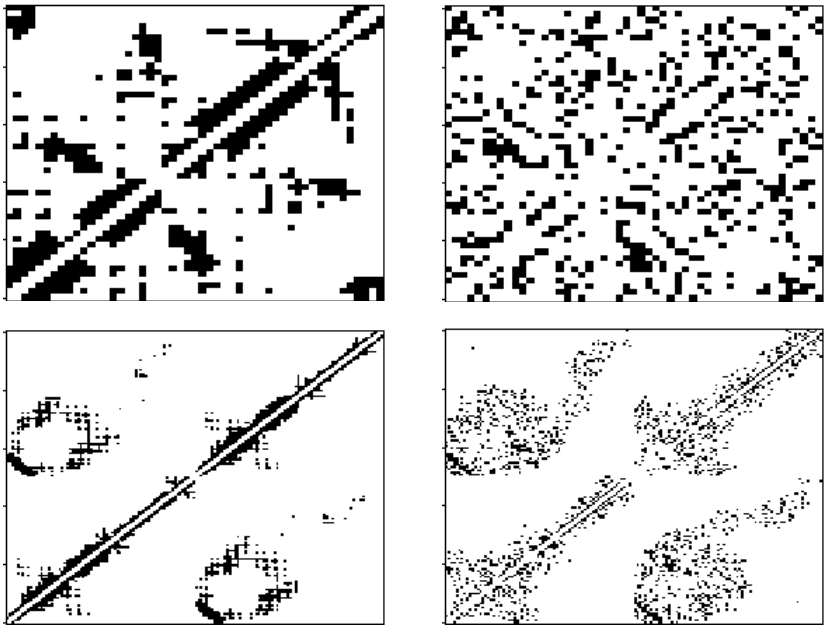


Fig. 2 Contact maps of two PRNs (left) and their respective MGEO4 networks (right). The top pair is for protein 1B19 which has 51 nodes and 282 edges. The bottom pair is 2ADL which has 144 nodes and 904 edges. A dark cell denotes an edge. A white cell denotes a non-edge.

have identical average node degree as their respective PRNs, and that their degree distributions are also Gaussian.

As expected, both the MGEO networks have significantly smaller clustering coefficients than their respective PRNs (Fig. 3b). The MGEO4 networks are also significantly less clustered than the MGEO2 networks. These differences in one-vertex clustering extend to dyadic or two-vertex clustering as demonstrated by the significant differences in *edge multiplicity* (Figs. 3c & 3d).

Edge multiplicity (*EM*) was introduced in [27] and used to quantify the organization of triangles in a network [28]. In its simplified form (details about the degree class of the endpoints of edges are ignored here), the multiplicity of an edge *e*, is the number of distinct triangles that passes through *e* (Fig. 4). $EM >> 1$ means triangles are packed onto shared edges, or alternatively, edges participate in many triangles. $EM \leq 1$ denotes that the triangles are disjoint. Networks with large *EM* values have strong transitivity, while those with small *EM* values have weak transitivity [27]. When the clustering coefficient (which is susceptible to the influence of large node degree variations in scale-free networks) of a network does not reflect the transitivity strength of a network, transitivity strength quantified by *EM* is a better indicator of the percolation properties of a network than its clustering coefficient [29]. *EM* is a measure of edge embeddedness, and its use here is equivalent to the number of direct neighbours common to the endpoints of an edge.

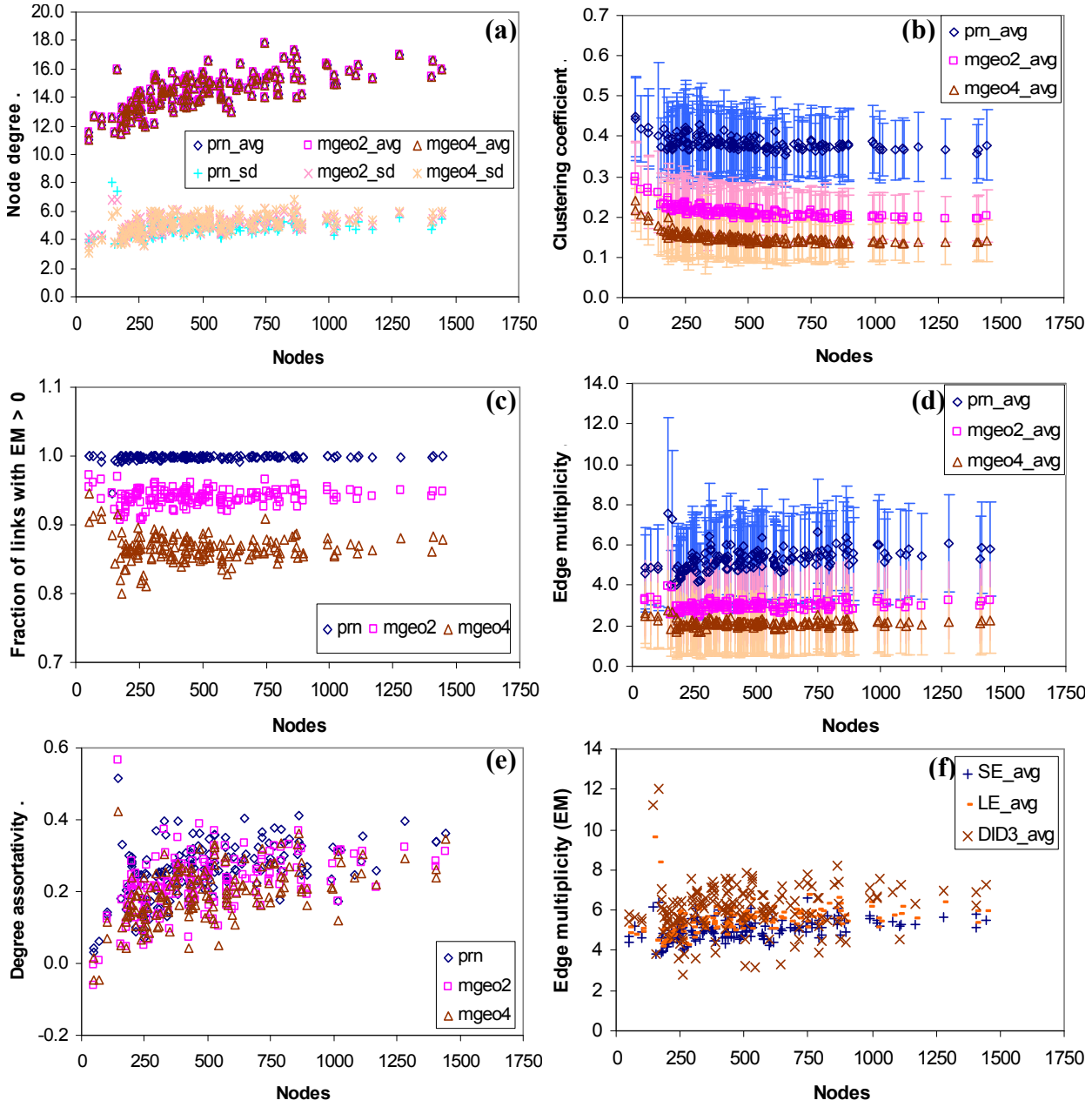


Fig. 3 Structural characteristics of PRNs and MGEO networks. (a) PRNs and MGEO networks have Gaussian degree distributions that peak at almost identical average node degrees. **(b)** The clustering coefficients (*C*) of PRNs are significantly larger than the *C* values of MGEO networks. MGEO2 networks have significantly higher clustering than MGEO 4 networks. **(c)** The fraction of PRNs links with EM > 0 is almost 1.0. A link with EM > 0 belongs to at least one triangle. The fraction of edges with EM > 0 is significantly smaller in the MGEO networks. **(d)** Edge multiplicity averaged over the edges of a PRN is significantly larger edge multiplicity averaged over thee dges of a MGEO network. **(e)** Both PRNs and MGEO networks are mildly positively associative by node degree. **(f)** Edge multiplicity by edge type. In the PRN networks, both long-range links (*LE_avg*) and 3did links (*DID3_avg*) have significantly larger average edge multiplicity than short-range links (*SE_avg*). Error bars in (b) and (d) denote the standard deviation.

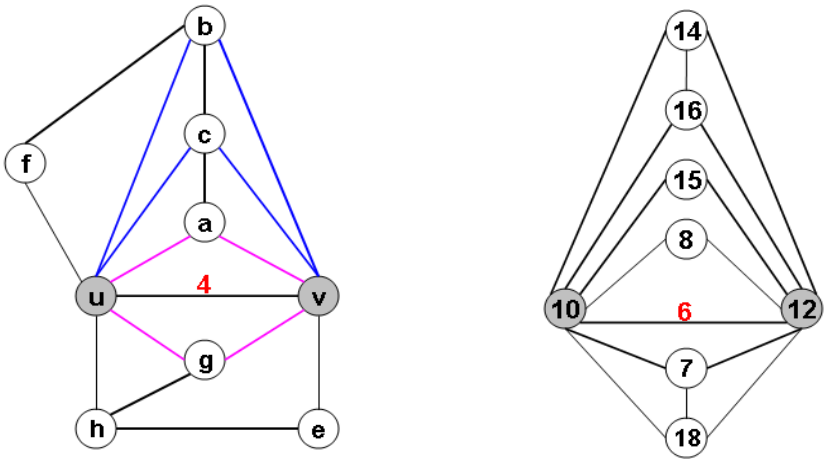


Fig. 4 Edge multiplicity (*EM*). EM of an edge is the number of triangles the edge completes [27]. In the diagram on the left, EM of edge (*u*, *v*) is 4 since nodes *u* and *v* have four direct or first neighbors in common: *b*, *c*, *a* and *g*. The blue and pink edges trace out two diamonds (cycles of length four) that are the result of triangles sharing (*u*, *v*). The diagram on the right is extracted from the 2FAC PRN (only the common first neighbors of nodes 10 and 12, and the links between them are shown). EM of edge (10, 12) is 6.

Almost all links in a PRN make up a leg of at least one triangle; the fraction is lower in MGEO networks but still above 80% (Fig. 3c). PRNs also have significantly larger *EM* values than MGEO networks (Fig. 3d). A large clustering coefficient signals an abundance of triangles and when triangles share edges, they stick together and have the potential to form longer cycles. Large *EM* values signal an abundance of diamond motifs. The abundance of triangle and diamond motifs induced by the strong transitivity in PRNs makes it feasible as done in [17] to utilize higher order concepts of local organization such as k-cliques and communities (overlapping cliques) to distinguish decoy structures from native ones. Ref. [30] reports a linear relationship between triangle and diamond motifs in their RINs, and interprets the presence of these local motifs as providing alternative pathways to ensure connectivity as links are made and destroyed when proteins undergo fluctuations at equilibrium. However, their RINs are constructed differently and the diamond motif is also defined differently. A diamond motif in [30] is a chordless four node cycle. By this definition, nodes *b*, *u*, *g*, and *v* in Fig. 4 (left) does not form a diamond motif because of the edge between *u* and *v*.

The multiplicity of an edge is limited by the degree of its endpoint nodes. Thus, networks whose nodes selectively link by degree, i.e. links are more likely between nodes with similar degrees than between nodes with dissimilar degrees, are more conducive to strong transitivity. Nodes in PRNs are mildly positively correlated by degree (Fig. 3e). Assortative mixing of nodes by degree has been reported in other RINs as well [31, 32].

RINs comprising only long-range links have significantly smaller *C* values than complete RINs [25]. In fact, much of the clustering in RINs has been attributed to short-range links [33]. But even though long-range links (*LE*) are less transitive than short-range links (*SE*) when considered separately, when *LE*

and *SE* are combine in a network, their interaction results in long-range links having significantly larger *EM* values than short-range links (Fig. 3f). Nonetheless, a strong linear correlation between sequence distance and *EM* was not observed. Edge transitivity gains significance when the properties of short-cut edges are investigated in section 3.6.

2.4 Expansion property of PRNs and MGEO4 networks

The PRNs have maximum inter-residue Euclidean distances averaged at 15.4Å (std. dev. = 1.3224), which is twice that normally considered in pure (no side-chain consideration) C_α-C_α or C_β-C_β contact maps. This raises the concern that the PRNs are not preserving topological cavities which are proxies for protein binding sites. Ref [34] examined the expansion factor of RINs and observed that most RINs, especially those of multi-domain proteins with more than 240 nodes, are not homogeneous networks (class I) but belong to the class of networks that exhibits modularity (class II). Class I networks are good expanders. An expander graph is sparse (bounded mean node degree) and its nodes are well-connected such that it is difficult to disconnect the graph, i.e. the cuts need to be large (see Appendix C for a more formal treatment). We can infer from this description that Class I networks will have poor modularity. Modularity describes an organizational structure where clusters of nodes that are more interconnected with each other than with other network nodes outside the cluster exist. Modular structure is an important feature for RINs to have for two complementary reasons: (i) modularity is a natural consequence of the architecture of larger multi-domain proteins, and (ii) the topological cavities or regions of lower connectivity between the clusters correspond to protein binding sites [34]. Using the spectral scaling method in [34], we determined that all but two of our PRNs belong to class II and are therefore modular and cavity preserving in principle (Fig. 5). The other two PRNs belong to class IV. For comparison, we constructed pure C_α-C_α contact maps with a 7.5 Å cutoff. Twelve of these networks belong to Class IV. This means that the PRN construction method presented in this paper allows much larger contact cutoffs, whilst maintaining the essential features of protein structure. Due to their more random construction, the MGEO4 networks are expected to have better expansion which they do. Fourteen MGEO4 networks belong to Class IV, and those MGEO4 networks that belong to Class II show smaller deviation from Class I behavior (Fig. 5d).

2.5 A Euclidean distance directed local search algorithm (EDS)

In [5, 6], Kleinberg describes a local search algorithm that does greedy routing. In this greedy local search algorithm, information used to decide the next move is confined to the neighbourhood within a small radius of the current node, and the search moves to a neighbouring node that is closest to the target node. As such a greedy local search has *directionality*, i.e. is anisotropic.

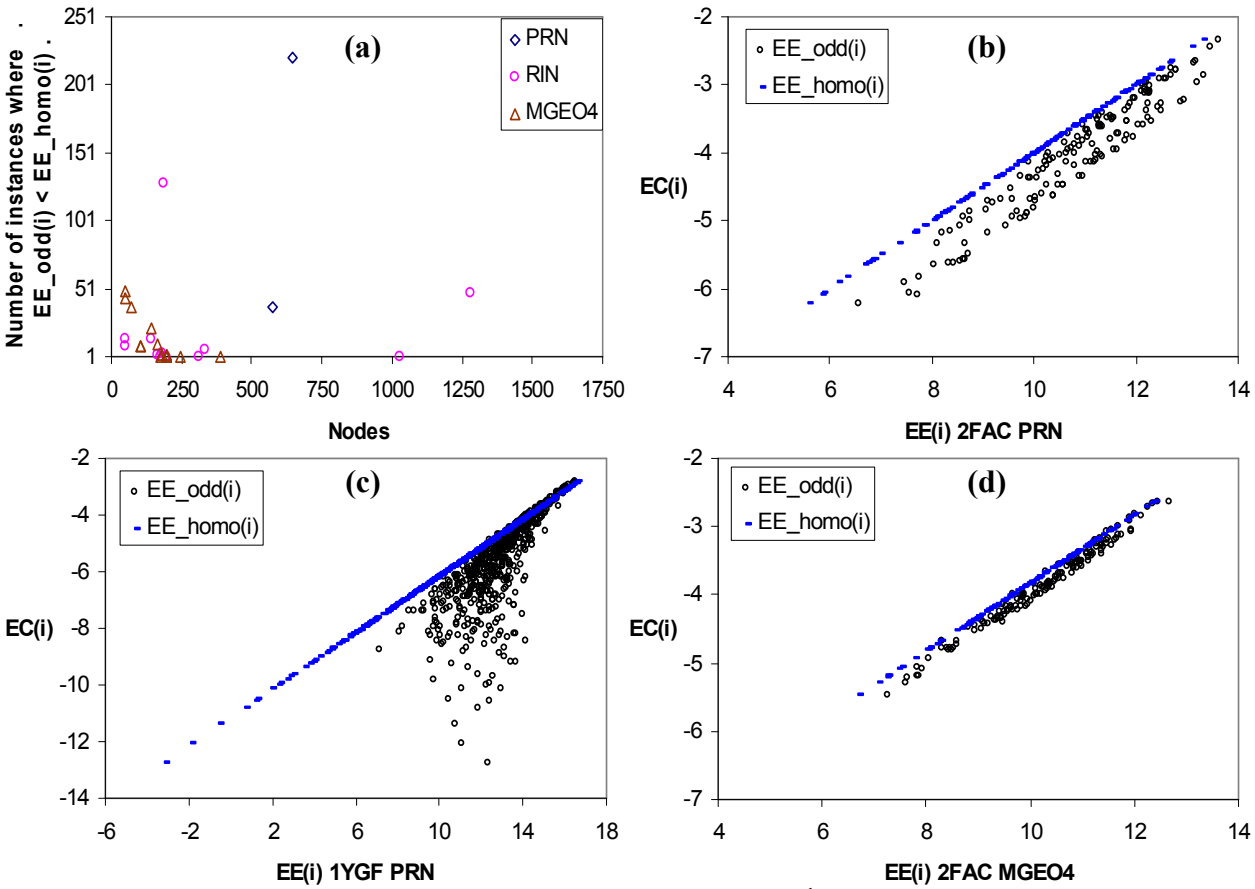


Fig. 5 Spectral scaling [34] results for the 166 PRNs. EC(i) denotes the i^{th} component of the principal eigenvector. EE(i) denotes the subgraph centrality measure for node i. EE_odd(i) denotes the odd-subgraph centrality measure for node i. When $\lambda_1 >> \lambda_2$, EE_odd(i) = EE_homo(i). The x- and y- axes in (b - d) are log base 2. **(a)** The number of nodes in a network where EE_odd(i) < EE_homo(i). A network with one or more such nodes (but not all nodes), belong to Class IV. The number of PRNs that belong to class IV is 2, compared with 12 for RINs, and 14 for MGEO4 networks. For this plot, RINs are protein contact maps built as pure C_α-C_α with 7.5 Å cutoff. **(b)** Spectral scaling results for a Class II network: the 2FAC PRN. All the EE_odd(i) points are to the right (>) of the EE_homo(i) points. **(c)** Spectral scaling results for one of the two PRNs (1YGF) that belong to class IV. There are EE_odd(i) points to the left and right of the line passing through the EE_homo(i) points. **(d)** The 2FAC MGEO4 network is also Class II, but compared with the 2FAC PRN (b), its EE_odd(i) points lie closer to the EE_homo(i) points.

Similar to Kleinberg's algorithm, the EDS algorithm does greedy routing based on proximity (Euclidean distance) to a target node. However, EDS differs in two main ways: it keeps a memory of all the nodes visited and enquired so far in the current search, and can therefore backtrack and re-route itself to another more promising path midway through the search. The information used is still local, but this information expands over time as more nodes are visited and their direct neighbours queried for their proximity to the target node; in other words, the search radius increases as the search progresses. To keep the search local, information gathered during the search for a path is completely forgotten once the path is found. Technically an EDS path is a walk since an EDS path may retrace edges and in the process revisit nodes.

Starting with the source node, EDS performs the steps outlined below for each node appended to a path until the target node is found. The main steps of the EDS algorithm to find a path from s to d are:

1. Get $N(x)$, the direct neighbors of the most recently visited node x. Initially, x is the source node s.
2. For each n in $N(x)$, compute the Euclidean distance between n and the target node d. (Proximity information for an n to d may already be computed in a previous iteration due to network clustering.)
3. If n is the target node d, stop searching and return path.
4. Otherwise, add the new proximity information to memory, i.e. all distance information gathered so far for this search.
5. Sort nodes in memory by proximity to find the next node to visit, y.
6. Move to y, which is an as yet unvisited node closest to the target. If necessary (when y is not directly reachable from the current node), backtrack (retrace the current path but also look at the immediate neighborhood of nodes retraced to find a bridge to y, i.e. a node neighboring y).
7. $x := y$. Go to 1.

The backtrack strategy in Step 6 can lead to unnecessary backtracking. However this apparent inefficiency ensures that the graph an EDS path traces is a tree, and increases the search space for an edge to y. A backtrack stops as soon as an edge to y is found. The search tree begins with a single node, s. In each iteration, the EDS algorithm adds to the existing search tree a single edge with a node not already in the search tree. The only way for EDS to revisit a node is by tracing the edges of the existing search tree. The EDS search is conducted on a finite connected graph and EDS terminates at d (The efficiency of EDS search depends on the structure of the network).

EDS paths run along the edges of the network. Edges may appear more than once in an EDS path if backtracking occurs. Edges in an EDS path may be classified as short-cut, backtrack, short-cut and back-track or neither short-cut nor backtrack. An EDS path may have zero or more short-cut and/or backtrack edges. The number of edges in an EDS path is its *length*. Fig. 6 depicts three EDS paths of length four from the 2FAC PRN. The *cost* of an EDS path is the number of unique nodes stored in memory for the search, and is bounded from above by the union of the direct neighbors of all nodes that appear on the path. The sequence of events to construct the rightmost EDS path in Fig. 6 is worked out in Appendix B.

Unlike BFS which guarantees $\lambda(i, j) = \lambda(j, i)$ by definition (undirected graph), an average of 53.50% (std. dev. 6.28%) of EDS path-pairs are not length invariant, i.e. $\lambda(i, j) \neq \lambda(j, i)$. We do not consider this a disadvantage. In fact, if network topology plays a role in determining paths and their lengths, and the topology is not homogeneous everywhere, then $\lambda(i, j) \neq \lambda(j, i)$ is to be expected. For results in this paper, BFS and EDS searches are run in both directions for each unique ordered node-pair. Average path length

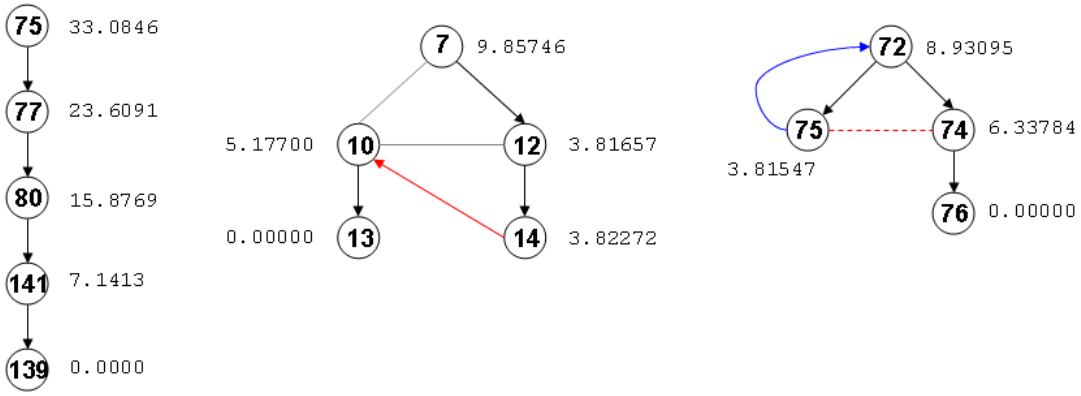


Fig. 6 Three EDS paths of length four found in the 2FAC PRN. PRN edges are undirected. The arrowheads are used to show the direction the edges are traversed in the respective paths. The leftmost path is ⟨75, 77, 80, 141, 139⟩ which is a straight-forward path. The EDS path in the middle is ⟨7, 12, 14, 10, 13⟩ and it has a *short-cut* edge indicated in red. Without edge (14, 10), EDS would have to backtrack to node 12 to get to node 10. A short-cut edge completes a *navigational cycle*, which in this instance is the triangle comprising nodes 14, 12, and 10. The rightmost EDS path is ⟨72, 75, 72, 74, 76⟩ and it has a *backtrack* edge indicated in blue. If edge (75, 74) existed, it would be a short-cut edge and the backtrack would be avoided. A non-existent short-cut edge marks a *navigational hole*, which in this instance is the two-edge path comprising nodes 75, 72 and 74. Short-cut and backtrack edges are defined more formally in the text. The real numbers beside each node are the Euclidean distances to the respective target nodes, which need not decrease monotonically as an EDS progresses. For example, nodes 12 and 14 are much closer to the target node 13 than node 10. The construction of the rightmost path is described in detail in Appendix B.

of a network is then redefined as: $L = \frac{1}{N(N-1)} \sum_{i \neq j}^{N} \lambda(i, j)$ where $\lambda(i, j)$ is the length (number of edges) of a path) from node i to node j. The total number of paths sampled by both BFS and EDS is $N(N\text{-}1)$. For BFS sampling two possibly different paths per node-pair increases the data available to compute path related statistics such as betweenness centrality. EDS paths that are identical either way, and/or identical to BFS paths exists. On average, the set of nodes along a BFS path from u to v is only 44% (std. dev. 6.03%) similar with the set of nodes along an EDS path from u to v. This similarity drops significantly to 41% (std. dev. 4.79%) for MGEO4 networks.

Short-cuts (SC) and navigational cycles: Define $G_{s,d} = (V', E')$ as the sub-graph of a PRN induced by V', the set of nodes on the EDS path from s to d in the PRN. E' is the subset of edges in the PRN that have both their endpoints in V'. While PRNs are not directed graphs, the edges in $G_{s,d}$ are oriented in the direction they are traversed by the EDS path from s to d. A bi-directional link denotes an edge that has been re-traced in a backtrack. The search tree $T_{s,d}$ of an EDS path from s to d is induced by all oriented edges in $G_{s,d}$ and is rooted on s. $T_{s,d}$ spans the nodes of $G_{s,d}$. Let $L^T(x)$ be a non-negative integer denoting the level of node x in a search tree T, and $\mathcal{N}(x)$ be the set of nodes in the connected graph (PRN) adjacent to x. $L^T(s) = 0$. Every n in $\mathcal{N}(x)$ which has not been previously assigned a level is assigned $L^T(x) + 1$. The search tree in Fig. 7 has 6 levels.

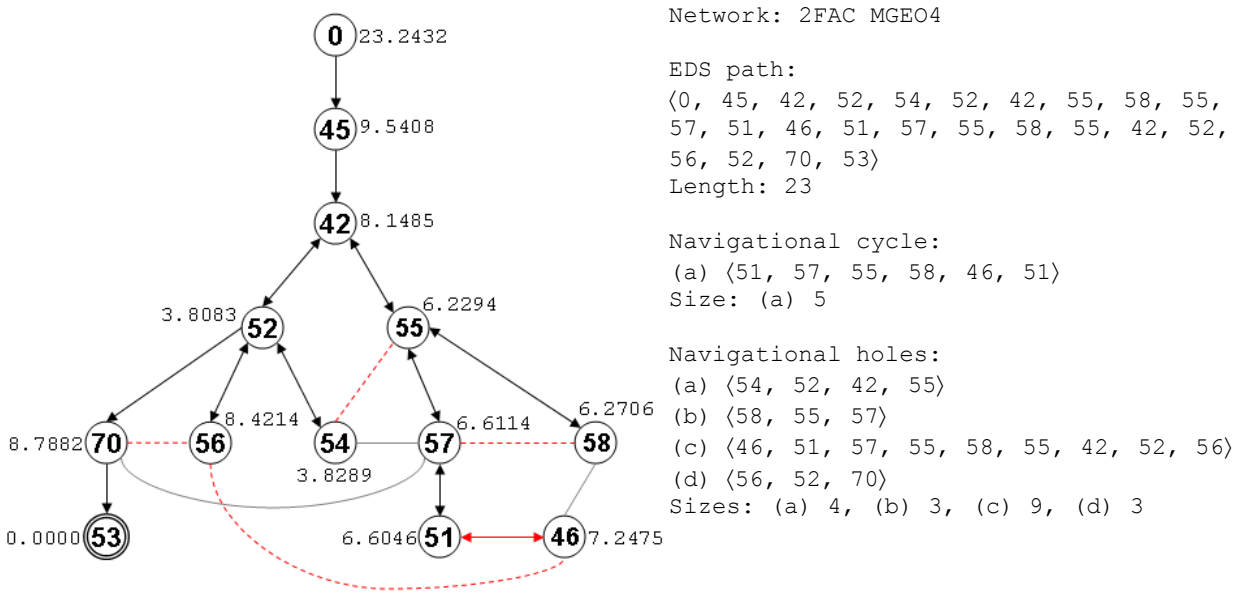


Fig. 7 The sub-graph $G_{0,53}$ for the EDS path from 0 to 53 on the 2FAC MGEO4 network. All lines except the red dashed ones are edges of $G_{0,53}$. Edges are oriented in the direction they are traversed by this EDS path. Bi-directional edges are backtrack edges. Un-oriented edges are not traversed but exist in the underlying MGEO4 network. The search tree for this EDS path is induced by all the oriented edges. The short-cut edge (51, 46) is marked in red. The red dashed lines indicate the non-existent short-cuts (NESC), i.e. edges that would be short-cuts if they existed in the underlying network, with everything else unchanged. The real numbers beside each node are the Euclidean distances to the target node 53. For comparison, the EDS path from 0 to 53 on the 2FAC PRN is a straight-forward path of length four ⟨0, 2, 71, 51, 53⟩.

Given the sub-graph G' and the search tree T for an EDS path from s to d: an edge (u, v) in T is a short-cut if $L^T(v) \leq L^T(u)$ and there exists an edge (w, v) in G' but not in T. Since $L^T(v) \leq L^T(u)$, there must be at least one other node in T besides u that is adjacent to v; let w be one of these nodes and W be the set of all such w nodes. To identify the specific w node, EDS retraces its step from u until it finds the *first* (x, v) edge where $x \in W$. The path $\langle u, \ldots, x, v\rangle$ together with the short-cut edge (u, v) forms a *navigational cycle*. Hence, short-cut edges complete navigational cycles, and are made possible by cycles in the underlying network.

The sub-graph in Fig. 7 contains one navigational cycle compromising nodes 51, 57, 55, 58, 46 which is completed by the short-cut edge (51, 46). The edge (46, 51) is backtracked by this EDS search to get to node 56. Suppose, with everything else unchanged, that edge (55, 46) exists in the sub-graph but not in the search tree in Fig. 7. Then the EDS path is unchanged and (51, 46) still forms a short-cut. But $L^T(46) = 4$, $W = \{55, 58\}$, and the navigational cycle associated with (51, 46) in this EDS search is now circumnavigated by a shorter path: ⟨51, 57, 55, 46, 51⟩.

The size of a navigational cycle is the number of edges it contains. The length of a short-cut is the size of the navigational cycle it closes less one. Since a short-cut edge may be part of one or more navigational cycles in different EDS searches, the length of a short-cut edge is context sensitive. Due to

differences in clustering and transitivity, short-cut edges are significantly more abundant and significantly shorter in PRNs than in MGEO networks (Figs. 9a & 9b).

Non-existent short-cuts (NESC) and navigational holes: If edge (51, 46) is removed from the sub-graph in Fig. 7, all things being equal, EDS would need to retrace its steps to node 58 to reach node 46 from node 51, and the path ⟨51, 57, 55, 58, 46⟩ would form a *navigational hole*. Essentially, backtracks travel along navigational holes until a direct neighbour of the node EDS needs to move to is met. The sub-graph in Fig. 7 has four navigational holes which could be closed by the red dashed edges if these edges exist in the underlying MGEO network. However, since these edges do not exist, they are called *non-existent short-cuts* (NESC).

A NESC is found when EDS needs to move from u to v but (u, v) is not an edge in the network. However, if we assume that (u, v) exists, then (u, v) is a short-cut as defined previously, and a special w node x can be located as described previously. But since (u, v) does not actually exist, EDS adds (x, v) to the search tree instead of (u, v), and the path $\langle u, \ldots, x, v\rangle$ forms a navigational hole. Both navigational cycles and holes can contain smaller cycles in them due to the backtracking that was used to find them. E.g.: the largest navigational hole in Fig. 7 contains a cycle ⟨55, 58, 55⟩.

The size of a navigational hole is the number of edges in an EDS path that circumnavigates it plus one. The length of a non-existent short-cut is the size of the navigational hole it would close if it existed less one. The length of a NESC also gives the search depth for an EDS search. As with short-cut edges, the length of NESCs may be context dependent. *NESC*s are significantly more abundant and significantly longer in MGEO networks than in PRNs (Figs. 9c & 9d).

2.6 Random short-cuts (*rsc*)

As part of this study, we investigate the properties of short-cut edges and their impact on PRN network statistics. These findings are compared against a set of edges chosen uniformly at random without replacement from all edges of a PRN such that there is a replacement edge for each short-cut. Let *SC* be the set of short-cuts found by EDS for a PRN, and *RSC* be a random short-cut set associated with the PRN. Then *RSC* is a random subset of the PRN's edges such that $|RSC| = |SC|$. Five such random subsets are generated, and they are denoted *rsc1*... *rsc5* in the figures. Thus, a PRN network without its short-cuts, PRN\SC, has the same number of edges as the PRN without its random short-cuts, PRN\RSC.

2.7 Molecular Dynamics simulation (MD) dataset

The MD dataset is obtained from the Dynameomics project [35, 36, 37]. *il*mm (in lucem Molecular Mechanics) is used to simulate the native and unfolding dynamics of the 2EZN protein which has 101 amino acids and is comprised mainly of beta strands interspersed with helices and loops. There are nine

MD runs in this dataset, each with a different number of snapshots (Table 1). The shorter unfolding runs were made to sample early unfolding events more thoroughly. Snapshots are taken at intervals of 1ps, except for the first 2 ns of the unfolding runs where snapshots are taken at 0.2 ps.

A PRN is constructed for every snapshot following the method described in section 2.1. The snapshots of a run capture the positions of all protein atoms and the solvent (water) atoms in the MD box at a point in time. Excluding the positions of the solvent atoms, gives one a file of x, y and z coordinates for the protein atoms from which PRNs can be constructed. The results we report are produced by averaging over all snapshots in a run. Network statistics for PRNs from the MD dataset are plotted against the *Radius of Gyration* (also obtained from the MD dataset), which is a measure of the compactness of a protein. Radius of Gyration (Rg) is a commonly used reaction coordinate to monitor progress during protein folding simulations. Protein folding is associated with more order and tighter packing, i.e. smaller Rg values.

Table 1 Definition of labels and colors to identify the different 2EZN MD runs.

Run type	Run label	Color in plots	Number of snapshots
Native dynamics (298K)	6250	Black	51,000
Non-native dynamics (498K)	6251	Brown	51,000
	6252	Purple	68,845
	6253	Blue	2,000
	6254	Cyan	2,000
	6255	Green	2,000
	6256	Magenta	10,000
	6257	Orange	10,000
	6258	Yellow	10,000

3. Results

3.1 Path-length

Both BFS and EDS search strategies found short paths on both PRNs and MGEO networks. The average length of both BFS and EDS paths increased logarithmically with increase in N (Fig. 8a), but BFS paths are significantly shorter than EDS paths. This is expected given that backtracking occurs in EDS paths and that BFS by definition should produce the shortest paths. The EDS paths are significantly more varied in length than the BFS paths (Fig. 8b). Large variation in path lengths is more congruent with the notion described in section 1 that intra-protein communication pathways are also varied in length. Fast, specific and reliable communication between some sites is crucial, but also some pathways exist to slow down communication as a way to absorb or localize the after effects of undesirable perturbations to maintain protein stability [14].

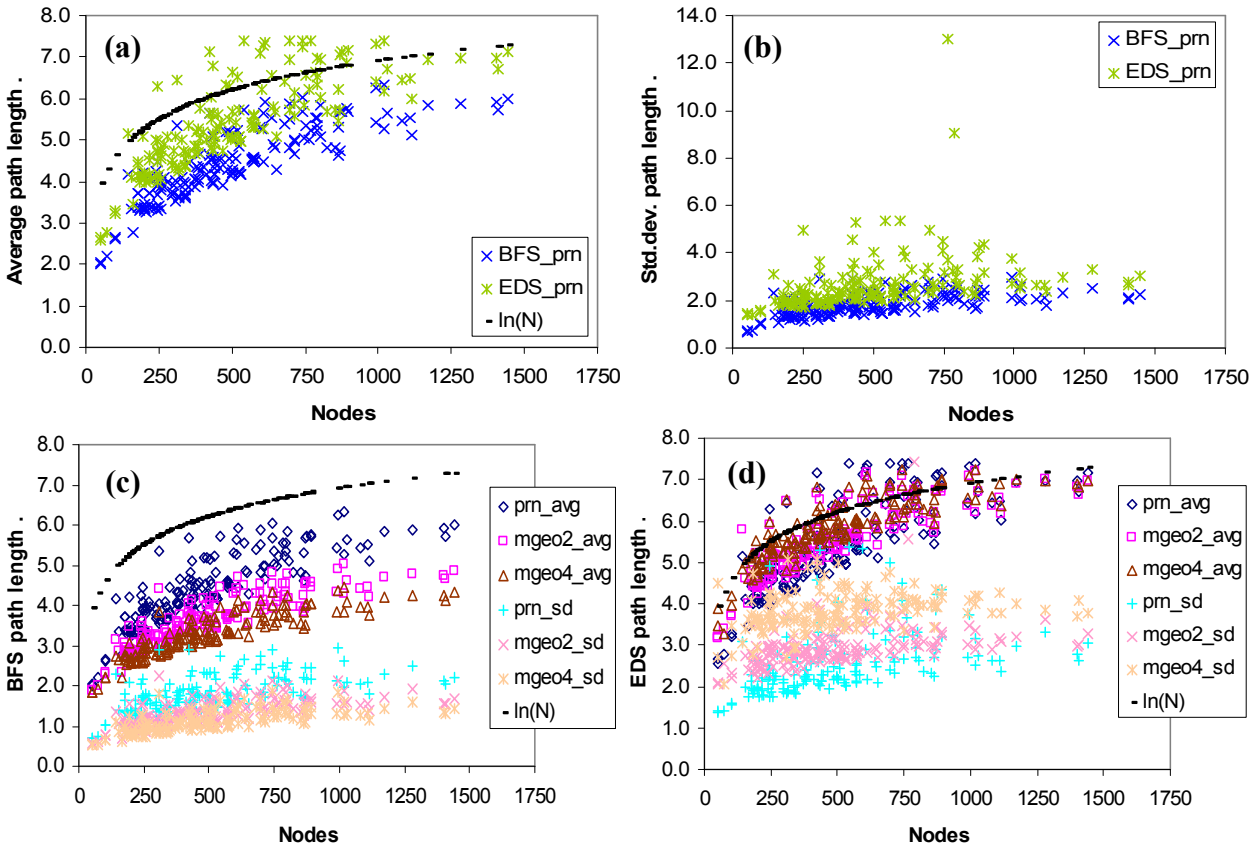


Fig. 8 BFS and EDS path length. (a) Both the average BFS path length (*BFS_prn*), and the average EDS path length (*EDS_prn*) on PRNs increase logarithmically with the number of nodes *N*. However, *EDS_prn* is significantly longer than *BFS_prn*. **(b)** The standard deviation of EDS path lengths (*EDS_prn*) is significantly larger than the standard deviation of BFS path lengths (*BFS_prn*) on PRNs. Therefore, EDS paths on PRNs are more varied in length than BFS paths on PRNs. **(c)** The more highly clustered PRNs have significantly longer average BFS path length (*prn_avg*) than the less highly clustered MGEO networks. The average BFS path length on MGEO2 and MGEO4 networks, denoted *mgeo2_avg* and *mgeo4_avg* respectively are significantly smaller than *prn_avg*. **(d)** Average EDS path length on PRNs denoted *prn_avg* is significantly shorter than both average EDS path length on MGEO2 networks (*mgeo2_avg*), and average EDS path length on MGEO4 networks (*mgeo4_avg*). Thus, average EDS path length decreases with increase in clustering.

MGEO networks have significantly shorter BFS paths than PRNs (Fig. 8c). In contrast, MGEO networks have significantly longer the EDS paths than PRNs (Fig. 8d). Thus, at the link density levels of the PRNs, the significantly weaker clustering of the MGEO networks reduces global path length (L_G) but increases local path length (L_W). This result suggests a topological reason for the high levels of clustering in PRNs. Clustering is a barrier to short paths found through global search (BFS), but becomes a facilitator for short paths found with local search (EDS). Given the importance of short intra-protein communication pathways, striving towards (folding) and maintaining a configuration with high clustering makes sense, but only from a local search perspective. The benefit of high clustering in a protein's native state, in terms of short intra-protein paths, may also act as a topological barrier preventing its unfolding.

Thus we observe a significant gap in C and L values between native protein dynamics and protein unfolding (Fig. 1).

The effect clustering has on BFS path length follows from the relationship: $L_{ER} \sim \ln(N) / \ln(K)$. Random graphs have little to no clustering and are locally tree-like. The average degree K in ER graphs then approximates the branching factor of a search tree. By introducing cycles into the search tree, some of the branches now loop back to a previous tree level. In other words, clustering reduces the effective branching factor, or effective K. Consequently, L_G increases. But an increasing average path-length concomitant with an increasing clustering coefficient does not fit the picture in Fig. 1.

An explanation for the effect clustering has on EDS path length follows. Unlike the BFS strategy of branch first then prune (when the target node is found), EDS prunes first (or branches conservatively) then if necessary (with the help of backtracking) figures out how to branch further. Define the number of edges retraced to branch onto a more promising path as the *depth* of a search (search depth is also the length of a NESC, section 2.5). The search depth for PRNs is significantly shallower than the search depth for MGEO networks (Fig. 9d), which means that on average, EDS retraces fewer edges when searching PRNs than MGEO networks. EDS also does significantly less backtracking when searching PRNs than MGEO networks (Fig. 9f). Thus, PRNs have significantly shorter average EDS path length than MGEO networks. The shallower search depths and fewer backtracking paths in PRNs are a consequence of their high levels of clustering and strong transitivity (Figs. 3b & 3d), which enrich PRNs with potential short-cut edges. In contrast, the lower levels of clustering and weaker transitivity of MGEO networks create significantly more navigational holes (Fig. 9c).

The short-cut edges which are significantly more abundant in PRNs than in MGEO networks (Fig. 9a) reduce the need for and the extent of backtracking, which in turn keeps the length of EDS paths in check. So while cycles pose a problem for global search (BFS) and increase L_G, they help local search (EDS) to decrease L_W. In the MD simulations on 2EZN, as the protein becomes more compact (Fig. 10), generally: (a) the number of short-cut edges increases, (b) the number of non-existent short-cuts decreases, (c) the fraction of EDS containing at least one short-cut increases, and (d) the fraction of EDS paths doing at least one backtrack decreases. From Fig. 1, we know that clustering and edge multiplicity increases while average path length decreases as the protein becomes more compact in these runs. Taken together, the observations from the MD simulations on 2EZN are aligned with the results of this section which are made for the 166 native PRNs – that clustering (and strong transitivity) facilitates navigation by increasing short-cuts and reducing backtracking.

The linear relationship between the number of short-cuts and the number of nodes in a PRN as ~2N reported in Fig. 9a is not an artifact of finding EDS paths both ways. When the analysis is repeated with

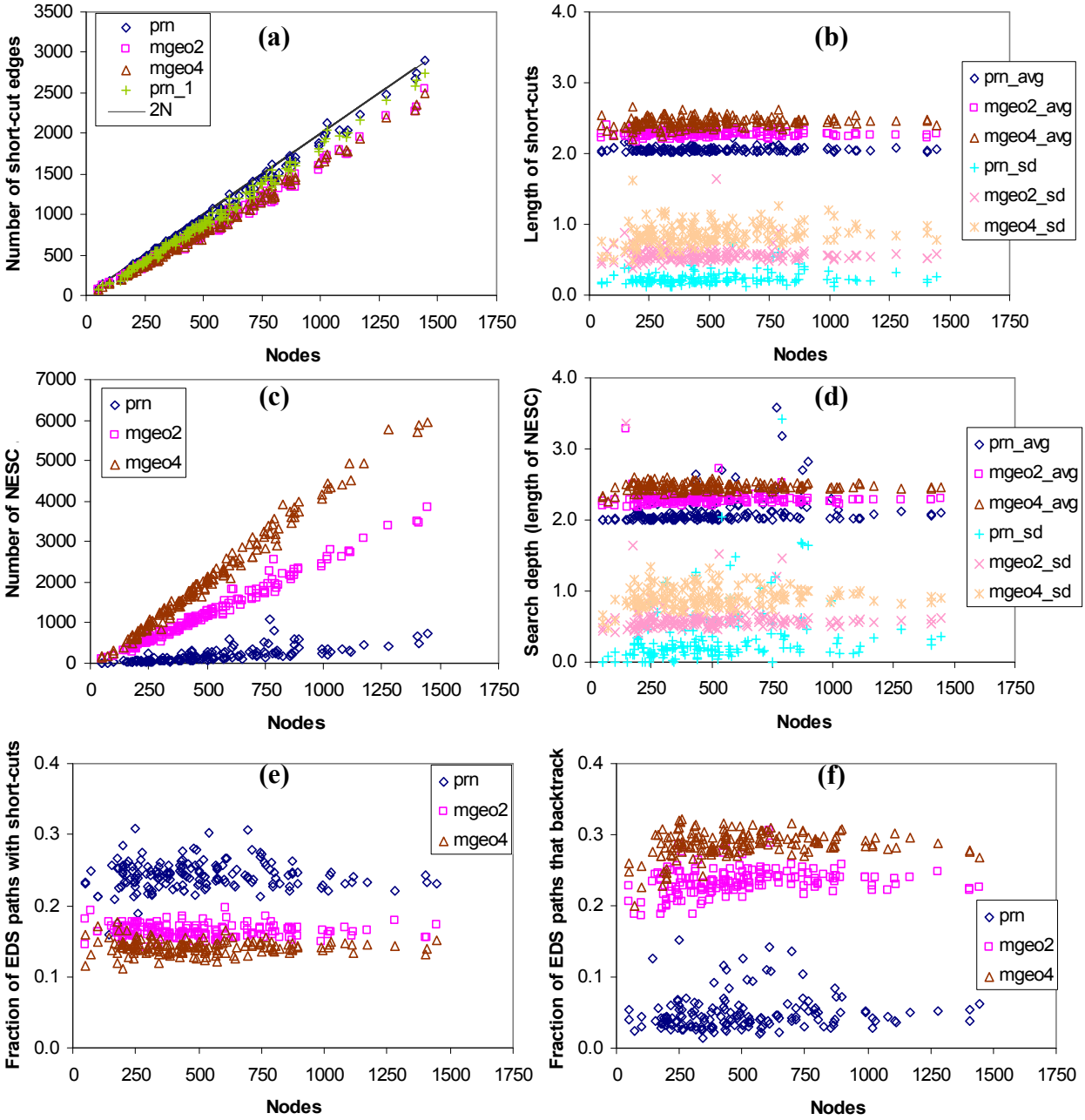


Fig. 9 Short-cuts and backtracking. **(a)** The number of short-cuts in PRNs is about twice the number of nodes in a network, $|SC| \sim 2N$. This approximation remains valid even when only one EDS path between each node-pair is used in the calculation (*prn_1*). MGEO networks have significantly fewer short-cuts than PRNs. **(b)** The average length of all short-cuts encountered by EDS on PRNs (*prn_avg*) is significantly shorter than the average length of all short-cuts encountered by EDS on both MGEO2 and MGEO4 networks, denoted *mgeo2_avg* and *mgeo4_avg* respectively. **(c)** The MGEO networks have significantly more non-existent short-cut edges (*NESC*) than PRNs. **(d)** The average length of all non-existent short-cuts encountered by EDS on PRNs (*prn_avg*) is significantly shorter than the average length of all non-existent short-cuts encountered by EDS on both MGEO2 and MGEO4 networks, denoted *mgeo2_avg* and *mgeo4_avg* respectively. **(e)** The fraction of EDS paths that traverses at least one short-cut edge is significantly larger for PRNs than MGEO networks. **(f)** The fraction of EDS paths that does backtracking is significantly smaller for PRNs than MGEO networks.

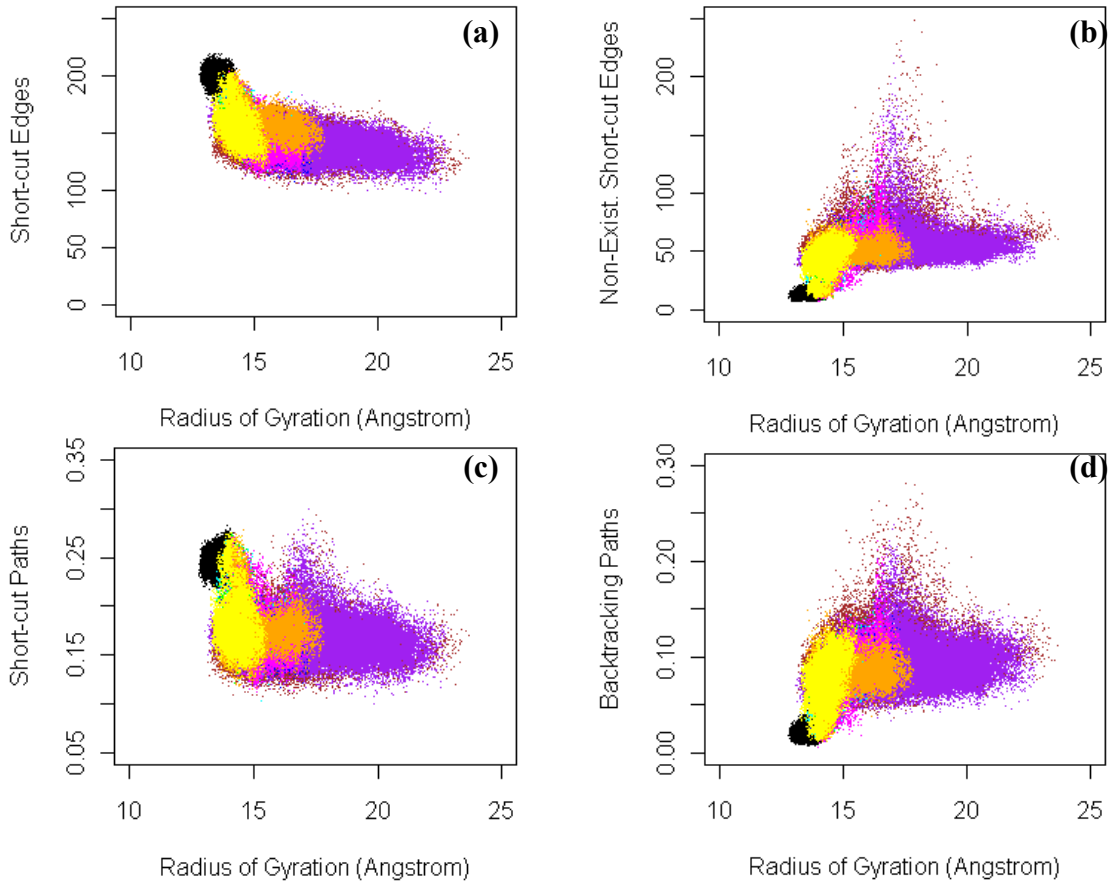


Fig. 10 Short-cuts and backtracks in PRNs from the 2EZN MD simulation. As the protein folds, generally: **(a)** the number of short-cut edges increases, **(b)** the number of non-existent short-cuts decreases, **(c)** the fraction of EDS paths containing at least one short-cut increases, and **(d)** the fraction of EDS paths doing at least one backtrack decreases. In the four plots, the points in black denote data from the native dynamics run 6250, while the non-black points come from the rest of the eight runs simulating non-native dynamics (section 2.7).

only one path chosen at random between every node-pair, the number of short-cuts still approximates $2N$ (prn_1 in Fig. 9a). The number of short-cuts found by EDS is however sensitive to how the PRNs are constructed. In a less dense PRN, say one constructed as pure C_α-C_α with 7.5 Å cutoff, the number of short-cuts are much fewer than $2N$.

When short-cut edges are removed from PRNs, there is a significant increase in both BFS and EDS average path length (Figs. 11a & 11b). This is understandable since the reduced network (PRN \ SC) has fewer links. However, the increase in average EDS path length is significantly larger than when a corresponding number of random edges are removed from PRNs (Fig. 11d). Hence, short-cut edges have a significant impact on average EDS path length and are a non-random subset of PRN links. In contrast, the increase in average BFS path length is significantly smaller than when a corresponding number of random edges are removed from PRNs (Fig. 11c). We revisit the effect of short-cuts on EDS path lengths in section 3.6.

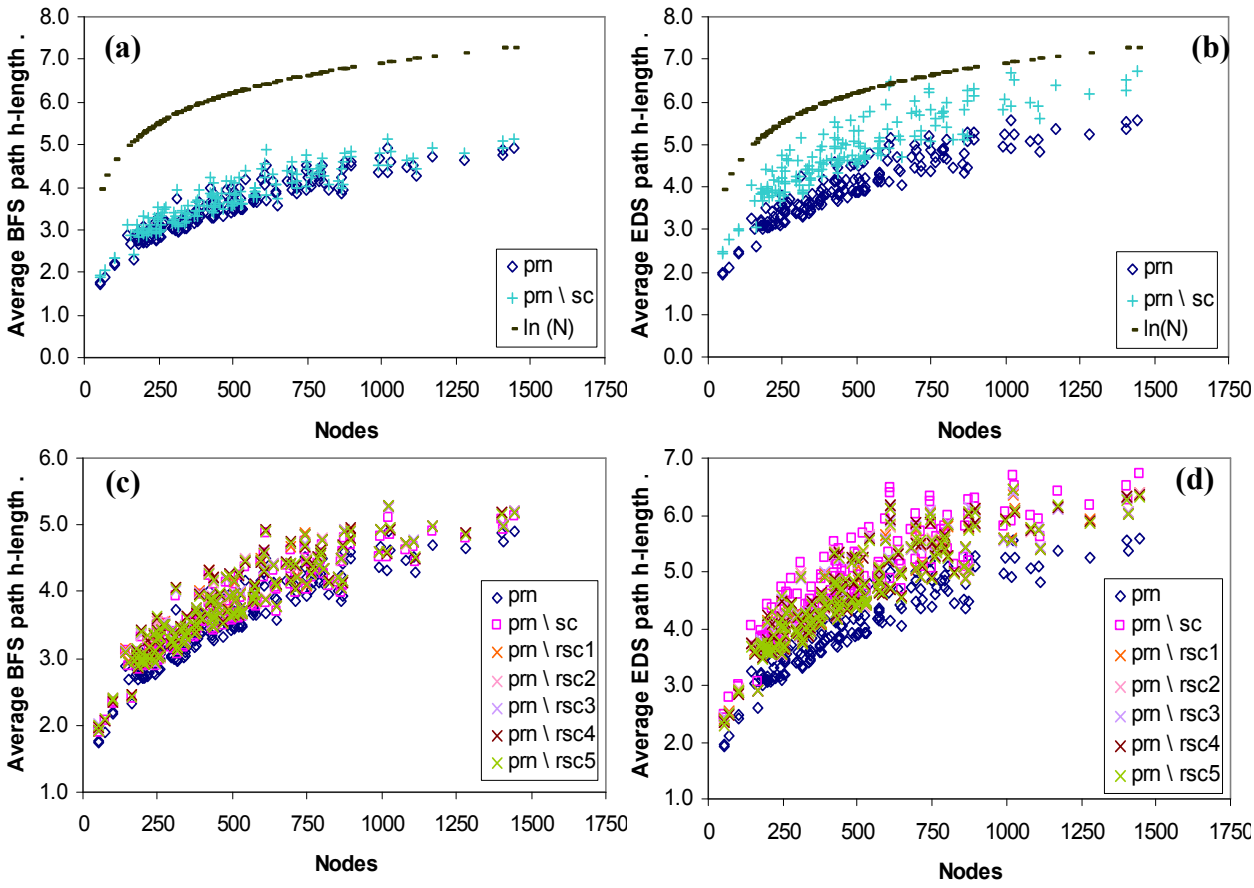


Fig. 11 Effect of removing short-cut edges on average path length. (a & b) Removal of short-cut edges from PRNs (denoted *prn\sc*) significantly increases the lengths of both BFS and EDS paths, but the effect is greater with EDS paths. **(c & d)** Removal of *rsc* (random short-cuts are described in section 2.6) from PRNs increased the average length of both BFS and EDS paths. However, *prn\sc* produces significantly longer EDS paths than *prn\rsc*. We revisit the difference in (d) in section 3.6 when more is known about short-cut edges. Since removing edges could potentially disconnect a PRN, the harmonic mean method of calculating average path length [8] (indicated by *h*) is used in these figures.

3.2 Search cost

Define the cost of finding a path from node *s* to node *d* as the number of unique nodes in the network that is touched by an algorithm when searching for the path. The search cost of a BFS path is the number of unique nodes visited by the BFS algorithm. The search cost of an EDS path is the number of unique nodes stored in memory (visited and enquired) by the EDS algorithm (see Appendix B for example). Due to the more diffusive nature of BFS, BFS paths rapidly become more expensive than EDS paths. BFS search cost increases linearly with network size, while EDS search cost scales logarithmically with *N* (Fig. 12).

Unlike BFS, EDS defers branching into other paths and only develops the path which appears most promising at the time. EDS touches only the direct neighbours of a node at each step of the way on a path. Thus, the cost of an EDS path *p* is at most the cardinality of the union of the direct neighbors of all nodes

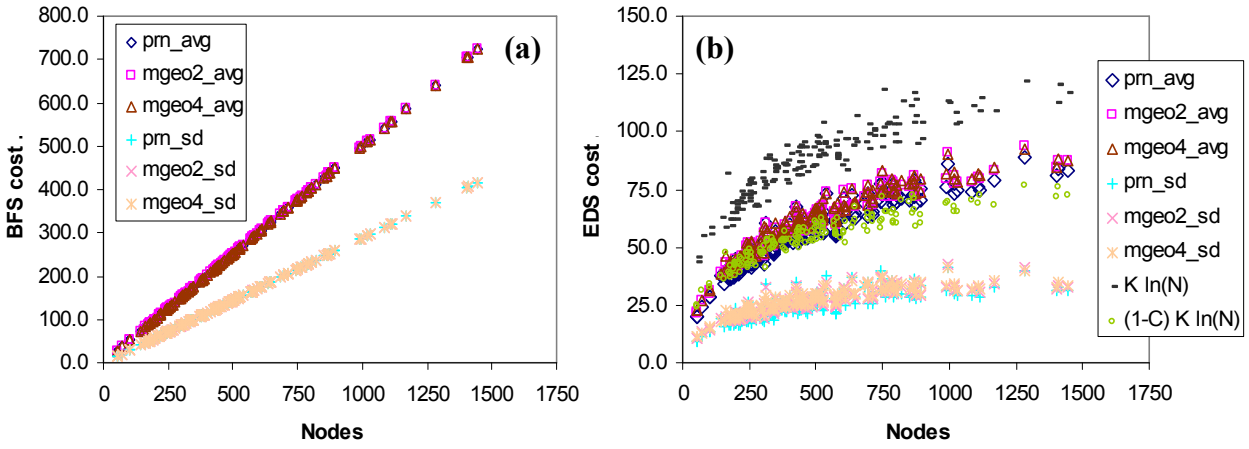


Fig. 12 Average search cost. (a) Average BFS search cost for all three network types increases linearly with the number of nodes N in a network, and is unaffected by changes in clustering. **(b)** *prn_avg*, *mgeo2_avg* and *mgeo4_avg* denote the average EDS search cost on PRNs, MGEO2 and MGEO4 networks respectively. Average EDS search cost on all three network types scales logarithmically with N, but *prn_avg* is significantly smaller than both *mgeo2_avg* and *mgeo4_avg*. *prn_avg* is approximately $(1\text{-}C)K \ln(N)$ where C is the clustering coefficient and K the mean node degree.

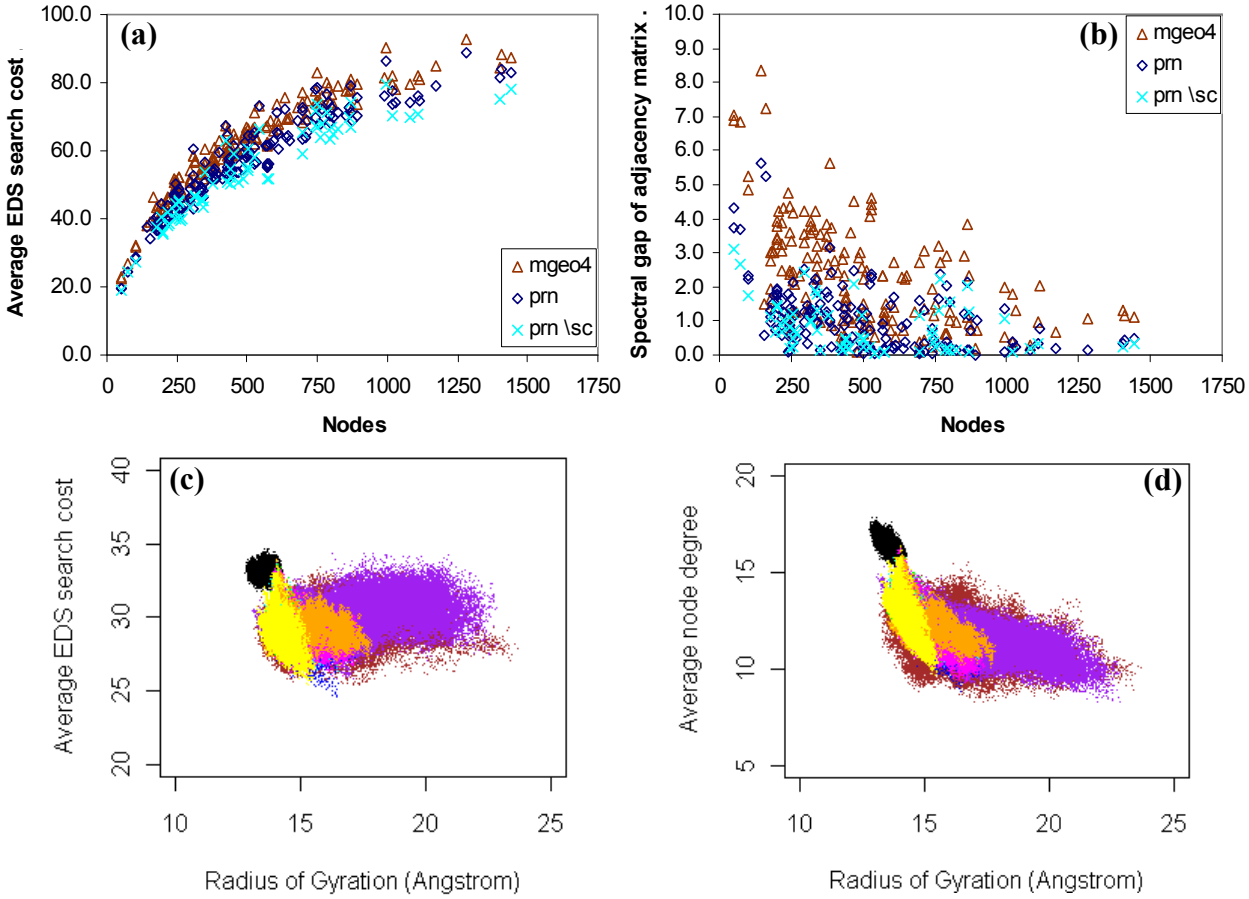


Fig. 13 Average EDS search cost and expansion of PRNs are positively related. (a & b) Networks with higher average EDS search cost also have larger spectral gaps. Large spectral gaps are associated with good expander graphs. **(c)** Average EDS search cost increases as the 2EZN protein folds. **(d)** Average node degree increases as the 2EZN protein folds.

on p, and is unaffected by backtracking. Due to heavy clustering, average EDS cost of a path is much less than $K\ ln(N)$ and approximates $(1\text{-}C)\ K\ ln(N)$ (Fig. 12b). Both mean node degree K and the clustering coefficient C are almost constant for PRNs and MGEO networks (Figs. 3a & 3b). BFS search cost is unaffected by changes in network clustering introduced via the MGEO networks (Fig. 12a). However, reduced clustering has an inflationary effect on EDS search cost. EDS paths on PRNs are significantly less costly to find than EDS paths on MGEO networks (Fig. 12b). Lower search cost indicates a less diffusive type of search, which may be preferable to reduce cross-talk when multiple searches are conducted simultaneously. It is also more reflective of how energy flows in proteins, which is anisotropic and sub-diffusive [12].

The low EDS search cost is hinted at by the expansion property of PRNs (section 2.4). The expansion factor of a graph with V nodes is determined by the smallest $|\mathcal{N}(S)| / |S|$ found over all sufficiently small ($< |V|/2$) node subsets of V (see Appendix C for more details). By definition, EDS search cost for a path p is also the size of the boundary for nodes in p, i.e. $|\mathcal{N}(S)|$ with S as the set of unique nodes in p. $|S|$ is $\leq N/2$ since EDS paths are short ($< \ln(N)$). Hence, by computing the EDS search cost for all paths, we are essentially calculating $|\mathcal{N}(S)|$ for a sample of node subsets which are sufficiently small. A direct relationship is observed between average EDS search cost and the spectral gap of the adjacency matrix. Networks with significantly higher average EDS search cost have significantly larger adjacency matrix spectral gaps (Figs. 13a & 13b).

As with average EDS search cost, the expansion factor (as indicated by the adjacency matrix spectral gap) for PRNs is sensitive to both clustering and average node degree. Both MGEO4 and PRN\SC networks have significantly weaker clustering than PRNs, but MGEO4 networks have significantly larger adjacency matrix spectral gaps than PRNs, while PRN\SC networks have significantly smaller adjacency matrix spectral gaps than PRNs. PRN\SC are worse expanders than PRNs because they have smaller average node degree. Average EDS search cost increases as the 2EZN protein folds, and is highest in the native state (Fig. 13c). This means that the increase in average EDS search cost due to the increase in average node degree as 2EZN folds (Fig. 13d), more than offsets the decrease in average EDS search cost due to the increase in average clustering (Fig. 1). Native PRNs may not be good expanders, but the process of protein folding improves their expansion property and this improvement, along with the increase in clustering, contributes to the navigability of PRNs.

3.3 Path analysis

Kleinberg describes his decentralized (local search) algorithm as homing in on a target node [6]. For PRNs however, even on EDS paths with no backtracking, Euclidean distance to target need not decrease monotonically. But for a fraction of EDS paths (and also BFS paths), what does change monotonically is

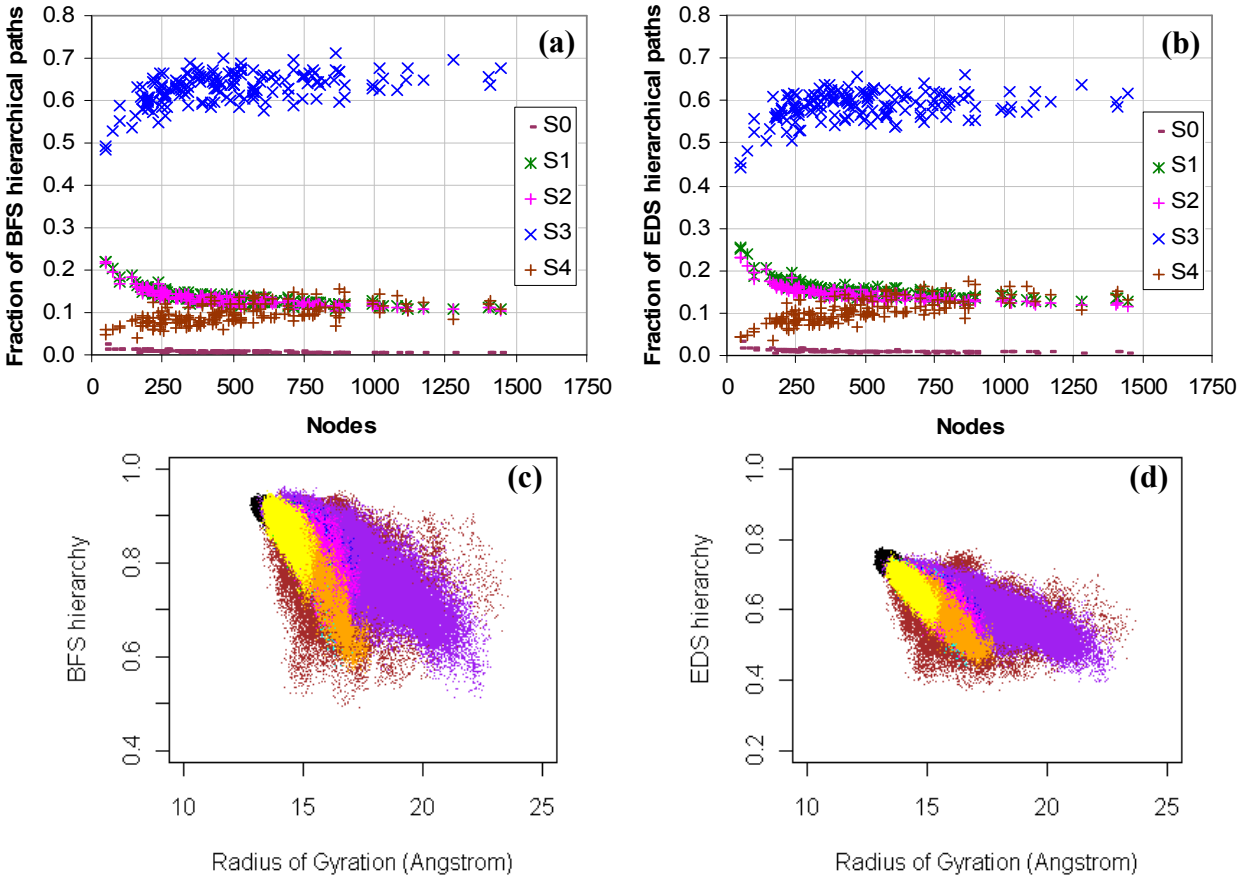


Fig. 14 Hierarchical paths. (a & b) Decomposition of BFS and EDS hierarchical paths by type. Hierarchical path types *S0…S3* are explained in the text. *S4* denotes paths with monotonically decreasing then monotonically increasing node degrees. *S4* paths are not typical hierarchical paths; they may even be considered *anti-hierarchy*. But we include them since *S4* paths also demonstrate structure. **(c & d)** The fraction of BFS paths and EDS paths which are hierarchical increases as the 2EZN protein folds.

node degree. Paths with monotonically increasing (type S1), monotonically decreasing (type S2) or monotonically increasing then decreasing node degrees (type S3) are hierarchical paths [38]. Implicit in this definition are paths with constant node degree (type S0). BFS and EDS hierarchical paths exhibit very similar hierarchical path decomposition (Fig. 14). The existence of both BFS and EDS hierarchical paths supports the notion that some hierarchical organization is present in PRNs. The fraction of hierarchical paths of a type is almost constant for $N > 250$ (larger multi-domain proteins). The majority of hierarchical paths are of type S3, which echoes the zoom-in zoom-out navigational pattern reported in other real-world networks [39], and is further evidence of the navigability of PRNs. Zoom-in correlates with increase in node degree and zoom-out with decrease in node degree.

3.4 Centrality

One of the first centrality measures to appear in the protein literature is *betweenness centrality*, defined as the number of shortest paths passing through a node [2]. Ref. [2] observed that betweenness centrality

changes as a protein folds to its unique native state and that residues with high (large) betweenness centrality during the transition state play a critical role in the folding process.

Given a set of paths *P*, define the betweenness centrality of a node *v* as the number of times paths in *P* passes through it, i.e. enter and exit, *v*. To observe the effect of local search on betweenness centrality in PRNs, we ranked the nodes according to their respective BFS and EDS betweenness centrality values and compared the two node rankings. We found a strong positive correlation between BFS betweenness centrality rank and EDS betweenness centrality rank, and this correlation is significantly stronger for PRNs than the MGEO networks. The average Pearson correlation for PRN, MGEO2 and MGEO4 networks respectively are 0.9018 (std. dev. = 0.0341), 0.8425 (std. dev. = 0.0431) and 0.8117 (std. dev. = 0.0580). A strong positive correlation between BFS betweenness centrality rank and EDS betweenness centrality rank means that nodes that are ranked highly by BFS betweenness centrality are also likely to be ranked highly by EDS betweenness centrality.

Another centrality measure that has proved insightful in protein research is *closeness centrality*, which measures how close nodes are to each other in a network. A node with high (large) closeness centrality has a low average distance to other nodes in the network. Closeness centrality of a node *i* is then inversely related to the length of the paths starting from node *i* to all the other *N*-1 nodes in the network: $CL_i = (N-1) \Big/ \sum_{j \neq i}^{N} \lambda(i, j)$. The average closeness centrality for a network with *N* nodes is: $CL = \frac{1}{N} \sum_{i}^{N} CL_i$. Since closeness centrality is a structural measure based on path-length, it too is influenced by clustering. Closeness centrality has been applied to identify possible protein folding nucleation sites [40] and to detect functional protein sites [41].

The decreased levels of clustering in the MGEO networks decreases the length of BFS paths and thus average BFS closeness is significantly higher in MGEO networks than PRNs. In contrast, EDS closeness centrality decreases significantly when clustering decreases. EDS closeness centrality responds to changes in clustering in a way that is more in keeping with the notion that as a protein becomes more compact and clustering increases, nodes or amino acids get closer to each other both in Euclidean distance and in graph distance. We found a strong positive correlation between BFS closeness centrality rank and EDS closeness centrality rank, and this correlation is significantly stronger for PRNs than the MGEO networks. The average Pearson correlation for PRN, MGEO2 and MGEO4 networks respectively are 0.9716 (std. dev. = 0.0470), 0.9541 (std. dev. = 0.0435) and 0.9329 (std. dev. = 0.0573). This means that nodes that are ranked highly by BFS closeness centrality are also likely to be ranked highly by EDS closeness centrality.

The strong and positive correlations between BFS and EDS betweenness centrality rank, and between BFS and EDS closeness centrality rank, are encouraging from a utilitarian perspective because it implies the transferability of existing research based on BFS, e.g. [40, 41, 42], and also the possibility of improving upon existing results with EDS.

3.5 Link usage

Short-range links are those with sequence distance $|u - v| \leq 10$ (section 2.1). Protein sequence distance is a familiar and meaningful metric in protein research. There is much discussion about the role of short- and long-range contacts in proteins [10]. There is a behavioral cost to the presence of long-range links in PRNs. Larger proteins with more long-range links tend to fold more slowly [43, 44, 45]. There is also an entropic cost to the formation of long-range links early in protein folding as the conformational possibilities of a sequence segment book-ended by a long-range contact is greatly reduced [46]. And indeed, the ratio of long-range to short-range links (LE/SE) increases for PRN edges as the 2EZN protein folds under MD simulation, and so does the LE/SE ratio for short-cut edges (Fig. 15). The bias towards the appearance of short-range links before long-range links accords with the framework model of protein folding which envisions protein folding as ascending a hierarchy of more complex forms starting from the secondary structure elements and progressing to higher order folds that comprise the organization of lower order elements. This process need not however exclude the necessity for feedback between the levels of organization [47]. In spatial networks, where nodes are located in a metric space e.g.: transportation and neural networks, links cover actual physical distances and as such longer links have a higher wiring cost [48]. For these reasons, relying more heavily on shorter links is deemed preferable.

Usage of an edge e is incremented by one with each traversal of e by a path. We observe that BFS paths are more biased towards long-range links than EDS paths (Fig. 16a). Since BFS paths are significantly shorter on average than EDS paths, link usage is normalized by average path length prior to statistical testing. After normalization, we found that: (i) BFS paths make significantly less use of short-range links than EDS paths; (ii) BFS paths make significantly greater use of long-range links than EDS paths; and (iii) BFS paths make significantly less use of 3did links than EDS paths (Fig. 16b).

Short-cut edges identified by EDS paths are predominantly short-range (Fig. 17). An average of 63.91% (std. dev. = 12.84%) of a short-cut edge set is composed of short-range links, while only 43.42% (std. dev=6.95%) of a PRN's links is short-range. Accordingly, the LE/SE ratio for short-cut edges is less than 1.0 in Fig. 15-right. The cutoff of 10 residues was chosen so that links within secondary structure

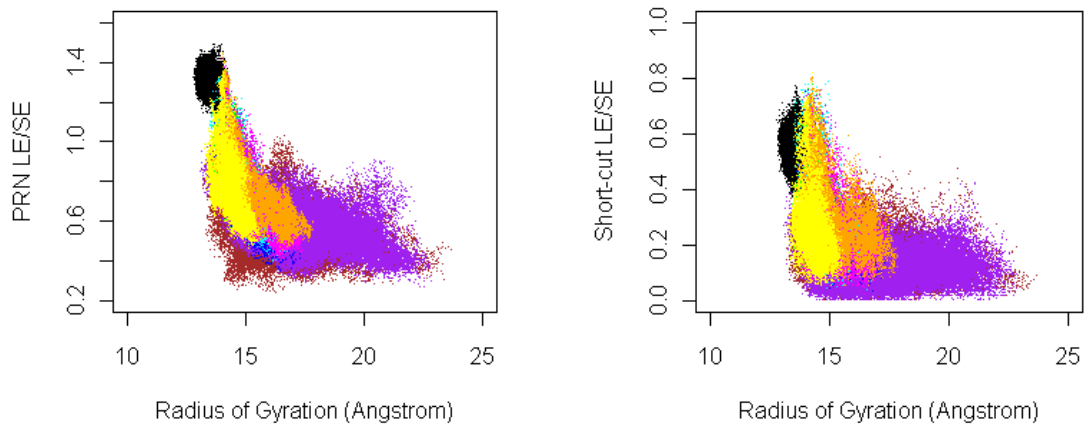


Fig. 15 Change in LE/SE, the ratio of long-range to short-range links, as 2EZN folds. LE/SE over all PRN edges (left) increases as the 2EZN protein folds under MD simulation, as does LE/SE for short-cut edges only (right). The LE/SE ratio for short-cut edges is less than 1.0, which means short-cut edges are predominantly short-range.

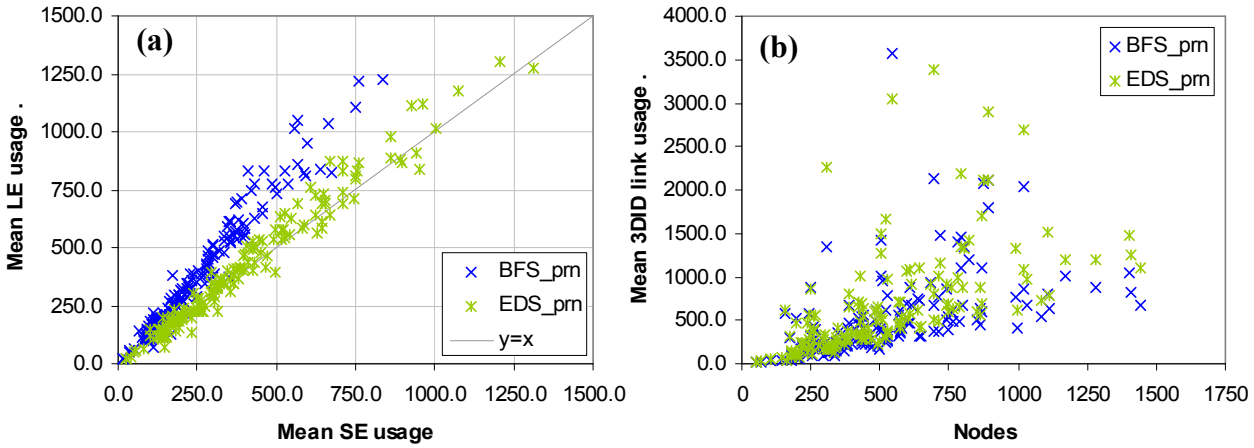


Fig. 16 Edge usage by type. (a) *BFS_prn* and *EDS_prn* respectively denote the mean BFS usage and mean EDS usage of long-range (LE) and short-range (SE) links for each of the 166 PRNs. Most of the *EDS_prn* and *BFS_prn* points lie above the x=y line, indicating a bias towards long-range links by both EDS and BFS. However, this bias is significantly stronger for BFS. **(b)** BFS paths make significantly less use of 3did links than EDS paths.

elements would be short-range (on average an α-helix is 11 residues in length and a β-strand, 6 residues) [25]. There is experimental evidence that energy in proteins is transported via secondary structures [49]. Energy transfer is faster in helixes (reaching the speed of sound if the helixes are rigid) and slower in the hydrogen bond network of beta sheets. And indeed, EDS paths in the 166 PRNs make heavier use of short-cut edges than non-short-cut edges (Fig. 18a). This is so not only for PRNs of configurations at equilibrium, but also for PRNs of configurations not at equilibrium (Fig. 18b). Average EDS edge usage, of both short-cuts and all edges, is significantly lighter (smaller) for PRNs in equilibrium which indicates that EDS edge usage becomes more balanced as a protein folds. A more detailed study of the EDS paths between allosterically linked protein binding sites may be fruitful. It may also be interesting to extend the idea of local search to protein complexes.

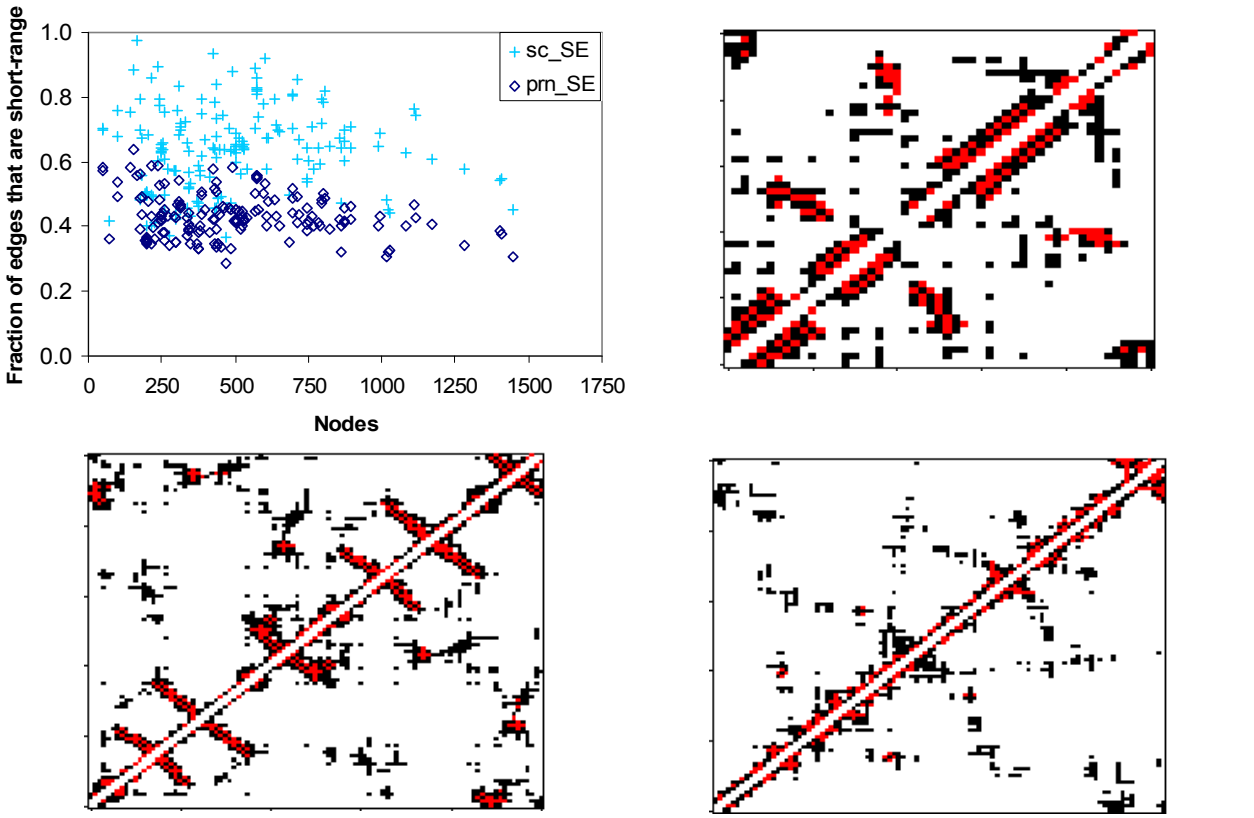


Fig. 17 Short-cuts are predominantly short-range. Top left: *sc_SE* and *prn_SE* respectively denote the fraction of short-cuts, and the fraction of all edges, that are short-range for each of the 166 PRNs. Since *sc_SE* is significantly larger than *prn_SE*, short-cuts are significantly enriched with short-range links. **Top right:** The 1B19 PRN in contact map form as in Fig 2 but with cells representing short-cut edges in red. The majority of the red cells hug the main diagonal where the protein's three alpha helix structures are located. In contact maps, alpha helix contacts are located along the main diagonal, and beta sheet contacts are situated either parallel with or perpendicular to the main diagonal. **Bottom left:** Contact map for a 2EZN (beta strand rich) native configuration with short-cuts indicated by red cells. **Bottom right:** Contact map for a 2EZN non-native configuration – the beta strands are not discernable yet at this stage.

3.6 Statistical signature of short-cut edges

As a consequence of being dominated by short-range links, short-cuts do not have significantly higher average edge multiplicity than non-short-cut edges. This follows from Fig. 3f which reports that short-range edges have significantly smaller average EM. However, against the backdrop of all edges, short-cut edges carry a distinctive statistical signature: they are situated in highly clustered areas, but have significantly weaker local community structure [50] themselves. Quantitatively this means that clustering amongst the first neighbors of short-cut links is significantly higher than clustering amongst the first neighbors of all edges, and clustering amongst the common first neighbors of short-cut links is significantly lower than clustering amongst the common first neighbors of all edges (Figs. 19a & 19b).

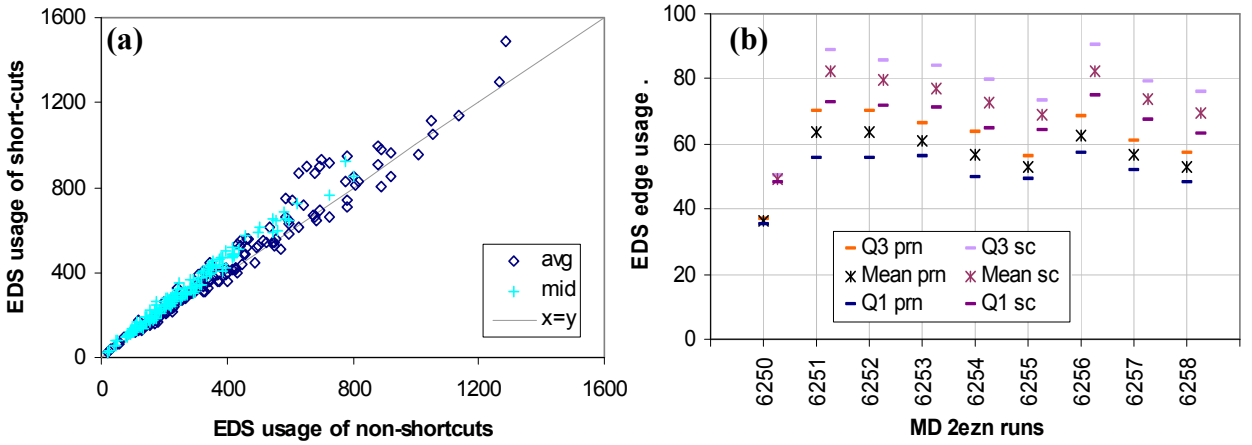


Fig. 18 Short-cut usage by EDS paths. (a) *avg* and *mid* respectively denote the average and median EDS usage of short-cuts and non-short-cuts for each of the 166 PRNs. Most of these points lie above the x=y line. Short-cut edges see significantly more EDS traffic than non short-cut edges. **(b)** EDS usage of short-cuts is compared with EDS usage of all edges in PRNs from the 2EZN MD dataset (section 2.7). For all MD runs (equilibrium and non-equilibrium), average EDS usage of short-cuts (*Mean sc*) is significantly heavier (larger) than average EDS usage of all edges (*Mean prn*). Average EDS edge usage, of both short-cuts and all edges, is significantly lighter (smaller) for PRNs in equilibrium which indicates that EDS edge usage becomes more evenly balanced as a protein folds.

Define clustering amongst the first neighbors of edge e as $C_{FN}(e)$, and clustering amongst the common (first) neighbors of edge e as $C_{CN}(e)$. The first neighbors of edge $e = (u, v)$ in Fig. 4(left) is the nine node set $\{a, b, c, e, f, g, h, u, v\}$. Since there are 17 links amongst the nodes of this set, $C_{FN}(e) = (2\times17)/(9\times8) = 0.47222$. The common neighbors edge (u, v) in Fig. 4(left) is the four node set $\{a, b, c, g\}$. As there are two links amongst the nodes of this set, $C_{CN}(e) = (2\times2)/(4\times3) = 0.33333$. The *FC* ratio for a given set of edges E is the number of edges in E where $C_{FN}(e) > C_{CN}(e)$ divided by the number of edges in E where $C_{FN}(e) \le C_{CN}(e)$. For each of the 166 PRNs, the *FC* ratio when E is the set of short-cuts (FC_{SC}) is significantly larger than the *FC* ratio when E is the set of all PRN edges (FC_{PRN}) (Fig. 19c). FC_{SC} is also significantly larger than the *FC* ratio when E is a set of random short-cuts. The ratio of FC_{SC} to FC_{PRN} increases as the 2EZN protein folds under MD simulation, signifying that the statistical signature of short-cut edges becomes stronger as PRNs approach equilibrium (Fig. 19d).

Local community structure around a link serves as a reservoir of alternate short routes should the link fail. This utility is evident from a global search or BFS perspective. But it applies as well for EDS on PRNs. Due to the relatively weaker local community structure surrounding short-cut links, EDS paths that connect node-pairs previously linked by short-cuts suffer greater positive dilation when short-cut links are removed, compared with EDS paths that do not connect node-pairs previously linked by short-cuts (Fig. 20a). This explains why the average EDS path length for *prn\sc* is a significantly longer than the average EDS path length for *prn\rsc* in Fig. 11d. Since $|SC| \sim 2N$, a node loses an average of four edges when short-cuts are removed from PRNs. This further weakens the local community structure around short-cut edges. Average EDS path dilation for short-cuts increases as the 2EZN protein folds, and is

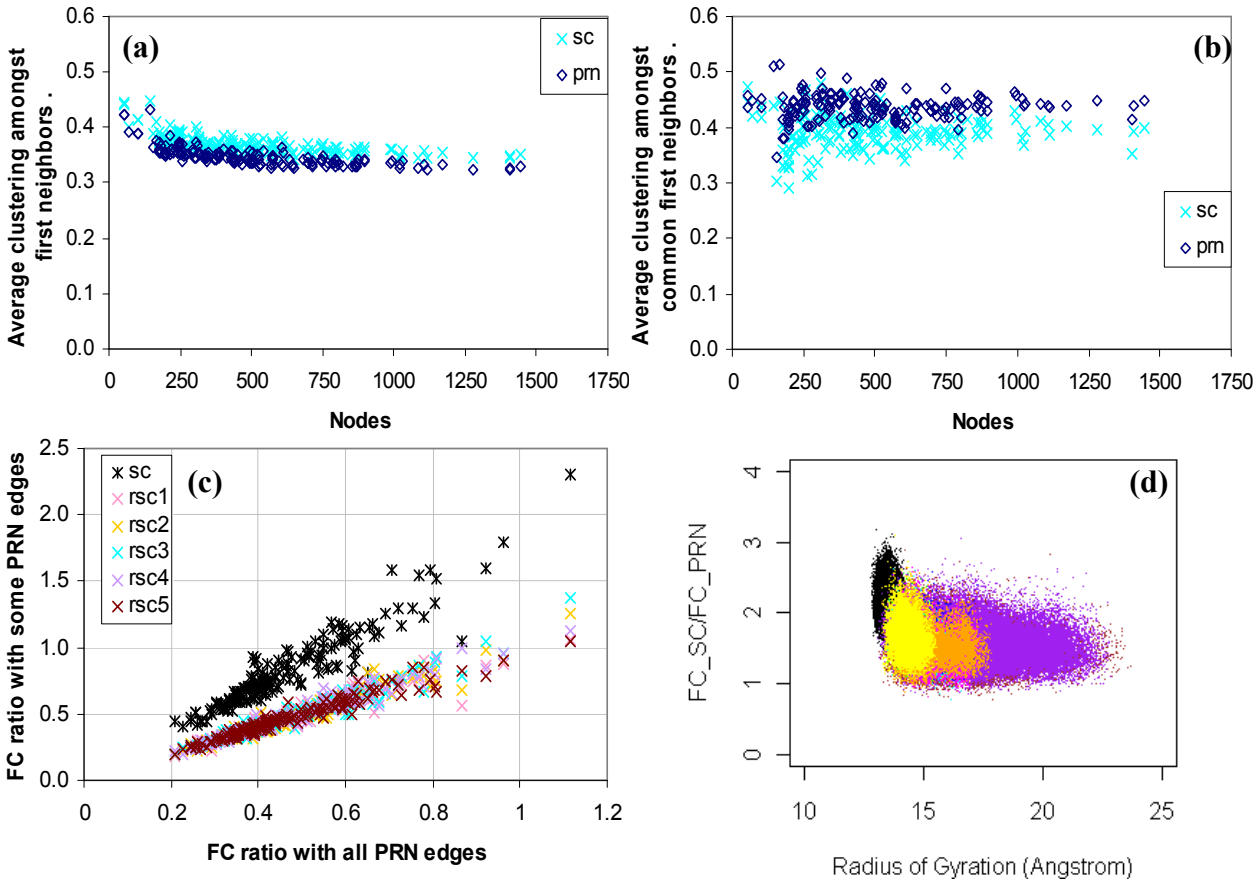


Fig. 19 Statistical signature of short-cuts. (a) Average clustering amongst first neighbors for short-cuts is significantly larger than average clustering amongst first neighbors for all edges. Short-cuts are more likely to be situated in highly clustered or dense areas in the network. **(b)** Average clustering amongst common first neighbors for short-cuts is significantly smaller than average clustering amongst common first neighbors for all edges. Short-cuts have weaker local community structure than non-shortcuts. **(c)** Short-cuts have significantly larger *FC* ratio than random short-cuts and than all edges. An edge set with a larger *FC* ratio has more edges whose first neighbor clustering is larger than its common first neighbor clustering, i.e. $C_{FN}(e) > C_{CN}(e)$. **(d)** The statistical signature of short-cuts strengthens as the 2EZN protein folds. FC_SC and FC_PRN are respectively the *FC* ratio for the set of short-cuts, and the set of all edges in a PRN.

significantly larger for native state PRNs (Fig. 20b). This behavior reflects the strengthening statistical signature of short-cut edges depicted in Fig. 19d. *Dilation* refers to the change in path length between a node-pair under two different circumstances. It is positive when the path is elongated and negative when the path is shortened. Some negative dilation did occur when short-cuts were removed from PRNs. Negative dilation applies only to paths between node-pairs that are not originally endpoints of edges. For Fig. 20, EDS path dilation between nodes u and v is $[\lambda_{PRN \setminus SC}(u, v) - \lambda_{PRN}(u, v)]$ or $[\lambda_{PRN \setminus RSC}(u, v) - \lambda_{PRN}(u, v)]$ where $\lambda_{PRN}(u, v)$, $\lambda_{PRN \setminus SC}(u, v)$ and $\lambda_{PRN \setminus RSC}(u, v)$ are respectively, the EDS path length between nodes u and v on a PRN, on a PRN without its short-cut edges, i.e. PRN\SC, and on a PRN without its random short-cuts, i.e. PRN\RSC. Average EDS path dilation for a set of edges E is

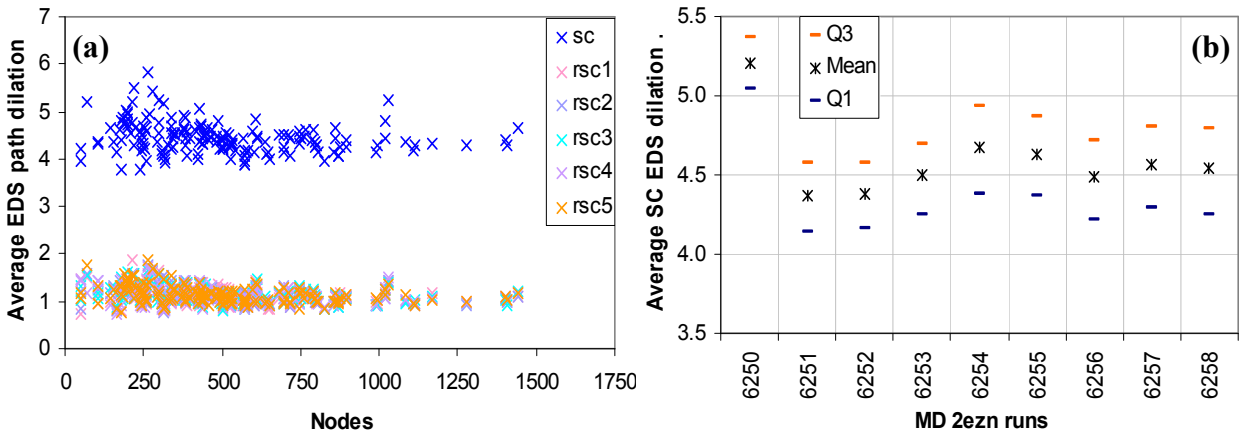


Fig. 20 EDS path dilation. (a) Short-cuts effect a significantly larger average EDS path length dilation than random short-cuts. **(b)** Average EDS path dilation for short-cuts is significantly larger for native state PRNs.

$\frac{1}{|E|}\sum_{i=1}^{|E|}\left[\lambda_{PRN\setminus E}(u,v)_i - \lambda_{PRN}(u,v)_i\right] \forall (u,v) \in E$ and $(x,y) \in E \Leftrightarrow (y,x) \in E$. Average SC EDS dilation for a MD run in Fig. 20b, is the average EDS dilation over the set of short-cut edges in a PRN (snapshot), averaged over all snapshots for that run.

4. Conclusion

Our examination of protein residue networks (PRNs) from a local search perspective was motivated by the need to understand why PRNs have small-world network structure. Specifically, why PRNs are highly clustered when having short paths is important for protein functionality. We found that by adopting a local search perspective, this conflict between form and function is resolved as increased clustering actually helps to reduce path length in PRNs. Further, the paths found via our EDS local search algorithm are more congruent with the characteristics of intra-protein communication. EDS paths are still short, i.e. their average path length still only increases logarithmically with network size (number of amino acids in a protein), and thus PRNs are navigable small-world networks. However, unlike paths found with a global search strategy like Breadth-First Search (BFS), EDS paths are more varied in length, their construction (search) cost is lower, and they make lighter use of long-range links. These differences make EDS paths a better model of energy flow in proteins, which is uneven, anisotropic, sub-diffusive and relies on secondary structures for transport. It would be instructive to compare EDS paths with actual pathways in proteins. In spite of the aforementioned differences, BFS and EDS paths share several similarities. BFS and EDS paths have nearly identical hierarchical decomposition profiles. Ranking PRN nodes by their BFS centrality values based on betweenness or closeness is not that different from ranking them by their corresponding EDS centrality values. This is encouraging because it implies that EDS centrality could be used instead of BFS centrality in existing methods, and the possibility of better results.

Clustering coupled with strong transitivity helps to keep EDS paths short on PRNs by creating a store of potential short-cut edges. Short-cut edges can be thought of as connecting sub-trees of an EDS search tree directly so that backtracking to the least common ancestor of the two sub-trees is avoided. However, this is only partially true, as short-cuts do not create cycles in an EDS search tree and an EDS search tree remains a tree even with short-cut edges (section 2.5). The number of short-cut edges scales linearly with protein size, and the short-cuts are dominated by short-range contacts (this is a measure based on sequence distance), experience heavier EDS usage (are more central), have stronger first neighbor clustering but weaker local community structure, and significantly impacts average EDS path length. The formation of short-cuts and the networks they create are subjects of our future study.

Throughout the paper, network statistics for PRNs generated from a molecular dynamics simulation of the native and unfolding dynamics of the 2EZN protein are reported. These dynamic PRN statistics help to support our results. They also provide a peek into the formation of the small-world network of PRNs. These dynamic PRN statistics tell us that native (equilibrium) PRN network statistics are significantly distinct from non-equilibrium PRN network statistics (the maintenance of this gap is explained by the structural advantages conferred to the protein by the native state); and that as the 2EZN protein folds (the protein becomes more compact), the following trends occur: clustering and transitivity increase, average path length decreases due to increase in the number of short-cuts and decrease in the fraction of EDS paths that backtrack, the average EDS search cost (expansion) increases due to increase in average node degree, the fraction of hierarchical paths increases, the ratio of long-range to short-range links increases, edges become less central (average edge usage becomes more balanced, which reduces the risk of bottlenecks), and the statistical signature of short-cut edges gain in strength.

Acknowledgements

This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca) and Compute/Calcul Canada. Funding was provided in part through a post-doctoral research position at Memorial University.

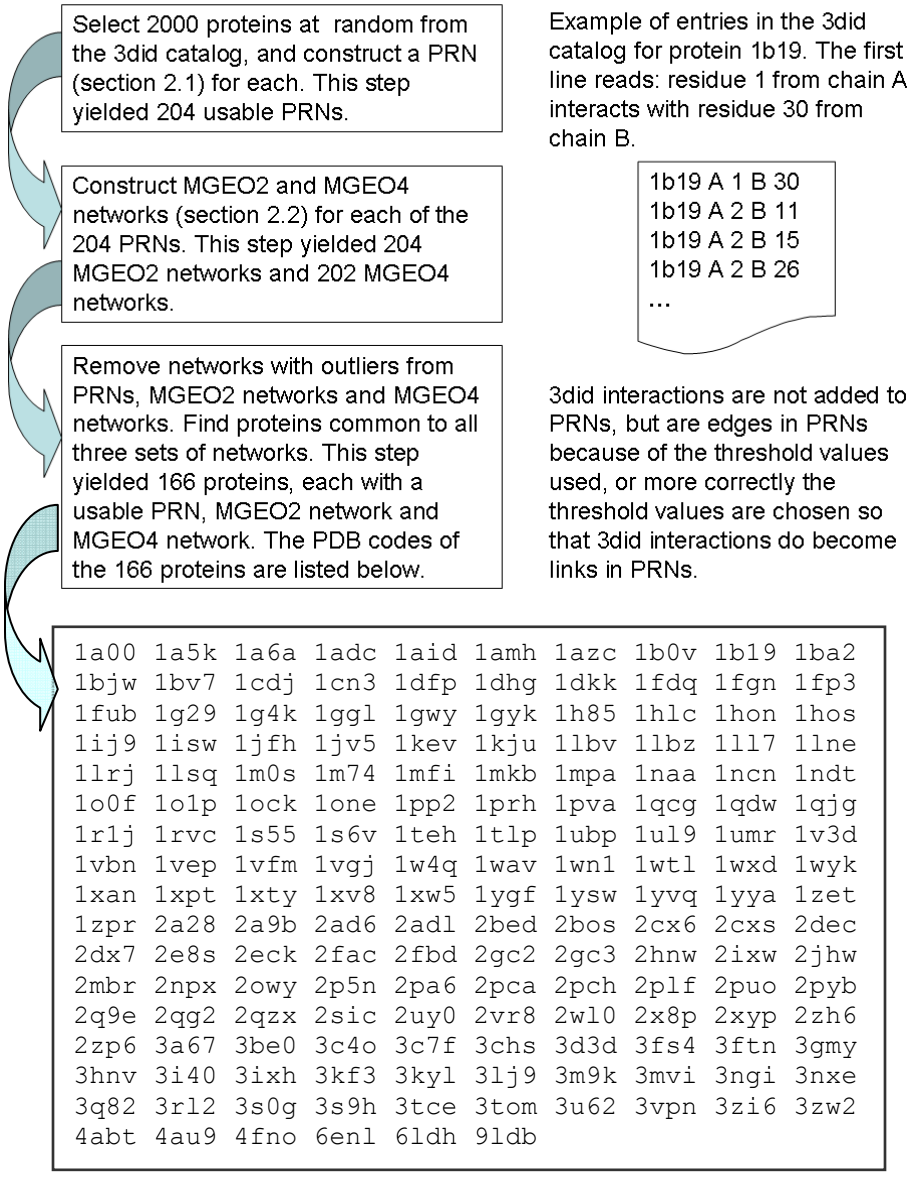
References

1 Watts DJ and Strogatz SH (1998) Collective dynamics of ‘small-world’ networks. Nature 393, 440-442.
2 Vendruscolo M, Dokholyan NV, Paci E and Karplus M (2002) Small-world view of the amino acids that play a key role in protein folding. Physical Review E 65 061910-1.
3 Bagler G and Sinha S (2005) Network properties of protein structures. Physica A: Statistics Mechanics and its Apps. 346(1-2) pp. 27—33.
4 Emerson IA and Gothandam KM (2012) Network analysis of transmembrane protein structures. Physica A 391:905-916.

5 Kleinberg J (2000) Navigation in a small world. *Nature* 406, 845
6 Kleinberg J (2000) The small-world phenomenon: an algorithmic perspective. Proc. of the 32nd Annual ACM Symposium on Theory of Computing, pp. 163-170.
7 Kasturirangan R (1999) Multiple scales in small-world graphs. cond-mat/9904055.
8 Newman MEJ (2003) The structure and function of complex networks. SIAM Review 45:167-256.
9 Gaci O and Balev S (2009) Node degree distribution in amino acid interaction networks. Computational Structural Bioinformatics Workshop, Washington DC, USA.
10 Whitley MJ and Lee AL (2009) Frameworks for understanding long-range intra-protein communication. Curr Protein Pept Sci. April, 10(2):116-127.
11 Jackson MB (2006) Molecular and cellular biophysics. Cambridge University Press.
12 Leitner DM (2008) Energy flow in proteins. Annu. Rev. Phys. Chem. 59:233-259.
13 Dokholyan NV, Li L, Ding F and Shakhnovich EI (2002) Topological determinants of protein folding. PNAS (13): 8637-8641.
14 Atilgan AR, Akan P and Baysal C (2004) Small-world communication of residues and significance for protein dynamics. Biophysical Journal 86:85-91.
15 Del Sol A, Fujihashi H, Amoros D and Nussinov R (2006) Residues crucial for maintaining short paths in network communication mediate signaling in proteins. Mol. Sys. Biol. Doi:10.1038/msb4100063
16 Atilgan AR, Turgut D and Atilgan C (2007) Screened non-bonded interactions in native proteins manipulate optimal paths for robust residue communication. Biophys. J. 92(9):3052-3062.
17 Chatterjee S, Ghosh S and Vishveshwara S (2013) Network properties of decoys and CASP predicted models: a comparison with native protein structures. Mol. BioSyst. 9:1774-1788.
18 Bhattacharyya M, Bhat, CR and Vishveshwara S (2013) An automated approach to network features of protein structure ensembles. Protein Science 22:1399-1416.
19 Park K and Kim D (2011) Modeling allosteric signal propagation using protein structure networks. BMC Bioinformatics 12. From The 9th Asia Pacific Bioinformatics Conference (APBC 2011) Inchon, Korea.
20 Li G, Magana D and Dyer RB (2014) Anisotropic energy flow and allosteric ligand binding in albumin. Nature Communications 5:3100.
21 Suel GM, Lockless SW, Wall MA and Ranganathan R (2003) Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nature Structural Biology 10(1):59-69.
22 Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE (2000) The Protein Data Bank. Nucleic Acids Research 28: 235-242. http://www.rcsb.org/pdb
23 Kannan N and Vishveshwara S (1999) Identification of side-chain clusters in protein structures by a graph spectral method. J. Mol. Biol. 292:441-464.
24 Mosca R, Ceol A, Stein A, Olivella R and Aloy P (2013) 3did: a catalog of domain-based interactions of known three-dimensional structure. Nucleic Acids Research 1-6.
25 Greene LH and Higman VA (2003) Uncovering network systems within protein structures. Journal of Molecular Biology 334:781-791.
26 Milenkovic T, Filippis I, Lappe M and Przulj N (2009) Optimized null model for protein structure networks. PLoS ONE 4(6): e5967.
27 Serrano MA and Boguna M (2006) Clustering in complex networks. I. General formalism. Phys. Rev. E 74, 056114
28 Colomer-de-Simon P, Serrano MA, Beiro MG, Alvarez-Hamelin I and Boguna M (2013) Deciphering the global organization of clustering in real complex networks. Scientific Reports 3:2517.
29 Serrano MA and Boguna M (2006) Clustering in complex networks. II. Percolation properties. Phys. Rev. E 74, 056115.
30 Atilgan AR and Atilgan C (2012) Local motifs in proteins combine to generate global functional moves. Briefings in Functional Genomics 2(6):479-488.

31 Bagler G and Sinha S (2007) Assortative mixing in Protein Contact Networks and protein folding kinetics. Structural Bioinformatics 23(14) pp. 1760—1767.
32 Turgut D, Atilgan AR and Atilgan C (2010) Assortative mixing in close-packed spatial networks. PLoS ONE 5(12):e15551
33 Bartoli L, Fariselli P and Casadio, R (2007) The effect of backbone on the small-world properties of Protein Contact Maps. Physical Biology 4:L1-L5.
34 Estrada E. (2010) Universality in protein residue networks. Biophysical Journal 98:890-900.
35 Van der Kamp MW, Schaeffer RD, Jonsson AL, Scouras AD, Simms AM, Toofanny RD, Benson NC, Anderson PC, Merkley ED, Rysavy S, Bromley D, Beck DAC and Daggett V (2010) Dynameomics: A comprehensive database of protein dynamics. Structure, 18: 423-435.
36 Beck DAC, Jonsson AL, Schaeffer RD, Scott KA, Day R, Toofanny RD, Alonso DOV and Daggett V (2008) Dynameomics: Mass Annotation of Protein Dynamics by All-Atom Molecular Dynamics Simulations. Protein Engineering Design & Selection 21: 353-368.
37 Jonsson AL, Scott KA and Daggett V (2009) Dynameomics: A consensus view of the protein unfolding/folding transition state ensemble across a diverse set of protein folds. Biophysical Journal 97:2958-2966.
38 Gao L (2001). On inferring autonomous system relationships in the Internet. IEEE/ACM Trans. On Networking 9: 733-745.
39 Boguna M, Krioukov D and Claffy KC (2008). Navigability of complex networks. Nature Physics 5:74-80.
40 Li J, Wang J and Wang W (2008) Identifying folding nucleus based on residue contact networks of proteins. Proteins 71: 1899 – 1907.
41 Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I and Pietrokovski S (2004) Network analysis of protein structures identifies functional residues. *Journal of Molecular Biology* 344 1135-1146.
42 Del Sol A, Fujihashi H, Amoros D and Nussinov R (2006) Residue centrality, functionally important residues and active site shape: Analysis of enzyme and non-enzyme families. Protein Science 15:2120-2128. Cold Spring Harbor Laboratory Press.
43 Plaxco KW, Simons KT and Baker D (1998) Contact order, transition state placement and the refolding rates of single domain proteins. J. Mol. Biol. 277:985-994.
44 Gromiha MM and Selvaraj S (2001) Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: Application of long-range order to folding rate prediction. J. Mol. Biol. 310:27-32.
45 Ivankov DN, Garbuzynskiy SO, Alm E, Plaxco KW, Baker D and Finkelstein AV (2003) Contact order revisited: Influence of protein size on the folding rate
46 Baker D (2000) A surprising simplicity to protein folding. Nature 405, 39-42.
47 Go N (1983) Theoretical Studies of Protein Folding. Ann. Rev. Biophys. Bioeng. 12:183-210.
48 Barthelemy M (2010) Spatial networks. arXiv:1010.0302 [cond-mat.stat-mech].
49 Botan V, et. al. (2007) Energy transport in peptide helices. PNAS 104(31):12749-12754.
50 Cannistraci CV, Alanis-Lobato G and Ravasi T (2013) From link-prediction in brain connectomes and protein interactomes to the local-community-paradigm in complex networks. Scientific Reports 3:1613. doi: 10.1038/srep01613.

Appendix A Overview of protein selection for PRN & MGEO network construction.



The 3did links are the intra- or inter-chain residue-residue interactions between contacting PFAM domains of a protein as listed in the 3did catalog [24]. A pair of PFAM domains in a protein is deemed able to interact with each other if they have at least five estimated contacts (hydrogen bonds, electrostatic or van der Waals interaction) between them.

Appendix B EDS search for node 76 (target) starting at node 72 (source).

At node 72, EDS inspects the 11 direct neighbors of node 72 in random order and computes their respective Euclidean distances to node 76. Since node 75 is the closest to node 76, EDS moves from node 72 to node 75. At this stage, EDS has visited nodes 72 and 75, and has memory of node 72 and its 11 direct neighbors. At node 75, EDS inspects its 11 direct neighbors and adds them to memory. From all the unvisited nodes in memory, EDS finds node 74 to be the closest to the target node. But EDS cannot move directly to node 74 since it is not a direct neighbor of node 75. To visit node 74, EDS must first move or backtrack to node 72. At node 74, EDS inspects its 11 direct neighbors in random order, and finds node 76. The maximum cost of this EDS search is 23 (at most 23 unique nodes are visited or inspected in the course of this search). The EDS path is ⟨72, 75, 72, 74, 76⟩ which has length four. This path is an example of an EDS path with backtracking that is also hierarchical (section 3.3) since its degree path which is ⟨11, 11, 11, 11, 4⟩ decreases monotonically. The BFS path is ⟨72, 74, 76⟩. In the reverse direction, the EDS path is ⟨76, 74, 72⟩ which is also hierarchical since its degree path increases monotonically.

Visited node n	Distance to target node	Inspected nodes (direct neighbors of n)	Level of n	Distance to target node
72 Level = 0 Degree = 11	8.93095	75	1	3.81547
		74	1	6.33784
		49	1	8.08382
		70	1	11.71590
		2	1	12.84950
		69	1	13.61700
		68	1	13.97110
		67	1	15.32500
		6	1	15.50050
		3	1	16.14170
		65	1	18.82700
75 Level = 1 Degree = 11	3.81547	51	2	7.76938
		50	2	7.80372
		49	2	8.08382
		73	2	8.90625
		72(x)	0	8.93095
		71	2	9.10019
		77	2	11.26160
		48	2	11.65110
		70	1	11.71590
		2	1	12.84950
		5	2	15.67090
74 Level = 1 Degree = 11	6.33784	76	2	0.00000
		51	2	7.76938
		50	2	7.80372
		49	2	8.08382
		72(x)	0	8.93095
		71	2	9.10019
		52	2	11.11780
		70	1	11.71590
		53	2	13.05500
		69	1	13.61700
		57	2	18.29190
76 Level = 2 Degree = 4	0.00000			

Appendix C Odd subgraph-centrality and network classes

The expansion factor γ of a graph G with $|V|$ nodes is the smallest $\frac{|\mathcal{N}(S)|}{|S|}$ ratio found over all node subsets S where $0 < |S| < |V|/2$, and the boundary set $\mathcal{N}(S)$ comprises all nodes in $V \setminus S$ that are first neighbors of nodes in S. Calculating γ directly from its definition is a hard problem, but techniques from spectral graph theory can be used to approximate. The expansion factor of a graph G denoted $\gamma(G)$, is related to the spectral gap $\Delta\lambda$, i.e. the difference between the largest (λ_1) and the second largest (λ_2) eigenvalues, of G's adjacency matrix by the Tanner-Alon-Milman inequality as $\frac{\Delta\lambda}{2} \le \gamma(G) \le \sqrt{2\lambda_1\Delta\lambda}$. A network with a large enough spectral gap is a good expander. Intuitively, good expanders are sparse (bounded node degree), well-connected (small diameter) and robust to small cuts. The spectral scaling method in [34] provides a way to evaluate the expansion property of graphs without defining "large enough". The method builds on the notion of odd-subgraph centrality which measures the weighted participation of a node in closed walks of odd lengths in a network.

Let $EE_{odd}(i)$ be the odd-subgraph centrality for node i, $EC(i)$ be the i^{th} component of the principal eigenvector (the eigenvector associated with λ_1), and $xj(i)$ be the i^{th} component of the eigenvector associated with the j^{th} eigenvalue λ_j. $EE_{odd}(i) = [EC(i)]^2 \sinh(\lambda_1) + \sum_{j=2}^{N} [xj(i)]^2 \sinh(\lambda_j)$ [Eq. 2 in 34]. When $\lambda_1 >> \lambda_2$, the first term dominates, yielding a power-law relationship between odd-subgraph centrality and the principal eigenvector. Let $EE_{homo}(i)$ represent $EE_{odd}(i)$ when $\lambda_1 >> \lambda_2$. Then a log-log plot of $EE_{homo}(i)$ vs. $EC(i)$ is a straight line with a slop of 0.5 and a y-intercept at $\log(\sinh^{-0.5}(\lambda_1))$. In graphs whose spectral gap is not "large enough", the odd-subgraph centrality values $EE_{odd}(i)$ will deviate from this straight line. The characterization of these deviations, i.e. whether $\log[EE_{homo}(i)/(EE_{odd}(i)]$ is positive or negative, forms the basis upon which the other three network classes are defined. These deviations arise due to the sign of the eigenvalues. Networks with zero deviations, i.e. $EE_{homo}(i) = EE_{odd}(i)$ for all i, are Class I networks

Class II networks are those with deviations but only of the negative kind, i.e. $EE_{homo}(i) < EE_{odd}(i)$. Class III networks are those with deviations but only of the positive kind, i.e. $EE_{homo}(i) > EE_{odd}(i)$. Class IV networks are those with deviations of both the positive and the negative kinds. Class II networks correspond to networks with modular structure, i.e. their nodes can be partitioned into two or more clusters such that intra-cluster connectivity is stronger than inter-cluster connectivity. Class III networks, in contrast, have a highly connected core of nodes that is only loosely connected to nodes outside the core, i.e. the periphery nodes. Class IV networks is a hybrid of Class II and Class III organizational patterns.