Graphlet-based Network Analysis of Protein Structures

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

von

Ioannis Filippis

Berlin
2012

Datum der Disputation:
26 June 2012

Gutachter:
Prof. Dr. Martin Vingron
Prof. Michael Schroeder

“Επιστήμη ποιητική ευδαιμονίας”

Πλάτων

Στον Παναγιώτη και στην Αργυρώ

Abstract

Network analysis of protein structures has provided valuable insight into protein folding and function. However, the lack of a unifying view in network modelling and analysis of protein structures and the unexploited advances in network theory prompted me to address three important challenges:

1. Rationalise the choice of network representation of protein structures.
2. Propose a well fitting null model for protein structure networks.
3. Develop a novel graph-based whole-residue empirical potential.

Graphlets, a recently introduced and powerful concept in graph theory, are a fundamental aspect of this thesis. The topological similarity between protein structure networks or individual residues was assessed using graphlet-based methods in order to propose an optimised null model and develop a novel potential.

Chapter 2 unifies the view of network representations by means of a controlled vocabulary and outlines the motivation behind the details of constructing such networks, and the popularity and optimality of the representations. In Chapter 3, an exhaustive set of 945 network representations is systematically analysed with respect to their similarity and fundamental network properties. The similarity between commonly used representations can be quite low and specific representations may exhibit high number of orphan residues and residues lying in "separate" components. Additionally, proteins with different secondary structure topologies have to be treated with caution in any network analysis. This work allows for a rational selection of a network representation based on certain principles, popularity, optimality and desired network properties and on its similarity to successfully utilised representations.

Chapter 4 shows that 3-dimensional geometric random graphs, that model spatial relationships between objects, provide the best fit to protein structure networks among several random graph models. The fit is overall better for a structurally diverse protein data set, various network representations and with respect to various topological properties. Geometric random graphs capture the network organisation better for larger proteins and proteins of low helical content and low thermostability. Choosing geometric random graphs as a null model results in the most specific identification of statistically significant subgraphs. This work has formed part of published literature.

In Chapter 5, a novel knowledge-based potential is developed by generalising the single-body contact-count potential to a whole-residue pure-topological one. The proposed scoring function outperforms the contact-count potential. The improved performance is consistent across various methods of generating decoys with respect to most performance metrics and is more prominent for the most successful fragment-based methods. The potential is also on par with a traditional four-body potential and exhibits strong complementarities with it, highlighting the capacity for further improvement.

Overall, this dissertation establishes the basis for the analysis of protein structures as networks and opens the door to new avenues in the quest for the perfect energy function.

Preface

Acknowledgements

This work was carried out in the Bioinformatics / Structural proteomics group of the Otto-Warburg Laboratory at the Max Planck Institute for Molecular Genetics in Berlin. I would like to express my gratitude to my supervisor Michael Lappe who gave me this opportunity and for his scientific support, guidance and patience. His faith in me and his knowledge and vision for the bioinformatics field contributed in large part to the successful completion of my thesis. Martin Vingron and Ernst-Walter Knapp, members of my PhD committee, naturally deserve special mention and thanks for their assistance, advice and counsel. I am grateful to Michael Schroeder for agreeing to be referee at such short notice and for all his help. I thank Nataša Pržulj for introducing me to the world of graphlets and for our fruitful collaboration. Over all these years in Berlin I had the pleasure of working with unforgettable colleagues. In particular I thank Dan Bolser, Henning Stehr, Jose Duarte and Konrad Paszkiewicz for stimulating discussions, support and good humour. Michael Sternberg and Sarah Butcher deserve also many thanks for their continuous support while I had and still have the pleasure working in their groups at Imperial College London.

Especially, I am grateful to my parents and my sister Vicky for their unshakeable belief in me and their continuous support that made this long journey possible. My best friends Dimitris, Antonis, Alessandro, Anna and Christos have been invaluable all these years. And last but not least I thank Katerina who made London the most special place to be over the last two years and whose constant encouragement helped me finishing this thesis.

Publications

A significant contribution of this thesis has been published in a peer reviewed journal. Specifically, Chapter 4 has been published in [191]:

Tijana Milenković*, Ioannis Filippis*, Michael Lappe, Nataša Pržulj. Optimized null model for protein structure networks (2009). PLoS ONE, 4(6), e5967, June 2009, *shared first authorship

Parts of Chapter 5 are in preparation for submission to a peer reviewed journal. The data set compiled in Chapter 3 has also been used in [78, 257] which I coauthored.

During my PhD, I co-authored the following publications [34, 78, 167, 257, 274] making significant contributions in [34, 274]. I participated in the 8th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction, published in [278].

Figures

This thesis reproduces four figures. I thank Dr. Nataša Pržulj for permission to adapt Figures 1.3 and 5.1 from [125, 240] as well as Dr. Michael Lappe for permission to adapt Figure 1.2A from [167]. Figure 1.1 is in the public domain and Figure 2.1 is reproduced under the Creative Commons Attribution-Share Alike 3.0 Unported license.

Contents

List of Figures

List of Tables

List of Abbreviations

CASP	Critical Assessment of techniques for protein Structure Prediction
CLUST	Random graph that preserves the degree distribution and the CLUSTering coefficient of all nodes of the real network
ER	Erdos-Renyi random graph
ER-DD	Erdos-Renyi random graph with the same Degree Distribution as the real network
GDD	Graphlet Degree Distribution
GDD-agreement	Graphlet Degree Distribution agreement
GDV	Graphlet Degree Vector
GDVS	Graphlet Degree Vector Similarity scoring function
GEO-3D	GEOmetric random graph using 3Dimensional euclidean boxes
MET	Random graph that preserves the degree distribution and the number of appearances of all 3-node subgraphs of the real network
MSDSD	Macromolecular Structure Data Search Database
PDB	Protein Data Bank
PISA	Protein Interfaces, Surfaces and Assemblies server
PQS	Protein Quaternary Structure database
RGF-distance	Relative Graphlet Frequency distance
RIG	Residue Interaction Graph
SCOP	Structural Classification of Proteins
SF-BA	Scale-free random graph with the Barabasi-Albert preferential attachment model
STICKY	STICKIness random graph
UA-ER-DD	Erdos-Renyi random graph with the same Degree Distribution as the real network (mfinder version)

Abbreviations regarding the RIG controlled vocabulary in Section 2.2.2

Chapter 1

Introduction

Summary

This first chapter introduces the field of network biology and in particular its impact on structural bioinformatics. General concepts in protein structure and graph theory are presented, a concise background on modelling protein structures as networks is provided, and the primary literature regarding the impact of network analysis of protein structures is reviewed. Graphlets, a new concept in graph theory and a fundamental aspect of this thesis, are explained in detail. Finally, an outline of the work presented in this thesis is provided.

1.1 Networks in biology

Biological research has been always driven by reductionism. Despite the inherent complexity of cellular organisation, the cell has been viewed as an un-coordinated bag of genes that act and can be studied in isolation. In the past decade high throughput experimental techniques have led to an exponential growth of molecular data that revolutionised biology with a new holistic perspective. Cellular components are not randomly connected together but rather exhibit certain functional interdependencies. Network science provided the means to describe and study the complexity of the functional dynamics of cells, organs and organisms. A network or a graph is a set of system components (nodes) that models their interactions as edges. In biology, nodes are genes, proteins, metabolites, RNA molecules, phenotypes. Protein-protein interaction networks model the physical interactions between proteins, signalling networks the capacity of molecules to (de)activate other molecules, regulatory networks the physical binding between transcription factors and regulatory elements, metabolic networks the biochemical reactions. Co-expression, phenotypic, and any other biological information can be further integrated to facilitate the study of the interplay of various effects.

Network analysis targets the laws and organising principles that biological networks follow. Quantitative description demonstrated that biological networks show manifestly non-random properties while modelling the networks helped to understand and explain the origin and evolution of their properties. Protein-protein interactions networks are scale-free; only a few proteins are highly connected while the majority has only a few interactions [19]. These networks are also ultra-small [57] in the sense that most of the cellular components (nodes) are only a few interactions away from any other component [314]. The scale-free nature has been shown to originate from gene duplication [229, 301]. Moreover, over-represented network patterns constitute essential units in regulatory networks [200]. Highly interconnectivity between proteins is indicative of their functional similarity and the formed modules carry specific cellular functions [134, 246, 329]. Topological centrality of proteins in a network may also unravel their biological importance, such as being disease related [133, 143, 279]. Finally, the architecture of biological networks provided insights into their robustness with respect to various perturbations [6, 20, 323].

The avalanche of research work in network biology inevitably influenced the view and understanding of protein structure space. Despite the fact that network analysis of protein structures has been extremely successful, limited analysis has been performed to establish the overarching principles of the network representations. The main aim of the work presented here is to address these overlooked aspects and to examine whether novel graph-based methods utilised in genomic networks offer the possibility of significant advances in structural bioinformatics.

1.2 Protein structure

Amino acids are small molecules containing three chemical groups and a hydrogen atom bound to a central α carbon. Amino group (NH_2) and carboxyl group ($COOH$) constitute the backbone while the R group, the side chain, defines the specific chemical properties of the amino acid. There are 20 amino acids that can be described as hydrophobic/hydrophilic, (a)polar, (un)charged. Additional sub-classifications refer to the size and reactivity.

Proteins are polypeptides, chains of amino acid residues that are linked to each other by a covalent bond. The distinct sequence of amino acids defines the three-dimensional structure and protein function is directly dependent on its structure. Apart from the covalent bonds between residues that are quite planar and rigid, numerous non-covalent forces impose constraints and influence the folded state of a protein. Hydrogen bonds are formed between backbone atoms as well as between the side-chain atoms and the surrounding aqueous medium. Hydrophobic interactions are entropically driven. Hydrophobic residues aggregate into the interior of the protein, while hydrophilic ones remain exposed to solvent [150]. Ionic bonds are either electrostatic repulsions between residues similarly charged or attractions between oppositely charged residues. Van der Waals forces summarise weak attractive and repulsive forces, a combination of dipole-dipole, dipole-induced dipole, and induced dipole-induced dipole forces.

There are four recognised levels of protein structure (Figure 1.1). Primary sequence is the linear sequence of amino acids. Secondary structure refers to local regularly occurring conformational units. α-helices and β-strands are stabilised through hydrogen bonds between the backbone atoms. β-strands are long and planar and occur in the core of the proteins, while α-helices are regular and spiral like structures and their hydrophobic and polar "faces" orient them towards both the interior and the surface of the protein. The complete three-dimensional structure of the protein is its tertiary structure. Often proteins contain more than one polypeptide chains. In such cases, quaternary structure is the way in which the folded monomer subunits form a complex.

Several classification schemes for the protein structure universe exist. Protein class is usually the highest level in the hierarchy and refers to the secondary structure composition. There are four major classes: *all-α* and *all-β* that are dominated by α-helices and β-strands respectively, $\alpha\,/\,\beta$ in which α-helices and β-strands alternate, and $\alpha + \beta$ in which α-helices and β-strands are rather segregated. In this work, the protein structural class as defined by Structural Classification of Proteins (SCOP) database [210] is frequently utilised to examine the generality or the specificity of our conclusions with respect to the structural architecture.

1.3 Network representation of protein structure

In network representation of protein structures, nodes represent amino acid residues and edges describe pair-wise contacts between residues. The resulting network is known as a Residue Interaction Graph (RIG) [8]. Different methods have been proposed to decompose protein structures into RIGs. These usually depend on the atoms selected to represent each residue, the definition of spatial proximity that denotes residues to interact and other properties of the interacting residues like their proximity in sequence or their secondary structure assignment. Figure 1.2 illustrates an example of interacting residues and of a RIG based on specific criteria. For example, protein 1l2y is decomposed into a network (Figure 1.2B) by considering all C_α atoms that are within 8Å and ignoring neighbouring residues in sequence. All networks considered in this work are undirected, unweighted graphs.

Although formal network representation of protein structures has been recently introduced, coarse-grained representations that are conceptually identical to RIGs, such as contact maps, have been widely used the last 40 years. In Chapter 2, we revise networks in structural bioinformatics field and we systemise the criteria used in modelling protein structures as networks.

1.4 Network properties

Several network measures are mentioned frequently throughout this work. Most of these measures have been used to quantify the importance of nodes and to characterise the overall topology of the network in network biology. In the following, these measures and properties are briefly introduced.

The *degree* of a node, its most elementary property, is the number of edges incident

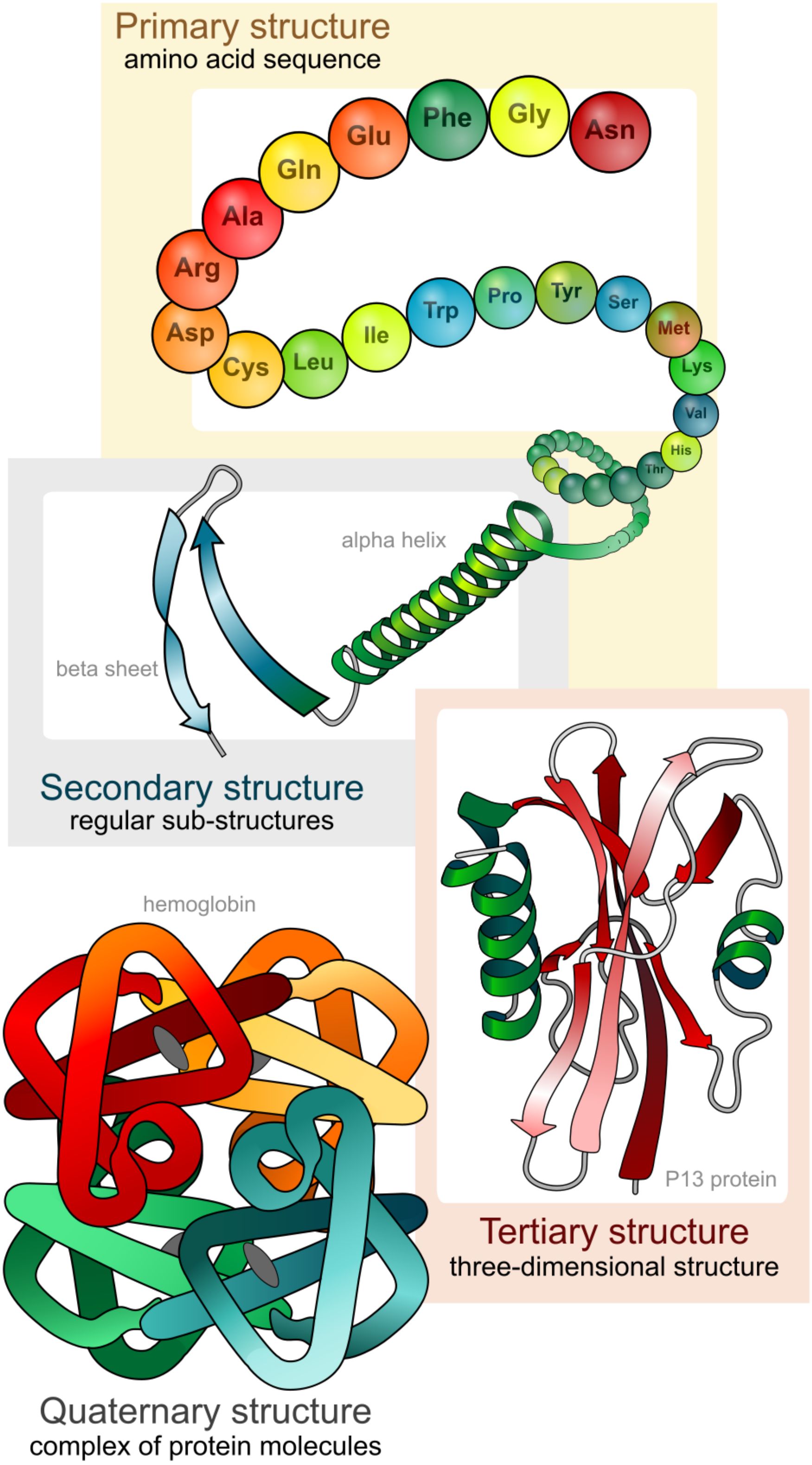


Figure 1.1: Protein structure levels.

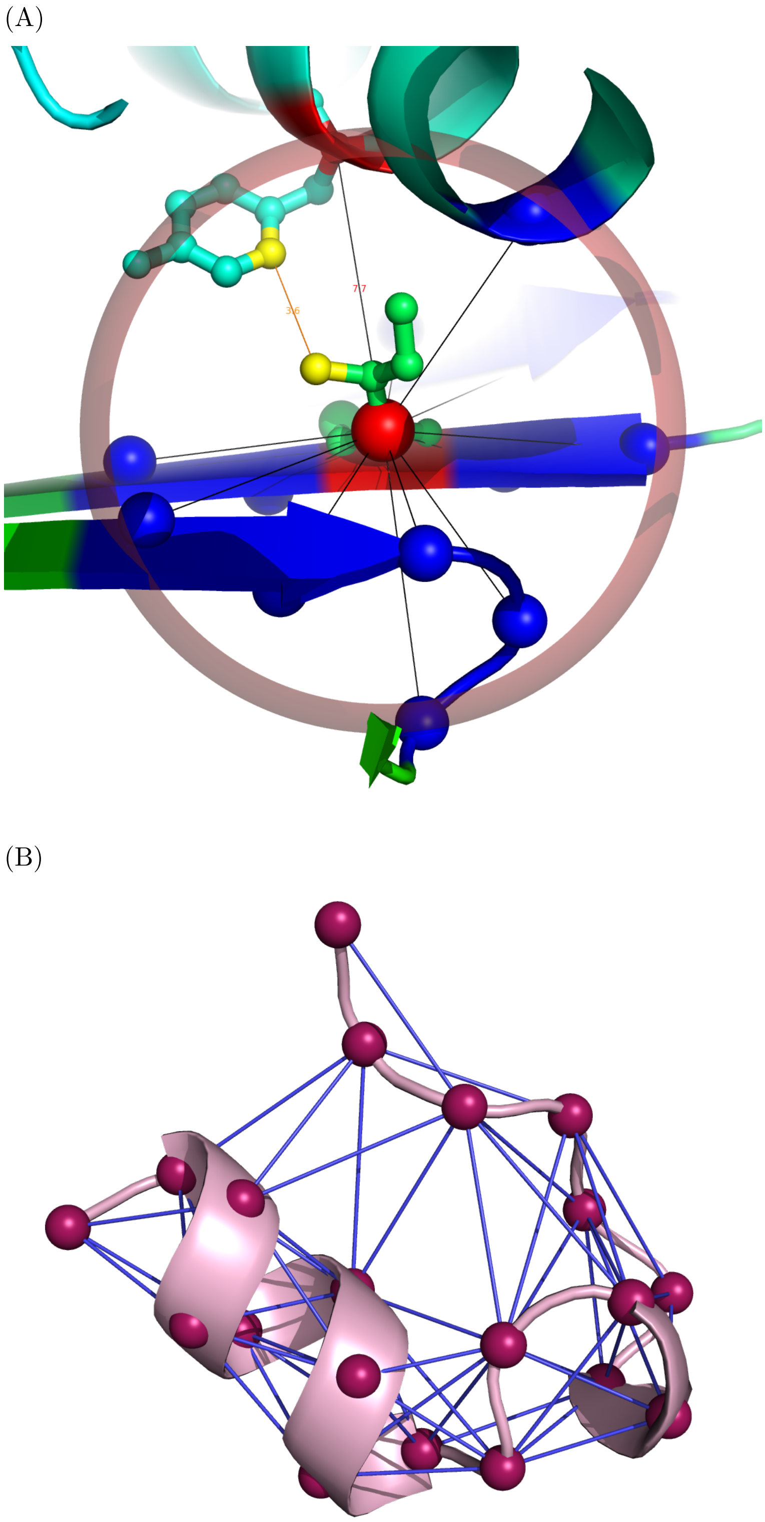


Figure 1.2: Residue-residue interactions (A) and Residue Interaction Graph (B). A. Large sphere denotes all blue coloured C_α atoms with 8Å from the central, red coloured C_α of Isoleucine 95 in 1a1m protein and lines denote the interatomic distances. Two residues may interact based on the atoms selected (e.g. yellow coloured C_β atoms or red coloured C_α atoms) and how spatially close they are. Reproduced from [167]. B. The network of interacting residues for 1l2y protein based on C_α atoms within 8Å and ignoring neighbouring residues in sequence. Residues (nodes) are modelled as spheres positioned at C_α atoms and residue-residue interactions as edges between spheres.

to the node. The *degree distribution P(k)* describes the probability that a node has degree k. Different network models may have distinct degree distributions. Random networks have Poisson degree distribution, while scale-free networks have a power-law degree distribution:

$$P(k) \propto k^{-\gamma}, \tag{1.1}$$

where γ is the degree exponent and usually is in the range (2, 3). In scale-free networks, only a few nodes are highly connected and are known as hubs, while the majority of the nodes have low degree. Many real-world networks have been shown to be scale-free [19].

The *clustering coefficient* C_z of node z in a network is the probability that two nodes i and j connected to the node z are themselves connected. The *clustering coefficient* of node z is defined as

$$C_z = \frac{2n_z}{N_z(N_z - 1)}, \tag{1.2}$$

where N_z and n_z are respectively the number of neighbouring nodes of z (degree) and the number of edges between the neighbours of z. The average of C_z over all nodes z of a network is the *clustering coefficient* C of the network. C indicates the propensity of nodes to form clusters and the resulting hierarchical nature of the network topology.

The smallest number of links that have to be traversed to get from node i to node j in a network is called the *distance* d_{ij} between nodes i and j and a path through the network that achieves this distance is called the *shortest path* between i and j. The average of shortest path lengths over all pairs of nodes in a network is called the *mean geodesic distance* l (also known as characteristic path length, average shortest path length, average network diameter). l is indicative of the "navigability" of a network.

Both *mean geodesic distance* l and *clustering coefficient* C are used to identify small-world networks [319], networks in which any node can be reached by any other node in a few steps. Small world networks have small *mean geodesic distance* l that grows proportionally to the logarithm of the number of nodes N in the network and high *clustering coefficient* C. Random graphs are small-world networks but scale-free networks have been shown to be ultra-small [57].

To identify central nodes in a network, two centrality measures have been developed based on the shortest paths. *Closeness centrality* of node i is the average of shortest path lengths from node i to every other node in the network. *Betweeness centrality* of node i is the fraction of shortest paths between residues other than i that pass through i. *Closeness centrality* indicates how fast a node can spread information to every other node in the network, while *betweeness centrality* the extent to which controls information within the network. The idea of centrality was originally introduced in 1948 [26].

Finally, a network is said to be *assortative* if nodes prefer to be connected to other nodes of similar characteristics [213]. A network may show assortative mixing on the node degree, i.e. high-degree nodes to favour connections with other high-degree nodes. When high-degree nodes prefer to be connected to low-degree ones, then the network is said to show disassortative mixing on the node degree. Pearson correlation coefficient of the degrees of the nodes connected measures the degree correlation. Assortative mixing has implications on network robustness [186].

1.5 Network analysis of protein structures

Vendruscolo et al. first constructed RIGs using only C_α atoms found within 8.5Å and demonstrated that RIGs with respect to the specific definition have small average shortest path length and high clustering coefficient [303]. The small world character of RIGs have been confirmed for other definitions: C_α atoms and 7Å distance cutoff [14], C_α atoms and 8Å distance cutoff [15], C_β atoms and 7Å or 8.5Å distance cutoffs [12], as well as all-atom representations and 4Å, 5Å, and 6Å cutoffs [106]. Although β-strands result in lower characteristic path length and clustering coefficient, secondary structure composition does not have any significant effect on the small-world nature of RIGs [14, 106]. As the distance cutoff increases for definitions based on C_β or all atoms, the networks exhibit weaker small-world character [12]. The small-world character was examined also with respect to the sequence separation of interacting residues. Greene and Higman showed that only short-range contacts, contacts between residues close in sequence, contribute significantly to the high clustering coefficient and thus, when only long-range interactions are considered, networks abolish their small-world character [106]. Bagler and Sinha argued that the addition of sequence connectivity in long-range networks is adequate for them to regain even partially their small-world-ness [14].

Greene and Higman also analysed the degree distribution for an all-atom representation, various distance cutoffs (4Å, 5Å, and 6Å) and various sequence separation thresholds (4, 6, 10, 14, 18) used for filtering out short-range contacts [106]. When considering all contacts independent of sequence separation, RIGs follow a bell-shaped Poisson distribution. For small sequence separation thresholds, the distribution is still Poisson. However, when only medium- and long- range contacts are considered (sequence separation ≥ 6), degree distribution is long-tail scale-free with an exponential cutoff. Secondary structure composition and distance cutoff have no significant effect on the form of degree distribution. These results have been verified for other definitions: C_α atoms and 7Å distance cutoff [14], C_α atoms and 8Å distance cutoff [15] and long-range version of the latter while preserving sequence connectivity [15].

Finally, Bagler and Sinha showed that RIGs have positive assortativity coefficient even when considering only long-range interactions [15]. This assortativity can be partially attributed to the observed degree distribution.

Despite the fact that previous network analyses of RIGs have provided valuable insight, our understanding of the impact of RIG definition upon network topology is far from complete. Each study utilises a different data set, focuses on a specific representation of a residue and examines a limited range of distance cutoffs or sequence separation thresholds. In Chapter 3, we set out a rigorous large-scale comparison of networks resulting from an exhaustive set of RIG definitions.

1.6 Applications in protein folding and function

Graph-theoretical measures have proved useful in the identification of critical residues for function as well as functional and viable split sites in proteins. Closeness centrality is one of the most prominent network properties utilised for prediction of functionally important residues [8, 51, 68, 283]. The precision in prediction can be increased by combining “local” network properties, such as the number of residues in direct contact or close in the network [274]. Closeness is also capable of identifying residues suitable for circular permutation [230] or highly flexible ones [12]. Interface residues with high betweenness centrality coincide with hot spots or are in direct contact with them [71]. Residues that mediate signalling significantly increase the average shortest path length upon removal [69], while shortest path analysis coupled with molecular dynamics reveals communication pathways [97]. The modular architecture of protein domains with respect to function has been also carefully investigated [40, 42, 49, 66, 67, 147, 247, 256, 258]. Decomposition of a protein domain into modules by clustering the corresponding RIG can successfully identify functional sites [66, 147] and binding regions [40, 42, 67, 256, 258], while modules are interconnected by residues important for allosteric communication [66].

Network analysis has provided also a broader insight into protein folding. Clustering [122, 147] as well identification of residues with many long-range contacts [112] can lead to putative nucleation centers. In transition state, nucleation centers can be identified as residues having high betweenness in the corresponding network [303, 308]. The average shortest path length has been also utilised as a topological determinant that discriminates between pre- and post-transition conformations [76]. Contact order, the average sequence separation over all interactions, calculated in the transition state ensemble has been shown to correlate well with folding rate showing that native-like topologies dominate the ensemble [226].

Other coarse-grained representations of protein structures that are conceptually identical to RIGs have been widely used in various aspects of structural bioinformatics. In Chapter 2, we revise the structural bioinformatics field from a network perspective and provide a thorough review of the importance of decomposing protein structures into networks.

1.7 Graphlets

Graphlet based network measures have been extensively applied to biological networks other than RIGs. Graphlets are small connected non-isomorphic induced subgraphs of large networks. They have been first introduced by Pržulj in [241]. Graphets differ from network motifs [200]. Motifs are partial subgraphs and thus may contain only some of the edges of the large network. They are network sub-patterns that occur in a network more or less frequently than expected at random. On the contrary, graphlets are induced and must contain all edges connecting its nodes as in the large network. Moreover, graphlets do not need to be over- or under- represented in the data compared with “randomised” networks. Motif approaches ignore subnetworks with "average" frequencies and thus only graphlets can be used for comparing two networks with respect to their full topology.

The number of graphlets on n nodes increases exponentially with n. Many real-world networks [319] and among them RIGs [303] have small world nature. Morover, the mean maximum shortest path length between any two nodes in the most common network representations over 60 proteins (Chapter 3) has value 9.8 ± 2.7. Therefore, using a relatively small graphlet size is sufficient to capture a large portion of a network without increasing the computational complexity unnecessarily. Figure 1.3 shows all 30 graphlets for 2- to 5-nodes, denoted by $G0$, $G1$, ..., $G29$. $G0$, the only 2-node graphlet, is actually the edge of a network.

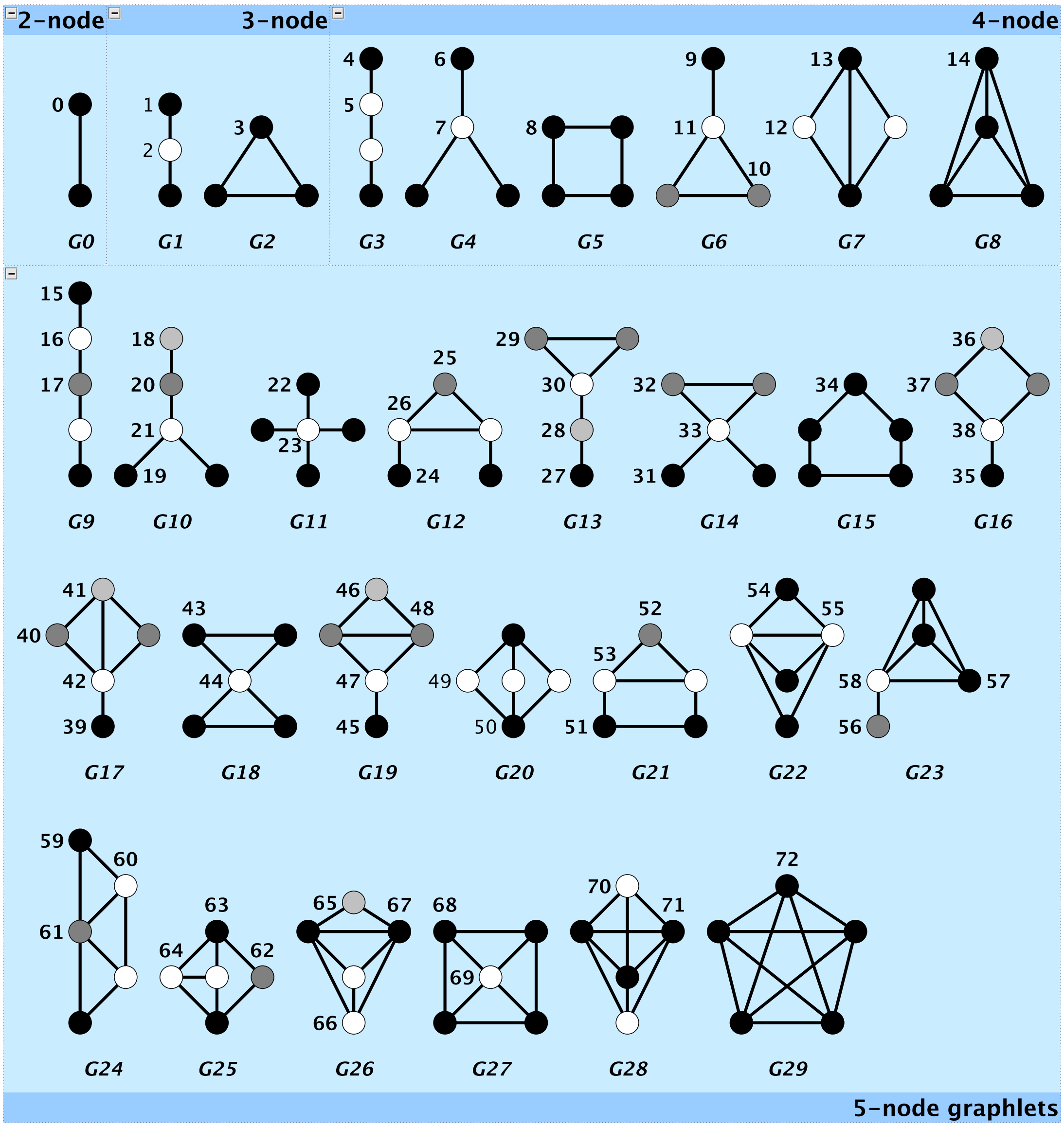


Figure 1.3: Graphlets and automorphism orbits. All 30 2- to 5- node graphlets and the corresponding 73 automorphism orbits. Nodes belonging to the same orbit have the same shade. Adapted from [240].

It is topological relevant to distinguish between nodes in a graphlet. For example, the two nodes at the periphery of $G1$ are identical from a topological point of view compared to the middle node. We have to take into account the symmetries of each graphlet. An isomorphism f from graph X to graph Y is a bijection between their sets of nodes such that any two nodes x and y in X and Y are connected if and only if $f(u)$ and $f(v)$ are also connected. Automorphism is an isomorphism of a graph to itself, a permutation of its nodes that preserves its structure, and the set of all automorphisms forms the automorphism group. The equivalence classes of the nodes of a graph under its automorphism group are called automorphism orbits. $G1$ has therefore two orbits distinguishing middle node from the end-ones. Figure 1.3 illustrates all 73 automorphism orbits (0, 1, . . . , 72) for 2- to 5-node graphlets. Nodes belonging to the same orbit have the same shade.

Graphlets and their automorphism orbits have been the basis for developing three graph theoretic measures. The relative graph frequency distance (RGF-distance) [241] and the graphlet degree distribution agreement (GDD-agreement) [240] are highly sensitive measures of local structural similarity between networks. The graphlet degree vector similarity can be used to identify topological similar nodes in a network or across networks [196]. All these measures have been designed by Pržulj and have been implemented in the network analysis software tool called GraphCrunch [163, 192].

RGF-distance and GDD-agreement have been used for modelling protein-protein interaction (PPI) networks [240, 241]. It has been also shown that neighbourhood topological similarity implies functional similarity in PPI networks and thus, graphlet degree vector similarity has been used to predict protein function [115, 196], identify cancer genes [193] and uncover melagonesis-related pathways [125]. Morever it has been adapted in a network alignment method called GRAAL [195] for topological alignments of biological networks [162, 195]. Recently a graphlet-based measure called graphlet degree centrality has been introduced to identify central genes in topologically complex and dense regions of PPI networks that are also biologically important [194].

Here, all these measures are employed in a novel way for protein structure networks. In Chapter 4, we use RGF-distance and GDD-agreement to identify an optimal null model for RIGs. In Chapter 5, we adapt the graphlet degree vector similarity to an empirical potential for discriminating native protein structures from decoys. All three measures are discussed in detail respectively in the fore-mentioned chapters.

1.8 Work presented

The primary focus of this work is to address three important problems.

Despite the fact that network analysis of protein structures has been extremely successful, there is lack of a unifying view. Different researchers adopt often arbitrarily different representations and usually neglect to demonstrate the impact of the representation itself upon results. Moreover, coarse-grained representations of protein structures that are conceptually identical to RIGs have been widely used the last 40 years. In Chapter 2, we revise the structural bioinformatics field from a network perspective and we systemise the network representations of protein structures by means of a controlled vocabulary. In Chapter 3, we systematically analyse the similarity of various representations and investigate the impact of the representation upon fundamental network properties. Both chapters rationalise the selection of the network representation, either based on the literature and the justification, popularity and optimality of a representation for a certain research problem or based on desired network properties and similarity to already successfully utilised representations.

A well fitting null model is also crucial in order to assess as accurately as possible the statistical significance of network properties. In Chapter 4, a reproduction of published work by this author, we examine the fit of various networks models to protein structure networks with respect to a multitude of local and global properties. We also illustrate the importance of the choice of the appropriate null model for motif analysis of protein structures.

Finally, a key element to successful protein structure prediction is an accurate energy function. In Chapter 5, we examine the efficacy of local network organisation as encoded in graphlets for discriminating native protein structures from decoys.

The overall aim of this research is to establish the basis for the analysis of protein structures as networks. The choice of a network representation, the similarity and network properties of various representations and an optimised null model are of fundamental scientific interest and are crucial for the further development of this research field. We also examine whether novel graph-based methods utilised in genomic networks offer the possibility of significant advances in structural bioinformatics by means of developing a novel knowledge-based potential.

Chapter 2

Systemising network representations of protein structures

Summary

Network analysis of protein structures is a relatively new field of structural bioinformatics. However, the same underlying principles apply when constructing a Residue Interaction Graph or modelling protein structure by means of other coarse-grained representations widely used the last 40 years. Thus far, despite the widespread applications of all these reduced representations, an overview of the criteria utilised to define a valid residue-residue interaction and subsequently a RIG has not been published. Here, we address the challenge of systemising network representations of protein structures. We manually collect 220 articles, annotate them with a controlled vocabulary and extract a manually curated data set of RIG definitions. Based on this data set, a detailed review of the research areas, the criteria used, the various selections for these criteria and their justification is provided. Additionally, cases of RIG definitions optimised for specific applications are discussed. Information provided here allows for an informed choice of network representation based on the current literature rather than arbitrarily adopting a representation. This chapter provided all essential background information to guide subsequent analysis in the following chapters. The proposed vocabulary will also contribute to a consistent notation of the RIG definitions and facilitate rigorous comparisons in the future.

2.1 Introduction

The view and understanding of protein structural space is three-dimensional per se. Yet, over the last 40 years it has been very popular to model protein structures as a set of spatially interacting residues. Phillips in 1970 [233] as well as Nishikawa et al. in 1972 [218] introduced the notion of distance matrix and contact map. Distance matrix is a matrix of all inter-residue distances while the contact map is a boolean matrix with non-zero values only for residues in contact. Similarly, Kannan and Vishveshwara defined the protein structure graph, a network where nodes represent residues and edges describe interactions between residues [147]. Later, Amitai et al. successfully introduced the Residue Interaction Graphs (RIGs) as a more accurate term for such networks [8]. There is a need to discriminate between RIGs and network representations of protein structures where nodes are atoms [137] or even secondary structure elements [107]. Distance matrices, contact maps and RIGs are almost equivalent concepts. The adjacency matrix of a RIG is actually a contact map that may be derived from a distance matrix.

Reduction of the three-dimensional structure of a protein to a two-dimensional coarse-grained representation simplifies analysis. Less computational resources are required and the representation is independent of the coordinate frame. Several methods have been developed to predict contact maps from sequence [53, 83, 86, 87, 100, 116, 136, 165, 178, 197, 222, 235, 238, 243, 263, 264, 267, 285, 288, 309, 313, 325]. There is a plethora of research areas where reduced representations have been applied successfully. Contact map facilitates structure alignment [46, 126, 138, 156, 270, 326] and automatic domain identification [2, 60, 79, 90, 95, 127, 272, 295, 327]. Knowledge-based potentials extracted from sets of interacting residues, are widely used for structure modelling and prediction of changes in stability [50, 88, 94, 119, 141, 151, 155, 159, 181, 187, 203–205, 209, 269, 271, 282, 307]. Network analysis identifies residues or regions critical for folding [112, 122, 147, 303, 308], stability [41, 47, 48, 54, 121, 122] and function [8, 12, 40, 42, 49, 51, 66–69, 71, 97, 147, 230, 247, 256, 258, 274, 283]. Topological properties calculated from a contact map or RIG, correlate well with folding kinetics [15, 75, 109, 114, 144, 146, 202, 234, 320, 335] and fold designability [80, 81, 262, 291].

To date, different methods with varying parameters have been proposed for decomposing protein structures into networks and contact maps. As the same principles can be applied to construct any of these reduced representations, hereafter there is no distinction between them: the term RIG is broadened to include contact maps or even simple sets of inter-residue interactions. Despite the widespread application of RIGs and the resulting advances in the structural bioinformatics field, limited analysis has been performed to provide an overview of the various RIG definitions: the different ways utilised so far for converting protein structures into RIGs.

In this review, we extract a manually curated data set of RIG definitions from 220 articles. This data set contains 70 basic definitions and additional criteria may extend the definitions up to 1,260. Unfortunately, no consistent notation is used in the literature neither for the definitions or even for the representations themselves. A lot of manual effort was necessary for the collection of the related articles, while manual curation was the only reliable option for the extraction of the definitions. Therefore, a large volume of text had to be analysed and the development of a controlled vocabulary was necessary for consistent annotation and subsequent analysis. These problems imposed certain limitations to the final number of publications analysed.

Based on the curated dataset, a detailed review of the research problems that are addressed utilising RIGs, is provided. The different RIG definitions are presented with emphasis given to the principal criteria that define a residue-residue interaction. The most popular choices for the criteria are outlined, while the justification of each choice is discussed if applicable.

2.2 Methodology

2.2.1 Literature search

Searching the literature via PubMed and using only keywords strictly related to the coarse grained representations did not yield satisfactory results. The search did not return any articles that refer to residues being in contact but without a specific reference to the reduced representation itself. Relaxing the keywords led to the retrieval of articles that do not actually provide any RIG definition. Identifying the definition within an article or the lack of it, is time consuming: no consistent notation is used and thus, the whole text has to be analysed. Therefore, the articles analysed were collected based on: a) the publications of researchers known to utilise reduced representations, b) a PubMed search using as keywords specific research areas and/or topological properties related to RIGs and c) articles that cite or are cited by the already collected ones. Articles that solely contain definitions under which residues in different chains are in contact were ignored.

2.2.2 Controlled vocabulary

During text analysis, a controlled vocabulary was developed for the efficient annotation and subsequent analysis of RIG definitions. Three different structural criteria called contact type, definition and range are used to define residue-residue interaction, while a fourth one called contact weight is also mentioned for the sake of completeness. Contact type (CT) defines the atoms that represent a residue, contact definition (CD) defines whether two residues are close enough to interact and contact range (CR) may exclude specific interactions based on properties of the interacting residues such as their sequence separation. A RIG definition is named basic when only CT and CD are specified. In general, a RIG definition will be notated as:

$$RIG = (CT)_{CD}^{CR}$$

In the following, each of the structural criteria is explained in detail.

Contact type

Two residues are considered to be in contact if they have at least a pair of specific atoms close in space. The atoms considered to interact or not and thus representing the residue, constitute the contact type. A residue can be represented from a single atom up to all its atoms in increasingly fine granularity. In all cases, non-heavy atoms (i.e. hydrogens) are not considered. Interactions between C_α atoms are denoted by C_α contact type. Similarly, C_β, BB, SC and ALL contact types are used to denote interactions mediated through C_β, backbone, side-chain and all atoms, respectively. A contact type may also refer to interactions between atoms of different type: e.g. interactions between C_α and C_β atoms are denoted by C_α/C_β. Moreover, sets of interacting atom types can be combined into a single contact type. For example, $C_\alpha + C_\beta$ denotes interactions between C_α atoms and interactions between C_β atoms, while $C_\alpha.C_\beta$ all possible interactions with C_α and C_β atoms (i.e. C_α, C_β and C_α/C_β).

A residue may also be represented by a single virtual atom, the centroid of the coordinates of a specific set of atoms. For example, SC_c refers to the centroid of side-chain atoms. The position of the centroid can be calculated as the geometric center or the center of mass of specific atoms. An “average” centroid per residue type may also be calculated as the average of the centroids for all residues of the same type in a representative data set. The cs subscript is used to denote representation by a special centroid. In the case of BB_{cs}, the backbone of a residue i is represented either by a sphere placed at the carbonyl atom [171] or by a point in the middle between the C_α atoms of the residues $i-1$ and $i+1$ [44]. For the SC_{cs} contact type, the side chain is represented by a point [44] or a sphere [171] centered at a specific distance from the C_α atom and in the direction of the $C_\alpha - C_\beta$ bond. Rarely, the residue is represented by a “functional atom” (FA), a single real atom that approximates the centre of functional activity [252, 316]. Residues with only hydrogen and carbon atoms in their side-chains are not considered as not commonly occurred in functional sites.

In the case of glycine and for contact types based on real side-chain atoms, the C_α atom is commonly utilised as a C_β. In rare cases, glycines are ignored [253] or a virtual C_β atom is constructed [3, 11, 44, 135, 208]. This atom can be placed at the hydrogen atom of the side chain [135] or its position might be calculated from the backbone [3, 44].

Contact definition

Contact definition specifies whether the atomic sites are spatially close enough, so that the corresponding residues are considered to interact. Two atoms may interact if they are within a certain distance cutoff. For example, when two residues are in contact if their C_α atoms are within 8.0Å, the resulting RIG is denoted as $(C_\alpha)_{8.0\text{Å}}$. An interaction may also be defined if the distance between their van der Waals spheres is less than a cutoff. In $(SC)_{\sum_{ij} r_{vdW}+0.5\text{Å}}$ RIG, the distance between the side-chain atoms of the interacting residues is less than the sum of their van der Waals radii plus 0.5Å. In case the residue is represented by a sphere, the radius of the sphere may be used instead (e.g. $(SC_{cs})_{\sum_{\mathrm{ij}} \mathrm{r}_{\mathrm{sphere}}+0.5\mathring{\mathrm{A}}}$). Among the previous definitions, only the definitions that do not take into account the radii of the van der Waals spheres or of other special spheres, will be considered hereafter as distance cutoff based. The subscript DC is used to denote this definition class (e.g. $(C_\alpha)_{\mathrm{DC}}$).

Delaunay tessellations [72] and Voronoi diagrams [311] have also been used to capture proximity relations between atoms in protein structural space. A Voronoi tessellation divides protein structure into convex polyhedra (called Voronoi cells), one per each atom (or site in general). All points in each cell are closer to the corresponding atom rather than to any other atom. In other words, space is decomposed into regions with the same set of closest neighbor sites. A topological dual to Voronoi partitioning, is the Delaunay tessellation that partitions space into tetrahedra called Delaunay simplices. A group of four atoms whose Voronoi polyhedra meet at a common point forms actually a simplex. The sphere defined by those four atoms/points contains no other points on its interior. Delaunay tessellation actually defines an ensemble of sets of four nearest neighbour atoms. A Voronoi and Delaunay tessellation in 2D is illustrated in Figure 2.1. Delaunay tessellation has been used to decompose protein structure into a set of interacting residues [131, 219] and especially for development of four-body potentials [50, 94, 159, 209, 269, 294]. Voronoi diagrams have been widely used to study packing and volume distributions [55, 89, 248]. Voronoi diagrams can also be used for RIG construction. It is possible to define contacting residues based on the contact area between the corresponding polyhedra [337]. Here, there is no discrimination between these two approaches and both are denoted by the subscript DT.

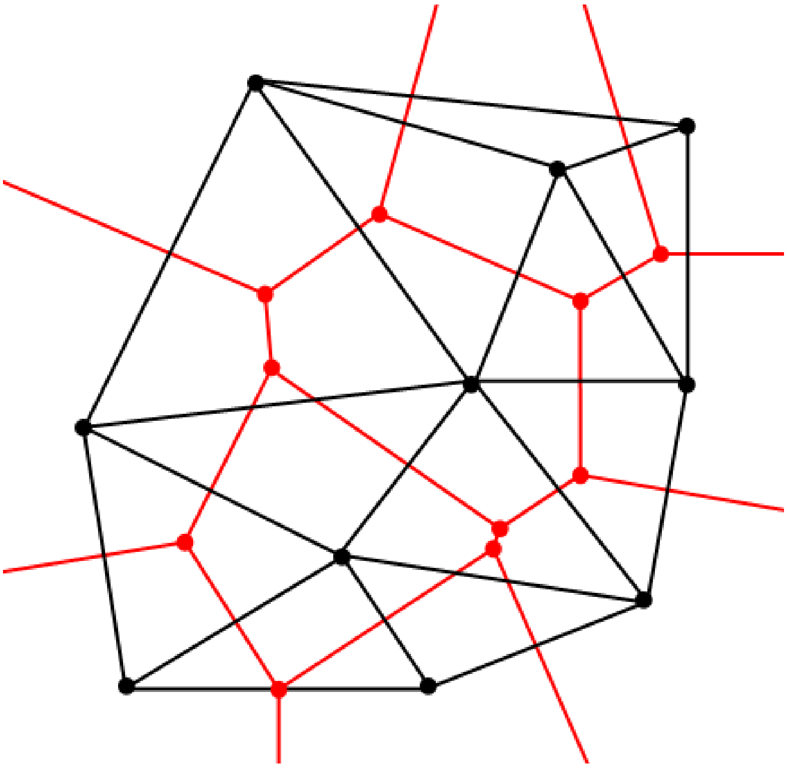


Figure 2.1: Voronoi/Delaunay tessellation in 2D space (Voronoi tessellation - red line, Delaunay tessellation - black line).

There are also three more contact types/definitions that do not comply with the vocabulary presented so far. CSU definition actually refers to the Contacts of Structural Units (CSU) software [276]. CSU defines contacts based on the contact surface area of putative interacting atoms and the atoms' biochemical properties. MC$_d$ and MC$_c$ are the discrete and continuous versions of a contact definition developed by Maiorov and Crippen [61, 181]. In this definition, only the backbone atoms (N, C', O) and C_β are considered. The following distance requirements $d(O, N) < 3.2$Å and $d(C', N) > 3.9$Å, $d(N\,or\,O, C_\beta) < 5.0$Å, and $d(C_\beta, C_\beta) < 9.0$Å define a backbone-backbone, a backbone-side-chain and a side-chain-side-chain contact, respectively. The last two types of contact are valid when there is no interfering atom: any atom closer than 1.4Å to the line that connects the interacting atoms. In the continuous case, lower and upper distance cutoff bounds are defined.

The choice of contact type and contact definition constitutes a basic RIG definition. For distance cutoff based contact definitions, the choice of the actual cutoff value is highly dependent on the choice of contact type. A basic RIG definition can be further parameterised by choosing a particular contact range.

Contact range

Contact range defines whether a residue-residue interaction is filtered out according to sequence separation or secondary structure assignment. Sequence separation is the absolute difference of the residues' sequence numbers and if the sequence separation is less than a threshold, the contact might not be taken into account. For example, in $(C_\alpha)^{\mathrm{|i-j|\geq 3}}_{8.0\mathrm{\mathring{A}}}$ RIG all interacting residues have sequence separation 3 or more. The motivation behind sequence separation filtering depends on the actual threshold used (see Section 2.3.5). The secondary structure assignment can also be used to filter explicitly contacts within the same secondary structure fragment (e.g. $(C_\alpha)^{\mathrm{s_i \neq s_j}}_{8.0\mathrm{\mathring{A}}}$). The all superscript denotes that no filtering is applied (e.g. $(C_\alpha)^{\mathrm{all}}_{8.0\mathrm{\mathring{A}}}$).

Contact weight

It must be pointed out that here all RIGs are treated as unweighted, undirected networks. However, in some articles the original RIGs are actually weighted. An interaction between residues i and j can be weighted by the number of atomic contacts [1, 327] or in a directed fashion by the number of atoms of residue j interacting with i normalised by the expected maximum atomic density (the average over a representative set of proteins of the maximum number of interacting atoms with a residue of same type with i) [121, 122]. The latter weighting scheme removes any bias due to large side-chains. It can be also modified so that weight is the same independent of direction. In the modified version (called interaction strength), the number of atomic contacts is used instead of the number of atoms of j interacting with i [147]. The sum of weighted atomic contacts can also be used to bias towards highly specific contacts: e.g. by weighting more the side-chain-side-chain atomic contacts than the backbone-side-chain ones [247]. Finally, another variation of contact weight is the use of a sigmoid or linear function of the corresponding distance in space that transforms boolean values to continuous ones [61, 95, 103, 122, 152, 154, 181].

2.2.3 Annotation and analysis

Contact type and contact definition are highly interrelated. On the contrary, the choice of contact range is per se independent of the choice of contact type and definition. However, there are cases where the selection of the contact range is result of optimising the RIG definition with respect to a specific application. Despite these cases, contact type and definition are annotated together while contact range separately. Moreover, sequence separation thresholds implicitly classify contacts to short-, medium-, and long-range ones. In addition to annotating the values of sequence separation thresholds, the corresponding contact classification is recorded as well. In the case of a single threshold, all contacts below the threshold are considered to be short-range contacts excluded from the RIG, while contacts above the threshold are long-range taken into account. In the case of a second threshold, the contacts between the two thresholds are short-range included in the RIG and in case of a third threshold, the contacts between the second and the third are medium-range. Any contact classification specified within an article supersedes the previous classification scheme.

Articles concerning distant-dependent potentials are ignored in general. However, if there is a maximum distance cutoff above which all interactions are ignored, then this cutoff is considered as the contact definition choice. Also in cases where sequence separation thresholds are explicitly used to discriminate between short-, medium-, and long-range contacts, contact range is annotated as well. Finally, for articles where the RIG definition is optimised with respect to a specific application, only the optimal definition is considered.

The number of occurrences of a RIG definition is calculated based on the number of last authors that use the specific definition and not on the actual number of articles annotated with that definition. In this way, RIG definitions can be analysed without any bias due to redundant articles with the same definition and published by the same research group. The name of the last author was preferred to discriminate among research groups as the last authors (128) are less than the first authors (164).

2.3 Results

2.3.1 Network perspective of structural bioinformatics

In total, we collected 220 articles and we extracted 70 basic RIG definitions and 18 contact ranges. The definitions along with the articles are listed in Appendix Tables A.1 and A.2. The research topics encountered cover every aspect of the protein structural bioinformatics. Protein structural alignment, domain decomposition, structure prediction and reconstruction, stability, function, folding, fold designability and contact map prediction are the main areas where coarse-grained representations have been extensively used.

The contact map was introduced as a manageable representation that outlines the tertiary topology in the form of a characteristic pattern of secondary elements [218, 233]. It was utilised to visually inspect putative nucleation centers [233] and to compare alternative or homologous conformations [215, 218]. Methods based on the distance matrix similarity [126, 270] and the contact map overlap [46, 138, 156, 326] were successfully developed for protein structure alignment and comparison. Recently, it was shown that it is possible to align structures using a vectorial representation that is based on contact map's eigenvalues [286].

Apart from protein comparison methods, approaches for automatic domain decomposition use contact maps and networks [2, 60, 79, 90, 95, 127, 272, 295, 327]. Evaluation of the putative domains is usually based on the underlying principle that domains are highly compact with intra-domain contacts dominating the network. Domain identification was successfully formulated as network problem solved by identifying the minimum cuts [327], maximising the cycle distributions [79], and clustering using a graph spectral method [272].

Although contact maps are reduced representations, they contain sufficient information to successfully reconstruct the three-dimensional structure. Distance geometry [62] based methods [11, 33, 117, 118, 238, 254], monte carlo approaches [172, 305, 306], heuristic methods [297, 299, 300], and discrete molecular dynamics [52] accomplish this task. Interestingly, it is possible to obtain a good reconstruction even from incomplete or noisy contact maps [52, 257, 299, 305] and to rationally identify a minimal subset of contacts for optimal structure recovery [257].

Vectorial representations extracted from contact maps also allow for accurate 3D reconstruction. The underlying structure can be recovered from: a) the contact number vector, a vector containing the number of contacts for each residue (known as contact number, degree, and coordination number) [254], b) the contact number vector combined with two other vectors that contain per each residue the average sequence separation over all its contacts (known as residue-wise contact order) and the secondary structure assignment [153], c) the contact map's principal eigenvector, the one corresponding to the maximum eigenvalue [239], and d) the effective connectivity profile that is a linear combination of the eigenvectors [324].

For over 30 years, pseudo-energy functions have been derived from empirical analysis of contact maps and have been widely applied in template based as well as template free structure modelling. Statistical analysis of the observed interactions is used for deriving such functions [119, 141, 155, 204, 205, 271, 282]. Alternatively, the energy parameters are optimised to select the native structure as the most energetically favourable among decoy structures [181, 203, 307]. Knowledge-based potentials usually are two-body in the sense that they define the energy of residue-residue interactions. Single-body potentials denote the propensity of a specific residue type to have a specific contact number. Such potentials correlate well with residue burial and hydrophobicity and can be utilised either alone or in combination with higher-order potentials. Four-body potentials [88, 94, 159, 209, 269] as well as whole-residue ones [171, 187] have been developed to allow for more cooperative models.

Contact potentials can also be applied to predict stability changes upon mutation [50, 151]. The contact number of the mutated residue correlates well with the change in stability [110, 111, 180]. Clustering [41, 121, 122], network analysis [54], as well as machine learning approaches [47, 48] are able to identify residues or regions critical for maintaining structural stability. Increased thermostability is also related to certain topological properties. Thermophilic proteins have more contacts at the solvent accessible surface [99], more highly connected residues [249], more long-range contacts [108], and higher average sequence separation over all interactions [249, 250], compared to mesophilic homologs.

The number of topological properties related to folding kinetics is overwhelming. Many network properties correlate well with folding rates: the average sequence separation over all interactions (contact order) [234], the fraction of short range contacts [202], the average number of long range contacts per residue (long range order) [114], the product of contact order and long range order (total contact distance) [335], the number of non-local contact clusters [146], the degree-degree correlation [15], and the fraction of residues with many long-range contacts [109]. Folding rates can be also predicted from a simple model based on the average shortest path over all contacting residues (effective contact order) [75, 320] and from a statistical mechanical model whose energy function is partially based on a contact map [220]. The impact of edge removal, that demonstrates the robustness of a network to maintain its average shortest path length, correlates with protein unfolding rates [144]. Contact order and long range order can successfully discriminate between two-state and multi-state proteins [177]. Moreover, $\alpha\,/\,\beta$ proteins have more contacts due to their compactness and thus, higher contact order and slower folding rates compared to proteins of the same size and of different structural class [93].

Graph-theoretical measures have proved useful in the identification of critical residues for function as well as functional and viable split sites in proteins. Network analysis has provided also a broader insight into protein folding. The importance of RIGs in understanding protein folding and function has been discussed in detail in Section 1.6.

The shape of sequence evolution has been related to topological characteristics of protein structural space [80, 81, 262, 291]. The sum of a series in traces of the powers of the contact matrix (i.e. the total number of length-2 closed loops, the total number of length-3 closed loops, etc.) determines fold designability. Structures with higher sum can be encoded by a higher number of sequences. This sum can be approximated by its first term that is actually the average contact number or by the maximum eigenvalue of the contact matrix.

It has become clear how important contact maps, networks and corresponding vectorial representations have been over the last 40 years. Inevitably, prediction of these reduced representations from sequence has grown into a field of its own importance. Contact map prediction is a separate category in the Critical Assessment of Techniques for Protein Structure Prediction (CASP) [136]. Many contact map prediction methods use correlated mutation analysis [83, 86, 87, 100, 116, 165, 222, 264, 267, 288, 309]: structurally and functionally important residues that are close in space should co-evolve to conserve the fold and the function, leading to highly co-varying positions in multiple sequence alignments. When structural information is included, correlated analysis is actually combined with empirically derived contact propensities [83, 267]. More sophisticated methods are based on machine-learning approaches resulting often in more accurate predictions [53, 86, 87, 116, 178, 197, 235, 238, 243, 263, 285, 313, 325]. Some of the methods use correlated mutations as training data [86, 87, 116]. Various methods have been also developed for the prediction of contact number [152, 154, 216, 236, 237, 331], residue-wise contact order [154, 277] and principal eigenvector [313] from sequence.

The forementioned methods are complementary rather than redundant with respect to structure prediction approaches. Given that structure can be reconstructed from contact maps, it is extremely important that there are cases where the best contact predictions are better than 3D model predictions and that consensus contacts can be derived from predicted models [136]. Predicted residue interactions [197] and predicted contact numbers [135] have also been used for model selection. Finally, folding rates can be estimated from predicted contact maps [244] as well as from predicted residue-wise contact orders.

2.3.2 Contact types

The most striking result from the analysis of contact types is that the top five utilised types cover more than 90% of all occurrences (Figure 2.2A). These are the C_α, ALL, C_β, SC, and SC_c in decreasing order of popularity. The SC_c contact type is almost exclusively used in developing potentials.

Single/dual atom residue representations are ∼1.6 times more encountered than multi-atom ones while the preference for more coarse-grained representations seems to be independent of publication year (Figure 2.2B). Single real atom contact types were originally preferred compared to multi-atom or centroid ones due to noticeable computational speedup and limited resources. Additionally, a considerable amount of the structures deposited in Protein Data Bank (PDB) [29] used to be of low quality, with only C_α coordinates reported or with many SC atomic coordinates missing. Although these restrictions do not apply anymore, single atom contact types still occur frequently in the literature. Contact map prediction methods predict C_α or C_β mediated contacts and thus, structure recovery methods focus on these contact types. Protein models submitted to CASP are mainly evaluated based on the C_α coordinates and developed model selection methods utilise the same coordinates as well.

The popularity of a contact type is not related to is utility. C_α is fully dependent on backbone conformation, while C_β is sensitive to side-chain directionality. However, C_β atoms are still fixed by residue's backbone. SC_c considers both directionality and differences in shape and volume of the side-chain. In general, backbone contacts are not sequence specific due to chemical identity of the backbone for each residue. Moreover, multi-atom contact types are more fine-grained representations of the protein structure compared to single-atom ones. These observations are reflected in the performance of different contact types in the context of potentials. C_β performs better than C_α [34, 148, 155], SC_c better than C_β or C_α [155, 280], SC better than SC_c [322] or C_α [166], and ALL better than C_α [23, 307].

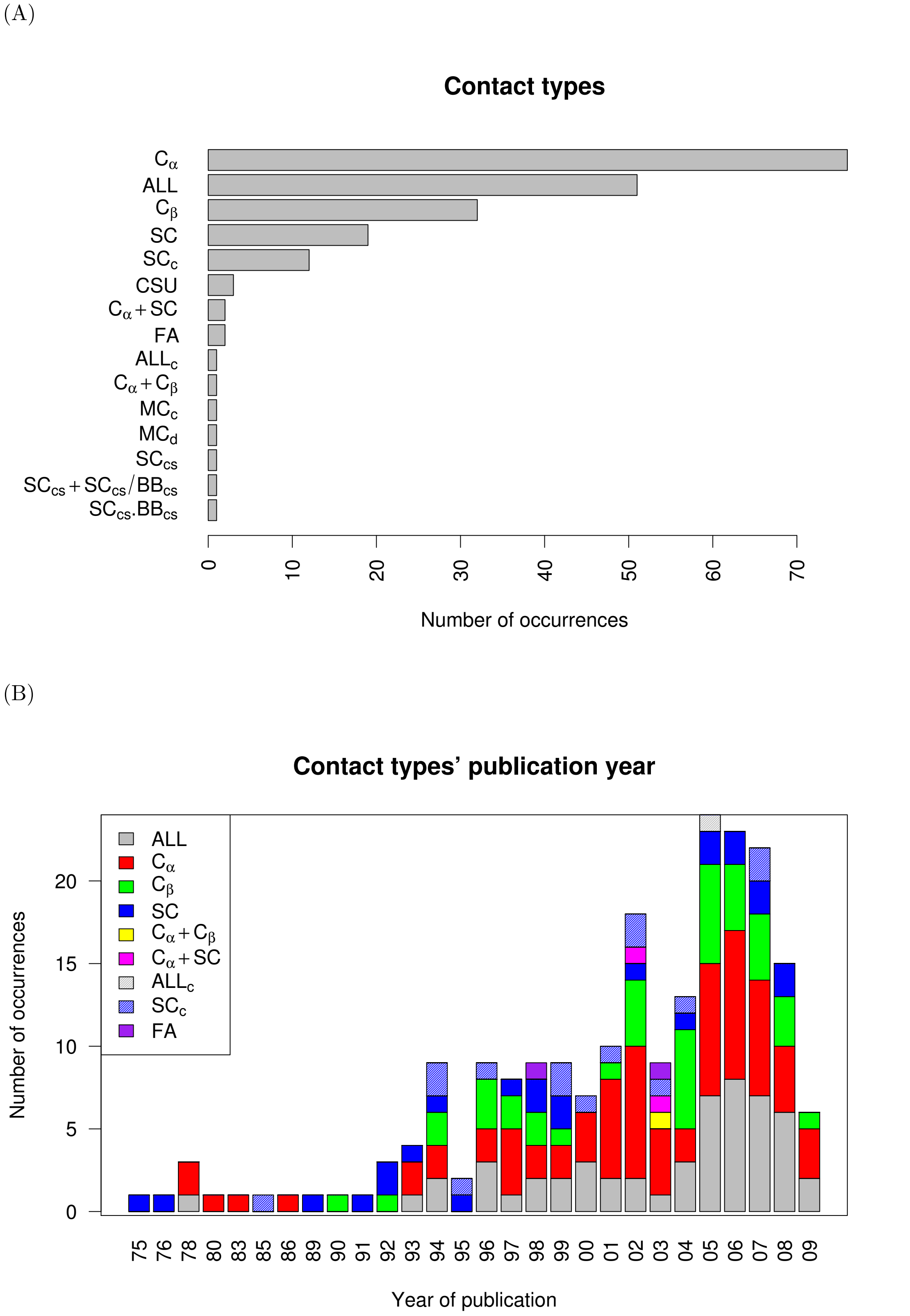


Figure 2.2: Distributions of the contact types (A) and their publication year (B).

2.3.3 Contact definitions

Surprisingly enough, distance cutoff (DC) based definitions dominate the literature (Figure 2.3A). As explained later in Section 2.3.4, van der Waals (r_{vdW}) definitions are more physically meaningful especially in the case of multi-atom contact types, while Delaunay (DT) definitions are based on a statistical geometry approach. Despite these facts, these definitions are not commonly deployed. The van der Waals based ones are encountered in only nine articles out of which seven are published before 1999. Moreover, in six out of the nine Delaunay related articles there is a common co-author.

Examining the distance cutoffs utilised, the dependence of the cutoff upon contact type becomes clear (Figure 2.3B). For multi-atom contact types, the cutoff is placed in the lower region [4.0, 6.0]Å and it peaks at 5.0Å. In contrast and for single/dual atom contact types, the cutoff peaks at 8.0Å and generally occurs in the upper region [6.0,12.0]Å. The most commonly used cutoffs for the top 5 populated contact types C_α, ALL, C_β, SC, and SC_c are 8.0Å, 5.0Å, 8.0Å, 4.5Å and 8Å respectively. Unusually high cutoffs are almost all result of optimising the RIG definition with respect to a specific application (see Section 2.3.6).

2.3.4 Basic RIG definitions

The distribution of all basic RIG definitions is presented in Figure 2.4 summarising all previous observations. All definitions with more than five occurrences in the literature are mainly based on C_α and ALL contact types (Figure 2.5). The top three most frequent ones are $(C_\alpha)_{8.0\text{Å}}$, $(ALL)_{5.0\text{Å}}$ and $(C_\beta)_{8.0\text{Å}}$.

The most common physically based motivation behind choosing a particular basic RIG definition is the exclusion of non specific solvent-mediated interactions. Among all contact definitions, the ones based on the van der Waals radii of the potentially interacting atoms are the most physically meaningful. In the rather strict case, two atoms are in contact if the distance between their van der Waals spheres is less than or equal to 0Å. This can be relaxed by increasing the cutoff to 0.5Å [295], 1.0Å [160, 268, 317] or 2.8Å [281, 282] which is the diameter of a water molecule. The increased cutoff allows for slight coordinate errors [268, 317] but it is too short to allow for a third atom to intervene. The overall definition ensures that all atoms in the first contact shell, atoms that are nearest-neighbours in space, are included, while atoms in the second shell are neglected. Similarly, C_α atoms in direct contact are within 6.5Å as the atom radius in this case is 2.0Å [248], with the water-mediated contacts to be placed in the [6.5, 9.5]Å distance interval [227]. In MC_d and MC_c contact types, special care is taken to ensure that there is no interfering third atom, while in the case of multi-atom RIGs such as ALL or SC, the 4.5Å threshold is utilised to guarantee that there is no other residue or solvent molecule between interacting residues [124, 201]. Moreover, the upper limit for attractive London-van der Waals forces is approximated by a 5.0Å cutoff [106, 292].

Miyazawa and Jernigan examined the residue packing in $(SC_c)^{|i-j|\geq 2}$ RIGs with respect to the distance cutoff [204]. The average number of contacts for residues in the interior

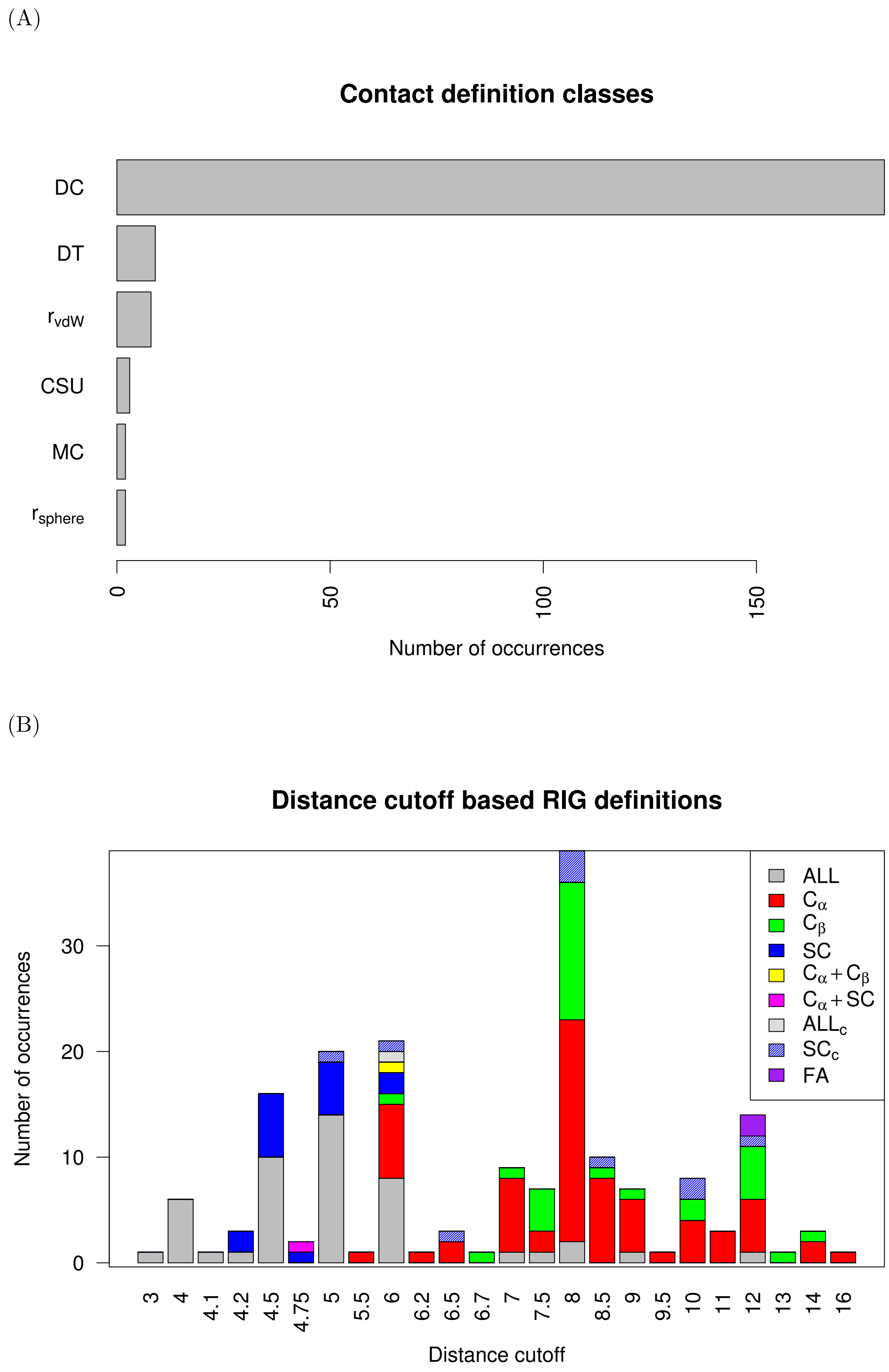


Figure 2.3: Distributions of the contact definitions (A) and the distance cutoffs (B). A. The number of occurrences in the literature of all contact definition classes: Distance cutoff (DC), Delaunay Tessellation (DT), van der Waals (r_{vdW}), Contacts of Structural Units (CSU), Maiorov and Crippen (MC), and sphere (r_{sphere}) based definitions.

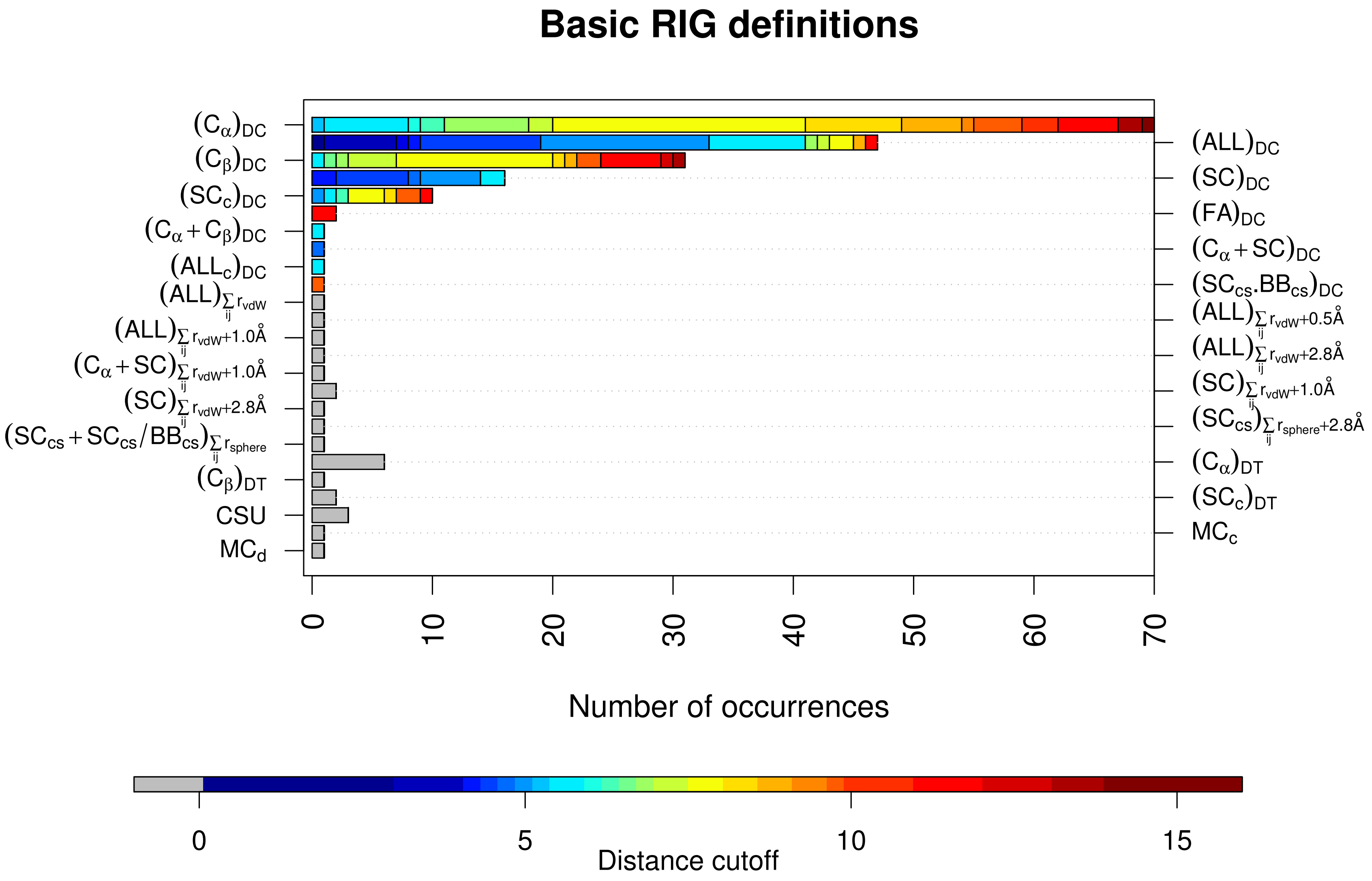


Figure 2.4: Distribution of the basic RIG definitions.

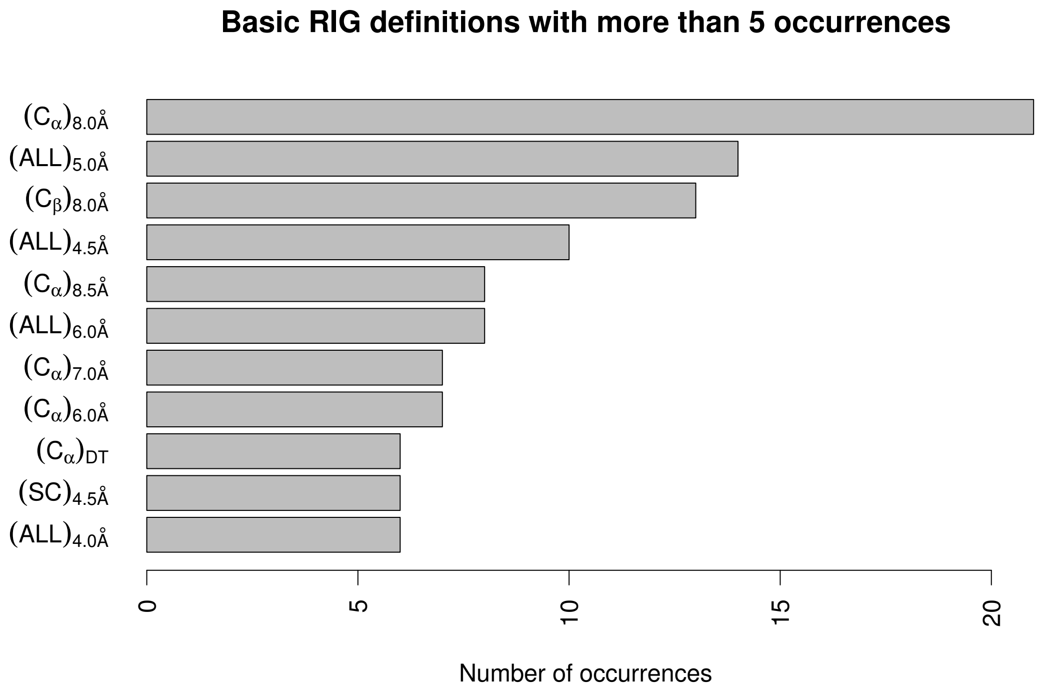


Figure 2.5: The top 11 most frequent basic RIG definitions with more than 5 occurrences.

of proteins peaks at 6.5Å and has the closest value to the expected number of contacts under the smoothed density assumption. For larger distances, the area outside protein surfaces is covered and the radial distribution becomes lower than one. The radial distribution function was studied as well by Atilgan et al. for $(C_\beta)^{\text{all}}$ networks [12]. The first local minimum or simply "hump" occurs at 6.7Å denoting the region of inter-residue contacts of the highest probability. Second, third, and fourth coordination shells were shown to be located at 8.5Å, 10.5Å, and 12.0Å, respectively. Chea and Livesay used 8.5Å threshold for C_α RIGs as the best approximation for the average side-chain size [51].

Manavalan and Ponnuswamy analysed the preferred environment of each of the 20 different amino acids and its non-polar nature [183]. Based on a $(C_\alpha)^{|i-j|\geq 4}$ reduced representation and varying distance threshold, they showed that the influence of each residue extends effectively up to 8.0Å: between 6.0Å and 8.0Å there is a clear preference for hydrophobic interactions while above 8.0Å energetically unfavourable interactions become statistically favourable.

Finally, the motivation behind preferring Delaunay tessellation over the other contact definitions is that it provides an objective and robust definition of nearest neighbours in three-dimensional space and it is not dependent on any specific parameterisation. Especially for large distance cutoffs, the number of interactions for a residue becomes larger than the number of actual geometric nearest neighbours.

2.3.5 Contact ranges

Filtering based on secondary structure assignment occurs only in two articles and thus, are excluded from the Figures in this section. In these articles, the RIGs are actually transformed to networks where nodes are secondary structure fragments [79, 156].

The distribution of sequence separation thresholds is presented in Figure 2.6A. In almost half of the cases, all inter-residue contacts are taken into account ($|\mathrm{i}-\mathrm{j}| \geq 1$). Sequence thresholds 2 and 3 are used to remove the backbone connectivity and contacts mainly defined by it. Higher thresholds 4 and 5 exclude effects of the helix/turn content. The repeating structural unit of an α-helix is 3.6 residues in length and α-helices are dominated by (i,i+3) and (i,i+4) inter-residue hydrogen bonds. At the same time, a turn is in general formed by 3 or 4 residues. Sequence separation of 10 is used as a conservative estimate based on the average length of an α-helix (11) and β-strand (6) [106]. In this way, interactions between residues within the same secondary structure fragment are removed and only contacts important for the tertiary structure are taken into account. Sequence separation thresholds 6, 12 and 24 are utilised for evaluation of contact map prediction in CASP. The probability of two residues to be in contact decreases with increasing sequence separation. Sequence-distant contacts are harder to predict and are expected to be more useful as constraints for structure prediction. Threshold 13 is result of optimising the RIG definition for prediction of folding rates [114], while the highest cutoff 31 is encountered in potentials.

Figure 2.6B shows for each RIG definition how many sequence separation thresholds are used if any, how these classify contacts into short-, medium-, and long-range ones and which contacts are excluded from the actual network. It becomes obvious that when filtering out contacts, a single threshold is commonly used to exclude the short-range ones. Rarely, contacts are classified into short-, medium-, and long- range contacts as for example in CASP (short: $6 \leq |\mathrm{i}-\mathrm{j}| < 12$, medium: $12 \leq |\mathrm{i}-\mathrm{j}| < 24$, long: $|\mathrm{i}-\mathrm{j}| \geq 24$).

2.3.6 Optimal RIG definitions

Apart from physically motivated definitions, the curated data set contains optimised definitions with respect to a specific research problem such as predicting folding rate and developing a potential. Such definitions may give rise to unusually high distance thresholds previously observed, while they heavily depend on the particular data sets and the overall methodology utilised. The optimisation usually refers to the choice of a distance threshold or the contact range with the remaining two criteria being fixed.

Plaxco et al. outlined that variation of the distance threshold in $(ALL)^{\mathrm{all}}$ RIGs (from 3.5Å to 8.0Å) does not significantly affect the correlation of the relative contact order with folding rate [234]. In contrast, Mirny and Shakhnovich showed that relative contact order has lower correlation when the cutoff is smaller than 6.5Å or larger than 9.0Å [202]. The highest correlation for the fraction of short-range contacts ($(ALL)^{|\mathrm{i}-\mathrm{j}|\geq 4}$) is also achieved at 7.0Å. Long range order calculated from $(C_\alpha)_{8.0\text{Å}}$ RIGs has the best correlation with folding rates when long range contacts are defined as the ones with

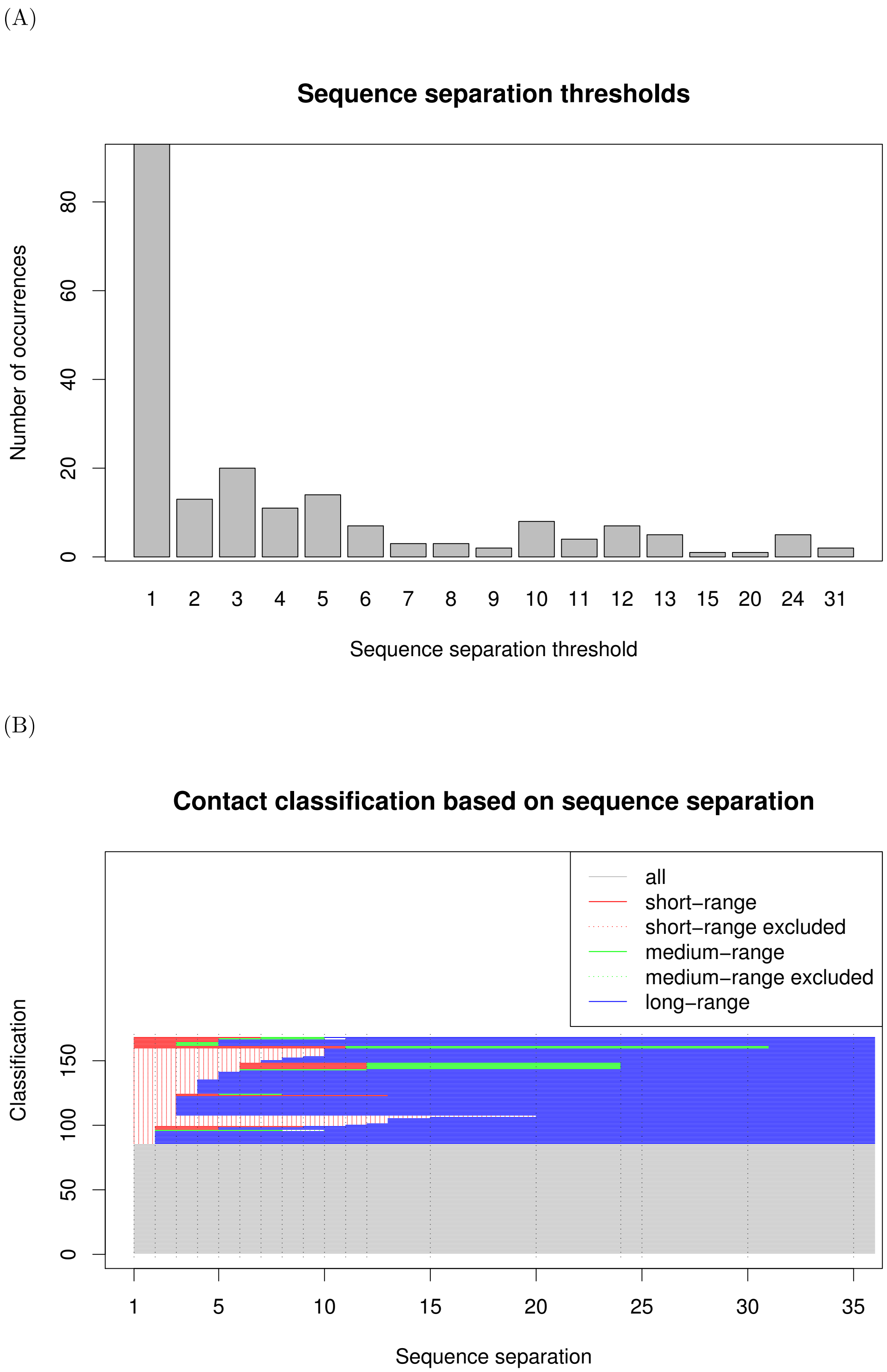


Figure 2.6: Distributions of the sequence separation thresholds (A) and the related contact classifications (B).

sequence separation at least 13 [114]. Among various $(C_\alpha)^{|\mathrm{i-j}|\geq 13}$ RIG definitions the one with 8.0Å cutoff gives the best correlation as well. The sequence separation is further optimised to 28, 45 and 11 for $all-\alpha$, $all-\beta$ and mixed proteins respectively. Gromiha also analysed the correlation for the number or fraction of well-connected residues in C_α RIGs with respect to both the distance cutoff and the contact range [109]. $(C_\alpha)^{|\mathrm{i-j}|\geq 13}_{7.5\mathring{\mathrm{A}}}$ and $(C_\alpha)^{|\mathrm{i-j}|\geq 4}_{6.5\mathring{\mathrm{A}}}$ RIGs gave the best correlation to folding rate for two-state and three-state proteins respectively. Zhou and Zhou utilised contacts of sequence separation at least 15 ($(ALL)^{|\mathrm{i-j}|\geq 15}_{6.0\mathring{\mathrm{A}}}$) for better prediction of folding rates based on the total contact order [335].

Kinjo et al. [152] and Yuan et al. [331] used 12.0Å as the optimal cutoff radius of $(C_\beta)^{|\mathrm{i-j}|\geq 3}$ RIGs for the prediction of contact numbers from sequence. The contact number of a residue shows the highest anti-correlation with its distance from the center of mass of a protein for 14.0Å distance cutoff in C_α RIGs [217]. Similarly, the residue burial expressed as the contact number yielded the best performance in fold recognition at 14.0Å cutoff for C_β RIGs [148]. Bolser et al. thoroughly examined different RIGs definitions with respect to contact type, distance cutoff and contact range, and the performance of the extracted single-body and two-body potentials in discriminating native structures from decoys [34]. C_α, C_β, $C_\alpha + C_\beta$ RIGs, including and excluding short range contacts (all and $|\mathrm{i}-\mathrm{j}| \geq 10$ contact ranges), and with varying distance threshold were compared. The best performance for the single-body and two-body potentials was obtained using $(C_\beta)^{\mathrm{all}}_{14.0\mathring{\mathrm{A}}}$ and $(C_\beta)^{\mathrm{all}}_{12.0\mathring{\mathrm{A}}}$ respectively. Vendruscolo et al. demonstrated that a two-body contact potential extracted from $(ALL)^{\mathrm{all}}$ RIGs best stabilises native proteins against decoys for 4.5Å distance cutoff [307]. In the case of $(C_\alpha)^{\mathrm{all}}$, the optimal threshold is 8.5Å. Bastolla et al. also optimised the RIG definition for an energy function [23]. $(ALL)^{|\mathrm{i-j}|\geq 3}_{4.5\mathring{\mathrm{A}}}$ and $(C_\alpha)^{|\mathrm{i-j}|\geq 3}_{11.0\mathring{\mathrm{A}}}$ guarantee the highest stability among various distance thresholds. Berrera et al. analysed the performance of two-body potentials in fold recognition utilising RIG definitions based on van der Waals radii [30]. Among the C_α, C_β, BB, $C_\alpha+SC$, and ALL contact types and with varying cutoff, $(C_\alpha + SC)^{\mathrm{all}}_{\sum_{\mathrm{ij}} \mathrm{r}_{\mathrm{vdW}}+1.0\mathring{\mathrm{A}}}$ was the optimal RIG definition.

Williams and Doherty demonstrated that pairwise contact propensities correlate well with evolutionary substitution costs [322]. The correlation is higher for SC (SC_c) potentials for 4.5Å (8.0Å) distance cutoff. Lin et al. trained a neural network to calculate the probability of a residue of a certain type to be in a given structural environment based on a $(SC_{cs}+SC_{cs}/BB_{cs})^{\mathrm{all}}_{\sum_{\mathrm{ij}} \mathrm{r}_{\mathrm{sphere}}}$ RIG definition [171]. The default radius for the side-chain sphere was set to 4.0Å to minimise the uncertainty about the type of the central residue given the residue's structural environment. Thomas et al. applied a distance-dependent potential based on C_β contacts to protein recognition tests with the optimal upper distance threshold found in the [11.0, 13.0]Å region [289]. Kuznetsov and Rackovsky tested the performance of distance-dependent C_β and SC_c potentials on the recognition of the residues' preferred environment [166]. The optimal upper bound for distance was found to be 12.0Å.

The contact number of a mutated residue as the number of methyl(ene) groups within 6.0Å, correlates well with the change in stability independent of the extent of burial of the mutated residue [180]. For $(C_\alpha)^{\mathrm{all}}$ RIGs the highest correlation of the contact number with stability changes occurs at 8.0Å, 6.0Å and 7.0Å for mutated residues in helical, strand and coil segments respectively [111]. Capriotti et al. developed machine learning approaches to predict whether a mutation is stabilising, destabilising or neutral based on the residue's structural environment [47, 48]. This environment as encoded in $(ALL)^{\text{all}}_{9.0\text{Å}}$ RIGs gives rise to the best prediction accuracy.

Crippen analysed the C_α RIG definition with respect to both the distance cutoff and the sequence separation for the protein structure decomposition into domains [60]. Domains are defined based on a hierarchical tree of sequence segments and for the $(C_\alpha)^{|i-j|\geq 7}_{9.0\text{Å}}$ definition, the segments best match the secondary structure assignment. Vassura et al. investigated how native-like the reconstructed protein structures from $(C_\alpha)^{\text{all}}$ RIGs are with respect to the distance threshold [300]. Reconstructions are more similar to the native structure when the threshold is within [10.0, 18.0]Å. Caprara et al. compared the automatic clustering of proteins into families based on $(ALL)^{\text{all}}$ contact maps and the contact map overlap, with the SCOP [210] classification [46]. 7.5Å was shown to be the best distance threshold. Finally, Cusack et al. showed that $(ALL)^{\text{all}}_{5.0\text{Å}}$ RIGs give rise to the most reliable prediction of functional residues compared to alternative distance thresholds [64].

2.4 Discussion

Collecting all articles where coarse-grained representations of protein structures are utilised, annotating the RIG definitions with a controlled vocabulary, and analysing the data is a daunting task. The lack of consistent annotation and the inevitable manual curation of articles may lead to a highly biased data set of RIG definitions. Here, the collected articles are relatively non-redundant with respect to the corresponding research groups. 259 researchers are cited either as first or last authors compared to the expected 431 ones for a totally non-redundant collection. The diverse research areas covered guarantee as well the generality of the data set. Although it is difficult to draw conclusions of general applicability, outlining the principal criteria used to define a residue-residue interaction and examining the selections so far and the motivation behind each one is crucial for understanding the ruling principles of reduced representations. This review and the related manually curated data set of RIG definitions is the only one publicly available so far. It can be used in order to make an informed choice of RIG definition instead of arbitrarily adopting a representation. The future extension of the controlled vocabulary to include information about the application field and the justification of the definition will facilitate a more rigorous comparison.

Chapter 3

Systematic comparison of network representations of protein structures

Summary

To date, limited analysis has been performed to rationalise the choice of the network representation of a protein structure. A randomly chosen definition will not necessarily exhibit identical network properties with certain other definitions and will not reproduce published results based on them. Here, we establish a unifying view for the network representations of protein structures. We compose a non-redundant, representative and of high quality data set of 60 proteins that ensures the generality of our analysis. We utilise an exhaustive data set of 945 RIG definitions and assess quantitatively the similarity of the resulting networks as well as fundamental properties for their connectivity. We demonstrate that the similarity between commonly used network representations can be in certain cases quite low. We investigate the impact of the RIG definition upon similarity and connectivity and in conjunction with protein structural class. In particular, when only long-range interactions are considered, RIG definitions usually exhibit lower similarities to each other, more residues do not have any interactions and residues may lie in "separate" components. The impact upon similarity and connectivity is more severe for *all-α* proteins than for *all-β* proteins as short-range interactions are dominant in helical structures. Based on the work presented here, researchers will be able to make an informed choice of representation necessary to achieve the desired network properties and to rationally compare results produced using different representations. Moreover, we provide open source software tools for converting protein structures to networks and for subsequent network analysis. To our knowledge, this is the first study that addresses the challenge of establishing the overarching principles of network representations by a large-scale, fine-grained comparison and analysis of RIGs.

3.1 Introduction

Despite the fact that network analysis of protein structures has been extremely successful, limited analysis has been performed to establish the overarching principles of the network representations. Different researchers adopt often arbitrarily different representations and often neglect to demonstrate the impact of the representation itself upon results. The impact of RIG definition has only been studied with respect to the form of the degree distribution and the small-world character and for a limited set of distance cutoffs, contact types and contact ranges (see Section 1.5).

The lack of a unifying view prompted a systematic analysis of various network representations. Here, we utilise an exhaustive data set of 945 RIG definitions and assess quantitatively the similarity of the resulting networks as well as fundamental properties for their connectivity. We investigate the impact of the RIG definition upon similarity and connectivity and in conjunction with protein structural class. Analysis in Chapter 2 facilitates the selection of a RIG definition among the ones utilised in the literature based on their justification, their popularity and their optimality for a certain research problem. Based on the work presented here, it is feasible to make an informed choice of representation necessary to achieve the desired network properties and to rationally compare results based on different representations.

To our knowledge, a large-scale analysis of RIG definitions with respect to their similarity and fundamental network properties has not been previously carried out.

3.2 Methodology

Here, we present how we perform a fine-grained analysis and comparison of RIGs. We discuss in detail how we select a representative and of high quality data set of 60 protein structures and 945 RIG definitions according to which 56,700 RIGs are constructed. We explain how we systematically assess the similarity of RIG definitions, how we provide general rules for the similarity of two contact types, and the network properties calculated for each RIG.

3.2.1 The data set of protein structures

To allow for a systematic analysis of a wide range of RIG definitions (see Section 3.2.2), a small, non-redundant, representative, and high quality data set of 60 protein structures was selected. A series of filtering and validation steps were performed based on various publicly available databases [29, 36, 39, 210] to eliminate any biases with respect to RIG parameters and their analysis, achieve quality goals, minimise redundancy and maximise coverage.

Initially, three publicly available databases were merged; Macromolecular Structure Data Search Database (MSDSD) [36] (MySQL version released on February 2006 containing PDB data as of April 2006), SCOP database (version 1.73 released on November 2007) and remediated PDB data (as of October 2007). Out of the 35,936

PDB entries (84,356 polymer chains) associated with valid biological units in MSDSD, 30,872 (66,757) remained after data integration. This was due to obsolete PDB entries, proteins without SCOP annotation and various inconsistencies across the databases.

The merged data set was filtered in several ways (Table 3.1). All protein chains stabilised by interactions that could not be possible taken into account in the RIG construction were not considered. Protein chains that according to MSDSD, their corresponding molecules can be assembled in non-monomeric biological units or other molecules are bound to them in monomeric assemblies, were removed. Disulphide-stabilised chains, multidomain chains, as well as folds not assigned to one of the main four structural classes, were filtered. For practical reasons, apart from the multidomain chains, monodomain ones consisted of more than one fragments, were removed as well.

To select structures of high quality, more filtering criteria were applied: (1) structures are solved by X-ray crystallography; (2) the resolution is better than 3.0Å; (3) the R factor is lower than 0.3; (4) sequences do not include any unknown or non-standard amino acids; (5) all backbone and side-chain atomic coordinates must be available for observed residues; (6) chains with any unobserved non-terminal residues are not allowed; (7) chains containing amino acids of multiple locations ("altLoc" field in ATOM records according to Protein Data Bank (PDB) [29] file format) are not allowed. Criteria 4-7 ensured additionally an unbiased analysis of RIGs. Missing or ambiguous conformational data affect the comparison of RIGs with respect to the representation of the residue and the corresponding atomic interaction sites. Moreover, the presence of unobserved residues bias the effect of filtering contacts based on residues' sequence separation. Out of the 30,872 PDB entries (66,757 chains), only 365 (534) were left after applying all criteria.

Table 3.1: Summary of the selection process of the data set of protein structures. The number of the PDB entries and chains in the merged data set that satisfy the corresponding criteria is given. The corresponding percentage is reported in parenthesis.

	Merged Data Set			
Selection	PDB Entries		PDB Chains	
All	30872	(100)	66757	(100)
Monomeric proteins not stabilised by any bound molecule	4797	(16)	5712	(9)
No disulphide-stabilised proteins	22989	(75)	50766	(76)
Single-fragment monodomains	24180	(78)	49366	(74)
Domains of one of the four main structural classes	27663	(90)	58539	(88)
X-ray proteins	26746	(87)	61408	(92)
Resolution better than 3.0Å	25071	(81)	54381	(82)
R-factor lower than 0.3	26110	(85)	59434	(89)
No unknown or non-standard amino acids	27789	(90)	59585	(89)
Coordinates for all atoms of observed residues	24027	(78)	49659	(74)
No unobserved non-terminal residues	26307	(85)	55160	(83)
No alternative coordinates	26056	(84)	57455	(86)
Remaining	365	(1)	534	(1)

The final selection was restricted to 60 proteins. Each structure was selected from different SCOP fold to ensure non-redundancy and conformational diversity. Although the SCOP database is depleted in *all-α* domains, a data set of 15 proteins for each one of the four main structural classes (*all-α*, *all-β*, $\alpha\,/\,\beta$, $\alpha + \beta$) was preferred. Such a balanced data set would facilitate RIG analysis with respect to structural class. From the folds of the remaining 534 domains and to maximise coverage, the 15 most populated domains were selected for each of the four structural classes. The population of a fold was defined based on the number of non redundant domains having that fold. This was calculated from the ASTRAL [39] set of domains having less than 40% sequence identity to each other. To break a tie, the number of domains of that fold with less than 95% sequence identity was utilised. One protein per fold was finally selected biasing towards: (1) similar CATH protein structure classification [63] with respect to the number of assigned domains; (2) better resolution; (3) non intrinsically disordered termini.

The oligomeric state of the selected folds was additionally manually verified by two other sources of data: the Protein Quaternary Structure (PQS) database [120] and the Protein Interfaces, Surfaces and Assemblies (PISA) server [161]. In cases whether the putative momeric state was not confirmed by both, the specific protein chain was removed from the data set and the selection procedure was repeated.

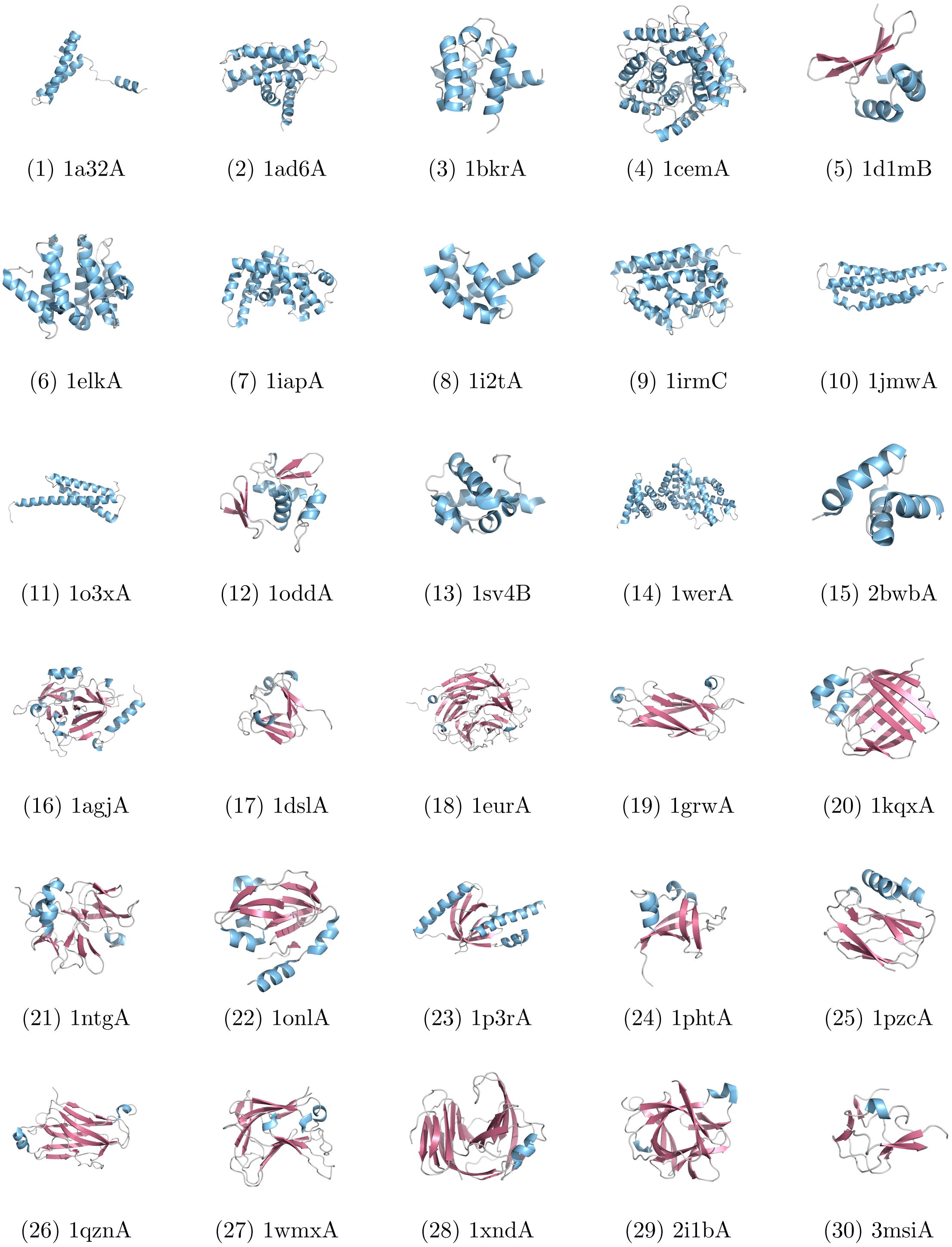
Summary

The final set of 60 protein chains encompasses a non-redundant, diverse in sequence and structure, set of monomeric, monodomain structures. The selected non homologous structures, each one of different fold, cover all main four structural classes equally. The preference for highly populated folds ensures that the selected folds are well studied in the literature, as well as a high coverage of the SCOP data set. The total number of all SCOP domains assigned to one of the four main classes and with pairwise sequence similarity less than 40% (95%) are 8,619 (13,676). The selected 60 folds cover 3,862 (6,503) domains representing more than 44% of all domains for both sequence similarity thresholds. Although the applied criteria for high structural quality are very strict and not commonly satisfied, the high coverage ensures the generality of this study.

The 60 protein folds, their population and the corresponding protein chains are listed in Appendix Table B.1. Ribbon drawings of the selected structures are presented in Figure 3.1. Appendix Figure C.1 shows the protein size distribution for certain size ranges and with respect to the four structural classes: $all-\alpha$, $all-\beta$, $\alpha\,/\,\beta$, and $\alpha+\beta$.

3.2.2 The data set of RIGs

In total, we construct 56,700 RIGs based on 945 RIG definitions for each one of the 60 proteins in the data set. RIG definitions are based on combinations of 9 contact types, 21 contact definitions and 5 contact ranges. The selected criteria cover almost all RIG definitions as occurring in the literature (see Chapter 2) and capture various different features of the protein structure.



(1) 1a32A (2) 1ad6A (3) 1bkrA (4) 1cemA (5) 1d1mB

(6) 1elkA (7) 1iapA (8) 1i2tA (9) 1irmC (10) 1jmwA

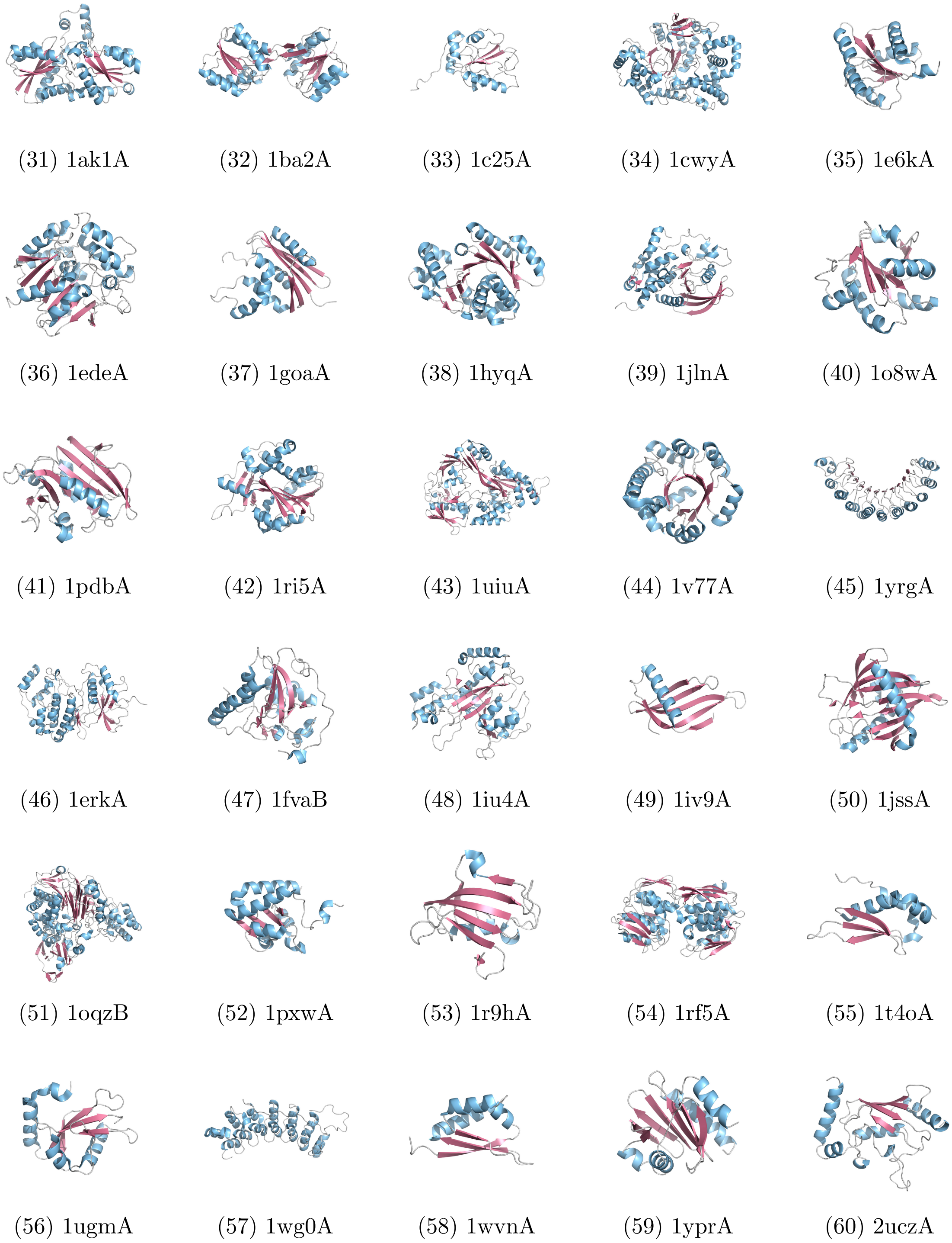
(11) 1o3xA (12) 1oddA (13) 1sv4B (14) 1werA (15) 2bwbA

(16) 1agjA (17) 1dslA (18) 1eurA (19) 1grwA (20) 1kqxA

(21) 1ntgA (22) 1onlA (23) 1p3rA (24) 1phtA (25) 1pzcA

(26) 1qznA (27) 1wmxA (28) 1xndA (29) 2i1bA (30) 3msiA

Figure 3.1: Ribbon drawings of the selected proteins. Proteins 1-15, 16-30, 31-45, and 46-60 are *all-α*, *all-β*, $\alpha\,/\,\beta$, and $\alpha + \beta$ respectively. Structures are labeled with their PDB identifier followed by the chain code. Drawn using distinct colours (blue-red-grey) to indicate different secondary structure conformations (helix-strand-other). (continued on next page)



(31) 1ak1A (32) 1ba2A (33) 1c25A (34) 1cwyA (35) 1e6kA

(36) 1edeA (37) 1goaA (38) 1hyqA (39) 1jlnA (40) 1o8wA

(41) 1pdbA (42) 1ri5A (43) 1uiuA (44) 1v77A (45) 1yrgA

(46) 1erkA (47) 1fvaB (48) 1iu4A (49) 1iv9A (50) 1jssA

(51) 1oqzB (52) 1pxwA (53) 1r9hA (54) 1rf5A (55) 1t4oA

(56) 1ugmA (57) 1wg0A (58) 1wvnA (59) 1yprA (60) 2uczA

Figure 3.1: continued from previous page

Contact types

The following nine contact types are selected: C_α, C_β, C_α/C_β, $C_\alpha.C_\beta$ ($C_{\alpha\beta}$), BB, SC, BB/SC, ALL, and $C_\alpha + SC$. These contact types include backbone-backbone (C_α, BB), backbone-sidechain (C_α/C_β, BB/SC), and sidechain-sidechain (C_β, SC) mediated interactions. Moreover, both single-/dual- atom (C_α, C_β, C_α/C_β, $C_\alpha.C_\beta$) and multi-atom (BB, SC, BB/SC, ALL) residue representations are analysed. $C_\alpha + SC$ has been shown to be the optimal among various contact types with respect to the performance of two-body potentials [30], and so it was included here.

Contact definitions

Although extremely low or high distance cutoffs are rarely used (Figure 2.3B), for the sake of completeness we set cutoffs to range from 2.5Å to 15.0Å. Increments of 0.5Å are used in the range [2.5, 10.0]Å and of 1.0Å in the less common [10.0, 15.0]Å.

Contact ranges

Based on the contact ranges frequently utilised in literature and the motivation behind their choice (see Section 2.3.5), we select five representative contact ranges: either all interactions are taken into account (*all* contact range) or interactions are filtered based on sequence separation thresholds 2, 4 and 10 ($|\mathrm{i}-\mathrm{j}| \geq \{2,4,10\}$) or interactions within the same secondary structure fragment are not considered ($\mathrm{s_i} \neq \mathrm{s_j}$). Non-filtering is far the most frequent choice. Sequence separation threshold 2 removes the sequence connectivity, threshold 4 excludes hydrogen bonds and effects of the helix/turn content and threshold 10 potentially excludes interactions within the same secondary structure based on the average length of α-helix and β-strand. The $\mathrm{s_i} \neq \mathrm{s_j}$ contact range is expected to provide an accurate implementation of the latter.

3.2.3 Similarity

The choice of a similarity measure is a contentious one. Here, we primarily assess the similarity of two RIGs using the *Tanimoto* coefficient T (also known as the extended *Jaccard* coefficient) [284]. The similarity T between two RIGs i and j is defined as:

$$T(i,j) = \frac{N_c}{N_i + N_j - N_c}, \tag{3.1}$$

where N_i and N_j are the number of edges of i and j RIGs respectively and N_c is the number of common edges. The *Tanimoto* coefficient has been widely used in chemoinformatics for comparison of molecules as well as for assessing the similarity of RIGs in CMview, an interactive software tool for RIG visualisation and analysis [302].

We also utilise the *Meet/Min* coefficient M (also known as *Simpson* coefficient), that is defined as:

$$M(i,j) = \frac{N_c}{\min\{N_i, N_j\}}. \tag{3.2}$$

Both coefficients are based on the size of the common edges. $Tanimoto$ normalises the intersection size over the size of the union while $Meet/Min$ over the size of the smaller RIG. Maximum Meet/Min similarity value of 1 means that one graph is subgraph of the other. $Meet/Min$ coefficient allows us to assess the similarity unbiased from any difference in size, i.e. the number of interactions.

Both similarity metrics consider only the presence of interactions. Another class of metrics like the $Hamming$ similarity treats equally the presence and absence of interactions. As RIGs are sparse graphs (for reasonable selected distance cutoffs), such metrics would lead to non-intuitive high similarities due to dominance of non-interacting residues. Moreover, $Tanimoto$ and $Meet/Min$ coefficients are "directional". For example, the distance cutoff value x that maximises the $Tanimoto$ similarity of $(C_\alpha)^{\text{all}}_{x\mathring{\text{A}}}$ with $(C_\beta)^{\text{all}}_{8.0\mathring{\text{A}}}$ does not necessarily imply that 8.0Å is the optimum value for distance cutoff y that maximises $(C_\beta)^{\text{all}}_{y\mathring{\text{A}}}$ with $(C_\alpha)^{\text{all}}_{x\mathring{\text{A}}}$. Although the intersection size is maximum for a specific pair of distance cutoffs, the denominator value in equations 3.1 and 3.2 changes based on the "reference" cutoff. Using the number of all possible interactions as denominator to address this issue, would lead to non-intuitive low similarities due to the sparseness of RIGs.

For each protein and for each contact range, we calculate both $Tanimoto$ and $Meet/Min$ similarity for each pair of RIGs with respect to all pairs of contact types and across all distance cutoffs. Specifically, for each of the five contact ranges and for each of the 36 pairs of contact types we compare 441 RIG definitions with respect to 21 different cutoffs. Therefore, 5 x 36 x 441 = 79,380 pairs of RIG definitions are compared for each protein. For each contact range and for each pair of RIG definitions we calculate the mean similarity over the 60 proteins and construct the similarity matrix. Each cell in this matrix contains the mean similarity for a specific pair of cutoffs, one for each contact type. The similarity matrix is not symmetric.

Best similarity

Defining a single similarity value that "best" describes the overall similarity of two contact types over all pairs of cutoffs and with respect to a certain contact range is a difficult decision. Similarly, the definition of the "best" distance-intercept, the distance difference of the cutoffs that provides the "best fit" between two contact types is far from trivial. Both definitions are important as they can provide general rules independent of the choice of cutoffs and facilitate the comparison of RIG definitions across pairs of contact types. Various factors influence the choice for these definitions. The "best" similarity and distance-intercept must be "undirectional" as opposed to the similarity metrics mentioned above. The similarity value is expected to increase as the distance cutoff increases and thus its maximum value does not best describe the overall similarity of two contact types. The "best" distance-intercept might also vary across different ranges of cutoffs. The "diagonal" in the similarity matrix that maximises the mean similarity might be biased by high similarities at high cutoffs or by outliers at low cutoffs.

We identify all local maxima in each similarity matrix S. A cell value $S(i, j)$ is a local maximum if it is the maximum value over all values in row i and all values in column j. In practise, local maxima are all "undirectional" similarities for specific pairs of cutoffs. Driven by the fact that the mean similarity does not increases monotonically as the cutoff increases (see Section 3.3.1), the "best" similarity is defined as the first local maximum over the set of local maxima ordered by increasing cutoff. In case the local maxima are monotonically increasing, the first local maximum value that corresponds to lowest cutoffs is selected. The "best" distance intercept is defined based on the pair of cutoffs for which the "best" similarity occurs. Although this approach is heuristic to some extent, the "best" similarity and distance-intercept selected are manually verified to be correct and correspond to reasonable cutoffs.

3.2.4 Network properties

The impact of RIG definition and in particular of contact range on connectivity, although fundamental, has not been assessed so far. For each RIG, we calculate the mean degree, the percentage of oprhan nodes, the number of connected components and the size of the giant component. Orphan (isolated) nodes are nodes without any edges. Connected components are induced subgraphs in which each node is reachable by every other node. The largest connected component is commonly referred to as the giant component. Both the number of connected components and the size of the giant component are important network properties. We also calculate the characteristic path length, the clustering coefficient and the assortativity coefficient for both the whole RIG and the giant component.

3.2.5 Implementation

All RIGs analysed in this work are constructed using OWL [225]. OWL is a Java library and a set of command line tools for the analysis of biological macromolecules. It provides functionality for analysing protein sequences and structures using built-in algorithms and interfaces to external tools and particular emphasis is given on analysis of RIGs. This author is among the main developers of this Java library. OWL is licensed under the GNU Lesser General Public License. The similarity calculation is integrated in OWL. The calculation of all network properties is implemented separately in a C++ standalone program based on the Boost Graph Library [265].

3.2.6 Remark

It is not feasible with respect to the scope of this thesis to present all results derived from this exhaustive analysis. Key novel aspects are discussed and illustrated in the results section. All network properties and all pair-wise RIG similarities calculated for the complete data set of RIGs as well as additional figures are publicly available as part of OWL [225].

3.3 Results and Discussion

3.3.1 Similarity

In the direction of rationalising the selection of a RIG definition, similarity matrices address two fundamental questions. Given two contact types and a specific contact range, it is feasible to quantify how similar the RIG definitions are for a certain pair of distance cutoffs, one for each contact type. Moreover, in a similarity matrix one may determine the distance cutoff that maximises the similarity of a certain contact type and range with a RIG definition of the same contact range, different contact type and fixed distance cutoff.

The mean Tanimoto similarity matrices over all proteins are illustrated for all pairwise comparisons between the four most frequent contact types (C_α, C_β, SC, ALL), for all distance cutoffs and with respect to all contact ranges. Figure 3.2 demonstrates the similarity between C_α and ALL RIGs, while Appendix Figures C.2 - C.6 cover the rest of the comparisons. Appendix Figure C.7 shows the similarity matrices for other selected pairs of contact types and for all distance cutoffs but with respect to a single contact range (*all*). It should be pointed out that all Tanimoto similarity matrices for the complete data set of RIGs are available in OWL [225]. For brevity, unless otherwise specified, the term similarity refers to the Tanimoto similarity.

Similarity matrices clearly show that different network representations of the same protein are never 100% similar. With respect to the RIG definitions compared and illustrated here and for *all* contact range, the lowest similarities are observed for the pairs BB-SC, C_α-SC and C_β-BB in order of increasing similarity. Their similarity does not exceed 70% and can be as low as 50% for reasonable selected cutoffs. On the contrary, the similarity of C_α and C_β with C_α/C_β and $C_{\alpha\beta}$ can be as high as 90%. The similarity is also quite robust across the different proteins for all network comparisons. The standard deviation of the similarity has a mean value of 2.39 ± 2.16 (%) based on all RIG comparisons.

Outliers in the similarity matrices are often observed in the lower left corner, i.e. for low distance cutoffs. For example in Figure 3.2A, $(C_\alpha)^{\text{all}}_{4.0\text{Å}}$ and $(ALL)^{\text{all}}_{2.5\text{Å}}$ have 98% similarity as at such low cutoffs the interactions simply reproduce the sequence connectivity. In general, similarity values for cutoffs less than 4.0Å for single-/dual-atom residue representations and less than 5.0Å for multi-atom residue representations are not meaningful.

Dependence on RIG definition

As the distance cutoffs increase and the networks become denser, similarity increases. This also means that the maximum similarity of a contact type x across all cutoffs with a certain RIG definition y increases as the distance cutoff for y increases as well (Appendix Figure C.8A). At the same time the distance cutoff for x that maximises its similarity with y also increases (Appendix Figure C.8B). However, it must be noted that the maximum similarity does not increase monotonically. From the similarity

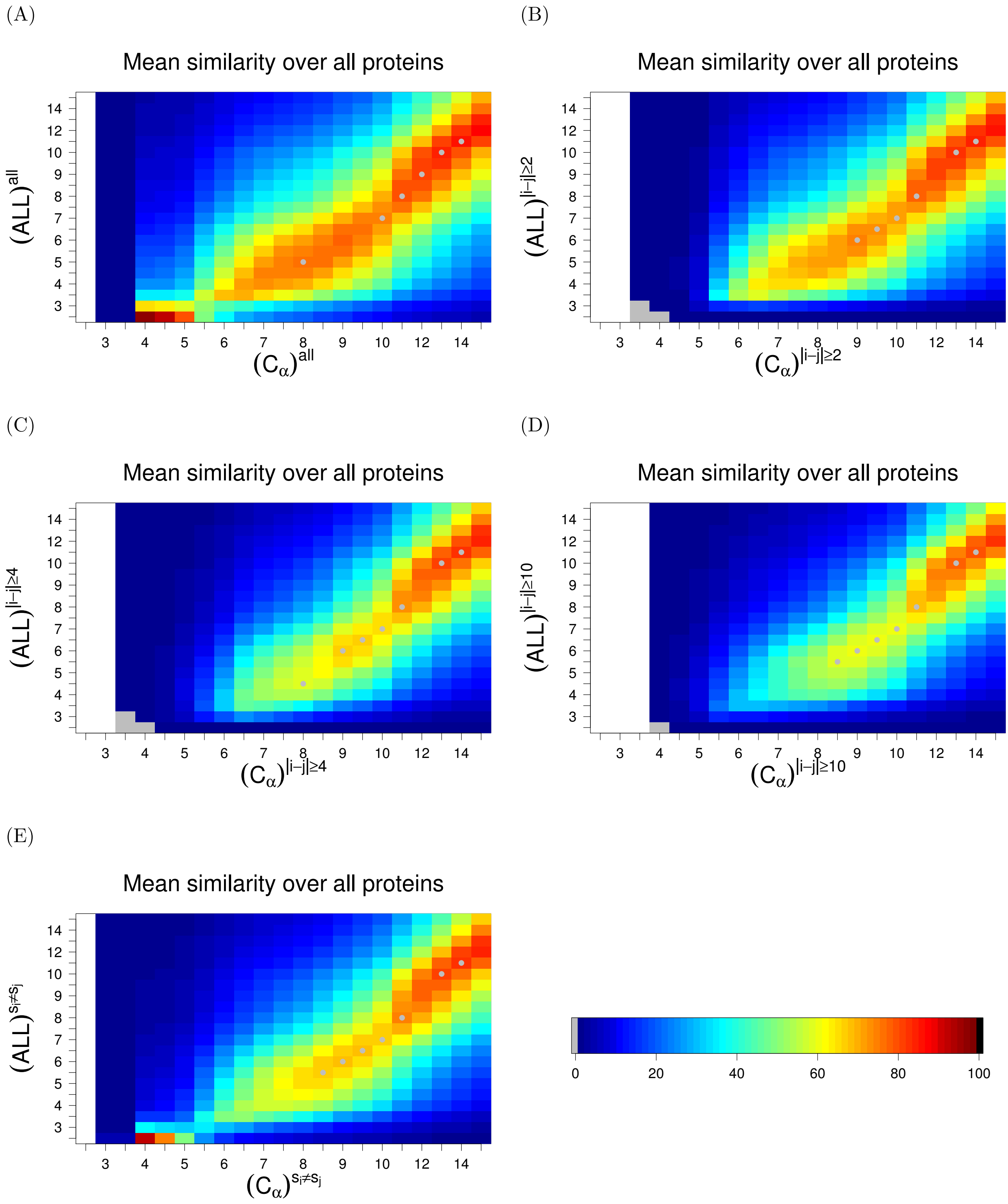


Figure 3.2: The mean similarity matrices between C_α and *ALL* RIGs, over all proteins, for all distance cutoffs and for contact ranges: A. all, B. $|i - j| \geq 2$, C. $|i - j| \geq 4$, D. $|i - j| \geq 10$, E. $s_i \neq s_j$. Grey filled circles correspond to local maxima.

matrix in Appendix Figure C.2A, it is obvious that the maximum mean similarity of $(C_\beta)^{\text{all}}$ with $(C_\alpha)^{\text{all}}$ increases up to 8.0Å for C_α, then decreases up to 9.5Å and increases again. This behaviour is recurrent in almost all similarity matrices and probably results from the first and second coordination shells occurring at different cutoffs for different contact types. Therefore, the network density does not increase at the same rate with distance cutoff for the contact types being compared.

When RIGs contain similar types of residue-residue interactions, then higher similarity is observed as expected. For example, C_β RIGs have higher similarity with SC RIGs compared to C_α ones (Appendix Figures C.3 and C.4) as both C_β and SC contain side-chain mediated interactions. Similarly, C_α RIGs have higher similarity with BB RIGs compared to C_β (Appendix Figure C.7, panels A and B). Moreover, the more fine-grained the residue representation in contact type selection, the lower similarity is observed. For example, C_α and C_β RIGs (Appendix Figure C.2A) have higher similarity with each other compared to the pair of BB and SC RIGs (Appendix Figure C.7C).

Contact range has significant effect in the observed similarity. Figure 3.3 shows the maximum mean similarities of C_α RIGs for each distance cutoff with respect to ALL RIGs of any cutoff, over all proteins, and for all contact ranges. The more interactions are filtered based on contact range, the lower the maximum similarity. The lowest similarity is almost always observed for sequence separation threshold of 10. Networks of interactions between secondary structure elements exhibit similar level of similarity with sequence separation thresholds 2 and 4. As the distance cutoff increases and especially at extremely high distance cutoffs, the effect of contact range diminishes. The impact of contact range upon similarity is identical for many pairs of RIG definitions.

However, there are cases that do not follow this rule. The impact of contact range is highly interrelated with the contact types being compared. For example, backbone-mediated interactions are primarily short-range ones as opposed to side-chain mediated interactions that are long-range. In Appendix Figure C.9A we plot the mean similarities of BB, SC, and BB/SC RIGs with ALL ones, over all proteins, for 5.0Å distance cutoff and for all contact ranges. The similarities of C_α, C_β, and C_α/C_β with $C_{\alpha\beta}$ for 8.0Å are also plotted (Appendix Figure C.9B). As the sequence separation of interacting residues being filtered increases, the lower the similarity of C_α and BB with $C_{\alpha\beta}$ and ALL respectively. On the contrary, the similarity of C_β and SC with $C_{\alpha\beta}$ and ALL respectively either remains constant or even increases. It is striking also that both ALL and $C_{\alpha\beta}$ RIGs are dominated by the backbone-sidechain mediated interactions and thus their similarity with BB/SC and C_α/C_β respectively is extremely robust with respect to contact range.

Contact range seems to have an effect on the standard deviation of the similarity. The more interactions that are filtered based on contact range, the higher the deviation. Deviation reaches its maximum mean value of 3.11 ± 2.68 for sequence separation threshold of 10.

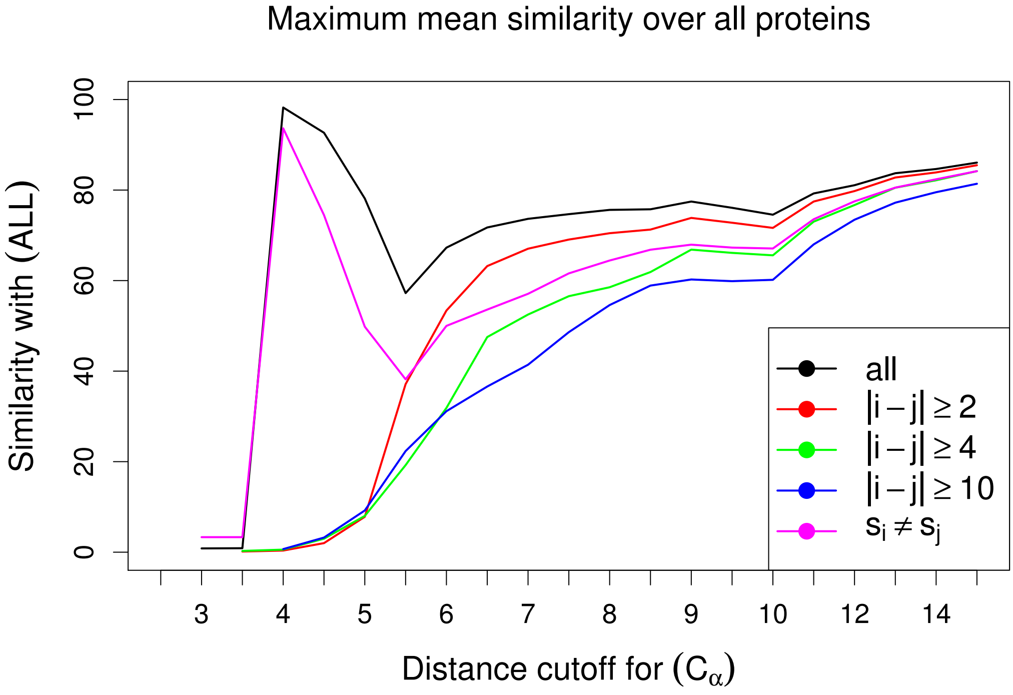


Figure 3.3: The maximum mean similarities of C_α RIGs for each distance cutoff with respect to ALL RIGs of any cutoff, over all proteins, for all contact ranges.

Dependence on structural class

Appendix Figure C.10 demonstrates the effect of structural class on the maximum mean similarity and with respect to contact range. $All-\alpha$ C_α RIGs have significantly higher maximum mean similarity with ALL RIGs compared to $all-\beta$ C_α RIGs for all and $|\mathrm{i}-\mathrm{j}| \geq 2$ contact ranges. However, at higher sequence separation threshold 10 similarity decreases significantly more for $all-\alpha$ RIGs compared to $all-\beta$ RIGs. An unpaired Wilcoxon rank sum test for the maximum mean similarity of $(C_\alpha)_{8.0\mathring{A}}$ with ALL RIGs of any cut-off for the $all-\alpha$ and for the $all-\beta$ proteins yielded p-values lower than 10^{-3} for the fore-mentioned differences. This is expected as in beta-sheet structures interactions are formed between residues that are well separated in sequence compared to alpha-helix structures that are dominated by short-range interactions. Moreover, the effect of contact range is independent of structural class. The effects of both contact range and structural class are summarised in Figure 3.4. The means of the maximum mean similarities for all combinations of RIG definitions over all proteins are plotted separately for each contact range and structural class.

Structural class also has an effect on the standard deviation itself. The standard deviation has increasing mean values 1.39±1.25, 1.63±1.28, 1.96±1.61 and 2.56±2.06 in the order of α/β, $all-\beta$, $\alpha+\beta$, $all-\alpha$ structural classes. This order remains consistent independent of contact range.

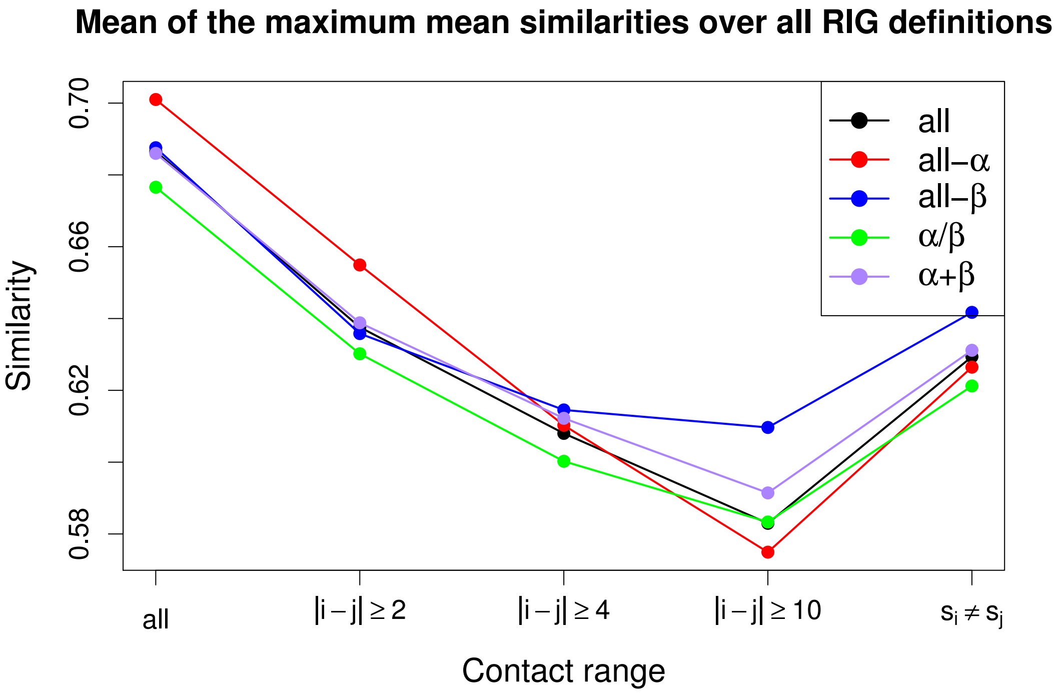


Figure 3.4: The means of the maximum mean similarities over all RIGs, for all proteins, and with respect to all contact ranges and structural classes.

Dependence on network density

High Tanimoto similarity is a result of high Meet/Min similarity and high agreement in network density. Appendix Figure C.11 shows matrices with the mean Tanimoto similarity, mean Meet/Min similarity and mean percentage difference in density between $(C_\beta)^{\mathrm{all}}$ and $(C_\alpha)^{\mathrm{all}}$ RIGs, over all proteins, and for all distance cutoffs. Distance cutoff value 7.5Å maximises the mean Tanimoto similarity of $(C_\alpha)^{\mathrm{all}}$ with $(C_\beta)^{\mathrm{all}}_{7.0\mathring{A}}$ to 67% having 91% mean Meet/Min similarity and 21% mean percentage difference in density. It is obvious that $(C_\alpha)^{\mathrm{all}}_{6.5\mathring{A}}$ has the highest agreement in density (96%). However, only 81% of the $(C_\alpha)^{\mathrm{all}}_{6.5\mathring{A}}$ RIGs overlaps on average with the $(C_\beta)^{\mathrm{all}}_{7.0\mathring{A}}$ RIGs leading to lower Tanimoto similarity of 65%. This is a clear example that the agreement in network density itself is necessary but not adequate condition for high Tanimoto similarity.

Similarity for commonly used RIG definitions

Due to the avalanche of information provide here and in order to facilitate the selection of RIG definition, we provide look up tables for the ten most frequent basic RIG definitions and the four most frequent contact types as determined in Sections 2.3.2 and 2.3.4. Table 3.2 provides the mean Tanimoto similarities for all pairs of the most frequent basic RIG definitions separately for each contact range. Their similarity lies in the range [18%, 80%] demonstrating the importance of the work presented here. In Table 3.3 we present for each one of four most frequent contact types (C_α, C_β, SC, and

Table 3.2: Mean Tanimoto similarity for all pairs of the ten most frequent basic RIG definitions for all contact ranges (all, $|\mathrm{i}-\mathrm{j}| \geq 2$, $|\mathrm{i}-\mathrm{j}| \geq 4$, $|\mathrm{i}-\mathrm{j}| \geq 10$, $\mathrm{s_i} \neq \mathrm{s_j}$).

RIG definition A	RIG definition B	all	$\lvert\mathrm{i}-\mathrm{j}\rvert \geq 2$	$\lvert\mathrm{i}-\mathrm{j}\rvert \geq 4$	$\lvert\mathrm{i}-\mathrm{j}\rvert \geq 10$	$\mathrm{s_i} \neq \mathrm{s_j}$
$(C_\alpha)_{6.0\mathring{A}}$	$(ALL)_{4.0\mathring{A}}$	62	50	32	30	44
$(C_\alpha)_{6.0\mathring{A}}$	$(ALL)_{4.5\mathring{A}}$	57	45	29	28	40
$(C_\alpha)_{6.0\mathring{A}}$	$(ALL)_{5.0\mathring{A}}$	52	40	26	26	36
$(C_\alpha)_{6.0\mathring{A}}$	$(ALL)_{6.0\mathring{A}}$	43	32	20	20	28
$(C_\alpha)_{7.0\mathring{A}}$	$(ALL)_{4.0\mathring{A}}$	73	65	52	41	57
$(C_\alpha)_{7.0\mathring{A}}$	$(ALL)_{4.5\mathring{A}}$	74	67	52	41	57
$(C_\alpha)_{7.0\mathring{A}}$	$(ALL)_{5.0\mathring{A}}$	71	64	49	41	56
$(C_\alpha)_{7.0\mathring{A}}$	$(ALL)_{6.0\mathring{A}}$	62	55	41	36	48
$(C_\alpha)_{8.0\mathring{A}}$	$(ALL)_{4.0\mathring{A}}$	70	62	54	46	56
$(C_\alpha)_{8.0\mathring{A}}$	$(ALL)_{4.5\mathring{A}}$	75	69	59	51	62
$(C_\alpha)_{8.0\mathring{A}}$	$(ALL)_{5.0\mathring{A}}$	76	70	58	53	64
$(C_\alpha)_{8.0\mathring{A}}$	$(ALL)_{6.0\mathring{A}}$	71	66	54	51	61
$(C_\alpha)_{8.5\mathring{A}}$	$(ALL)_{4.0\mathring{A}}$	65	58	49	44	52
$(C_\alpha)_{8.5\mathring{A}}$	$(ALL)_{4.5\mathring{A}}$	72	66	57	52	60
$(C_\alpha)_{8.5\mathring{A}}$	$(ALL)_{5.0\mathring{A}}$	75	71	61	56	64
$(C_\alpha)_{8.5\mathring{A}}$	$(ALL)_{6.0\mathring{A}}$	75	70	61	58	66
$(C_\beta)_{8.0\mathring{A}}$	$(ALL)_{4.0\mathring{A}}$	73	66	58	51	60
$(C_\beta)_{8.0\mathring{A}}$	$(ALL)_{4.5\mathring{A}}$	79	74	67	62	69
$(C_\beta)_{8.0\mathring{A}}$	$(ALL)_{5.0\mathring{A}}$	78	74	67	66	70
$(C_\beta)_{8.0\mathring{A}}$	$(ALL)_{6.0\mathring{A}}$	71	66	59	60	63
$(C_\beta)_{8.0\mathring{A}}$	$(C_\alpha)_{6.0\mathring{A}}$	56	44	29	28	39
$(C_\beta)_{8.0\mathring{A}}$	$(C_\alpha)_{7.0\mathring{A}}$	75	69	57	47	60
$(C_\beta)_{8.0\mathring{A}}$	$(C_\alpha)_{8.0\mathring{A}}$	80	75	69	63	70
$(C_\beta)_{8.0\mathring{A}}$	$(C_\alpha)_{8.5\mathring{A}}$	77	73	67	65	69
$(SC)_{4.5\mathring{A}}$	$(ALL)_{4.0\mathring{A}}$	36	41	52	57	48
$(SC)_{4.5\mathring{A}}$	$(ALL)_{4.5\mathring{A}}$	40	45	60	68	55
$(SC)_{4.5\mathring{A}}$	$(ALL)_{5.0\mathring{A}}$	35	39	50	58	47
$(SC)_{4.5\mathring{A}}$	$(ALL)_{6.0\mathring{A}}$	29	31	36	42	36
$(SC)_{4.5\mathring{A}}$	$(C_\alpha)_{6.0\mathring{A}}$	19	19	20	18	21
$(SC)_{4.5\mathring{A}}$	$(C_\alpha)_{7.0\mathring{A}}$	24	25	29	27	27
$(SC)_{4.5\mathring{A}}$	$(C_\alpha)_{8.0\mathring{A}}$	26	28	34	35	32
$(SC)_{4.5\mathring{A}}$	$(C_\alpha)_{8.5\mathring{A}}$	26	27	32	34	31
$(SC)_{4.5\mathring{A}}$	$(C_\beta)_{8.0\mathring{A}}$	33	36	44	47	42

Table 3.3: Distance cutoffs that maximise mean Tanimoto similarity for the four most frequent contact types (C_α, C_β, SC, and ALL) and with respect to the ten most frequent basic RIG definitions. Basic RIG definitions are shown in column 1 while distance cutoffs are provided with respect to all 5 contact ranges (all, $|\mathrm{i}-\mathrm{j}| \geq 2$, $|\mathrm{i}-\mathrm{j}| \geq 4$, $|\mathrm{i}-\mathrm{j}| \geq 10$, $\mathrm{s_i} \neq \mathrm{s_j}$). Cutoffs for identical contact types compared are not shown as trivial.

RIG definition	all				$\lvert\mathrm{i}-\mathrm{j}\rvert \geq 2$				$\lvert\mathrm{i}-\mathrm{j}\rvert \geq 4$				$\lvert\mathrm{i}-\mathrm{j}\rvert \geq 10$				$\mathrm{s_i} \neq \mathrm{s_j}$			
	C_α	C_β	SC	ALL	C_α	C_β	SC	ALL	C_α	C_β	SC	ALL	C_α	C_β	SC	ALL	C_α	C_β	SC	ALL
$(C_\alpha)_{6.0\mathring{A}}$	-	7.5	6.5	3.5	-	7.5	7.5	3.5	-	6.0	6.5	3.5	-	6.0	6.5	3.5	-	6.5	6.5	3.5
$(C_\alpha)_{7.0\mathring{A}}$	-	8.0	7.5	4.5	-	8.0	7.5	4.5	-	7.5	6.5	4.0	-	7.0	6.5	5.0	-	7.5	7.0	4.5
$(C_\alpha)_{8.0\mathring{A}}$	-	8.0	7.5	5.0	-	8.0	7.5	5.0	-	8.0	7.0	4.5	-	8.0	7.0	5.5	-	8.0	7.0	5.5
$(C_\alpha)_{8.5\mathring{A}}$	-	8.5	7.5	5.5	-	8.5	7.5	5.5	-	8.5	7.0	5.5	-	8.0	7.0	5.5	-	8.5	7.5	5.5
$(C_\beta)_{8.0\mathring{A}}$	8.0	-	7.5	4.5	8.0	-	7.5	4.5	8.0	-	7.0	5.0	8.5	-	7.0	5.0	8.0	-	7.0	5.0
$(SC)_{4.5\mathring{A}}$	8.0	7.0	-	4.5	8.0	6.5	-	4.5	8.0	7.0	-	4.5	8.0	7.0	-	4.5	8.0	7.0	-	4.5
$(ALL)_{4.0\mathring{A}}$	7.0	7.5	6.5	-	7.0	7.5	6.5	-	7.5	7.5	6.5	-	8.0	7.5	4.0	-	7.5	7.5	6.5	-
$(ALL)_{4.5\mathring{A}}$	7.5	8.0	6.5	-	8.0	8.0	7.0	-	8.0	7.5	6.5	-	8.5	7.5	6.0	-	8.0	8.0	6.5	-
$(ALL)_{5.0\mathring{A}}$	8.0	8.0	7.5	-	8.5	8.0	7.5	-	8.5	8.0	6.5	-	8.5	8.0	6.0	-	8.5	8.0	6.5	-
$(ALL)_{6.0\mathring{A}}$	9.0	9.0	7.5	-	9.0	9.0	7.5	-	9.0	9.0	7.0	-	9.0	8.5	7.0	-	9.0	9.0	7.5	-

ALL) the distance cutoffs that maximise their similarity with the ten most frequent RIG definitions, separately for each contact range. This table allows researchers to select the optimal cutoff for their favourite contact type and with respect to published work based on a certain RIG definition.

3.3.2 Best similarity

Best similarity matrices summarise in a higher, more abstract level the pair-wise similarities between all contact types for a certain contact range independent of distance cutoffs. Each cell in the upper triangular part of the matrix contains the "best" similarity, the value that describes "best" the overall similarity of two contact types over all pairs of cutoffs. The lower triangular part contains the cutoffs at which the "best" similarities occurs and their difference or else distance-intercept. The values of the matrix are calculated as described in Section 3.2.3. Figure 3.5 shows the best similarity matrix for all pairs of contact types with respect to *all* contact range and all proteins. The matrices for the other contact ranges as well as structural class specific matrices are publicly available in OWL [225].

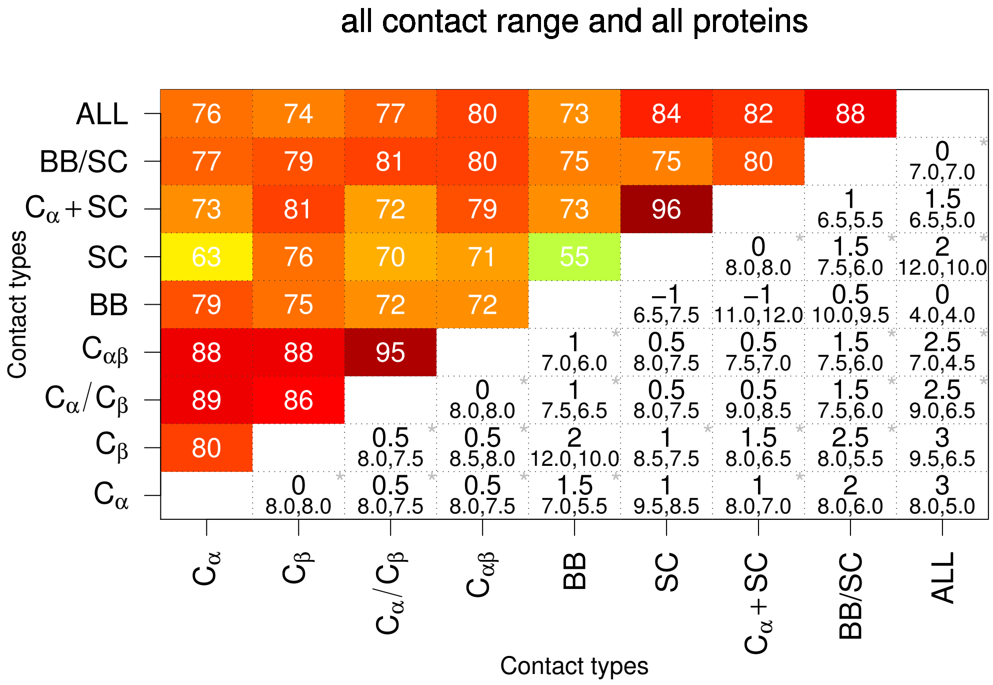


Figure 3.5: "Best" mean similarities and distance-intercepts between all contact types, over all proteins, and for *all* contact range. The upper triangular part of the matrix contains the best similarities. The lower triangular contains the cutoffs at which the best similarity occurs and their difference. For star denoted cells, the cutoffs are selected based on the first local maximum over the local maxima in the similarity matrix. Otherwise, the first local maximum in the similarity matrix is used.

The "best" similarities are in agreement with previous observations. For *all* contact range and according to Figure 3.5, the lowest "best" similarities are observed for the pairs BB-SC and C_α-SC, while C_α-C_α/C_β, C_α-$C_{\alpha\beta}$, C_β-C_α/C_β, C_β-$C_{\alpha\beta}$, C_α/C_β-$C_{\alpha\beta}$, BB/SC-ALL, and SC-Ca+SC pairs have the highest "best" similarities. The highest similarity of 96% occurs between SC and $Ca + SC$ and the lowest similarity of 55% between BB and SC RIGs.

Although the approach for the calculation of the "best" similarity matrices is heuristic to some extent, the "best" similarity and distance-intercept values have been manually verified to be correct and to correspond to reasonable cutoffs. For example, the best distance intercept between $(C_\alpha)^{\mathrm{all}}$ and $(C_\beta)^{\mathrm{all}}$ RIGs is 0, i.e. identical distance cutoffs provide the best similarity. This can be confirmed by visual inspection of the similarity matrix in Appendix Figure C.2A. Other approaches such as selecting the diagonal with the highest average mean similarity lead to non optimal distance-intercepts.

3.3.3 Connectivity

Although the effect of the contact range on network properties like the degree distribution, the characteristic path length and the clustering coefficient has been previously studied, the effect on fundamental properties that capture the overall network connectivity has not been investigated. Here, we calculate four network properties, the mean degree, the mean percentage of orphan residues, the mean number of connected components and the mean giant component size as the percentage of the protein length, over all proteins and for each contact range.

Table 3.4 displays the mean values for the ten most frequent basic RIG definitions and for all contact ranges. For the highest sequence separation threshold, the mean degree is decreased significantly, the percentage of orphan residues lies in the range [12%, 55%], there are up to eight connected components, and the giant component contains from 33% up to 88% of all residues. These observations are of dual importance. First, any network analysis of long-range RIGs that aims to unravel important residues for protein function, stability and folding must take into consideration the effect of orphan residues in the sensitivity of the approach. Second, residues may lie in separate components in the network and thus rigorous calculation of network properties must treat with caution non reachable nodes. Network analysis on the whole network might yield different results compared to analysis on individual connected components.

We further examine the effect of distance cutoff, contact range and structural class on the fore-mentioned network properties. Appendix Figures C.12 - C.15 show the mean values over all proteins for C_α RIGs and all distance cutoffs and with respect to each contact range and each structural class separately. Distance cutoff and contact range have opposite effects. As the distance cutoff increases, the mean degree increases, the number of orphan residues decreases, the number of connected components decreases and the size of the giant component increases. However, as the sequence separation of interacting residues being filtered increases, the exactly opposite effect occurs. Most important and as similarly observed for similarity, the effect of the contact range is more severe for $all-\alpha$ proteins than for $all-\beta$ proteins. $All-\alpha$ proteins have significantly higher percentage of orphan residues and lower giant component size compared

Table 3.4: Effect of contact range on connectivity for the ten most frequent basic RIG definitions for all contact ranges: all (all), $|\mathrm{i}-\mathrm{j}| \geq 2$ (2), $|\mathrm{i}-\mathrm{j}| \geq 4$ (4), $|\mathrm{i}-\mathrm{j}| \geq 10$ (10), $\mathrm{s_i} \neq \mathrm{s_j}$ (s), and with respect to four network properties: mean degree (Degree), mean percentage of orphan residues (PoOR), mean number of connected components (NoCC) and mean giant component size (GCS).

RIG definition	Degree					PoOR					NoCC					GCS				
	all	2	4	10	s	all	2	4	10	s	all	2	4	10	s	all	2	4	10	s
$(C_\alpha)_{6.0\mathring{A}}$	5	4	2	2	3	0	4	36	55	23	1	2	13	8	6	100	90	44	33	59
$(C_\alpha)_{7.0\mathring{A}}$	8	6	3	3	4	0	0	9	40	12	1	1	4	6	2	100	100	84	50	82
$(C_\alpha)_{8.0\mathring{A}}$	10	8	5	4	6	0	0	4	27	7	1	1	2	2	1	100	100	93	70	92
$(C_\alpha)_{8.5\mathring{A}}$	11	9	6	5	7	0	0	3	21	5	1	1	1	1	1	100	100	97	78	95
$(C_\beta)_{8.0\mathring{A}}$	10	8	5	4	6	0	0	4	20	6	1	1	2	1	1	100	100	94	80	94
$(SC)_{4.5\mathring{A}}$	4	4	3	3	3	5	10	18	28	16	3	4	5	4	3	91	85	76	67	80
$(ALL)_{4.0\mathring{A}}$	8	6	4	3	5	0	0	6	26	8	1	1	3	3	1	100	99	90	72	92
$(ALL)_{4.5\mathring{A}}$	9	7	5	4	6	0	0	5	21	6	1	1	2	1	1	100	100	93	78	94
$(ALL)_{5.0\mathring{A}}$	10	8	6	5	6	0	0	3	19	5	1	1	1	1	1	100	100	96	81	95
$(ALL)_{6.0\mathring{A}}$	13	11	8	6	8	0	0	1	12	2	1	1	1	1	1	100	100	98	88	97

to $all-\beta$ for $(C_\alpha)^{|i-j|\geq 10}_{8.0\mathring{A}}$ and $(C_\alpha)^{s_i \neq s_j}_{8.0\mathring{A}}$ RIGs and significantly higher number of connected components for $(C_\alpha)^{|i-j|\geq 10}_{8.0\mathring{A}}$ RIGS. Moreover and independent of contact range, the mean degree in $all-\alpha$ $(C_\alpha)_{8.0\mathring{A}}$ RIGs is significantly lower compared to $all-\beta$ $(C_\alpha)_{8.0\mathring{A}}$ ones. An unpaired Wilcoxon rank sum test for all network properties of $(C_\alpha)_{8.0\mathring{A}}$ RIGs for the $all-\alpha$ and for the $all-\beta$ proteins yielded p-values lower than 10^{-3} in the fore-mentioned cases. Figure 3.6 clearly illustrates the difference in the number of non-orphan residues and in the giant component size between $all-\alpha$ and $all-\beta$ proteins for $(C_\alpha)^{|i-j|\geq 10}$ RIGs.

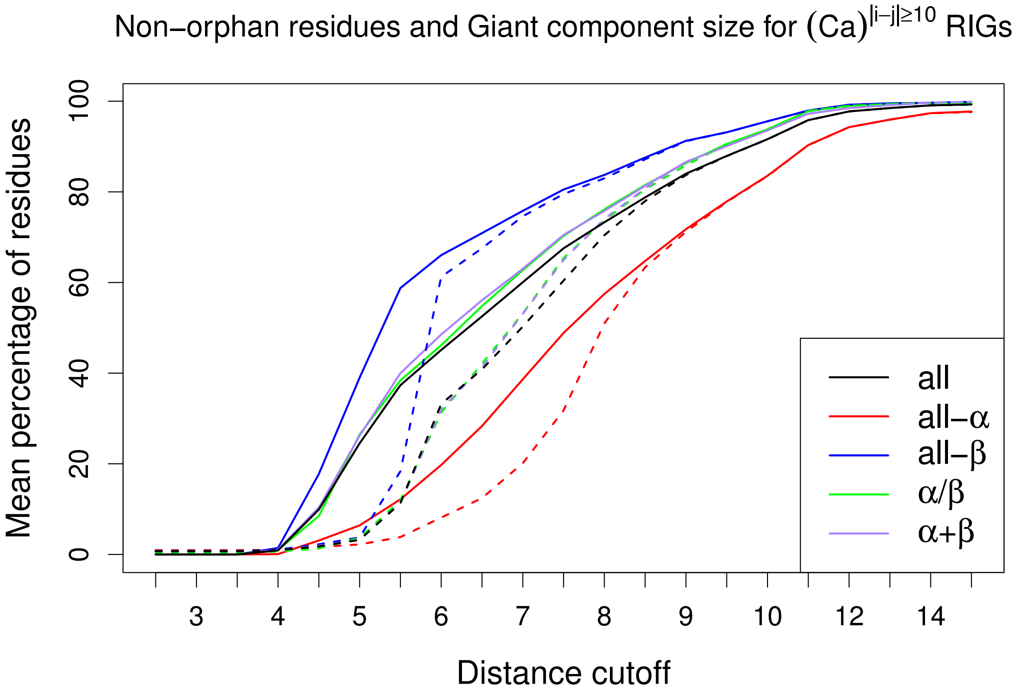


Figure 3.6: Mean number of non-orphan residues (solid lines) and giant component size (dashed lines) for $(C_\alpha)^{|i-j|\geq 10}$ RIGs, over all proteins, with respect to all structural classes.

All network properties mentioned in Section 3.2.4 have been calculated for all proteins and are publicly available in OWL [225].

3.3.4 Small world character

Greene and Higman [106] have shown that long-range RIGs exhibit no small world character. Here, we examine whether this conclusion is valid also with respect to the giant component of long-range RIGs. As in [106], we calculate the characteristic path length L and clustering coefficient C for $(ALL)^{|i-j|\geq 10}_{5.0\mathring{A}}$ RIGs as well as for random and regular graphs with the same number of nodes N and the same mean degree k as the RIGs. The C and L values for random and regular graphs are calculated based on the following formulas [303]:

$$
\begin{aligned}
C_{random} &\sim \frac{k}{N} \\
L_{random} &\sim \frac{ln(N)}{ln(k)} \\
C_{regular} &\sim \frac{3(k-2)}{4(k-1)} \\
L_{regular} &\sim \frac{N(N+k-2)}{2k(N-1)}
\end{aligned}
$$

The results are shown in Figure 3.7, confirming that long-range interaction networks and their giant components essentially do not differ from random networks. The choice of RIG definition has no effect on the non small-world character of the giant components. In Figure 3.7, giant components in both $(ALL)^{|i-j|\geq 10}_{5.0\mathring{A}}$ and $(C_\alpha)^{|i-j|\geq 10}$ RIGs have small clustering coefficient.

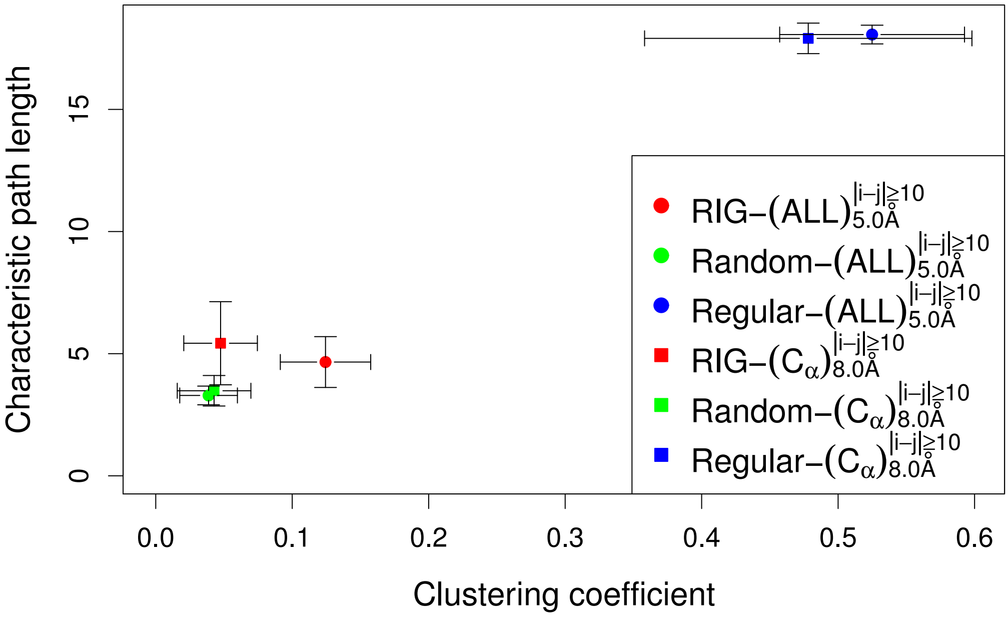


Figure 3.7: Mean characteristic path length and clustering coefficient for giant components in $(ALL)^{|i-j|\geq 10}_{5.0\mathring{A}}$ RIGs (red circles) and $(C_\alpha)^{|i-j|\geq 10}_{8.0\mathring{A}}$ RIGs (red squares), and for random (green circles/squares) and regular (blue circles/squares) networks with equal number of nodes and edges as the RIGs, over all proteins. Values for random and regular networks are obtained based on a theoretical model [318]. Error bars represent the standard deviations over all 60 proteins.

3.3.5 Comparison with related work

Bartoli et al. [21] have shown that $(C_\alpha)^{\mathrm{all}}_{7.0\mathring{A}}$ RIGs are most similar to $(ALL)^{\mathrm{all}}_{5.0\mathring{A}}$ ones with respect to the mean Hamming similarity over a large set of 1,753 non-redundant protein structures. The distance cutoffs for the C_α contact type analysed ranged from 6.0Å to 9.0Å in increments of 0.5Å. Moreover, they demonstrated that the proposed cutoff maximises their similarity with respect to clustering coefficient C and characteristic path length L and that actually $(C_\alpha)^{\mathrm{all}}_{7.0\mathring{A}}$ and $(ALL)^{\mathrm{all}}_{5.0\mathring{A}}$ RIGs have exactly the same values: $C_{C_\alpha} = 0.56 \pm 0.03$, $L_{C_\alpha} = 5.3 \pm 1.4$, $C_{ALL} = 0.56 \pm 0.04$ and $L_{ALL} = 5.3 \pm 1.4$.

However, our calculations demonstrate that $(C_\alpha)^{\mathrm{all}}_{8.0\mathring{A}}$ RIGs have the maximum mean Tanimoto and Hamming similarity with $(ALL)^{\mathrm{all}}_{5.0\mathring{A}}$ RIGs (Appendix Figure C.16A). The 8.0Å distance cutoff also maximises their similarity with respect to their small-world character (Appendix Figure C.16B). However, the C and L values are not identical: $C_{C_\alpha} = 0.56 \pm 0.03$, $L_{C_\alpha} = 4.15 \pm 0.93$, $C_{ALL} = 0.53 \pm 0.03$ and $L_{ALL} = 3.71 \pm 0.79$. We hypothesise that the difference in the results is due to the difference in the data sets of protein structures and especially with respect to their structural class composition. The lack of missing or ambiguous coordinates in our high quality data set ensures the most accurate assessment of the similarity between RIGs of different contact types.

Heringa and Argos [121] as well as Brinda and Vishveshwara [41] have also studied connected components that emerge after applying a filtering threshold on weighted edges of RIGs that represent the strength of a residue-residue interaction. However, this analysis differs from connected components that arise from applying sequence separation thresholds on commonly used network representations of protein structures and thus, comparison was not feasible.

3.4 Conclusion

The choice of the network representation of a protein structure is not trivial. Here, we rationalise this choice by comparing and analysing 56,700 RIGs resulting from 945 RIG definitions for a representative data set of 60 protein structures.

We assess quantitatively the similarity of all RIG definitions and we demonstrate that the similarity between commonly used network representations can be in certain cases quite low. The observed similarity levels re-enforce the importance of the choice of RIG definition. An arbitrarily chosen definition will not necessarily reproduce published results based on certain other definitions. Furthermore, the quest for an optimal definition is of great importance and can lead to results of higher sensitivity and specificity with respect to a certain application.

We also analyse the effect of contact type, distance cutoff, contact range, and structural class on similarity between RIGs. Decrease in the granularity of residue representations, increase of the distance cutoffs and less filtering of interactions with respect to contact range lead to higher similarity. Increased helical content also bias the level of similarity, being high when short-range interactions are included and low when excluded. Short-range interactions are dominant in helical structures compared to beta-sheet structures where long-range interactions are formed.

Contact range and structural class affect the connectivity of RIGs. Network properties like the number of orphan residues, the number of connected components, and the giant component size, have never been systematically studied before. As the sequence separation threshold increases, more residues are excluded from the RIG and residues may lie in "separate" components. The impact is more severe for $all-\alpha$ proteins than for $all-\beta$ proteins. Orphan residues might significantly affect the sensitivity of network-based methods, while pairs of residues that cannot be reached from each other have to be treated with caution for certain network analyses.

The data set analysed contains highly populated folds; folds that are well studied and that provide high coverage of the SCOP database. The high coverage as well as the low standard deviations of the observed similarities across all proteins ensure the generality of this work. A preliminary investigation of similarity and connectivity based on a data set of nine different protein folds (see Data Set 1 in Section 4.2.3) leads to qualitatively identical conclusions. However, it must be pointed out that the structural class composition of a data set as well as the selection of proteins that are stabilised by interactions not commonly modelled in RIGs may bias any network analysis.

Overall, we establish a unifying view for the network representations of protein structures. We provide researchers with open source software tools for converting protein structures to networks and for subsequent network analysis. The proposed data set can be utilised in studies that aim to optimise the RIG definition for a certain application [78]. Our large-scale analysis allows the rational selection of RIG definition as well as the rational comparison of results produced using different definitions. In future we will extend our work to multi-domain proteins as well as to protein complexes.

Acknowledgments

OWL [225] was created by researchers of the Bioinformatics / Structural Proteomics Group at the Max Planck Institute for Molecular Genetics. This author along with Henning Stehr and Jose Duarte have been the main developers. The development was initiated by Michael Lappe who also provided strategic guidance.

Chapter 4

Optimized null model and motif detection for protein structure networks

Summary

Finding a well-fitting null model is crucial for assessing the statistical significance of topological features observed in networks. Thus far, the challenge of finding an optimised null model for Residue Interaction Graphs has not been addressed. Degree-preserving randomised models have been widely used for analysis of other biological networks. However, such a single summary network property may not be detailed enough to capture the complex topological characteristics of protein structures and their network counterparts. Here, we investigate a variety of topological properties to find a well fitting null model for RIGs. Two graphlet-based properties are highly constraining measures of the topological similarity between two networks. The RIGs are derived from a structurally diverse protein data set at various distance cutoffs and for different contact types. We compare the network structure of RIGs to several random graph models. We show that 3-dimensional geometric random graphs, that model spatial relationships between objects, provide the best fit to RIGs. We investigate the relationship between the strength of the fit and various structural features. Geometric random graphs capture the network organisation of RIGs better for larger proteins and for proteins of the same size, the fit is better when helical content is lower. The tighter packing of the solvent accessible surface in thermostable proteins leads to a worse fit, while the quaternary association has no significant impact. A null model has important implications for finding statistically significant subgraphs (motifs) that play an important role in protein folding, stability and function. We demonstrate that choosing geometric graphs as a null model results in the most specific identification of motifs. To our knowledge, this is the first study that addresses the challenge of finding an optimised null model for RIGs.

4.1 Introduction

Much attention has recently been given to the statistical significance of topological features observed in biological networks. In order to assess the statistical significance, modelling biological networks and finding a well-fitting null model is of crucial importance. A good model should generate graphs that resemble real data as closely as possible across a wide range of network properties. Only a well-fitting network model that precisely reproduces the network structure and laws through which the network has emerged can enable us to understand and replicate the underlying biological processes. A good null model can be used to guide biological experiments in a time- and cost-optimal way and to predict the structure and behaviour of a system. Since incorrect models lead to incorrect predictions, it is vital to have as accurate a model as possible.

Thus far, graph null models that take into account the network size and the overall degree distribution have been formulated in the field of protein-protein interaction networks [186, 200]. These random models were utilised as the reference state to identify interaction patterns that are over-represented in the experimentally observed networks [200] and to compare the behaviour of certain topological properties [186]. It has been argued that such a realistic but simple approach for defining a null model might wrongly identify as significant the motifs that result from other topological features not taken into account by the null model [10].

As already discussed, network analyses of protein structures have been mainly focused on the degree distribution. It has been shown that the Poisson probability model best describes the degree distribution of RIGs [12, 70, 106]. However, when only long-range interactions are considered, an exponential distribution with a single-scale, fast decaying tail is observed. This distribution exhibits, to some extent, scale-free properties [106]. Moreover, a random rewiring of RIGs, that keeps the number of contacts of each residue fixed, affects the characteristic path length and clustering coefficient and thus such random networks loose the observed small-world character of RIGs [12].

Despite the fact that previous network analyses of RIGs have provided valuable insight, a null model that captures the network organisation of protein structures has not been established. Here, we address this important challenge of finding an appropriate null model for protein structure networks. Degree-preserving null models may not be detailed enough to capture the complex topological characteristics of protein structures. In this direction, we utilise two sensitive graph theoretic measures based on graphlets for assessing the topological similarity between two networks. These measures, the relative graphlet frequency distance and the graphlet degree distribution agreement, have been successfully applied to modelling protein-protein interaction networks [240, 241]. Moreover, we illustrate the importance of choosing an appropriate null model in motif detection. The only related work suggested that a coarser representation of protein structures, in which nodes correspond to secondary-structure elements, has the same network motifs as does a variant of geometric graphs [199].

4.2 Methodology

Here, we present how we determine which random graph keeps the observed topological characteristics of RIGs. We compare each RIG with five different random graph models. To overcome the limitations introduced by using a single network property (such as the degree distribution), we perform a fine-grained analysis of RIGs that is based on a multitude of network properties. We perform systematic analysis on various RIG definitions and on various fold types to ensure the generality of our analysis. In total, we compare 1,973 RIGs to 295,950 network models. Also, we examine how protein size, structural class, protein thermostability, and quaternary structure affect the strength of the fit for the best null model. Finally, we perform network motif search in RIGs with different random graph models to demonstrate the importance of choosing our proposed null model. In the following, the graph models and properties and all individual steps are explained in detail.

4.2.1 Network models

For each RIG, we evaluated the fit of five different random graph models. In Erdös-Rényi random graphs ("ER"), edges between pairs of nodes are added uniformly at random with the same probability p. "ER" graphs are generated by using the LEDA [188] implementation of $G_{n,m}$, a random graph G with n nodes and m edges. Random graphs with the same degree distribution as the data ("ER-DD") are generated by using the "stubs" method. "Stubs" are attributed to the nodes based on the degree distribution of the real network and each edge is created at random between nodes with stubs [214]. Scale-free networks ("SF-BA") are generated by using the Barabási-Albert preferential attachment model [18]. Geometric random graphs are defined as follows: nodes correspond to uniformly randomly distributed points in a metric space and edges are created between pairs of nodes if the corresponding points are close enough in the metric space according to some distance norm. A variant of geometric random graphs in this study uses 3-dimensional Euclidean boxes and the Euclidean distance norm ("GEO-3D"). Finally, "stickiness network model" ("STICKY") is based on stickiness indices, numbers that summarise node connectivities [242].

Model networks were generated and compared to RIGs using GraphCrunch [192]. For all random graph models, parameters are chosen in such way that each of the generated model networks that corresponds to a RIG has the same number of nodes and the number of edges within 1% of those in the RIG. We generated 30 networks per random graph model for each of the 1,973 RIGs. Thus, in addition to analysing 1,973 RIGs, we also analysed $5 \times 30 \times 1,973 = 295,950$ model networks corresponding to the RIGs and compared them to the RIGs.

4.2.2 Network properties

Exact comparisons of large networks are computationally infeasible due to NP completeness of the underlying subgraph isomorphism problem [58]. Thus, to evaluate the fit of the data to the model networks, we compare the RIGs to the model networks with respect to easily computable *network properties*. We use GraphCrunch [192] to evaluate the fit of different models to the data. RIGs are compared to the corresponding model networks with respect to two graphlet-based *local* and five standard *global* network properties.

Local network properties

We used the following two measures of local structural similarities between two networks: relative graphlet frequency distance (*RGF-distance*) [241] and graphlet degree distribution agreement (*GDD-agreement*) [240]. They have been introduced by Pržulj and are based on graphlets. Since the number of graphlets on n nodes increases exponentially with n, the RGF-distance and GDD-agreement computations are based on 2- to 5-node graphlets (Figure 1.3).

RGF-distance compares the frequencies of the appearance of all 30 2- to 5-node graphlets in two networks. The *RGF-distance* between two networks G and H is defined as:

$$RGF(G,H) = \sum_{i=0}^{29} |F_i(G) - F_i(H)|, \tag{4.1}$$

where

$$F_i(G) = -log \frac{N_i(G)}{\sum_{i=0}^{29} N_i(G)} \tag{4.2}$$

and $N_i(G)$ is the number of graphlets of type i in graph G. The distance is based on the differencies between the relative frequencies of the graphlets and logarithmic function ensures that the distance is not dominated by the most frequent graphlets. The smaller the *RGF-distance*, the more similar two networks are.

GDD-agreement generalises the notion of the degree distribution to the spectrum of *graphlet degree distributions (GDDs).* The degree distribution measures the number of nodes of degree k, i.e. the number of nodes "touching" k edges, for each value of k, where an *edge* is the only graphlet with two nodes. GDDs generalise the degree distribution to other graphlets: they measure for each graphlet on 2 to 5 nodes, the number of nodes "touching" k graphlets *at a particular node.* The "symmetries" between nodes of a graphlet need to be taken into account. This is summarised by the 73 *automorphism orbits* for 2- to 5-node graphlets (see Section 1.7). For each of the 73 orbits j and for graph G, we measure the j^{th} GDD or else the sample distribution $d_G^j(k)$, i.e. the distribution of the number of nodes "touching" k times the corresponding graphlet at orbit j. The distribution is normalised and defined as

$$N_G^j(k) = \frac{\frac{d_G^j(k)}{k}}{\sum_{k=1}^{\infty} \frac{d_G^j(k)}{k}}. \tag{4.3}$$

Dividing by k decreases the contribution of more frequent orbits, while the denominator normalises the distribution to the fraction of the total area under the curve for specific k. The normalised j^{th} GDDs of two networks for all j orbits are compared and combined into the GDD-agreement of two networks. Specifically, the agreement between networks G and H for a specific orbit j is defined as:

$$A^j(G,H) = 1 - D^j(G,H) = 1 - \frac{1}{\sqrt{2}} \left(\sum_{k=1}^{\infty} \left[N_G^j(k) - N_H^j(k) \right]^2 \right)^{1/2}. \qquad (4.4)$$

The agreement is in [0, 1], where 1 means that the two networks are identical with respect to j^{th} GDD-agreement. The total *GDD-agreement* is the arithmetic or the geometric average of the j^{th} GDD-agreements for all j. Since *GDD-agreement* encompasses the fit of each of the 73 GDDs of the networks being compared, it is a very strong measure of structural similarity between two networks. The larger the agreement, the more similar the networks.

As our analysis is consistent with respect to both the arithmetic and geometric versions of GDD-agreement, hereafter we present results only for the geometric GDD-agreement.

Global network properties

We used the following global network properties: the degree distribution, the average clustering coefficient, the clustering spectrum, the average network diameter (shortest path length), and the spectrum of shortest path lengths. These properties have been explained in detail in Section 1.4. The distribution of the average clustering coefficients of degree k nodes is the *clustering spectrum C(k).* The *spectrum of shortest path lengths* is the distribution of shortest path lengths between all pairs of nodes in a network.

4.2.3 Data sets

We analyse three data sets of RIGs for identifying the best-fitting null model and for assessing the quality of its fit with respect to various structural features. In total, we analyse 1,973 RIGs corresponding to 1,469 proteins.

Best-fitting null model

First, we analyse single chain RIGs for nine structurally diverse proteins with the following PDB [28] codes, followed by the chain identifier whenever applicable: 1agd:B, 1fap:B, 1ho4:A, 1i1b, 1mjc, 1rbp, 1sha:A, 2acy and 3eca:A. Atomic coordinates were taken from the Macromolecular Structure Data Search Database (MSDSD) [36]. All structures are solved by X-ray crystallography and their resolution lies in the range [1.5, 2.7] Å. These proteins are a subset of the non-redundant data set that Greene and Higman examined [106]. Specifically, they analysed 65 proteins that cover nine different protein folds, all structural classes and the three kingdoms of life. Moreover, these proteins are diverse in terms of protein sequence and function as well. Since our main concern is our results to be applicable for structurally diverse proteins, we selected one structure from each protein fold randomly for further analysis.

We perform systematic analysis on multiple RIG definitions. We consider three multi-atom residue representations, BB, SC and ALL contact types, and we do not apply any filtering with respect to contact range. We set distance cutoffs to range from 4.0Å to 9.0Å in increments of 0.5Å. Most of the studies that involve RIGs with multi-atom residue representation use distance cutoffs that lie in the range [4.0, 5.0] Å [106, 122, 201, 230, 267]. This is also evident from the analysis in Section 2.3.3 (Figure 2.3B). Therefore, in this range, we choose a finer increment of 0.1Å. For low distance cutoffs of 2Å, 2.5Å, 3Å, and 3.5Å, RIGs with BB contact type reproduce the polypeptide chain connectivity while SC RIGs become highly disconnected and sparse. We excluded from our analysis all RIGs defined with distance cutoff less than 4.0Å to ensure that in all networks at least 80% of the residues have non-covalent contacts. In total and with respect to this data set, we analyse $9 \times 19 \times 3 = 513$ RIGs for the nine proteins, for 19 distance cutoffs and the three contact types (BB, SC, and ALL). Henceforth, we refer to this data set as *Data Set 1*.

Next, to ensure that our results are applicable to a wide range of proteins, we analyse an additional data set of 1,272 RIGs corresponding to 1,272 proteins. These RIGs are constructed with the most commonly used multi-atom basic RIG definition, $(ALL)^{\text{all}}_{5.0\mathring{A}}$ (Figure 2.5). This non-redundant, representative set of X-ray structures from the PDB was pre-compiled by the PISCES server [315]. All proteins have resolution better than 1.8Å, reliability factor (R-factor) less than or equal to 0.25, and their pairwise sequence similarity does not exceed 20%. Henceforth, we refer to this data set as *Data Set 2*.

Quality of the fit

We examine whether the strength of the fit of the best-fitting null model to RIGs changes with respect to protein size and structural class. We use *Data Set 2* and we analyse 744 out of the 1,272 proteins that consist of domains with identical structural class and for which SCOP [210] annotation covers more than 90% of the residues. Out of 744 proteins, 141 are $all-\alpha$, 161 are $all-\beta$, 221 are $\alpha\,/\,\beta$, and 221 are $\alpha+\beta$. The distribution of protein size for the analysed proteins with respect to their structural classes is presented in Appendix Figure D.36A. Only 47 out of the 744 proteins are multi-domain ones and thus are unlikely to bias our analysis.

Furthermore, we analyse the relationship between the strength of the fit of GEO-3D to RIGs and the quaternary structure of the corresponding proteins. Out of the 744 proteins, we examine 75 pairs of monodomain monomeric and monodomain multimeric proteins. Proteins within a pair are of equal size and belong to the same structural class, while proteins across pairs may differ in size and class. Proteins in different pairs have from 64 to 390 residues, with average size of 157 ± 77 residues. 13 protein pairs are $all-\alpha$, 12 are $all-\beta$, 21 are $\alpha\,/\,\beta$, and 29 are $\alpha+\beta$.

Finally, we study the relation between structural features of thermostable proteins and the degree of fitting. We analyse a high quality data set of 94 pairs of *T. maritima* proteins, a representative of thermophiles, and their close homologs from mesophilic species [249]. Although these mesophilic homologs are distinguished to 62 orthologs and 32 paralogs, the statistically significant differences for structural features responsible for thermostability are consistent in both cases [249]. Therefore, we analyse all

94 pairs as a single data set. We construct $(ALL)^{\mathrm{all}}_{4.5\mathring{A}}$ RIGs. Although the same basic RIG definition was used by Robinson-Rechavi et al. [249], our RIG definition is different with respect to contact range; we do not filter out interactions between residues that are less than four residues apart in the primary sequence. Henceforth, we refer to this data set as *Data Set 3*.

4.2.4 Motif detection

To illustrate the importance of the choice of the appropriate null model for a network-based analysis of protein structures, we examine the issue of identifying network motifs in RIGs. Since motifs (anti-motifs) are over-represented (under-represented) subgraphs that appear in a real-world network at frequencies that are much higher (lower) than those of their randomised counterparts [200], motif discovery requires comparing real-world networks with randomised ones, i.e. with model networks. Thus, using an inadequate model may identify as over-represented (under-represented) subgraphs that otherwise would not have been identified as motifs (anti-motifs).

We use mfinder [149] to search for all undirected subgraphs on 3, 4, and 5 nodes (Appendix Figure D.39) in nine $(ALL)^{\mathrm{all}}_{5.0\mathring{A}}$ RIGs corresponding to the nine proteins of *Data Set 1*. In addition to our five network models, we use the three standard models supported by mfinder. We denote these three models as follows: “UA-ER-DD” is the random graph model that preserves the degree distribution of a real-world network, while “CLUST” and “MET” network models, in addition to the degree distribution, preserve the clustering coefficient of all nodes and the number of appearances of all 3-node subgraphs of a real network, respectively.

We detect the statistically significant subgraphs according to their P-values, absolute Z-scores, and absolute M-factors, the motif selection criteria proposed by Milo et al. [200] and Kashtan et al. [149]. P-value is defined as:

$$P_i = Prob[(Nreal_i \leq Nrand_i) \cup (Nreal_i > Nrand_i)] < 0.01, \tag{4.5}$$

where $Nrand_i$ is the number of appearances of the pattern i in a randomised network, and $Nreal_i$ is the number of its appearances in the real network. If the subgraph i is over-represented (under-represented) in the real network with respect to randomised networks with probability lower than 0.01, then the subgraph is a motif (anti-motif). To estimate the empirical p-value, we generate 1,000 networks per random graph model.

For plotting purposes as well as for the significance profiles described below, we use Z-scores instead of P-values. Z-score is defined as:

$$|Z_i| = \frac{|Nreal_i - \bar{N}rand_i|}{sd(Nrand_i)}, \tag{4.6}$$

where $\bar{N}rand_i$ is the mean number of appearances of the pattern i in the randomised networks, and $sd(Nrand_i)$ is their standard deviation.

M-factor is defined as:

$$|M_i| = \frac{|Nreal_i - \bar{N}rand_i|}{\bar{N}rand_i} > 0.1. \tag{4.7}$$

The percentage difference between the number of appearances in the real and the randomised networks must be higher than 0.1. M-factor ensures that subgraphs with just small standard deviation will not be considered misleadingly as significant.

We do not consider the third criteria for network motif selection – uniqueness. Uniqueness is the number of times a subgraph appears in the real network with disjoint set of nodes. Even if uniqueness is less than 4, which is the default threshold in mfinder, there is no reason to reject such subgraphs as non-significant. On the contrary, in RIGs we do not expect motifs that are biologically important to occur many times with completely different set of residues. When a subgraph does not appear in randomised networks of a specific network model, we exclude that subgraph from further motif analysis for that network model.

Also, we address the question of whether different random graph models attribute similar significance to the subgraphs, independent of the magnitude of the significance itself. Similar to the significance profile method [199], we construct 29-dimensional vectors of absolute Z-scores corresponding to 29 3- to 5-node subgraphs where each coordinate represents the Z-score for a given subgraph. For each RIG, we define these vectors with respect to each of the eight network models. Thus, we construct eight vectors of Z-scores for each RIG. Then, we compute Pearson correlation coefficients between all pairs of Z-score vectors for a given RIG. Since the network size is constant in each comparison, there is no need to normalise the Z-scores [199]. High Pearson correlation coefficients between Z-score vectors that correspond to two different network models for the same RIG would indicate that both network models assign similar significance, independent of the magnitude of the significance, and thus, by adjusting the Z-score threshold, the same (anti-)motifs would be identified.

4.2.5 Implementation

All RIGs are constructed using OWL [225] (see Section 3.2.5). Model networks are generated and compared to RIGs using GraphCrunch [192]. We use mfinder [149] for motif detection. The accessible surface area and the volume are calculated using the programs calc-surface and calc-volume [312]. Secondary structure is assigned using the program DSSP [145]. Then, the 8-states of DSSP are converted to three secondary structure states according to EVA conversion scheme [251]. Structural class assignment is based on SCOP release 1.73 [210]. Quaternary structure is predicted by PISA server [161], version 1.14.

4.3 Results

4.3.1 Best-fitting null model

RIG-definition wide topological analysis

We find that all network properties offer support to the superiority of the GEO-3D network model to a large number of RIGs constructed using various contact types and a wide range of distance cutoffs. Given that *BB* and *SC* RIGs are quite different from one another with respect to the set of interacting residues, the robustness of our result across various RIG definitions is quite surprising.

For all of the RIGs in Data Set 1, RGF-distances and GDD-agreements between the RIGs and the model networks strongly favour geometric random graphs. Based on RGF-distances (with minimal exceptions described below), the fit of the GEO-3D model is the best for all nine proteins, all three contact types and all of the distance cutoffs; the exceptions are the lowest distance cutoffs ([4.0, 4.2]Å) for *SC* contact type for four out of nine proteins only. GDD-agreement favours GEO-3D model for all proteins, all three contact types, and all of the distance cutoffs between 4.0Å and 9.0Å, except for RIGs of *ALL* contact type and distance cutoffs higher than 6.5Å. Above a certain distance threshold, residues that are not physically interacting are defined to be in contact. Thus, *ALL* RIGs, that contain more interactions than *BB* and *SC* RIGs, become more "random-like" for large distance cutoffs. For this reason, GDD-agreement rarely favours GEO-3D graphs for such high distance cutoffs. On the contrary, in *BB* and *SC* RIGs, that consider fewer atoms in inter-residue interactions, the direct neighbourhood of a residue does not encompass as many not physically interacting residues at higher distance cutoffs compared to *ALL* RIGs. Therefore, even above 6.5Å, GEO-3D provides a good fit for most of the *BB* and *SC* networks. Illustrations showing GDD-agreements and RGF-distances of 1i1b protein with the five network models are presented in Figure 4.1. The fit of the network models to the other eight proteins with respect to these two network properties is presented in Appendix Figures D.1 to D.8.

The magnitude of GDD-agreement between RIGs and GEO-3D graphs seems to be related to protein size. The two smallest proteins, 1mjc and 1fap, have GDD-agreements of up to around 0.7, while the largest protein, 3eca, has much higher GDD-agreements of up to 0.85. Following this observation, in Section 4.3.2, we analyse the effect of protein size on the strength of the fit of GEO-3D to RIGs in more detail. Moreover, the RGF-distances between the RIGs and the geometric random graphs are usually higher (meaning worse fit) for *SC* networks compared to networks of other contact types. Since side-chains are more mobile compared to the rigid backbone [174], we expect that *SC* networks form more complex interaction patterns compared to networks that contain backbone interactions. There is also a general trend that RGF-distance decreases with increasing distance cutoff, independent of the network model. Equivalently, GDD-agreement increases as the distance cutoff increases for most of the models. Since both the smaller RGF-distance and the larger GDD-agreement indicate improved fit of the network model to RIG, these observations might suggest that for

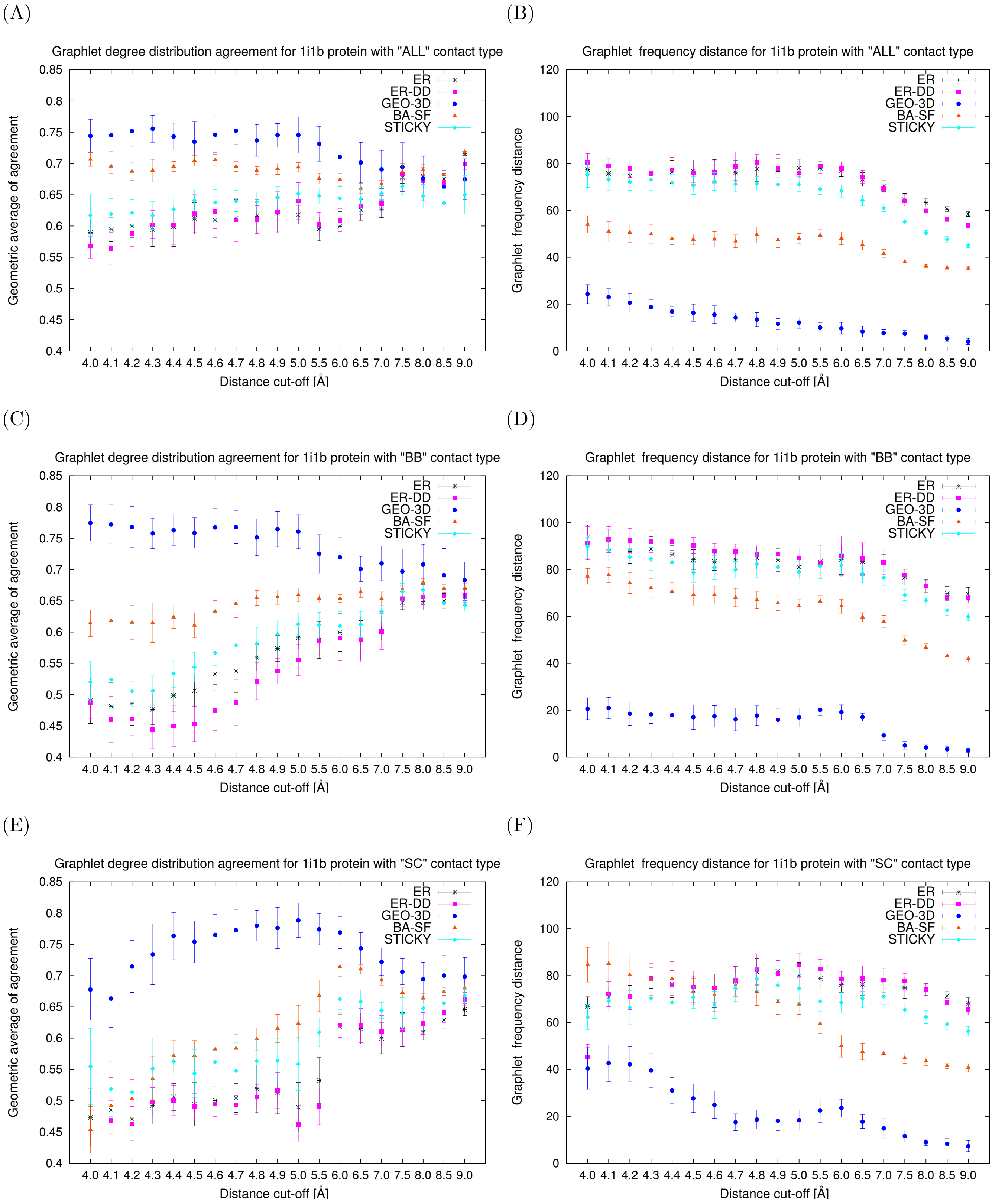


Figure 4.1: GDD-agreements and RGF-distances between model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1i1b protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** GDD-agreement for *ALL* contact type. **B.** RGF-distance for *ALL* contact type. **C.** GDD-agreement for *BB* contact type. **D.** RGF-distance for *BB* contact type. **E.** GDD-agreement for *SC* contact type. **F.** RGF-distance for *SC* contact type. The larger the GDD-agreement in panels A, C, and E the better the fit. The smaller the RGF-distance in panels B, D, and F the better the fit.

higher distance cutoffs, graphlets of higher order are needed to improve the quality of the fit to the data.

We also examine the fit of the network models to the RIGs with respect to five standard network properties. Illustrations showing Pearson's correlation coefficients between the degree distributions of 513 RIGs constructed for the nine proteins and the corresponding model networks are presented in Appendix Figures D.9 to D.17. Note that the perfect fit of the degree distributions of ER-DD model networks to those of RIGs is trivial, since ER-DD networks are constructed to have exactly the same degree distribution as the data. Similarly, STICKY model networks are constrained to have the expected degree distribution of real networks [242]. ER and GEO-3D model networks have Poisson degree distributions, and they both reproduce the degree distributions of all of the 513 RIGs of Data set 1. BA-SF model networks have scale-free degree distributions and they do not reproduce the degree distributions of any of the 513 RIGs that we analysed.

Also, GEO-3D model networks reproduce well the clustering spectra of the RIGs for distance cutoffs smaller than 8Å (Appendix Figures D.9 to D.17). Similarly, the average clustering coefficients of almost all of the 513 RIGs are generally best reproduced by GEO-3D networks (Appendix Figures D.18 to D.26). There exist very few exceptions to this observation. For a very small number of distance cutoffs lower than 5.0Å in the *SC* RIGs of five proteins, the clustering coefficients of BA-SF networks describe the best those of the corresponding RIGs. Interestingly, all small proteins with size less than 105 residues (1agd, 1fap, 1mjc, 1sha, 2acy) are included in the set of these five proteins. Also, we notice the trend that for all proteins and all contact types, the higher the cutoff, the better the fit of clustering coefficient between the GEO-3D model and the data. The average diameters of all RIGs are best reproduced by the GEO-3D networks for all distance cutoffs of *BB* and *ALL* contact types (Appendix Figures D.18 to D.26). The same is true for almost all of the RIGs of *SC* contact type; only for the lowest distance cutoffs of several proteins, ER and ER-DD models provide a better fit. Note also that for these *SC* RIGs of low distance cutoffs, the diameters of the RIGs are close to being within one standard deviation of the average diameters of GEO-3D networks. Finally, GEO-3D model provides the best fit to RIGs with respect to shortest path length spectra. This is true for all nine proteins, all three contact types, and all 19 distance cutoffs with the exception of the lowest distance cut-offs for *SC* contact type (Appendix Figures D.27 to D.31).

Protein-fold wide topological analysis

To examine the fit of model networks to RIGs corresponding to a larger number of proteins, we analyse Data Set 2. We summarise the results of the fit of each of the five network models to these 1,272 RIGs with respect to each of the above described network properties, by measuring the percentage of RIGs for which a given network model is the best-fitting null model for a given property, the percentage of RIGs for which a given network model is the second best-fitting null model for a given property, etc. (Figure 4.2). GEO-3D is the best-fitting null model for almost all RIGs with respect to all network properties except for the degree distribution. With respect to

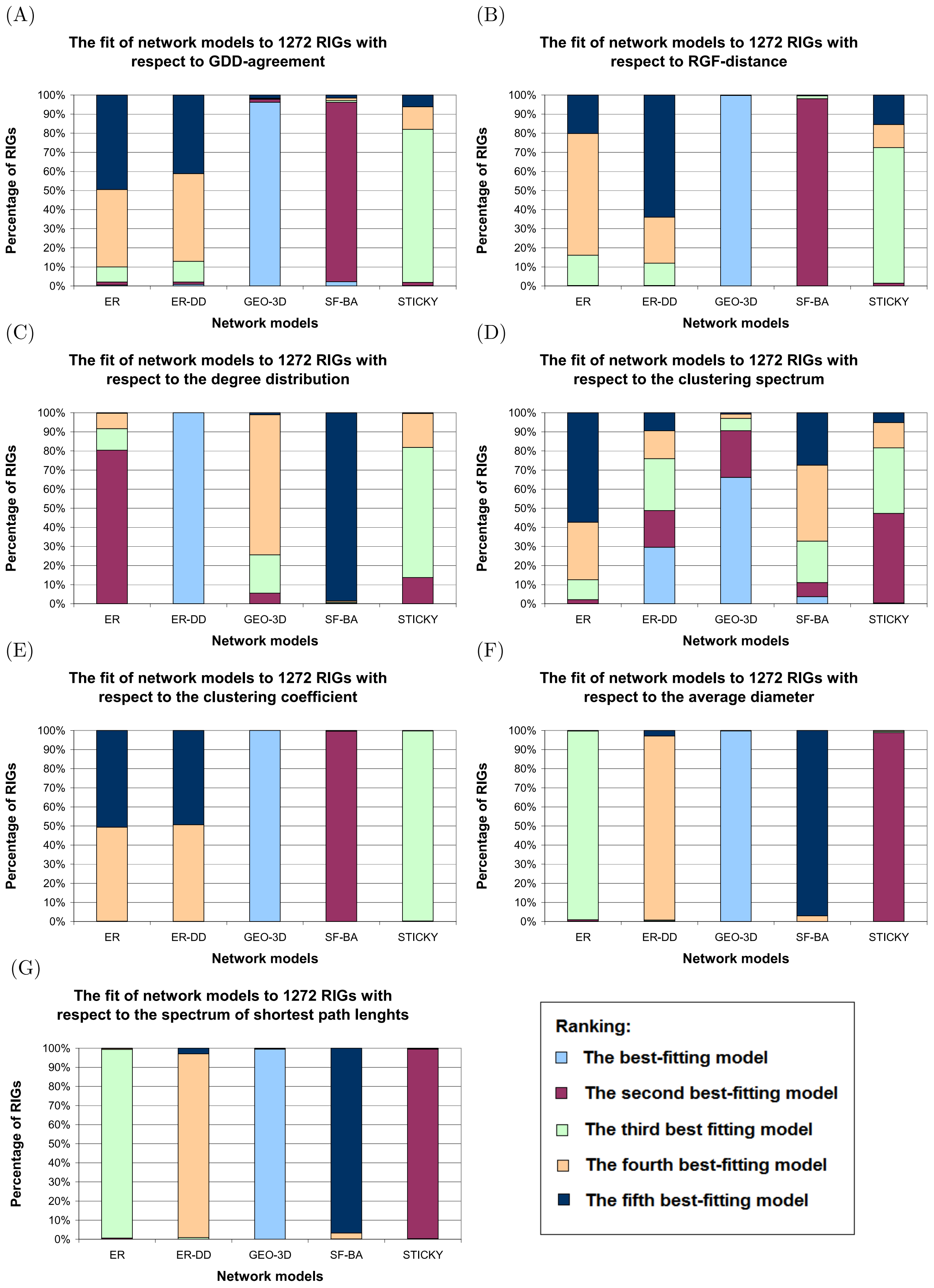


Figure 4.2: The ranking of five network models (ER, ER-DD, GEO-3D, SF-BA, and STICKY) for 1,272 $(ALL)^{\text{all}}_{5.0\mathring{\text{A}}}$ RIGs based on GDD-agreements (A), RGF-distances (B), and agreements between degree distributions (C), clustering spectra (D), clustering coefficients (E), average diameters (F) and spectra of shortest path lengths (G).

this property, ER-DD is by definition the best-fitting null model for all RIGs (Figure 4.2C). Since STICKY networks are defined to have the expected degree distributions of real-world networks, and ER and GEO-3D networks have Poisson degree distributions as all-atom RIGs do [12, 70, 106], all three of these models reproduce the degree distributions of all RIGs. As expected, BA-SF model networks that have power-law degree distributions do not reproduce the degree distributions of any of the 1,272 RIGs (Figure 4.2C). Table 4.1 summarises the results for the fit of geometric random graphs.

Table 4.1: The ranking of GEO-3D graphs for 1,272 $(ALL)^{\text{all}}_{5.0\mathring{A}}$ RIGs based on seven network properties. Rank, the most frequent rank of GEO-3D among five network models (ER, ER-DD, GEO-3D, SF-BA, and STICKY); Freq., the percentage of RIGs for which GEO-3D has the specific rank. The effective rank with respect to degree distribution is given in parentheses; ER-DD and STICKY are expected to have the degree-distribution of RIGs.

Property	Rank	Freq.(%)
GDD-agreement	1	96
RGF-distance	1	100
Pearson correlation between Degree Distributions	4(2)	73
Percentage difference of Clustering Coefficients	1	100
Pearson correlation between Clustering Spectra	1	66
Percentage difference of Average Diameters	1	100
Pearson Correlation between Shortest path lengths Spectra	1	99

Similar results are obtained for all RIGs in Data set 3. GEO-3D is the best-fitting null model for almost all RIGs corresponding to both thermophilic (Appendix Figure D.32) and mesophilic proteins (Appendix Figure D.33). This is true for all network properties, with the exception of the degree distribution, which behaves as explained above.

4.3.2 The quality of the fit of geometric random graph model

Here, we examine how protein size, structural class, protein thermostability, and quaternary structure affect the strength of the fit of geometric random graphs.

Protein size

We first analyse whether the strength of the fit of GEO-3D to RIGs changes with RIG size. Here, we consider all 1,272 RIGs from Data Set 2. Our data points are network property values describing the agreement of a RIG of a given size and the

GEO-3D model. If there exist more than one RIG of a given size, we average the network property value over all such RIGs. We find that the fit of GEO-3D is strongly correlated with RIG size and that this correlation can be expressed as a power-law function $f = a * x^b + c$. We find such function that fits the data in the least-squares sense, minimising the sum of squares due to error (also called the summed square of residual), for each of the network properties (Figure 4.3 and Appendix Figure D.34). We quantify the goodness of fit of each of the power-law functions to the observed correlation data with R-Square (RS) measure (Appendix Figures D.34 and D.35). R-Square illustrates how successful the fit is in explaining the variation of the data; it takes values between 0 and 1, with larger values indicating a better fit. The fit is good for almost all network properties (RS values above 0.76). The only exceptions are RGF-distance (RS of 0.43), the clustering spectrum, and the spectrum of shortest path lengths (RS values of about 0.17).

As protein size increases, the fit also noticeably increases with respect to GDD-agreement and degree-distribution (Figure 4.3, panels A and C). This trend is also observed, in somewhat less pronounced way, with respect to RGF-distance (Figure 4.3B). Surface residues are less well packed compared to buried residues, leading to a heterogeneous density distribution. However, for larger proteins, the percentage of buried residues, as well as the packing density of the solvent-exposed residues increase [91]. Therefore, for larger proteins, the degree distribution and the interaction patterns of the residues become more homogeneous, and thus, the network topology is better reproduced by the geometric random graphs. In these cases where the fit of the GEO-3D graphs to RIGs correlates well with size, the fit itself improves rapidly up to approximately 200 residues and then it slowly converges (Figure 4.3). This behaviour has also been observed in the average protein packing as a function of the size and has been attributed to the size distribution of mono-domain proteins [91].

Average diameters of both RIGs and GEO-3D graphs increase with protein size, while clustering coefficients slightly decrease (Figure 4.3, panels E and F). The fit of GEO-3D to RIGs with respect to these properties is independent of protein size. The small world character of the RIGs is not severely affected by increase in protein size [12]. This is also why the “small-worldness” of GEO-3D graphs agrees almost equally well with that in RIGs, independent of size. Similarly, the fit of GEO-3D shows no correlation with protein size with respect to clustering spectrum and spectrum of shortest path lengths (Figure 4.3, panels D and G). It must be pointed out that as protein size increases, GEO-3D graphs have slightly higher clustering coefficient compared to the RIGs. This means that the random graphs are more compact than expected and probably more “spherical”.

Fold class and secondary structure

We also examine whether the strength of the fit of GEO-3D depends on the protein secondary structure. We analyse RIGs in Data Set 2 that belong to the four structural classes $all-\alpha$ (a), $all-\beta$ (b), $\alpha\,/\,\beta$ (c), and $\alpha+\beta$ (d). Since GDD-agreement is not only the most constraining network property, but also encompasses all other network properties [240], we perform this analysis with respect to GDD-agreement only. First,

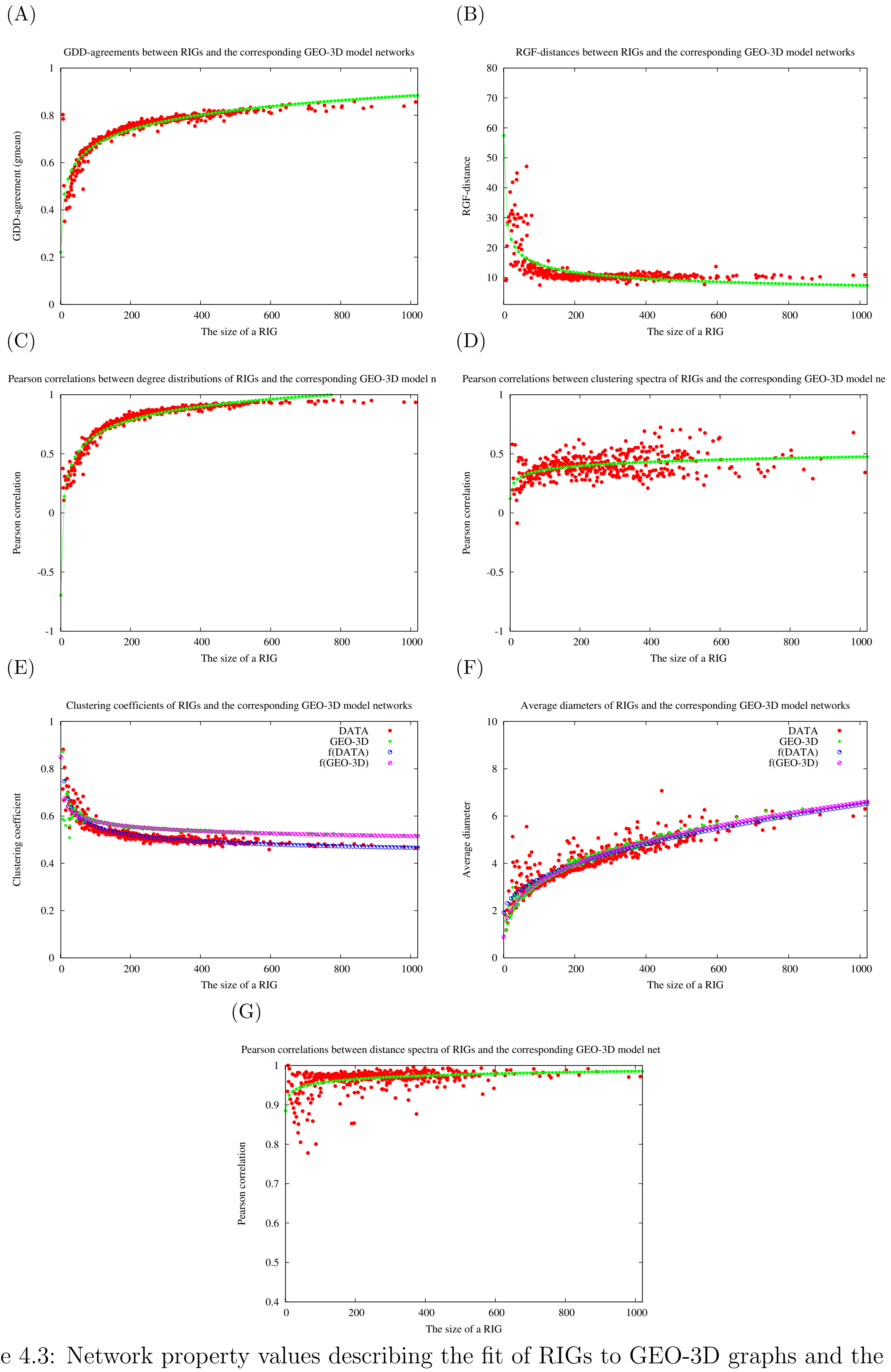


Figure 4.3: Network property values describing the fit of RIGs to GEO-3D graphs and the fitted power-law functions with respect to protein size. 1,272 RIGs from Data Set 2 were analysed with respect to: (A) GDD-agreements, (B) RGF-distances, (C) agreements between degree distributions, (D) agreements between clustering spectra, (E) clustering coefficients of RIGs and the corresponding GEO-3D model networks, (F) average diameters of RIGs and the corresponding GEO-3D model networks, and (G) agreements between spectra of shortest path lengths.

we calculate the mean GDD-agreement, density and volume-to-surface ratio per structural class and for various ranges of protein size (Appendix Figure D.36). We verify for our data set, that $\alpha\,/\,\beta$ proteins are more compact compared to proteins of equal size from other structural classes, similar to [93]. Despite our previous observation that the change in volume to surface area as protein size increases leads to better fitting for larger proteins, the fit of GEO-3D graphs is not higher for more compact proteins of equal size.

We evaluate the statistical significance of the difference of the fit of GEO-3D across structural classes. As before, we find that within each structural class, there exists a strong correlation between the fit of GEO-3D and protein size and that its correlation can be expressed as a power-law function (Appendix Figure D.34). We remove any bias that might exist due to differences in the distribution of protein size for different classes in the following way. We compare the power-law functions that were fitted to the four classes with respect to GDD-agreement (Appendix Figure D.34 and D.35). The functions are evaluated on the RIG size interval that is common to all classes, with protein size ranging from 87 to 501 residues. We assess the statistical significance of the difference between two functions by performing ANOVA statistical test, with p-values close to 0 strongly suggesting that the values of two functions on a given RIG size interval are drawn from different populations. That is, low p-values indicate that the fit of GEO-3D to proteins of a given size belonging to the classes being compared is significantly different.

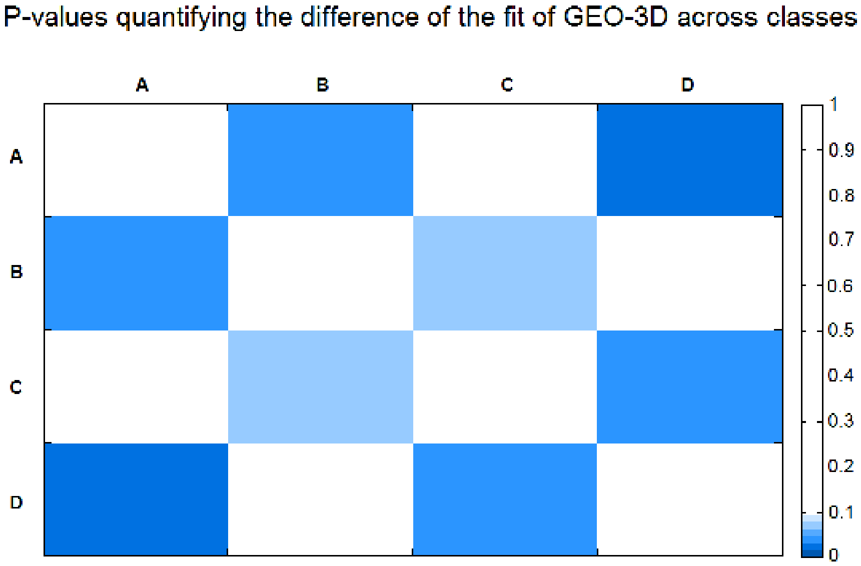
The p-values illustrating the differences in the fit of GEO-3D over all class pairs are presented in Figure 4.4A. The difference in the fit is statistically significant over all class pairs (p-values < 0.077) apart from $all-\alpha$, $\alpha\,/\,\beta$ (p-value of 0.74) and $all-\beta$, $\alpha+\beta$ (p-value of 0.87) pairs. In $\alpha\,/\,\beta$ proteins, the percentage of residues that are in α-helices is higher than the percentage of residues that are in β-strands compared to $\alpha+\beta$ proteins (Appendix Figure D.37A). Thus, $all-\alpha$ and $\alpha\,/\,\beta$ proteins have higher helical content than $all-\beta$ and $\alpha+\beta$ proteins. This further validates the correctness of our GEO-3D model that successfully distinguishes between structurally different classes.

Overall, for proteins of the same size the fit is better for proteins with low helical content, i.e. $all-\beta$ and $\alpha+\beta$ proteins. For large proteins with more than 300-350 residues, the fit of GEO-3D is the highest for $all-\beta$ proteins followed by $\alpha+\beta$ (Figure 4.4B). However, for smaller proteins with less than 300-350 residues, the fit of GEO-3D to $\alpha+\beta$ proteins is higher than to $all-\beta$ proteins, even though they are less compact and have lower β-strand content compared to $all-\beta$ proteins. This could be attributed to the higher percentage of non-regular secondary structural elements, being neither helix nor strand (e.g. loop), in $all-\beta$ proteins of small size (Appendix Figure D.37B).

Protein thermostability

Thermophilic proteins are on average shorter and have higher average connectivity and clustering coefficient compared to mesophilic ones [249]. Moreover, the increase in packing density is observed only for already highly connected residues [249] and for solvent-exposed ones [99]. After verifying that GEO-3D is the best fitting model for al-

(A)



(B)

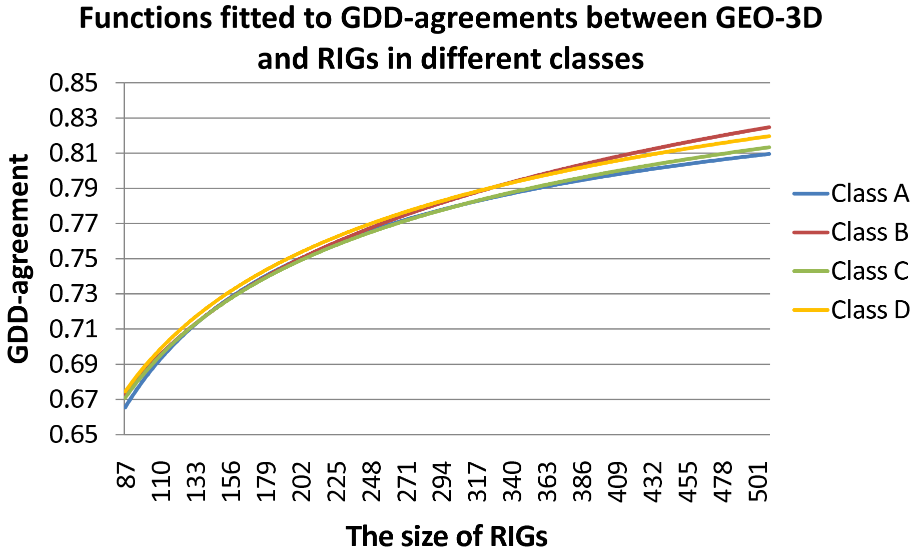


Figure 4.4: The fit of GEO-3D to RIGs in different structural classes. (A) *P*-values quantifying the difference in the fit of GEO-3D to proteins of a given size belonging to four different structural classes, $all-\alpha$ (A), $all-\beta$ (B), α/β (C), and $\alpha+\beta$ (D), with respect to GDD-agreement. Low p-values indicate that the difference in the fit between two classes is significant. (B) The functions that are fitted to GDD-agreements between GEO-3D and RIGs in different classes, from which these p-values are computed.

most all RIGs corresponding to both mesophilic and thermophilic proteins, we evaluate the effect of the structural features responsible for protein thermostability on the degree of fitting. We observe difference in the fit of GEO-3D graphs to thermophilic and the corresponding mesophilic proteins (Appendix Figure D.38). We examine the statistical significance of the difference with respect to all network properties, using Student's pairwise t-test. The results are presented in Table 4.2. We observe significantly higher fitting for mesophilic proteins with respect to GDD-agreement (*diff* $= 0.0087$, $p < 10^{-4}$), degree distribution (*diff* $= 0.0125$, $p = 0.0006$) and clustering coefficient (*diff* $= 0.5484$, $p = 0.0054$). Consistent to our results described above, it is possible that the higher fit of GEO-3D to mesophilic proteins is partially due to their larger size. However, clustering coefficient tends to decrease as protein size increases and thus, we conclude that the observed difference could be also attributed to the tighter packing of highly connected, solvent-exposed residues in thermophilic proteins. We also verify that thermophilic proteins are shorter and have higher average connectivity compared to mesophilic proteins. Although we use different contact range, this is in agreement with the original study [249].

Table 4.2: Pairwise comparison of the fitting of GEO-3D to RIGs and of feature of RIGs, between thermophilic proteins and their mesophilic homologs. Tma, mean value for thermophilic (*T.maritima*) proteins; Meso, mean value for mesophilic proteins; Difference, the mean paired difference between thermophilic and mesophilic values; p-value, p-value for Student's paired t-test; Statistically significant differences are shown in bold and are determined with a threshold of p=0.05/9=0.0055 (including the Bonferroni correction over 9 tests for 5% significance threshold).

Property	Tma	Meso	Difference	p-value
GDD-agreement	0.7586	0.7673	**-0.0087**	$\boldsymbol{p < 10^{-4}}$
RGF-distance	12.1262	12.1483	-0.0221	$p = 0.4379$
Pearson correlation between degree distributions	0.8244	0.8369	**-0.0125**	$\boldsymbol{p = 0.0006}$
Percentage difference of clustering coefficients	9.0433	8.4949	**0.5484**	$\boldsymbol{p = 0.0054}$
Pearson correlation between clustering spectra	0.4966	0.5282	-0.0316	$p = 0.0176$
Percentage difference of average diameters	6.1568	6.5928	-0.436	$p = 0.1183$
Pearson correlation between shortest path lengths spectra	0.964	0.9628	0.0013	$p = 0.2625$
Protein size	252.6702	261.8936	**-9.2234**	$\boldsymbol{p = 0.0029}$
Average degree	4.6679	4.6029	**0.065**	$\boldsymbol{p < 10^{-4}}$

Quaternary structure

We determine the effect of the quaternary structure to the fit of GEO-3D to RIGs. The network topology on the surface of a protein is expected to differ between monomers and multimers. Protein-protein interfaces tend to be more hydrophobic than the non-interface surface, while interface residues are more well packed [142]. We analyse 75 pairs of monomeric and multimeric proteins from Data Set 2. Proteins in each pair have equal size and belong to the same structural class, eliminating any bias due to these structural features. We compare the strength of the fit of GEO-3D to monomers with its fit to the corresponding multimers using Student's pairwise t-test over all pairs. We do this with respect to each of the network properties. Additionally, we compare clustering coefficients and average diameters of monomers with those of multimers using the same test. The results are presented in Table 4.3. Although monomers have significantly higher number of contacts per residue and lower average diameter compared to multimers, we observe no significant difference in the fit of GEO-3D between monomers and multimers, with respect to any of the network properties.

Table 4.3: Pairwise comparison of the fitting of GEO-3D to RIGs and of feature of RIGs, between monomeric and multimeric proteins. Mono, mean value for monomeric proteins; Multi, mean value for multimeric proteins; Difference, the mean paired difference between monomers and multimers; p-value, p-value for Student's paired t-test; Statistically significant differences are shown in bold and are determined with a threshold of p=0.005 (including the Bonferroni correction over 10 tests for 5% significance threshold).

Property	Mono	Multi	Difference	p-value
GDD-agreement	0.7171	0.7153	0.0018	$p = 0.2584$
RGF-distance	11.2578	11.2188	0.039	$p = 0.4377$
Pearson correlation between degree distributions	0.7285	0.73	-0.0014	$p = 0.3964$
Percentage difference of clustering coefficients	7.4296	7.323	0.1067	$p = 0.3865$
Pearson correlation between clustering spectra	0.3736	0.4062	-0.0325	$p = 0.0288$
Percentage difference of average diameters	5.0247	5.8833	-0.8586	$p = 0.1416$
Pearson correlation between shortest path lengths spectra	0.9718	0.967	0.0048	$p = 0.1081$
Number of edges	813.2733	798.9467	**14.3267**	$\boldsymbol{p = 0.0038}$
Clustering coefficient	0.5317	0.5364	-0.0048	$p = 0.0226$
Average diameter	3.4331	3.6439	**-0.2108**	$\boldsymbol{p = 0.0002}$

4.3.3 Application to motif detection

We perform network motif search in nine $(ALL)^{\text{all}}_{5.0\text{Å}}$ RIGs corresponding to the nine proteins of Data Set 1. We compare the frequencies of all 3- to 5- node subgraphs in RIGs to eight network models. Since we have already shown that GEO-3D networks provide the best fit to RIGs with respect to graphlet-based measures, subgraphs exhibit low Z-scores and low M-factors when RIGs are compared against geometric random graphs, as expected. On the contrary, with all other network models, a large number of subgraphs have exceptionally high Z-scores and M-factors. Therefore, GEO-3D model exhibits the highest "specificity" in the selection of network motifs. The absolute Z-scores and absolute M-factors of all 3- to 5- node subgraphs are presented in Appendix Figures D.40 to D.48.

The number of motifs and anti-motifs identified in the nine RIGs with respect to the eight network models are presented in Figure 4.5. We used the same P-value and M-factor thresholds for detection of all (anti-)motifs in all RIGs and with respect to all network models. In all nine RIGs, the fewest number of subgraphs are identified as (anti-)motifs when GEO-3D graphs are used as the null model for (anti-)motif detection. Interestingly, in the case of GEO-3D only anti-motifs are identified for all nine proteins. The CLUST model, that preserves the clustering coefficient of all residues, and the GEO-3D model exhibit the lowest and second lowest number of anti-motifs. We hypothesise that the majority of significant anti-motifs emerge due to null models that do not capture well the packing density of proteins.

Furthermore, we compare the statistical significance the network models assign to subgraphs, independent of the magnitude of the significance itself. We compute Pearson correlation coefficients between all pairs of Z-score vectors corresponding to all pairs of network models for each RIG. We observed Pearson correlation coefficients lower than 0.5 between vectors corresponding to GEO-3D model and vectors corresponding to all other network models (Figure 4.6). Therefore, the results obtained by using geometric network model can not be reproduced with other network models by simple adjustment of the motif selection criteria.

In conclusion, random graph models that preserve only some topological properties of RIGs tend to identify as significantly (under-)over-represented *almost all* analysed subgraphs (Figure 4.5). That is, the statistical significance of the majority of the subgraphs is markedly reduced when being assessed against GEO-3D graphs, especially when compared to commonly used null models that only preserve the size and the degree distribution of a network. Therefore, it is questionable whether non-geometric network models could be used to accurately assess the statistical significance of biologically relevant subgraphs in RIGs.

4.4 Discussion

In summary, we tackle the important issue of finding a well fitting null model for protein structure networks. From the above analyses, we conclude that GEO-3D graphs are the best-fitting null model for various graph representations of protein structures. Our result concurs with previous studies focusing on degree distributions and network properties of protein structure networks [12, 70, 106]: GEO-3D graphs have Poisson degree distributions and exhibit small-world character. We also analyse in detail the relationship between the strength of the fit of GEO-3D to RIGs and protein size, secondary structure, and compactness as observed in thermostable proteins. We have found that structural features of proteins such as the high surface area in comparison

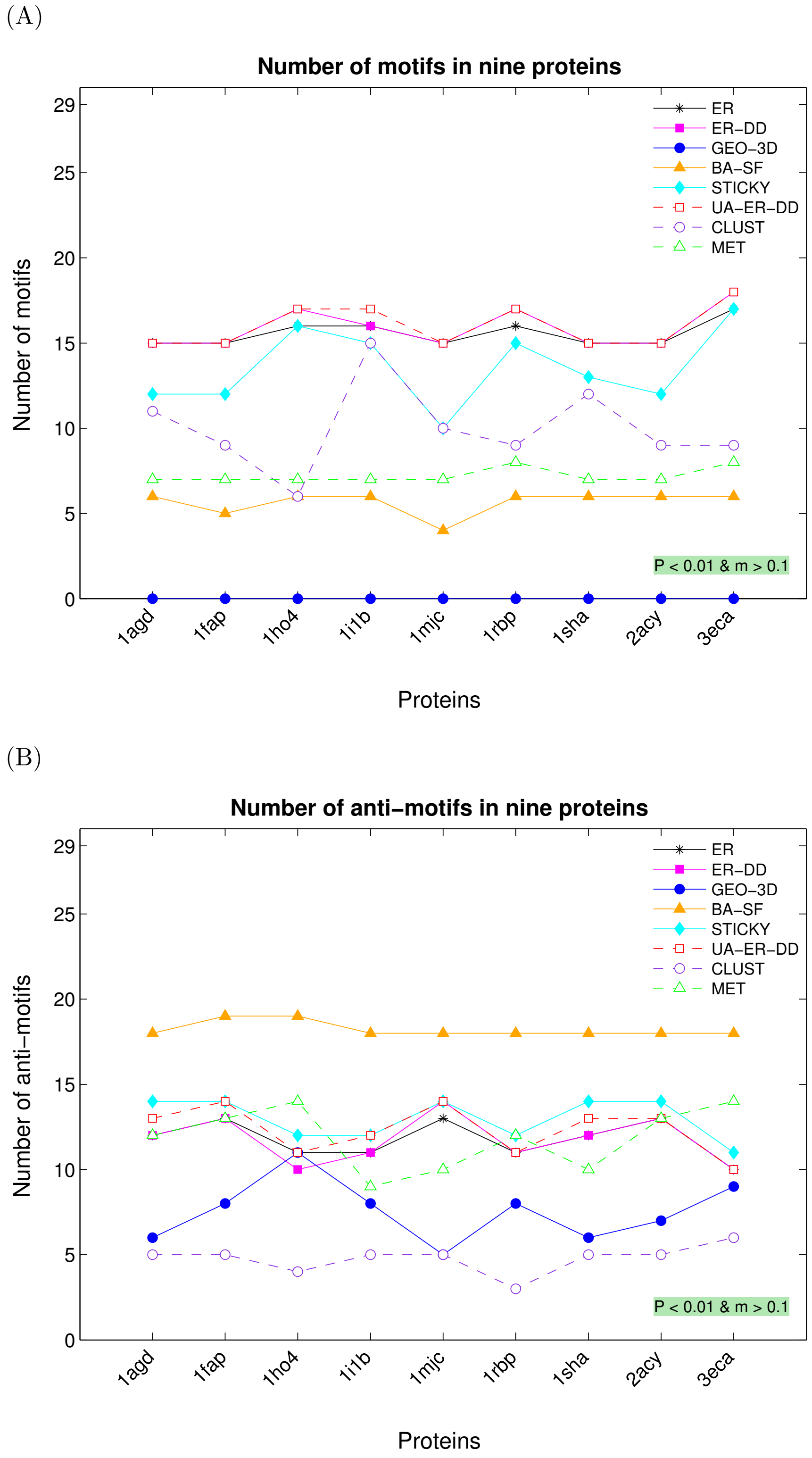


Figure 4.5: The total number of (A) motifs and (B) anti-motifs identified in nine $(ALL)^{\text{all}}_{5.0\mathring{A}}$ RIGs corresponding to the nine proteins (1adg, 1fap, 1ho4, 1i1b, 1mjc, 1rbp, 1sha, 2acy, and 3eca). The motifs and anti-motifs were identified with respect to eight network models (ER, ER-DD, GEO-3D, BA-SF, STICKY, UA-ER-DD, CLUST, and MET). The threshold values used for motif selection (P-value lower than 0.01 and M-factor greater than 0.1) are displayed within the colored textbox.

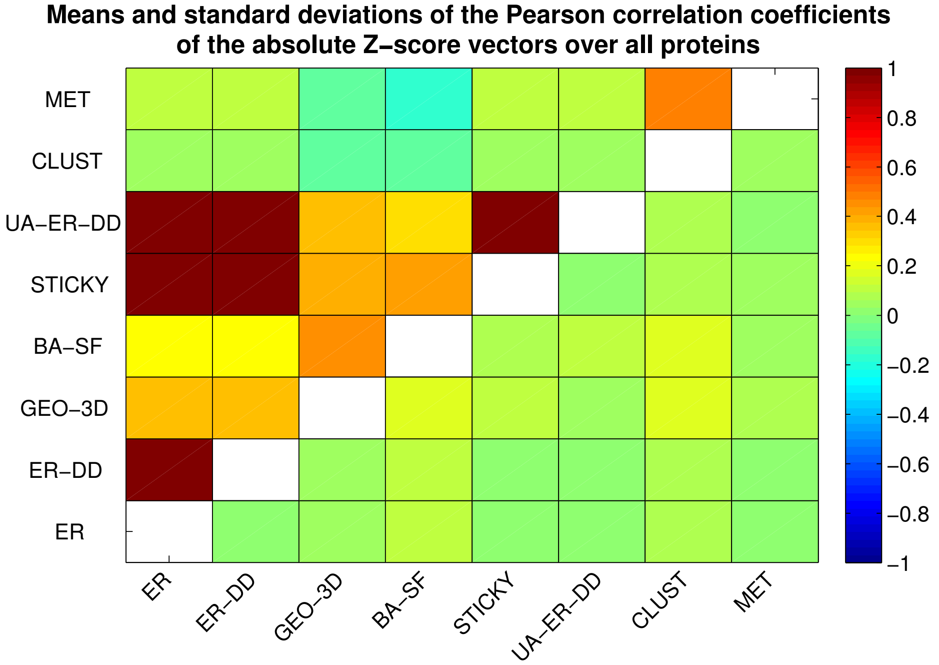


Figure 4.6: Means (upper triangle) and standard deviations (lower triangle) of the Pearson correlation coefficients of the vector of absolute Z-scores of all subgraphs for all possible pairwise combinations of eight network models (ER, ER-DD, GEO-3D, BA-SF, STICKY, UA-ER-DD, CLUST, and MET) over nine $(ALL)^{\text{all}}_{5.0\mathring{A}}$ RIGs. RIGs correspond to proteins 1adg, 1fap, 1ho4, 1i1b, 1mjc, 1rbp, 1sha, 2acy, and 3eca.

to the buried core in small proteins, the high helical content, and the high packing density for solvent exposed residues give rise to topological features that are not well captured by the GEO-3D graphs. Although our proposed null model is far from ideal, its fit is overall much better than the other investigated network models. GEO-3D graphs model spatial relationships of objects, and therefore, they are expected to mimic well the underlying nature of packed residues in proteins. Our result is even more encouraging considering how specific the geometric random graphs are in the identification of (anti-)motifs.

Our geometric random graph null model will facilitate further graph-based studies of protein conformation. This analysis may also have important implications for protein structure comparison and prediction. For example, Contact Map Overlap (CMO) problem [105] measures protein structural similarity based on a graph alignment of contact maps. A correct random graph model could provide means of assessing the statistical significance of contact map similarity. Additionally, our results may facilitate the structural motif discovery, even in the absence of homologs. Instead of comparing the protein of interest against existing structures, it might be sufficient to compare the observed structural network against the randomised counterparts. Finally, it would be interesting to investigate to what extent our analysis could contribute to reliable discriminatory functions that can distinguish near-native conformations from non-native ones. Graph properties of RIGs have been already utilised in this direction [298]. Similarly, the strength of the fit of geometric random graphs to the RIG of a predicted conformation might indicate how native-like the specific protein conformation is.

The null model proposed here is only topologically similar to protein structure networks. A possible area for improvement is to refine it based on additional biophysical properties. According to the model, nodes correspond to points in space distributed uniformly at random and without any preference. In reality, two residues prefer to be connected based on their sequence separation, their residue type, their secondary structure, or even their neighbourhood. Moreover, the chain connectivity imposes constraints that are currently neglected. Thus, further refinements of the geometric model, that would incorporate these biological properties, are expected to yield an even better fitting null model for protein structure networks.

Acknowledgments

The work presented in this chapter with respect to the fitting of the null models (Section 4.3.1) and the dependence of the quality of the fit of GEO-3D graphs on protein size and structural class was performed in collaboration with Nataša Pržulj and Tijana Milenković [191].

Chapter 5

A graphlet-based whole-residue empirical potential for discriminating native structures from decoys

Summary

The quest for an accurate energy function, the key element to successful structure prediction, remains open after over 35 years of research. Recently, it has been shown that the single-body contact-count potential is as effective as the two-body residue-residue interaction potential. Here, we develop a novel knowledge-based potential that generalises the single-body contact-count potential by considering the extended neighbourhood topology of a residue. In this direction, we utilise a recently developed graphlet-based, highly constraining measure of the similarity of the topology of two nodes and their near vicinities in a network. The "nativeness" of a residue conformation is determined by comparing its neighbourhood topology against ~300,000 residues in a non-redundant, representative data set of 1,473 native proteins. Using a large and standard set of protein decoys we investigate in-depth the performance of our whole-residue potential. It performs sufficiently well outperforming the contact-count potential and exhibiting at least 1.8-fold improvement in the mean performance. This improvement is consistent across various methods of generating decoys with respect to two out of three performance metrics and is more prominent in the most successful fragment-based methods. We show also that our potential can be as competitive as a traditional four-body potential and most important exhibits certain strong complementarities with it. Although the overall performance is far from ideal, this novel investigation open new avenues in the research field and could eventually lead to a significantly improved potential by optimising the RIG definition and by incorporating the strong points of other multi-body approaches.

5.1 Introduction

A key element to successful protein structure prediction is an accurate energy function. The function must capture the relationship between sequence and structure and have the global free energy minimum at the native state [9]. Although a plethora of energy functions has been developed over the last 35 years, the quest for the perfect energy function continues and novel approaches have become crucial.

There are two main classes of energy functions, the physics-based and the knowledge-based approaches. In physics-based approaches, a molecular mechanics force field treats the protein at the atomic level summarising information regarding electrostatics, van der Waals interactions, bonds, and torsion angles. Such force fields as ECEPP [212, 338], MM [4, 5], AMBER [59, 321], CHARM [43, 179, 275] and GROMOS [173] have been developed from small molecular structural data. As physics-based approaches are extremely time consuming, knowledge-based potentials have been developed. The key concept in such potentials is extracting features from known protein structures and performing statistical analysis with respect to random predictions [119, 141, 155, 204, 205, 271, 282]. Maximising the discrimination between a native structure and a decoy may also be used to develop an empirical potential [181, 203, 307]. Knowledge-based potentials are defined either at residue [17, 204, 282] or at atomic level [84, 176, 189] and refer to various features of protein structures such as bond angles [73], solvent accessibility [65, 336], and distance between residues [17, 141, 204] or atoms [176, 189].

Most knowledge-based potentials are based on the sum of two-body interactions [204, 273, 282]. Residue-residue interaction potentials encode the probability of two residues being in contact, e.g. interactions between hydrophobic residues are preferable compared to interactions between hydrophobic and hydrophilic ones. Two-body potentials ignore the protein/solvent boundary [102, 290, 334] and thus, usually are combined with single-body potentials [141, 205] that correlate well with residue burial and hydrophobicity. Still, it has been concluded that two-body potentials are not sufficient for reliable protein structure prediction [31, 187, 304]. The assumption that interactions are independent is inaccurate and the cooperativity of residue contacts needs to be modelled. In this direction, multi-body potentials have been developed. Three-body [169], four-body [88, 94, 159, 209, 269], and even whole-residue potentials [171, 187] may take into account the residue types of all residues in contact, their sequence separation and even the solvent accessibility and have often demonstrated an advantage compared to pair-wise methods.

Recently, it has been shown that the single-body contact-count potential that simply denotes the propensity of a residue type to have a specific number of contacts, is as effective as the two-body residue-residue interaction potentials [34]. It even carries more information in a statistical sense and thus should be studied more. Driven by these observations, we address the challenge of developing a knowledge-based potential that generalises the contact-count energy function and attempts to take into account the cooperativity of residue contacts. We utilise a sensitive graph theoretic measure for quantifying the topological similarity of the extended neighbourhood of two nodes. The graphlet degree vector [196], a generalisation of the node degree, describes the interconnectedness of the neighbourhood of a node up to 4 hops. Therefore, it enables the identification of topologically similar nodes and equivalently of residues that have similar conformation in protein structures. This graph theoretic measure has been successfully applied to predict the biological function of a protein in PPI networks [115, 196], to identify new cancer genes in melanogenesis-related pathways [125, 193] and to facilitate network alignments of biological networks [162, 195].

5.2 Methodology

Here, we present how we tailor the graphlet degree vector to a knowledge-based potential and how we assess its performance and compare it with the contact-count potential. Our approach follows this sequence: in order to implement the scoring function, we construct a library of "native" graphlet degree vectors from 1,437 $(C_\alpha)^{\text{all}}_{8.0\mathring{A}}$ RIGs, corresponding to a non-redundant data set of 1,437 proteins. The scoring function is tested on multiple decoy data sets. Each decoy is converted to $(C_\alpha)^{\text{all}}_{8.0\mathring{A}}$ RIG and each residue is scored by assessing the similarity of its graphlet degree vector against the library. To measure the performance of the scoring function three different performance metrics are used. In the following Sections, the graph theoretical measures and all individual steps are explained in detail.

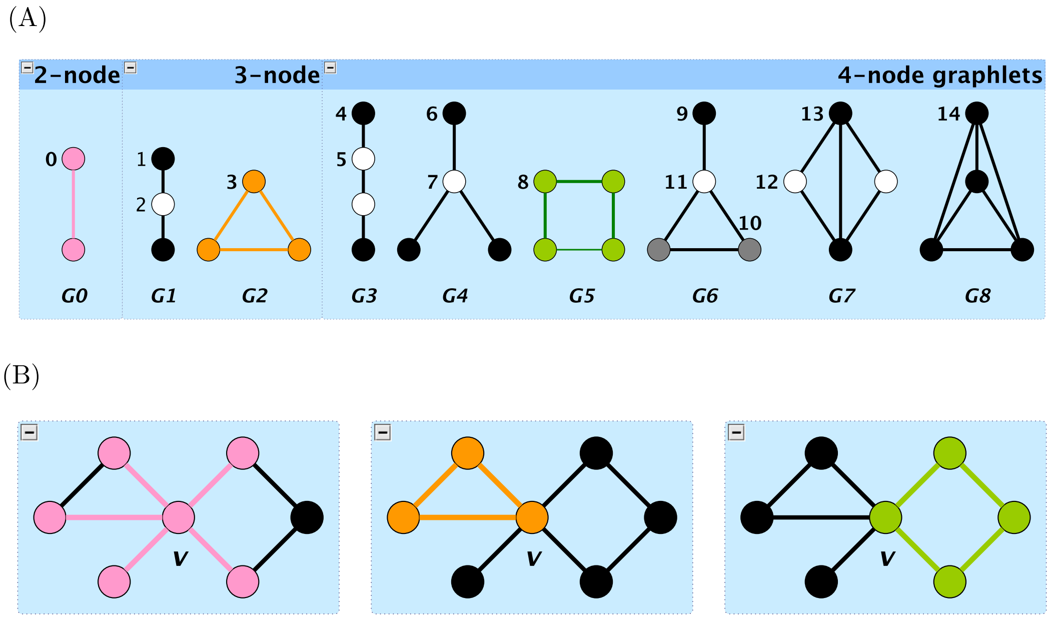
5.2.1 Graphlet degree vector and similarity

The graphlet degree vector and its similarity with another vector have been introduced by Pržulj [196]. The graphlet degree vector (or signature) (GDV) of a node describes the topology of a node and its neighbourhood up to distance 4. It counts the number of all 2- to 5- node graphlets the node "touches", taking into account the corresponding 73 different automorphsim orbits. Figure 5.1 shows an example for the first 14 automorphism orbits. GDV is a 73-dimensional vector where each coordinate corresponds to an orbit and with different weights assigned to each orbit. An orbit can essentially affect other orbits: e.g. occurrences of orbit 0 affect the occurrences of all other orbits and similarly orbit 15 depends on occurrences of orbits 0, 1 and 4. In order to remove this bias, the more an orbit is affected by other orbits, the less important it is and lower weight must be assigned to its occurrences. The weight w_i of an orbit i is defined as

$$w_i = 1 - \frac{log(o_i)}{log(73)}, \tag{5.1}$$

where o_i is the number of orbits affecting orbit i. For example, $o_{15} = 4$ taking into account that each orbit affects itself. Logarithmic function of o_i in the formula increases the weight for more important orbits (those with lower o_i), dividing by $log(73)$ scales the weight to [0, 1] and subtracting from 1 correctly assigns higher weight to more important orbits. Maximum weight 1 is assigned to an orbit that is not affected by any other orbit, i.e. orbit 0.

Comparing the GDV of two nodes is a highly constraining measure of the similarity of their local neighbourhood. The distance $D_i(u, v)$ between nodes u and v for a specific



Orbit i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
GDV(v)[i]	5	2	9	1	0	6	0	7	1	0	0	3	0	0	0

Figure 5.1: Graphlet degree vector example based on 2- to 4- node graphlets. A. All 9 2- to 4- node graphlets and the corresponding 15 automorphism orbits. B. Table corresponds to the graphlet degree vector for a particular node v, i.e. the number of graphelts the node v "touches" for each automorphism orbit. Graphlets 0, 2 and 5 that have a single orbit and that node v "touches" are illustrated in pink, orange and green respectively. Adapted from [125].

orbit i is defined as

$$D_i(u,v) = w_i * \frac{|log(u_i+1) - log(v_i+1)|}{log(max\{u_i, v_i\}+2)}, \tag{5.2}$$

where w_i the weight for orbit i and $u_i(v_i)$ the occurrences of orbit i for node $u(v)$. By using the logarithm of the occurrences of an orbit, the distance measure cannot be dominated by cases where occurrences differ by several orders of magnitude. The relative difference of the occurrences is scaled to [0, 1) and weighted by the orbit importance. The logarithmic function imposes the addition of 1 or 2 to the number of occurrences to prevent it from being infinite or 0.

Consequently, the signature similarity $S(u,v)$ between nodes u and v is defined as:

$$S(u,v) = 1 - D(u,v) = 1 - \frac{\sum_{i=0}^{72} D_i(u,v)}{\sum_{i=0}^{72} w_i}, \tag{5.3}$$

where $D(u,v)$ is their total distance. The similarity as expected is in (0, 1].

5.2.2 The data sets of native proteins and protein decoys

A diverse set of 1,437 proteins was selected to develop the pseudo-potential. A non-redundant, representative set of X-ray structures from the PDB was pre-compiled by the PISCES server [315]. All proteins have resolution better than 1.5Å, reliability factor (R-factor) less than or equal to 0.3, and their pairwise sequence similarity does not exceed 30%. All proteins with less than 30 observed standard amino acids were removed. For each protein, the solvent accessible surface area was calculated using the program NACCESS [132] and each residue was classified to buried when having accessibility lower or equal to 25%, otherwise as exposed. Secondary structure was assigned using the program DSSP [145] and the 8-states of DSSP were converted to three secondary structure states according to EVA conversion scheme [251]. Each protein was converted to RIG using "Ca" contact type and distance cutoff of 8.0Å. For each residue in the dataset, the graphlet degree vector was calculated forming the library of native GDVs.

The Decoys 'R' Us database [255] contains 10 multiple decoy sets where multiple alternative, non-native conformations are provided for each native structure for a set of proteins. In total, this dataset contains 153 proteins and 125,404 decoys. Decoys are generated using widely different methods, one per decoy set. The sets hg-structal, ig-structal and ig-structal-hires are generated using homology modelling and have the lowest median RMSD across all sets, as most of the models are very close to the native. The sets 4state-reduced and lattice-ssfit are based on exhaustive conformational enumeration on lattice models and subsequent filtering. The sets lmds and vhp-mcmd have been obtained by energy minimisation after randomisation (lmds) or molecular dynamics simulation (vhp-mcmd). The set semfold has been assembled using fragment insertion method, while the sets fisa and fisa-casp based on fragments and simulated annealing. The semfold set is the largest one having on average 12,900 decoys per protein. It is obvious that such a diverse data set is sufficient for evaluating energy functions for protein structures in a rigorous way.

5.2.3 Scoring function

Any knowledge-based potential converts the observed properties of known structures and the corresponding randomly expected propensities into an energy-like quantity. For example, the residue contact-count potential is defined as

$$S_{an} = log\left(\frac{P_{an}^{obs}}{P_{an}^{exp}}\right), \tag{5.4}$$

where P_{an}^{obs} and P_{an}^{exp} are the observed and expected probabilities of a residue of amino acid type a having n contacts. P_{an}^{obs} is defined as

$$P_{an}^{obs} = \frac{number\ of\ residues\ of\ type\ a\ and\ degree\ n}{number\ of\ residues\ of\ degree\ n}, \tag{5.5}$$

while P_{an}^{exp} is defined as

$$P_{an}^{exp} = \frac{number\ of\ residues\ of\ type\ a}{number\ of\ all\ residues}. \tag{5.6}$$

The expected values assume the amino acid composition to be independent of degree. The free energy or score of a protein can be calculated as the sum of the score for each residue.

The scoring function of a residue of type a and of GDV g can be defined in a similar way. The observed value of P_{ag} is defined as

$$P_{ag}^{obs} = \frac{number\ of\ residues\ of\ type\ a\ and\ with\ GDV\ similar\ to\ g}{number\ of\ residues\ with\ GDV\ similar\ to\ g}, \tag{5.7}$$

while the expected value as previously is the fraction of residues of type a. In order to identify residues with similar GDV, the signature similarity is utilised.

The scoring matrix for S_{an} can be easily calculated based on a data set of native structures. In order to calculate a similar matrix for S_{ag} the residues in the native structures must be clustered based on their GDV similarities. Then, a residue in a decoy structure can be scored after identifying the cluster it belongs to and based on the scoring matrix. Here, instead of pre-clustering the "native" residues and pre-calculating a scoring matrix, scoring is performed "simultaneously" with a signature threshold-based clustering [196] approach. For each residue in the decoy structures, we identify all residues in the library of native structures that have signature similarity with it above a certain threshold. Then, the residue is scored based on P_{ag}^{obs} as calculated for the set of similar residues. We set thresholds to range from 0.7 to 1.0 in increments of 0.01 and assess the performance separately for each threshold.

Undersampling

The problem of limited number of observations may arise due to calculating probabilities based on a background data set. For example, the number of residues of a specific residue type a and degree n in the data set might be low or even 0 leading to potentially false probabilities. Similarly, the instances of residues of type a and with similar signatures may become rare, especially after a certain similarity threshold. To address this issue of sparse data, we consider two approaches. All cases with less than 5 observed or expected instances are discarded as in [34] and thus, the corresponding residue is neither penalised nor rewarded. In the second approach, we select to penalise such cases with the minimum score for any residue of the examined structure at a specific threshold.

We also address the undersampling issue by introducing two variations of the scoring function. Instead of calculating the score for a specific threshold, we calculate it for the range of thresholds examined and we keep either the maximum score or the score of the maximum threshold for which there are sufficient instances. In the first case, we select the most favourable score for a residue. In the second case, we assume that with increasing threshold the cluster of residues of similar neighbourhood topology becomes more indicative of the nativeness of the conformation of the residue under scoring.

Validation

In order to evaluate the quality of an energy function, various criteria and corresponding performance metrics are important [98]. It must be pointed out that the structure with the lowest energy corresponds to the structure with the maximum score. The native structure should have the lowest energy among all conformations and thus a Rank of 1 when conformations ordered by decreasing score. Second, an energy function should clearly discriminate between the native structure and all decoys. The Z-score of the score of the native structure compared to the average score of decoys should be large and positive. Third, the energy of a conformation should decrease as its structural similarity to native increases. High Spearman rank correlation of the energy with RMSD ensures that the quality of the energy function is independent of the variation in the structural similarity of decoys to native. Here, we use all three performance metrics, Rank, Z-score and Spearman rank correlation, to assess the performance of our scoring function.

5.2.4 Implementation

All RIGs were constructed using OWL [225] (see Section 3.2.5). The code written for performing all calculations reported in this chapter was adapted from software provided by Pržulj [196]. This was an early version of the implementation of graphlet degree vector and signature similarity calculation and of signature threshold-based clustering in GraphCrunch2 [163]. The code was modified so that GDV calculation is integrated in OWL and scoring was implemented separately in a C++ standalone program.

5.3 Results

5.3.1 The "native" graphlet degree vector library

Each structure in the data set of native proteins is converted to a $(C_\alpha)^{\text{all}}_{8.0\mathring{A}}$ RIG and the GDV is calculated for each residue. The resulting "native" GDV library contains 309,085 GDVs. The degree distribution is important in studying networks. Similarly and in order to understand the extent of topological similarity between residues in the data set and select reasonable signature similarity thresholds, we undertake the demanding challenge of calculating the signature similarity for all ∼48 billion pairs of GDVs. The distribution of the signature similarities is plotted in Figure 5.2. The peak is observed in the range [0.80, 0.81) while the number of residues with signature similarities greater than 0.8 decreases exponentially. For our scoring, we select threshold to range from 0.7 to 1.0 in increments of 0.01. Although 42% of residue pairs have similarity less than 0.7, every residue in the data set has signature similarity greater than 0.7 with at least one other residue and on average with ∼90k other residues. Thus, at 0.7 the signature similarities are expected to be uninformative.

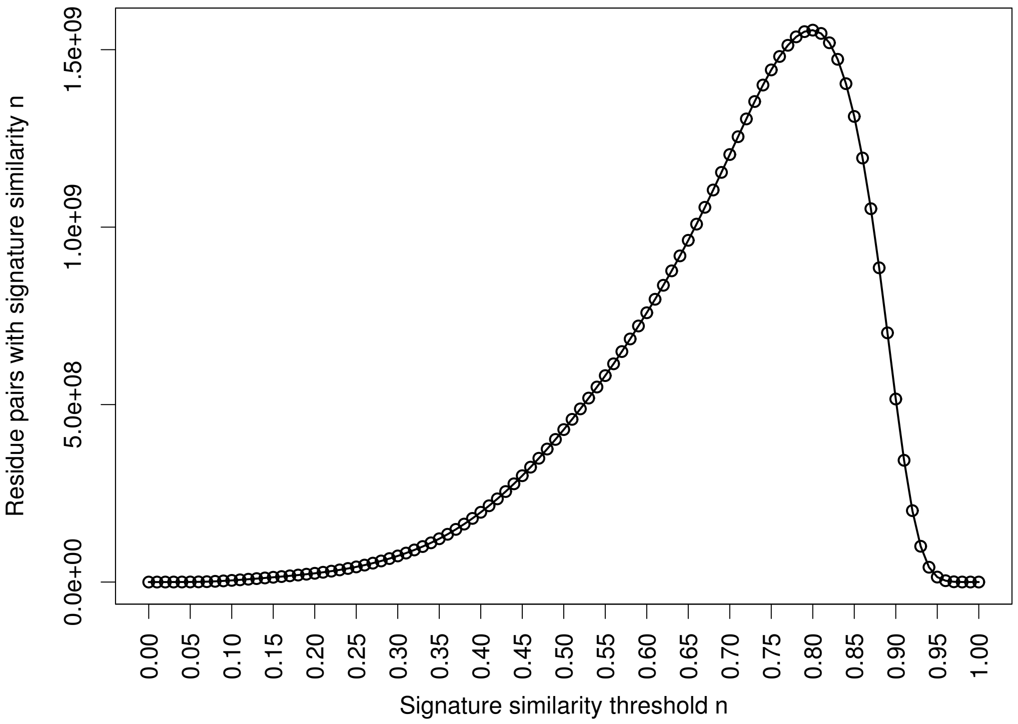


Figure 5.2: Distribution of signature similarities for all pairs of residues in the data set of native structures.

The specificity and statistical "power" of a scoring function is related to the degree to which observed propensities deviate from random. If any residues can have highly similar neighbourhood topology at random and independent of residue type, then the score would be equal to zero. In order to assess the consistency of residue types within clusters of residues of high signature similarity, we perform signature threshold-based clustering of the native GDV library at various thresholds and with a leave-one-out approach. For each residue we identify the cluster of all other residues that have signature similarity with it above a certain threshold. For each cluster i we calculate its redundancy R [232] with respect to a classification scheme s as

$$R_i = 1 - \left(\frac{-\left(\sum_{s=1}^{n} p_s log_2 p_s\right)}{log_2 n} \right), \tag{5.8}$$

where n is the number of classes (possible values) of s and p_s is the relative frequency of the class. R values are in range [0, 1] and the most consistent the classification of residues in a cluster the higher the value of R. We examine four classification schemes: the amino acid type (20 classes), the secondary structure assignment (3 classes), the solvent accessibility (2 classes as buried or exposed) and the degree (23 classes). The average redundancy for each classification scheme and for similarity thresholds in range [0.7, 1] with increments of 0.01 is plotted in Figure 5.3. As expected, as the similarity threshold increases, the redundancy of the clusters increases as well. Clusters are more consistent with respect to the degree of the residues followed by the solvent accessibility, the secondary structure and the amino acid type. Redundancy scores heavily depend on the number of the classes of a scheme. Moreover, amino acids of different type but for example of same hydrophobicity may have similar topology. Despite these, the redundancy values for amino acid type are surprisingly low.

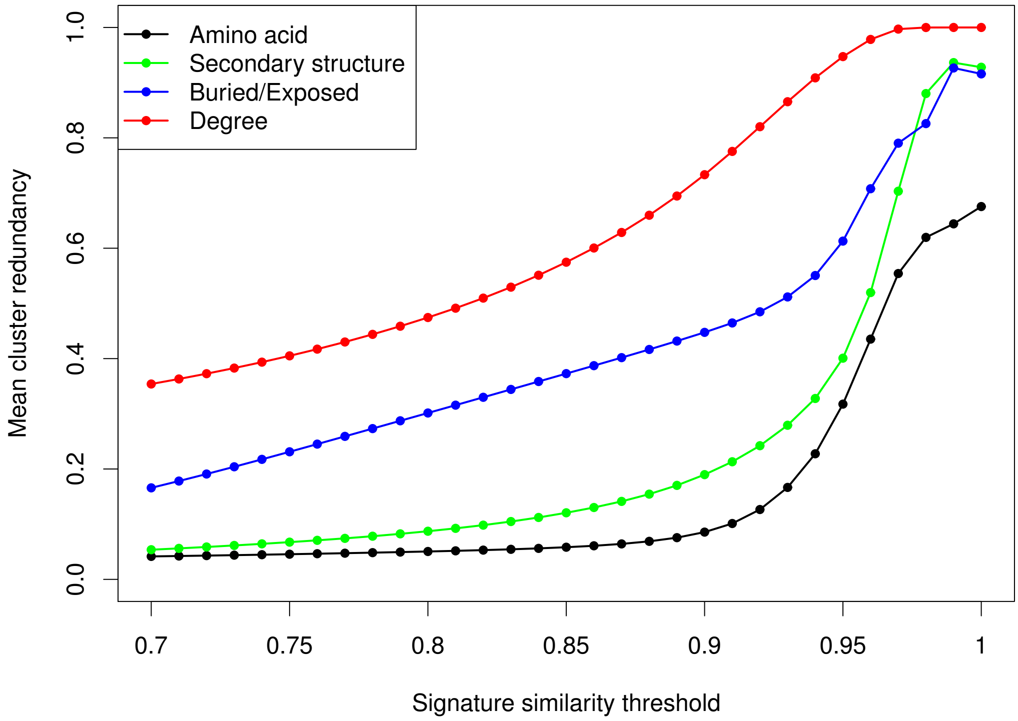


Figure 5.3: Mean cluster redundancy for four classification schemes and with varying signature similarity threshold.

We further investigate the range of signature similarities between residues in the same class and residues in different classes. Figure 5.4A shows, for each residue type, the boxplots of signature similarities with residues of same type, with hydrophobic residues and with hydrophilic residues. Residues are classified to hydrophobic (C, M, F, I, L, V, W, Y) and to hydrophilic (A, G, T, S, N, Q, D, E, H, R, K, P) as in [168]. Residues of same type have much higher similarity when both hydrophobic compared to when hydrophilic. Residues' similarity is on average higher but not notably different for residues of the same type compared to hydrophobic or hydrophilic residues. There is also a trend that a hydrophilic residue has slightly higher similarities with other hydrophilic residues than with hydrophobic ones, while a hydrophobic residue is more similar to hydrophilic residues rather than other hydrophobic ones. It seems that hydrophilic residues, tending to be on the surface of the protein, have a wide variability in neighbourhood interconnectedness independent of residue type. On the contrary, hydrophobic residues have limited variability in neighbourhood topology and rather specific topology to residue type. Figure 5.4B shows the boxplots of signature similarities with respect to the secondary structure. The highest similarities are observed when both residues are in extended conformations, while the lowest when any residue is in loop regions. As expected, regarding solvent accessibility (Figure 5.4C), buried residues have much higher similarity with each other compared to any other case.

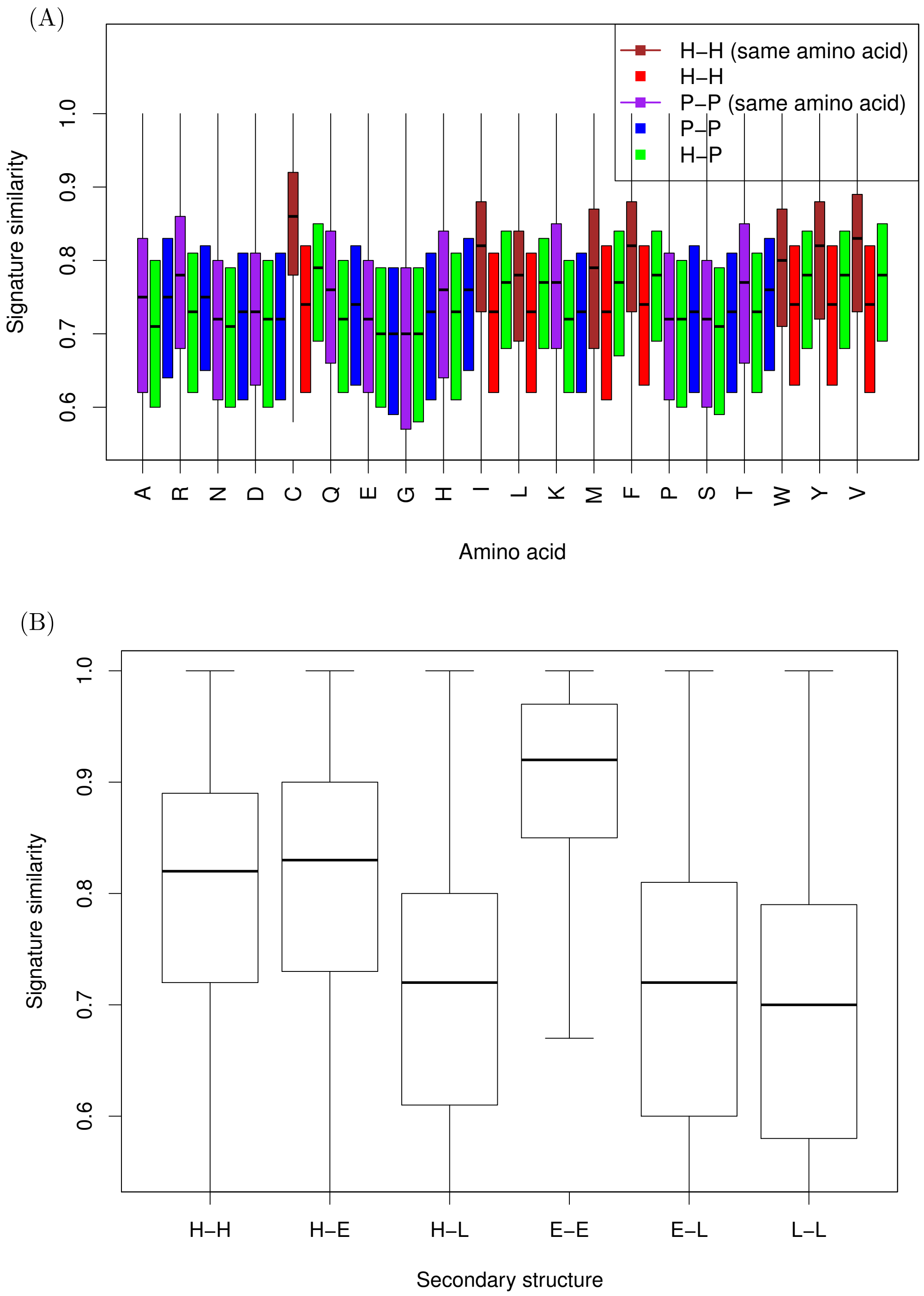


Figure 5.4: Boxplots of signature similarities for all pairs of residues in the data set of native structures with respect to amino acid type (A), secondary structure (B) and solvent accessibility (C). A. For each residue type the boxplots of signature similarities with residues of same type, with hydrophobic (H) residues and with hydrophilic (P) residues are shown. Whiskers are plotted only for residues of same type to visually discriminate between residue types. B. Regarding secondary structure assignment, H, E, and L denote residues in Helical, Extended and Loop conformation. C. Regarding solvent accessibility, B and E denote Buried and Exposed residues. (continued on next page)

(C)

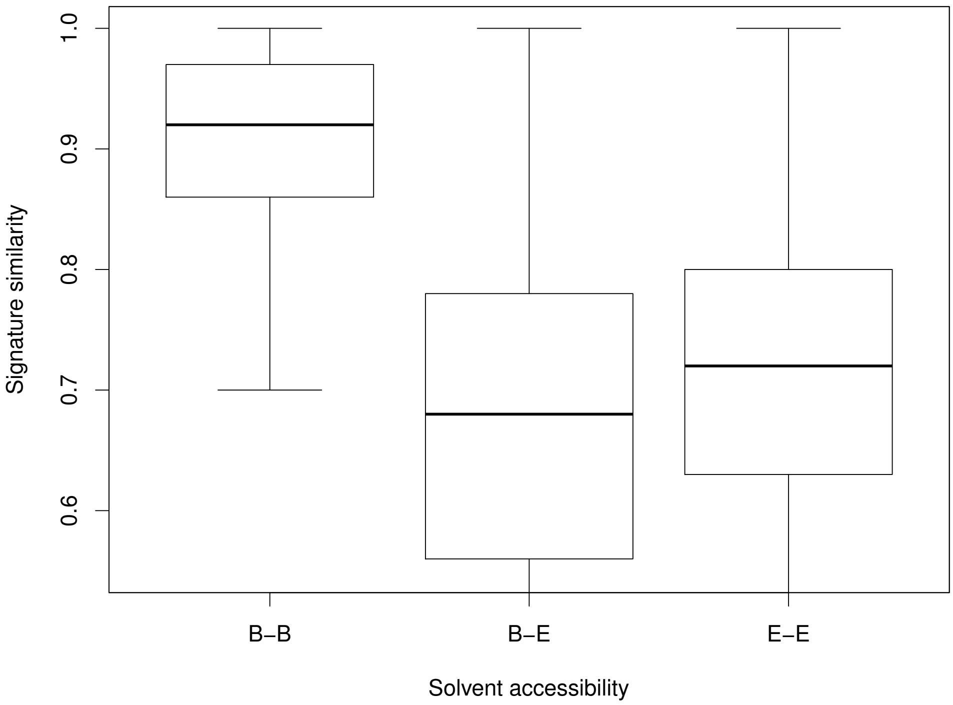


Figure 5.4: continued from previous page

5.3.2 Scoring decoy structures

Each decoy structure is converted to $(C_\alpha)^{\text{all}}_{8.0\mathring{A}}$ RIG and each residue is scored after calculating its GDV and identifying all residues in the native library with signature similarity above a certain threshold. 65 different scores are calculated for each protein. 64 of them are variations of the proposed energy function denoted hereafter as GDVS (i.e. Graphlet Degree Vector Similarity). 31 scores correspond to each one of the thresholds in the range [0.7, 1.0]. Two different strategies in handling sparse data, either discarding such cases (GDVS$_{thresh}$) or penalising (GDVS$^{pen}_{thresh}$) them doubles the total number of scoring functions. Two additional approaches are independent of specific similarity threshold and consider either the maximum (GDVS$_{max}$) or the last (GDVS$_{last}$) valid score as described in Methods Section 5.2.3. The residue contact-count potential was also implemented to allow for a fair comparison with GDVS, unbiased from the choice of the data set of native structures.

In the following section we examine: i) the issue of sparse data in scoring with respect to the choice of similarity threshold, ii) the comparative mean performance of all 65 scoring functions across all decoy sets, and iii) the consistency of the best performing GDVS variations per decoy set.

Undersampling

The choice of the signature similarity threshold can have a large and significant effect on the number of residues not scored due to sparse data. Figure 5.5 clearly demonstrates that the mean percentage of residues not scored per structure increases significantly as the similarity threshold increases as well. At 0.93 almost 60% of the residues are not scored on average per decoy structure. At such high level of sparse data, the scoring function is expected not to perform well. This disagrees with the need for high similarity thresholds where clusters are expected to become more homogeneous. As expected, more residues are not scored on average in decoys rather in the native conformations of the decoy sets.

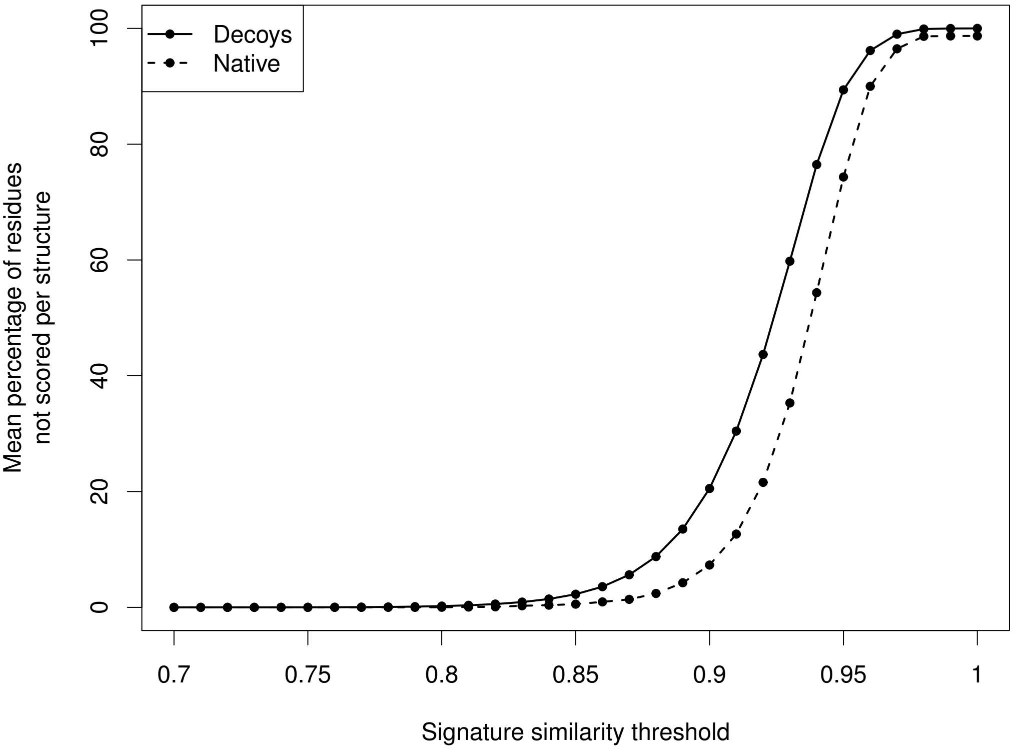


Figure 5.5: Mean percentage of residues not scored per structure in the decoys sets due to undersampling and with varying signature similarity threshold. Solid line corresponds to the decoy structures while dashed line to the native conformations.

The mean performance

The score performance is benchmarked using three different metrics: Rank (R), Z-score (Z) and Spearman rank correlation (S). The decoy quality is independently measured based on the Ca RMSD of the decoy to its corresponding native structure. For each scoring function and for each decoy set we calculate the mean value of the metric over all native proteins in the set. The overall mean performance corresponds to the mean of the mean values for decoy sets. In this way, we assess the performance independently of the method used to generate a decoy.

Figure 5.6 shows the mean performance for all 65 scoring functions based on all three metrics. The trend in performance with respect to similarity threshold is the same for all three performance metrics. As the threshold increases, our scoring function performs better up to a specific threshold value, after which performance deteriorates.

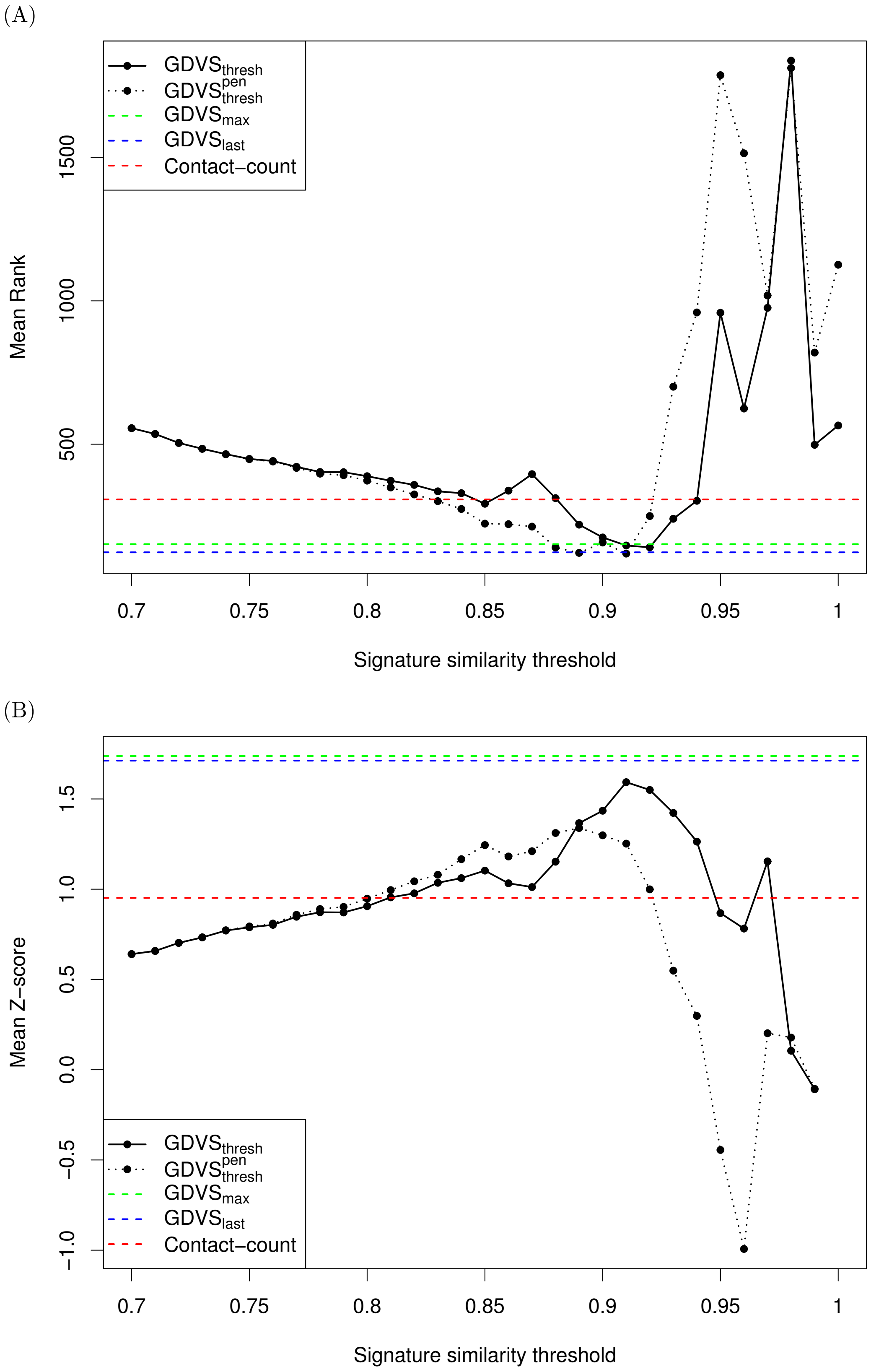


Figure 5.6: Mean performance over all decoy sets. The mean value of Rank (A), Z-score (B) and Spearman rank correlation (C) over all decoy sets using GDVS and contact-count for scoring. The mean is actually the mean of the mean values per decoy set. Dashed lines correspond to threshold independent scoring functions. (continued on next page)

(C)

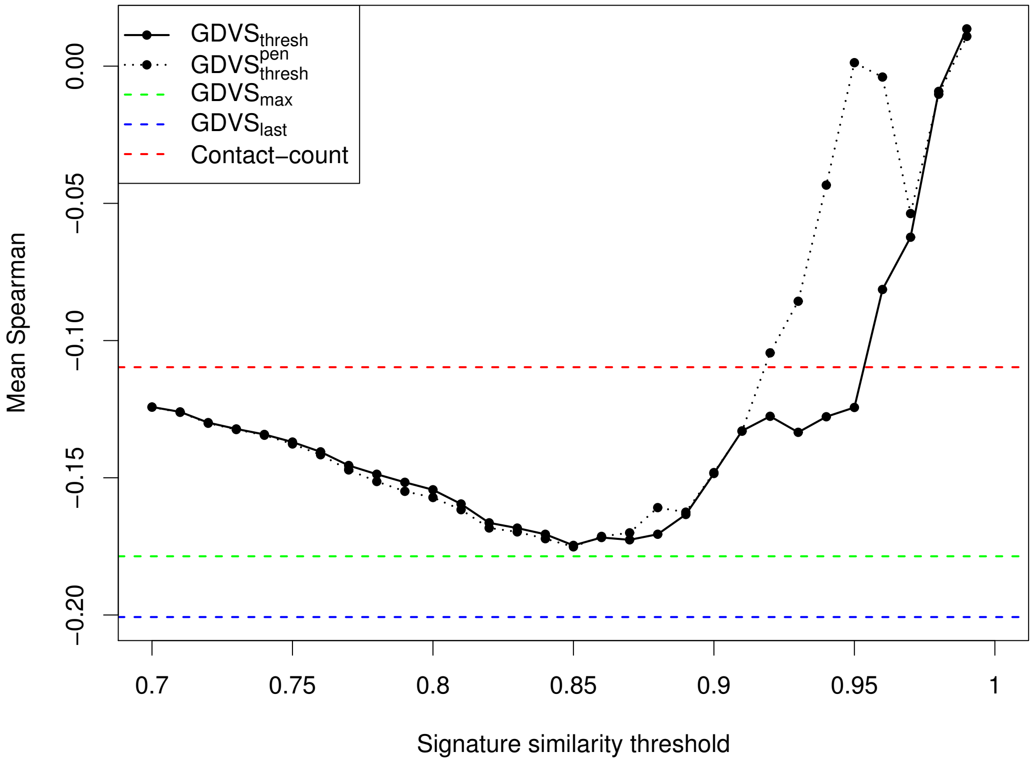


Figure 5.6: continued from previous page

This is independent of the handling of the sparse data. GDVS$^{pen}_{thresh}$ performs slightly better than GDVS$_{thresh}$ in the region where performance improves, then much worse. Penalising sparse cases with the minimum score for a decoy is a reasonable choice only as long as most of the residues for a decoy are scored. The threshold values for which performance peaks for GDVS$_{thresh}$ (GDVS$^{pen}_{thresh}$) are 0.92 (0.91), 0.91 (0.89) and 0.85 (0.85) with respect to R, Z and S. GDVS$_{max}$ and GDVS$_{last}$ are independent of similarity threshold and always perform better or equal to threshold-specific scores for any threshold. The best performing functions are GDVS$^{pen}_{0.91}$ (followed by GDVS$_{last}$) with respect to Rank, GDVS$_{max}$ (followed by GDVS$_{last}$) with respect to Z-score, and GDVS$_{last}$ with respect to Spearman rank correlation. All GDVS$_{thresh}$ scoring functions (for reasonable thresholds) as well as GDVS$_{max}$ and GDVS$_{last}$ perform better than the residue contact-count potential. GDVS$_{last}$, the GDVS scoring function with the best overall mean performance, outperforms residue contact-count potential. Rank, Z-score and Spearman show on average 2.5-fold, 1.8-fold and 1.8-fold improvement in performance of GDVS$_{last}$ compared to contact-count.

Performance per decoy set

It is important that scoring functions perform equally well on all decoy sets. Here, we select four scoring functions, GDVS$^{pen}_{0.89}$, GDVS$_{max}$, GDVS$_{last}$ and contact-count and evaluate how consistent their mean performances per decoy set. Figure 5.7A shows the mean Rank for each decoy set for each one of the four functions. Contact-count

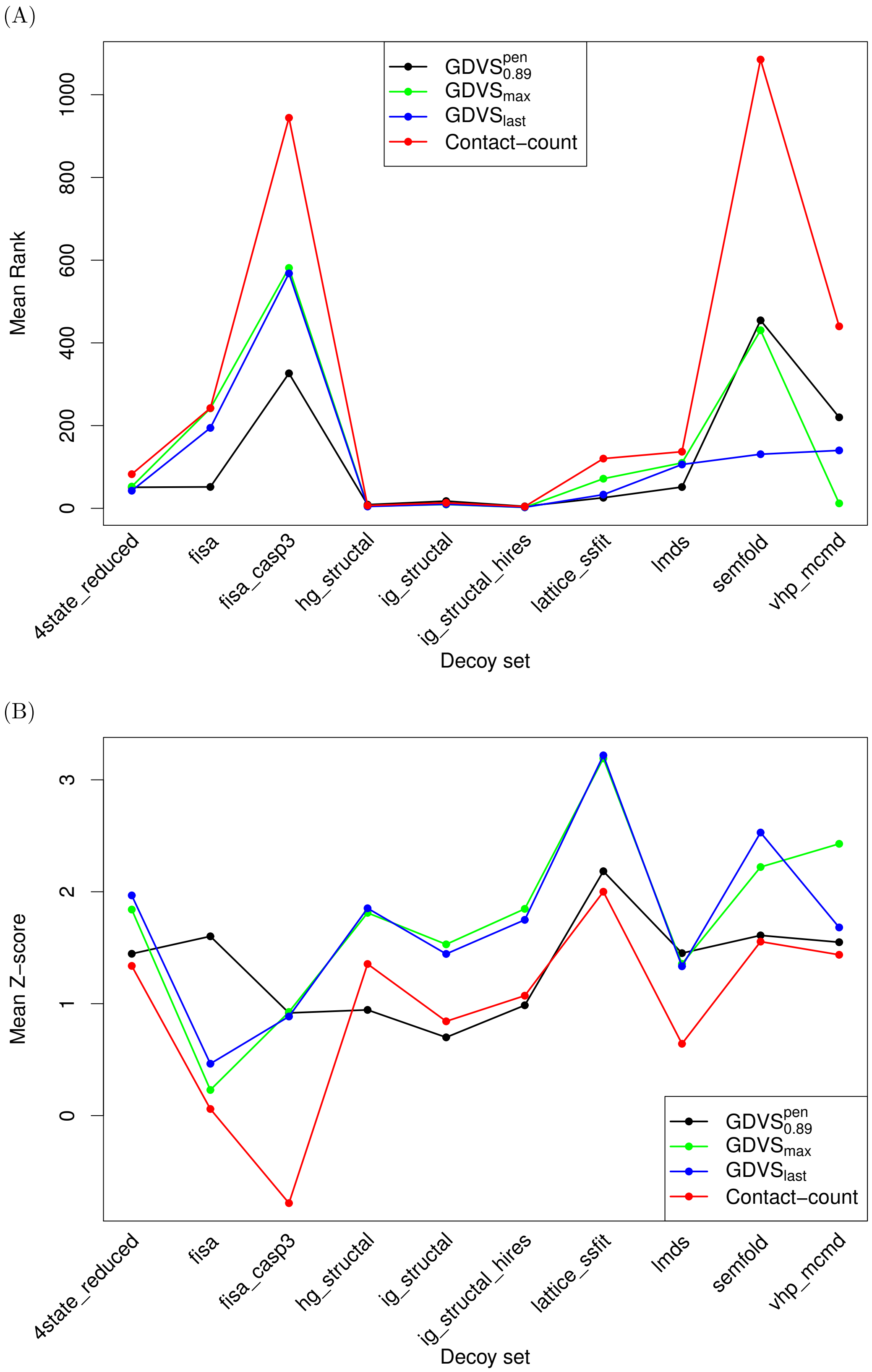


Figure 5.7: Mean performance per decoy set. The mean value of Rank (A), Z-score (B) and Spearman rank correlation (C) per decoy set using GDVS and contact-count for scoring.

(C)

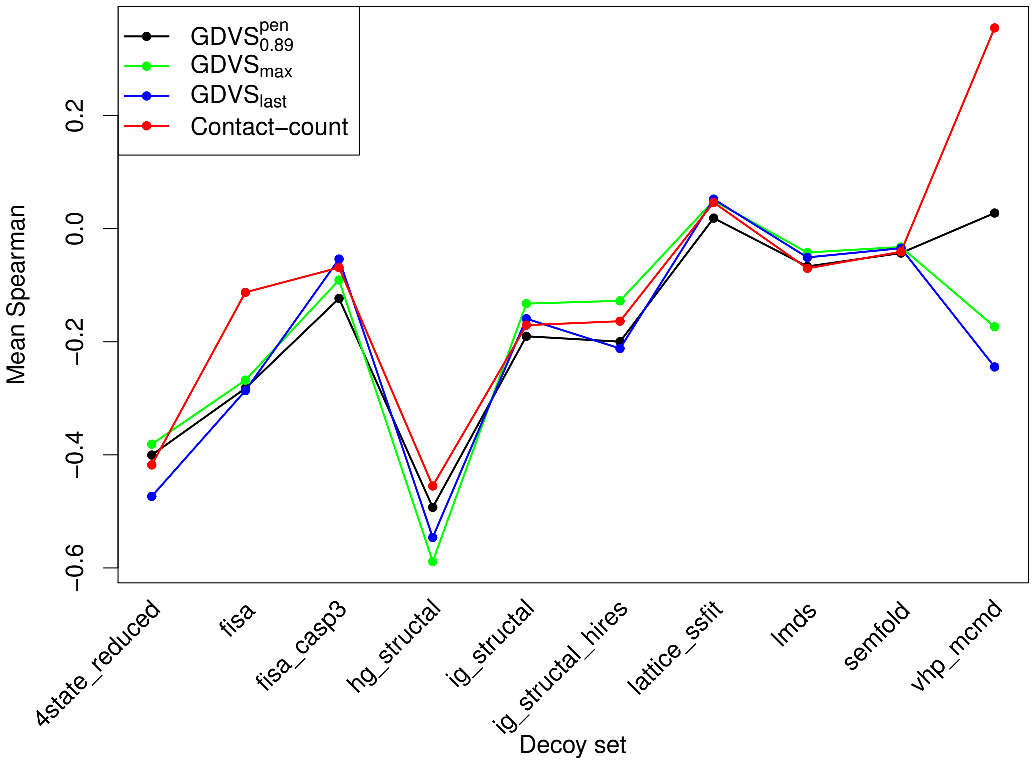


Figure 5.7: continued from previous page

performance is competitive to GDVS only for the homology modelling decoy sets (hg-structal, ig-structal, ig-structal-hires). In these sets, most of the decoys are near-native structures. Interestingly, GDVS performs much better for fragment-based decoy sets (fisa, fisa-casp3, semfold). GDVS$_{max}$ and GDVS$_{last}$ have consistently higher mean Z-score than the contact-count independent of the decoy set (Figure 5.7B). Regarding the Spearman rank, contact-count performs equally well with GDVS functions (Figure 5.7C). In this case, the 1.8-fold improvement is biased by GDVS performance for two decoy sets, fisa and vhp-mcmd.

5.3.3 Comparison with four-body potential

GDVS scoring function performs consistently much better than the contact-count potential with respect to Rank and Z-score. Despite this, its performance its far from ideal. Comparison of GDVS with other potentials that are known to perform better than contact-count or contact-type is difficult. For a fair comparison, the potentials must all be derived from the same data set of native structures and must be applied to the same data set of decoy structures. Moreover, certain features that are integrated in the score and do not lie within the objectives of the comparison, must be handled exactly in the same ways, e.g. RIG definition or the undersampling issue.

Here, we compare GDVS with a four-body potential developed by Tropsha [159]. This potential is based on Delaunay tetrahedra occurring in tessellation of protein structures. Five classes of tetrahedra are defined based on the sequence separation of the interacting residues. The observed amino acid types of all residues in a tetrahedron are utilised in the scoring function. It has to be pointed out that a tetrahedron is conceptually identical to graphlet 8 and automorphism orbit 14. Only Rank and Z-scores for 4state-reduced, lattice-ssfit and lmds decoy sets are provided in [159]. Thus, the comparison does not include decoy sets for which the improved performance of GDVS$_{last}$ compared to contact-count is prominent.

In Table 5.1, we compare the performance of GDVS$_{last}$ with the four-body potential. Both mean Rank and Z-score are better for the four-body potential for the 4state-reduced decoy set and worse for the ldms one. For the lattice-ssfit, GDVS$_{last}$ performs better with respect to the Rank and worse with respect to the Z-score. It must be pointed out that the performance is also strongly complementary. In 10 out of the 12 cases where the Rank for the four-body potential is greater than 6, GDVS performs much better. In 10 cases, four-body performs well and much better than GDVS. Only in 5 cases, both GDVS and four-body perform equally well.

5.4 Discussion

The importance of single-body residue environment potentials has been recently revised in a positive way [34]. As the quest for an accurate energy function remains open after over 35 years of work, we investigate a novel way of generalising the contact-count potential into a whole-residue one. This has been made feasible based on a recently developed graph theoretic measure that describes the topology of a residue and its near vicinity in a network. Here, we implement a novel scoring function, we attempt to understand its biological significance and the relations to structural features, we identify potential issues in scoring and propose alternative strategies to address them and we assess the scoring performance based on a data set of decoy structures.

We have clearly demonstrated that the major bottleneck in the performance of our scoring function is the undersampling issue. As the similarity threshold increases, the clusters of residues of similar signatures become more homogeneous and the scoring function carries more information in a statistical sense. At the same time, the number of sparse clusters increases and derived probabilities for such cases will be inaccurate. At threshold value of 0.9, there are few, if any, observations for more than 20% of the residues in a decoy on average. Therefore, corrections for sparse data are of crucial importance. Penalising based on scores specific to each decoy is preferable to ignoring such cases but only for decoys with a low percentage of penalised residues. The threshold independent strategies improve the mean performance significantly, but not consistently for each decoy set. There are cases where threshold based scores are better than threshold independent ones.

The calculation of a traditional scoring matrix based on pre-clustering of the native library of GDVs would help in this direction. Clustering approaches such as hierarchical clustering, k-medoids, and Markov Cluster Algorithm [82, 296] can be utilised to perform the clustering based on a signature distance matrix [193]. First, such clustering methods might allow for more biological relevant residues to be clustered

Table 5.1: Comparison of performance of GDVS potential with a four-body potential. Four-body potential has been developed by Tropsha based on Delaunay Tessellation [159]. The number of decoys, the Rank and Z-score are shown for each native structure in the 4state-reduced, lattice-ssfit and lmds multiple decoy sets from Decoys 'R' Us database [255]. The mean values per decoy set are also provided and shown in bold.

		GDVS		Four-body	
Proteins	Number of Decoys	Rank	Z-score	Rank	Z-score
1ctf	630	1	3.57	7	2.62
1r69	675	48	1.60	3	2.90
1sn3	660	46	1.50	113	1.04
2cro	674	174	0.64	1	3.04
3icb	653	52	1.49	1	2.90
4pti	687	8	2.49	1	3.18
4rxn	677	39	1.60	5	1.60
4state-reduced	**665**	**53**	**1.84**	**19**	**2.61**
1beo	2000	3	3.49	1	5.35
1ctf	2000	1	5.33	1	4.18
1dkt-A	2000	1	3.74	89	1.67
1fca	2000	385	0.83	1	4.91
1nkl	2000	1	3.67	1	4.38
1pgb	2000	1	3.69	14	2.58
1trl-A	2000	179	1.42	1179	-0.23
4icb	2000	2	3.36	1	5.47
lattice-ssfit	**2000**	**72**	**3.19**	**161**	**3.54**
1shf-A	437	1	3.06	28	1.48
1b0n-B	497	47	1.22	488	-1.93
1bba	500	95	0.84	205	0.20
1ctf	500	1	3.54	1	2.63
1dkt	215	167	-0.85	4	2.06
1fc2	500	239	0.01	372	-0.71
1igd	501	1	4.58	189	0.32
2cro	500	428	-1.04	1	3.88
2ovo	352	103	0.48	46	0.99
4pti	343	17	1.71	7	1.98
lmds	**434**	**110**	**1.36**	**134**	**1.09**

together as a fixed threshold might be too stringent [193]. Second, a scoring matrix enables the usage of existing correction methods for sparse data. For example, the background probability of any residue in a specific conformation (cluster) can be used for an unreliable residue-specific probability [271].

Based on the redundancy analysis for clusters and the range of similarity values with respect to various classification schemes, one would expect that the proposed scoring function will not distinguish well between the 20 amino acid types placed in the same environment, i.e. having similar GDVs. On the contrary, we have shown that our whole-residue potential outperforms the contact-count potential exhibiting at least 1.8-fold improvement in the mean performance. This improvement is consistent across various methods of generating decoys with respect to two out of three performance metrics. It is also evident that GDVS can be as competitive as a traditional four-body potential and most important exhibits certain strong complementarities with it. The four-body potential uses extra information about the protein structure like the amino acid types of all residues in contact and their sequence separation. Integration of such information will enhance the performance of our scoring function.

Interestingly, our scoring function outperforms contact-count especially in fragment-based decoy sets. This has been shown for both Rank and Z-score metrics. The most successful structure prediction method over the last years has been a fragment-based called Rosetta [266]. Rosetta uses a library of three- and nine- residue fragments to assemble various conformations based on these fragments using fragment-insertion method. Each new conformation is evaluated using a low-resolution scoring function based on protein-like features and nonlocal interactions. Recently, it has been argued that the current limitations of Rosetta should be attributed to the low-resolution scoring functions that could be improved [37]. Our scoring function could help in this direction. We hypothesise that the improved performance in fragment-based decoys is attributable to the assessment of nativeness of the spatial neighbourhood of a residue up to 4 hops. Such a deep neighbourhood traverses short structural fragments and captures their interconnectedness. Our scoring function can be also adapted to ignore all interactions between residues within the same fragment.

The proposed scoring function is far from optimal. There are several questions that need investigation and several ways that this work can be extended. The performance of GDVS should be assessed across various RIG definitions. Single-body and two-body potentials have been shown to perform better at higher distance cutoffs, for side-chain based contact-types and when ignoring non-specific short-range interactions [23, 27, 30, 34, 148, 166, 190]. It would be also interesting to assess the independent contribution of each automorphism orbit and the performance with respect to the choice of neighbourhood size in GDV. As already discussed, pre-clustering can potentially improve the accuracy. It will also certainly improve the computationally speed, as the residue preferred conformation will be identified based on the signature similarity with all cluster-centroids and the corresponding score will be selected from the scoring matrix. Incorporation of more structural features such as residue burial, secondary structure and more amino acid types may improve the accuracy of our scoring function. For example, instead of calculating graphlet degree vectors for each residue and defining 20 categories based on amino acid type, we could extend the score to a function of backbone triplets of specific amino acid and solvent accessibility composition [88]. Extending the current work to protein-protein docking will be an important task for the future. It remains to be seen whether GDVs across interfaces have a diverse neighbourhood topology suitable for scoring.

The objective of this work has not been to provide an energy function that outperforms all existing ones but rather to perform a novel investigation that could potentially open new avenues in this research area. The proposed scoring function has been shown to perform sufficiently well and in future it should be feasible to extend it to an even better potential by incorporating the strong points of other multi-body approaches, optimising the RIG definition and addressing more efficiently the undersampling issue.

Chapter 6

Summary and outlook

At this point in time, genomics data are being generated at an unprecedented pace. At the same time, and after 50 years of intensive research, our understanding of protein folding and function is far from complete. Novel and efficient approaches in structural bioinformatics have never been more important. In the past decade, the network view of protein structure space has provided valuable insight. Further development is crucial in order to fully understand the biological significance behind the overwhelming complexity in network representations of protein structures.

To make progress, we need to address some fundamental aspects of the network analysis of protein structures and exploit the avalanche of network theory advances in other research fields and in particular other biological networks. This dissertation tackles three important problems that have not been addressed before:

1. Rationalising the choice of network representation of protein structures and the comparison between various representations.
2. Proposing a well fitting null model for protein structure networks.
3. Developing a novel knowledge-based potential by means of generalising the single-body contact-count potential to a whole-residue pure-topological potential.

The methodology that addresses the last two challenges benefits from recently introduced graph theoretic measures. These are highly constraining measures of the topological similarity either between whole networks or between individual nodes based on their extended neighbourhood. These measures have been successfully applied to protein-protein interaction networks and are based on graphlets, small connected induced subgraphs of large networks. Here, we apply them for the first time to protein structures.

In Chapter 2, the field of structural bioinformatics was revised from a network perspective. All representations, including non-formal network ones, that treat protein structures as sets of interacting residues have been unified by means of a controlled vocabulary. The principal criteria that define a residue-residue interaction have been outlined and the choices for these criteria, their popularity, the motivation behind each one and their optimality with respect to specific research areas were thoroughly discussed. In Chapter 3, the similarity and fundamental network properties for an exhaustive set of network representations were assessed. It has been shown that similarity between different commonly used representations can be quite low. The details of the network representation can have significant impact upon the similarity and the connectivity of the resulting networks, leading often to disconnected components and residues without any interactions. Additionally, we showed that proteins with different secondary structure topologies have to be treated with caution in any network analysis. Overall, it has been clearly demonstrated that an arbitrarily chosen representation is not necessarily similar to other representations, does not exhibit identical network properties, and will not reproduce all results. This work highlighted the importance of selecting an optimal representation for a given application. Both Chapters 2 and 3, allow researchers to rationally select a network representation, either based on the literature and the justification, popularity and optimality of a representation or based on desired network properties and similarity to successfully utilised representations.

Chapter 4 shows that 3-dimensional geometric random graphs, that model spatial relationships between objects, provide the best fit to protein structure networks among several random graph models. The fit has been assessed for a structurally diverse protein data set, various network representations, and with respect to various topological properties. The multitude of local and global network properties, including the graphlet-based highly constraining measures of local topology, overcome the common bias of using a single summary network property for defining a null model. The relationship between the strength of the fit and various structural features has been also investigated. Geometric random graphs capture better the network organisation of larger proteins. Between proteins of equal size, the fitting improves for proteins with low helical content, while the tighter packing of the solvent accessible surface in thermostable proteins leads to worse fitting. Interestingly, quaternary association of proteins plays no significant role in the fitting. Overall, the well fitting null model presented here can be used to accurately assess the statistical significance of topological features observed in networks of protein structures. In this direction, it was shown that choosing geometric random graphs as a null model results in the most specific identification of motifs.

The final chapter in this work reports a novel knowledge-based potential for discriminating native structures from decoys. This potential generalises the single-body contact-count potential to a pure topological whole-residue one. The contact-count potential has been recently shown to perform as well as two-body ones and to be even more informative from a statistical point of view. Its generalisation has been feasible based on graphlet-based measures that describe the topology of a residue and its extended neighbourhood and assess their similarity. The results clearly demonstrate that the proposed scoring function outperforms the contact-count potential. Most important, the improvement in the performance is independent of the methodology of generating decoys and with respect to most performance metrics. The improvement is more prominent in the successful fragment-based methods of generating decoys, making our methodology even more appealing. Although the overall performance is far from ideal, the whole-residue potential has been shown to be on par with a traditional four-body potential and exhibits strong complementarities with it. This highlights that our potential can be further improved since it does not utilise the residue types of interacting residues and their sequence separation as multi-body potentials do.

The analysis performed throughout this thesis has prompted the development of a Java library that focuses on the conversion of protein structures to networks based on various definitions. This library also facilitates subsequent network analysis of protein structures. It is available as Open Source Software and will support further research in this field.

The work presented here can be extended in numerous ways. Network analyses of this thesis have been focused on monomeric, monodomain proteins. The null model has been shown to be optimal, independently of the oligomerisation order of proteins. However, the rational selection of network representation can be extended to multi-domain proteins as well as to protein complexes. Moreover, we plan to extend our scoring function for protein-protein docking. It remains to be seen whether residues in interfaces have more diverse neighbourhood topologies leading to better performance for the potential.

In its present form, the null model proposed in Chapter 4 is topologically accurate. A possible area for improvement is to refine it based on additional biophysical properties. Chain connectivity imposes certain constraints and residues prefer to be connected based on their sequence separation, their residue type, and their secondary structure. Refinements of the geometric model, that would incorporate these biological properties, are expected to yield an even better fitting null model. Furthermore, motif application in Chapter 4 has been utilised only to unravel the specificity of null models. Further analysis of statistically significant under-represented and over-represented subgraphs may reveal biologically significant building blocks for protein folding, function and stability.

There are also three promising avenues for further improvement of our scoring function. Our potential has been examined with respect to the most popular network representation. As has been clearly demonstrated in this work, it is imperative to assess the impact of the representation upon performance and identify the optimal network representation. Alternative strategies that identify clusters of topologically similar residues and a-priori clustering in the set of native proteins may improve the accuracy and certainly the computational speed of the proposed scoring function. Further refinement of the scoring function to include structural features as the majority of multi-body approaches do, is expected to lead to better performance. Apart from adapting the scoring function for protein-protein docking, another topic of future research is to develop our methodology into a low-resolution energy function for Rosetta, the most successful structure prediction method to date.

All in all, the aim of this dissertation has been to establish the basis for the analysis of protein structures as networks. The choice of a network representation, the similarity of various representations, the optimised null model and the respective tools developed here are of fundamental scientific importance and are crucial for further development of this research field. This work also opens the door to new avenues in the quest for the perfect energy function.

Appendix A

Supplementary tables for Chapter 2

Table A.1: Citations for the basic RIG definitions

RIG definition	**Citations**
$(C_\alpha)_{8.0\mathring{A}}$	[2, 15, 33, 45, 53, 95, 108, 110–114, 116, 144, 164, 177, 184, 187, 216, 235–238, 263, 285, 313, 325]
$(ALL)_{5.0\mathring{A}}$	[1, 21, 23, 46, 49, 64, 66–71, 106, 123, 146, 207, 283, 287, 330]
$(SC)_{4.5\mathring{A}}$	[40–42, 96, 97, 121, 147, 223, 256, 258, 259, 272, 273, 310, 322, 333]
$(C_\beta)_{8.0\mathring{A}}$	[35, 56, 86, 87, 135, 136, 178, 197, 221, 222, 243, 253, 257, 261, 288]
$(ALL)_{4.5\mathring{A}}$	[22–25, 32, 85, 124, 201, 203, 239, 249, 250, 286, 307]
$(C_\alpha)_{8.5\mathring{A}}$	[51, 76, 144, 172, 303, 304, 307, 308, 324, 336]
$(C_\alpha)_{6.0\mathring{A}}$	[111, 130, 182, 235–237, 249, 250, 309]
$(ALL)_{6.0\mathring{A}}$	[75, 93, 99, 198, 226, 234, 309, 335]
$(C_\beta)_{7.5\mathring{A}}$	[12, 52, 74, 77, 151, 208, 262, 291]
$(ALL)_{4.0\mathring{A}}$	[90, 127, 198, 220, 228, 230, 327]
$(C_\alpha)_{10.0\mathring{A}}$	[172, 235–237, 254, 299, 300]
$(C_\alpha)_{12.0\mathring{A}}$	[38, 235–237, 285, 297, 313]
$(C_\alpha)_{7.0\mathring{A}}$	[14, 16, 21, 111, 129, 326, 332]
$(C_\alpha)_{\mathrm{DT}}$	[94, 131, 159, 209, 219, 269, 294]
$(C_\beta)_{12.0\mathring{A}}$	[34, 152–154, 166, 277, 331]
$(C_\alpha)_{9.0\mathring{A}}$	[60, 175, 245, 257, 305, 306]
$(SC)_{5.0\mathring{A}}$	[101, 103, 104, 122, 185, 253]
CSU	[8, 83, 247, 276]
$(C_\alpha)_{11.0\mathring{A}}$	[23, 156, 211]
$(C_\alpha)_{6.5\mathring{A}}$	[109, 231, 272]
$(SC_c)_{10.0\mathring{A}}$	[139, 140, 293]
$(SC_c)_{6.5\mathring{A}}$	[204–206]
$(SC_c)_{8.0\mathring{A}}$	[88, 140, 322]
$(ALL)_{8.0\mathring{A}}$	[93, 99]
$(ALL)_{9.0\mathring{A}}$	[47, 48]
$(C_\alpha)_{14.0\mathring{A}}$	[217, 254]
$(C_\alpha)_{7.5\mathring{A}}$	[80, 109]
$(C_\beta)_{10.0\mathring{A}}$	[141, 266]
$(FA)_{12.0\mathring{A}}$	[252, 316]
MC$_d$	[61, 181]
$(SC)_{4.2\mathring{A}}$	[157, 274]

Continued on next page

Table A.1 – continued from previous page

RIG definition	Citations
$(SC)_{6.0\text{Å}}$	[165, 180]
$(SC)_{\sum_{\text{ij}} \text{r}_{\text{vdW}}+1.0\text{Å}}$	[160, 268]
$(SC)_{\sum_{\text{ij}} \text{r}_{\text{vdW}}+2.8\text{Å}}$	[281, 282]
$(SC_c)_{\text{DT}}$	[50, 159]
$(ALL)_{12.0\text{Å}}$	[330]
$(ALL)_{3.0\text{Å}}$	[330]
$(ALL)_{4.1\text{Å}}$	[138]
$(ALL)_{4.2\text{Å}}$	[274]
$(ALL)_{7.0\text{Å}}$	[202]
$(ALL)_{7.5\text{Å}}$	[46]
$(ALL)_{\sum_{\text{ij}} \text{r}_{\text{vdW}}}$	[260]
$(ALL)_{\sum_{\text{ij}} \text{r}_{\text{vdW}}+0.5\text{Å}}$	[295]
$(ALL)_{\sum_{\text{ij}} \text{r}_{\text{vdW}}+1.0\text{Å}}$	[317]
$(SC)_{\sum_{\text{ij}} \text{r}_{\text{vdW}}+2.8\text{Å}}$	[224]
$(ALL_c)_{6.0\text{Å}}$	[309]
$(C_\alpha)_{16.0\text{Å}}$	[33]
$(C_\alpha)_{5.5\text{Å}}$	[328]
$(C_\alpha)_{6.2\text{Å}}$	[79]
$(C_\alpha)_{9.5\text{Å}}$	[131]
$(C_\alpha + C_\beta)_{6.0\text{Å}}$	[320]
$(C_\alpha + SC)_{4.75\text{Å}}$	[267]
$(C_\alpha + SC)_{\sum_{\text{ij}} \text{r}_{\text{vdW}}+1.0\text{Å}}$	[30]
$(C_\beta)_{13.0\text{Å}}$	[289]
$(C_\beta)_{14.0\text{Å}}$	[148]
$(C_\beta)_{6.0\text{Å}}$	[309]
$(C_\beta)_{6.7\text{Å}}$	[13]
$(C_\beta)_{7.0\text{Å}}$	[3]
$(C_\beta)_{8.5\text{Å}}$	[12]
$(C_\beta)_{9.0\text{Å}}$	[27]
$(C_\beta)_{\text{DT}}$	[337]
MC_c	[181]
$(SC)_{4.75\text{Å}}$	[158]
$(SC_c)_{12.0\text{Å}}$	[166]
$(SC_c)_{5.0\text{Å}}$	[11]
	Continued on next page

Table A.1 – continued from previous page

RIG definition	**Citations**
$(SC_c)_{6.0\mathring{A}}$	[264]
$(SC_c)_{8.5\mathring{A}}$	[7]
$(SC_{cs})_{\sum_{ij} r_{sphere}+2.8\mathring{A}}$	[170]
$(SC_{cs}+SC_{cs}/BB_{cs})_{\sum_{ij} r_{sphere}}$	[171]
$(SC_{cs}.BB_{cs})_{10.0\mathring{A}}$	[44]

Table A.2: Citations for the contact ranges

Contact range	**Citations**
all	[1–3, 7, 8, 12–16, 21, 27, 30, 33, 34, 38, 46–51, 64, 66–71, 74–77, 80, 83, 88, 90, 92–95, 99, 101, 103, 104, 106, 108, 110–113, 119, 121, 122, 124, 127, 130, 131, 135, 138–141, 144, 148, 151, 157, 159, 160, 164, 170–172, 180, 184, 185, 187, 198, 201, 207–209, 216, 217, 219, 221, 222, 224, 230, 234–238, 245, 247, 249, 250, 252–254, 257, 260, 262, 266, 267, 269, 272, 273, 276, 280, 281, 283, 285, 287, 289, 291, 293, 294, 297, 299, 300, 303, 305–307, 310, 316, 320, 324, 326–328, 330, 336, 337]
$\mathrm{\lvert i-j \rvert \geq 2}$	[32, 40–42, 96, 97, 147, 155, 166, 172, 198, 202–206, 223, 226, 256, 258, 259, 274, 317, 322, 333]
$\mathrm{\lvert i-j \rvert \geq 3}$	[22, 23, 25, 35, 52, 61, 108, 112, 113, 146, 152–154, 158, 164, 175, 181, 220, 227, 231, 239, 277, 286, 304, 308, 331]
$\mathrm{\lvert i-j \rvert \geq 4}$	[24, 85, 109, 128, 129, 228, 249, 263, 268, 288, 295, 332]
$\mathrm{\lvert i-j \rvert \geq 5}$	[3, 11, 44, 45, 56, 61, 108, 112, 113, 116, 164, 181, 198, 201, 202, 281, 282]
$\mathrm{\lvert i-j \rvert \geq 6}$	[53, 136, 165, 197, 243, 313, 325]
$\mathrm{\lvert i-j \rvert \geq 7}$	[60, 86, 87, 92]
$\mathrm{\lvert i-j \rvert \geq 8}$	[61, 178, 181, 309]
$\mathrm{\lvert i-j \rvert \geq 9}$	[155, 261]
$\mathrm{\lvert i-j \rvert \geq 10}$	[66, 92, 106, 123, 166, 211, 257, 280]
$\mathrm{\lvert i-j \rvert \geq 11}$	[119, 141, 264, 281]
$\mathrm{\lvert i-j \rvert \geq 12}$	[53, 136, 182, 197, 243, 313, 325]
$\mathrm{\lvert i-j \rvert \geq 13}$	[15, 109, 114, 177, 227]
$\mathrm{\lvert i-j \rvert \geq 15}$	[335]
$\mathrm{\lvert i-j \rvert \geq 20}$	[280]
$\mathrm{\lvert i-j \rvert \geq 24}$	[53, 136, 243, 313, 325]
$\mathrm{\lvert i-j \rvert \geq 31}$	[119, 141]
$\mathrm{s_i \neq s_j}$	[79, 156]

Appendix B

Supplementary tables for Chapter 3

Table B.1: Selected folds and the corresponding proteins. Domains40 (Domains95) is the number of SCOP domains of a specific fold with pairwise sequence similarity less than 40% (95%). The corresponding percentage with respect to the whole structural class is reported in parenthesis.

Fold	**Domains40**		**Domains95**		**PDB id**	**Chain**
all-α						
DNA/RNA-binding 3-helical bundle	280	(16)	386	(14)	1odd	A
alpha-alpha superhelix	94	(5)	121	(4)	1elk	A
SAM domain-like	67	(4)	89	(3)	1sv4	B
Four-helical up-and-down bundle	66	(4)	93	(3)	1jmw	A
RuvA C-terminal domain-like	43	(2)	56	(2)	2bwb	A
alpha/alpha toroid	42	(2)	49	(2)	1cem	A
Spectrin repeat-like	40	(2)	50	(2)	1o3x	A
lambda repressor-like DNA-binding domains	25	(1)	43	(2)	1d1m	B
Cyclin-like	17	(1)	18	(1)	1ad6	A
Heme oxygenase-like	13	(1)	18	(1)	1irm	C
CH domain-like	12	(1)	19	(1)	1bkr	A
Regulator of G-protein signaling, RGS	6	(0)	9	(0)	1iap	A
GTPase activation domain, GAP	6	(0)	6	(0)	1wer	A
PABP domain-like	5	(0)	6	(0)	1i2t	A
S15/NS1 RNA-binding domain	4	(0)	8	(0)	1a32	A
all-β						
Immunoglobulin-like beta-sandwich	369	(18)	1160	(31)	1grw	A
OB-fold	126	(6)	177	(5)	1ntg	A
SH3-like barrel	124	(6)	202	(5)	1pht	A
PH domain-like barrel	72	(4)	98	(3)	1p3r	A
Concanavalin A-like lectins/glucanases	57	(3)	122	(3)	1xnd	A
Galactose-binding domain-like	50	(2)	59	(2)	1wmx	A
Cupredoxin-like	47	(2)	97	(3)	1pzc	A

Continued on next page

Table B.1 – continued from previous page

Fold	Domains40		Domains95		PDB id	Chain
Trypsin-like serine proteases	44	(2)	114	(3)	1agj	A
beta-Trefoil	42	(2)	73	(2)	2i1b	A
Common fold of diphtheria toxin/transcription factors/cytochrome f	41	(2)	67	(2)	1qzn	A
Lipocalins	32	(2)	69	(2)	1kqx	A
6-bladed beta-propeller	27	(1)	30	(1)	1eur	A
beta-clip	24	(1)	38	(1)	3msi	A
Barrel-sandwich hybrid	18	(1)	31	(1)	1onl	A
gamma-Crystallin-like	12	(1)	26	(1)	1dsl	A
α / β						
TIM beta/alpha-barrel	322	(13)	519	(14)	1cwy	A
P-loop containing nucleoside triphosphate hydrolases	232	(9)	409	(11)	1hyq	A
Flavodoxin-like	124	(5)	166	(4)	1e6k	A
Ribonuclease H-like motif	120	(5)	176	(5)	1goa	A
Thioredoxin fold	108	(4)	177	(5)	1o8w	A
S-adenosyl-L-methionine-dependent methyltransferases	94	(4)	120	(3)	1ri5	A
alpha/beta-Hydrolases	70	(3)	108	(3)	1ede	A
Periplasmic binding protein-like II	52	(2)	80	(2)	1uiu	A
Leucine-rich repeat, LRR (right-handed beta-alpha superhelix)	25	(1)	31	(1)	1yrg	A
Rhodanese/Cell cycle control phosphatase	21	(1)	23	(1)	1c25	A
(Phosphotyrosine protein) phosphatases II	20	(1)	28	(1)	1jln	A
7-stranded beta/alpha barrel	19	(1)	24	(1)	1v77	A
Chelatase-like	16	(1)	19	(1)	1ak1	A
Periplasmic binding protein-like I	15	(1)	18	(0)	1ba2	A

Continued on next page

Table B.1 – continued from previous page

Fold	Domains40		Domains95		PDB id	Chain
Dihydrofolate reductase-like	14	(1)	15	(0)	1pdb	A
$\alpha + \beta$						
Ferredoxin-like	302	(13)	430	(12)	1fva	B
beta-Grasp (ubiquitin-like)	117	(5)	211	(6)	1ugm	A
Cystatin-like	49	(2)	58	(2)	1iv9	A
Protein kinase-like (PK-like)	44	(2)	77	(2)	1erk	A
TBP-like	38	(2)	56	(2)	1jss	A
Cysteine proteinases	37	(2)	75	(2)	1iu4	A
Ntn hydrolase-like	37	(2)	59	(2)	1oqz	B
Profilin-like	35	(2)	50	(1)	1ypr	A
Bacillus chorismate mutase-like	28	(1)	52	(2)	1pxw	A
FKBP-like	25	(1)	45	(1)	1r9h	A
dsRBD-like	21	(1)	31	(1)	1t4o	A
IF3-like	19	(1)	27	(1)	1rf5	A
Eukaryotic type KH-domain (KH-domain type I)	19	(1)	25	(1)	1wvn	A
UBC-like	17	(1)	39	(1)	2ucz	A
beta-hairpin-alpha-hairpin repeat	17	(1)	21	(1)	1wg0	A

Appendix C

Supplementary figures for Chapter 3

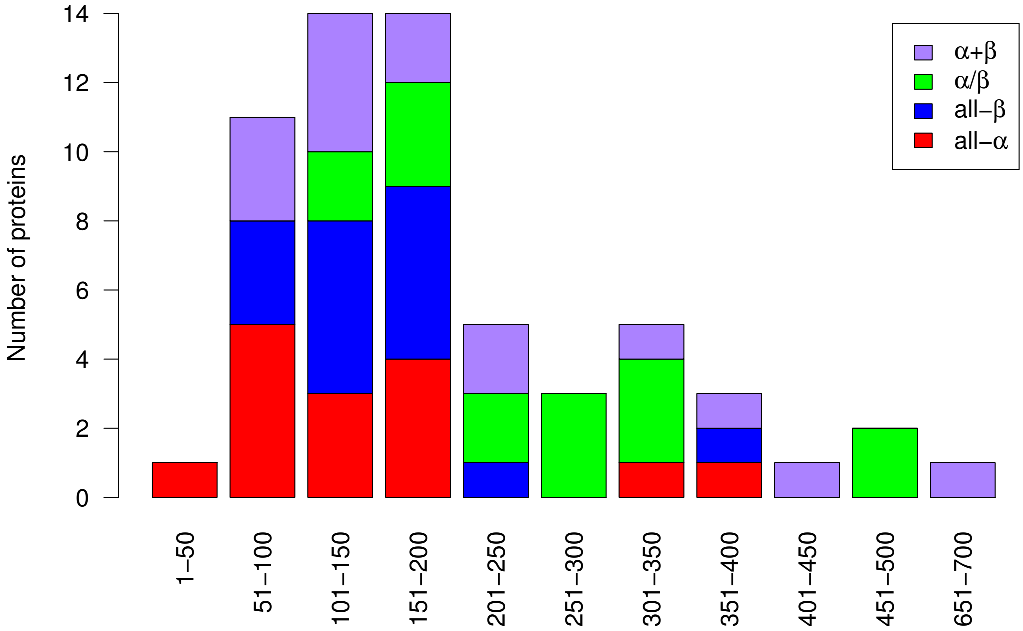


Figure C.1: Protein size distribution with respect to structural class.

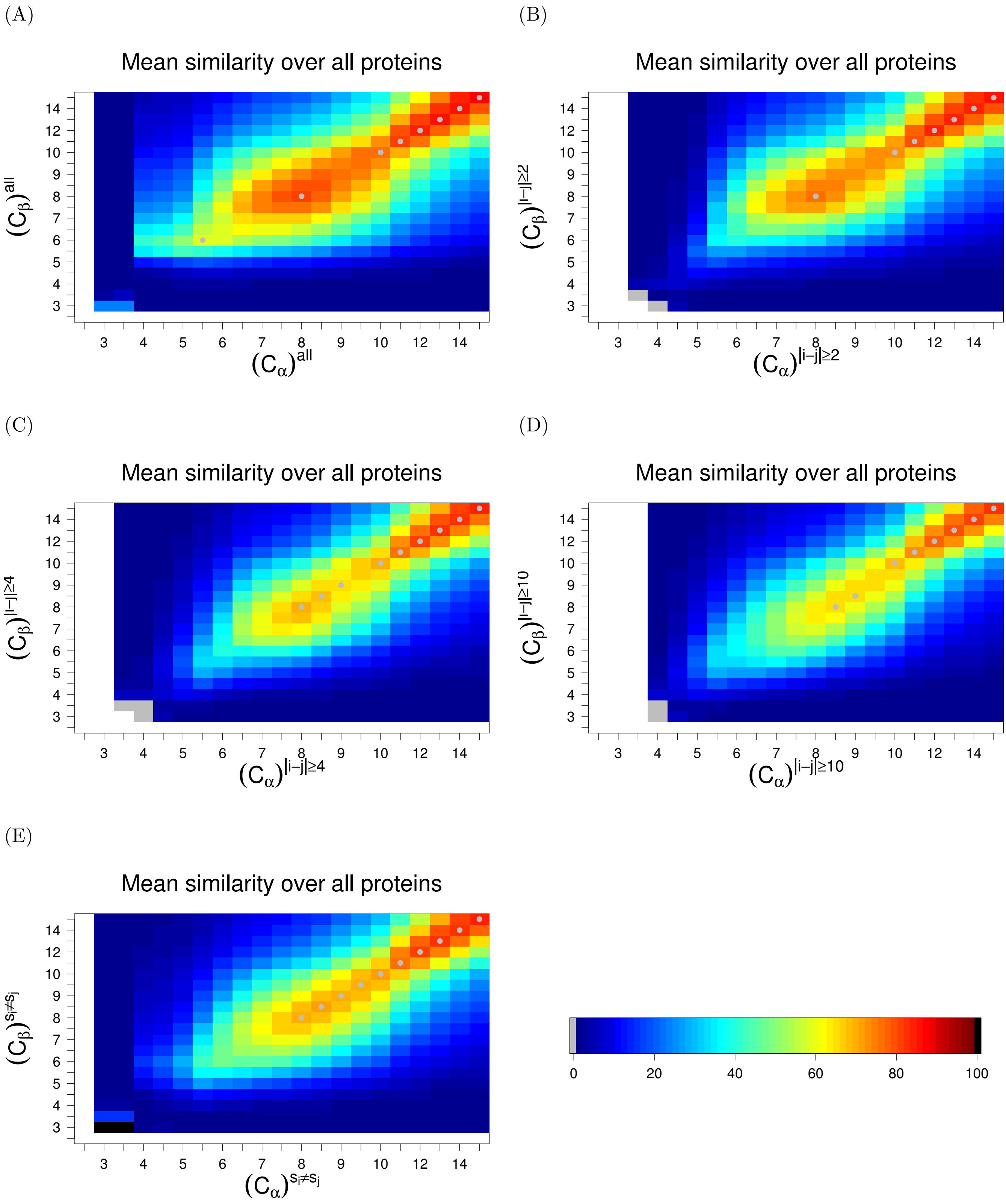


Figure C.2: The mean similarity matrices between C_α and C_β RIGs, over all proteins, for all distance cutoffs and for contact ranges: A. all, B. $|\mathrm{i}-\mathrm{j}| \geq 2$, C. $|\mathrm{i}-\mathrm{j}| \geq 4$, D. $|\mathrm{i}-\mathrm{j}| \geq 10$, E. $\mathrm{s_i} \neq \mathrm{s_j}$. Grey filled circles correspond to local maxima.

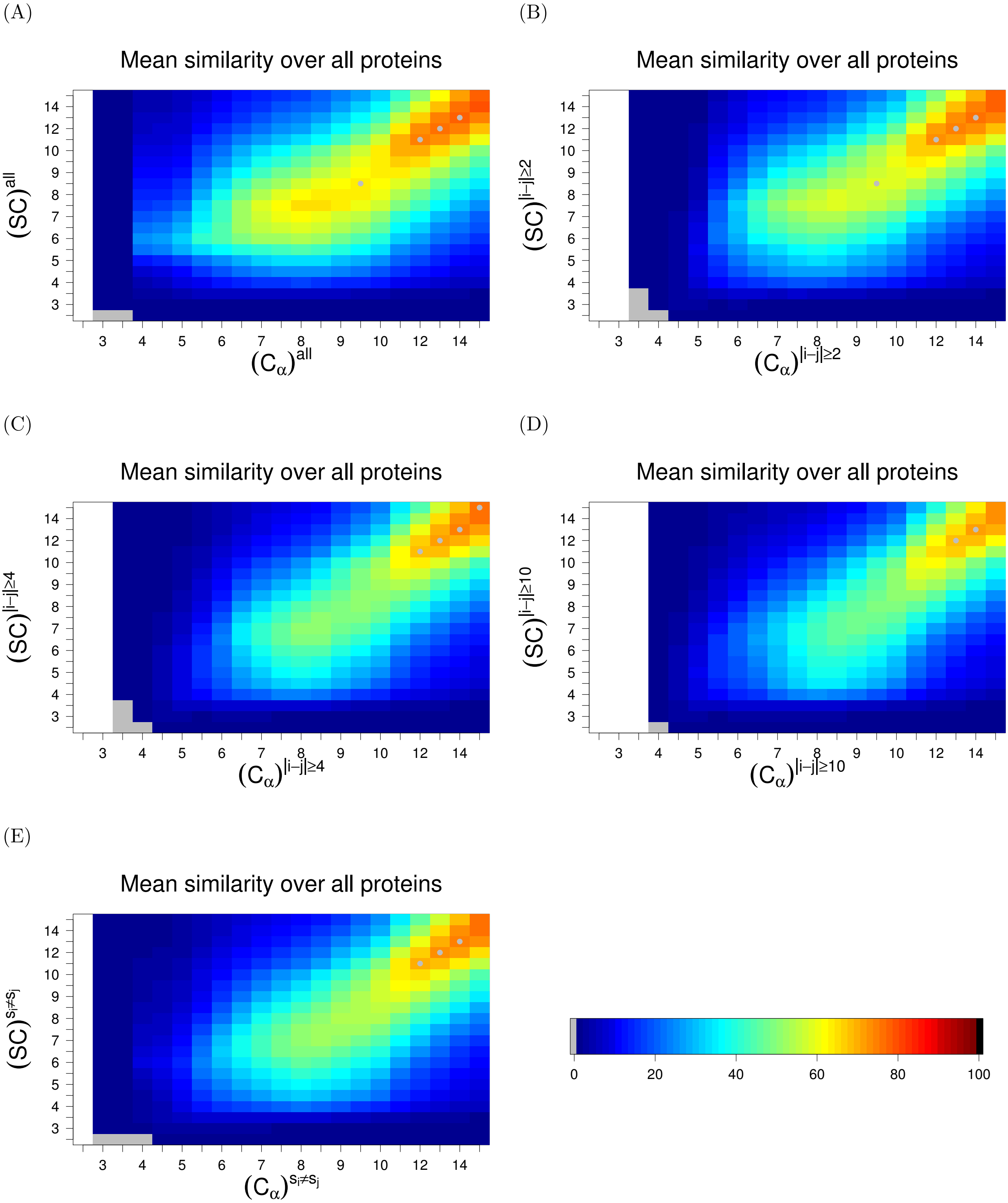


Figure C.3: The mean similarity matrices between C_α and SC RIGs, over all proteins, for all distance cutoffs and for contact ranges: A. all, B. $|\mathrm{i}-\mathrm{j}| \geq 2$, C. $|\mathrm{i}-\mathrm{j}| \geq 4$, D. $|\mathrm{i}-\mathrm{j}| \geq 10$, E. $\mathrm{s_i} \neq \mathrm{s_j}$. Grey filled circles correspond to local maxima.

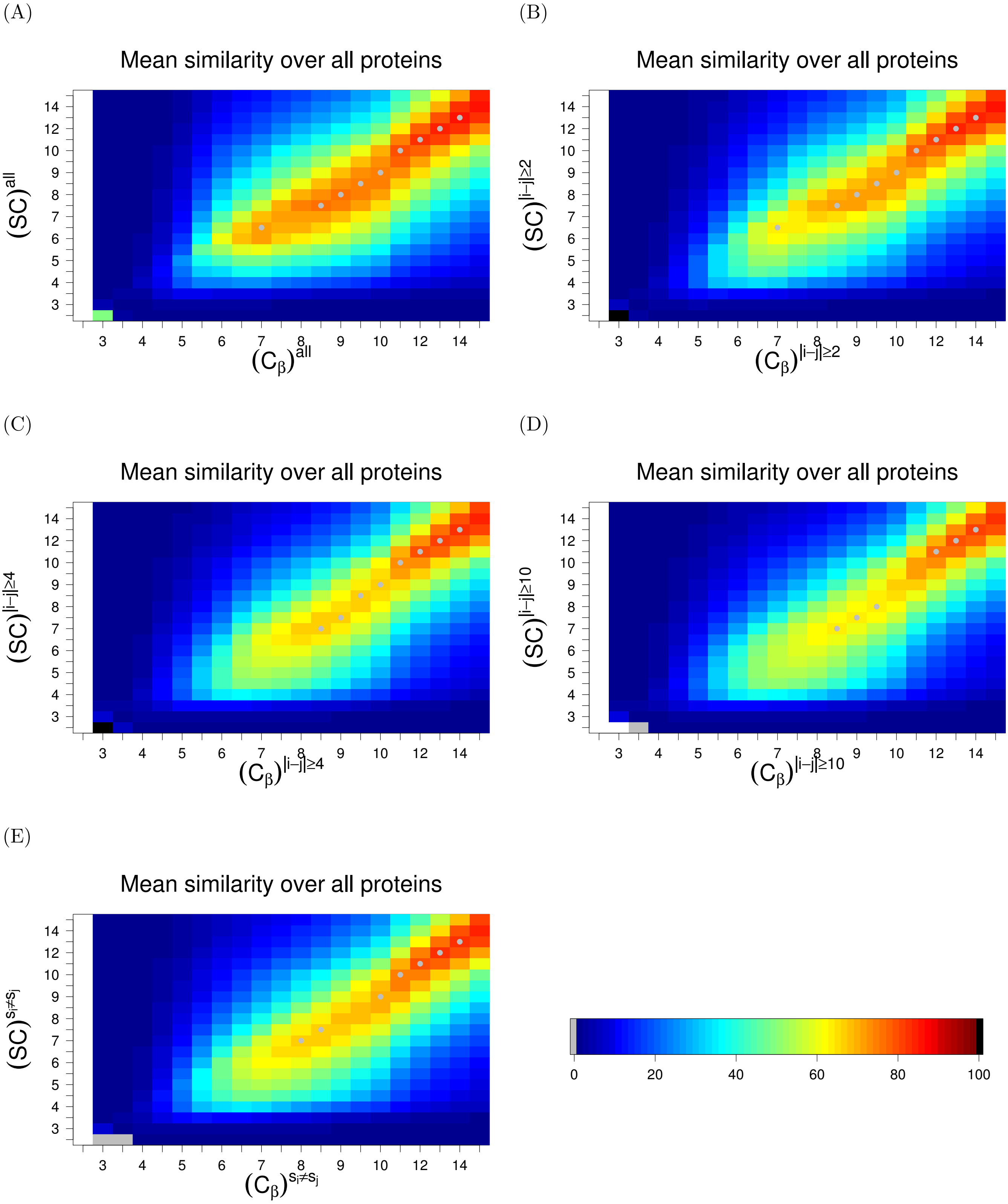


Figure C.4: The mean similarity matrices between C_β and SC RIGs, over all proteins, for all distance cutoffs and for contact ranges: A. all, B. $|i - j| \geq 2$, C. $|i - j| \geq 4$, D. $|i - j| \geq 10$, E. $s_i \neq s_j$. Grey filled circles correspond to local maxima.

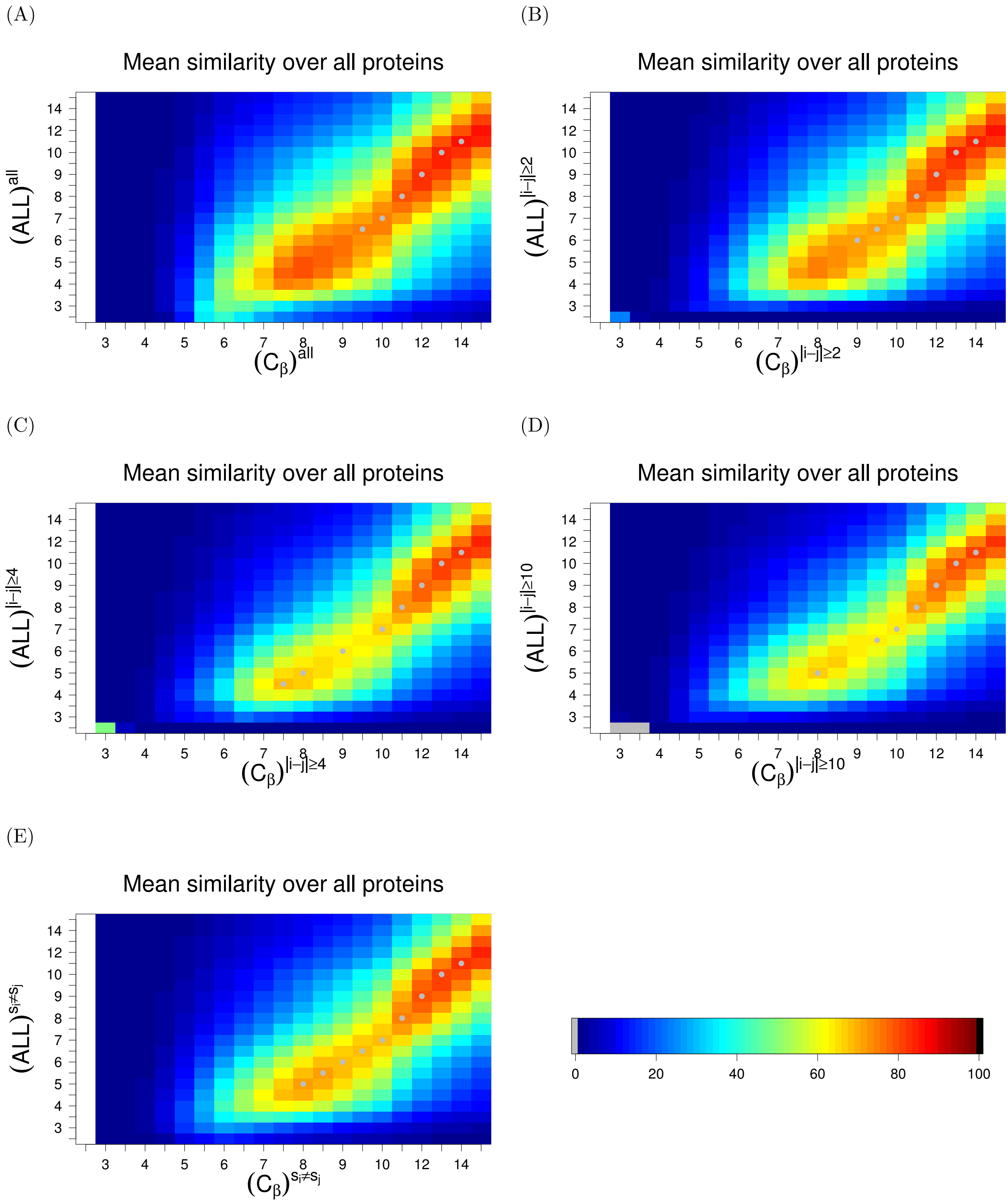


Figure C.5: The mean similarity matrices between C_β and ALL RIGs, over all proteins, for all distance cutoffs and for contact ranges: A. all, B. $|i-j| \geq 2$, C. $|i-j| \geq 4$, D. $|i-j| \geq 10$, E. $s_i \neq s_j$. Grey filled circles correspond to local maxima.

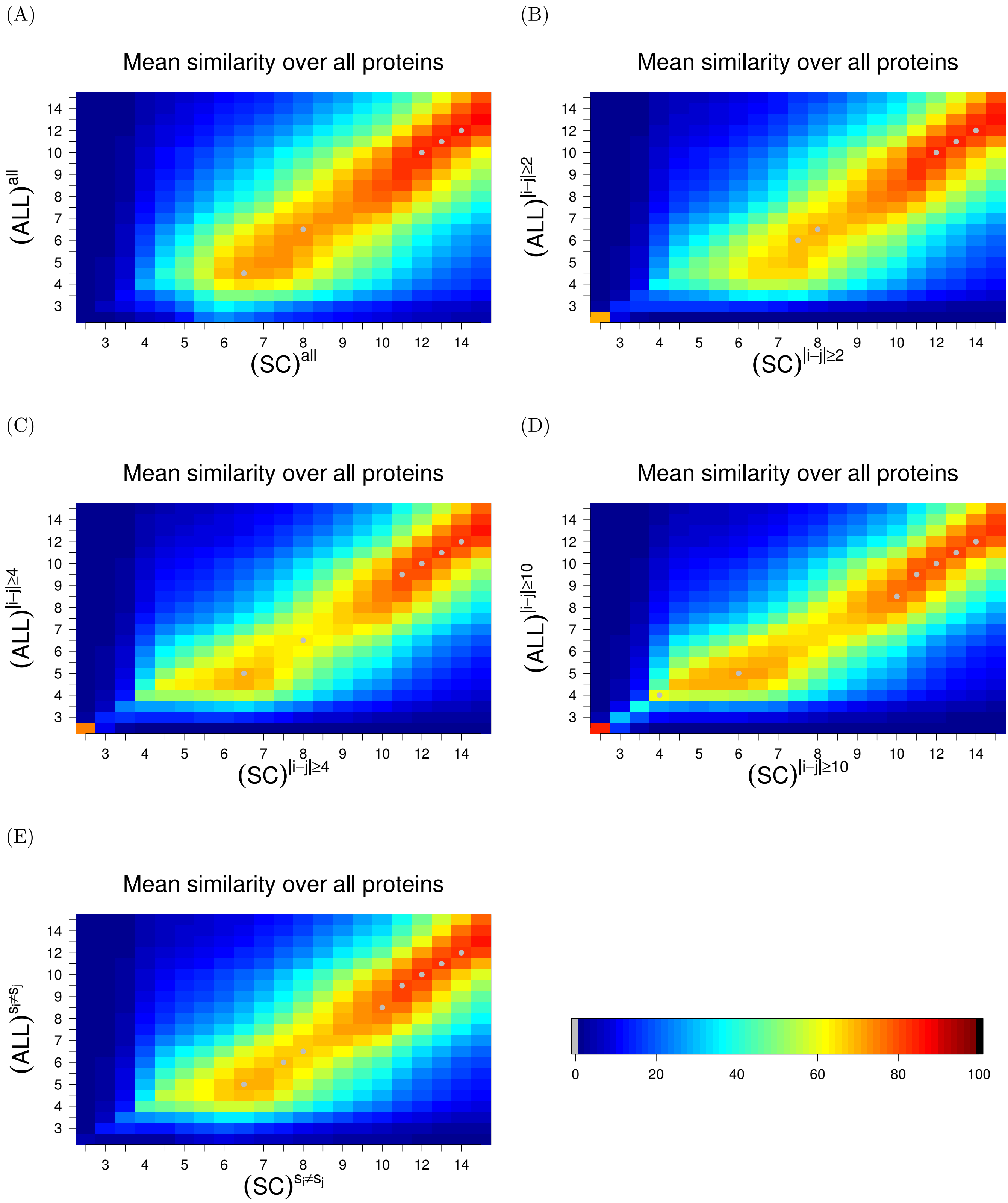


Figure C.6: The mean similarity matrices between *SC* and *ALL* RIGs, over all proteins, for all distance cutoffs and for contact ranges: A. all, B. $|i-j| \geq 2$, C. $|i-j| \geq 4$, D. $|i-j| \geq 10$, E. $s_i \neq s_j$. Grey filled circles correspond to local maxima.

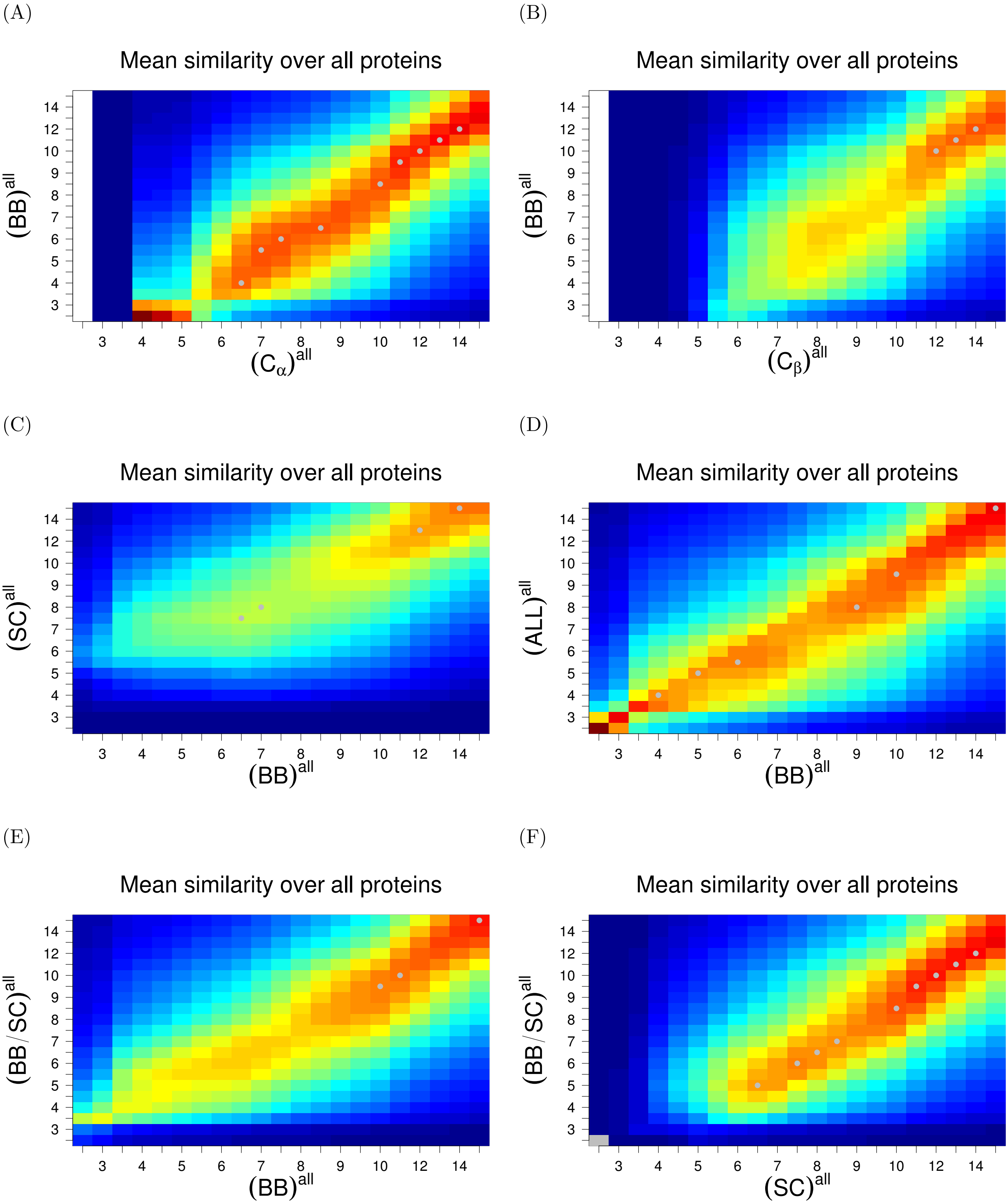


Figure C.7: The mean similarity matrices over all proteins, for all distance cutoffs and *all* contact range between: A. C_α and BB, B. C_β and BB, C. BB and SC, D. BB and ALL, E. BB and BB/SC, F. SC and BB/SC, G. C_α and C_α/C_β, H. C_α and $C_{\alpha\beta}$, I. C_β and C_α/C_β, J. C_β and $C_{\alpha\beta}$, K. C_α and BB/SC, L. C_β and BB/SC. Grey filled circles correspond to local maxima. (continued on next page)

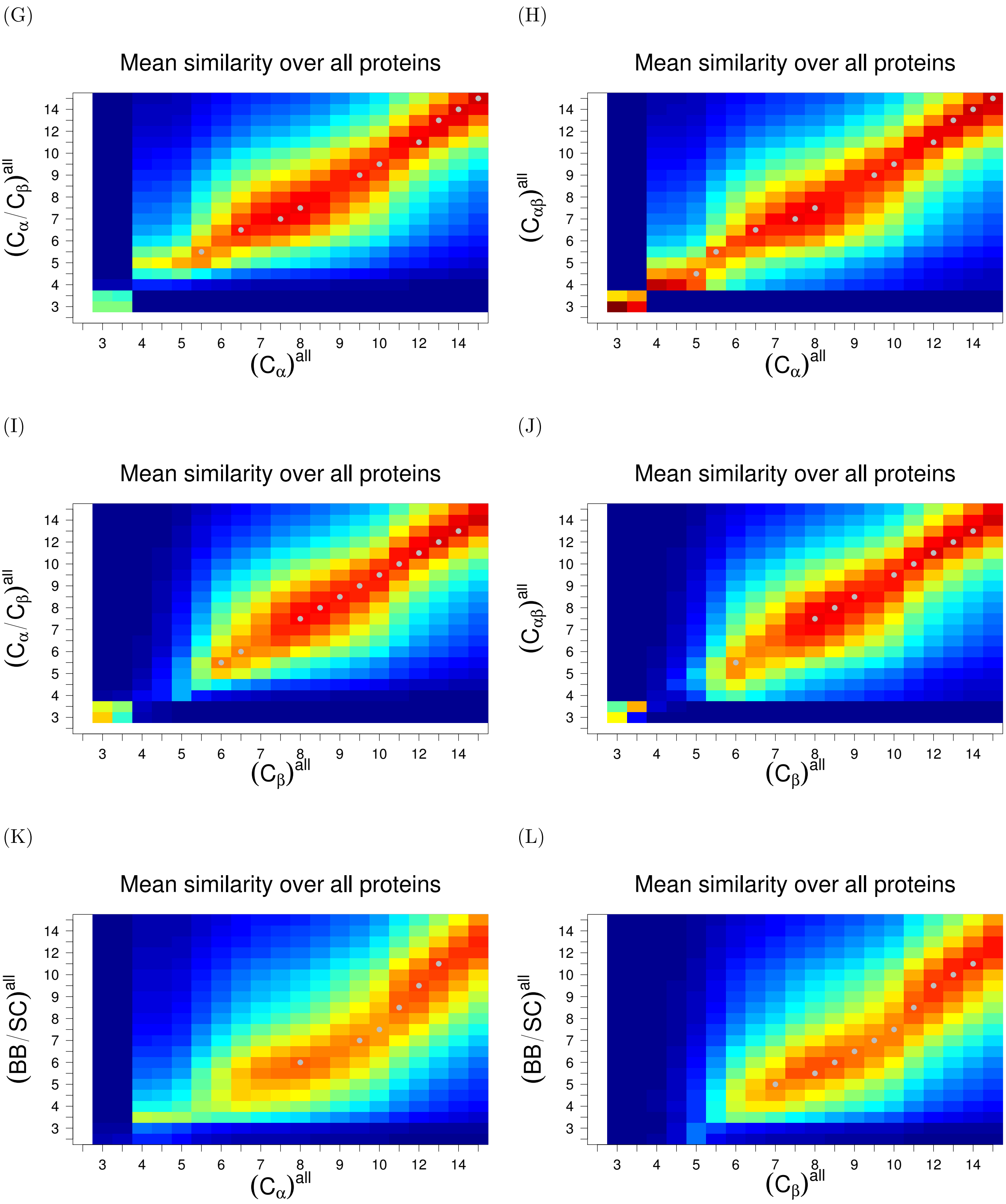
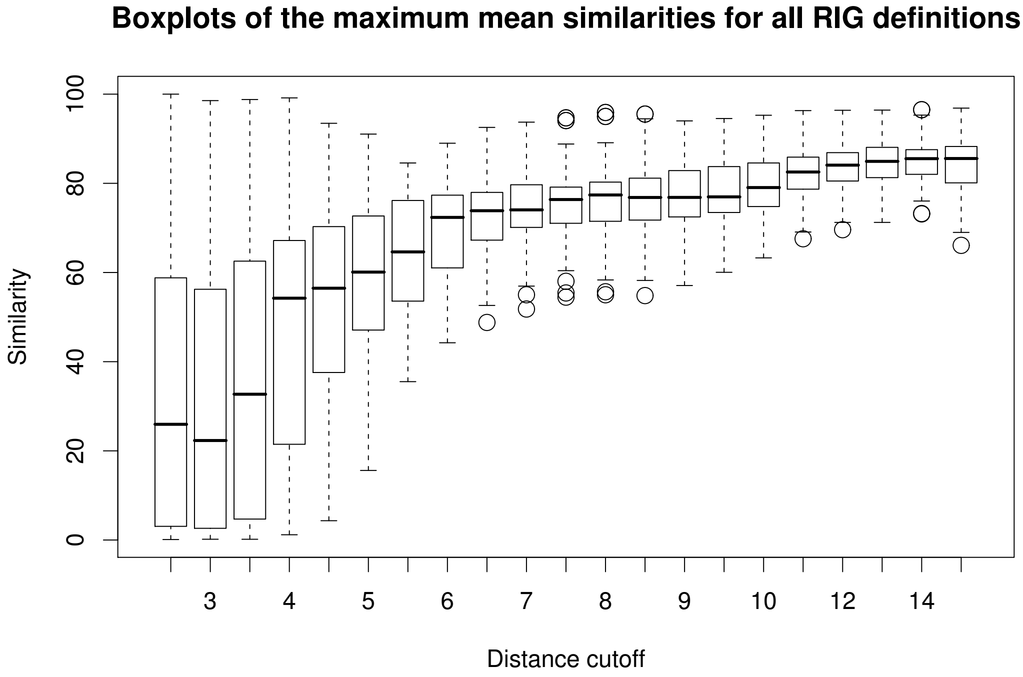


Figure C.7: continued from previous page

(A)



(B)

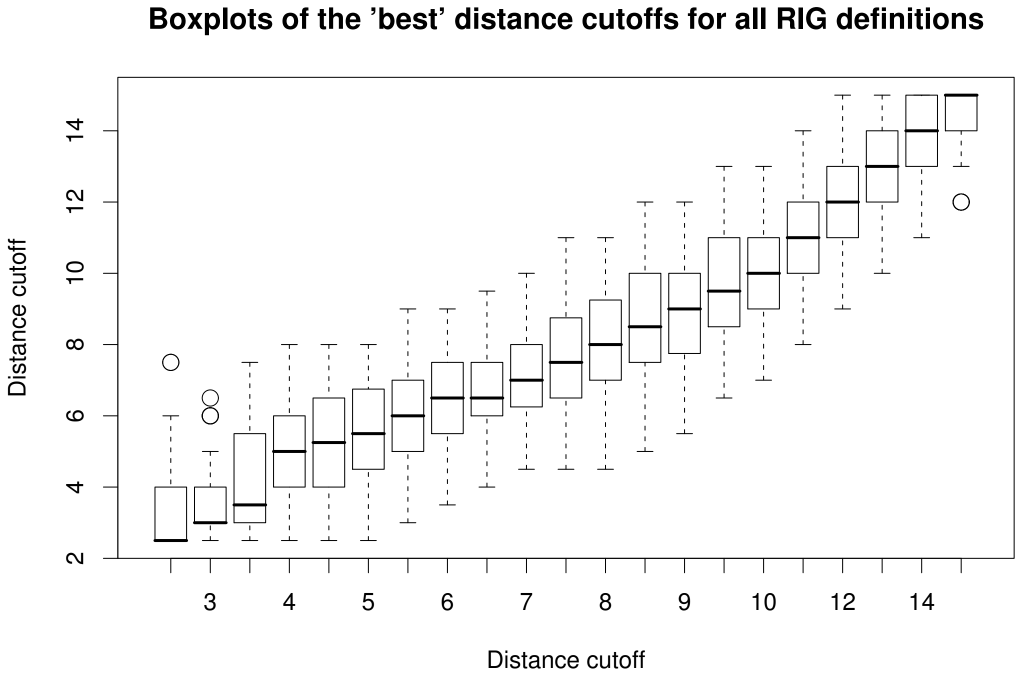
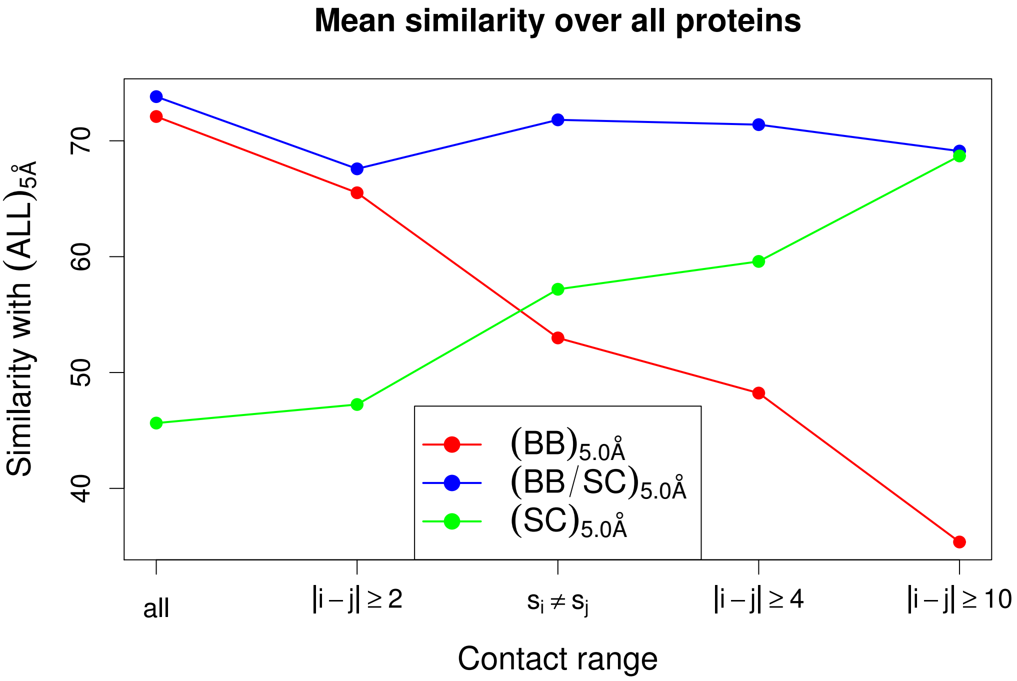


Figure C.8: Boxplots of the maximum mean similarities (A) and the corresponding distance cutoffs (B) over all RIGs for all distance cutoffs.

(A)



(B)

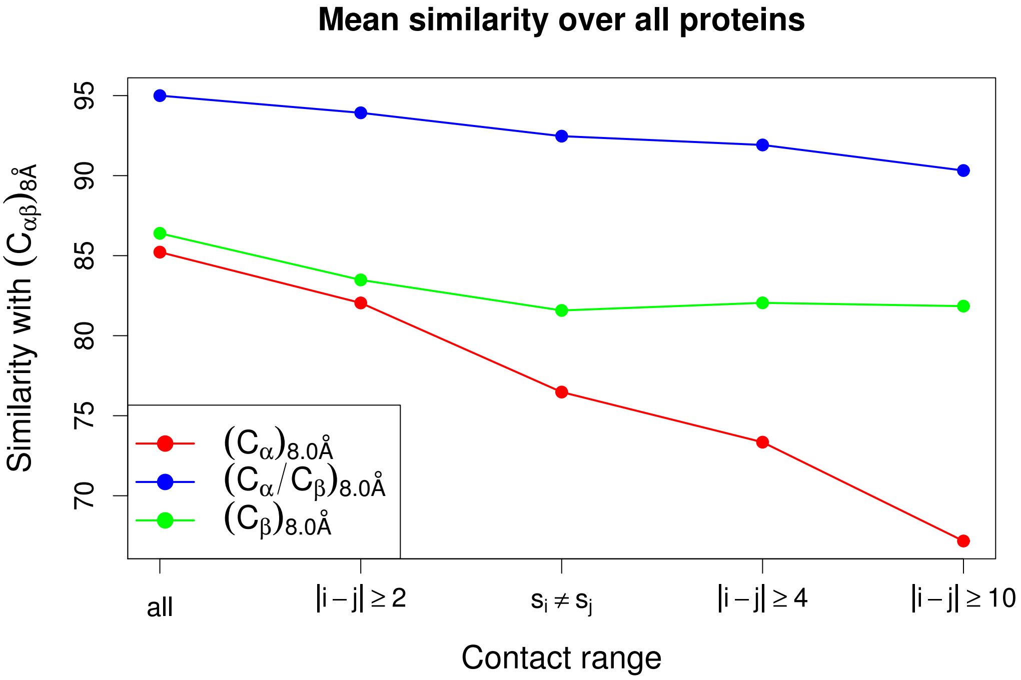


Figure C.9: A. The mean similarities of BB, SC and BB/SC RIGs with ALL ones, over all proteins, for 5.0Å distance cutoff and for all contact ranges. B. The mean similarities of C_α, C_β and C_α/C_β RIGs with $C_{\alpha\beta}$ ones, over all proteins, for 8.0Å distance cutoff and for all contact ranges.

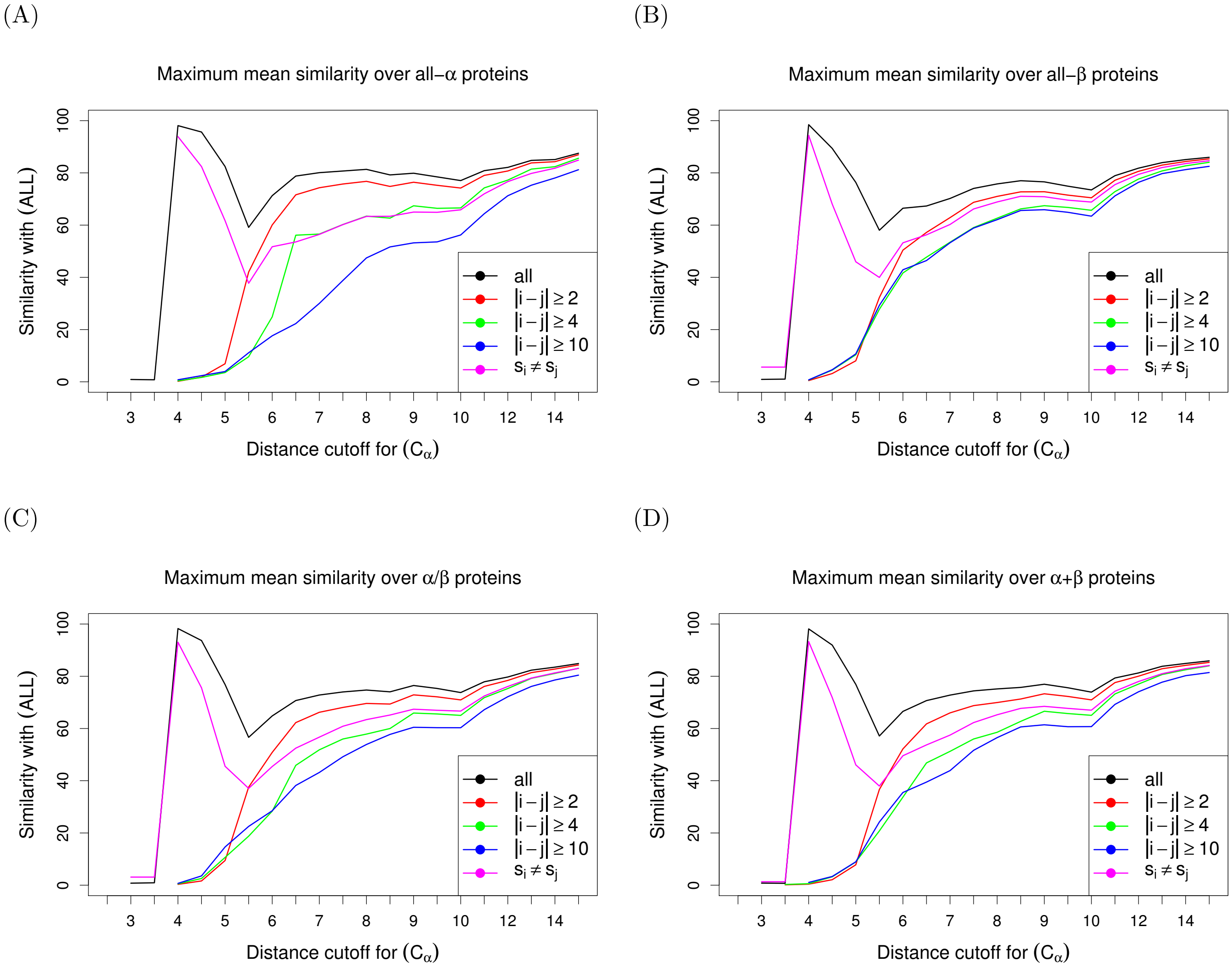


Figure C.10: The maximum mean similarities of C_α RIGs for each distance cutoff with respect to *ALL* RIGs of any cutoff, over all proteins, for all contact ranges and for all structural classes.

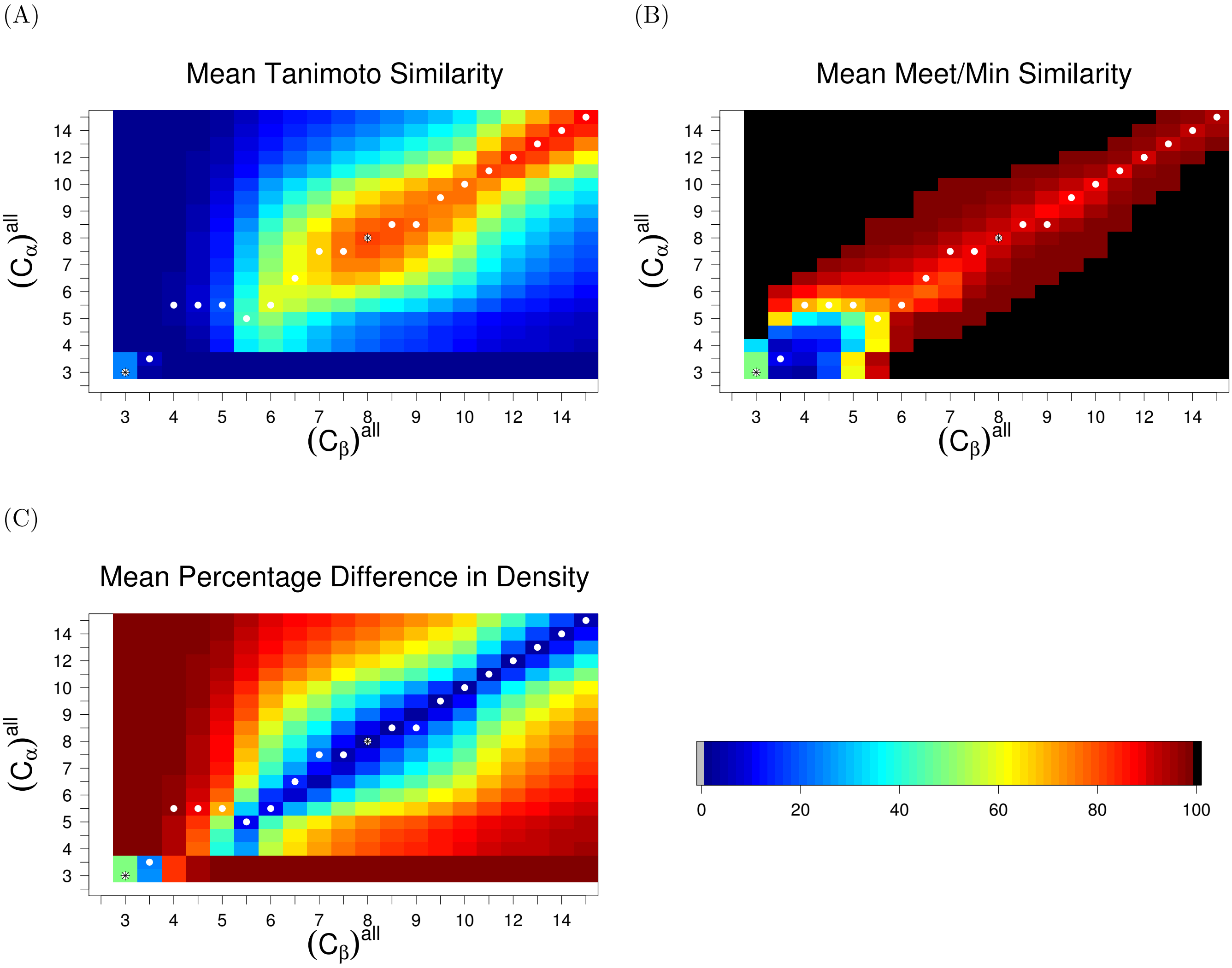


Figure C.11: The mean Tanimoto (A) and Meet/Min (B) similarity matrices, and the mean percentage difference in density (C) between C_β and C_α RIGs, over all proteins, for all distance cutoffs and for *all* contact range. White filled circles denote the position of the maximum mean Tanimoto similarity per column. Circles with overlapping black stars correspond to the local maxima of the maximum mean Tanimoto similarities.

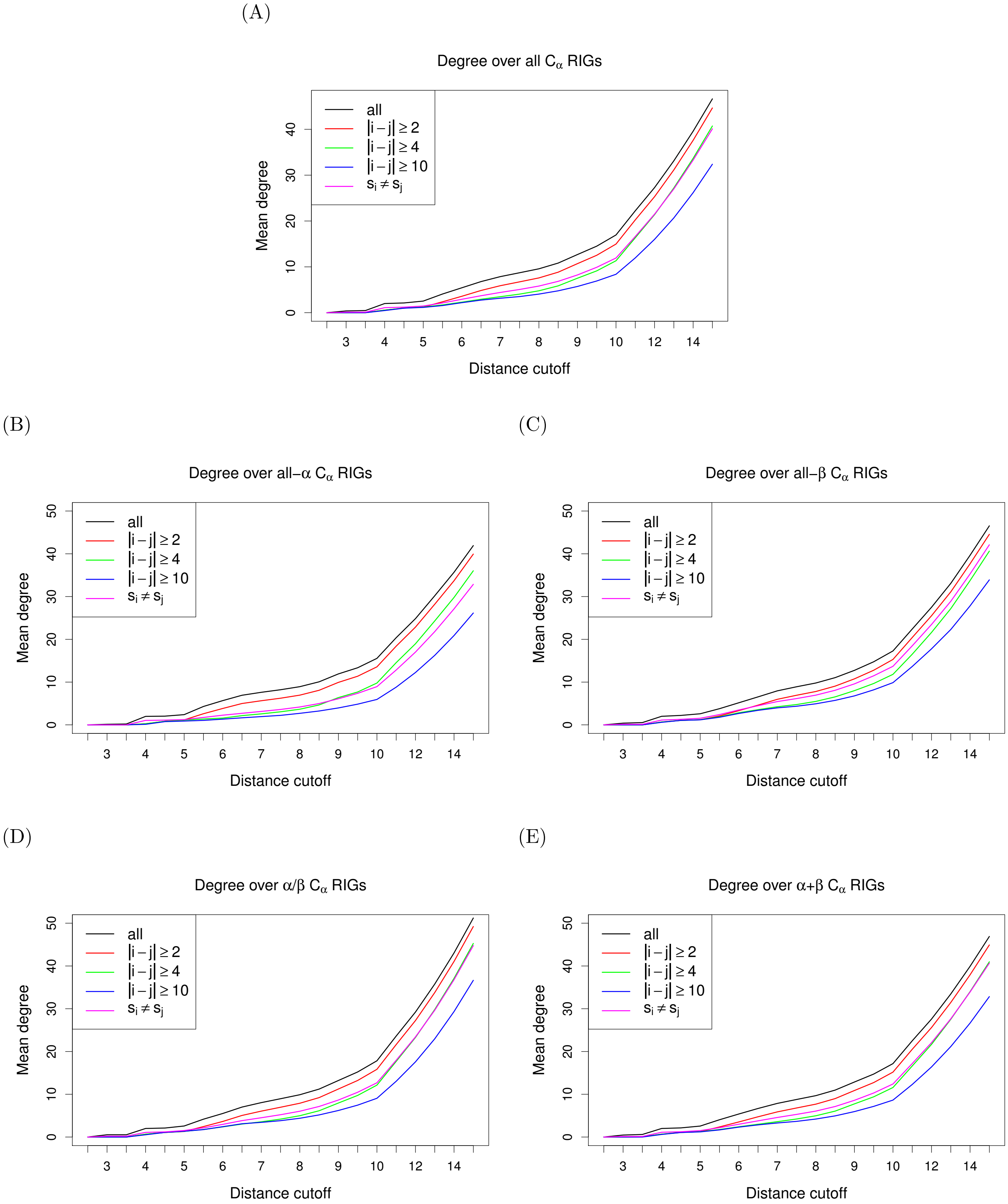


Figure C.12: The mean degree of C_α RIGs for all distance cutoffs and contact ranges, over all proteins, with respect to all structural classes.

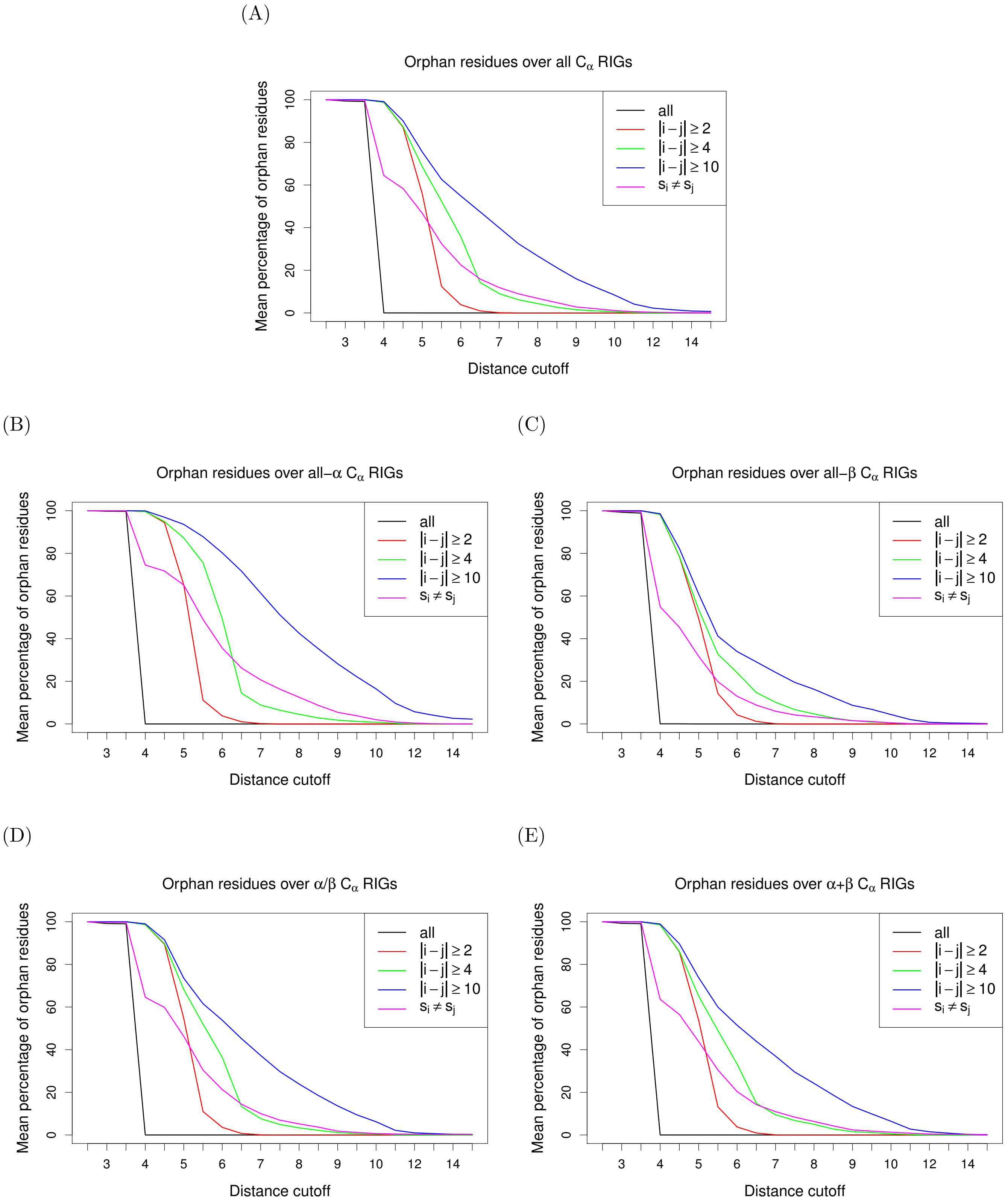


Figure C.13: The mean percentage of orphan residues of C_α RIGs for all distance cutoffs and contact ranges, over all proteins, with respect to all structural classes.

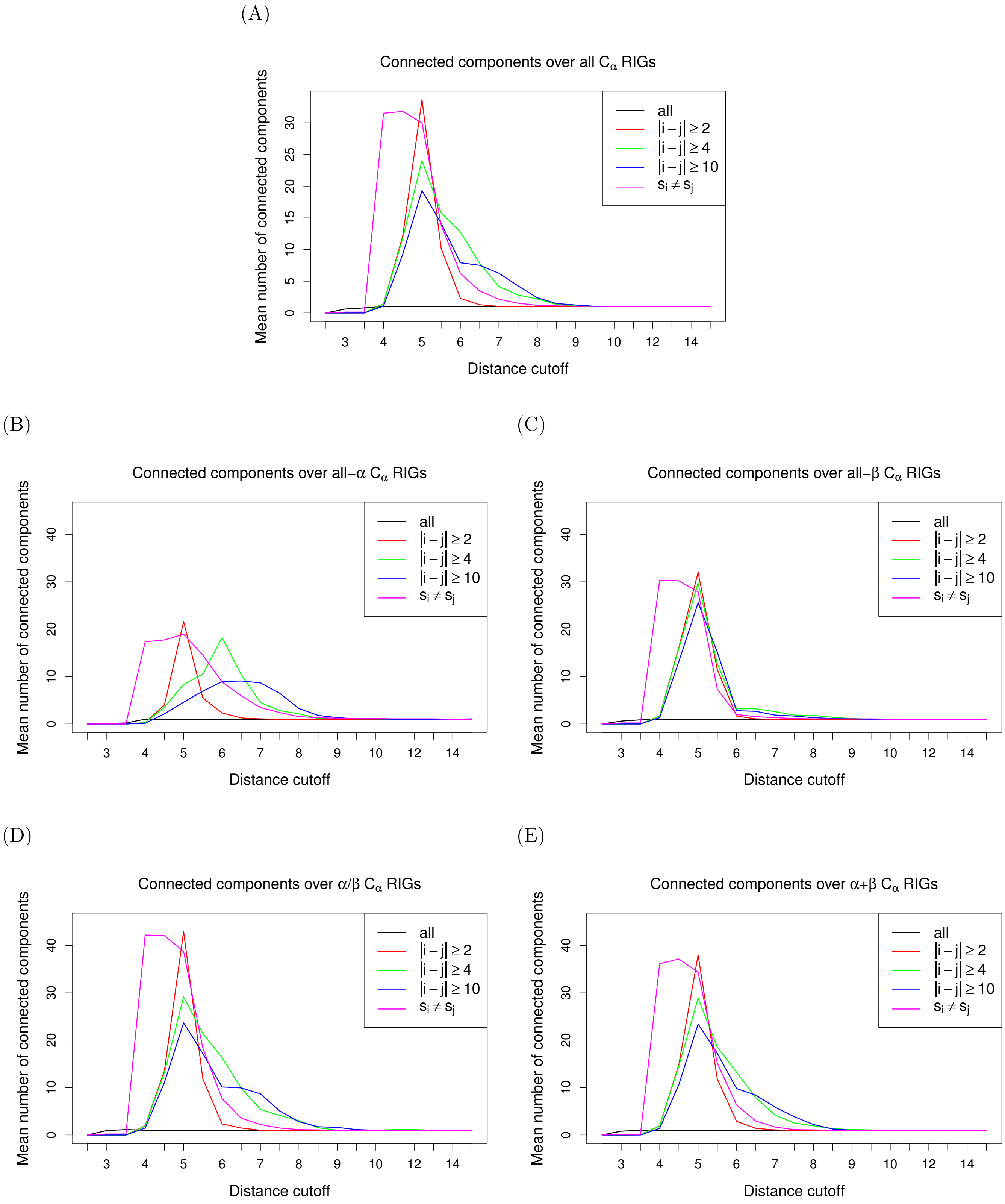


Figure C.14: The mean number of connected components of C_α RIGs for all distance cutoffs and contact ranges, over all proteins, with respect to all structural classes.

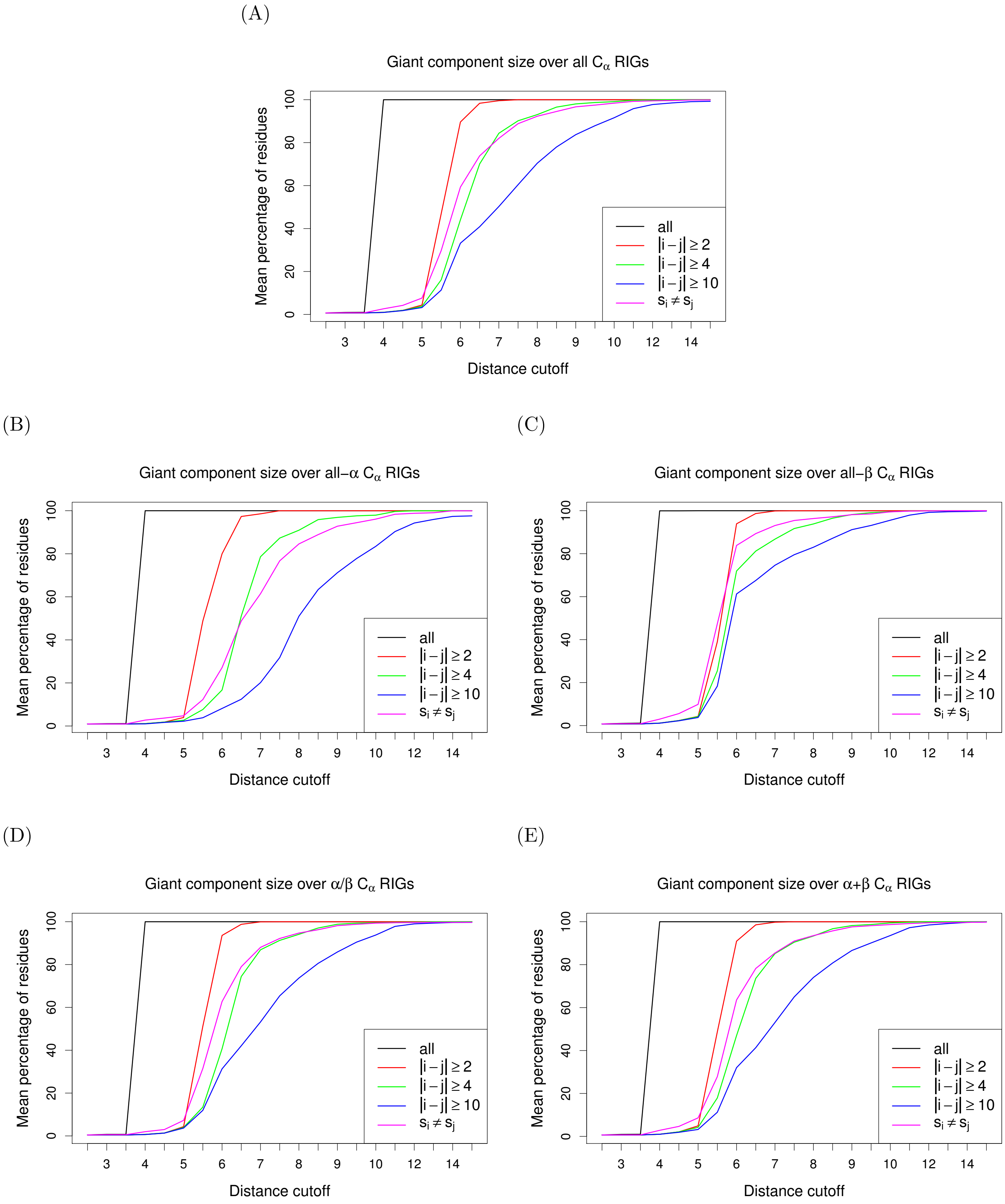
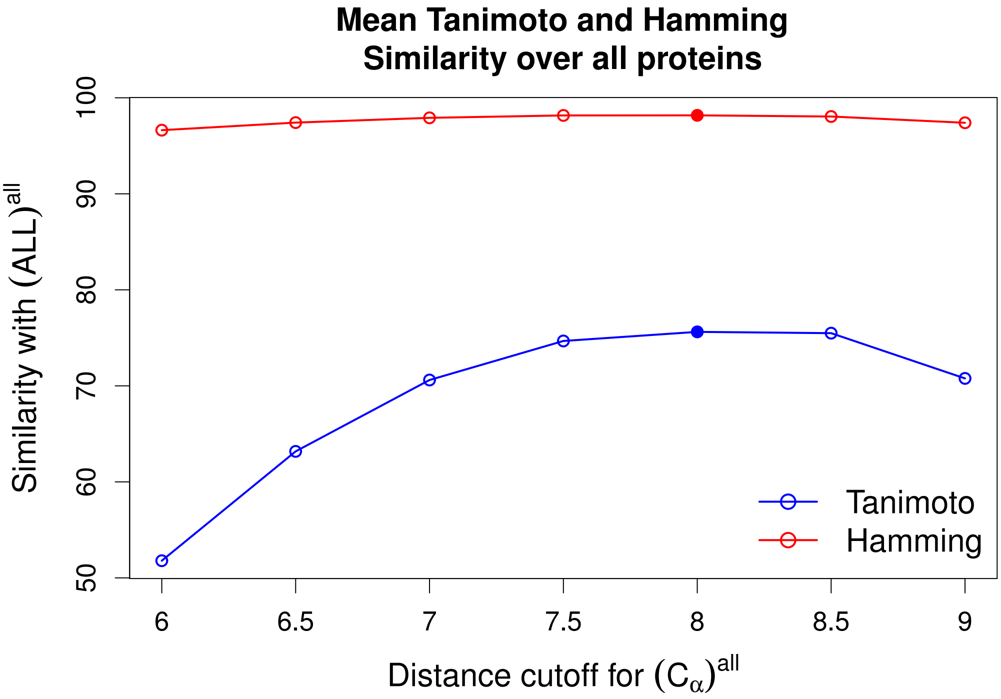


Figure C.15: The mean giant component size of C_α RIGs for all distance cutoffs and contact ranges, over all proteins, with respect to all structural classes.

(A)



(B)

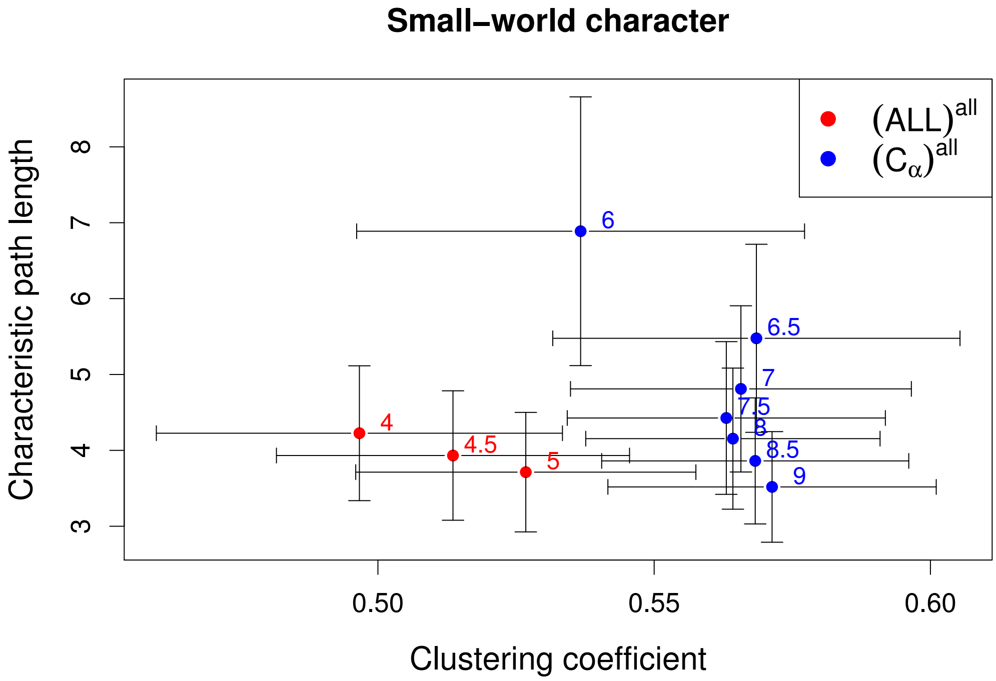


Figure C.16: Comparison of $(C_\alpha)^{all}$ with $(ALL)^{all}$ RIGs with respect to their similarity (A) and their small-world character (B) for a certain range of distance cutoffs. A. Mean Tanimoto and Hamming similarity between $(ALL)^{all}_{5.0\mathring{A}}$ RIGs and $(C_\alpha)^{all}$ RIGs with distance cutoff in the range [6.0, 9.0]Å over all proteins. B. Mean characteristic path length and clustering coefficient for $(C_\alpha)^{all}$ RIGs with distance cutoff in range [6.0, 9.0]Å and $(ALL)^{all}$ RIGs with distance cutoff in range [4.0, 5.0]Å, over all proteins. Error bars represent the standard deviations.

Appendix D

Supplementary figures for Chapter 4

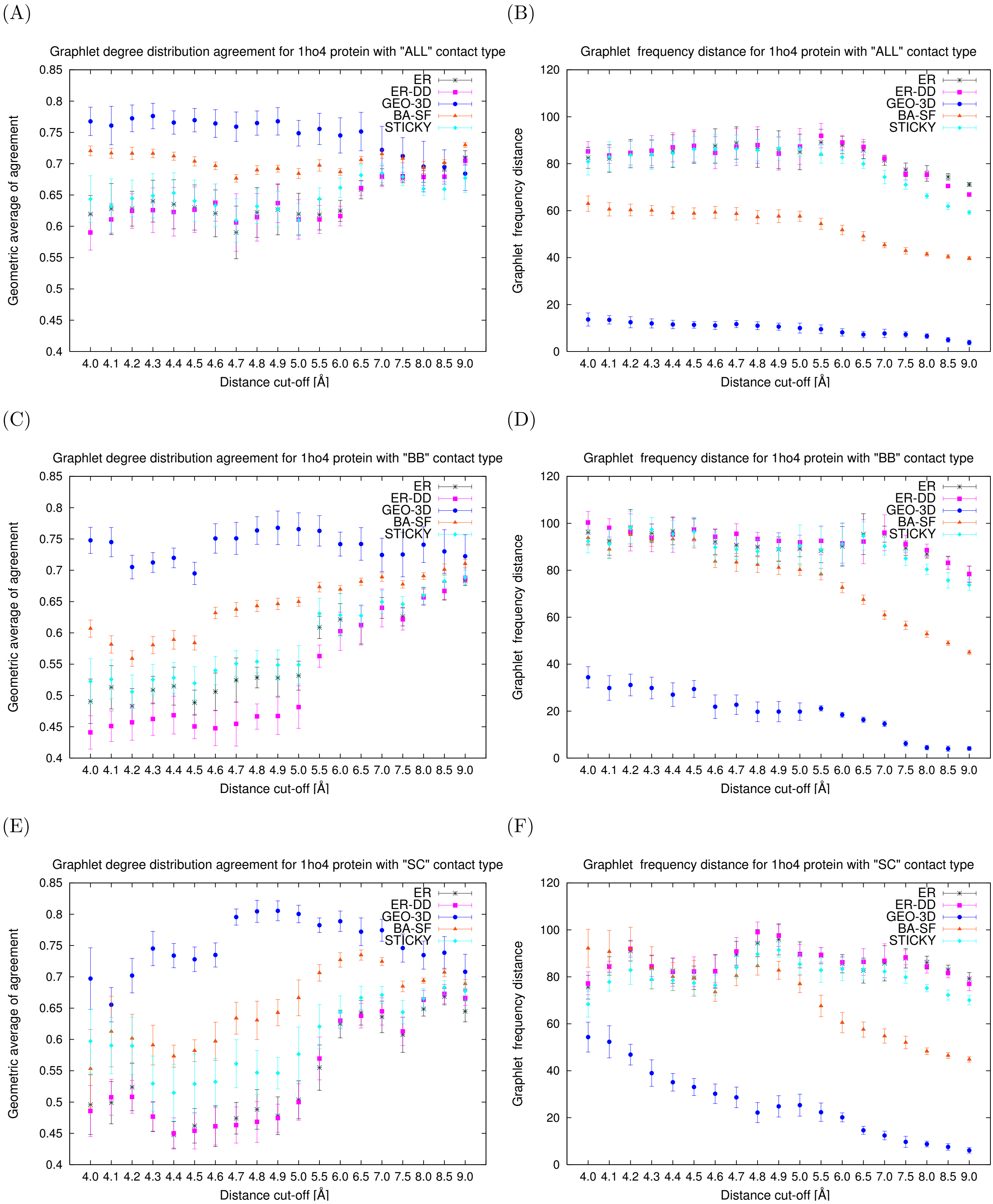


Figure D.1: GDD-agreements and RGF-distances between model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1ho4 protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** GDD-agreement for *ALL* contact type. **B.** RGF-distance for *ALL* contact type. **C.** GDD-agreement for *BB* contact type. **D.** RGF-distance for *BB* contact type. **E.** GDD-agreement for *SC* contact type. **F.** RGF-distance for *SC* contact type. The larger the GDD-agreement in panels A, C, and E the better the fit. The smaller the RGF-distance in panels B, D, and F the better the fit.

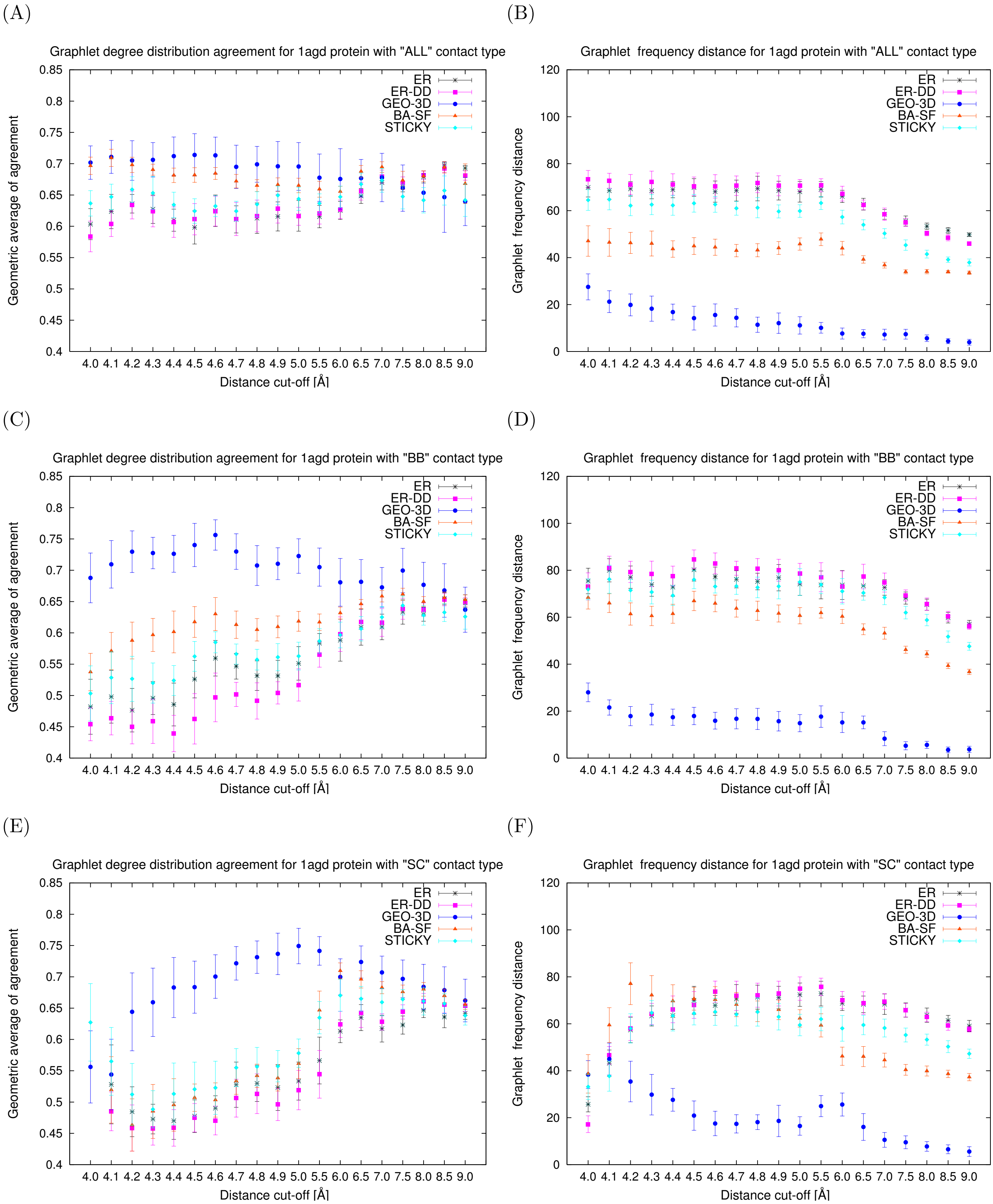


Figure D.2: GDD-agreements and RGF-distances between model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1agd protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** GDD-agreement for *ALL* contact type. **B.** RGF-distance for *ALL* contact type. **C.** GDD-agreement for *BB* contact type. **D.** RGF-distance for *BB* contact type. **E.** GDD-agreement for *SC* contact type. **F.** RGF-distance for *SC* contact type. The larger the GDD-agreement in panels A, C, and E the better the fit. The smaller the RGF-distance in panels B, D, and F the better the fit.

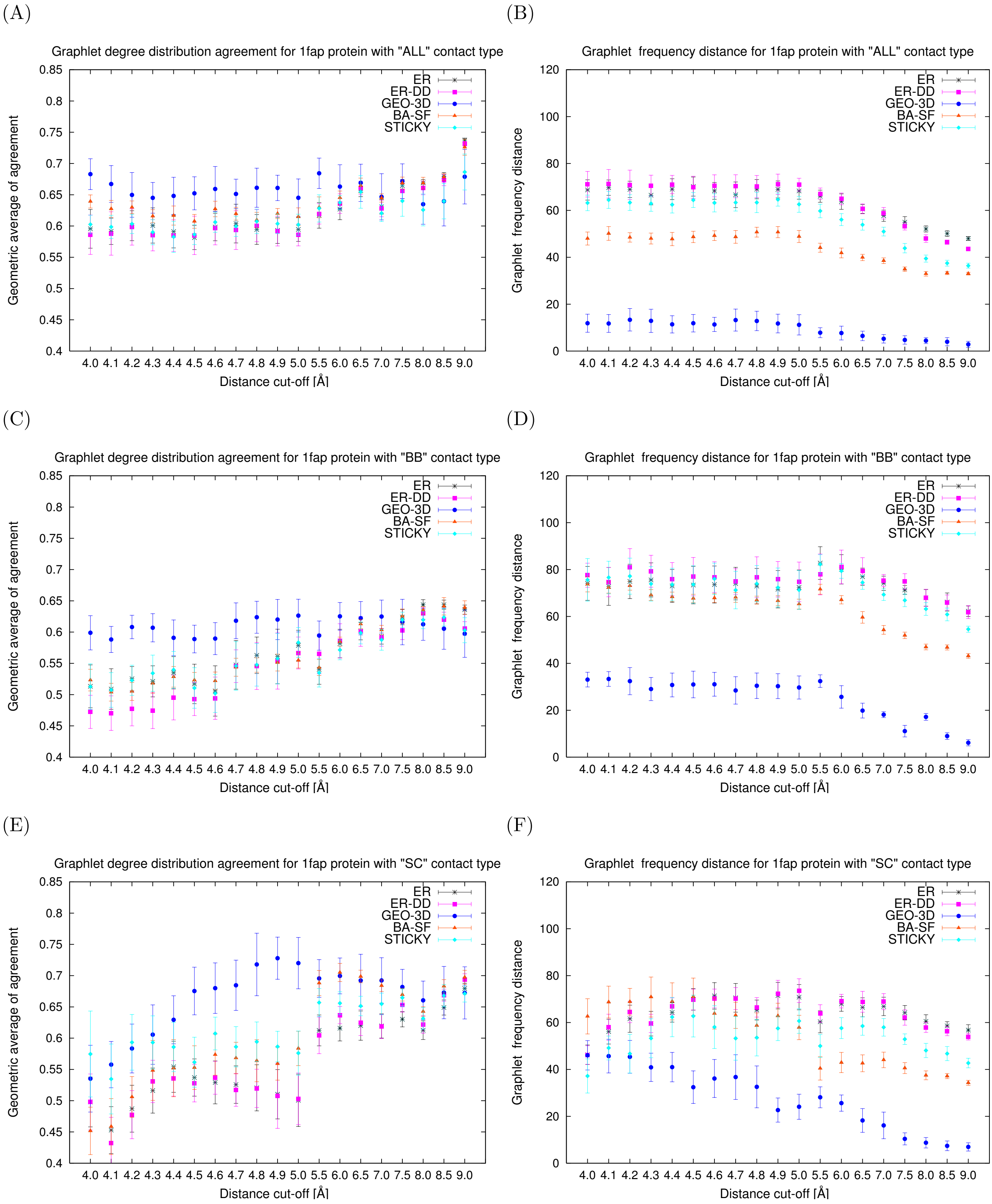


Figure D.3: GDD-agreements and RGF-distances between model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1fap protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** GDD-agreement for *ALL* contact type. **B.** RGF-distance for *ALL* contact type. **C.** GDD-agreement for *BB* contact type. **D.** RGF-distance for *BB* contact type. **E.** GDD-agreement for *SC* contact type. **F.** RGF-distance for *SC* contact type. The larger the GDD-agreement in panels A, C, and E the better the fit. The smaller the RGF-distance in panels B, D, and F the better the fit.

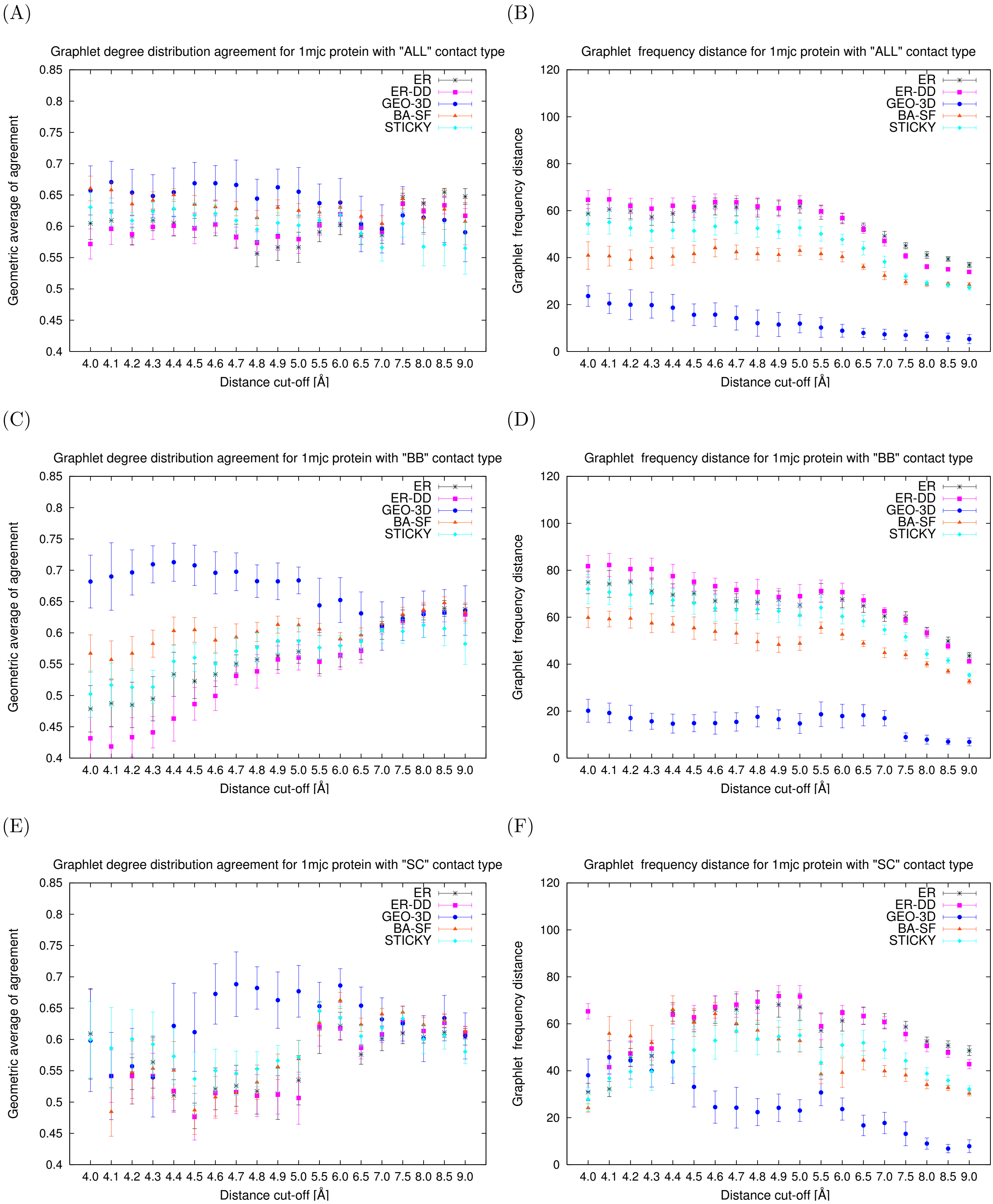


Figure D.4: GDD-agreements and RGF-distances between model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1mjc protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** GDD-agreement for *ALL* contact type. **B.** RGF-distance for *ALL* contact type. **C.** GDD-agreement for *BB* contact type. **D.** RGF-distance for *BB* contact type. **E.** GDD-agreement for *SC* contact type. **F.** RGF-distance for *SC* contact type. The larger the GDD-agreement in panels A, C, and E the better the fit. The smaller the RGF-distance in panels B, D, and F the better the fit.

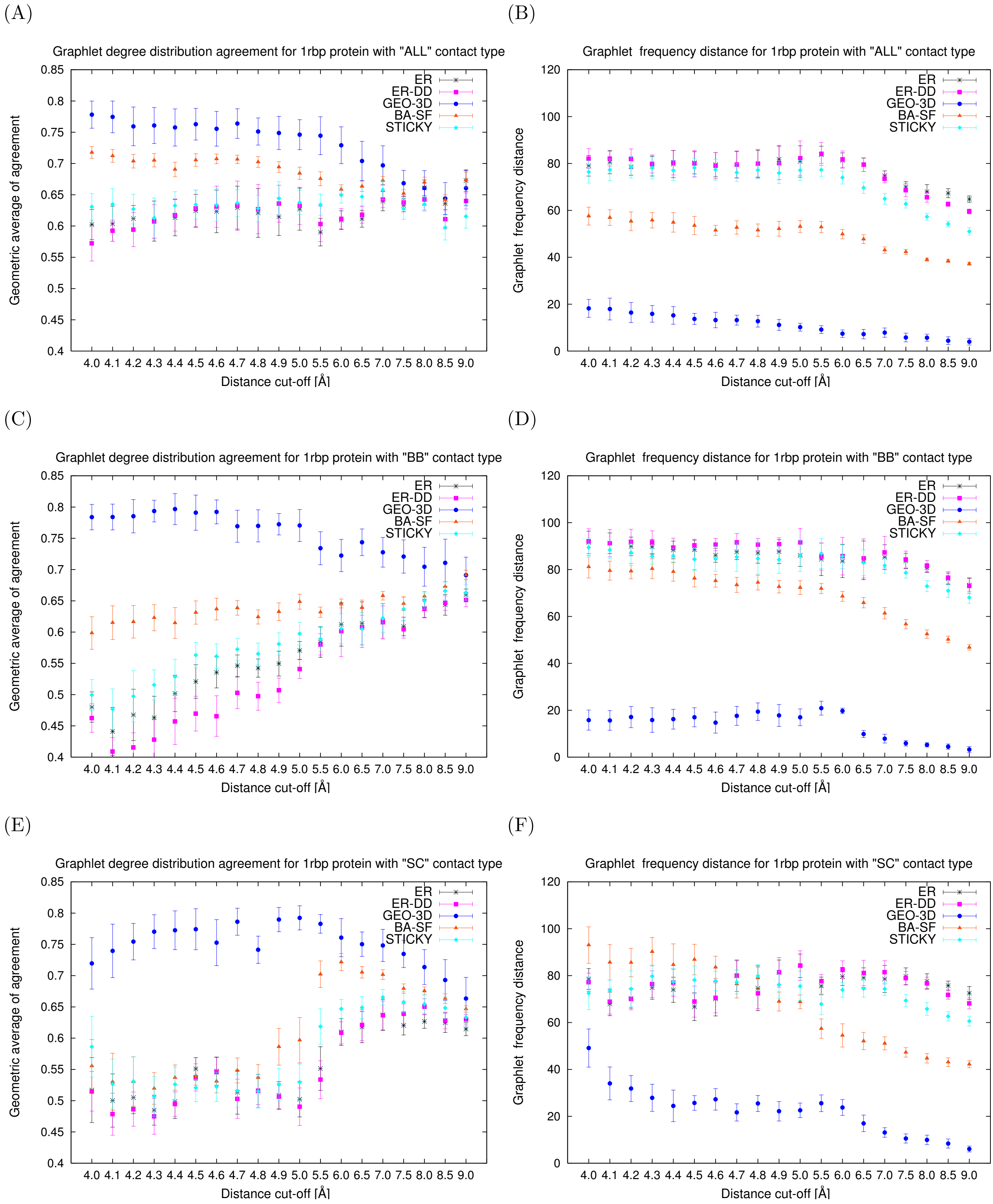


Figure D.5: GDD-agreements and RGF-distances between model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1rbp protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** GDD-agreement for *ALL* contact type. **B.** RGF-distance for *ALL* contact type. **C.** GDD-agreement for *BB* contact type. **D.** RGF-distance for *BB* contact type. **E.** GDD-agreement for *SC* contact type. **F.** RGF-distance for *SC* contact type. The larger the GDD-agreement in panels A, C, and E the better the fit. The smaller the RGF-distance in panels B, D, and F the better the fit.

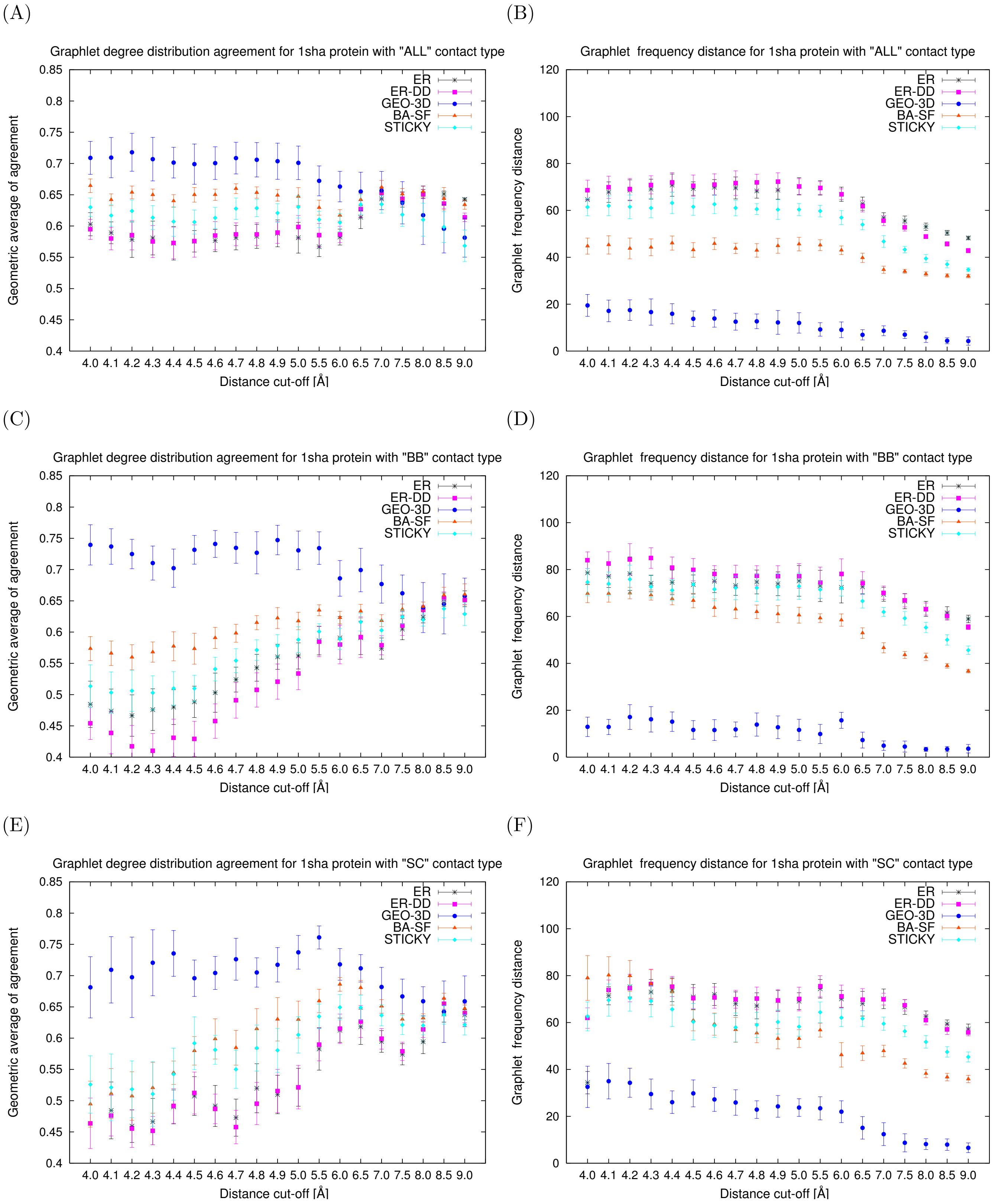


Figure D.6: GDD-agreements and RGF-distances between model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1sha protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** GDD-agreement for *ALL* contact type. **B.** RGF-distance for *ALL* contact type. **C.** GDD-agreement for *BB* contact type. **D.** RGF-distance for *BB* contact type. **E.** GDD-agreement for *SC* contact type. **F.** RGF-distance for *SC* contact type. The larger the GDD-agreement in panels A, C, and E the better the fit. The smaller the RGF-distance in panels B, D, and F the better the fit.

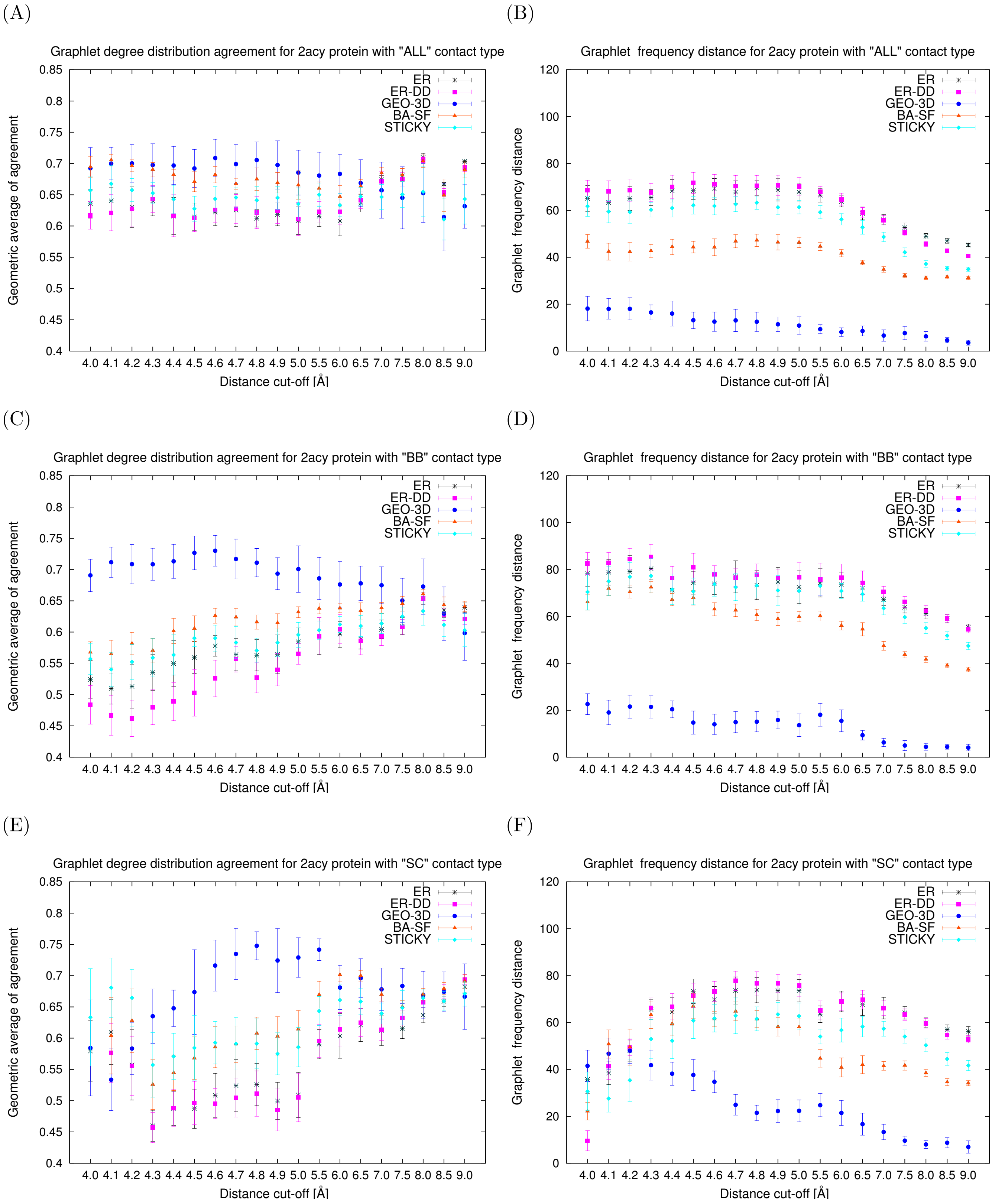


Figure D.7: GDD-agreements and RGF-distances between model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 2acy protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** GDD-agreement for *ALL* contact type. **B.** RGF-distance for *ALL* contact type. **C.** GDD-agreement for *BB* contact type. **D.** RGF-distance for *BB* contact type. **E.** GDD-agreement for *SC* contact type. **F.** RGF-distance for *SC* contact type. The larger the GDD-agreement in panels A, C, and E the better the fit. The smaller the RGF-distance in panels B, D, and F the better the fit.

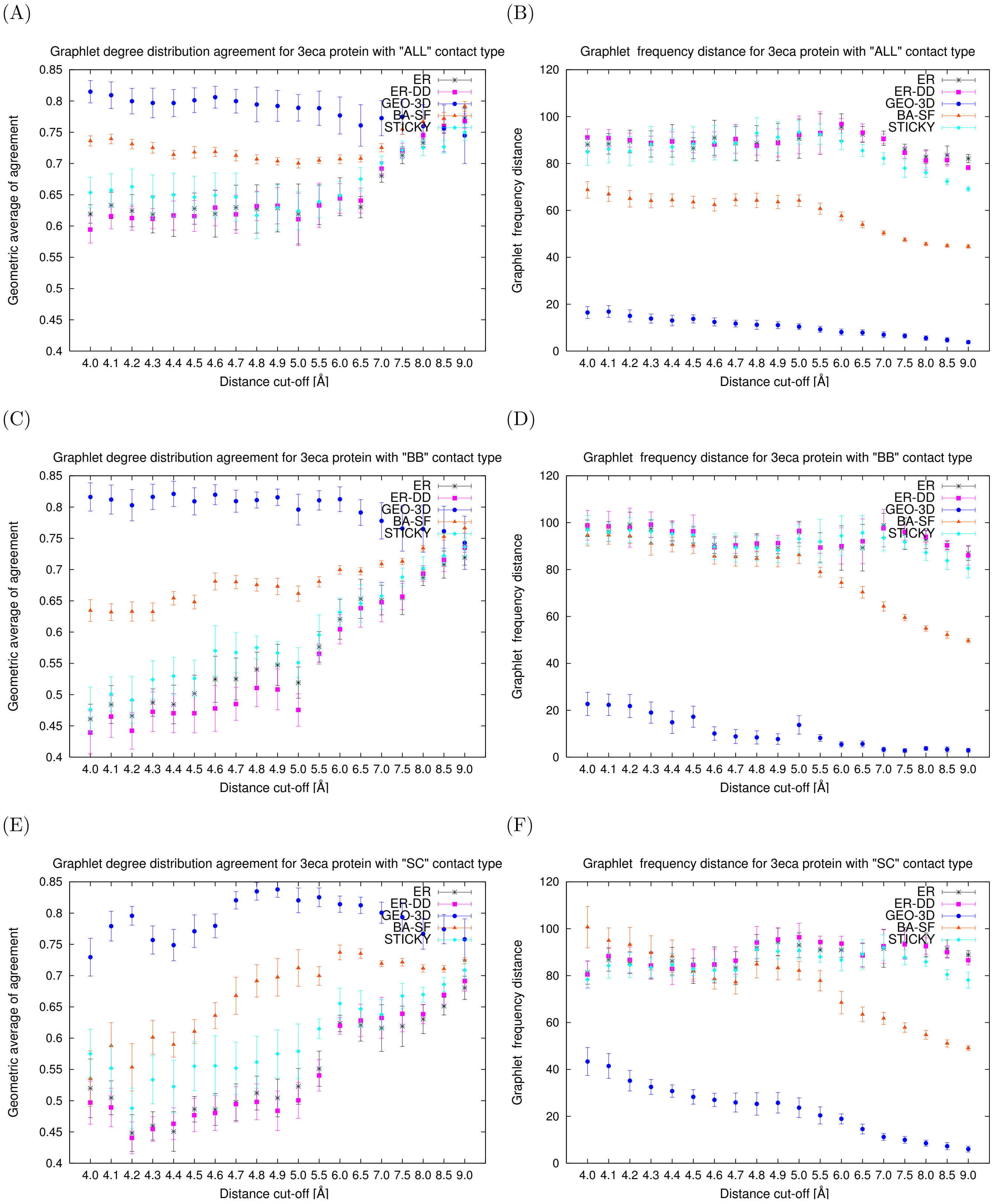
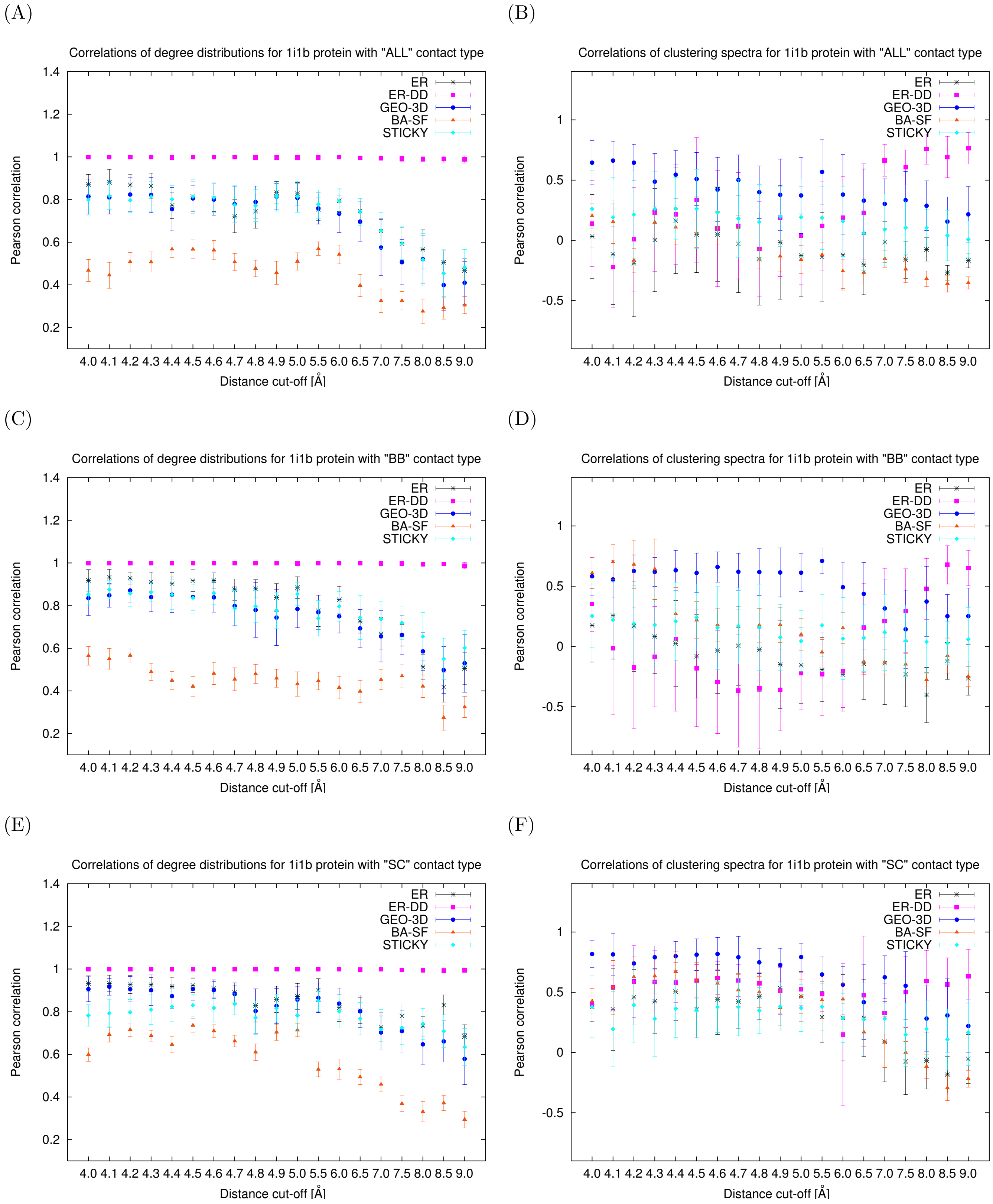


Figure D.8: GDD-agreements and RGF-distances between model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 3eca protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** GDD-agreement for *ALL* contact type. **B.** RGF-distance for *ALL* contact type. **C.** GDD-agreement for *BB* contact type. **D.** RGF-distance for *BB* contact type. **E.** GDD-agreement for *SC* contact type. **F.** RGF-distance for *SC* contact type. The larger the GDD-agreement in panels A, C, and E the better the fit. The smaller the RGF-distance in panels B, D, and F the better the fit.



Correlations of degree distributions for 1i1b protein with "ALL" contact type

Correlations of clustering spectra for 1i1b protein with "ALL" contact type

Correlations of degree distributions for 1i1b protein with "BB" contact type

Correlations of clustering spectra for 1i1b protein with "BB" contact type

Correlations of degree distributions for 1i1b protein with "SC" contact type

Correlations of clustering spectra for 1i1b protein with "SC" contact type

Figure D.9: The Pearson correlation coefficients of degree distributions and clustering spectra of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1i1b protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** degree distribution for *ALL* contact type. **B.** clustering spectrum for *ALL* contact type. **C.** degree distribution for *BB* contact type. **D.** clustering spectrum for *BB* contact type. **E.** degree distribution for *SC* contact type. **F.** clustering spectrum for *SC* contact type.

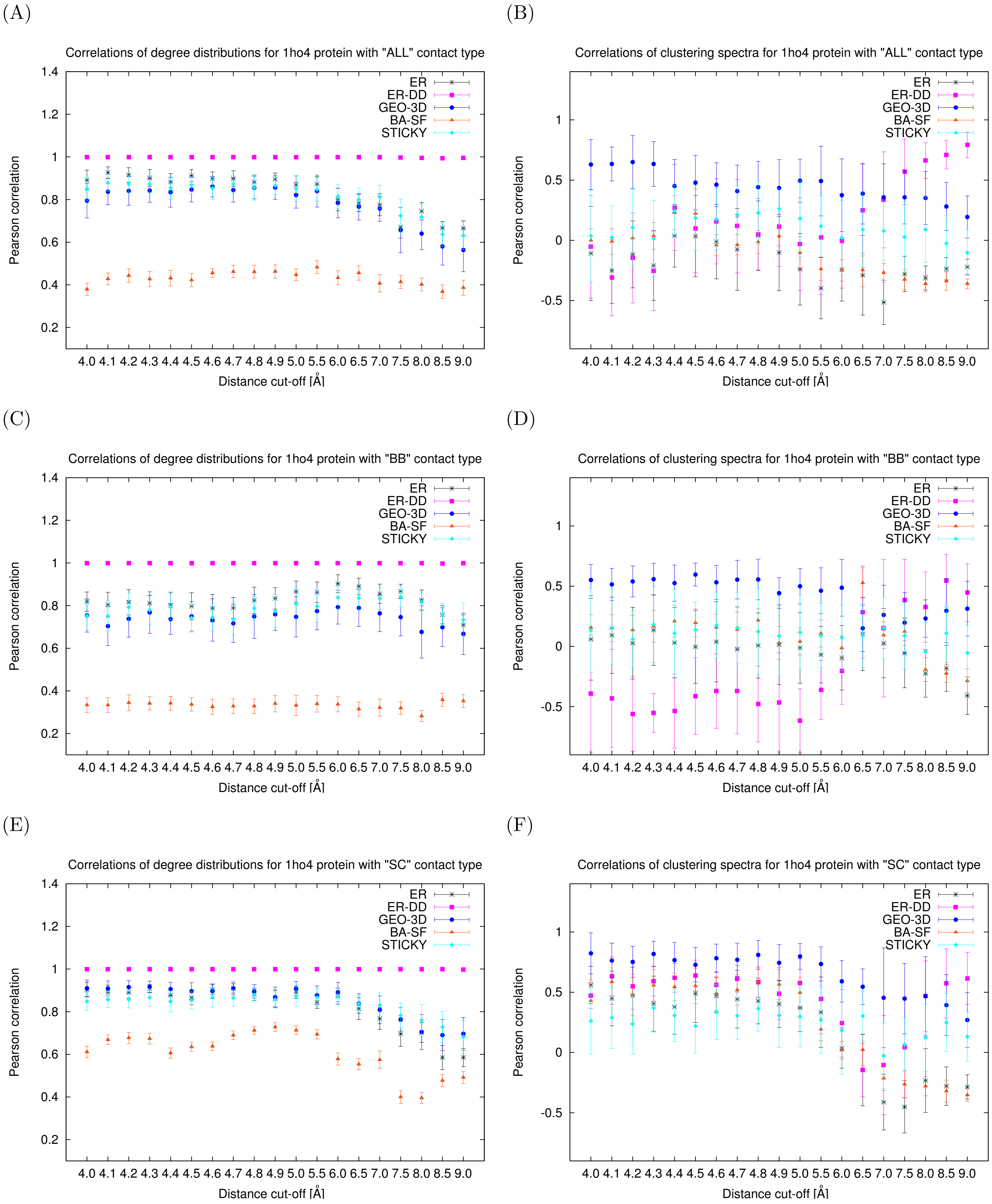


Figure D.10: The Pearson correlation coefficients of degree distributions and clustering spectra of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1ho4 protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** degree distribution for *ALL* contact type. **B.** clustering spectrum for *ALL* contact type. **C.** degree distribution for *BB* contact type. **D.** clustering spectrum for *BB* contact type. **E.** degree distribution for *SC* contact type. **F.** clustering spectrum for *SC* contact type.

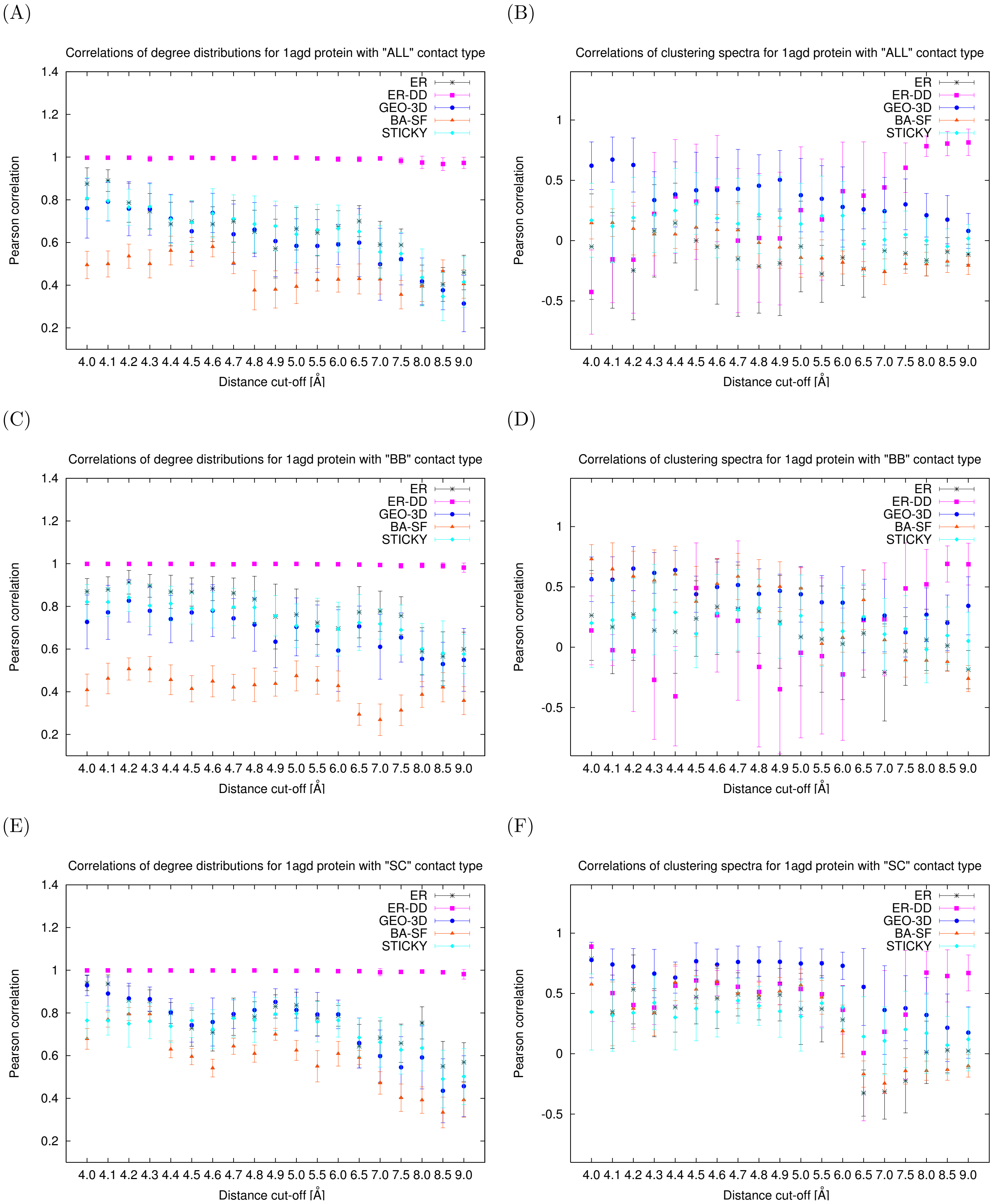


Figure D.11: The Pearson correlation coefficients of degree distributions and clustering spectra of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1agd protein that are constructed for each of the three contact types (ALL, BB and SC) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** degree distribution for ALL contact type. **B.** clustering spectrum for ALL contact type. **C.** degree distribution for BB contact type. **D.** clustering spectrum for BB contact type. **E.** degree distribution for SC contact type. **F.** clustering spectrum for SC contact type.

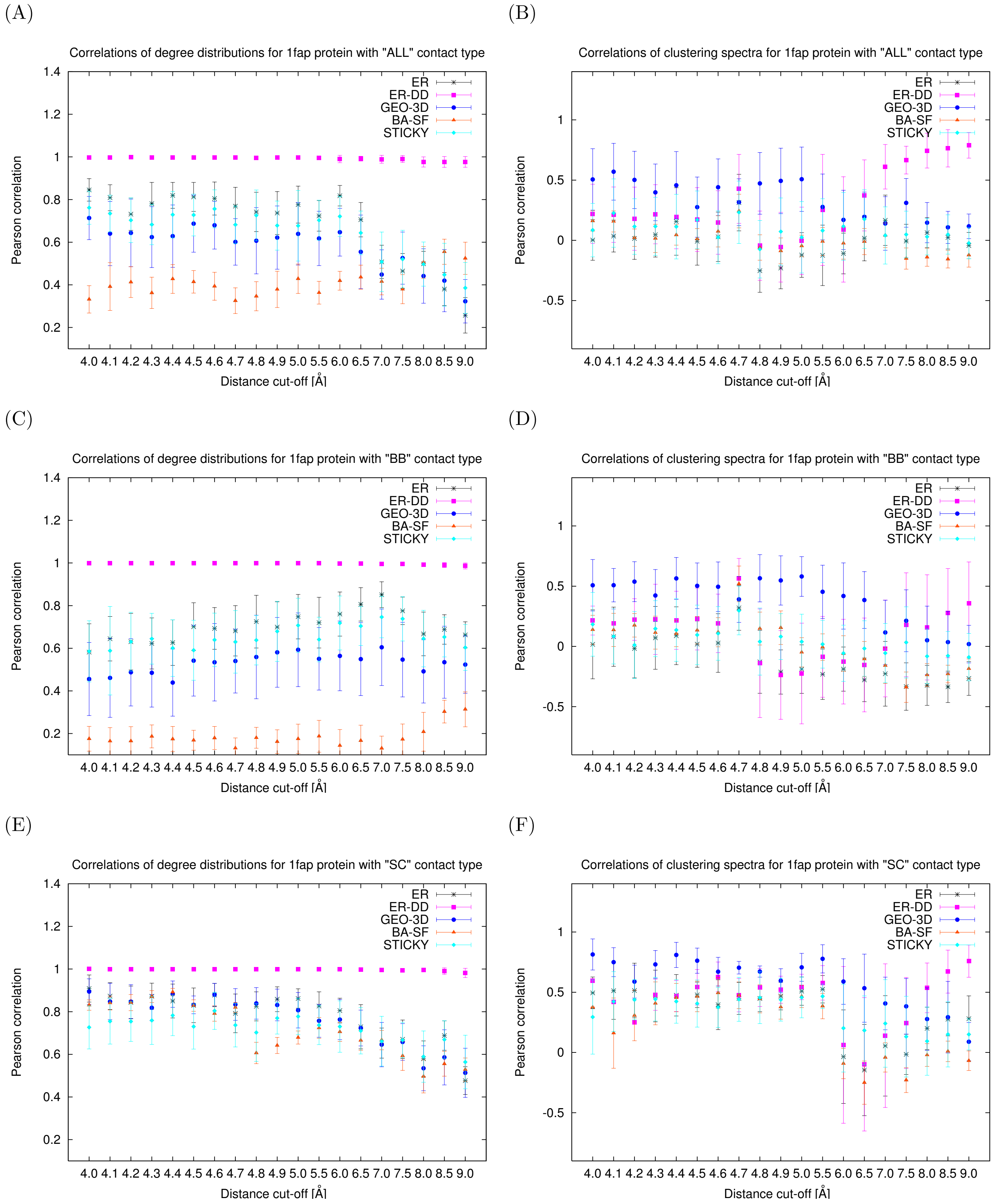


Figure D.12: The Pearson correlation coefficients of degree distributions and clustering spectra of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1fap protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** degree distribution for *ALL* contact type. **B.** clustering spectrum for *ALL* contact type. **C.** degree distribution for *BB* contact type. **D.** clustering spectrum for *BB* contact type. **E.** degree distribution for *SC* contact type. **F.** clustering spectrum for *SC* contact type.

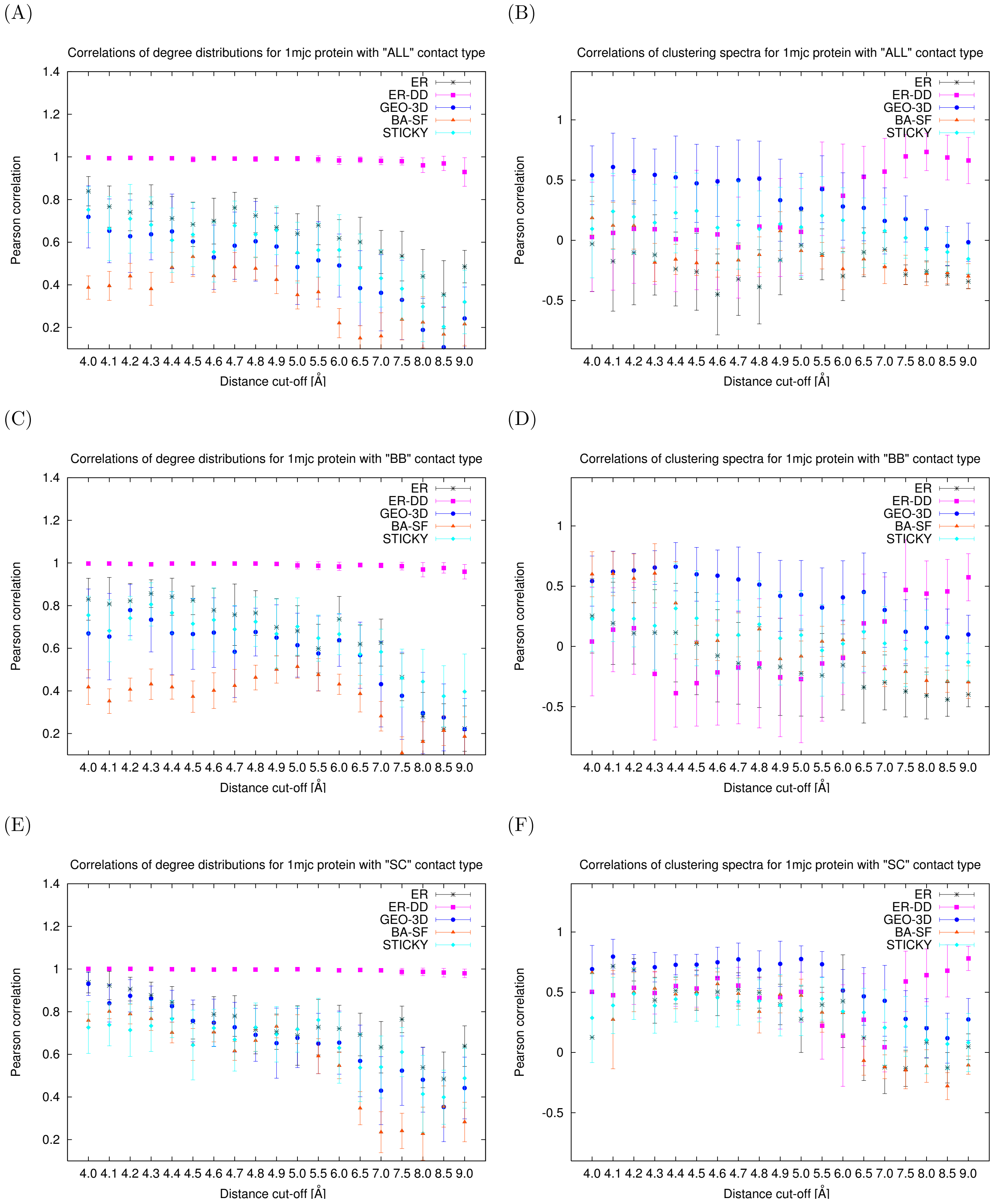


Figure D.13: The Pearson correlation coefficients of degree distributions and clustering spectra of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1mjc protein that are constructed for each of the three contact types (ALL, BB and SC) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** degree distribution for ALL contact type. **B.** clustering spectrum for ALL contact type. **C.** degree distribution for BB contact type. **D.** clustering spectrum for BB contact type. **E.** degree distribution for SC contact type. **F.** clustering spectrum for SC contact type.

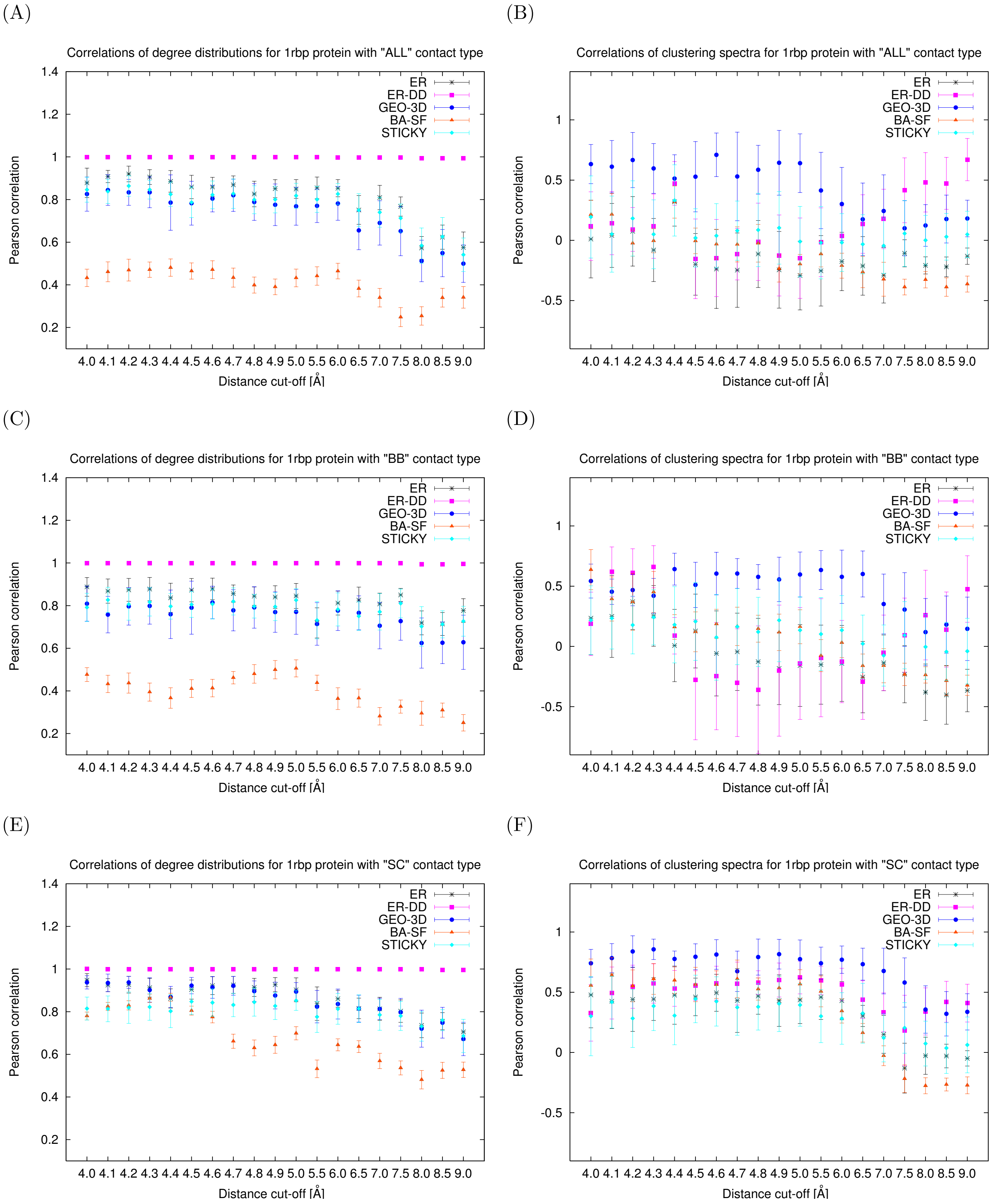


Figure D.14: The Pearson correlation coefficients of degree distributions and clustering spectra of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1rbp protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** degree distribution for *ALL* contact type. **B.** clustering spectrum for *ALL* contact type. **C.** degree distribution for *BB* contact type. **D.** clustering spectrum for *BB* contact type. **E.** degree distribution for *SC* contact type. **F.** clustering spectrum for *SC* contact type.

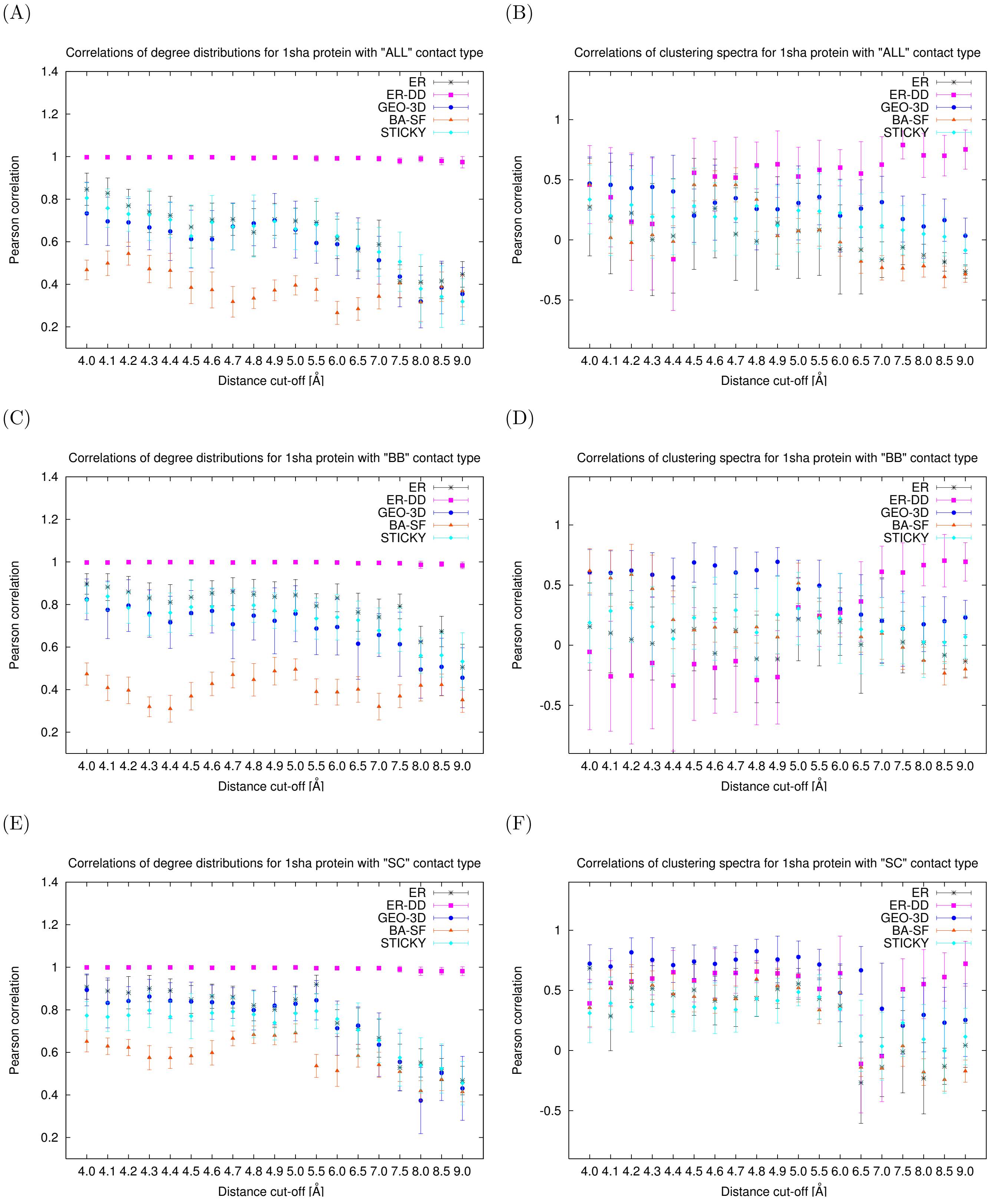


Figure D.15: The Pearson correlation coefficients of degree distributions and clustering spectra of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1sha protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** degree distribution for *ALL* contact type. **B.** clustering spectrum for *ALL* contact type. **C.** degree distribution for *BB* contact type. **D.** clustering spectrum for *BB* contact type. **E.** degree distribution for *SC* contact type. **F.** clustering spectrum for *SC* contact type.

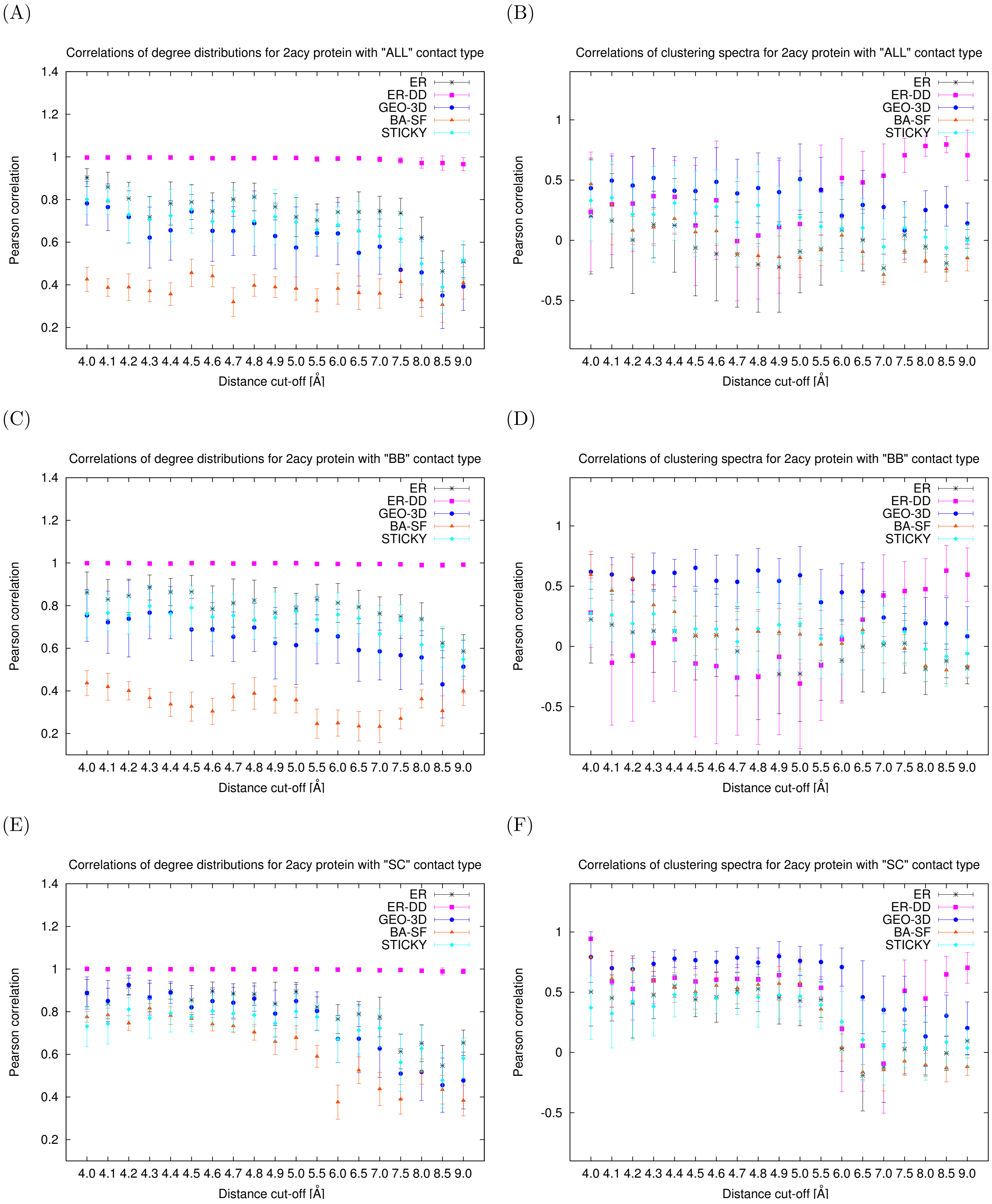


Figure D.16: The Pearson correlation coefficients of degree distributions and clustering spectra of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 2acy protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** degree distribution for *ALL* contact type. **B.** clustering spectrum for *ALL* contact type. **C.** degree distribution for *BB* contact type. **D.** clustering spectrum for *BB* contact type. **E.** degree distribution for *SC* contact type. **F.** clustering spectrum for *SC* contact type.

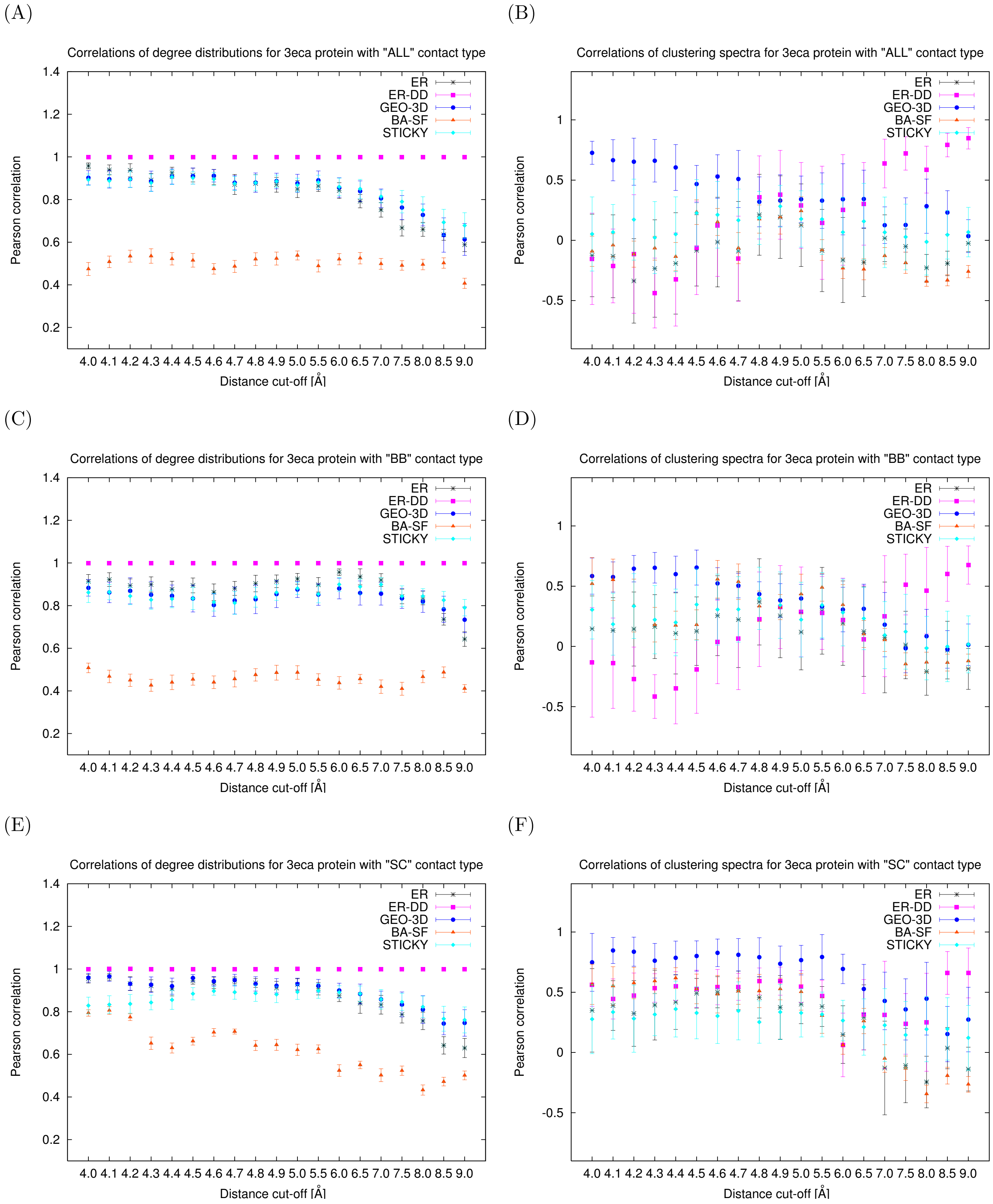


Figure D.17: The Pearson correlation coefficients of degree distributions and clustering spectra of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 3eca protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** degree distribution for *ALL* contact type. **B.** clustering spectrum for *ALL* contact type. **C.** degree distribution for *BB* contact type. **D.** clustering spectrum for *BB* contact type. **E.** degree distribution for *SC* contact type. **F.** clustering spectrum for *SC* contact type.

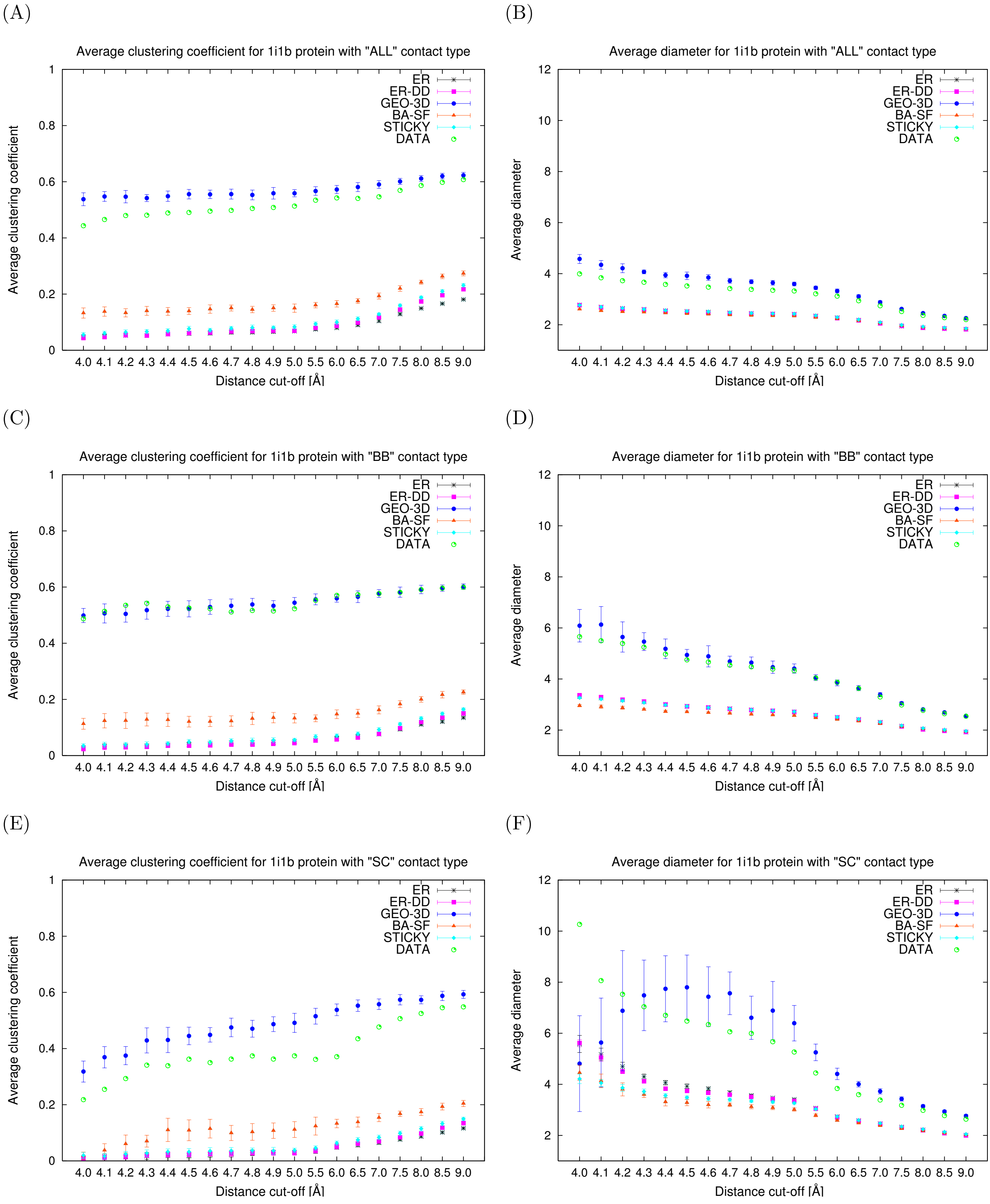


Figure D.18: The agreements of average clustering coefficients and diameters of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1i1b protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** average clustering coefficients for *ALL* contact type. **B.** average diameters for *ALL* contact type. **C.** average clustering coefficients for *BB* contact type. **D.** average diameters for *BB* contact type. **E.** average clustering coefficients for *SC* contact type. **F.** average diameters for *SC* contact type.

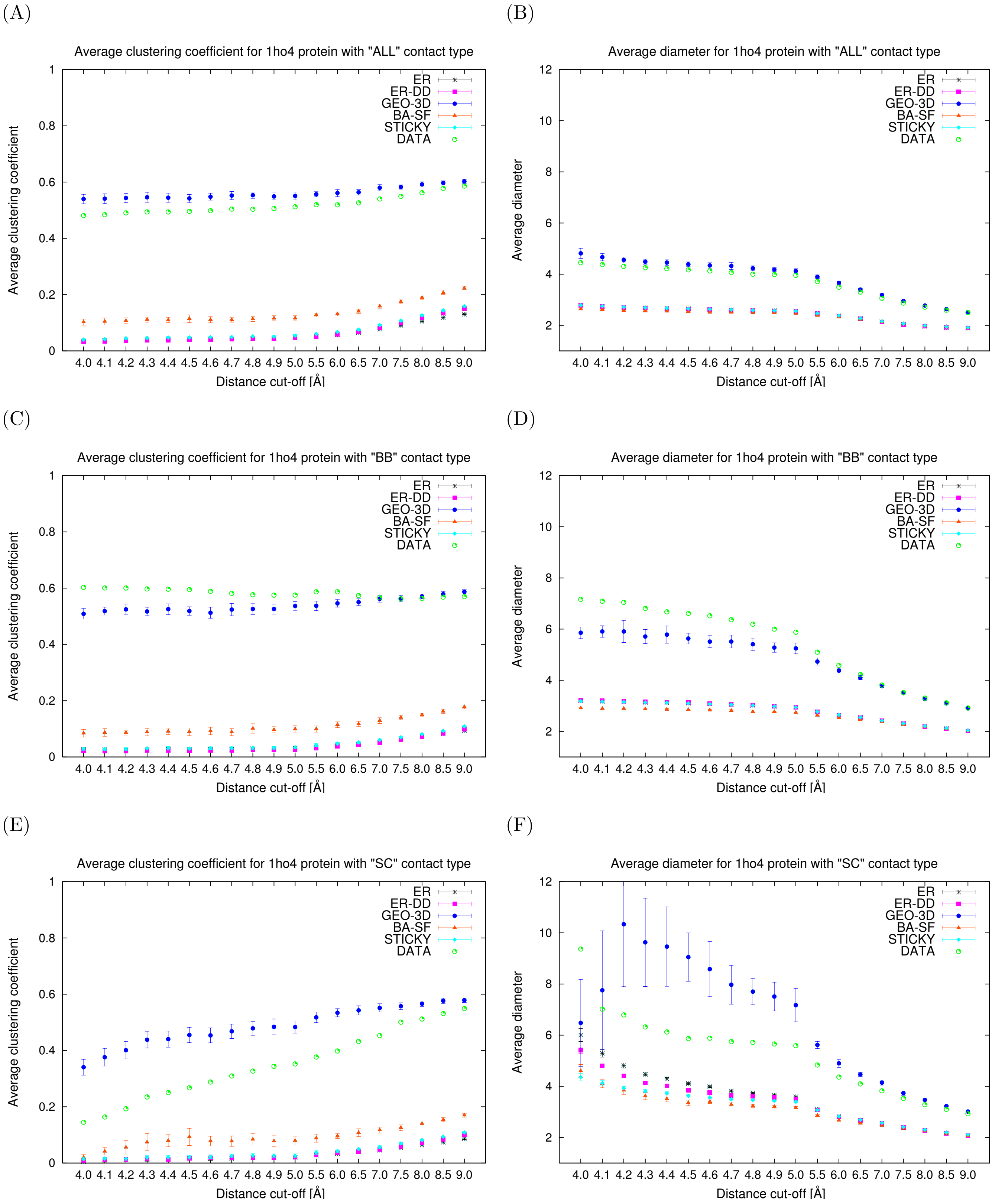


Figure D.19: The agreements of average clustering coefficients and diameters of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1ho4 protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** average clustering coefficients for *ALL* contact type. **B.** average diameters for *ALL* contact type. **C.** average clustering coefficients for *BB* contact type. **D.** average diameters for *BB* contact type. **E.** average clustering coefficients for *SC* contact type. **F.** average diameters for *SC* contact type.

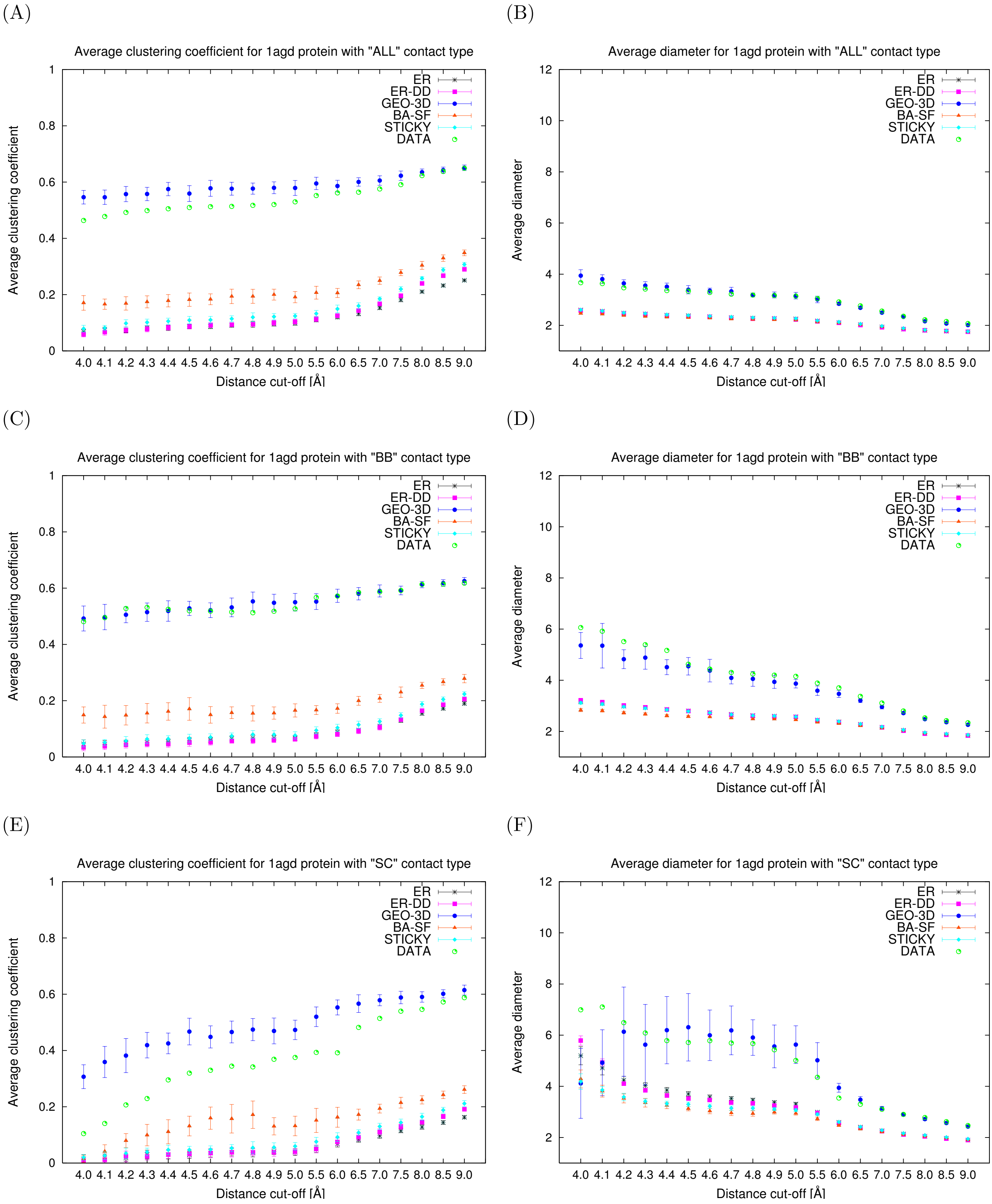


Figure D.20: The agreements of average clustering coefficients and diameters of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1agd protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** average clustering coefficients for *ALL* contact type. **B.** average diameters for *ALL* contact type. **C.** average clustering coefficients for *BB* contact type. **D.** average diameters for *BB* contact type. **E.** average clustering coefficients for *SC* contact type. **F.** average diameters for *SC* contact type.

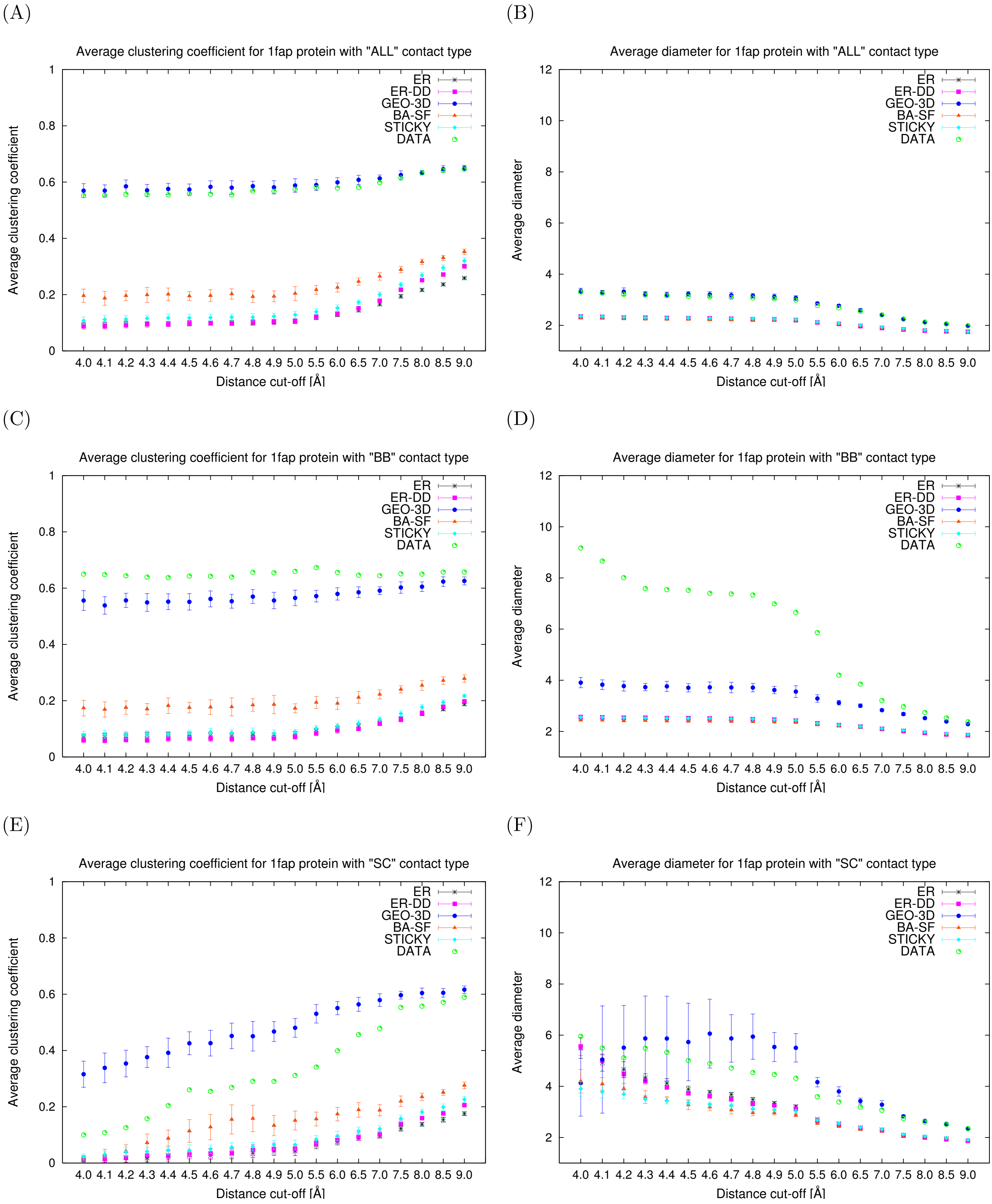


Figure D.21: The agreements of average clustering coefficients and diameters of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1fap protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** average clustering coefficients for *ALL* contact type. **B.** average diameters for *ALL* contact type. **C.** average clustering coefficients for *BB* contact type. **D.** average diameters for *BB* contact type. **E.** average clustering coefficients for *SC* contact type. **F.** average diameters for *SC* contact type.

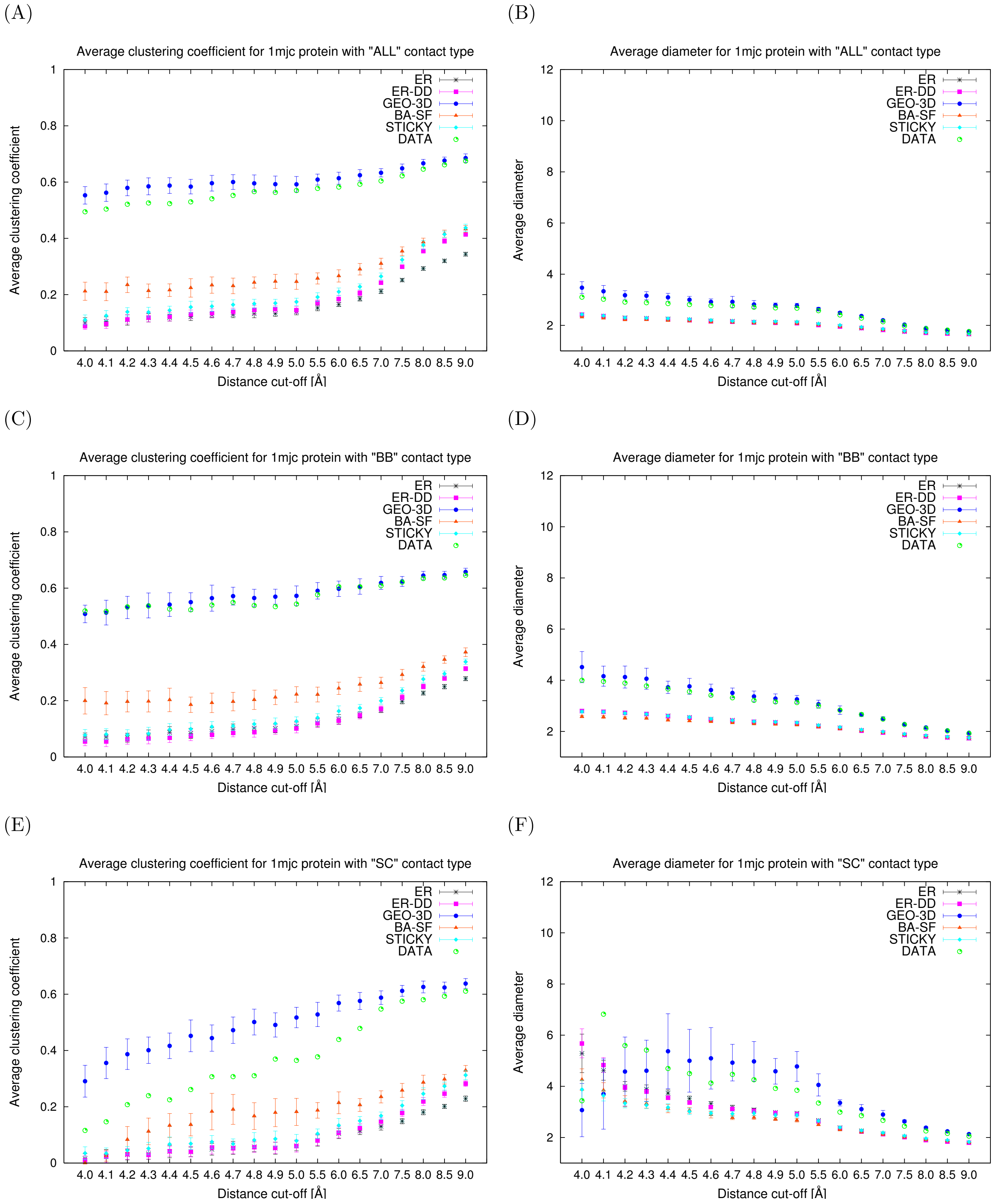


Figure D.22: The agreements of average clustering coefficients and diameters of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1mjc protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** average clustering coefficients for *ALL* contact type. **B.** average diameters for *ALL* contact type. **C.** average clustering coefficients for *BB* contact type. **D.** average diameters for *BB* contact type. **E.** average clustering coefficients for *SC* contact type. **F.** average diameters for *SC* contact type.

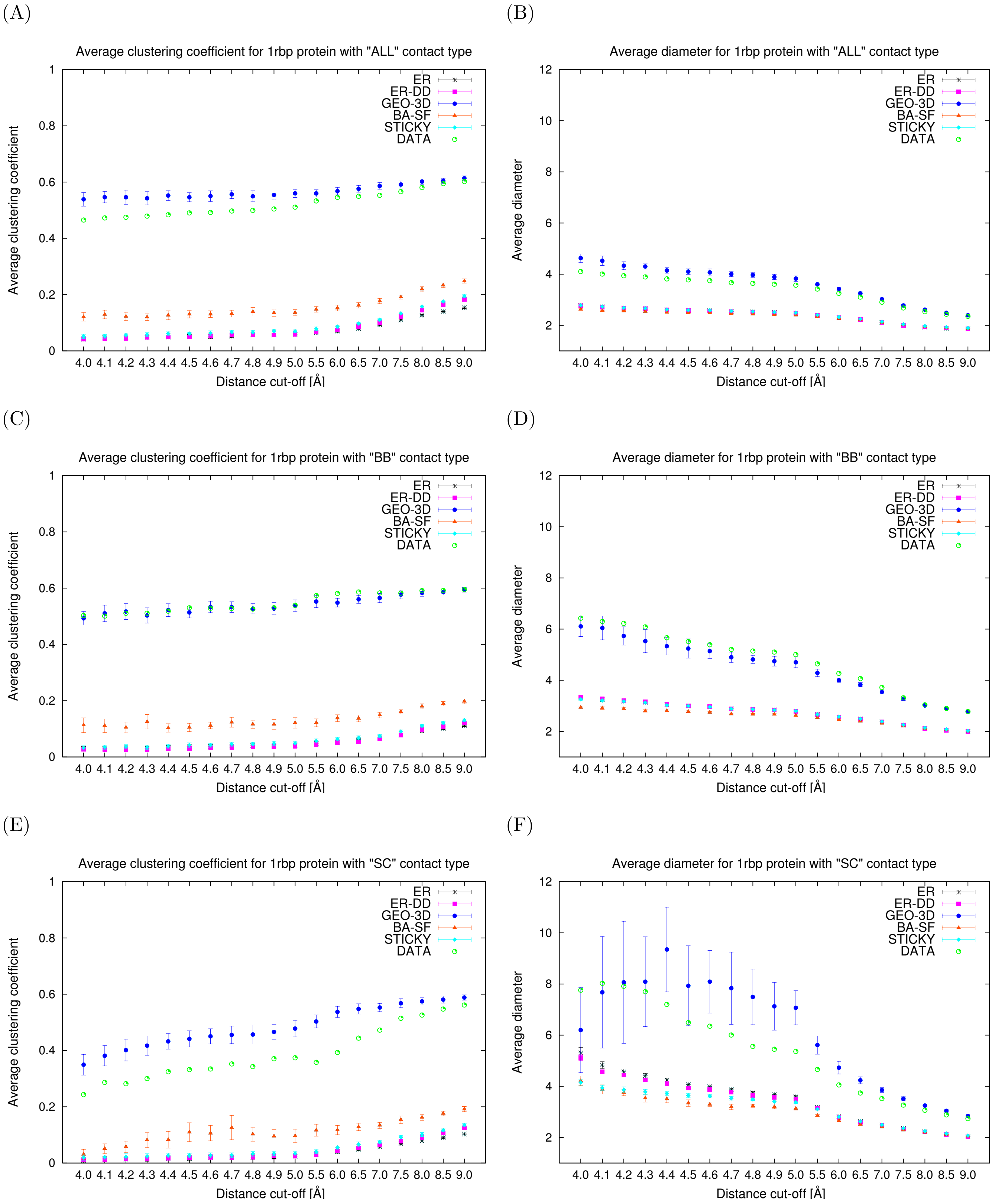


Figure D.23: The agreements of average clustering coefficients and diameters of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1rbp protein that are constructed for each of the three contact types (ALL, BB and SC) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** average clustering coefficients for ALL contact type. **B.** average diameters for ALL contact type. **C.** average clustering coefficients for BB contact type. **D.** average diameters for BB contact type. **E.** average clustering coefficients for SC contact type. **F.** average diameters for SC contact type.

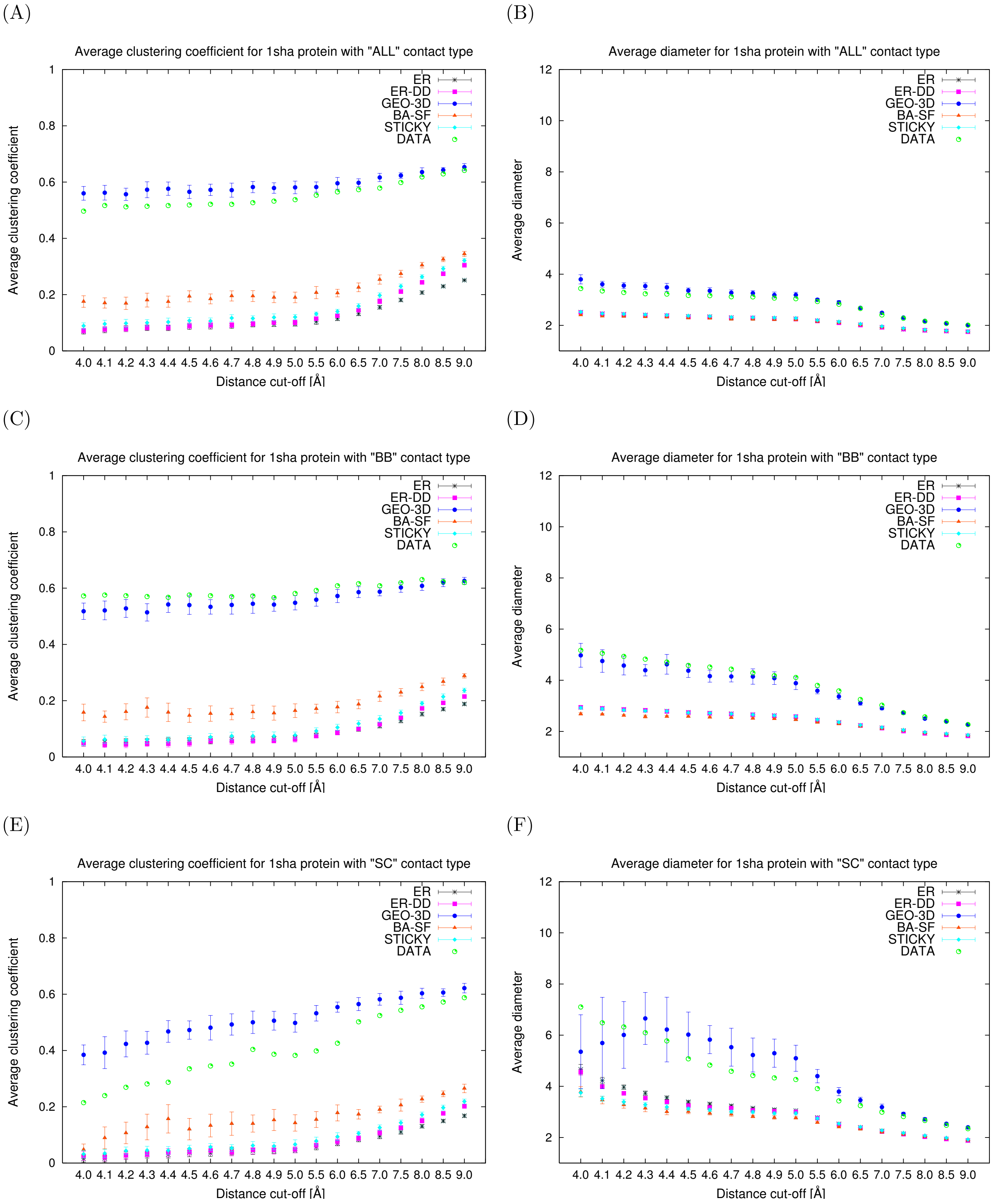


Figure D.24: The agreements of average clustering coefficients and diameters of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1sha protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** average clustering coefficients for *ALL* contact type. **B.** average diameters for *ALL* contact type. **C.** average clustering coefficients for *BB* contact type. **D.** average diameters for *BB* contact type. **E.** average clustering coefficients for *SC* contact type. **F.** average diameters for *SC* contact type.

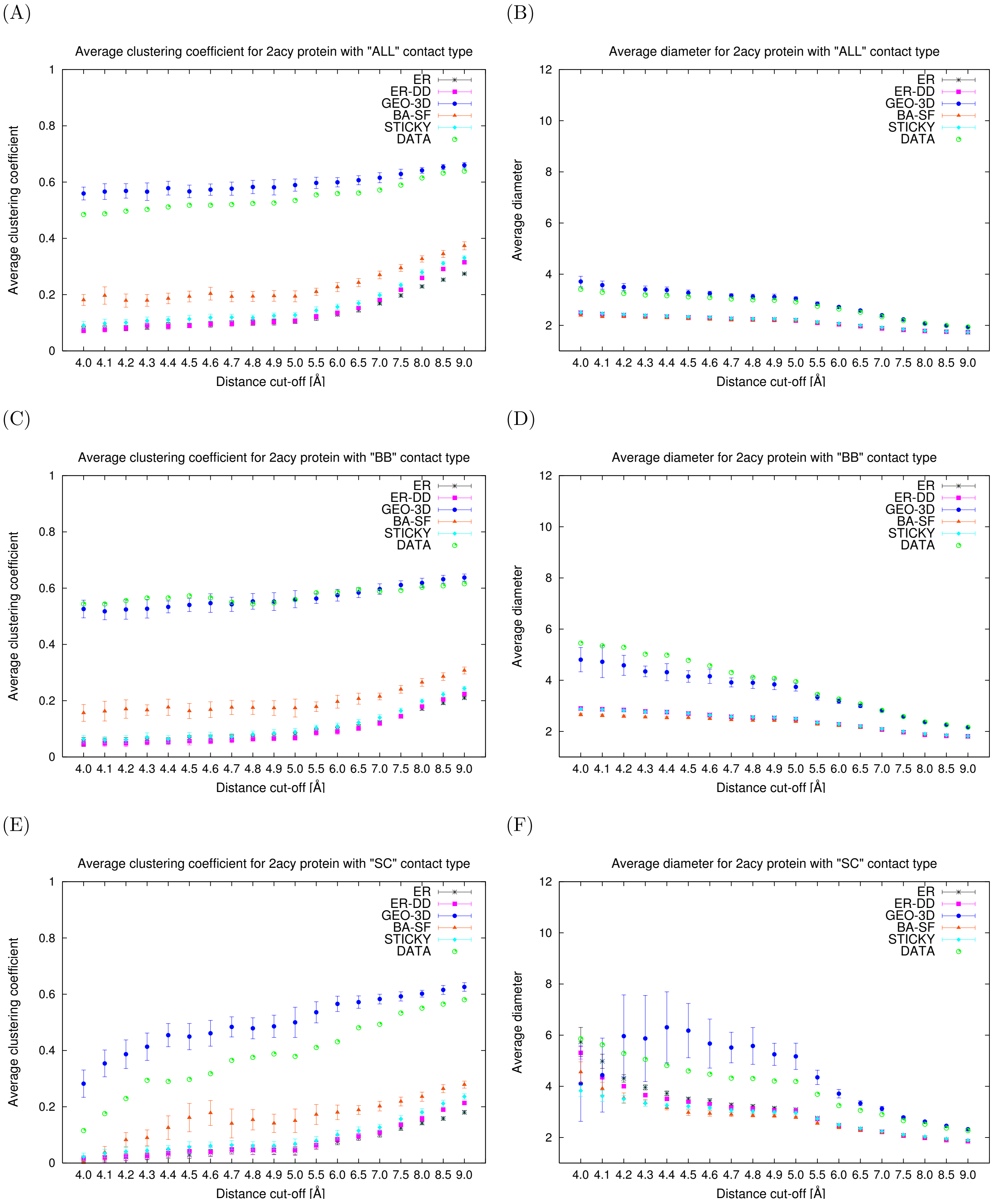


Figure D.25: The agreements of average clustering coefficients and diameters of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 2acy protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** average clustering coefficients for *ALL* contact type. **B.** average diameters for *ALL* contact type. **C.** average clustering coefficients for *BB* contact type. **D.** average diameters for *BB* contact type. **E.** average clustering coefficients for *SC* contact type. **F.** average diameters for *SC* contact type.

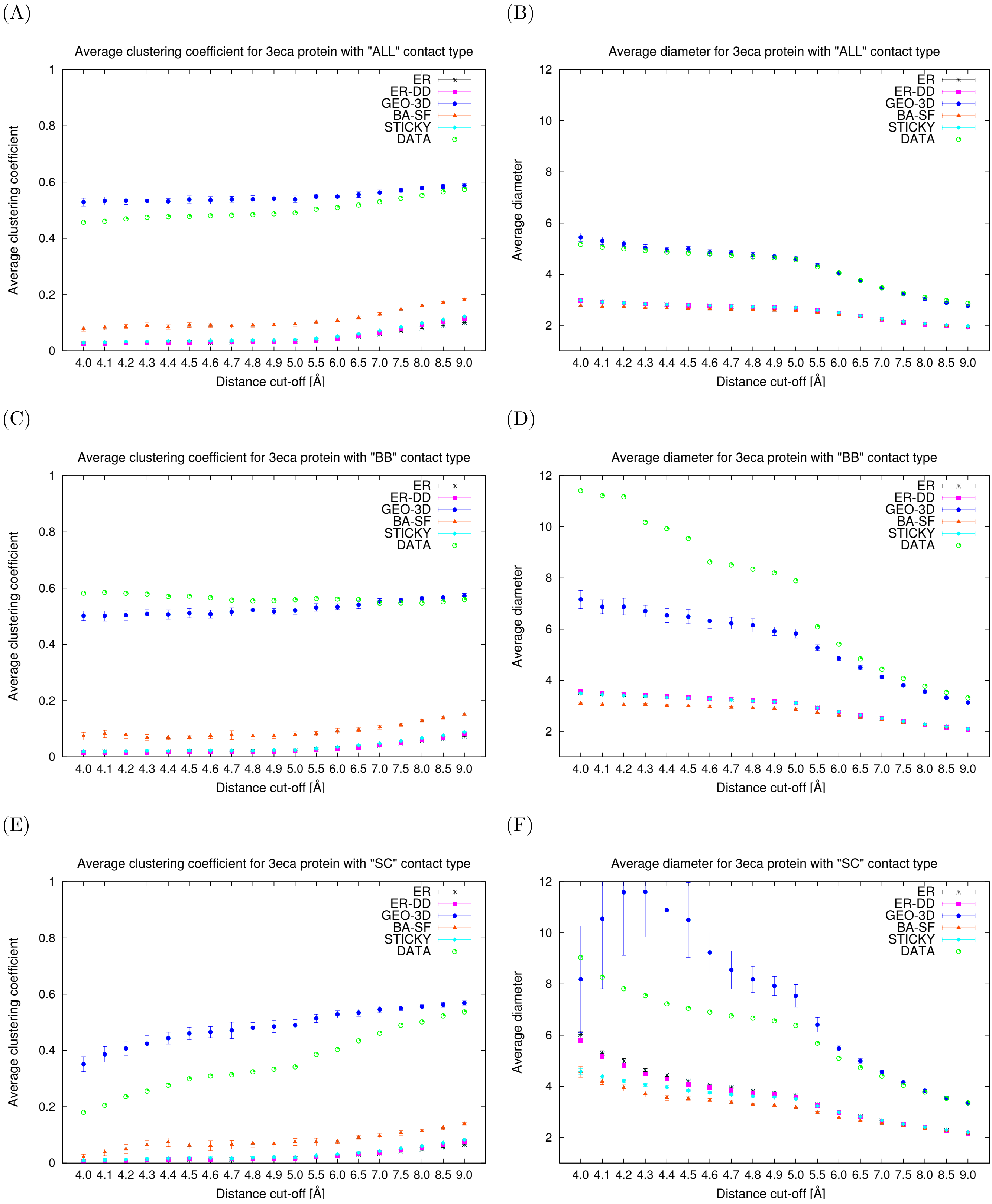


Figure D.26: The agreements of average clustering coefficients and diameters of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 3eca protein that are constructed for each of the three contact types (*ALL*, *BB* and *SC*) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** average clustering coefficients for *ALL* contact type. **B.** average diameters for *ALL* contact type. **C.** average clustering coefficients for *BB* contact type. **D.** average diameters for *BB* contact type. **E.** average clustering coefficients for *SC* contact type. **F.** average diameters for *SC* contact type.

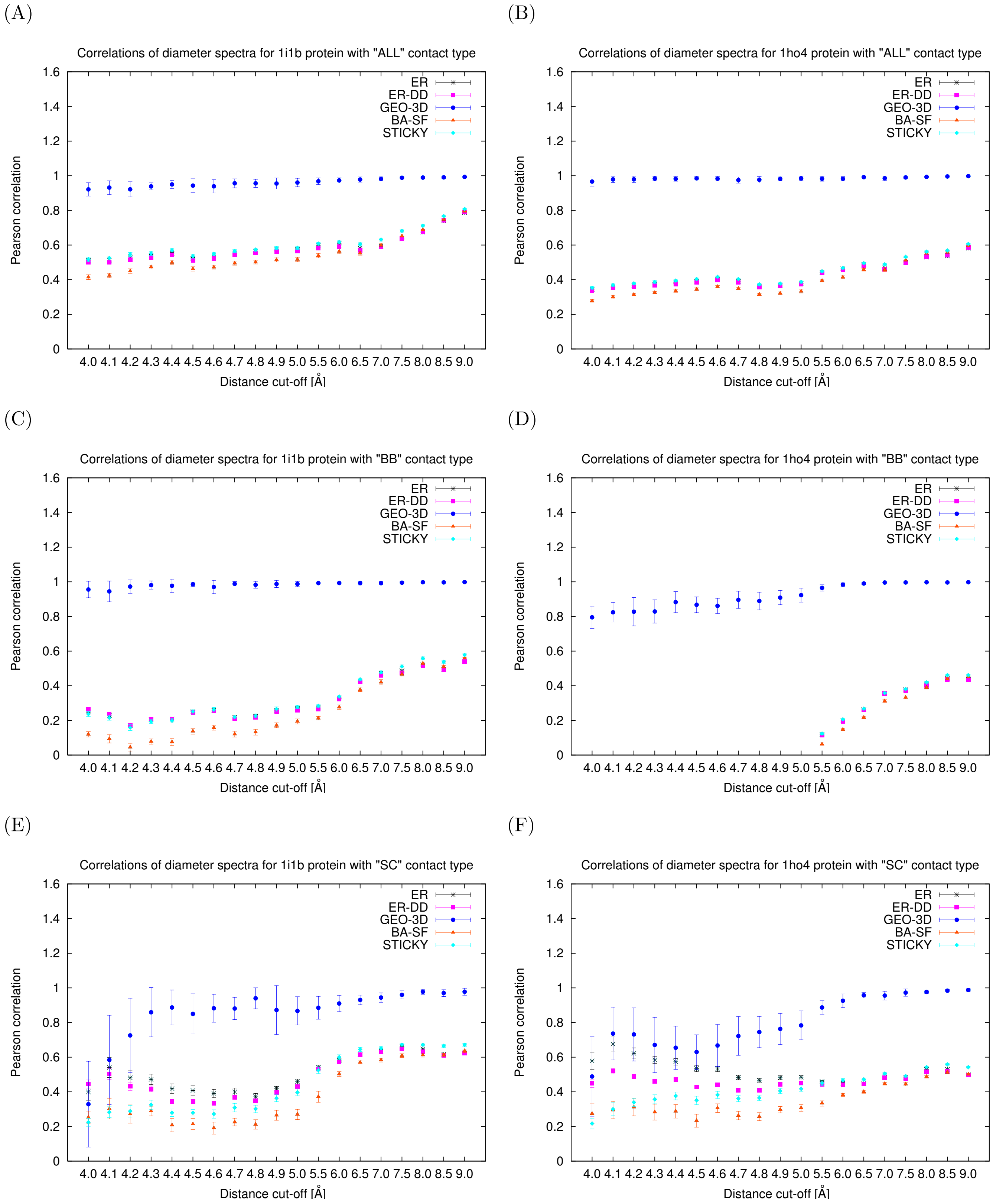


Figure D.27: The Pearson correlation coefficients of the shortest path lengths spectra of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1i1b and 1ho4 proteins that are constructed for each of the three contact types (“BB”, “ALL”, and “SC”) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** spectra of shortest path lengths for *ALL* contact type and 1i1b. **B.** spectra of shortest path lengths for *ALL* contact type and 1ho4. **C.** spectra of shortest path lengths for *BB* contact type and 1i1b. **D.** spectra of shortest path lengths for *BB* contact type and 1ho4. **E.** spectra of shortest path lengths for *SC* contact type and 1i1b. **F.** spectra of shortest path lengths for *SC* contact type and 1ho4.

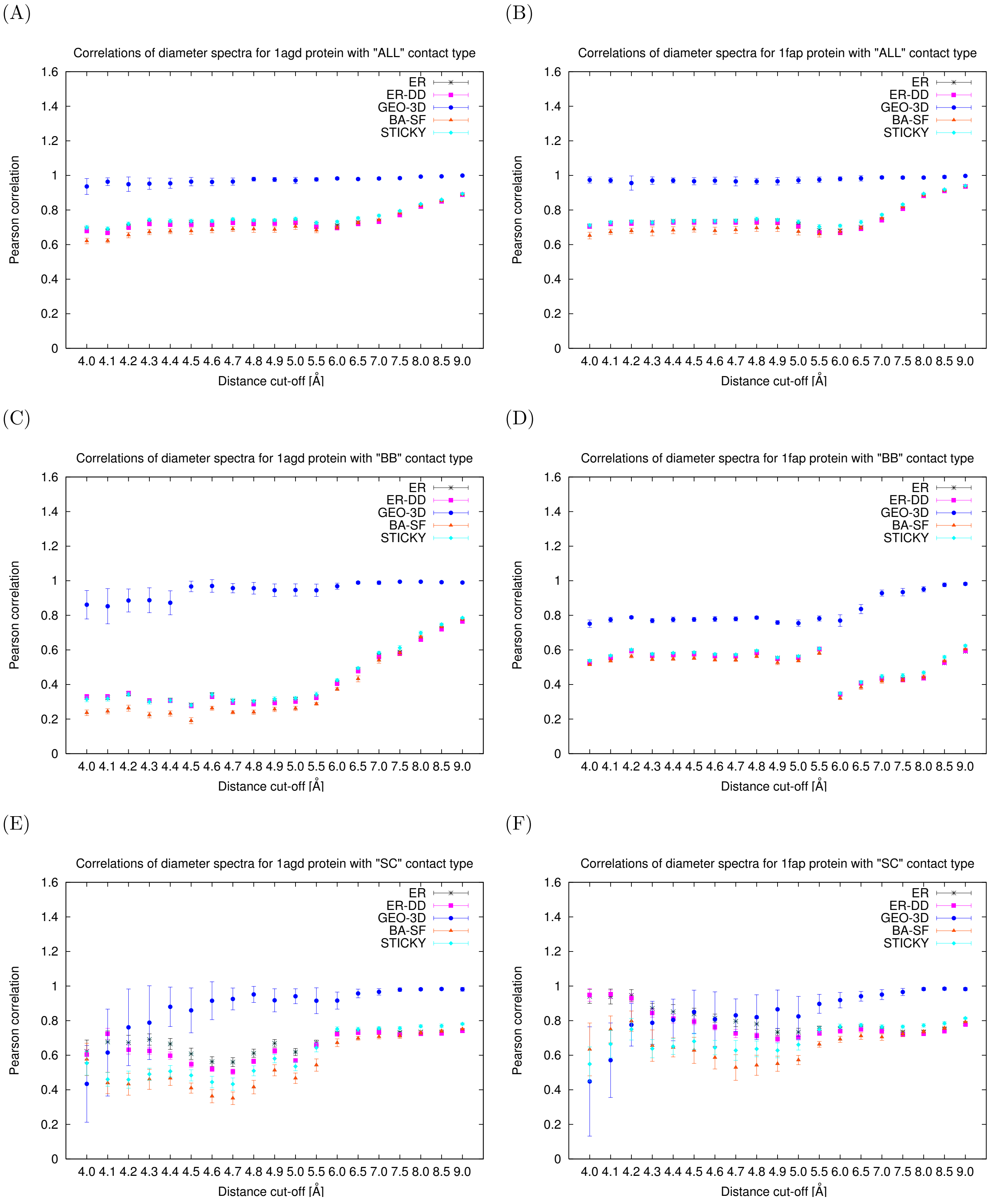


Figure D.28: The Pearson correlation coefficients of the shortest path lengths spectra of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1agd and 1fap proteins that are constructed for each of the three contact types (“BB”, “ALL”, and “SC”) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** spectra of shortest path lengths for *ALL* contact type and 1agd. **B.** spectra of shortest path lengths for *ALL* contact type and 1fap. **C.** spectra of shortest path lengths for *BB* contact type and 1agd. **D.** spectra of shortest path lengths for *BB* contact type and 1fap. **E.** spectra of shortest path lengths for *SC* contact type and 1agd. **F.** spectra of shortest path lengths for *SC* contact type and 1fap.

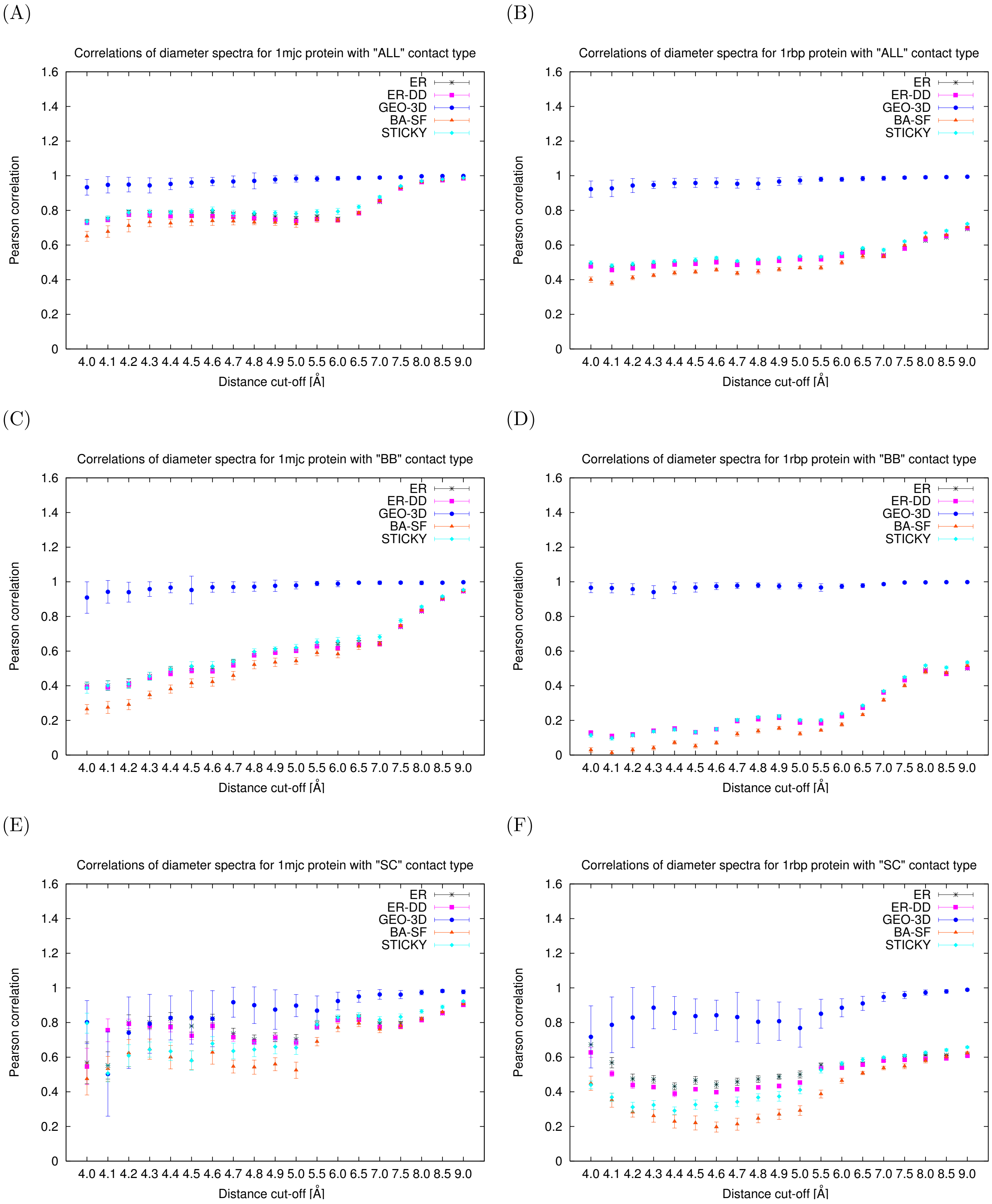
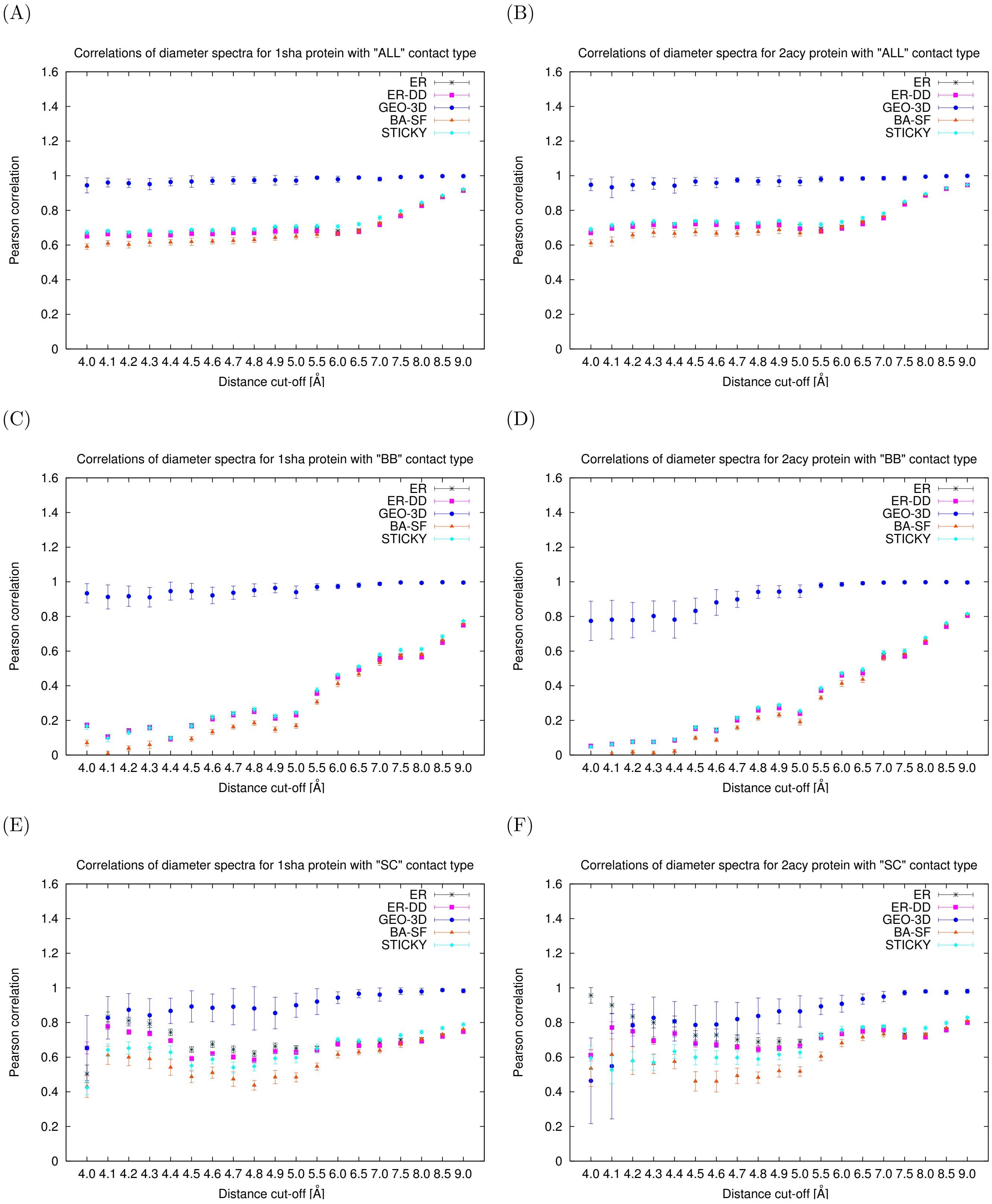


Figure D.29: The Pearson correlation coefficients of the shortest path lengths spectra of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1mjc and 1rbp proteins that are constructed for each of the three contact types (“BB”, “ALL”, and “SC”) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** spectra of shortest path lengths for *ALL* contact type and 1mjc. **B.** spectra of shortest path lengths for *ALL* contact type and 1rbp. **C.** spectra of shortest path lengths for *BB* contact type and 1mjc. **D.** spectra of shortest path lengths for *BB* contact type and 1rbp. **E.** spectra of shortest path lengths for *SC* contact type and 1mjc. **F.** spectra of shortest path lengths for *SC* contact type and 1rbp.


(A) Correlations of diameter spectra for 1sha protein with "ALL" contact type

(B) Correlations of diameter spectra for 2acy protein with "ALL" contact type

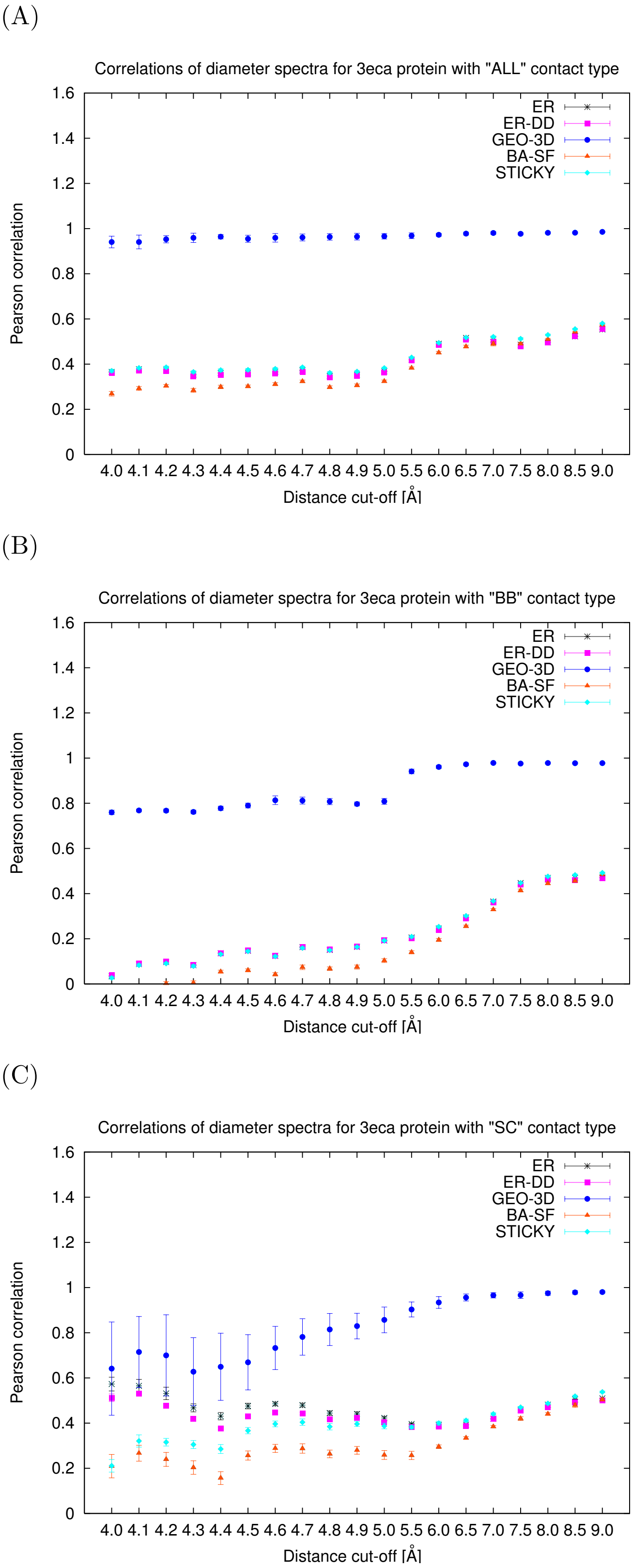
(C) Correlations of diameter spectra for 1sha protein with "BB" contact type

(D) Correlations of diameter spectra for 2acy protein with "BB" contact type

(E) Correlations of diameter spectra for 1sha protein with "SC" contact type

(F) Correlations of diameter spectra for 2acy protein with "SC" contact type

Figure D.30: The Pearson correlation coefficients of the shortest path lengths spectra of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 1sha and 2acy proteins that are constructed for each of the three contact types (“BB”, “ALL”, and “SC”) and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** spectra of shortest path lengths for *ALL* contact type and 1sha. **B.** spectra of shortest path lengths for *ALL* contact type and 2acy. **C.** spectra of shortest path lengths for *BB* contact type and 1sha. **D.** spectra of shortest path lengths for *BB* contact type and 2acy. **E.** spectra of shortest path lengths for *SC* contact type and 1sha. **F.** spectra of shortest path lengths for *SC* contact type and 2acy.


(A)

(B)

(C)

Figure D.31: The Pearson correlation coefficients of the shortest path lengths spectra of model networks (ER, ER-DD, GEO-3D, SF-BA, and STICKY) and RIGs corresponding to 3eca protein that are constructed for each of the three contact types ("BB", "ALL", and "SC") and a series of distance cut-off values between 4.0 and 9.0 Angstroms: **A.** spectra of shortest path lengths for *ALL* contact type. **B.** spectra of shortest path lengths for *BB* contact type. **C.** spectra of shortest path lengths for *SC* contact type.

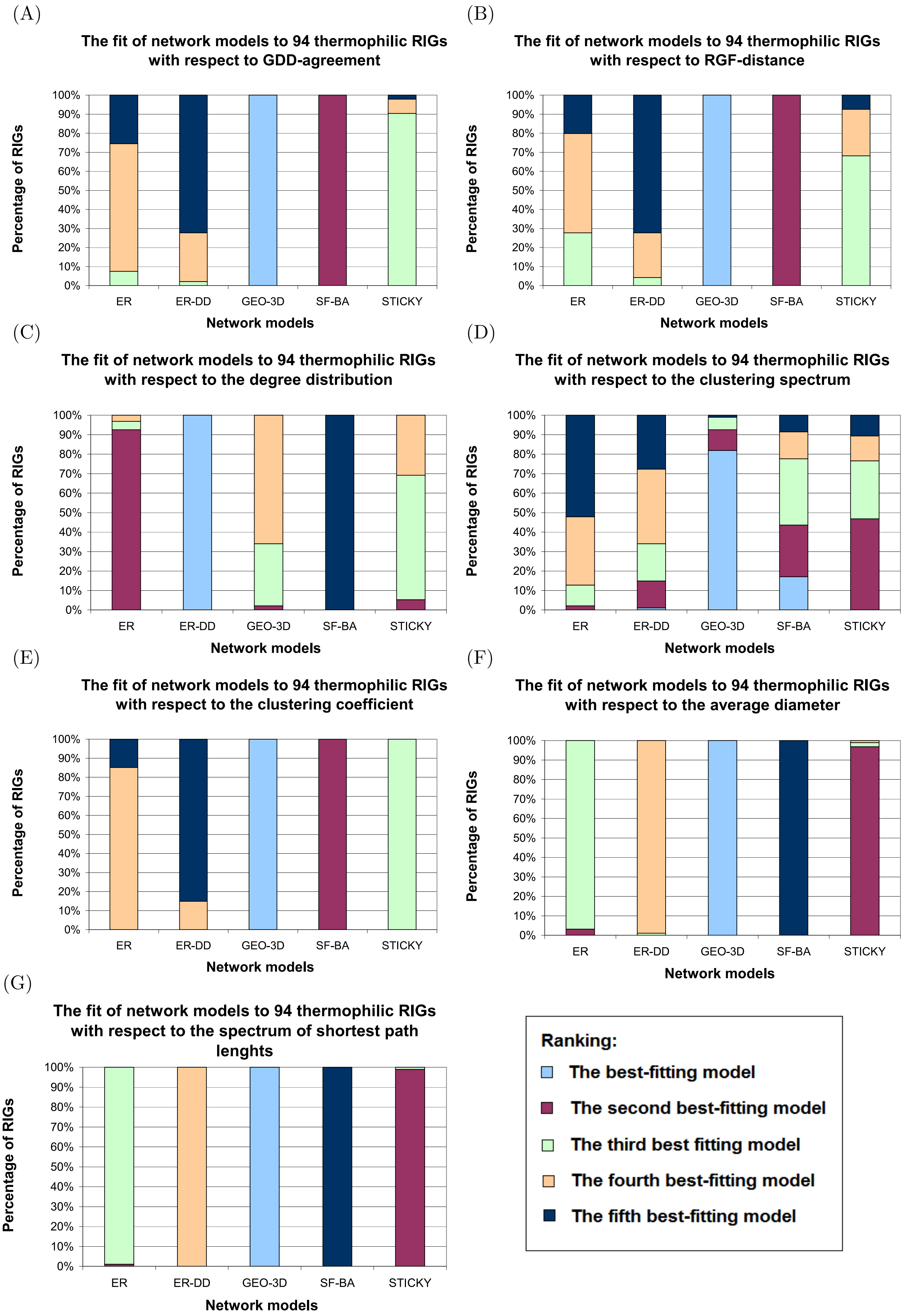


Figure D.32: The ranking of five network models (ER, ER-DD, GEO-3D, SF-BA, and STICKY) for 94 $(ALL)^{\mathrm{all}}_{4.5\mathring{A}}$ RIGs corresponding to the 94 thermophilic proteins. Ranking is based on GDD-agreements (A), RGF-distances (B), and agreements between degree distributions (C), clustering spectra (D), clustering coefficients (E), average diameters (F) and spectra of shortest path lengths (G).

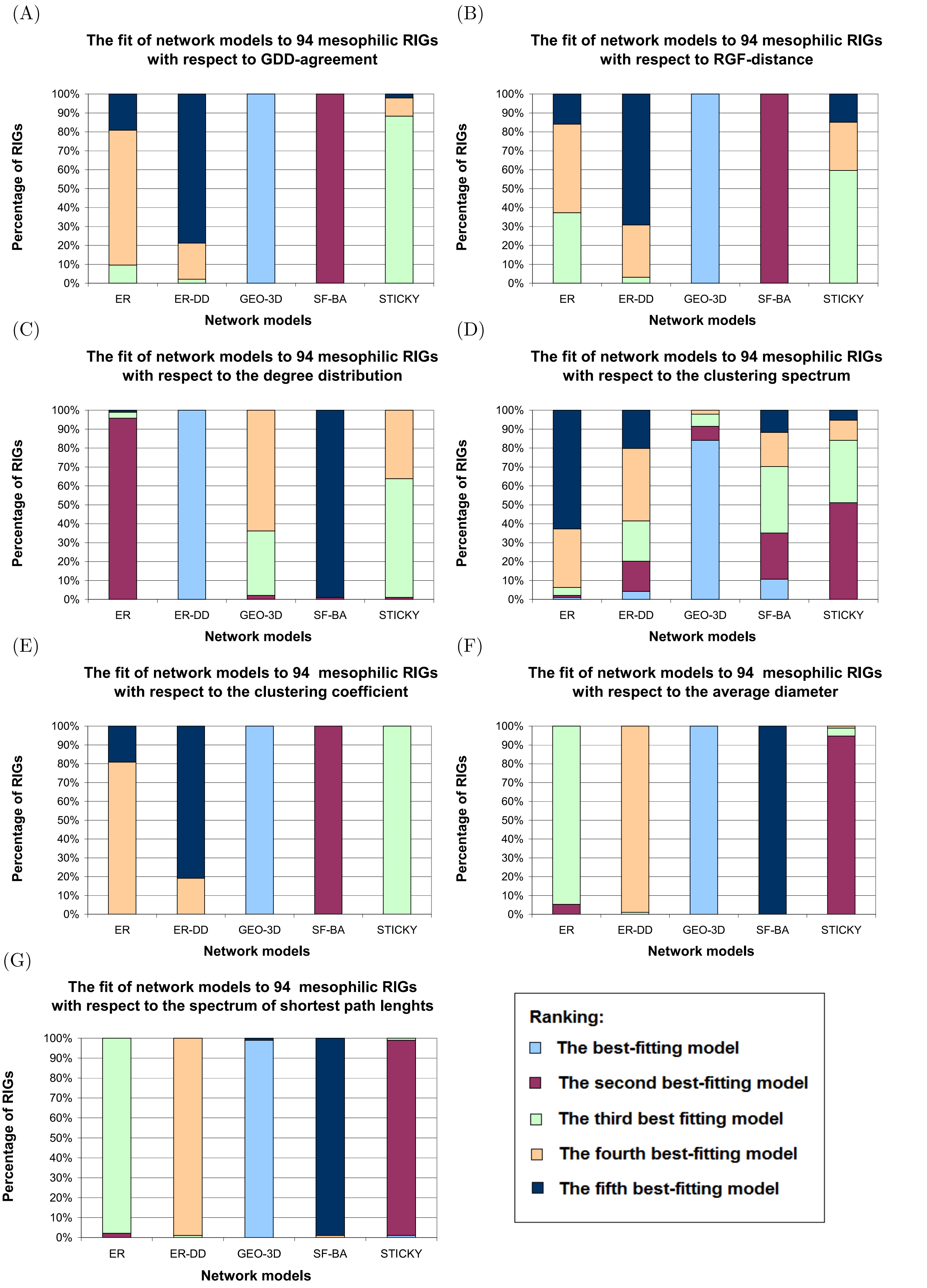
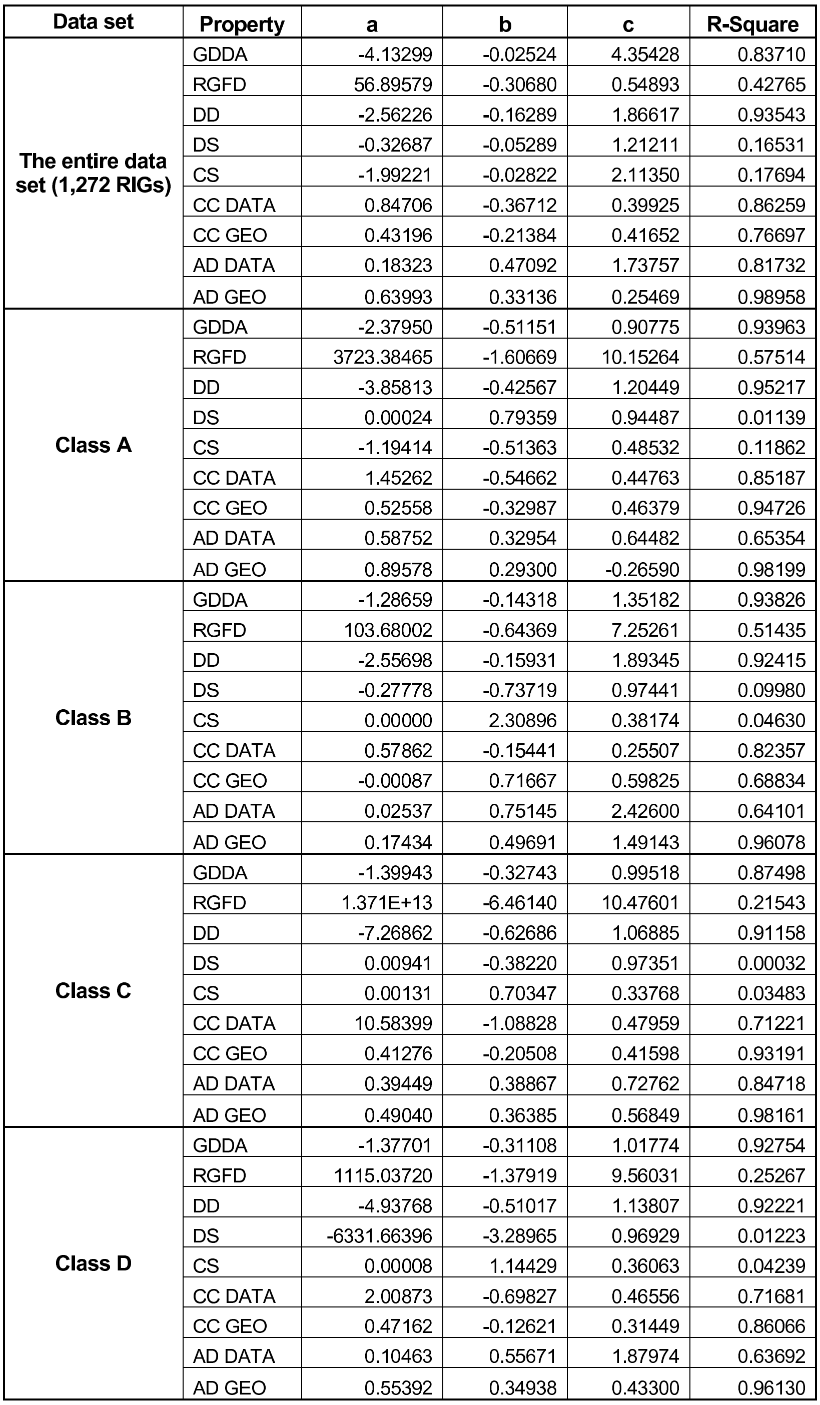


Figure D.33: The ranking of five network models (ER, ER-DD, GEO-3D, SF-BA, and STICKY) for 94 $(ALL)^{\text{all}}_{4.5\mathring{A}}$ RIGs corresponding to the 94 mesophilic proteins. Ranking is based on GDD-agreements (A), RGF-distances (B), and agreements between degree distributions (C), clustering spectra (D), clustering coefficients (E), average diameters (F) and spectra of shortest path lengths (G).



Data set	Property	a	b	c	R-Square
The entire data set (1,272 RIGs)	GDDA	-4.13299	-0.02524	4.35428	0.83710
	RGFD	56.89579	-0.30680	0.54893	0.42765
	DD	-2.56226	-0.16289	1.86617	0.93543
	DS	-0.32687	-0.05289	1.21211	0.16531
	CS	-1.99221	-0.02822	2.11350	0.17694
	CC DATA	0.84706	-0.36712	0.39925	0.86259
	CC GEO	0.43196	-0.21384	0.41652	0.76697
	AD DATA	0.18323	0.47092	1.73757	0.81732
	AD GEO	0.63993	0.33136	0.25469	0.98958
Class A	GDDA	-2.37950	-0.51151	0.90775	0.93963
	RGFD	3723.38465	-1.60669	10.15264	0.57514
	DD	-3.85813	-0.42567	1.20449	0.95217
	DS	0.00024	0.79359	0.94487	0.01139
	CS	-1.19414	-0.51363	0.48532	0.11862
	CC DATA	1.45262	-0.54662	0.44763	0.85187
	CC GEO	0.52558	-0.32987	0.46379	0.94726
	AD DATA	0.58752	0.32954	0.64482	0.65354
	AD GEO	0.89578	0.29300	-0.26590	0.98199
Class B	GDDA	-1.28659	-0.14318	1.35182	0.93826
	RGFD	103.68002	-0.64369	7.25261	0.51435
	DD	-2.55698	-0.15931	1.89345	0.92415
	DS	-0.27778	-0.73719	0.97441	0.09980
	CS	0.00000	2.30896	0.38174	0.04630
	CC DATA	0.57862	-0.15441	0.25507	0.82357
	CC GEO	-0.00087	0.71667	0.59825	0.68834
	AD DATA	0.02537	0.75145	2.42600	0.64101
	AD GEO	0.17434	0.49691	1.49143	0.96078
Class C	GDDA	-1.39943	-0.32743	0.99518	0.87498
	RGFD	1.371E+13	-6.46140	10.47601	0.21543
	DD	-7.26862	-0.62686	1.06885	0.91158
	DS	0.00941	-0.38220	0.97351	0.00032
	CS	0.00131	0.70347	0.33768	0.03483
	CC DATA	10.58399	-1.08828	0.47959	0.71221
	CC GEO	0.41276	-0.20508	0.41598	0.93191
	AD DATA	0.39449	0.38867	0.72762	0.84718
	AD GEO	0.49040	0.36385	0.56849	0.98161
Class D	GDDA	-1.37701	-0.31108	1.01774	0.92754
	RGFD	1115.03720	-1.37919	9.56031	0.25267
	DD	-4.93768	-0.51017	1.13807	0.92221
	DS	-6331.66396	-3.28965	0.96929	0.01223
	CS	0.00008	1.14429	0.36063	0.04239
	CC DATA	2.00873	-0.69827	0.46556	0.71681
	CC GEO	0.47162	-0.12621	0.31449	0.86066
	AD DATA	0.10463	0.55671	1.87974	0.63692
	AD GEO	0.55392	0.34938	0.43300	0.96130

Figure D.34: The coefficients a, b, and c for the fitted power-law functions $a*x^b+c$ and R-Square values measuring the goodness of fit with respect to the following network properties: GDD-agreement (GDDA), RGF-distance (RGFD), agreements between degree distributions (DD), agreements between clustering spectra (CS), clustering coefficients of RIGs (CC DATA) and the corresponding GEO-3D model networks (CC GEO), average diameters of RIGs (AD DATA) and the corresponding GEO-3D model networks (AD GEO), and agreements between spectra of shortest path lengths (DS). The statistics are presented for the entire Data Set 2 of 1,272 RIGs, as well as for individual groups of proteins belonging to the four structural classes: $all-\alpha$ (class A), $all-\beta$ (class B), $\alpha\,/\,\beta$ (class C), and $\alpha+\beta$ (class D).

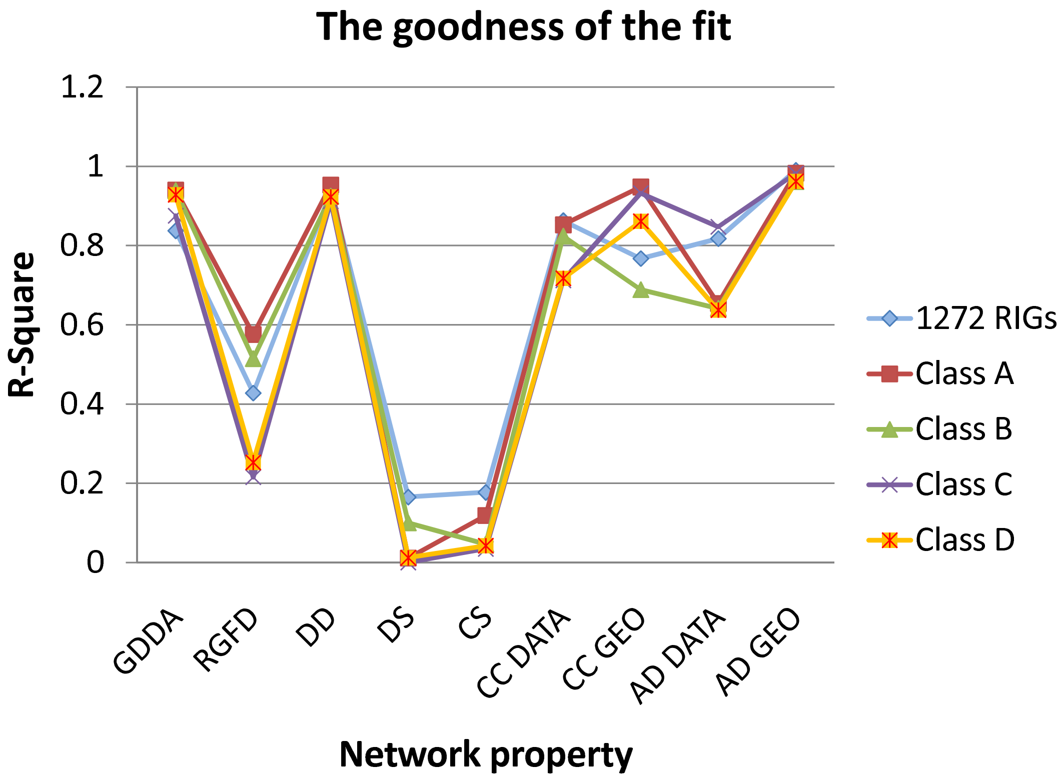


Figure D.35: The goodness of fit of the fitted power-law functions, expressed in terms of R-squares, for the following network properties: GDD-agreement (GDDA), RGF-distance (RGFD), agreements between degree distributions (DD), agreements between clustering spectra (CS), clustering coefficients of RIGs (CC DATA) and the corresponding GEO-3D model networks (CC GEO), average diameters of RIGs (AD DATA) and the corresponding GEO-3D model networks (AD GEO), and agreements between spectra of shortest path lengths (DS). The entire Data Set 2 of 1,272 RIGs was analysed, as well as for individual groups of proteins belonging to the four structural classes: $all-\alpha$ (A), $all-\beta$ (B), $\alpha\,/\,\beta$ (C), and $\alpha+\beta$ (D).

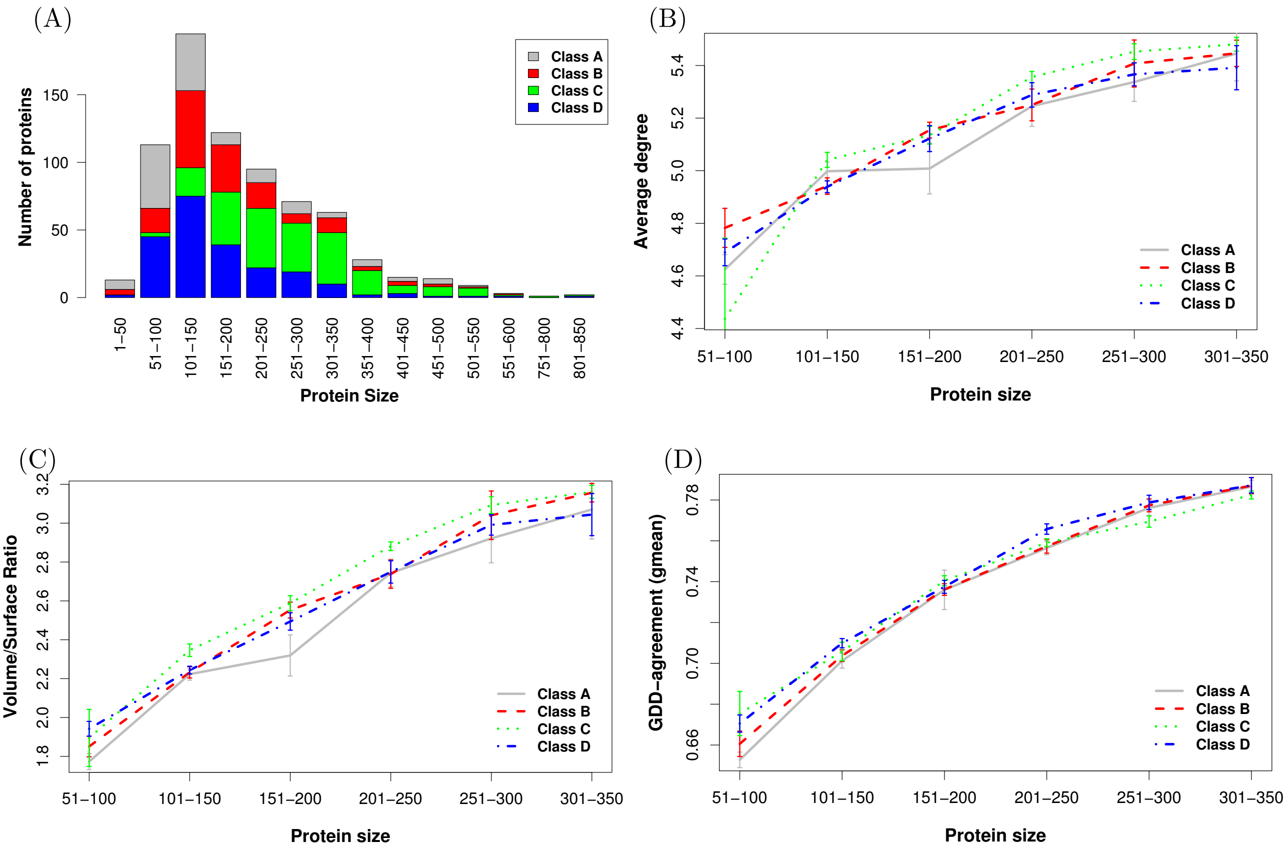
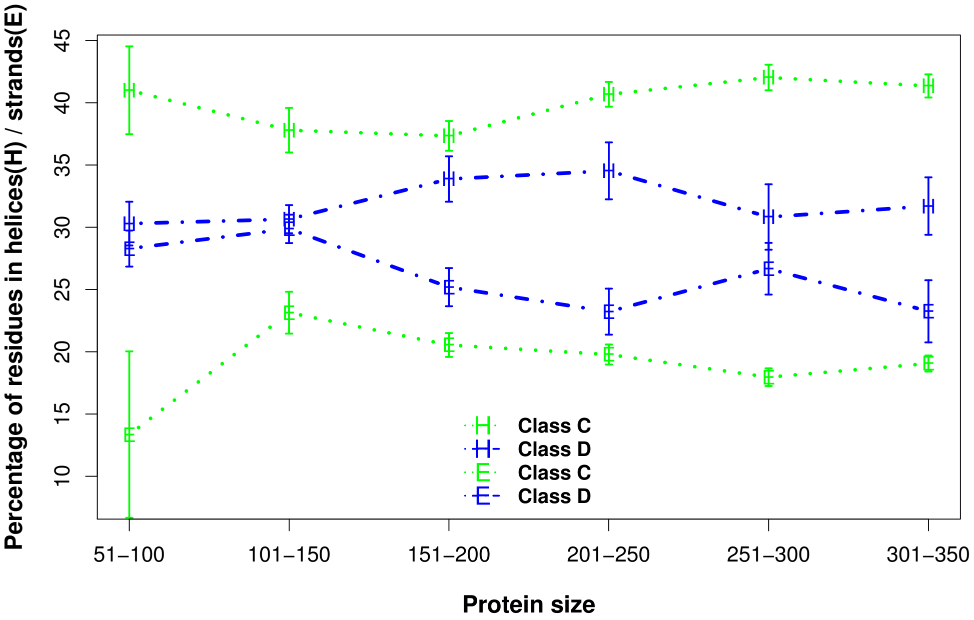


Figure D.36: (A) Distribution of protein size for 744 proteins from Data Set 2 that belong to one of the four structural classes: $all-\alpha$ (A), $all-\beta$ (B), $\alpha\,/\,\beta$ (C), and $\alpha+\beta$ (D). (B) Average degree, (C) average volume-to-surface ratio, and (D) average GDD-agreement between GEO-3D graphs and the corresponding RIGs, with respect to each size range and each of the four structural classes. The standard error of the mean is plotted.

(A)



(B)

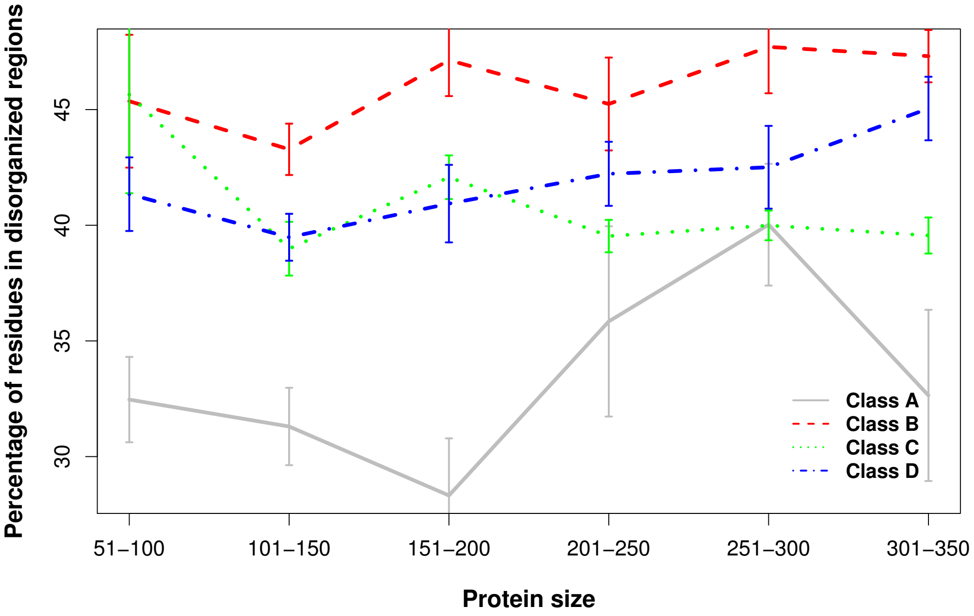


Figure D.37: (A) Average percentage of residues in α-helices and β-strands for all $\alpha\,/\,\beta$ (C), and $\alpha+\beta$ (D) proteins in Data Set 2, with respect to each protein size range. (B) Average percentage of residues in loop regions for 744 proteins in Data Set 2 belonging to the four structural classes, $all-\alpha$ (A), $all-\beta$ (B), $\alpha\,/\,\beta$ (C), and $\alpha+\beta$ (D), with respect to each size range and each class. The standard error of the mean is plotted.

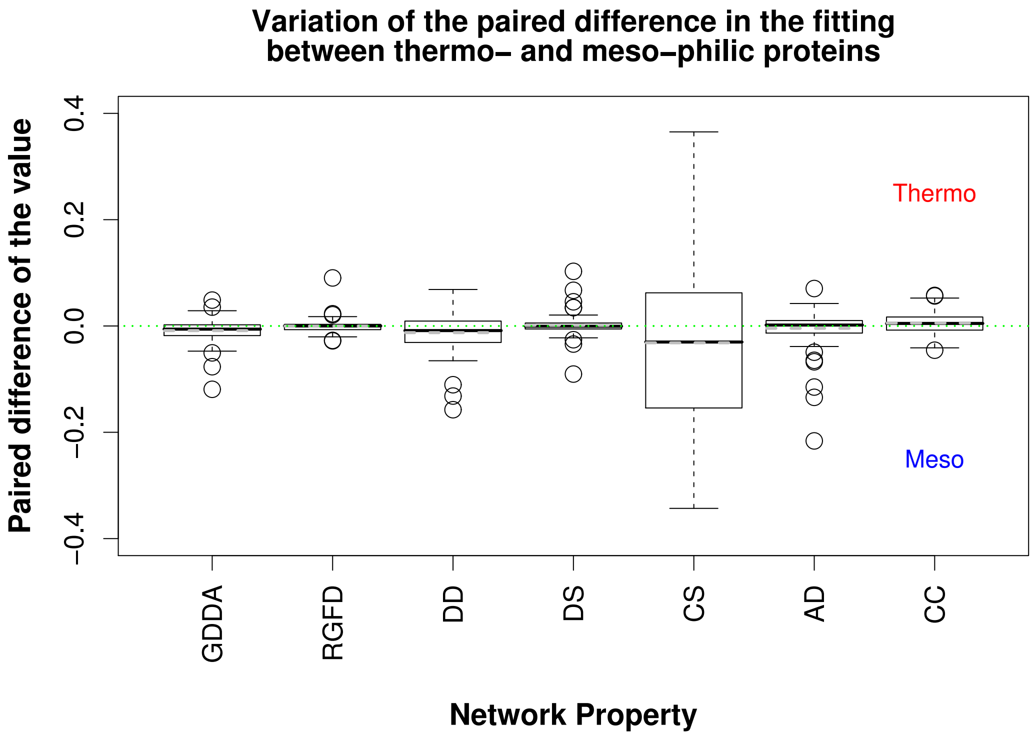


Figure D.38: Boxplots of the paired difference, between thermophilic and mesophilic proteins, in the degree of fitting of GEO-3D graphs to RIGs, with respect to following network properties: GDD-agreement (GDDA), RGF-distance (RGFD), agreements between degree distributions (DD), agreements between spectra of shortest path lengths (DS), agreements between clustering spectra (CS), agreements between average diameters (AD), and agreements between clustering coefficients (CC). Grey dashed lines denote the mean, while values above (below) the green dotted line denote that the value is higher for thermophiles (mesophiles). The values of agreements between RGF-distances, average diameters and clustering coefficients have been scaled to range from zero to one.

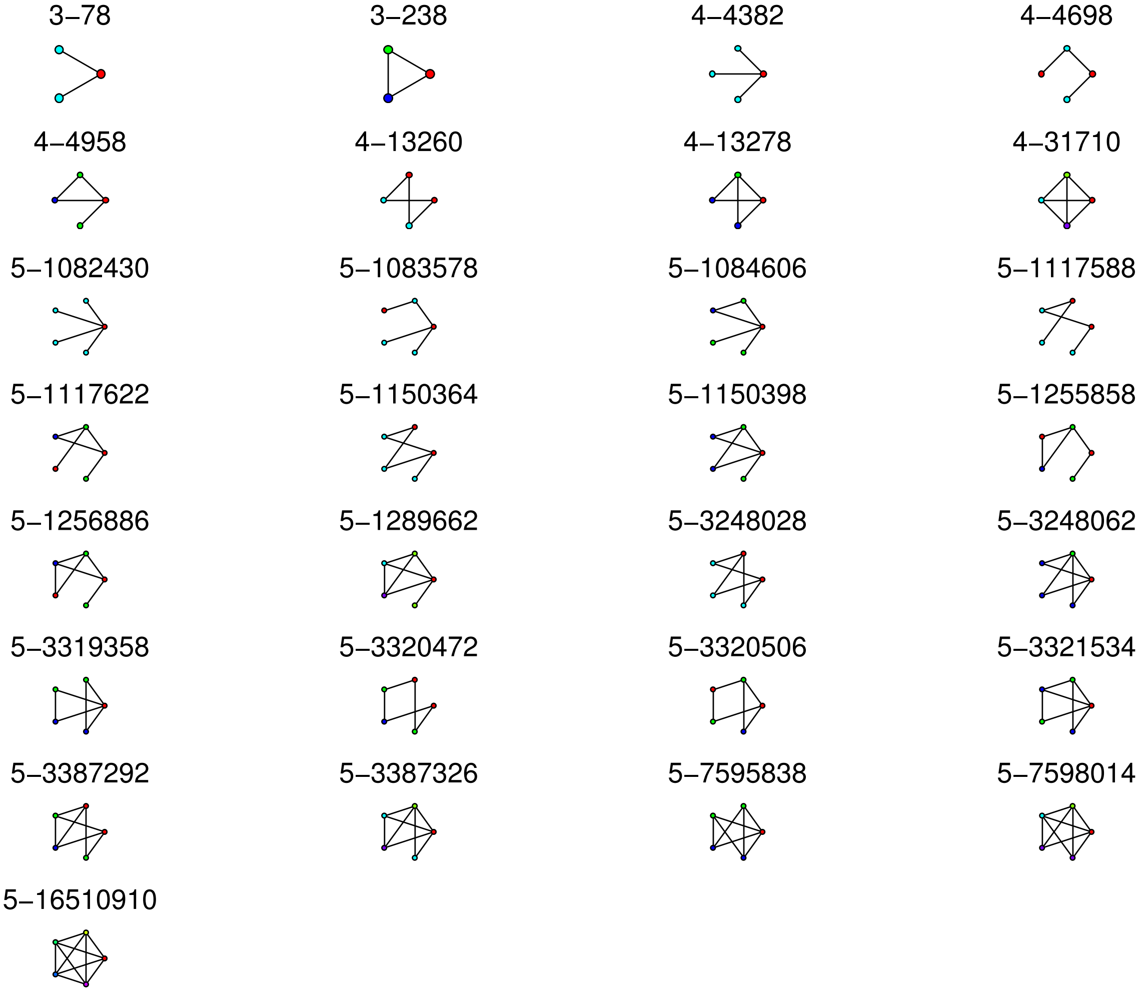
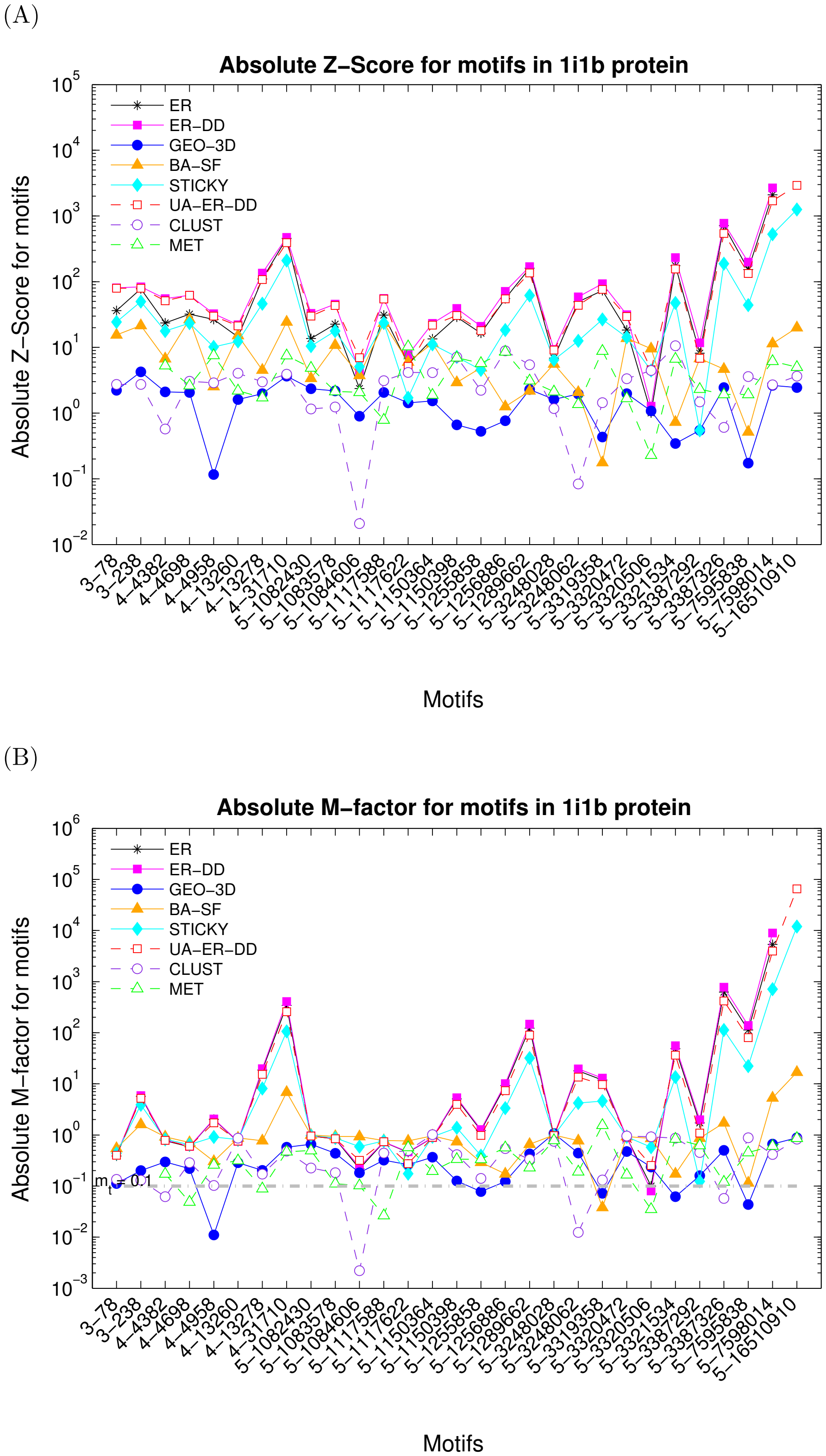


Figure D.39: Motif dictionary: all 3- to 5-node subgraphs. Each subgraph is labelled with its size followed by its ID according to mfinder.


(A)

(B)

Figure D.40: Absolute (A) Z-scores and (B) M-factors for all 3- to 5-node subgraphs in the $(ALL)^{\text{all}}_{5.0\mathring{A}}$ RIG corresponding to 1i1b protein. These statistics were computed with respect to eight network models (ER, ER-DD, GEO-3D, BA-SF, STICKY, UA-ER-DD, CLUST, and MET). Y-axis is shown in a logarithmic scale to facilitate the comparison of different models. The threshold value used for motif selection (M-factor greater than 0.1) is displayed as the grey dash-dot line.

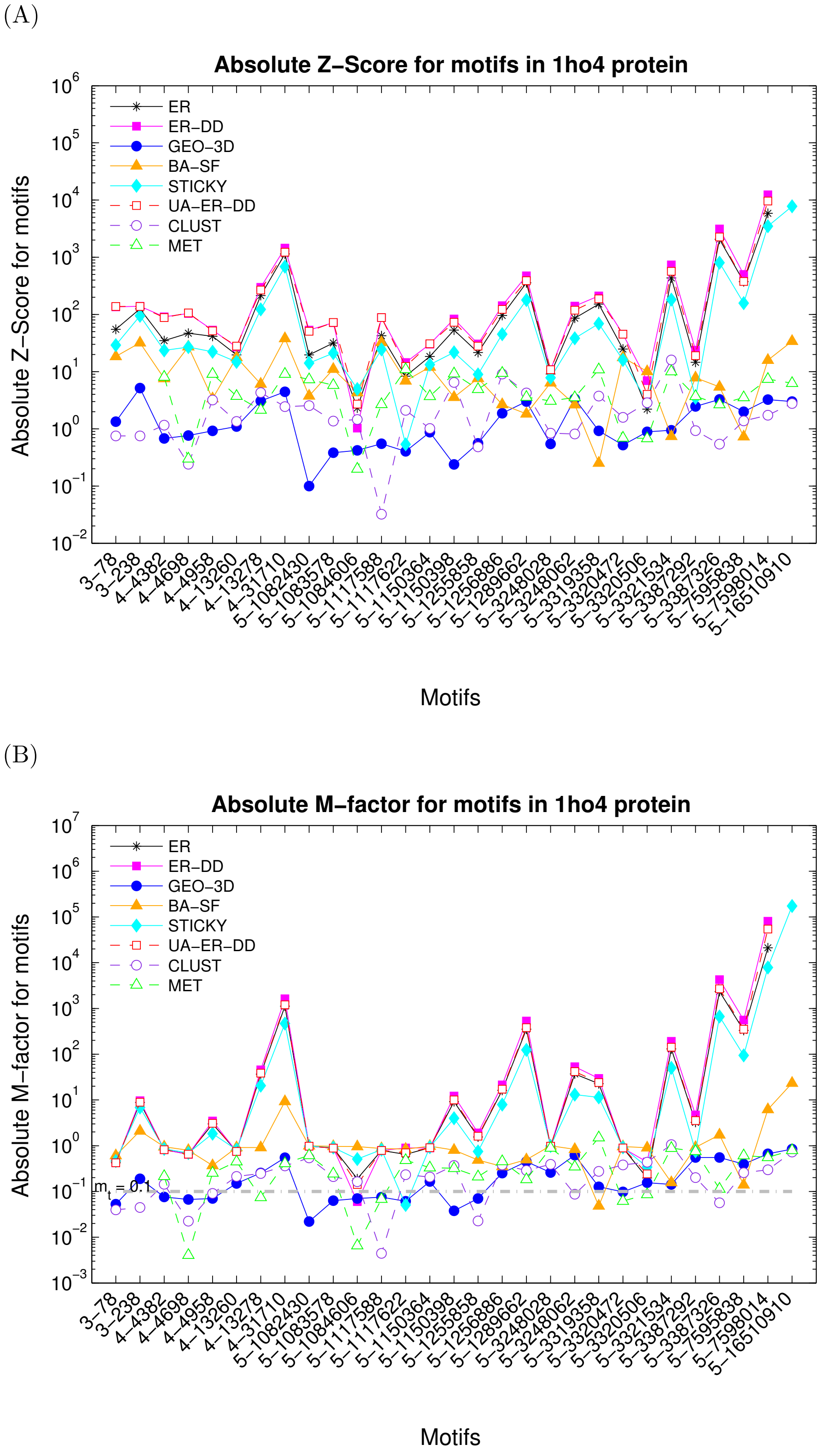


Figure D.41: Absolute (A) Z-scores and (B) M-factors for all 3- to 5-node subgraphs in the $(ALL)^{\text{all}}_{5.0\text{Å}}$ RIG corresponding to 1ho4 protein. These statistics were computed with respect to eight network models (ER, ER-DD, GEO-3D, BA-SF, STICKY, UA-ER-DD, CLUST, and MET). Y-axis is shown in a logarithmic scale to facilitate the comparison of different models. The threshold value used for motif selection (M-factor greater than 0.1) is displayed as the grey dash-dot line.

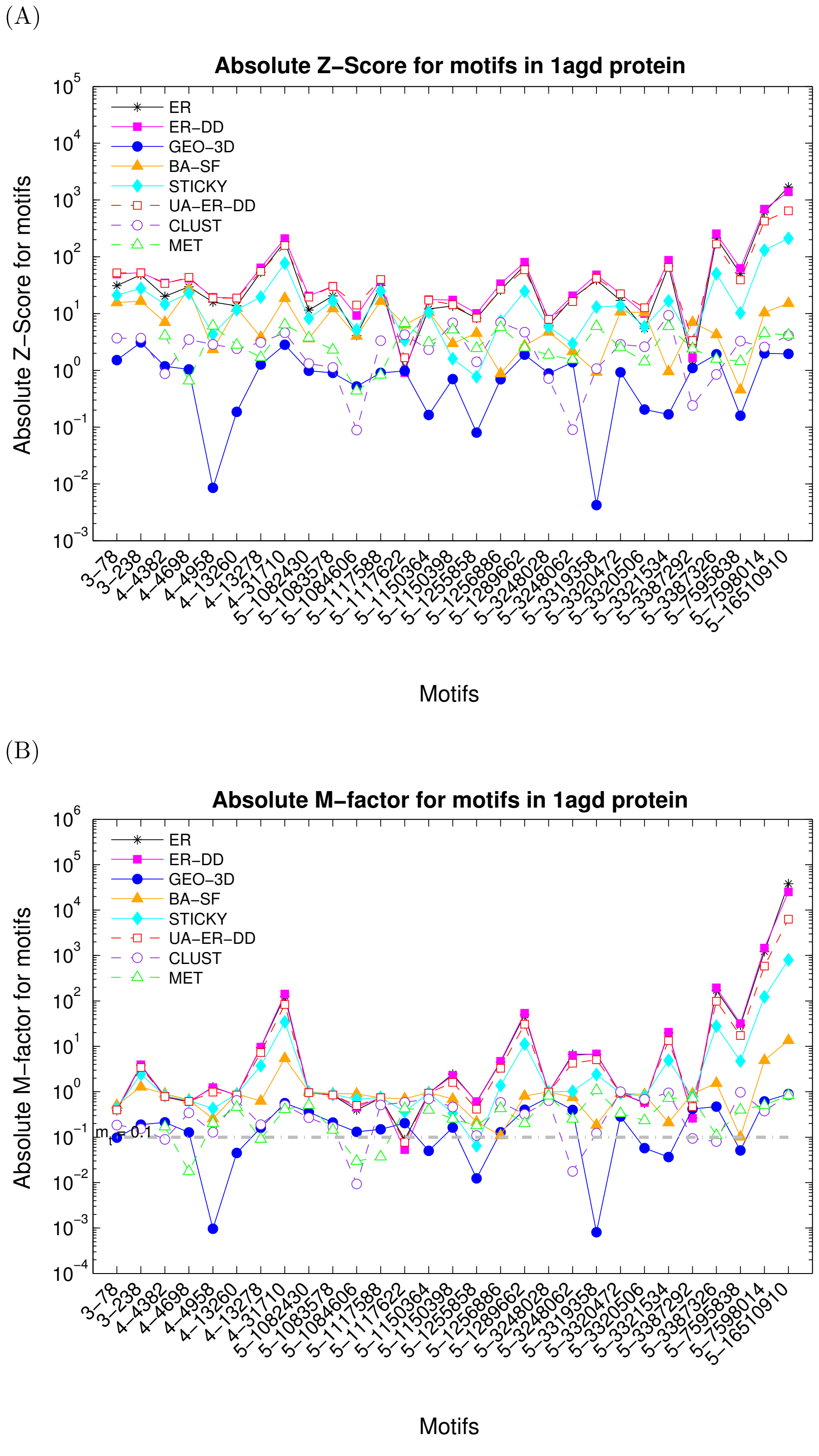
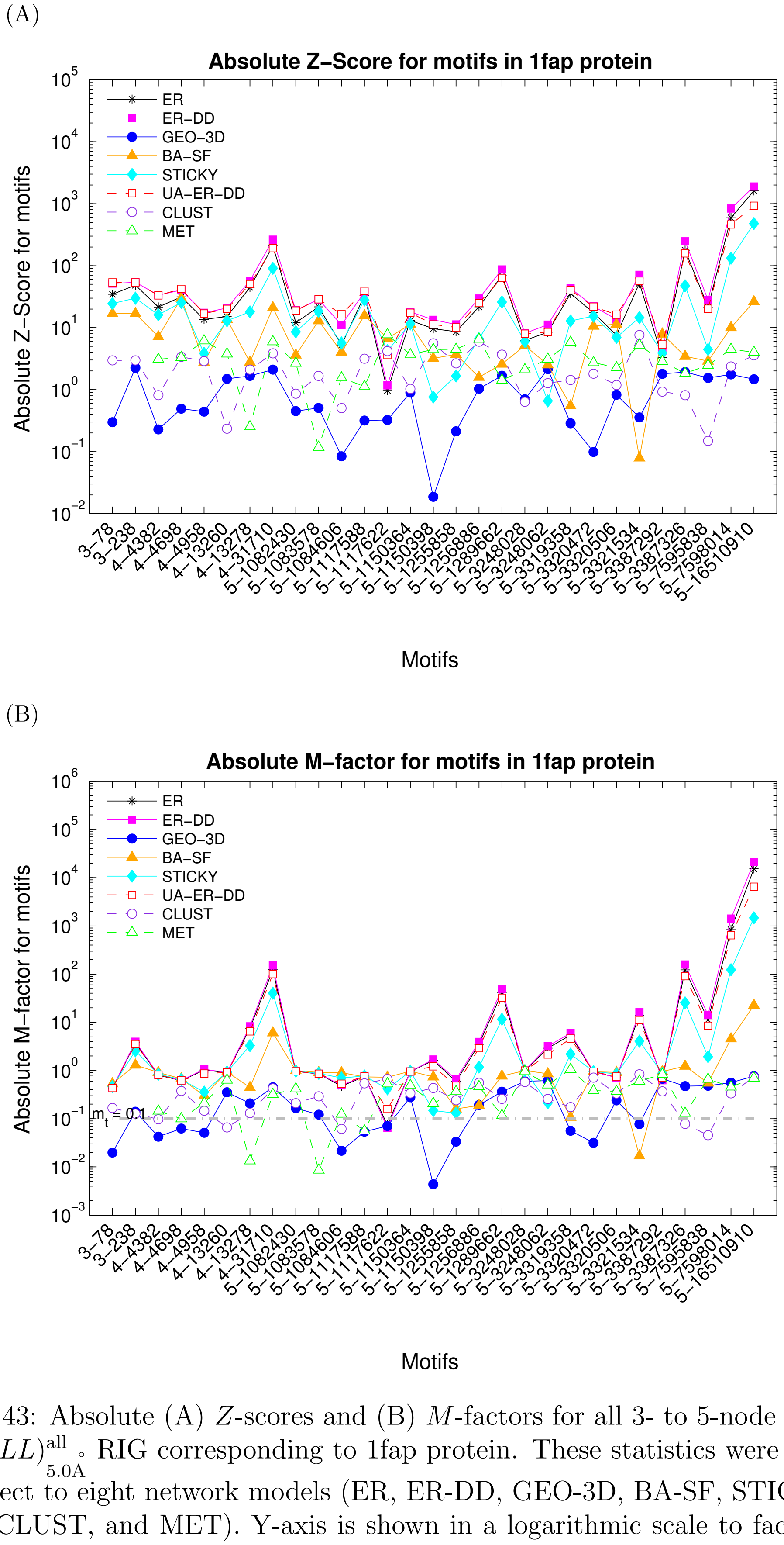


Figure D.42: Absolute (A) Z-scores and (B) M-factors for all 3- to 5-node subgraphs in the $(ALL)^{\mathrm{all}}_{5.0\mathring{\mathrm{A}}}$ RIG corresponding to 1agd protein. These statistics were computed with respect to eight network models (ER, ER-DD, GEO-3D, BA-SF, STICKY, UA-ER-DD, CLUST, and MET). Y-axis is shown in a logarithmic scale to facilitate the comparison of different models. The threshold value used for motif selection (M-factor greater than 0.1) is displayed as the grey dash-dot line.


(A)

(B)

Figure D.43: Absolute (A) Z-scores and (B) M-factors for all 3- to 5-node subgraphs in the $(ALL)^{\mathrm{all}}_{5.0\mathring{A}}$ RIG corresponding to 1fap protein. These statistics were computed with respect to eight network models (ER, ER-DD, GEO-3D, BA-SF, STICKY, UA-ER-DD, CLUST, and MET). Y-axis is shown in a logarithmic scale to facilitate the comparison of different models. The threshold value used for motif selection (M-factor greater than 0.1) is displayed as the grey dash-dot line.

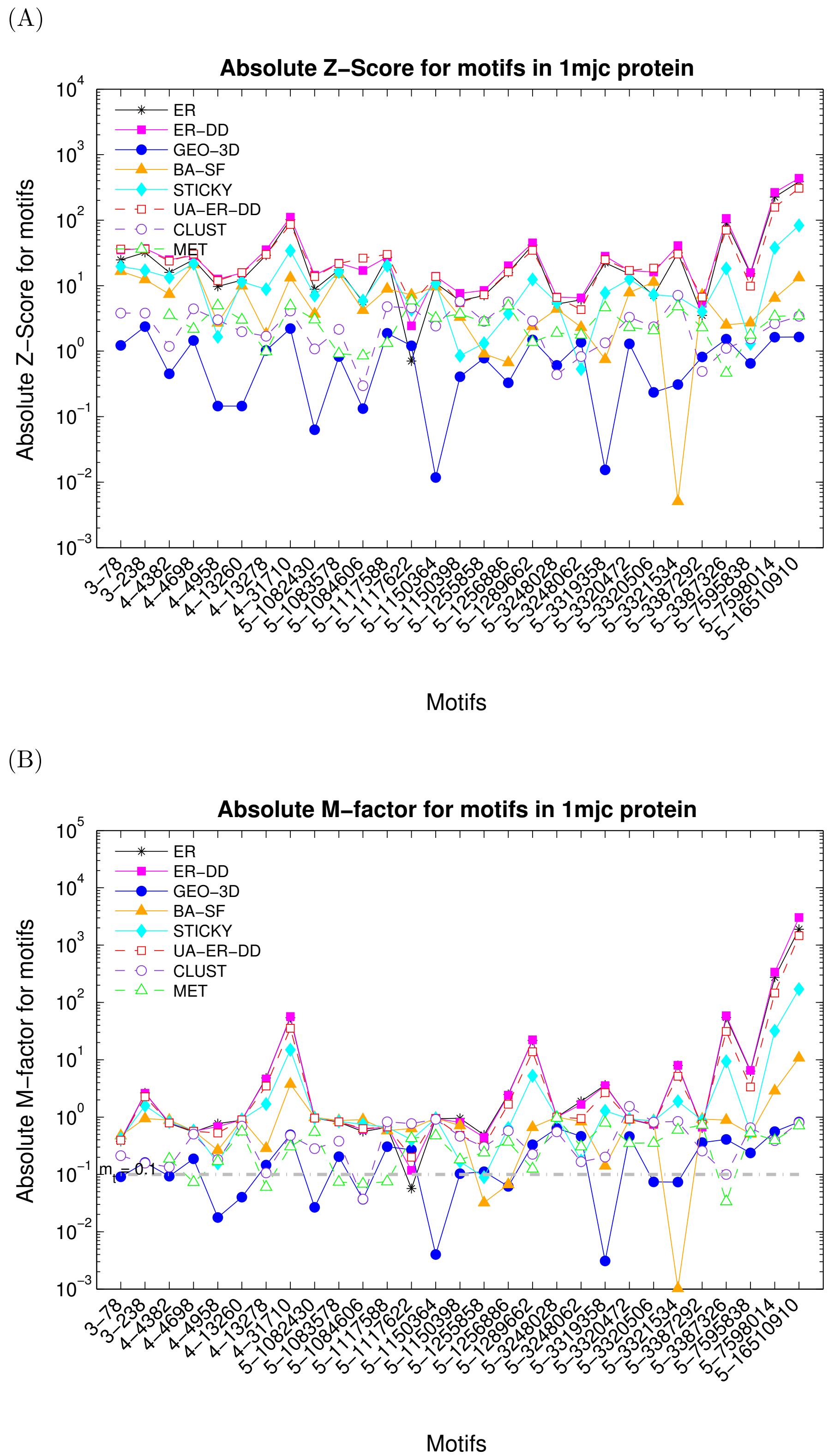


Figure D.44: Absolute (A) Z-scores and (B) M-factors for all 3- to 5-node subgraphs in the $(ALL)^{\text{all}}_{5.0\text{Å}}$ RIG corresponding to 1mjc protein. These statistics were computed with respect to eight network models (ER, ER-DD, GEO-3D, BA-SF, STICKY, UA-ER-DD, CLUST, and MET). Y-axis is shown in a logarithmic scale to facilitate the comparison of different models. The threshold value used for motif selection (M-factor greater than 0.1) is displayed as the grey dash-dot line.

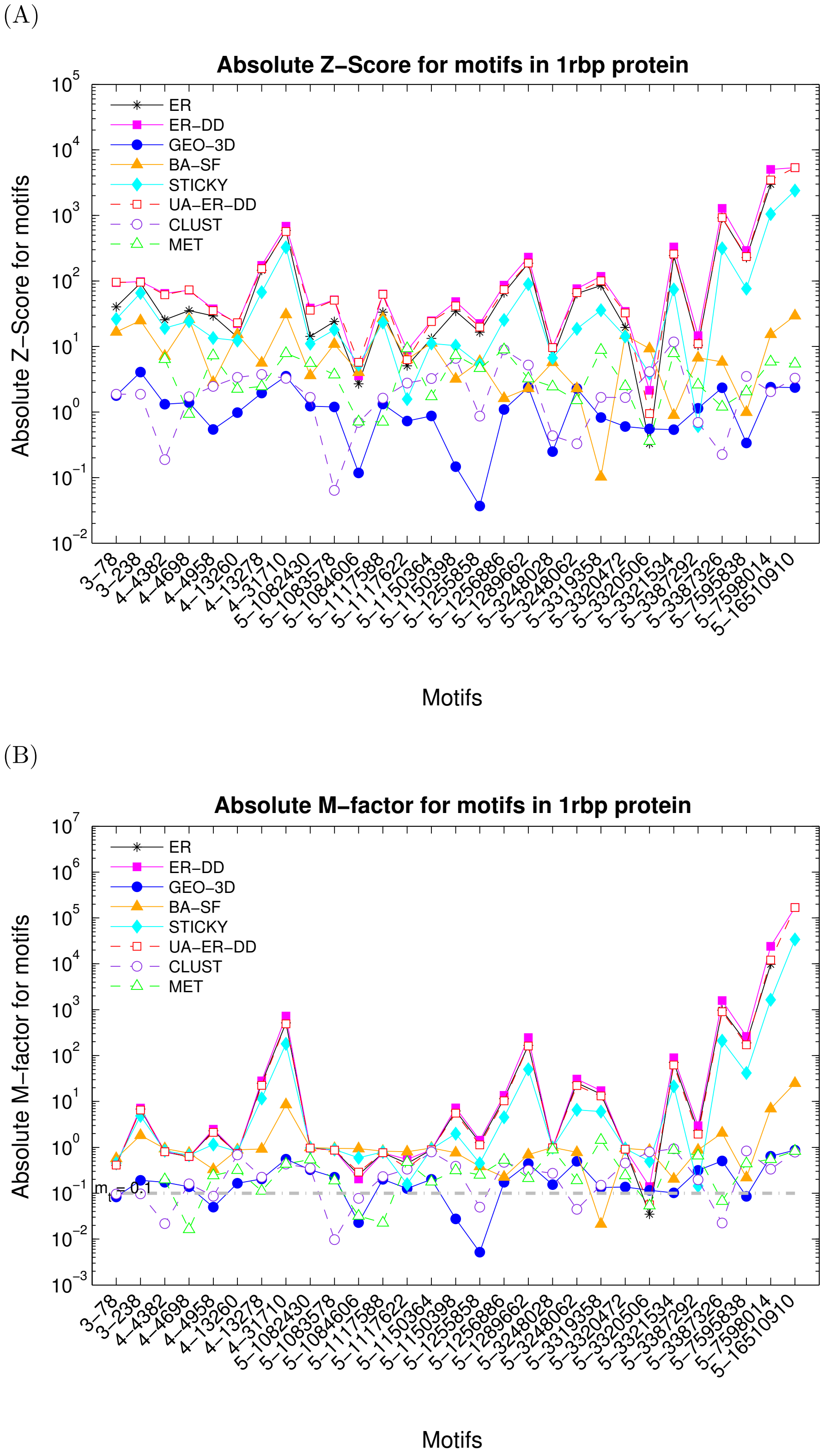
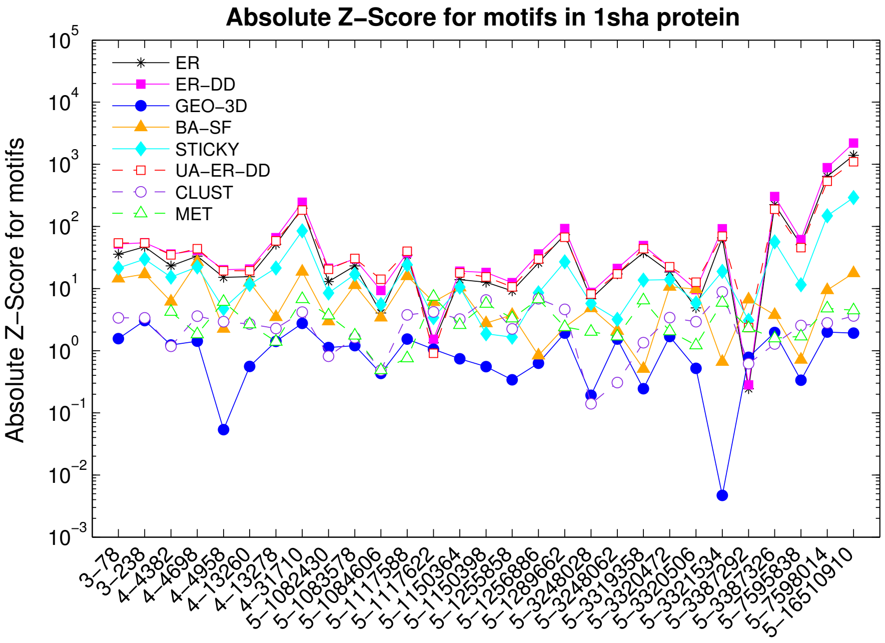


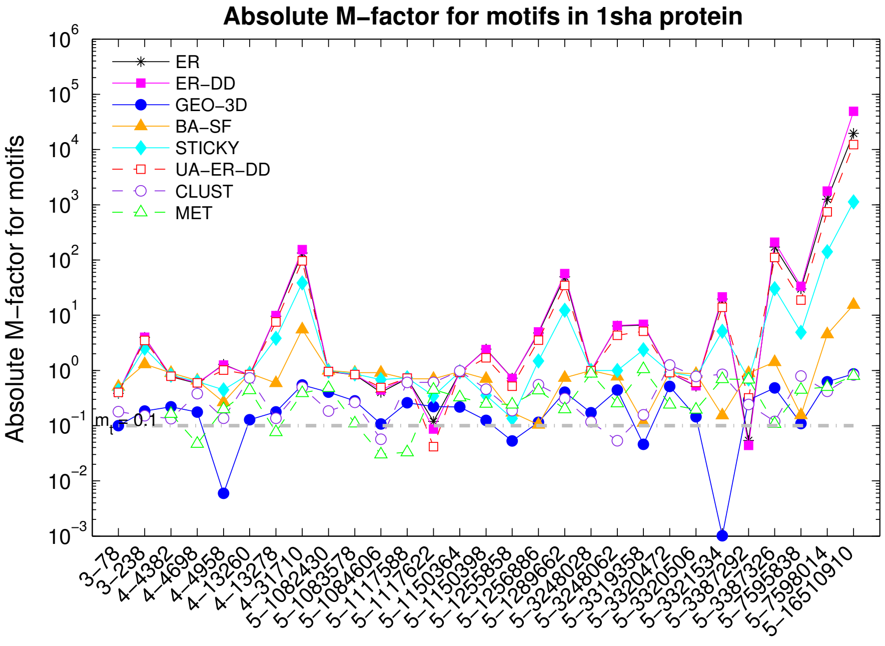
Figure D.45: Absolute (A) Z-scores and (B) M-factors for all 3- to 5-node subgraphs in the $(ALL)^{\text{all}}_{5.0\text{Å}}$ RIG corresponding to 1rbp protein. These statistics were computed with respect to eight network models (ER, ER-DD, GEO-3D, BA-SF, STICKY, UA-ER-DD, CLUST, and MET). Y-axis is shown in a logarithmic scale to facilitate the comparison of different models. The threshold value used for motif selection (M-factor greater than 0.1) is displayed as the grey dash-dot line.

(A)



Motifs

(B)



Motifs

Figure D.46: Absolute (A) Z-scores and (B) M-factors for all 3- to 5-node subgraphs in the $(ALL)^{\text{all}}_{5.0\text{A}^{\circ}}$ RIG corresponding to 1sha protein. These statistics were computed with respect to eight network models (ER, ER-DD, GEO-3D, BA-SF, STICKY, UA-ER-DD, CLUST, and MET). Y-axis is shown in a logarithmic scale to facilitate the comparison of different models. The threshold value used for motif selection (M-factor greater than 0.1) is displayed as the grey dash-dot line.

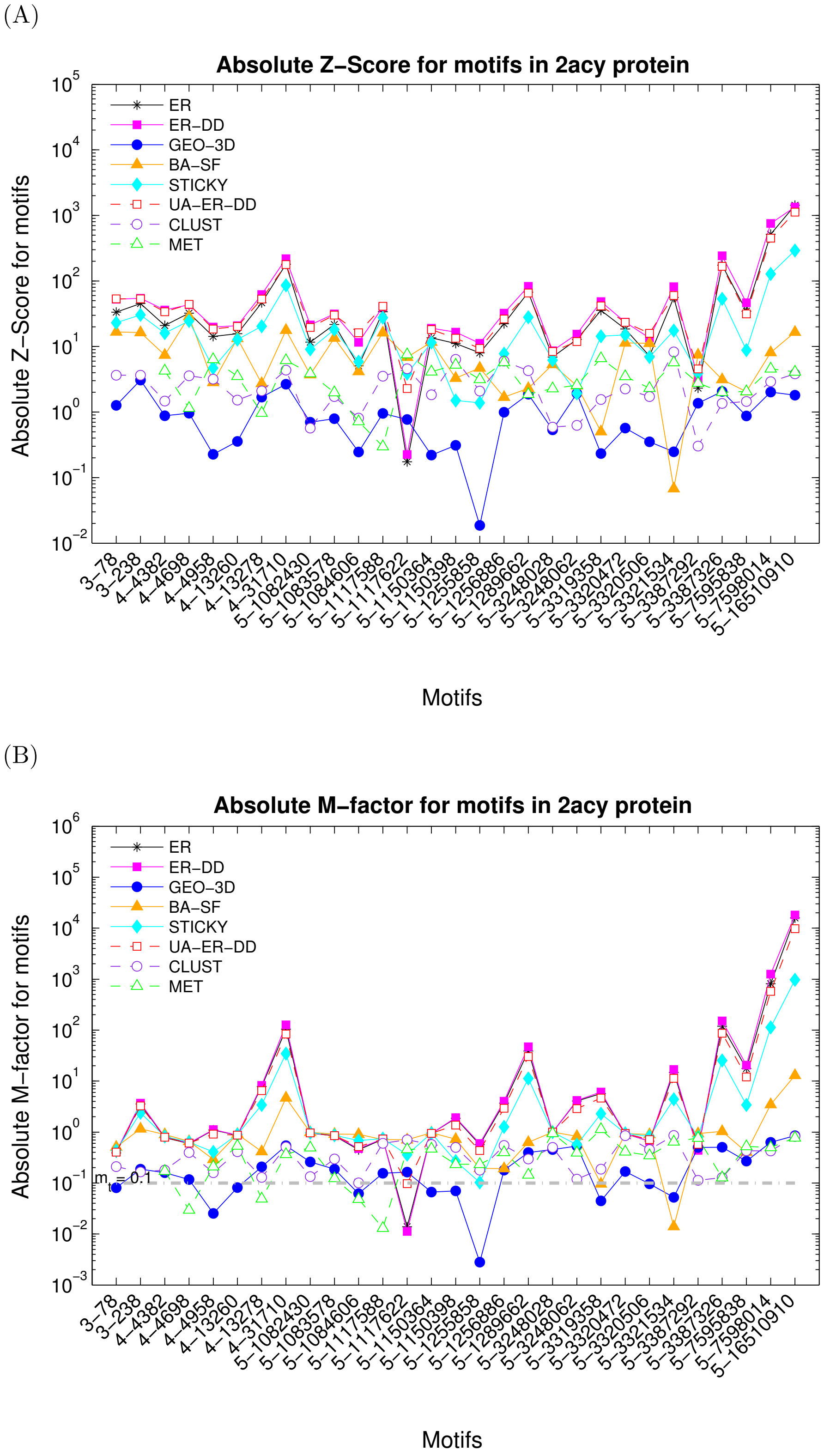


Figure D.47: Absolute (A) Z-scores and (B) M-factors for all 3- to 5-node subgraphs in the $(ALL)^{\text{all}}_{5.0\text{A}^{\circ}}$ RIG corresponding to 2acy protein. These statistics were computed with respect to eight network models (ER, ER-DD, GEO-3D, BA-SF, STICKY, UA-ER-DD, CLUST, and MET). Y-axis is shown in a logarithmic scale to facilitate the comparison of different models. The threshold value used for motif selection (M-factor greater than 0.1) is displayed as the grey dash-dot line.

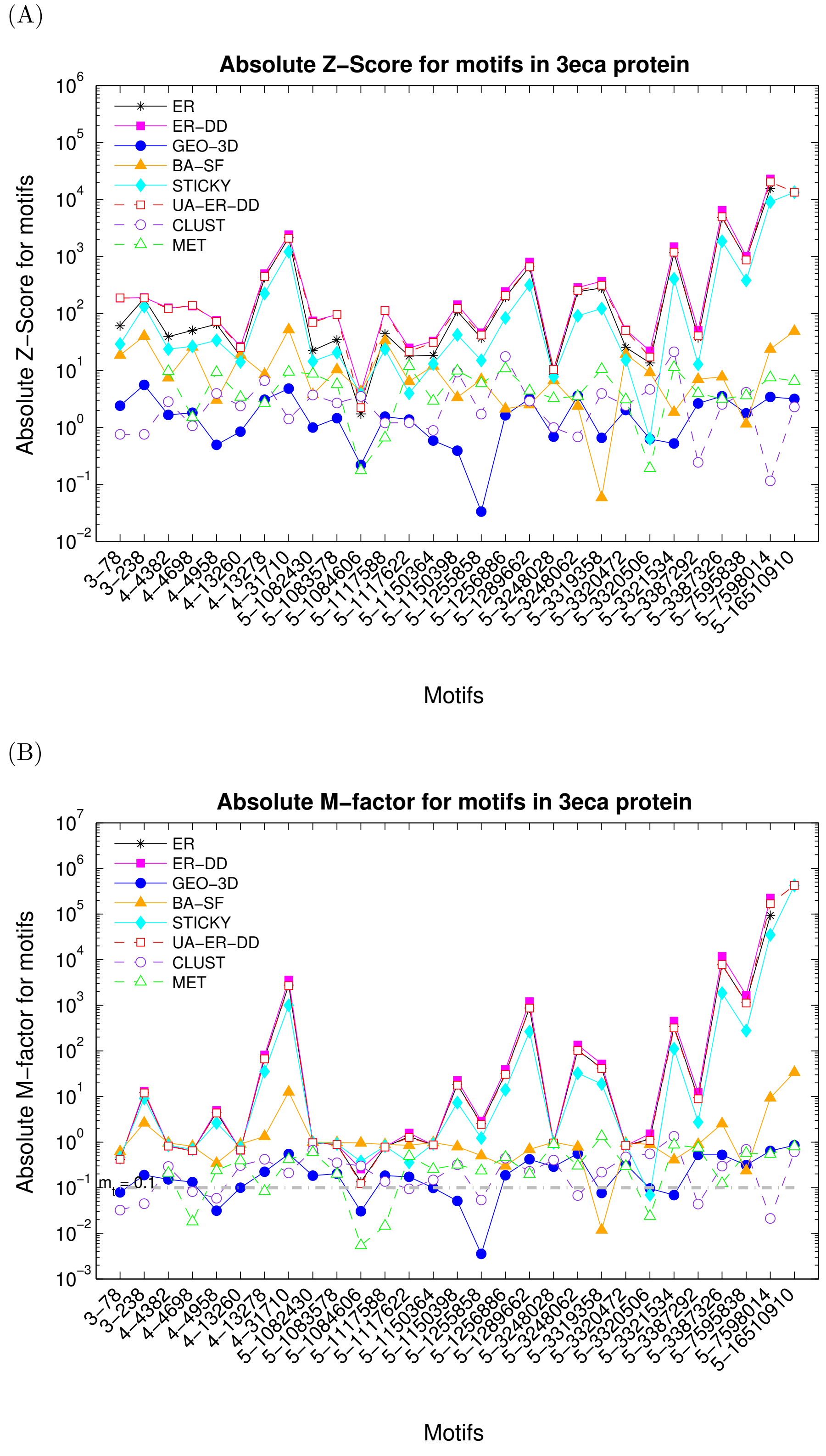


Figure D.48: Absolute (A) Z-scores and (B) M-factors for all 3- to 5-node subgraphs in the $(ALL)^{\text{all}}_{5.0\mathring{A}}$ RIG corresponding to 3eca protein. These statistics were computed with respect to eight network models (ER, ER-DD, GEO-3D, BA-SF, STICKY, UA-ER-DD, CLUST, and MET). Y-axis is shown in a logarithmic scale to facilitate the comparison of different models. The threshold value used for motif selection (M-factor greater than 0.1) is displayed as the grey dash-dot line.

Appendix E

Zusammenfassung

Zusammenfassung

Netzwerkanalysen von Proteinstrukturen erlauben wichtige Einblicke in Proteinfaltung und -funktion. Da bisher eine einheitliche Sichtweise auf die Netwerkmodellierung und Analyse von Proteinstrukturen fehlt und neuere Erkenntnisse der Netzwerktheorie bislang unberücksichtigt blieben, werden in dieser Arbeit die folgenden wichtigen Ziele bearbeitet:

1. Die rationale Auswahl geeigneter Netzwerkdarstellungen von Proteinstrukturen
2. Die Ausarbeitung eines optimierten Nullmodells für Proteinstrukturnetzwerke
3. Die Entwicklung einer neuen graphenbasierten empirischen Potentialfunktion

Die Theorie der Graphlets, ein kürzlich eingeführtes, mächtiges Konzept in der Graphentheorie bildet die Grundlage dieser Arbeit. Mit Hilfe der Graphlets werden die topologischen Ähnlichkeiten verschiedener Netwerkdarstellungen untersucht. Dies führt zu einem optimierten Nullmodell und schließlich zu einer neuen empirischen Potentialfunktion.

Kapitel 2 vereint verschiedene Netzwerkdarstellungen über ein kontrolliertes Vokabular. Dabei werden die Details der Netzwerk-Konstruktion motiviert sowie deren Popularität und Optimalität erläutert. In Kapitel 3 wird ein umfassender Satz von insgesamt 945 verschiedenen Netzwerkdarstellungen systematisch hinsichtlich ihrer Ähnlichkeit und ihrer grundlegenden Netzwerkeigenschaften analysiert. Es wird gezeigt, dass verschiedene häufig verwendete Darstellungen eine geringe Ähnlichkeit zueinander aufweisen. Zudem tauchen in einigen Darstellungen mehrere Zusammenhangskomponenten und nichtverbundene Knoten auf. Insbesondere Vergleiche zwischen Proteinen mit unterschiedlichen Sekundärstrukturtopologien sollten mit Vorsicht gezogen werden. Dieser Teil der Arbeit legt die Grundlagen für eine rationale Auswahl nach Kriterien wie Häufigkeit, Optimalität wünschenswerter Netzwerkeigenschaften sowie Ähnlichkeit zu bereits erfolgreich eingesetzten Darstellungen.

In Kapitel 4 wird gezeigt, dass unter einer Reihe von Zufallsgraphmodellen die three-dimensionalen geometrischen Zufalls-Graphen am besten den Eigenschaften von Proteinstrukturnetzwerken entsprechen. Die Übereinstimmung, gemessen an einem strukturell diversen Datensatz, bleibt unter den verschiedensten Darstellungen und den verschiedenen topologischen Eigenschaften erhalten. Geometrische Zufallsgraphen entsprechen in ihrer Netzwerkstruktur am ehesten großen Proteinen, Strukturen mit einem geringen Anteil an alpha-helices oder solchen mit geringer Thermostabilität. Die Wahl geometrischer Zufalls-Graphen als Nullmodell erlaubt die sehr spezifische Identifikation statistisch signifikanter Teilgraphen. Dieser Teil der Arbeit wurde bereits erfolgreich publiziert.

In Kapitel 5 wird eine neue empirische Potentialfunktion entwickelt, indem die Kontaktzahl als Potentialfunktion in eine rein topologische und residuen-basierte Form verallgemeinert und verbessert wird. Die verbesserten Eigenschaften sind konsistent und robust gegenüber verschiedenen Methoden zur Generierung von Decoys und verschiedenen Qualitätsmaßen. Die Ergebnisse liegen insgesamt etwa gleich auf mit denen vorhandener Vier-Körper-Potentiale, verhalten sich jedoch im Einzelfall mitunter stark komplementär zueinander. Dies deutet auf weiteres Entwicklungspotential hin.

Insgesamt werden mit dieser Arbeit die Grundlagen für die systematische Analyse von Proteinstrukturen als Netzwerke gelegt und neue Ansatzmöglichkeiten für die Suche nach einer optimalen Energiefunktion eröffnet.

Bibliography

[1] M. Aftabuddin and S. Kundu. Weighted and unweighted network of amino acids within protein. *Physica A: Statistical Mechanics and its Applications*, 369(2):895 – 904, 2006.

[2] N. Alexandrov and I. Shindyalov. Pdp: protein domain parser. *Bioinformatics*, 19(3):429–30, Feb 2003.

[3] N. N. Alexandrov, R. Nussinov, and R. M. Zimmer. Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. *Pacific Symposium on Biocomputing*, pages 53–72, Jan 1996.

[4] N. L. Allinger. Conformational analysis. 130. mm2. a hydrocarbon force field utilizing v1 and v2 torsional terms. *Journal of the American Chemical Society*, 99(25):8127–8134, 1977.

[5] N. L. Allinger, Y. H. Yuh, and J. H. Lii. Molecular mechanics. the mm3 force field for hydrocarbons. 1. *Journal of the American Chemical Society*, 111(23):8551–8566, 1989.

[6] U. Alon, M. G. Surette, N. Barkai, and S. Leibler. Robustness in bacterial chemotaxis. *Nature*, 397(6715):168–171, Jan 1999.

[7] N. A. Alves and A. S. Martinez. Inferring topological features of proteins from amino acid residue networks. *Physica A: Statistical Mechanics and its Applications*, 375(1):336 – 344, 2007.

[8] G. Amitai, A. Shemesh, E. Sitbon, M. Shklar, D. Netanely, I. Venger, and S. Pietrokovski. Network analysis of protein structures identifies functional residues. *J Mol Biol*, 344(4):1135–46, Dec 2004.

[9] C. B. Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(96):223–230, Jul 1973.

[10] Y. Artzy-Randrup, S. J. Fleishman, N. Ben-Tal, and L. Stone. Comment on “Network motifs: Simple building blocks of complex networks” and “Superfamilies of evolved and designed networks”. *Science*, 305:1107c, 2004.

[11] A. Aszódi, M. J. Gradwell, and W. R. Taylor. Global fold determination from a small number of distance restraints. *J Mol Biol*, 251(2):308–26, Aug 1995.

[12] A. R. Atilgan, P. Akan, and C. Baysal. Small-world communication of residues and significance for protein dynamics. *Biophys J*, 86(1 Pt 1):85–91, Jan 2004.

[13] A. R. Atilgan, D. Turgut, and C. Atilgan. Screened nonbonded interactions in native proteins manipulate optimal paths for robust residue communication. *Biophys J*, 92(9):3052–62, May 2007.

[14] G. Bagler and S. Sinha. Network properties of protein structures. *Physica A: Statistical Mechanics and its Applications*, 346(1-2):27 – 33, 2005.

[15] G. Bagler and S. Sinha. Assortative mixing in protein contact networks and protein folding kinetics. *Bioinformatics*, 23(14):1760–7, Jul 2007.

[16] I. Bahar, A. R. Atilgan, and B. Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des*, 2(3):173–181, 1997.

[17] I. Bahar and R. L. Jernigan. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J Mol Biol*, 266(1):195–214, Feb 1997.

[18] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[19] A.-L. Barabási and Z. N. Oltvai. Network biology: Understanding the cell's functional organization. *Nature Reviews*, 5:101–113, 2004.

[20] N. Barkai and S. Leibler. Robustness in simple biochemical networks. *Nature*, 387(6636):913–917, Jun 1997.

[21] L. Bartoli, P. Fariselli, and R. Casadio. The effect of backbone on the small-world properties of protein contact maps. *Phys Biol*, 4(4):L1–L5, Dec 2007.

[22] U. Bastolla and L. Demetrius. Stability constraints and protein evolution: the role of chain length, composition and disulfide bonds. *Protein Eng Des Sel*, 18(9):405–15, Sep 2005.

[23] U. Bastolla, J. Farwer, E. W. Knapp, and M. Vendruscolo. How to guarantee optimal stability for most representative structures in the protein data bank. *Proteins*, 44(2):79–96, Aug 2001.

[24] U. Bastolla, A. R. Ortíz, M. Porto, and F. Teichert. Effective connectivity profile: a structural representation that evidences the relationship between protein structures and sequences. *Proteins*, 73(4):872–88, Dec 2008.

[25] U. Bastolla, M. Porto, H. E. Roman, and M. Vendruscolo. Principal eigenvector of contact matrices and hydrophobicity profiles in proteins. *Proteins*, 58(1):22–30, Jan 2005.

[26] A. Bavelas. A mathematical model for group structure. *Applied Anthropology*, 7:16–30, 1948.

[27] P. Benkert, S. C. E. Tosatto, and D. Schomburg. Qmean: A comprehensive scoring function for model quality assessment. *Proteins*, 71(1):261–77, Apr 2008.

[28] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig,

I. Shindyalov, and P. Bourne. The protein data bank. *Nucleic Acids Research*, 28:235–242, 2000.

[29] H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook, and C. Zardecki. The protein data bank. *Acta Crystallogr D Biol Crystallogr*, 58(Pt 6 No 1):899–907, 2002. 0907-4449 Journal Article.

[30] M. Berrera, H. Molinari, and F. Fogolari. Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. *BMC Bioinformatics*, 4:8, Feb 2003.

[31] M. R. Betancourt and D. Thirumalai. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci*, 8(2):361–369, Feb 1999.

[32] J. D. Bloom, D. A. Drummond, F. H. Arnold, and C. O. Wilke. Structural determinants of the rate of protein evolution in yeast. *Molecular Biology and Evolution*, 23(9):1751–61, Sep 2006.

[33] J. Bohr, H. Bohr, S. Brunak, R. M. Cotterill, H. Fredholm, B. Lautrup, and S. B. Petersen. Protein structures from distance inequalities. *J Mol Biol*, 231(3):861–9, Jun 1993.

[34] D. M. Bolser, I. Filippis, H. Stehr, J. Duarte, and M. Lappe. Residue contact-count potentials are as effective as residue-residue contact-type potentials for ranking protein decoys. *BMC Struct Biol*, 8:53, Jan 2008.

[35] R. Bonneau, I. Ruczinski, J. Tsai, and D. Baker. Contact order and ab initio protein structure prediction. *Protein Sci*, 11(8):1937–44, Aug 2002.

[36] H. Boutselakis, D. Dimitropoulos, J. Fillon, A. Golovin, K. Henrick, A. Hussain, J. Ionides, M. John, P. A. Keller, E. Krissinel, P. McNeil, A. Naim, R. Newman, T. Oldfield, J. Pineda, A. Rachedi, J. Copeland, A. Sitnov, S. Sobhany, A. Suarez-Uruena, J. Swaminathan, M. Tagari, J. Tate, S. Tromm, S. Velankar, and W. Vranken. E-msd: the european bioinformatics institute macromolecular structure database. *Nucleic Acids Res*, 31(1):458–462, Jan 2003.

[37] G. R. Bowman and V. S. Pande. Simulated tempering yields insight into the low-resolution rosetta scoring functions. *Proteins*, 74(3):777–788, Feb 2009.

[38] W. Braun. Representation of short and long-range handedness in protein structures by signed distance maps. *J Mol Biol*, 163(4):613–621, Feb 1983.

[39] S. E. Brenner, P. Koehl, and M. Levitt. The astral compendium for protein structure and sequence analysis. *Nucleic Acids Res*, 28(1):254–256, Jan 2000.

[40] K. V. Brinda, A. Surolia, and S. Vishveshwara. Insights into the quaternary association of proteins through structure graphs: a case study of lectins. *Biochem J*, 391(Pt 1):1–15, Oct 2005.

[41] K. V. Brinda and S. Vishveshwara. A network representation of protein structures: implications for protein stability. *Biophys J*, 89(6):4159–70, Dec 2005.

[42] K. V. Brinda and S. Vishveshwara. Oligomeric protein structure networks: insights into protein-protein interactions. *BMC Bioinformatics*, 6:296, Jan 2005.

[43] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, 1983.

[44] S. H. Bryant and C. E. Lawrence. An empirical energy function for threading protein sequence through the folding motif. *Proteins*, 16(1):92–112, May 1993.

[45] C. Bystroff and Y. Shao. Fully automated ab initio protein structure prediction using i-sites, hmmstr and rosetta. *Bioinformatics*, 18 Suppl 1:S54–61, Jan 2002.

[46] A. Caprara, R. Carr, S. Istrail, G. Lancia, and B. Walenz. 1001 optimal pdb structure alignments: integer programming methods for finding the maximum contact map overlap. *J Comput Biol*, 11(1):27–52, Jan 2004.

[47] E. Capriotti, P. Fariselli, and R. Casadio. A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics*, 20 Suppl 1:i63–i68, Aug 2004.

[48] E. Capriotti, P. Fariselli, I. Rossi, and R. Casadio. A three-state prediction of single point mutations on protein stability changes. *BMC Bioinformatics*, 9 Suppl 2:S6, 2008.

[49] P. Carbonell, R. Nussinov, and A. del Sol. Energetic determinants of protein binding specificity: insights into protein interaction networks. *Proteomics*, 9(7):1744–53, Apr 2009.

[50] C. W. Carter, B. C. LeFebvre, S. A. Cammer, A. Tropsha, and M. H. Edgell. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. *J Mol Biol*, 311(4):625–38, Aug 2001.

[51] E. Chea and D. R. Livesay. How accurate and statistically robust are catalytic site predictions based on closeness centrality? *BMC Bioinformatics*, 8:153, Jan 2007.

[52] Y. Chen, F. Ding, and N. V. Dokholyan. Fidelity of the protein structure reconstruction from inter-residue proximity constraints. *The journal of physical chemistry B*, 111(25):7432–8, Jun 2007.

[53] J. Cheng and P. Baldi. Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, 8:113, Jan 2007.

[54] T. M. K. Cheng, Y.-E. Lu, M. Vendruscolo, P. Lio', and T. L. Blundell. Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms. *PLoS Comput Biol*, 4(7):e1000135, 2008.

[55] C. Chothia. Structural invariants in protein folding. *Nature*, 254(5498):304–308, Mar 1975.

[56] M. S. Cline, K. Karplus, R. H. Lathrop, T. F. Smith, R. G. Rogers, and D. Haussler. Information-theoretic dissection of pairwise contact potentials. *Proteins*, 49(1):7–14, Oct 2002.

[57] R. Cohen and S. Havlin. Scale-free networks are ultra small. *Physical Review Letters*, 90:058701, 2003.

[58] S. A. Cook. The complexity of theorem-proving procedures. In *Proc. 3rd Ann. ACM Symp. on Theory of Computing*, pages 151–158. Assosiation for Computing Machinery, 1971.

[59] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

[60] G. M. Crippen. The tree structural organization of proteins. *J Mol Biol*, 126(3):315–32, Dec 1978.

[61] G. M. Crippen. Prediction of protein folding from amino acid sequence over discrete conformation spaces. *Biochemistry*, 30(17):4232–7, Apr 1991.

[62] G. M. Crippen and T. F. Havel. *Distance Geometry and Molecular Conformation (Chemometrics Series).* Research Studies Pr, 1988.

[63] A. L. Cuff, I. Sillitoe, T. Lewis, O. C. Redfern, R. Garratt, J. Thornton, and C. A. Orengo. The cath classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res*, 37(Database issue):D310–D314, Jan 2009.

[64] M. P. Cusack, B. Thibert, D. E. Bredesen, and G. del Rio. Efficient identification of critical residues based only on protein structure by network analysis. *PLoS ONE*, 2(5):e421, Jan 2007.

[65] Y. Dehouck, D. Gilis, and M. Rooman. A new generation of statistical potentials for proteins. *Biophys J*, 90(11):4010–4017, Jun 2006.

[66] A. del Sol, M. J. Araúzo-Bravo, D. Amoros, and R. Nussinov. Modular architecture of protein structures and allosteric communications: potential implications for signaling proteins and regulatory linkages. *Genome Biol*, 8(5):R92, Jan 2007.

[67] A. del Sol and P. Carbonell. The modular organization of domain structures: insights into protein-protein binding. *PLoS Comput Biol*, 3(12):e239, Dec 2007.

[68] A. del Sol, H. Fujihashi, D. Amoros, and R. Nussinov. Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families. *Protein Sci*, 15(9):2120–8, Sep 2006.

[69] A. del Sol, H. Fujihashi, D. Amoros, and R. Nussinov. Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Mol Syst Biol*, 2:2006.0019, Jan 2006.

[70] A. del Sol, H. Fujihashi, and P. O'Meara. Topology of small-world networks of protein-protein complex structures. *Bioinformatics*, 21(8):1311–5, Apr 2005.

[71] A. del Sol and P. O'Meara. Small-world network approach to identify key residues in protein-protein interaction. *Proteins*, 58(3):672–82, Feb 2005.

[72] B. Delaunay. Sur la sphére vide. a la memoire de georges voronoi. *Izv. Akad. Nauk SSSR, Otdelenie Matematicheskih i Estestvennyh Nauk*, 7:793–800, 1934.

[73] R. S. DeWitte and E. I. Shakhnovich. Pseudodihedrals: simplified protein backbone representation with knowledge-based energy. *Protein Sci*, 3(9):1570–1581, Sep 1994.

[74] F. Ding, K. C. Prutzman, S. L. Campbell, and N. V. Dokholyan. Topological determinants of protein domain swapping. *Structure*, 14(1):5–14, Jan 2006.

[75] P. D. Dixit and T. R. Weikl. A simple measure of native-state topology and chain connectivity predicts the folding rates of two-state proteins with and without crosslinks. *Proteins*, 64(1):193–7, Jul 2006.

[76] N. V. Dokholyan, L. Li, F. Ding, and E. I. Shakhnovich. Topological determinants of protein folding. *Proc Natl Acad Sci USA*, 99(13):8637–41, Jun 2002.

[77] N. V. Dokholyan, L. A. Mirny, and E. I. Shakhnovich. Understanding conserved amino acids in proteins. *Physica A: Statistical Mechanics and its Applications*, 314(1-4):600 – 606, 2002.

[78] J. M. Duarte, R. Sathyapriya, H. Stehr, I. Filippis, and M. Lappe. Optimal contact definition for reconstruction of contact maps. *BMC Bioinformatics*, 11:283, 2010.

[79] F. Emmert-Streib and A. Mushegian. A topological algorithm for identification of structural domains of proteins. *BMC Bioinformatics*, 8:237, Jan 2007.

[80] J. L. England, B. E. Shakhnovich, and E. I. Shakhnovich. Natural selection of more designable folds: a mechanism for thermophilic adaptation. *Proc Natl Acad Sci USA*, 100(15):8727–31, Jul 2003.

[81] J. L. England and E. I. Shakhnovich. Structural determinant of protein designability. *Phys Rev Lett*, 90(21):218101, May 2003.

[82] A. J. Enright, S. V. Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, 30(7):1575–1584, Apr 2002.

[83] E. Eyal, M. Frenkel-Morgenstern, V. Sobolev, and S. Pietrokovski. A pair-to-pair amino acids substitution matrix and its applications for protein structure prediction. *Proteins*, 67(1):142–53, Apr 2007.

[84] Q. Fang and D. Shortle. Protein refolding in silico with atom-based statistical potentials and conformational search using a simple genetic algorithm. *J Mol Biol*, 359(5):1456–1467, Jun 2006.

[85] P. Fariselli and R. Casadio. A neural network based predictor of residue contacts in proteins. *Protein Eng*, 12(1):15–21, Jan 1999.

[86] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio. Prediction of contact maps with neural networks and correlated mutations. *Protein Eng*, 14(11):835–43, Nov 2001.

[87] P. Fariselli, O. Olmea, A. Valencia, and R. Casadio. Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins*, Suppl 5:157–62, Jan 2001.

[88] Y. Feng, A. Kloczkowski, and R. L. Jernigan. Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys. *Proteins*, 68(1):57–66, Jul 2007.

[89] J. L. Finney. Volume occupation, environment and accessibility in proteins. the problem of the protein surface. *J Mol Biol*, 96(4):721–732, Aug 1975.

[90] K. F. Fischer and S. Marqusee. A rapid test for identification of autonomous folding units in proteins. *J Mol Biol*, 302(3):701–12, Sep 2000.

[91] P. J. Fleming and F. M. Richards. Protein packing: dependence on protein size, secondary structure and amino acid composition. *Journal of molecular biology*, 299(2):487–498, 2000.

[92] E. Furuichi and P. Koehl. Influence of protein structure databases on the predictive power of statistical pair potentials. *Proteins*, 31(2):139–149, May 1998.

[93] O. V. Galzitskaya, D. C. Reifsnyder, N. S. Bogatyreva, D. N. Ivankov, and S. O. Garbuzynskiy. More compact protein globules exhibit slower folding rates. *Proteins*, 70(2):329–32, Feb 2008.

[94] H. H. Gan, A. Tropsha, and T. Schlick. Lattice protein folding with two and four-body statistical potentials. *Proteins*, 43(2):161–74, May 2001.

[95] J.-C. Gelly, A. G. de Brevern, and S. Hazout. 'protein peeling': an approach for splitting a 3d protein structure into compact fragments. *Bioinformatics*, 22(2):129–33, Jan 2006.

[96] A. Ghosh, K. V. Brinda, and S. Vishveshwara. Dynamics of lysozyme structure network: probing the process of unfolding. *Biophys J*, 92(7):2523–35, Apr 2007.

[97] A. Ghosh and S. Vishveshwara. A study of communication pathways in methionyl- trna synthetase by molecular dynamics simulations and structure network analysis. *Proc Natl Acad Sci USA*, 104(40):15711–6, Oct 2007.

[98] D. Gilis. Protein decoy sets for evaluating energy functions. *J Biomol Struct Dyn*, 21(6):725–736, Jun 2004.

[99] A. V. Glyakina, S. O. Garbuzynskiy, M. Y. Lobanov, and O. V. Galzitskaya. Different packing of external residues can explain differences in the thermostability of proteins from thermophilic and mesophilic organisms. *Bioinformatics*, 23(17):2231–8, Sep 2007.

[100] U. Göbel, C. Sander, R. Schneider, and A. Valencia. Correlated mutations and residue contacts in proteins. *Proteins*, 18(4):309–17, Apr 1994.

[101] A. Godzik, A. Kolinski, and J. Skolnick. Topology fingerprint approach to the inverse protein folding problem. *J Mol Biol*, 227(1):227–38, Sep 1992.

[102] A. Godzik, A. Koliński, and J. Skolnick. Are proteins ideal mixtures of amino acids? analysis of energy parameter sets. *Protein Sci*, 4(10):2107–2117, Oct 1995.

[103] A. Godzik and C. Sander. Conservation of residue interactions in a family of ca-binding proteins. *Protein Eng*, 2(8):589–96, Aug 1989.

[104] A. Godzik and J. Skolnick. Sequence-structure matching in globular proteins: application to supersecondary and tertiary structure determination. *Proc Natl Acad Sci USA*, 89(24):12098–102, Dec 1992.

[105] A. Godzik and J. Skolnick. Flexible algorithm for direct multiple alignment of protein structures and sequences. *Comput Appl Biosci*, 10(6):587–596, Dec 1994.

[106] L. H. Greene and V. A. Higman. Uncovering network systems within protein structures. *J Mol Biol*, 334(4):781–91, Dec 2003.

[107] H. M. Grindley, P. J. Artymiuk, D. W. Rice, and P. Willett. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. *J Mol Biol*, 229(3):707–721, Feb 1993.

[108] M. M. Gromiha. Important inter-residue contacts for enhancing the thermal stability of thermophilic proteins. *Biophysical Chemistry*, 91(1):71 – 77, 2001.

[109] M. M. Gromiha. Multiple contact network is a key determinant to protein folding rates. *Journal of chemical information and modeling*, 49(4):1130–5, Apr 2009.

[110] M. M. Gromiha, M. Oobatake, H. Kono, H. Uedaira, and A. Sarai. Role of structural and sequence information in the prediction of protein stability changes: comparison between buried and partially buried mutations. *Protein Eng*, 12(7):549–55, Jul 1999.

[111] M. M. Gromiha, M. Oobatake, H. Kono, H. Uedaira, and A. Sarai. Importance of surrounding residues for protein stability of partially buried mutations. *J Biomol Struct Dyn*, 18(2):281–295, Oct 2000.

[112] M. M. Gromiha and S. Selvaraj. Influence of medium and long range interactions in different structural classes of globular proteins. *Journal of biological physics*, 23(3):151–162, 1997.

[113] M. M. Gromiha and S. Selvaraj. Importance of long-range interactions in protein folding. *Biophys Chem*, 77(1):49–68, Mar 1999.

[114] M. M. Gromiha and S. Selvaraj. 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(1):27–32, Jun 2001.

[115] C. Guerrero, T. Milenković, N. Pržulj, P. Kaiser, and L. Huang. Characterization of the proteasome interaction network using a qtax-based tag-team strategy and protein interaction network analysis. *Proc Natl Acad Sci U S A*, 105(36):13333–13338, Sep 2008.

[116] N. Hamilton, K. Burrage, M. A. Ragan, and T. Huber. Protein contact prediction using patterns of correlation. *Proteins*, 56(4):679–84, Sep 2004.

[117] T. F. Havel, G. M. Crippen, and I. D. Kuntz. Effects of distance constraints on macromolecular conformation. ii. simulation of experimental results and theoretical predictions. *Biopolymers*, 18(1):73–81, 1979.

[118] T. F. Havel and K. Wüthrich. An evaluation of the combined use of nuclear magnetic resonance and distance geometry for the determination of protein conformation in solution. *J Mol Biol*, 182::281–294, 1985.

[119] M. Hendlich, P. Lackner, S. Weitckus, H. Floeckner, R. Froschauer, K. Gottsbacher, G. Casari, and M. J. Sippl. Identification of native protein folds amongst a large number of incorrect models. the calculation of low energy conformations from potentials of mean force. *J Mol Biol*, 216(1):167–80, Nov 1990.

[120] K. Henrick and J. M. Thornton. PQS: a protein quaternary structure file server. *Trends Biochem Sci*, 23(9):358–361, Sep 1998.

[121] J. Heringa and P. Argos. Side-chain clusters in protein structures and their role in protein folding. *J Mol Biol*, 220(1):151–71, Jul 1991.

[122] J. Heringa, P. Argos, M. R. Egmond, and J. de Vlieg. Increasing thermal stability of subtilisin from mutations suggested by strongly interacting side-chain clusters. *Protein Eng*, 8(1):21–30, Jan 1995.

[123] V. A. Higman and L. H. Greene. Elucidation of conserved long-range interaction networks in proteins and their significance in determining protein topology. *Physica A: Statistical Mechanics and its Applications*, 368(2):595 – 606, 2006.

[124] D. A. Hinds and M. Levitt. Exploring conformational space with a simple lattice model for protein structure. *J Mol Biol*, 243(4):668–82, Nov 1994.

[125] H. Ho, T. Milenković, V. Memisević, J. Aruri, N. Pržulj, and A. K. Ganesan. Protein interaction network topology uncovers melanogenesis regulatory network components within functional genomics datasets. *BMC Syst Biol*, 4:84, 2010.

[126] L. Holm and C. Sander. Protein structure comparison by alignment of distance matrices. *J Mol Biol*, 233(1):123–38, Sep 1993.

[127] L. Holm and C. Sander. Parser for protein folding units. *Proteins*, 19(3):256–68, Jul 1994.

[128] J. B. Holmes and J. Tsai. Characterizing conserved structural contacts by pairwise relative contacts and relative packing groups. *Journal of Molecular Biology*, 354(3):706–721, Dec 2005.

[129] J. Hu, X. Shen, Y. Shao, C. Bystroff, and M. J. Zaki. Mining protein contact

maps. In *In The 2nd ACM SIGKDD Workshop on Data Mining in Bioinformatics (BIOKDD)*, 2002.

[130] Z. Hu, D. Bowen, W. M. Southerland, A. del Sol, Y. Pan, R. Nussinov, and B. Ma. Ligand binding and circular permutation modify residue interaction network in dhfr. *PLoS Comput Biol*, 3(6):e117, Jun 2007.

[131] J. Huan, D. Bandyopadhyay, W. Wang, J. Snoeyink, J. Prins, and A. Tropsha. Comparing graph representations of protein structure for mining family-specific residue-based packing motifs. *J Comput Biol*, 12(6):657–671, 2005.

[132] S. Hubbard and J. Thornton. Naccess computer program. *University College London Department of Biochemistry and Molecular Biology*, 1993.

[133] T. Ideker and R. Sharan. Protein networks in disease. *Genome Res*, 18(4):644–652, Apr 2008.

[134] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai. Revealing modular organization in the yeast transcriptional network. *Nat Genet*, 31(4):370–377, Aug 2002.

[135] T. Ishida, S. Nakamura, and K. Shimizu. Potential for assessing quality of protein structure based on contact number prediction. *Proteins*, 64(4):940–7, Sep 2006.

[136] J. M. G. Izarzugaza, O. G. na, M. L. Tress, A. Valencia, and N. D. Clarke. Assessment of intramolecular contact predictions for casp7. *Proteins*, 69 Suppl 8:152–8, Jan 2007.

[137] D. J. Jacobs, A. J. Rader, L. A. Kuhn, and M. F. Thorpe. Protein flexibility predictions using graph theory. *Proteins*, 44(2):150–165, Aug 2001.

[138] B. J. Jain and M. Lappe. Joining softassign and dynamic programming for the contact map overlap problem. In *BIRD*, pages 410–423, 2007.

[139] I. Jonassen, I. Eidhammer, D. Conklin, and W. R. Taylor. Structure motif discovery and mining the pdb. *Bioinformatics*, 18(2):362–7, Feb 2002.

[140] I. Jonassen, I. Eidhammer, and W. R. Taylor. Discovery of local packing motifs in protein structures. *Proteins*, 34(2):206–19, Feb 1999.

[141] D. T. Jones, W. R. Taylor, and J. M. Thornton. A new approach to protein fold recognition. *Nature*, 358(6381):86–9, Jul 1992.

[142] S. Jones and J. M. Thornton. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 93(1):13–20, 1996.

[143] P. F. Jonsson and P. A. Bates. Global topological features of cancer proteins in the human interactome. *Bioinformatics*, 22(18):2291–2297, Sep 2006.

[144] J. Jung, J. Lee, and H.-T. Moon. Topological determinants of protein unfolding rates. *Proteins*, 58(2):389–95, Feb 2005.

[145] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[146] K. Kamagata and K. Kuwajima. Surprisingly high correlation between early and late stages in non-two-state protein folding. *J Mol Biol*, 357(5):1647–54, Apr 2006.

[147] N. Kannan and S. Vishveshwara. Identification of side-chain clusters in protein structures by a graph spectral method. *J Mol Biol*, 292(2):441–64, Sep 1999.

[148] R. Karchin, M. Cline, and K. Karplus. Evaluation of local structure alphabets based on residue burial. *Proteins*, 55(3):508–18, May 2004.

[149] N. Kashtan, S. Itzkovitz, R. Milo, and U. Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20:1746–1758, 2004.

[150] W. Kauzmann. Some factors in the interpretation of protein denaturation. *Adv Protein Chem*, 14:1–63, 1959.

[151] J. Khatun, S. D. Khare, and N. V. Dokholyan. Can contact potentials reliably predict stability of proteins? *J Mol Biol*, 336(5):1223–38, Mar 2004.

[152] A. R. Kinjo, K. Horimoto, and K. Nishikawa. Predicting absolute contact numbers of native protein structure from amino acid sequence. *Proteins*, 58(1):158–65, Jan 2005.

[153] A. R. Kinjo and K. Nishikawa. Recoverable one-dimensional encoding of three-dimensional protein structures. *Bioinformatics*, 21(10):2167–70, May 2005.

[154] A. R. Kinjo and K. Nishikawa. Crnpred: highly accurate prediction of one-dimensional protein structures by large-scale critical random networks. *BMC Bioinformatics*, 7:401, Jan 2006.

[155] J. P. Kocher, M. J. Rooman, and S. J. Wodak. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J Mol Biol*, 235(5):1598–613, Feb 1994.

[156] B. Kolbeck, P. May, T. Schmidt-Goenner, T. Steinke, and E.-W. Knapp. Connectivity independent protein-structure alignment: a hierarchical approach. *BMC Bioinformatics*, 7:510, Jan 2006.

[157] A. Kolinski, A. Godzik, and J. Skolnick. A general method for the prediction of the three dimensional structure and folding pathway of globular proteins: Application to designed helical proteins. *The Journal of Chemical Physics*, 98(9):7420–7433, 1993.

[158] A. Kolinski and J. Skolnick. Discretized model of proteins. i. monte carlo study of cooperativity in homopolypeptides. *The Journal of Chemical Physics*, 97(12):9412–9426, 1992.

[159] B. Krishnamoorthy and A. Tropsha. Development of a four-body statistical

pseudo-potential to discriminate native from non-native protein conformations. *Bioinformatics*, 19(12):1540–8, Aug 2003.

[160] A. Krishnan, A. Giuliani, J. P. Zbilut, and M. Tomita. Network scaling invariants help to elucidate basic topological principles of proteins. *J Proteome Res*, 6(10):3924–34, Oct 2007.

[161] E. Krissinel and K. Henrick. Inference of macromolecular assemblies from crystalline state. *Journal of molecular biology*, 372(3):774–797, 2007.

[162] O. Kuchaiev, T. Milenković, V. Memisevic, W. Hayes, and N. Pržulj. Topological network alignment uncovers biological function and phylogeny. *J R Soc Interface*, 7(50):1341–1354, Sep 2010.

[163] O. Kuchaiev, A. Stevanović, W. Hayes, and N. Pržulj. Graphcrunch 2: Software tool for network modeling, alignment and clustering. *BMC Bioinformatics*, 12:24, 2011.

[164] T. S. Kumarevel, M. M. Gromiha, S. Selvaraj, K. Gayatri, and P. K. R. Kumar. Influence of medium- and long-range interactions in different folding types of globular proteins. *Biophys Chem*, 99(2):189–198, Oct 2002.

[165] P. J. Kundrotas and E. G. Alexov. Predicting residue contacts using pragmatic correlated mutations method: reducing the false positives. *BMC Bioinformatics*, 7:503, Jan 2006.

[166] I. B. Kuznetsov and S. Rackovsky. Discriminative ability with respect to amino acid types: assessing the performance of knowledge-based potentials without threading. *Proteins*, 49(2):266–84, Nov 2002.

[167] M. Lappe, G. Bagler, I. Filippis, H. Stehr, J. M. Duarte, and R. Sathyapriya. Designing evolvable libraries using multi-body potentials. *Curr Opin Biotechnol*, 20(4):437–446, Aug 2009.

[168] T. Li, K. Fan, J. Wang, and W. Wang. Reduction of protein sequence complexity by residue grouping. *Protein Eng*, 16(5):323–330, May 2003.

[169] X. Li and J. Liang. Geometric cooperativity and anticooperativity of three-body interactions in native proteins. *Proteins*, 60(1):46–65, Jul 2005.

[170] K. Lin, J. Kleinjung, W. R. Taylor, and J. Heringa. Testing homology with contact accepted mutation (cao): a contact-based markov model of protein evolution. *Comput Biol Chem*, 27(2):93–102, May 2003.

[171] K. Lin, A. C. W. May, and W. R. Taylor. Threading using neural network (tune): the measure of protein sequence-structure compatibility. *Bioinformatics*, 18(10):1350–7, Oct 2002.

[172] M. Lin, H.-M. Lu, R. Chen, and J. Liang. Generating properly weighted ensemble of conformations of proteins from sparse or indirect distance constraints. *J. Chem. Phys.*, 129(9):094101, Sep 2008.

[173] E. Lindahl, B. Hess, and D. Van Der Spoel. Gromacs 3.0: a package for molecular

simulation and trajectory analysis. *Journal of Molecular Modeling*, 7(8):306–317, 2001.

[174] K. Lindorff-Larsen, R. B. Best, M. A. Depristo, C. M. Dobson, and M. Vendruscolo. Simultaneous determination of protein structure and dynamics. *Nature*, 433(7022):128–132, Jan 2005.

[175] A. M. Lisewski and O. Lichtarge. Rapid detection of similarity in protein structure and function through contact metric distances. *Nucleic Acids Res*, 34(22):e152, Jan 2006.

[176] H. Lu and J. Skolnick. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins*, 44(3):223–232, Aug 2001.

[177] B.-G. Ma, L.-L. Chen, and H.-Y. Zhang. What determines protein folding type? an investigation of intrinsic structural properties and its implications for understanding folding mechanisms. *J Mol Biol*, 370(3):439–48, Jul 2007.

[178] R. M. MacCallum. Striped sheets and protein contact prediction. *Bioinformatics*, 20 Suppl 1:i224–31, Aug 2004.

[179] A. D. MacKerell, D. Bashford, Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B*, 102(18):3586–3616, 1998.

[180] E. R. Main, K. F. Fulton, and S. E. Jackson. Context-dependent nature of destabilizing mutations on the stability of fkbp12. *Biochemistry*, 37(17):6145–6153, Apr 1998.

[181] V. N. Maiorov and G. M. Crippen. Contact potential that recognizes the correct folding of globular proteins. *J Mol Biol*, 227(3):876–88, Oct 1992.

[182] D. E. Makarov, C. A. Keller, K. W. Plaxco, and H. Metiu. How the folding rate constant of simple, single-domain proteins depends on the number of native contacts. *Proc Natl Acad Sci USA*, 99(6):3535–9, Mar 2002.

[183] P. Manavalan and P. K. Ponnuswamy. A study of the preferred environment of amino acid residues in globular proteins. *Arch Biochem Biophys*, 184(2):476–487, Dec 1977.

[184] P. Manavalan and P. K. Ponnuswamy. Hydrophobic character of amino acid residues in globular proteins. *Nature*, 275(5681):673–4, Oct 1978.

[185] L. C. Martin, G. B. Gloor, S. D. Dunn, and L. M. Wahl. Using information theory to search for co-evolving residues in proteins. *Bioinformatics*, 21(22):4116–4124, Nov 2005.

[186] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002.

[187] S. Mayewski. A multibody, whole-residue potential for protein structures, with testing by monte carlo simulated annealing. *Proteins*, 59(2):152–69, May 2005.

[188] K. Mehlhorn and S. Naher. *Leda: A platform for combinatorial and geometric computing*. Cambridge University Press, 1999.

[189] F. Melo and E. Feytmans. Novel knowledge-based mean force potential at atomic level. *J Mol Biol*, 267(1):207–222, Mar 1997.

[190] F. Melo, R. Sánchez, and A. Sali. Statistical potentials for fold assessment. *Protein Sci*, 11(2):430–448, Feb 2002.

[191] T. Milenković, I. Filippis, M. Lappe, and N. Pržulj. Optimized null model for protein structure networks. *PLoS One*, 4(6):e5967, 2009.

[192] T. Milenković, J. Lai, and N. Pržulj. Graphcrunch: a tool for large network analyses. *BMC Bioinformatics*, 9(70), 2008.

[193] T. Milenković, V. Memisevic, A. K. Ganesan, and N. Pržulj. Systems-level cancer gene identification from protein interaction network topology applied to melanogenesis-related functional genomics data. *J R Soc Interface*, 7(44):423–437, Mar 2010.

[194] T. Milenković, V. Memišević, A. Bonato, and N. Pršulj. Dominating biological networks. *PLoS One*, 6(8):e23016, 2011.

[195] T. Milenković, W. L. Ng, W. Hayes, and N. Pržulj. Optimal network alignment with graphlet degree vectors. *Cancer Inform*, 9:121–137, 2010.

[196] T. Milenković and N. Pržulj. Uncovering biological network function via graphlet degree signatures. *Cancer Inform*, 6:257–273, 2008.

[197] C. S. Miller and D. Eisenberg. Using inferred residue contacts to distinguish between correct and incorrect protein models. *Bioinformatics*, 24(14):1575–82, Jul 2008.

[198] E. J. Miller, K. F. Fischer, and S. Marqusee. Experimental evaluation of topological parameters determining protein-folding rates. *Proc Natl Acad Sci USA*, 99(16):10359–63, Aug 2002.

[199] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of evolved and designed networks. *Science*, 303:1538–1542, 2004.

[200] R. Milo, S. S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298:824–827, 2002.

[201] L. Mirny and E. Domany. Protein fold recognition and dynamics in the space of contact maps. *Proteins*, 26(4):391–410, Dec 1996.

[202] L. Mirny and E. Shakhnovich. Protein folding theory: from lattice to all-atom models. *Annual review of biophysics and biomolecular structure*, 30:361–96, Jan 2001.

[203] L. A. Mirny and E. I. Shakhnovich. How to derive a protein folding potential? a new approach to an old problem. *J Mol Biol*, 264(5):1164–79, Dec 1996.

[204] S. Miyazawa and R. L. Jernigan. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, 18(3):534–552, 1985.

[205] S. Miyazawa and R. L. Jernigan. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol*, 256(3):623–44, Mar 1996.

[206] S. Miyazawa and R. L. Jernigan. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins*, 34(1):49–68, Jan 1999.

[207] H. M. M. Molina, C. Millán-Pacheco, N. Pastor, and G. del Rio. Computer-based screening of functional conformers of proteins. *PLoS Comput Biol*, 4(2):e1000009, Feb 2008.

[208] S. D. Mooney, M. H.-P. Liang, R. DeConde, and R. B. Altman. Structural characterization of proteins using residue environments. *Proteins*, 61(4):741–747, Dec 2005.

[209] P. J. Munson and R. K. Singh. Statistical significance of hierarchical multibody potentials based on delaunay tessellation and their application in sequence-structure alignment. *Protein Sci*, 6(7):1467–81, Jul 1997.

[210] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–40, 1995.

[211] O. G. na, D. Baker, R. M. MacCallum, J. Meiler, M. Punta, B. Rost, M. L. Tress, and A. Valencia. Casp6 assessment of contact prediction. *Proteins*, 61 Suppl 7:214–24, Jan 2005.

[212] G. Nemethy, K. D. Gibson, K. A. Palmer, C. N. Yoon, G. Paterlini, A. Zagari, S. Rumsey, and H. A. Scheraga. Energy parameters in polypeptides. 10. improved geometrical parameters and nonbonded interactions for use in the ecepp/3 algorithm, with application to proline-containing peptides. *The Journal of Physical Chemistry*, 96(15):6472–6484, 1992.

[213] M. E. J. Newman. Assortative mixing in networks. *Phys Rev Lett*, 89(20):208701, Nov 2002.

[214] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[215] K. Nishikawa and T. Ooi. Comparison of homologous tertiary structures of proteins. *Journal of Theoretical Biology*, 43(2):351 – 374, 1974.

[216] K. Nishikawa and T. Ooi. Prediction of the surface-interior diagram of globular proteins by an empirical method. *Int J Pept Protein Res*, 16(1):19–32, Jul 1980.

[217] K. Nishikawa and T. Ooi. Radial locations of amino acid residues in a globular protein: correlation with the sequence. *J Biochem*, 100(4):1043–1047, Oct 1986.

[218] K. Nishikawa, T. Ooi, Y. Isogai, and N. Saito. Tertiary structure of proteins. i. representation and computation of the conformations. *Journal of the Physical Society of Japan*, 32:1331–1337, 1972.

[219] T. S. Norcross and T. O. Yeates. A framework for describing topological frustration in models of protein folding. *J Mol Biol*, 362(3):605–21, Sep 2006.

[220] V. M. noz and W. A. Eaton. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc Natl Acad Sci USA*, 96(20):11311–6, Sep 1999.

[221] O. Olmea, B. Rost, and A. Valencia. Effective use of sequence correlation and conservation in fold recognition. *J Mol Biol*, 293(5):1221–39, Nov 1999.

[222] O. Olmea and A. Valencia. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Folding & design*, 2(3):S25–32, Jan 1997.

[223] A. R. Ortiz, A. Kolinski, P. Rotkiewicz, B. Ilkowski, and J. Skolnick. Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins*, Suppl 3:177–85, Jan 1999.

[224] C. Ouzounis, C. Sander, M. Scharf, and R. Schneider. Prediction of protein structure by evaluation of sequence-structure fitness. aligning sequences to contact profiles derived from three-dimensional structures. *J Mol Biol*, 232(3):805–25, Aug 1993.

[225] OWL. The Otto Warburg Java Library of Structural Bioinformatics - http://www.bioinformatics.org/owl/. 2011.

[226] E. Paci, K. Lindorff-Larsen, C. M. Dobson, M. Karplus, and M. Vendruscolo. Transition state contact orders correlate with protein folding rates. *J Mol Biol*, 352(3):495–500, Sep 2005.

[227] G. A. Papoian, J. Ulander, M. P. Eastwood, Z. Luthey-Schulten, and P. G. Wolynes. Water in protein structure prediction. *Proc Natl Acad Sci USA*, 101(10):3352–7, Mar 2004.

[228] B. Park and M. Levitt. Energy functions that discriminate x-ray and near native folds from well-constructed decoys. *J Mol Biol*, 258(2):367–92, May 1996.

[229] R. Pastor-Satorras, E. Smith, and R. V. Sole. Evolving protein interaction networks through gene duplication. *Journal of Theoretical Biology*, 222:199–210, 2003.

[230] K. H. Paszkiewicz, M. J. E. Sternberg, and M. Lappe. Prediction of viable circular permutants using a graph theoretic approach. *Bioinformatics*, 22(11):1353–8, Jun 2006.

[231] S. M. Patra and S. Vishveshwara. Backbone cluster identification in proteins by a graph theoretical method. *Biophys Chem*, 84(1):13–25, Feb 2000.

[232] J. B. Pereira-Leal, A. J. Enright, and C. A. Ouzounis. Detection of functional modules from protein interaction networks. *Proteins*, 54(1):49–57, Jan 2004.

[233] D. C. Phillips. The development of crystallographic enzymology. *Biochem Soc Symp*, 30:11–28, Jan 1970.

[234] K. W. Plaxco, K. T. Simons, and D. Baker. Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol*, 277(4):985–94, Apr 1998.

[235] G. Pollastri and P. Baldi. Prediction of contact maps by giohmms and recurrent neural networks using lateral propagation from all four cardinal corners. *Bioinformatics*, 18 Suppl 1:S62–70, Jan 2002.

[236] G. Pollastri, P. Baldi, P. Fariselli, and R. Casadio. Improved prediction of the number of residue contacts in proteins by recurrent neural networks. *Bioinformatics*, 17 Suppl 1:S234–42, Jan 2001.

[237] G. Pollastri, P. Baldi, P. Fariselli, and R. Casadio. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins*, 47(2):142–53, May 2002.

[238] G. Pollastri, A. Vullo, P. Frasconi, and P. Baldi. Modular dag-rnn architectures for assembling coarse protein structures. *J Comput Biol*, 13(3):631–50, Apr 2006.

[239] M. Porto, U. Bastolla, H. E. Roman, and M. Vendruscolo. Reconstruction of protein structures from a vectorial representation. *Phys Rev Lett*, 92(21):218101, May 2004.

[240] N. Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, Jan 2007.

[241] N. Pržulj, D. G. Corneil, and I. Jurisica. Modeling interactome: Scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, 2004.

[242] N. Pržulj and D. Higham. Modelling protein-protein interaction networks via a stickiness index. *Journal of the Royal Society Interface*, 3(10):711–716, 2006.

[243] M. Punta and B. Rost. Profcon: novel prediction of long-range contacts. *Bioinformatics*, 21(13):2960–8, Jul 2005.

[244] M. Punta and B. Rost. Protein folding rates estimated from contact predictions. *J Mol Biol*, 348(3):507–512, May 2005.

[245] R. Rajgaria, S. R. Mcallister, and C. A. Floudas. A novel high resolution ca-ca distance dependent force field based on a high quality decoy set. *Proteins*, 65(3):726–741, Nov 2006.

[246] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–5, 2002.

[247] D. Reichmann, O. Rahat, S. Albeck, R. Meged, O. Dym, and G. Schreiber. The

modular architecture of protein-protein binding interfaces. *Proc Natl Acad Sci USA*, 102(1):57–62, Jan 2005.

[248] F. M. Richards. The interpretation of protein structures: total volume, group volume distributions and packing density. *J Mol Biol*, 82(1):1–14, Jan 1974.

[249] M. Robinson-Rechavi, A. Alibés, and A. Godzik. Contribution of electrostatic interactions, compactness and quaternary structure to protein thermostability: lessons from structural genomics of thermotoga maritima. *J Mol Biol*, 356(2):547–57, Feb 2006.

[250] M. Robinson-Rechavi and A. Godzik. Structural genomics of thermotoga maritima proteins shows that contact order is a major determinant of protein thermostability. *Structure*, 13(6):857–60, Jun 2005.

[251] B. Rost and V. A. Eyrich. Eva: large-scale analysis of secondary structure prediction. *Proteins*, Suppl 5:192–199, 2001.

[252] R. B. Russell. Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. *J Mol Biol*, 279(5):1211–27, Jun 1998.

[253] R. B. Russell and G. J. Barton. Structural features can be unconserved in proteins with similar folds. an analysis of side-chain to side-chain contacts secondary structure and accessibility. *J Mol Biol*, 244(3):332–50, Dec 1994.

[254] S. Saitoh, T. Nakai, and K. Nishikawa. A geometrical constraint approach for reproducing the native backbone conformation of a protein. *Proteins*, 15(2):191–204, Feb 1993.

[255] R. Samudrala and M. Levitt. Decoys 'r' us: a database of incorrect conformations to improve protein structure prediction. *Protein Sci*, 9(7):1399–1401, Jul 2000.

[256] R. Sathyapriya, K. V. Brinda, and S. Vishveshwara. Correlation of the side-chain hubs with the functional residues in dna binding protein structures. *Journal of chemical information and modeling*, 46(1):123–9, Jan 2006.

[257] R. Sathyapriya, J. M. Duarte, H. Stehr, I. Filippis, and M. Lappe. Defining an essence of structure determining residue contacts in proteins. *PLoS Comput Biol*, 5(12):e1000584, Dec 2009.

[258] R. Sathyapriya, M. S. Vijayabaskar, and S. Vishveshwara. Insights into protein-dna interactions through structure network analysis. *PLoS Comput Biol*, 4(9):e1000170, Jan 2008.

[259] R. Sathyapriya and S. Vishveshwara. Structure networks of e. coli glutaminyl-trna synthetase: effects of ligand binding. *Proteins*, 68(2):541–50, Aug 2007.

[260] J. Selbig and P. Argos. Relationships between protein sequence and structure patterns based on residue contacts. *Proteins*, 31(2):172–85, May 1998.

[261] G. Shackelford and K. Karplus. Contact prediction using mutual information and neural nets. *Proteins*, 69 Suppl 8:159–64, Jan 2007.

[262] B. E. Shakhnovich, E. Deeds, C. Delisi, and E. Shakhnovich. Protein structure and evolutionary history determine sequence space topology. *Genome Res*, 15(3):385–92, Mar 2005.

[263] Y. Shao and C. Bystroff. Predicting interresidue contacts using templates and pathways. *Proteins*, 53 Suppl 6:497–502, Jan 2003.

[264] I. N. Shindyalov, N. A. Kolchanov, and C. Sander. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? *Protein Eng*, 7(3):349–58, Mar 1994.

[265] J. G. Siek, L.-Q. Lee, and A. Lumsdaine. *The Boost Graph Library: User Guide and Reference Manual.* Addison-Wesley, 2002.

[266] K. T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of Molecular Biology*, 268(1):209–25, Apr 1997.

[267] M. S. Singer, G. Vriend, and R. P. Bywater. Prediction of protein residue contacts with a pdb-derived likelihood matrix. *Protein Eng*, 15(9):721–5, Sep 2002.

[268] J. Singh and J. M. Thornton. *Atlas of Protein Side-Chain Interactions.* IRL Press at Oxford University Press (Oxford, New York), 1992.

[269] R. K. Singh, A. Tropsha, and I. I. Vaisman. Delaunay tessellation of proteins: four body nearest-neighbor propensities of amino acid residues. *J Comput Biol*, 3(2):213–21, Jan 1996.

[270] M. J. Sippl. On the problem of comparing protein structures. development and applications of a new method for the assessment of structural similarities of polypeptide conformations. *J Mol Biol*, 156(2):359–388, Apr 1982.

[271] M. J. Sippl. Calculation of conformational ensembles from potentials of mean force. an approach to the knowledge-based prediction of local structures in globular proteins. *J Mol Biol*, 213(4):859–883, Jun 1990.

[272] R. K. Sistla, B. K. V, and S. Vishveshwara. Identification of domains and domain interface residues in multidomain proteins from graph spectral method. *Proteins*, 59(3):616–26, May 2005.

[273] J. Skolnick, L. Jaroszewski, A. Kolinski, and A. Godzik. Derivation and testing of pair potentials for protein folding. when is the quasichemical approximation correct? *Protein Sci*, 6(3):676–688, Mar 1997.

[274] P. Slama, I. Filippis, and M. Lappe. Detection of protein catalytic residues at high precision using local network properties. *BMC Bioinformatics*, 9:517, Jan 2008.

[275] J. C. Smith and M. Karplus. Empirical force field study of geometries and conformational transitions of some organic molecules. *Journal of the American Chemical Society*, 114(3):801–812, 1992.

[276] V. Sobolev, A. Sorokine, J. Prilusky, E. E. Abola, and M. Edelman. Automated analysis of interatomic contacts in proteins. *Bioinformatics*, 15(4):327–32, Apr 1999.

[277] J. Song and K. Burrage. Predicting residue-wise contact orders in proteins by support vector regression. *BMC Bioinformatics*, 7:425, Jan 2006.

[278] H. Stehr, J. Duarte, I. Filippis, S. Rajagopal, K. Syal, S. Risbud, L. Holm, and M. Lappe. Struppi: comparative modeling using consensus information from multiple templates and physics-based refinement. *In Abstracts book, 8th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction Edited by: Moult J, Fidelis K, Kryshtafovych A, Rost B, Tramontano A.*, 2008.

[279] J. Sun and Z. Zhao. A comparative study of cancer proteins in the human protein-protein interaction network. *BMC Genomics*, 11 Suppl 3:S5, 2010.

[280] S. R. Sunyaev, F. Eisenhaber, P. Argos, E. N. Kuznetsov, and V. G. Tumanyan. Are knowledge-based potentials derived from protein structure sets discriminative with respect to amino acid types? *Proteins*, 31(3):225–46, May 1998.

[281] S. Tanaka and H. A. Scheraga. Model of protein folding: inclusion of short-, medium-, and long-range interactions. *Proc Natl Acad Sci USA*, 72(10):3802–6, Oct 1975.

[282] S. Tanaka and H. A. Scheraga. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, 9(6):945–50, Jan 1976.

[283] Y.-R. Tang, Z.-Y. Sheng, Y.-Z. Chen, and Z. Zhang. An improved prediction of catalytic residues in enzyme structures. *Protein Eng Des Sel*, 21(5):295–302, May 2008.

[284] T. Tanimoto. Ibm internal report 17th nov. 1957. Technical report, 1957.

[285] A. Tegge, Z. Wang, J. Eickholt, and J. Cheng. Nncon: improved protein contact map prediction using 2d-recursive neural networks. *Nucleic Acids Res*, May 2009.

[286] F. Teichert, U. Bastolla, and M. Porto. Sabertooth: protein structural alignment based on a vectorial structure representation. *BMC Bioinformatics*, 8:425, Jan 2007.

[287] B. Thibert, D. E. Bredesen, and G. del Rio. Improved prediction of critical residues for protein function based on network and phylogenetic analyses. *BMC Bioinformatics*, 6:213, Jan 2005.

[288] D. J. Thomas, G. Casari, and C. Sander. The prediction of protein contacts from multiple sequence alignments. *Protein Eng*, 9(11):941–8, Nov 1996.

[289] P. D. Thomas and K. A. Dill. An iterative method for extracting energy-like quantities from protein structures. *Proc Natl Acad Sci USA*, 93(21):11628–33, Oct 1996.

[290] P. D. Thomas and K. A. Dill. Statistical potentials extracted from protein structures: how accurate are they? *J Mol Biol*, 257(2):457–469, Mar 1996.

[291] G. Tiana, B. E. Shakhnovich, N. V. Dokholyan, and E. I. Shakhnovich. Imprint of evolution on protein structures. *Proc Natl Acad Sci USA*, 101(9):2846–51, Mar 2004.

[292] J. I. Tinoco, K. Sauer, and J. Wang. *Physical Chemistry: Principles and Applications in Biological Sciences*. Prentice-Hall, New Jersey, 3rd edition, 1995.

[293] D. Tobi and R. Elber. Distance-dependent, pair potential for protein folding: results from linear optimization. *Proteins*, 41(1):40–6, Oct 2000.

[294] A. Tropsha, R. K. Singh, I. I. Vaisman, and W. Zheng. Statistical geometry analysis of proteins: implications for inverted structure prediction. *Pac Symp Biocomput*, pages 614–623, 1996.

[295] C. J. Tsai and R. Nussinov. Hydrophobic folding units derived from dissimilar monomer structures and their interactions. *Protein Sci*, 6(1):24–42, Jan 1997.

[296] S. Van Dongen. Graph clustering via a discrete uncoupling process. *SIAM J. Matrix Anal. Appl.*, 30:121–141, February 2008.

[297] M. Vassura, L. Margara, P. di Lena, F. Medri, P. Fariselli, and R. Casadio. Fault tolerance for large scale protein 3d reconstruction from contact maps. In *WABI*, pages 25–37, 2007.

[298] M. Vassura, L. Margara, P. Fariselli, and R. Casadio. A graph theoretic approach to protein structure selection. In *WILF*, volume 4578 of *Lecture Notes in Computer Science*, pages 497–504. Springer, 2007.

[299] M. Vassura, L. Margara, P. D. Lena, F. Medri, P. Fariselli, and R. Casadio. Ftcomar: fault tolerant three-dimensional structure reconstruction from protein contact maps. *Bioinformatics*, 24(10):1313–5, May 2008.

[300] M. Vassura, L. Margara, F. Medri, P. di Lena, P. Fariselli, and R. Casadio. Reconstruction of 3d structures from protein contact maps. In *ISBRA*, pages 578–589, 2007.

[301] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani. Modeling of protein interaction networks. *ComPlexUs*, 1:38–44, 2001.

[302] C. Vehlow, H. Stehr, M. Winkelmann, J. M. Duarte, L. Petzold, J. Dinse, and M. Lappe. Cmview: interactive contact map visualization and analysis. *Bioinformatics*, 27(11):1573–1574, Jun 2011.

[303] M. Vendruscolo, N. V. Dokholyan, E. Paci, and M. Karplus. Small-world view of the amino acids that play a key role in protein folding. *Physical review E, Statistical, nonlinear, and soft matter physics*, 65(6 Pt 1):061910, Jun 2002.

[304] M. Vendruscolo and E. Domany. Pairwise contact potentials are unsuitable for protein folding. *The Journal of Chemical Physics*, 109(24):11101–11108, 1998.

[305] M. Vendruscolo and E. Domany. Protein folding using contact maps. *Vitam Horm*, 58:171–212, 2000.

[306] M. Vendruscolo, E. Kussell, and E. Domany. Recovery of protein structure from contact maps. *Folding & design*, 2(5):295–306, Jan 1997.

[307] M. Vendruscolo, R. Najmanovich, and E. Domany. Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? *Proteins*, 38(2):134–148, Feb 2000.

[308] M. Vendruscolo, E. Paci, C. M. Dobson, and M. Karplus. Three key residues form a critical contact network in a protein folding transition state. *Nature*, 409(6820):641–5, Feb 2001.

[309] S. Vicatos, B. V. B. Reddy, and Y. Kaznessis. Prediction of distant residue contacts with the use of evolutionary information. *Proteins*, 58(4):935–49, Mar 2005.

[310] S. Vishveshwara, K. V. Brinda, and N. Kannan. Protein structure: Insights from graph theory. *Journal of Theoretical and Computational Chemistry (JTCC)*, 1:187 – 211, 2002.

[311] G. Voronoi. Nouvelles applications des paramétres continus á la théorie des formes quadratiques, deuxiéme memoire, recherche sur les parallelloédres primitifs. *Journal für die Reine und Angewandte Mathematik*, 134:198–287, 1908.

[312] N. R. Voss and M. Gerstein. Calculation of standard atomic volumes for rna and comparison with proteins: Rna is packed more tightly. *Journal of molecular biology*, 346(2):477–492, 2005.

[313] A. Vullo, I. Walsh, and G. Pollastri. A two-stage approach for improved prediction of residue contact maps. *BMC Bioinformatics*, 7:180, Jan 2006.

[314] A. Wagner and D. Fell. The small world inside large metabolic networks. *Proc. Roy. Soc. London Series B*, 268:1803–1810, 2001.

[315] G. Wang and R. L. J. Dunbrack. PISCES: a protein sequence culling server. *Bioinformatics*, 19(12):1589–1591, Aug 2003. Comparative Study.

[316] P. P. Wangikar, A. V. Tendulkar, S. Ramya, D. N. Mali, and S. Sarawagi. Functional sites in protein families uncovered via an objective and automated graph theoretic approach. *J Mol Biol*, 326(3):955–78, Feb 2003.

[317] P. K. Warme and R. S. Morgan. A survey of atomic interactions in 21 proteins. *Journal of Molecular Biology*, 118(3):273–87, Jan 1978.

[318] D. J. Watts. *Small Worlds: The Dynamics of Networks between Order and Randomness*, volume 107. Princeton University Press, 1999.

[319] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, Jun 1998.

[320] T. R. Weikl and K. A. Dill. Folding rates and low-entropy-loss routes of two-state proteins. *J Mol Biol*, 329(3):585–98, Jun 2003.

[321] S. J. Weiner, P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society*, 106(3):765–784, 1984.

[322] G. Williams and P. Doherty. Inter-residue distances derived from fold contact propensities correlate with evolutionary substitution costs. *BMC Bioinformatics*, 5:153, Oct 2004.

[323] E. A. Winzeler, D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, R. Benito, J. D. Boeke, H. Bussey, A. M. Chu, C. Connelly, K. Davis, F. Dietrich, S. W. Dow, M. E. Bakkoury, F. Foury, S. H. Friend, E. Gentalen, G. Giaever, J. H. Hegemann, T. Jones, M. Laub, H. Liao, N. Liebundguth, D. J. Lockhart, A. Lucau-Danila, M. Lussier, N. M. Rabet, P. Menard, M. Mittmann, C. Pai, C. Rebischung, J. L. Revuelta, L. Riles, C. J. Roberts, P. Ross-MacDonald, B. Scherens, M. Snyder, S. Sookhai-Mahadeo, R. K. Storms, S. Veronneau, M. Voet, G. Volckaert, T. R. Ward, R. Wysocki, G. S. Yen, K. Yu, K. Zimmermann, P. Philippsen, M. Johnston, and R. W. Davis. Functional characterization of the s. cerevisiae genome by gene deletion and parallel analysis. *Science*, 285(5429):901–906, Aug 1999.

[324] K. Wolff, M. Vendruscolo, and M. Porto. Stochastic reconstruction of protein structures from effective connectivity profiles. *PMC Biophys*, 1(1):5, Jan 2008.

[325] S. Wu and Y. Zhang. A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. *Bioinformatics*, 24(7):924–31, Apr 2008.

[326] W. Xie and N. V. Sahinidis. A branch-and-reduce algorithm for the contact map overlap problem. In *RECOMB*, pages 516–529, 2006.

[327] Y. Xu, D. Xu, H. N. Gabow, and H. Gabow. Protein domain decomposition using a graph-theoretic approach. *Bioinformatics*, 16(12):1091–104, Dec 2000.

[328] D. P. Yee, H. S. Chan, T. F. Havel, and K. A. Dill. Does compactness induce secondary structure in proteins? a study of poly-alanine chains computed by distance geometry. *J Mol Biol*, 241(4):557–73, Aug 1994.

[329] S.-H. Yook, Z. N. Oltvai, and A.-L. Barabási. Functional and topological characterization of protein interaction networks. *Proteomics*, 4:928–942, 2004.

[330] T. Yu, X. Zou, S.-Y. Huang, and X.-W. Zou. Cutoff variation induces different topological properties: a new discovery of amino acid network within protein. *J Theor Biol*, 256(3):408–413, Feb 2009.

[331] Z. Yuan. Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. *BMC Bioinformatics*, 6:248, Jan 2005.

[332] M. J. Zaki, S. Jin, and C. Bystroff. Mining residue contacts in proteins using local structure predictions. *IEEE transactions on systems, man, and cybernetics*

Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society, 33(5):789–801, Jan 2003.

[333] L. Zhang and J. Skolnick. How do potentials derived from structural databases relate to "true" potentials? *Protein Sci*, 7(1):112–22, Jan 1998.

[334] H. Zhou and Y. Zhou. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci*, 11(11):2714–2726, Nov 2002.

[335] H. Zhou and Y. Zhou. Folding rate prediction using total contact distance. *Biophys J*, 82(1 Pt 1):458–63, Jan 2002.

[336] H. Zhou and Y. Zhou. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins*, 55(4):1005–1013, Jun 2004.

[337] R. Zimmer, M. Wöhler, and R. Thiele. New scoring schemes for protein fold recognition based on voronoi contacts. *Bioinformatics*, 14(3):295–308, Jan 1998.

[338] S. S. Zimmerman, M. S. Pottle, G. Némethy, and H. A. Scheraga. Conformational analysis of the 20 naturally occurring amino acid residues using ecepp. *Macromolecules*, 10(1):1–9, 1977.

Short Curriculum Vitae

For reasons of data protection, the Curriculum vitae is not published in the online version.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Declaration

I hereby declare that the thesis submitted herewith is my own work and that only the sources and aids listed have been used.

Berlin, Dezember 2011 Ioannis Filippis