Spring 2020

Network Approaches to Elucidate the Determinants of Protein Topology and Stability

Zeinab Haratipour
Old Dominion University, zhara001@odu.edu

Recommended Citation
Haratipour, Zeinab. "Network Approaches to Elucidate the Determinants of Protein Topology and Stability" (2020). Doctor of Philosophy (PhD), Thesis, Chemistry & Biochemistry, Old Dominion University, DOI: 10.25777/83d0-ez36
https://digitalcommons.odu.edu/chemistry_etds/49

NETWORK APPROACHES TO ELUCIDATE THE DETERMINANTS OF PROTEIN TOPOLOGY AND STABILITY

by

Zeinab Haratipour
B.S. July 2007, Islamic Azad University, Iran
M.S. May 2011, Islamic Azad University, Iran
M.S. May 2018, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

CHEMISTRY

OLD DOMINION UNIVERSITY
May 2020

Approved by:

Lesley Greene (Director)

Jing He (Member)

Chris Osgood (Member)

Jennifer Poutsma (Member)

ABSTRACT

NETWORK APPROACHES TO ELUCIDATE THE DETERMINANTS OF PROTEIN TOPOLOGY AND STABILITY

Zeinab Haratipour
Old Dominion University, 2020
Director: Dr. Lesley Greene

Predicting three-dimensional structures of proteins from sequence information alone, remains one of the most profoundly challenging and intensely studied problems in basic science. It has uniquely garnered the interdisciplinary efforts of biologists, biochemists, computer scientists, mathematicians and physicists. The advancement of computational methods to study fundamental features of proteins also enables insights that are either difficult to explore experimentally or complimentary to further interpret experimental data. In the present research and through the combined development and application of molecular dynamics and network science approaches we aimed to elucidate the role of geographically important amino acids and evolutionarily conserved long-range interactions which are proposed to be key to protein stability and topology. Using a model system of nine proteins that share a Greek-key topology, the proteins were unfolded under high temperature with molecular dynamics simulations. The unfolded trajectories were analyzed by calculating root-mean-square-deviation, contact distances, root-mean-square-fluctuation and fraction of remaining contacts. The results indicated that the conserved long-range interactions are significantly more persistent over time than the non-conserved long-range interactions thus dominant contributors to topological stability. The behavior of the conserved long-range interactions in the folding of our model proteins was also tested using simulated annealing and the formation of giant network clusters. The results demonstrated that the conserved interactions play a dominant role in folding by governing the native topology and facilitating rapid formation of the native network. In a third study, the role of the residues with high betweeness centrality scores in maintaining the protein network and in governing the Greek-key topology were examined by fragmentation and diameter tests. Here we found a subset of selected residues in similar geographical positions in all model proteins, which demonstrates the role of these specific residues and regions in governing the Greek-key topology from a network perspective. In conclusion, we can say that the determination of protein topology in terms of a network structure will facilitate predicting the folding and stability of proteins.

This dissertation is dedicated to my husband for his support, encouragement, and constant love during the challenges of graduate school and life. I am truly thankful for having you in my life.

ACKNOWLEDGMENTS

Throughout my PhD dissertation I have received a prodigious deal of support and assistance. I would first like to express my sincere gratitude to my advisor Dr. Lesley Greene for her invaluable support throughout all these challenging years, for her strong motivations whenever I got exhausted, for her thoughtful guidance when I got lost during my research, for being so patience with me to do my last year of Ph.D. remotely, while being with my husband far from ODU, for all the long hours of Facetime calls and finally, thank you for always leaving your office door open for me to reach out whenever I ran into a trouble or had a question about my research or writing.

I would like to whole-heartedly thank my dissertation committee members, Dr. Jennifer Poutsma for all the advanced training and expert advice in the computer lab, Dr. Jing He and Dr. Chris Osgood for their great support and insightful comments. I thank my fellow labmates, John and Cherelle for all the scientific discussions, technical collaborations, conference trips, and for all the fun we have had in the last five years. My sincere thanks also go to the Department of Chemistry and Biochemistry at Old Dominion University, who provided me an opportunity to join their program and gave access to all the laboratories and research facilities. Thank you for the financial supports by offering me several teaching assistant positions to cover my tuition and expenses during these years. Without your precious support it would not be possible to conduct this research. Last but not least, I would like to express my deepest gratitude to my parents, who have always loved me unconditionally and whose encouragement and prayers have guided me throughout my life.

NOMENCLATURE

BC	Betweeness Centrality
CC	Closeness Centrality
CHARMM	Chemistry at HARvard Macromolecular Mechanics
CNS	Crystallography & NMR System
CO	Contact Order
Ig	Immunoglobulin
MD	Molecular Dynamics
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect
PDB	Protein Data Bank
RMSD	Root Mean Square Deviation
RMSF	Root Mean Square Fluctuation
SA	Simulated Annealing
SCOP	Structural Classification of Proteins database
SD	Standard Deviation
TM-Score	Template Modeling Score
3D	Three-Dimensional

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Protein Structure

The protein structure can be described at four hierarchical levels of complexity (Fig. 1); 1) Primary structure is a linear sequence of amino acid residues connected to one another by a peptide bond. 2) Secondary structure is the local arrangement of polypeptide backbone atoms without including the side chains. Secondary structure is found in the three forms of α-helix, β-sheet and loop. 3) Tertiary structure is three-dimensional (3D) arrangement of an entire polypeptide containing its side chain. 4) Quaternary structure is the arrangement of two or more polypeptide chains in a protein [1].

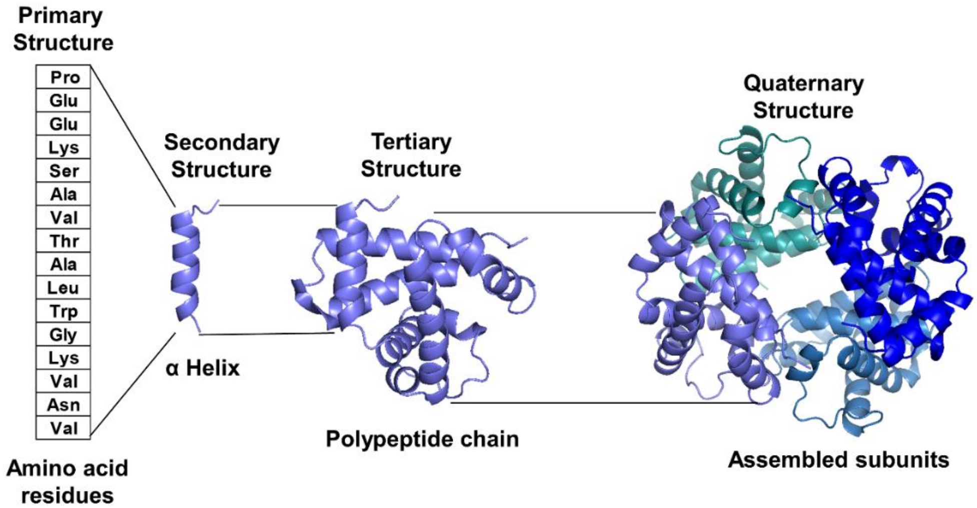


Fig. 1. Four hierarchical level of proteins structure; primary, secondary, tertiary, and quaternary structure. The protein structures were visualized using Pymol (version 2.1.1). Figure adapted from [1].

This dissertation was formatted based on the *Journal of Molecular Biology*

Two prominent protein structure classification databases are SCOP and CATH that classify proteins based on their structural domains. CATH sorts protein domains into four hierarchical levels including, class, architecture, topology or fold, and homologous superfamily [2, 3]. Major classification levels of SCOP are class, fold, superfamily, and family [4, 5]. Despite the differences between these two databases, both classify proteins into three classes according to their secondary structures; all-α, all-β, and mixed α/β (Fig. 2). Then, these three classes of proteins can be further subdivided by their topology, that is according to how their secondary structural elements are connected and packed in space [6-8]. Further, the domains in a fold are grouped into superfamilies that have at least one distant common ancestor and then in SCOP they are further clustered into families that have a more recent common ancestor [7].

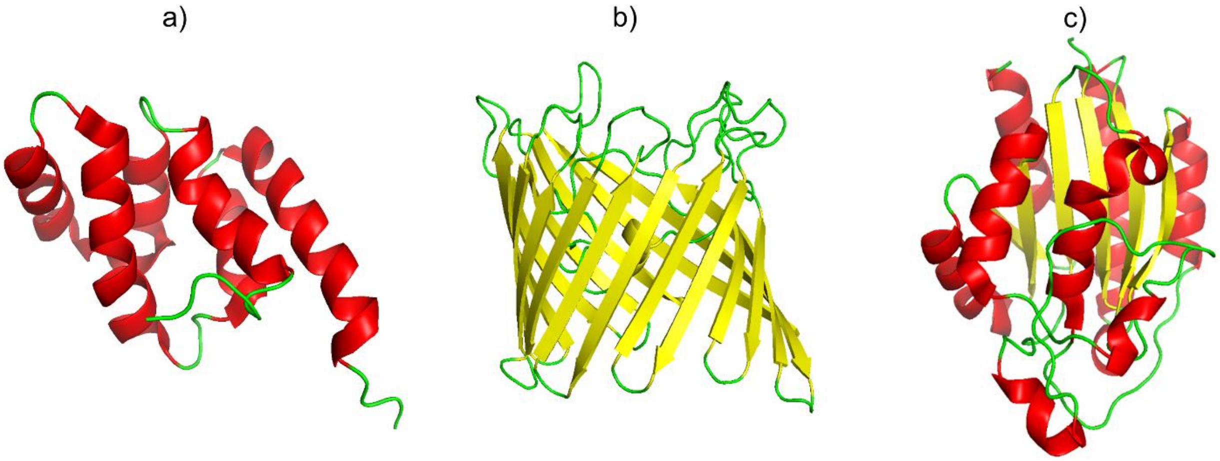


Fig. 2. Three classes of proteins. (a) all-α. (b) all-β. (c) mixed α/β. α-helices, β-strands and loops are shown in red, yellow and green, respectively.

Protein Interactions

The overall three-dimensional structure of proteins is stabilized by several forces such as, hydrophobic interactions, hydrogen bonds, salt bridges, van der Waals interactions, and disulfide bonds (Fig. 3). The non-bonded interactions can be classified by the distance between the interacting residues. Local- or short-range interactions are between residues that are close in both primary sequence and 3D space. Non-local or long-range interactions are defined as interactions between residues that are close in 3D space but distant in the primary structure. The local interactions, also known as short-range interactions are more significant in stabilizing the secondary structures, while the long-range interactions play a more important role in global protein stability and defining the native structure [9-18]. Several investigations have contributed to understanding the role of these interactions in the folding and stability of proteins [9-13, 15, 17, 19]. In 1975, Seiji Tanaka performed a Monte Carlo simulation of protein folding for bovine pancreatic trypsin inhibitor and showed the significance of the long-range interactions in the folding of this protein [19]. Not long after, Nobuhiro Go investigated a lattice model of a protein with the Monte Carlo simulation method and demonstrated that specific long-range interactions were essential for highly cooperative stabilization of the native conformation and that short-range interactions contribute to the acceleration of the folding and unfolding transitions [20]. In the late twentieth century, Michael Gromiha showed the importance of long-range interactions in 150 different globular proteins in terms of residue distances and, later, in 2001, the significance of long-range interactions in commonly occurring folds of globular proteins [16, 21]. Recently, several studies showed that the formation of long-range interactions early in folding can smooth the energy landscape, facilitate productive folding, guide its topology and may also prevent aggregation [22-24]. Such a prospect served as an enticing gain on the protein folding front, as well as a great stride in understanding mechanisms of diseased states.

Studying these interactions not only helps to understand protein folding and stability but also can calculate the proteins folding rate. Contact order (CO), which reflects the importance of the short-range and long-range interaction in the protein native structure, is the most important characterized determinant of protein rate [25-29]. CO is defined as the average sequence separation between contacting residues normalized by the total sequence length;

$$CO = \frac{1}{L*N} \sum_{i=0}^{N} \Delta S_{i,j} \qquad (1)$$

where N is the total number of contacts, $\Delta S_{i,j}$ is the sequence separation between interacting amino acids $i$and j, and L is the total number of amino acids in the protein [25]. Proteins with higher CO exhibit larger long-range networks, slower folding rates and more well-ordered transition states [25-29]. Formation of the long-range interactions in the early steps of folding could slow down the folding process and provide more time for the protein to arrange itself into a better structure.

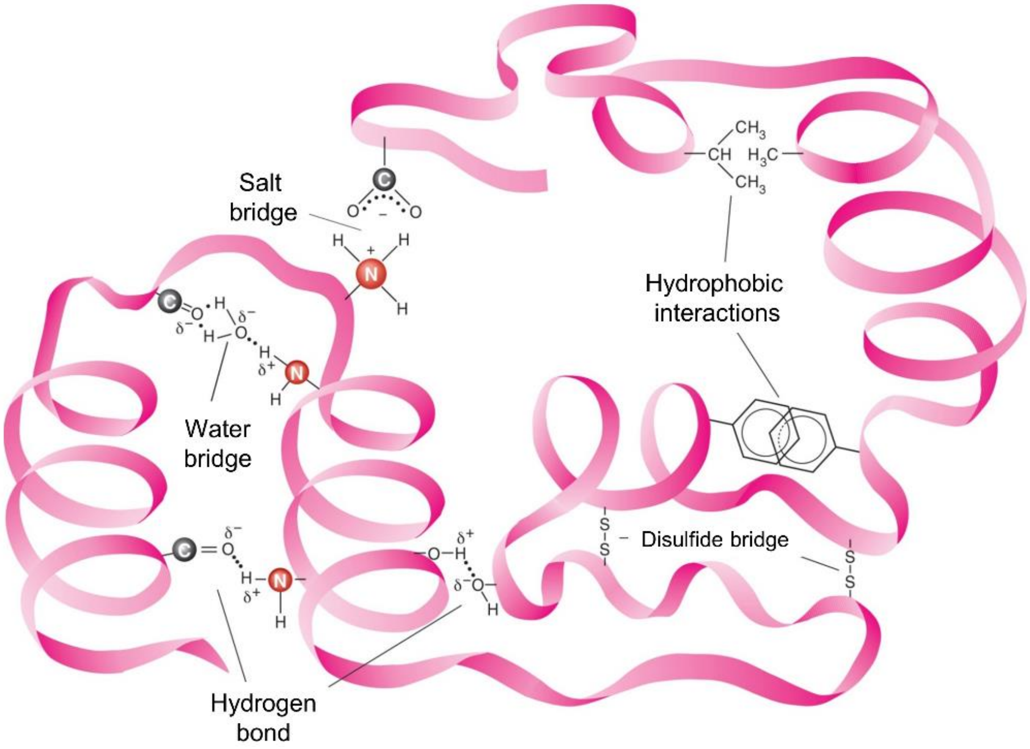


Fig. 3. Interactions which stabilize the protein 3D structure include hydrophobic interactions, hydrogen bonds, salt bridges, and disulfide bonds. Figure reproduced from [30].

Protein Folding

A comprehensive understanding of the mechanisms which underlie the folding of a primary structure into its native three-dimensional conformation and the *de novo* prediction of these structures *in silico* has not fully been resolved. Three folding models have been suggested; first one is the framework model in which formation of the secondary structural elements occurs before assembly of these elements into the final 3D structure. Second is the hydrophobic collapse model in which the folding reaction is initiated by a hydrophobic collapse in the interior of the protein molecule and finalizes with the growth of the secondary structural elements. In the third model called the nucleation-condensation mechanism, formation of a local nucleus of a small amount of secondary and tertiary structure forms and acts as a scaffold for further folding (Fig. 4) [31, 32].

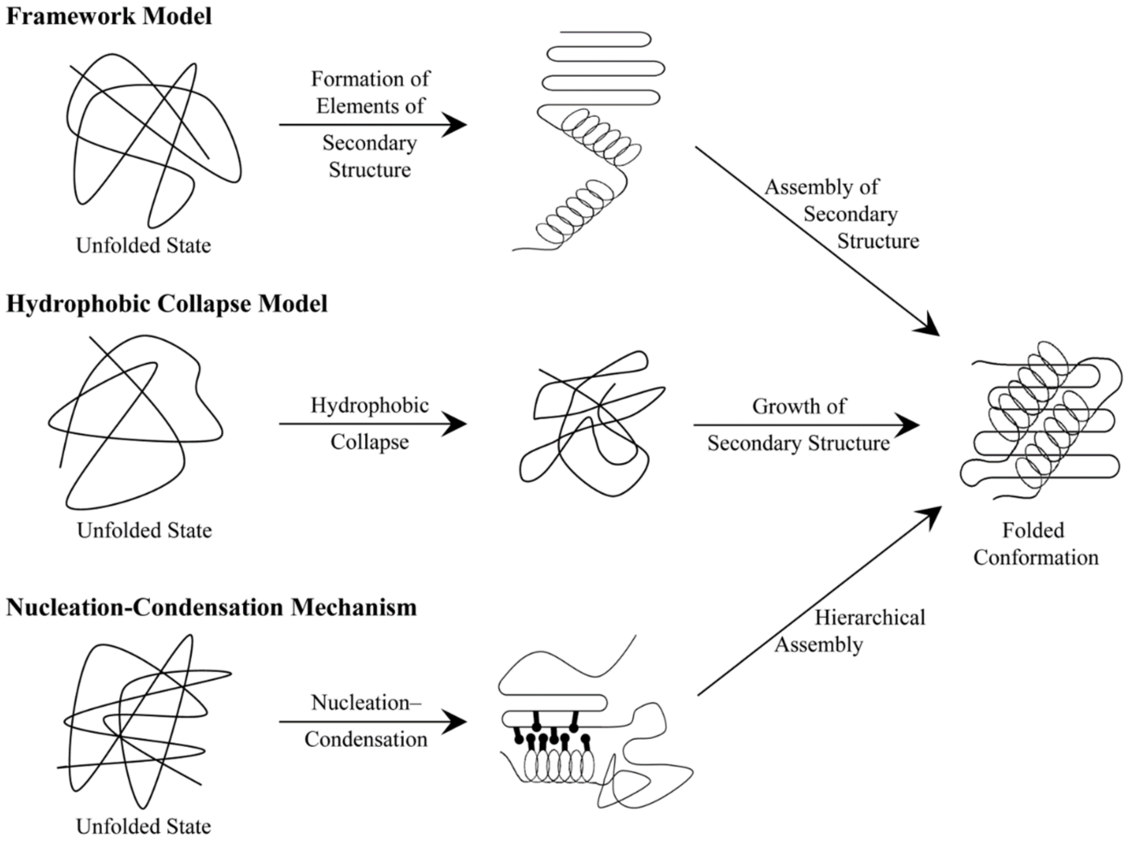

Framework Model
Formation of Elements of Secondary Structure
Unfolded State
Assembly of Secondary Structure
Hydrophobic Collapse Model
Hydrophobic Collapse
Growth of Secondary Structure
Unfolded State
Folded Conformation
Nucleation-Condensation Mechanism
Hierarchical Assembly
Nucleation–Condensation
Unfolded State

Fig. 4. Three suggested folding models; framework model, hydrophobic collapse model and nucleation-condensation mechanism. Figure reproduced from [32].

The nucleation-condensation mechanism which was pioneered by Alan Fersht, and co-workers at the university of Cambridge has been the focus of numerous experimental studies [33, 34]. Fersht and co-workers supported the nucleation mechanism by applying the experimental protein engineering technique on the 64-residue chymotrypsin inhibitor 2 (CI2). He showed a nucleation site developed in the transition state (TS) of CI2 as it folded. The nucleus consists of an α-helix, stabilized by long-range interactions forming as the rest of the protein collapses around it. The nucleus was determined by a protein engineering technique called ϕ-value analysis. It indirectly characterizes the structures of the TS in protein folding. In this analysis, interactions involving an amino acid are reduced or removed completely and kinetic and equilibrium measurements are performed on the unfolding and refolding of the resulting mutant to determine the extent of the interactions at different stages on the folding pathway [33, 34]. ϕ-value analysis is now a well-known method for studying the fleeting transition states of proteins, which are very difficult to study by NMR and impossible by x-ray crystallography. ϕ-value is defined as the ratio of changes to the free energy of activation for folding ($\Delta\Delta G_{\ddagger-D}$) and to the equilibrium free energy of folding ($\Delta\Delta G_{N-D}$,) shown by the following equation [35];

$$\phi_F = \Delta\Delta G_{\ddagger-D} \,/\, \Delta\Delta G_{N-D} \qquad (2)$$

ϕ-value range from 0 to 1. ϕ=0 indicates that the mutation did not affect the transition state and that interaction does not form in the TS. Conversely, ϕ=1 shows that the mutation disturbed the TS and thus is formed in the TS (Fig. 5). A newer method, PSI-value analysis has been developed to extend the ability to characterize the TS structure [36].

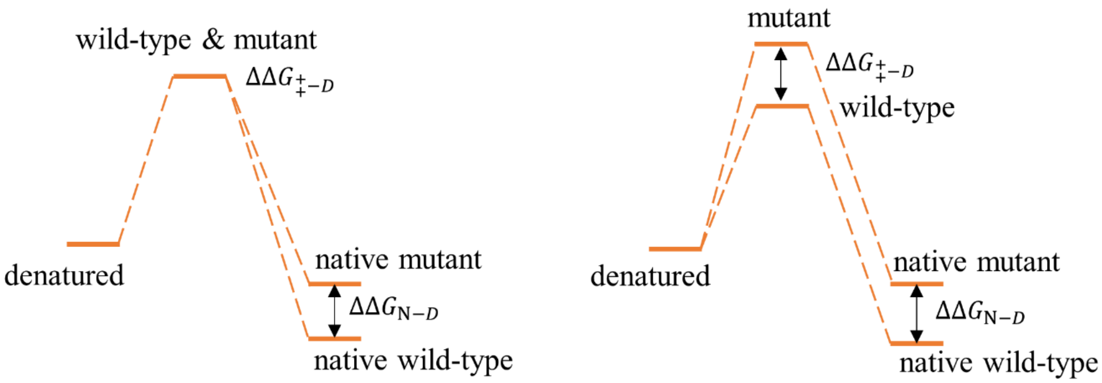

wild-type & mutant
$\Delta\Delta G_{\ddagger-D}$
denatured
native mutant
$\Delta\Delta G_{\mathrm{N}-D}$
native wild-type
mutant
$\Delta\Delta G_{\ddagger-D}$
wild-type
denatured
native mutant
$\Delta\Delta G_{\mathrm{N}-D}$
native wild-type

Fig. 5. Two extreme values of ϕ. The diagram on the left shows ϕ = 0, when the mutation did not affect the TS. The diagram on the right indicates ϕ = 1, when the mutation affects the TS. Figure adapted from [37].

In all folding models, a protein progresses from a high energy high-entropy state to a low energy-low-entropy state. This energy-entropy relationship is known as the folding energy landscape funnel (Fig. 6) [38]. The unfolded protein which has high energy, high entropy and low stability can take many different conformations and it is located on the top of the folding funnel. As the protein folds, it goes down through the funnel and the number of the possible conformations decreases as well as the energy and the entropy. At the bottom of the funnel, is the native structure of the protein with low energy, low entropy and high stability.

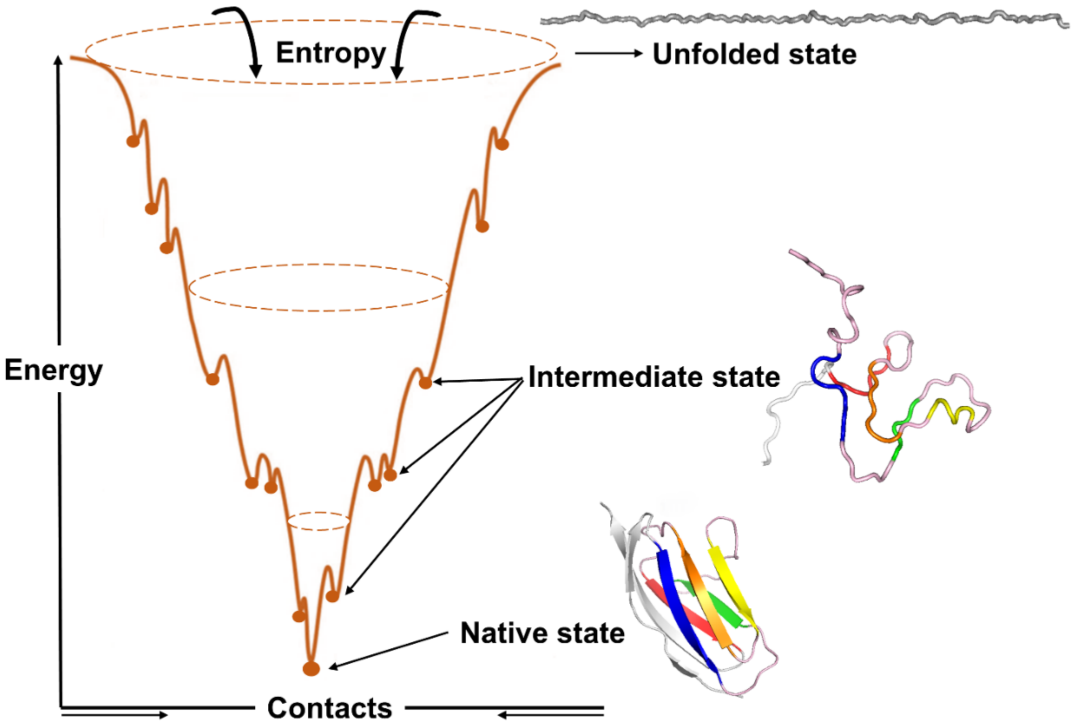


Fig. 6. The schematic of the folding energy landscape funnel. The folding of 1TIT is shown here. Figure adapted from [39].

Folding on the Ribosome and the Chaperone/Chaperonin System

The ribosome is the site of protein synthesis and is found in most living cells. This macromolecular machine is a ribonucleoprotein particle made of two subunits: the small subunit that read the messenger RNA and the large subunit that attach amino acids residues to one another to make a polypeptide chain [40]. The folding process can start co-translationally when a newly synthesized peptide is still in the ribosome [41]. The main questions are: When does a peptide start to fold? How does the ribosome affect the co-translational folding process? During the protein synthesis, the growing peptide passes through the ribosome’s exit tunnel. Small proteins can fold inside the tunnel, but the tunnel is too small for large molecules to make long-range interactions and fold inside. Therefore, formation of large tertiary structures should take place out of the ribosome and is unaffected by the ribosome,which is in agreement with the correct folding of the isolated proteins in vitro systems [41-46].

Proteins need to fold correctly to be functional. Misfolding not only alters the protein function but also causes many different diseases such as Alzheimer's and Parkinson's diseases. Molecular chaperones are a group of proteins that assist in protein folding. They can recognize and bind the non-native proteins to promote correct folding and prevent misfolding and aggregation. Chaperones also have the ability to unfold misfolded proteins and help them form the correct fold [40] (Fig. 7). Two major molecular chaperones are chaperonins and the Hsp70 chaperone system. The Hsp70 chaperone system facilitates the folding of denatured proteins in the ATP hydrolysis-dependent reaction cycle. Chaperonins form a double ring structure with a central cavity where the folding protein occurs. The formation of this chaperonin cage which isolate the denatured protein, can prevent aggregation. The space restriction in the chaperonin cage could help accelerate protein folding [47].

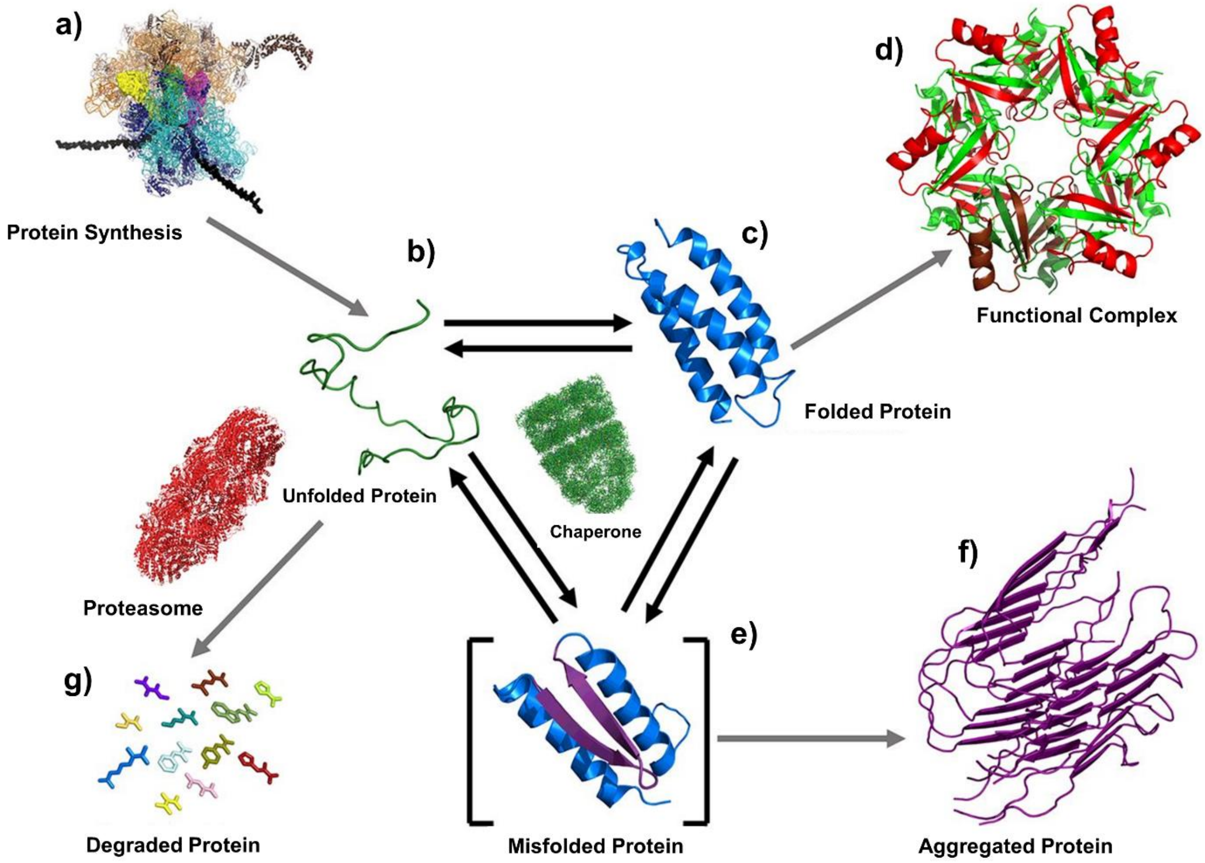


Fig. 7. Protein folding inside the cell. Proteins are synthesized in the ribosome. The newly synthesized polypeptide needs to fold properly to be functional. Chaperones assist the unfolded protein to fold correctly. Chaperones also can unfold the misfolded protein to help it to correct its fold. The misfolded proteins that cannot be fixed by chaperons, will be degraded by proteasome or will aggregate. Figure reproduced from [48].

Computational Approaches

Dynamics play an important role in the formation and functionality of macromolecules. Significant features of macromolecules such as protein folding can be understood only when dynamic properties are considered [49]. Proteins and nucleic acids need to undergo significant conformational changes while performing their biological functions. For instance; DNA must change its conformation and adjust to the protein surface while binding to the transcription factors [50, 51]. An enzyme's conformation is altered by binding to an allosteric regulator in the allosteric regulation process [52, 53]. In hemoglobin, the dome-shaped heme becomes planer after binding to an oxygen and causes a similar shift in the other three hemes to assist them to bind to more oxygen [54, 55]. Studying the macromolecular conformational changes *in vitro*, is a very complicated, expensive, and time-consuming process. Nuclear magnetic resonance (NMR) and X-ray crystallography are powerful experimental methods that can map the atomic positions. However, the molecules such as proteins and nucleic acids have picosecond level motions which are too fast for either X-ray crystallography and NMR spectroscopy to capture. Computer simulations have been designed to model such a quick movements that are difficult to access experimentally [56, 57].

Molecular dynamics (MD) is the most powerful simulations for modeling the physical movements of particles. In this simulation, particles are given velocities assigned to a selected temperature and allowed to move in response to all the forces acting on them in paths determined by Newton's equation of motion [57]. CHARMM (Chemistry at HARvard Macromolecular Mechanics) is one of the popular molecular simulations programs. This is a very flexible computer program that applies empirical energy functions to model macromolecular systems including proteins, peptides, lipids, nucleic acids, carbohydrates and small molecule ligands, crystals, and membrane environments [58, 59]. This program perform MD simulations by following steps: read or model the initial structure, perform energy minimizations by first- or second-derivative techniques, neutralize systems, perform equilibration, and simulates the motion of the system by numerically integrating Newton's second low of motion [59].

The following drives the dynamic equation of motion used for MD simulation. Given the atoms initial positions, xi (t_0), and their respective velocities, vi (t_0) at time t_0 and the position can be propagated forward using equation 3

$$x_i\,(t_1) = x_i\,(t_0) + v_i\,(t_0)\Delta t \tag{3}$$

The new velocities can be calculated from the old ones by equation 4

$$v_i\,(t_1) = v_i\,(t_0) + \Delta v_i\,(t_0) \tag{4}$$

Newton's equation ($F = ma$ *or* $F = mdV/dt$) can be used to calculate the change in velocity using equations 5 and 6

$$\Delta v_i\,(t_0) = \frac{F_{i\,(t_0)}}{m_i}\Delta t \tag{5}$$

$$v_i\,(t_1) = v_i\,(t_0) + \frac{F_{i\,(t_0)}}{m_i}\Delta t \tag{6}$$

where Fi is the sum of the forces acting on the ith particle, Thus

$$F(r) = -\nabla U(r) \tag{7}$$

$$U(r) = \sum U_{bonded}\,(r) + \sum U_{nonbonded}\,(r) \tag{8}$$

$$U_{bonded} = U_{bond} + U_{angle} + U_{dihedral} + U_{improper} \tag{9}$$

$$U_{nonbonded} = U_{LJ} + U_{elec} \tag{10}$$

$$U(r) = \sum\nolimits_{bonds} K_b\,(b - b_0)^2 + \sum\nolimits_{angles} K_\theta\,(\theta + \theta_0)^2 + \tag{11}$$

$$\sum\nolimits_{dihedral} K_\phi[1 + cos\,(n\phi - \delta)] + \sum\nolimits_{impropers} K_\omega\,(\omega - \omega_0\,)^2 +$$

$$\sum\nolimits_{Urey-Bradly} K_\mu\,(\mu + \mu_0)^2 + \sum\nolimits_{nonbonded} \varepsilon\,[(\frac{R_{min,ij}}{r_{rj}})^{12} - \left(\frac{R_{min,jj}}{r_{rj}})^6\right] + \sum\nolimits_{nonbonded} \frac{q_i q_j}{\varepsilon r_{ij}}$$

In the first term of equation 11, K_b is the bond force constant and b-b_0 is the distance from equilibrium for a given bond. In the second term K_θ is the angle force constant and $\theta + \theta_0$ is the degrees from equilibrium for a given angle. In the third term, K_ϕ is the dihedral force constant, n is the multiplicity, ϕ is the dihedral angle and δ is the phase shift. In the fourth term, K_ω is the force constant and $\omega - \omega_0$ is the out of plane angle. In the fifth term, K_μ is the force constant and $\mu + \mu_0$ is the distance from equilibrium of the 1,3-nonbonded interactions. In the sixth term, ε is the electric permittivity constant, r_{ij} is the distance between two nonbonded atoms in the configuration and $R_{min,ij}$ is the constant distance at which the potential is zero. In the last term, q_i and q_j are partial charges of atoms i and j. Figure 8 describes some of these terms.

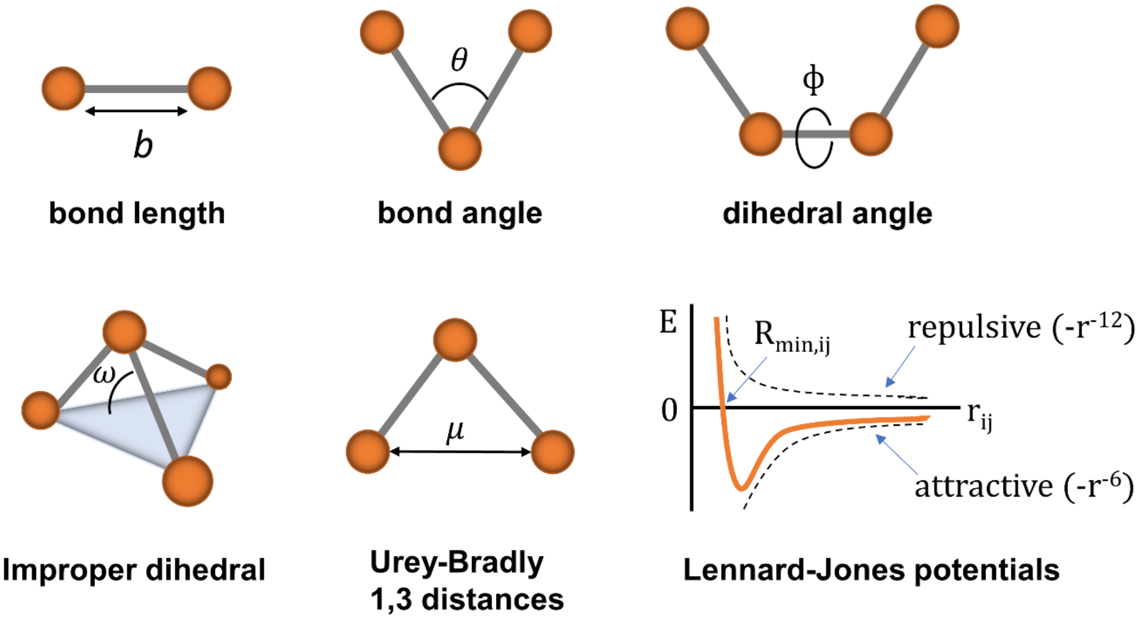


Fig. 8. Potential energy function terms of equation 10.

Significant advancements have been made toward improving and accelerating molecular dynamics simulations. The Anton supercomputer which is a parallel machine, has been designed to execute a millisecond-scale MD simulation with a macromolecule with millions of atoms [60].

Anton consists of 512 processing nodes which contain a specialized MD computation engine. The machine applies both special-purpose logic and novel parallel algorithms to accelerate the time dependent calculations in MD simulation [60, 61]. Minimizing the MD simulation timescale has been essential to making progress in the fields of biology and chemistry for studying the structures and motions of macromolecules.

Simulated annealing is another simulation method for studying macromolecular structures. Simulated annealing is so powerful that it can convert an extended protein structure into a well-defined 3D structure through the application of Nuclear Overhauser Effect (NOE) distance restraints [62]. CNS (Crystallography & NMR System) is program that can perform simulated annealing. The software is used to determine the structure of macromolecule from X-ray crystallography data or solution nuclear magnetic resonance (NMR) spectroscopy data [62, 63]. In the simulated annealing protocol, the starting structure is energy minimized and then heated at several intervals until the system gains enough energy to cross energy barriers. The atomic positions at the end of a simulations step are determined from their starting positions, as well as from their velocities and accelerations, which in turn are both derived from the starting positions using equations 1 and 2. Velocities are calculated from the Maxwell distribution at a given temperature and accelerations are determined by Newton's equation of the force field [64]. In the next step, the temperature is gradually reduced in intervals to develop a system under the influence of the potential field. By cooling down the system, the number of possible conformations, energy and entropy decrease until a minimum energy protein structure is reached. After each simulation step, the energy potential is recalculated for the new atomic positions and a further simulation step follows.

$$V = E_{empirical} + E_{effective} \tag{12}$$

with:

$$E_{effective} = E_{NOE} + E_{torsion} \tag{13}$$

and

$$E_{empirical} = E_{bond} + E_{angle} + E_{dihedral} + E_{vdw} + E_{elect} \tag{14}$$

$E_{empirical}$ contains all information about the primary structure of the protein and data about the topology and the bonds in the protein and $E_{effective}$ takes the experimentally determined constraints into account.

The application of network science is also a valuable approach to study protein structures. Proteins can be considered as a network where amino acids are nodes and interactions between them are edges (Fig.9) [65]. Network theory has been increasingly applied to describe the stability, folding, dynamics and function of proteins [66]. The nature of networks is well designed to determine the residues and interactions that are the major determinants in protein topology and stability [67-73].

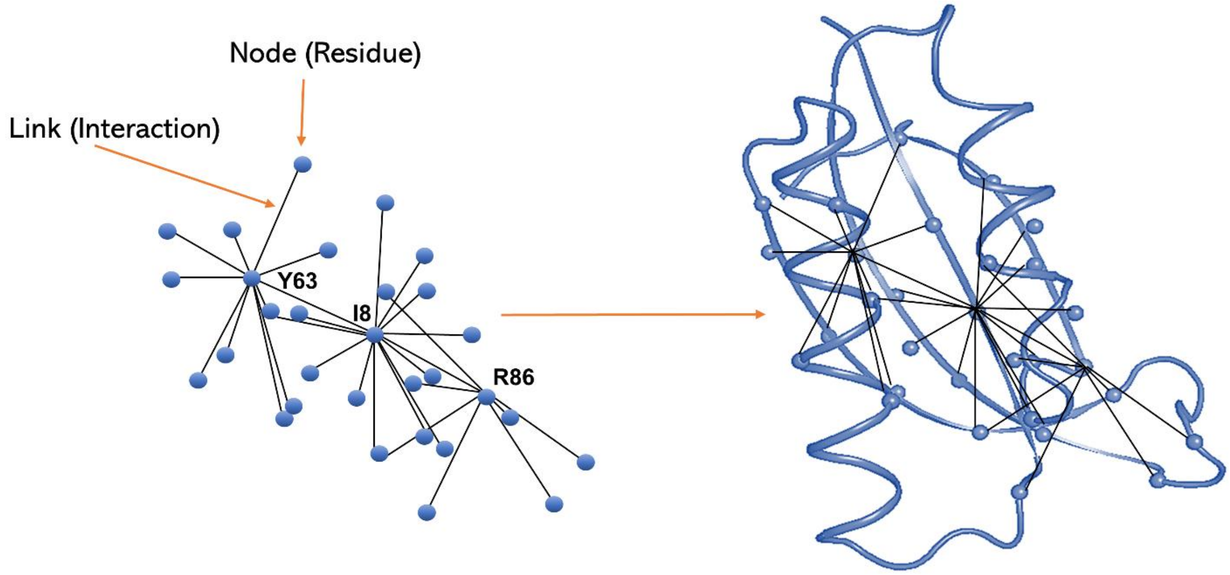


Fig. 9. Subset of the long-range interaction network in 1RIS. Amino acid residues are considered as nodes and long-range interactions as links. The long-range network is shown both inside and outside the protein. the names of highly connected residues are shown in the network.

In the past few decades, computational structure prediction of proteins has greatly advanced in order to address the large fraction of sequences whose structures cannot be determined experimentally [74, 75]. There are two major classes of protein structure prediction: comparative modeling and *de novo* methods. Comparative modeling predicts the structure based on the detectable similarity with a known structure. Second, *de novo* or *ab initio* methods predict the structure based only on the amino acid sequence [76, 77]. To date, knowledge-based methods, which extract information from solved protein structures has been more reliable. However, the yearly evaluation, Critical Assessment of Techniques for the Protein Structure Prediction (CASP) has shown a significant improvement in the *ab initio* methods [78, 79]. For instance, the folding of larger proteins (> 100 residues) have been performed successfully by applying the co-evolution-based contact map predictions [77]. To increase progress in *ab initio* methods, parallel improvement of accurate potential energy functions and efficient optimization methods are both needed [77].

Protein structure prediction includes four levels: 1D prediction of secondary structures, 2D prediction of the spatial relationship between amino acids such as distances, 3D prediction of tertiary structure, and 4D prediction of quaternary structures. The 1D prediction has been studied the most and it has a critical role in the development of protein structure prediction methods. The main goal is the predication of the 3D structure of proteins, so 1D and 2D predictions are applied as input for 3D coordinate predictors [80].

Machine learning methods, which are an important class of tools to automatically extract useful information from the protein data bank, have been widely used in all levels (1D to 4D) of protein structure prediction [80]. In the latest critical assessment of CASP, machine learning methods, including neural networks, self-organizing maps , and support vector machines have shown great improvement in protein structure prediction, but progress remains to be made in both the accuracy and scope of these methods [80-82]. The latest developments and progress in protein structure prediction reported by CASP includes the following: 1) new techniques for predicting 3D contacts can create the impressive template free models, however the template-based models are still the most accurate; 2) more focus on modeling the quaternary structure of proteins, in collaboration with the Critical Assessment of Predicted Interactions (CAPRI), demonstrated good results, but there is still room for improvement in this area; 3) refinement of primary protein models now improves nearly all models; 4) the application of sparse NMR constraints can completely advance the accuracy of models; 5) there is a great advancement in methods that estimate the accuracy of models [74, 75, 78]. Some of the well-known and accurate protein structure predication algorithms are Amber [83], CHARMM [59], UNRES [84], ROSETTA [85], I-TASSER [86].

Research Aims

The aim of this dissertation project was to conduct a comprehensive investigation to determine which amino acid residues and long-range interactions have a critical role in the topological determination and structural stability of proteins. To achieve this goal, five different computational approaches were applied to study the structures of a group of proteins that share a common Greek-key topology (Fig.10) but differ in sequence, secondary structure and function. Nine proteins constitute our model system and were selected from the following three superfamilies: the death domains, α/β-plaits, and immunoglobulins, which are classified as all-α, mixed α/β, and all-β, respectively. These proteins are listed by superfamily, species, name and PDB code: Death domains [human death domain of the FAS-associated death domain -1E3Y [87]; human death effector domain - 1A1W [88]; human iceberg - 1DGN [89]], α/β plaits [human fourth metal-binding domain of the Menkes copper-transporting ATPase - 2AW0 [90]; *Thermus thermophilus* ribosomal S6 - 1RIS [91]; bovine acylphosphatase - 2ACY [92]] and Immunoglobulins [human titin - 1TIT [93]; turkey telokin - 1TLK [94]; human tenascin - 1TEN [95]]. These three superfamilies were chosen because they share the highly populated Greek-key topology. While this topology was originally attributed to the Ig fold, it was shown that this topology also underlies the α/β-plait fold and death domain fold [96].

In aim one, the role of the evolutionary conserved long-range interaction networks in the structural stability of protein models was tested using molecular dynamics simulations. Proteins were unfolded under high temperature conditions and then the unfolded trajectories were analyzed by four different methods. First, the proteins structural stability was examined by calculating the root-mean-square deviation (RMSD) at different temperatures. Then, the distances and fluctuations of the conserved contacts was compared to those of the non-conserved contacts as the protein unfolds. In the last analysis, the fraction of the contacts remaining was calculated for both the conserved and non-conserved contacts during the unfolding process.

The principle aim of the second study was to examine the role of the evolutionary conserved network of interactions in the topological determination and folding of the protein models by two computational methods. Initially, the extended form of each protein was folded by employing conserved interactions as the physical constraints using the simulated annealing method to form the gross native-like topology. Then, the giant cluster method was used to confirm the importance of these conserved interactions in the rapid formation of the protein's network structure. The last aim was designed to test the idea that the residues with high betweenness centrality scores are potentially significant in maintaining the protein network and in governing the Greek-key topology. This hypothesis is tested by two different computational methods: a fragmentation test and an analysis of diameter impacts.

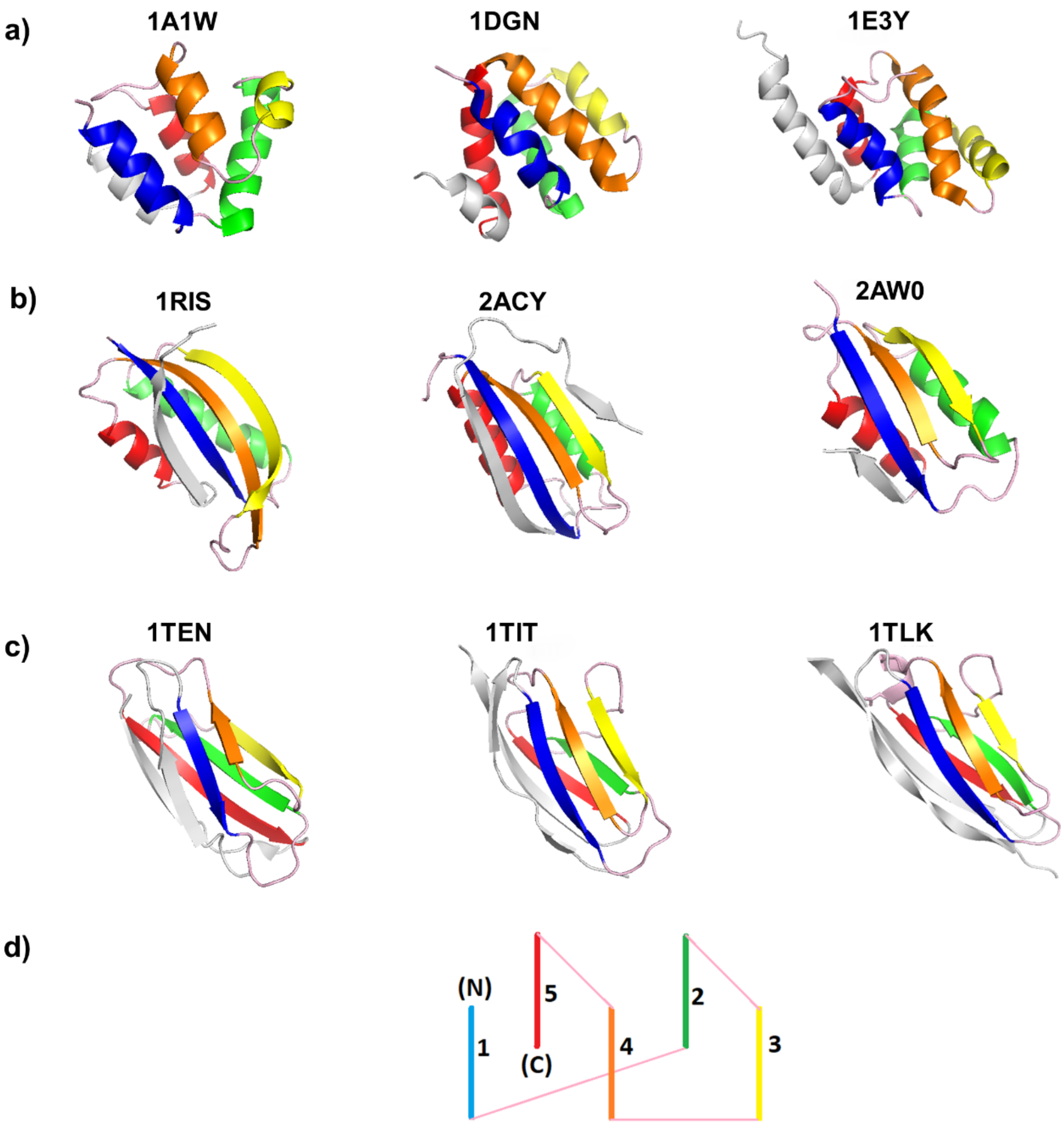


Fig 10. 3D structure of proteins color coded based on the schematic of the Greek-key topology. (a) Death domains: 1E3Y, 1DGN, 1A1W. (b) α/β plaits: 1RIS, 2ACY, 2AW0. (c) Immunoglobulins: 1TEN, 1TIT, 1TLK. (d) Schematic of the Greek-key topology.

CHAPTER II

ROLE OF THE CONSERVED LONG-RANGE INTERACTION NETWORKS IN THE STRUCTURAL STABILITY OF PROTEINS

INTRODUCTION

Protein stability stands at the nexus of structure, folding and function. Many different factors including alteration of solvent, temperature, PH, and forces, mutation, ligand binding, ion binding and many other factors, affect the protein folding process and so affecting the structural stability and biological function of proteins[13, 97-106]. Several computational and experimental approaches have been designed to predict the proteins structure and calculate their structural stability in terms of different factors [107-120]. The focus of this study is the computational analysis of protein structural stability considering the effect of the non-covalent interactions.

Protein structures are predominantly composed of a network of relatively weak short- and long-range non-covalent interactions between amino acids. Short-range interactions form and stabilize the secondary structures and long-range interactions organize and stabilize the arrangement of secondary elements in space to form the stable native topology [19, 20]. The significant role of the long-range interactions in folding and structural stability of proteins have been investigated by several experimental and computational studies [15, 16, 19, 20, 96, 121-123]. Long-range interactions have been found as a well-arranged network that govern and stabilize the proteins native structure [65, 96, 121, 123]. This long-range network includes all the non-bonded interactions from the hydrophobic interactions in the interior of the protein to the charge-charge interactions on the surface of the protein[9, 99, 124-136]. The long-range interactions have been also found in the denatured state or early stage of folding that it provides a smooth energy landscape and so productive folding [22-24, 137-140]. The role of the long-range interactions in protein folding is also defined in terms of contact order (CO), when CO is the average sequence separation between contacting residues normalized by the total sequence length [25]. Proteins with higher CO exhibited predominant long-range network, lower folding rate and more well-ordered transition states [25-29]. Therefore, the formation of the long-range interactions in the early steps of folding, slow the process and provide more time for protein to arrange itself into the more stable topology. All these observations show the important role of the long-range interactions in understanding the mechanism of protein folding, the basis of protein stability and the pathways of misfolding and aggregation.

In the present study, molecular dynamics simulations were applied to investigate protein unfolding and test the hypothesis that these conserved interactions (also referred to as contacts) are key determinants of structural stability. MD can uniquely probe stability at the level of individual amino acids and interactions in a protein as it is unfolding. This can provide atomic-level resolution of the major determinants of conformational stability [141]. For our studies, we used a group of proteins that share the Greek-key topology (Fig. 10), but differ in sequence, secondary structure and function. Three proteins were selected from three different superfamilies; the all α-helical death domains, the mixed α/β-plaits and all β-sheet immunoglobulins. The selected proteins were unfolded under high temperature using MD simulations to examine the stability and persistence of the conserved contacts in comparison to the non-conserved contacts (Fig. 11). Using three methods that analyze contact distance, fluctuation, and fraction of remaining contacts it has become evident that the conserved interactions are more persistent and thus, key determinants of structural stability.

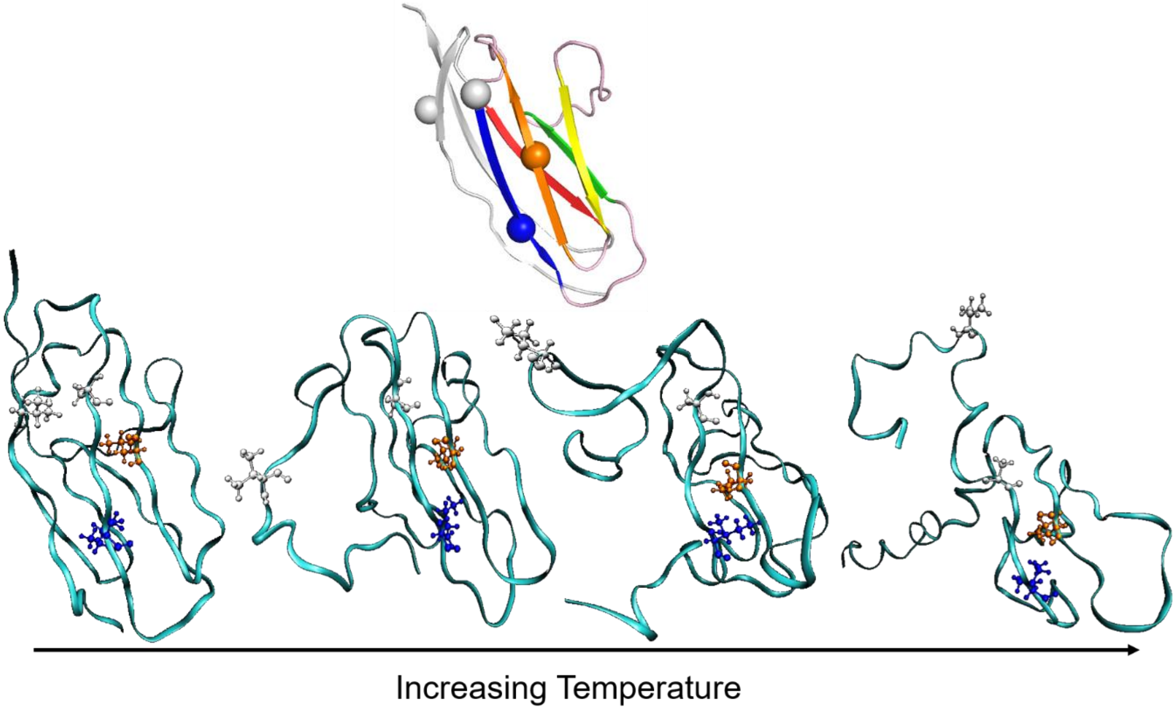


Fig. 11. Unfolding of titin using Charmm [58, 59]. 3D native structure of titin shown at the top. Orange and blue residues make a conserved contact and two white residues make a non-conserved contact. The conserved contact is still present at the end of the simulation indicating that this contact is more stable and persistent than the broken non-conserved contact.

MATERIAL AND METHODS

Molecular Dynamics Simulation Protocol

To test the role of the conserved contacts (Appendix A and E) in the structural stability of proteins, nine proteins from a group of 28 were selected to study using molecular dynamics simulations. The proteins are listed by superfamily, species, name and PDB code: Death domains [human death domain of the FAS-associated death domain -1E3Y [87]; human death effector domain - 1A1W [88]; human iceberg - 1DGN [89]], α/β plaits [human fourth metal-binding domain of the Menkes copper-transporting ATPase - 2AW0 [90]; *Thermus thermophilus* ribosomal S6 - 1RIS [91]; bovine acylphosphatase - 2ACY [92]] and Immunoglobulins [human titin - 1TIT [93]; turkey telokin - 1TLK [94]; human tenascin - 1TEN [95]].

In this study, proteins were examined under high-temperature conditions to see if their conserved interactions contain some inherent stability over other non-conserved contacts (Fig. 11) and, thus; are major stabilizing determinants of the Greek-key topology. First the system is energy minimized for 500 steps until obtain an energy tolerance of 0.001 kcal/mol. Second the system is solvated by placing in the octahedral box of water (Table 1) and then the complex of protein and water molecules are minimized for 100 steps. Next, the complex is neutralized by replacing the water molecules with ions such as Na^+ and Cl^-. Dynamics occurs in the isothermal–isobaric ensemble (NPT) under periodic boundary conditions. The electrostatic potential across the periodic boundaries is managed by Ewald algorithm. A switching function is applied to cover the van der Waals potential from 8Å to 12Å. The SHAKE algorithm is applied to restrain the lengths of all bonds involving hydrogen bonds. Equilibration is performed for 200ps to relax the system while it is heated. Dynamics is initiated from cooled structure and heated to 298K over 10K steps (20ps) of dynamics. All simulations are extended for 80 nanoseconds of production dynamics, using a 2-fs time-step. By saving coordinates every 1000 step, 40,000 frames per trajectory can be prepared for analysis. The MD simulations were run for each protein at several different temperatures (300K, 350K, 400K, 450K, 500K, 550K) using CHARMM (Chemistry at HARvard Macromolecular Mechanics) [58, 59]. The highest temperature is selected to be enough to unfold the protein. Matthew Robinson and Megan Barnes also collaborated in this study by performing MD simulation of three proteins for 20 nanoseconds out of 80 nanoseconds. The CHARMM simulations were run on Old Dominion University’ s Turing that is a high-performance computing clusters for parallel programming applications. The Turing cluster contains 258 multi-core compute nodes, each containing between 16 and 32 cores and 128 GB of RAM.

Table 1. Number of the water molecules applied in the solvation.

Proteins (PDB Code)	# of water molecules
1A1W	6329
1DGN	6324
1E3Y	8535
1RIS	9140
2ACY	6284
2AW0	6473
1TEN	6330
1TIT	6861
1TLK	6296

RMSD Calculation

In the first analysis, RMSD was calculated for each protein at different temperatures to test the protein conformational stability during the simulation. RMSD is known as one of the most common quantitative measures of the similarity between two superimposed atomic coordinate systems (usually the backbone atoms). RMSD values are mainly used for analyzing the stability and predicting conformational changes of a system [142]. RMSD values are presented in Å and calculated by

$$RMSD = \sqrt{1/n \sum_{i=1}^{n} d_i^2} \tag{15}$$

where n is the number of the atoms and d_i is the distance between the two atoms in the i^{th} pair.

Calculated RMSDs of the selected proteins were graphed vs. simulation time using the xmgrace plotting tool [143].

Analysis

Three sets of conserved contacts (D, T, TN) were assessed to see how they behave differently than the non-conserved contacts in the protein. Three different tests were applied to ascertain if the conserved contacts play a more important role in the protein structural stability than the non-conserved contacts. Initially, three sets of randomly selected non-conserved long-range contacts were formed to match each set of the conserved contacts (D, T, TN). Only random contacts that were as long as or shorter than the longest contact in each set were chosen.

In the first test, the distances for the conserved contacts and the randomly selected non-conserved contacts were calculated for the different temperature simulations. The shortest distance between non-hydrogen atoms in each residue was used as the contact distance. The contact distances of each set were averaged at each temperature for the whole simulation for both conserved and non-conserved contacts and then plotted using Microsoft excel 365. Temperatures at which the protein completely unfolded were not used for this analysis.

In the second test, RMSF was calculated for the conserved and non-conserved contact distances. The RMSFs for each contact set were averaged for each temperature simulation and plotted using Microsoft excel 365. RMSF is the fluctuation observed between the residues or atoms of a macromolecule. The atomic fluctuation shows the level of flexibility of a system during a simulation [144]. This calculation was performed for a set of conserved and three equivalent sets of non-conserved contacts for each network method (D, T, TN) independently. As with the contact distances, the RMSFs are only calculated for temperatures before the protein unfolds.

In the third test, the fraction of the conserved contacts remaining is compared with the fraction of the non-conserved contacts that are remaining as the protein unfolds. In each set of the conserved and non-conserved contacts, the fraction of the contacts remaining is the number still in contact divided by the original number of the contacts in the set. The ratio of the remaining contacts was plotted vs. simulation time using xmgrace plotting tool [143].

RESULTS AND DISCUSSION

RMSD Evaluation

As shown in Figure 12, root-mean-square-deviation (RMSD) is plotted vs. the simulation time for the selected proteins. RMSD increases with temperature because increasing temperature means increasing the kinetic energy, hence more fluctuations. In 1A1W, 1DGN, 2AW0 and 1TIT, there is hardly any changes in the RMSD at each temperature except 450K that is the unfolding trajectory. The RMSD stays almost flat at lower temperatures 300 and 350K for 1RIS, 2ACY, 1TEN and 1TLK, and it steadily increases at the higher temperatures. In 1E3Y, the RMSD start to increase at 450K and its almost unchanged for the lower temperatures. The highest temperature that unfolds the protein is indicated by an RMSD $\geq$ 10Å in all protein models. According to the simulation results, 1DGN, 1A1W, 2AW0, and 1TIT are unfolded at 450K. 1E3Y, 2ACY, 1TEN and 1TLK are unfolded at 500K and 1RIS is unfolded at 550K. The latter unfolds at higher temperature because it comes from a thermophile which is consistent with what we expected and additionally validates the methodology. As expected, the structure of all protein models remains close to the native structure at lower temperatures and they start to lose the topology as the temperature increase. Surprisingly, the RMSD of 1A1W and 1RIS at 350K is lower than 300K. These two proteins resist at 350K, but a little thermal shaking may adjust their conformations to be closer to the native structure. This unexpected result may also be explained by the limitations of RMSD calculation. Since the RMSD measures the distances between all amino acids pairs equally, a small number of local structural change could result in a high RMSD, even when the overall topologies of the compared structures are similar.

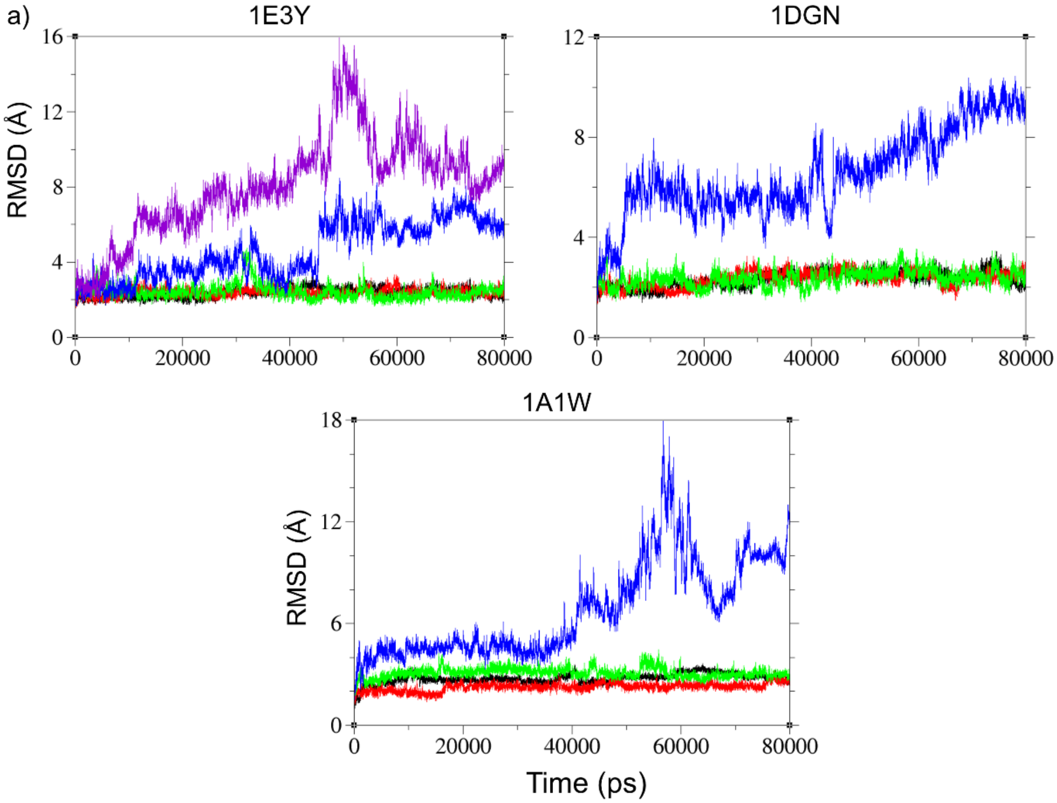
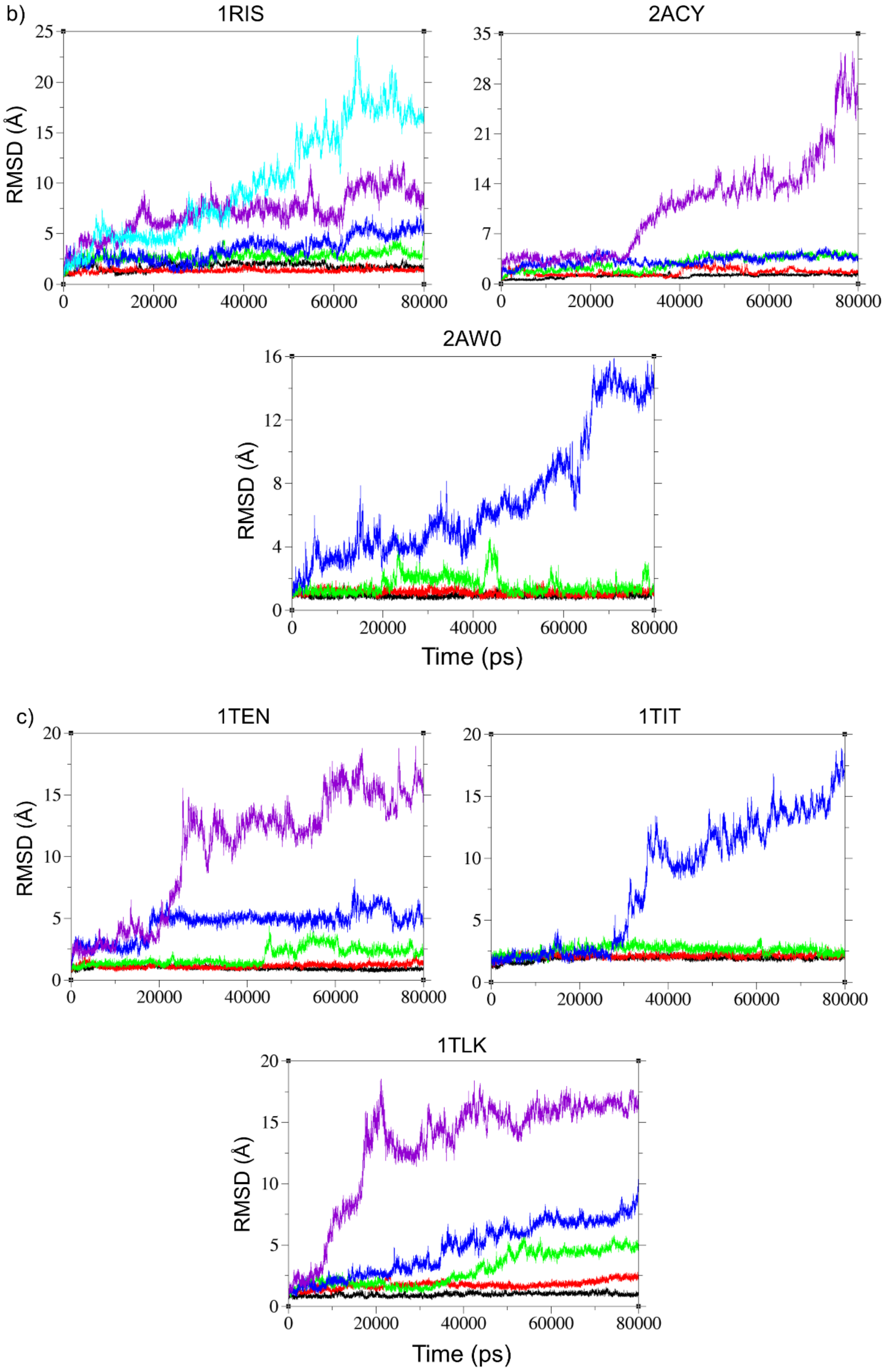


Fig. 12. Backbone RMSD vs. time at different temperatures ;300K, 350K, 400K, 450K, 500K and 550K are shown in black, red, green, blue, violet and cyan respectively. (a) Death domains: 1A1W, 1DGN, 1E3y (b) α/β plaits: 1RIS, 2ACY, 2AW0. (c) Immunoglobulins: 1TEN, 1TIT, 1TLK.

Figure 12 Continued



Calculating Contact Distances

In the first contact analysis, the contact distances were calculated as the protein unfolds. Each graph in Figure 13 displays the average contact distances at different temperature for a set of conserved contacts and three sets of randomly selected non-conserved contacts. To determine the rate of increase of the contact distances with temperature, linear least squares regression lines are fitted to the data. The results of this analysis agree with the RMSD calculations. For 1A1W, 1DGN, 2AW0 and 1TIT, the contacts distances increase smoothly at temperatures 300-400K, however there is a larger increase for 2AW0 at 400K. In 2ACY, 1TEN and 1TLK, there is small changes in the contacts distances from 300 to 350K but a larger increase at 400K and then more at 450K. The contacts distances of 1E3Y at 300-400K are close to each other and increases slightly at 450K. In 1RIS, there is a small increase in the contact distances at 300-450K and larger increase at 500K. The contacts distances of 1A1W at 350 is smaller than 300K that can be explained by the unexpected RMSD results.

As shown in the graphs, the slope of the conserved lines is less than those of the non-conserved lines. The slower increase in separation in the conserved contact distances indicates that they fall apart later which suggests that they play a more important role in protein stability than the non-conserved contacts. To quantify this comparison, the line’s slope of the randomly selected non-conserved contacts distances is divided by the line’s slope of the conserved contacts distances and defined as R/C (Table 2). The greater the R/C value over 1, the greater the stability of the conserved contacts over the random contacts. These results quantitatively confirm the stability of the conserved contacts in comparison to the non-conserved contacts. The average slope of the random lines is also calculated and divided by the slope of the conserved line to easier compare the types of networks (T, TN, D) in each protein. The direct network of 1E3Y, 1A1W, 1RIS, and

1TLK, and the TN network of 1DGN, 2ACY, 2AW0, and 1TEN, indicated the larger R/C value and more stability of conserved contacts. 1TIT is the only protein that show a higher R/C value for toggle network. According to these observations the direct and TN conserved networks show more stability during the simulation compared to toggle network. To easier compare the R/C value over the proteins, the average of R/C values of all three networks is calculated in each protein. The range of the calculated value for 1E3Y, 1DGN, 1A1W, 2AW0, 1TIT, and 1TLK, is 1.4-2.7 and for 1RIS and 2ACY is 3.4 and 4.5 respectively. 1TEN demonstrate the highest value around 7.5. According to this result, the conserved network of 1TEN shows the highest stability compared to the other proteins, but the slope of the conserved lines doesn’t confirm that. By comparing the slopes of the conserved and non-conserved lines in all nine proteins, it can be concluded that the R/C value of 1TEN is higher than the others because of the high slopes of the non-conserved lines, not the low slope of the conserved line. The difference of the R/C value between all nine proteins may be explained by the characterization of the randomly selected contacts. Since the non-conserved contacts are selected randomly, they may have different stability in different proteins. However, the conserved contacts demonstrated more stability compared to all the randomly selected non-conserved contacts in all proteins.

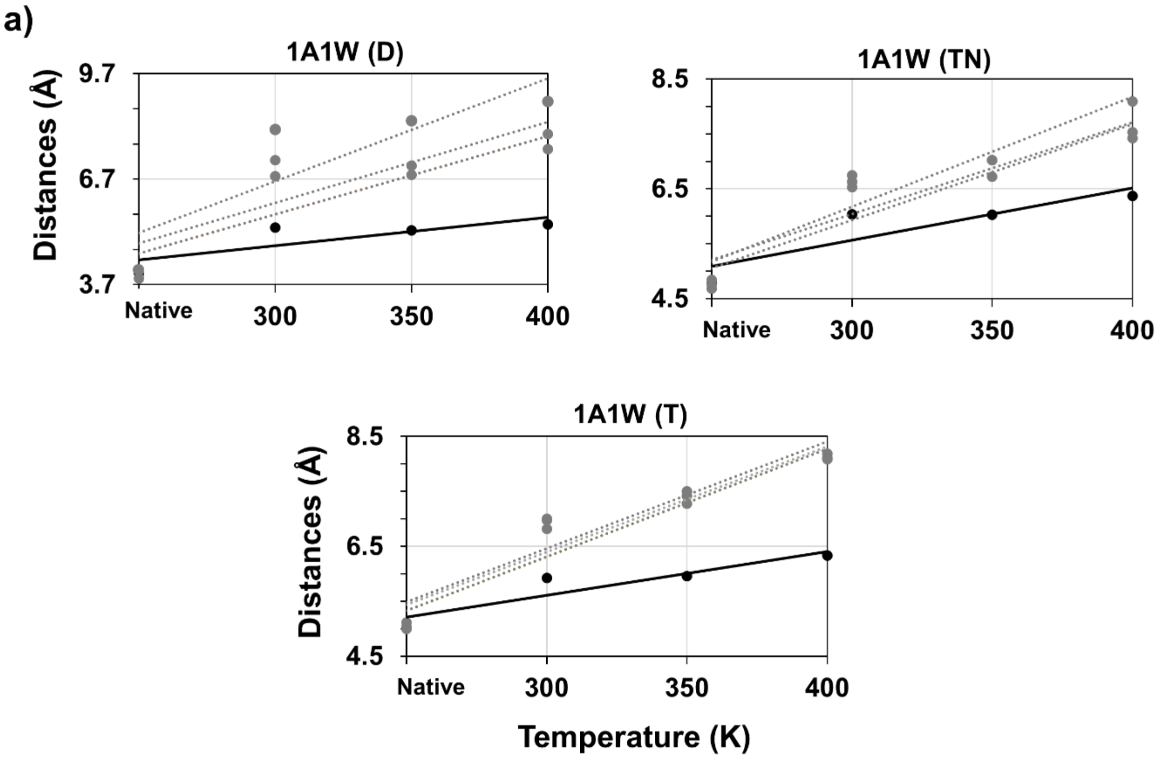


Fig. 13. Average contact distances vs. temperature. Solid and dashed lines display conserved and randomly selected non-conserved contacts, respectively. (a) Death domains: 1A1W, 1DGN, 1E3Y (b) α/β-plaits: 1RIS, 2ACY, 2AW0. (c) Immunoglobulins: 1TEN, 1TIT, 1TLK. D, T and TN are abbreviations of the terms Direct, Toggle and Toggle without acidic and basic residues.

Figure 13 Continued

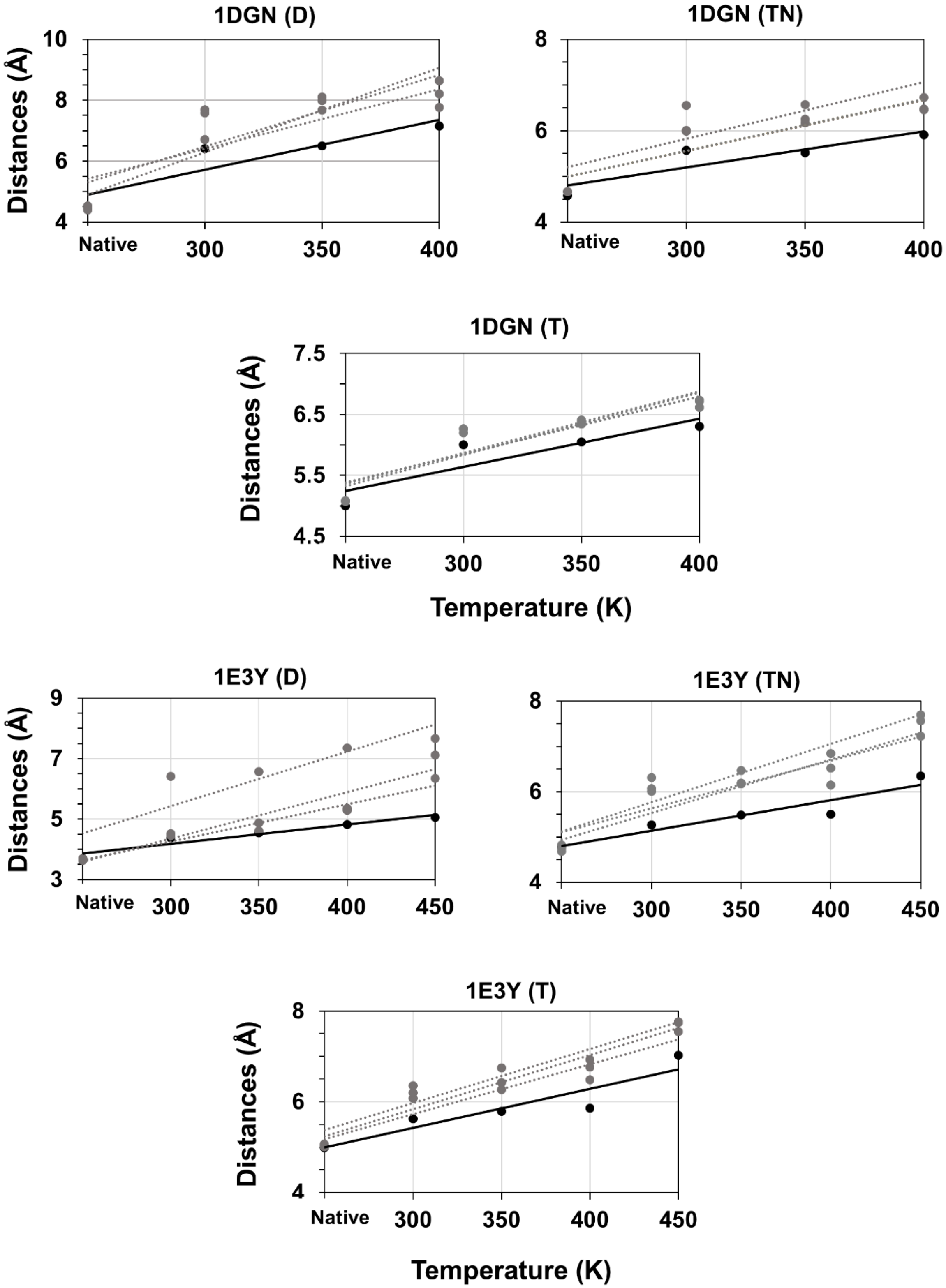


Figure 13 Continued

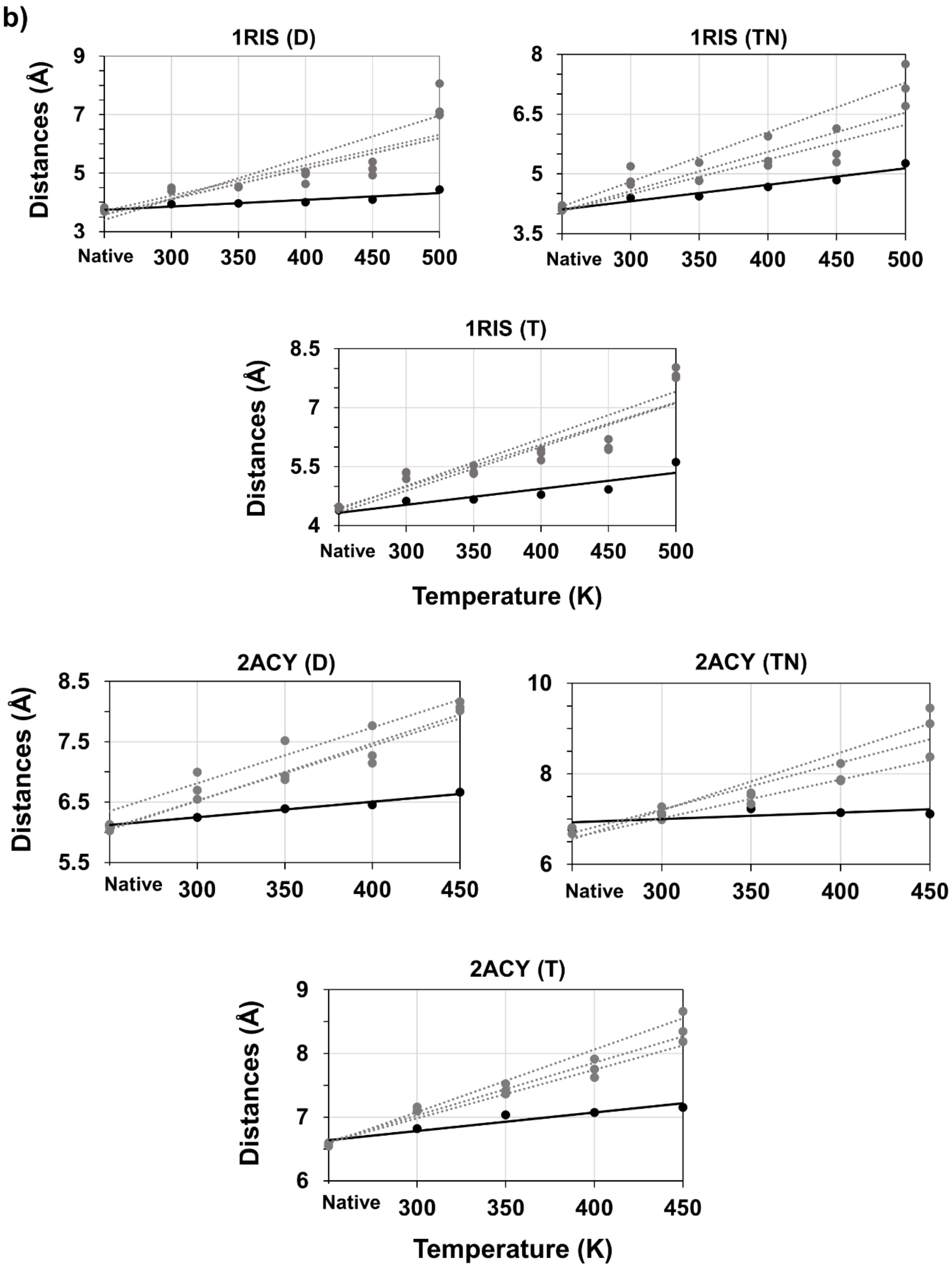


Figure 13 Continued

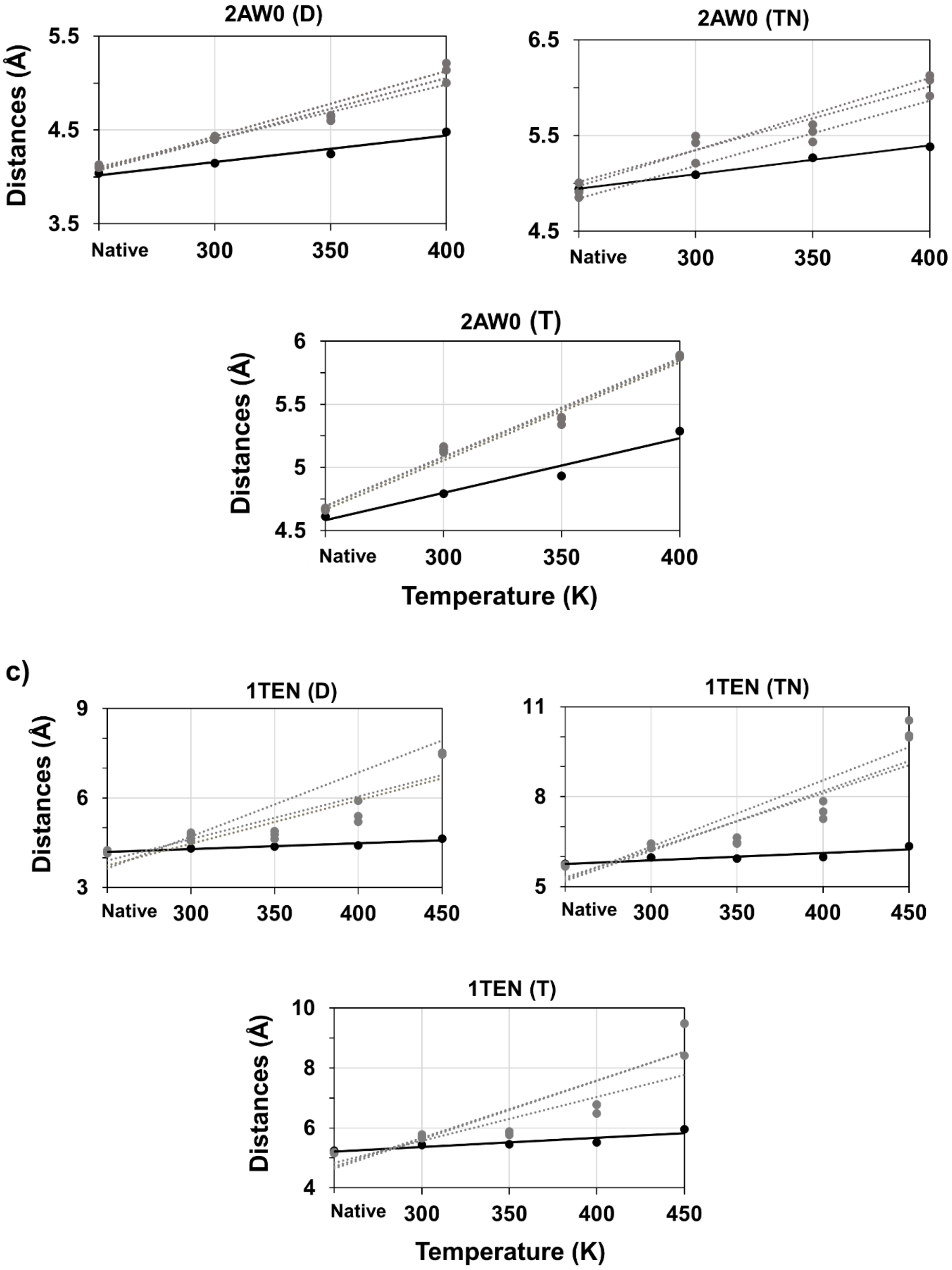


Figure 13 Continued

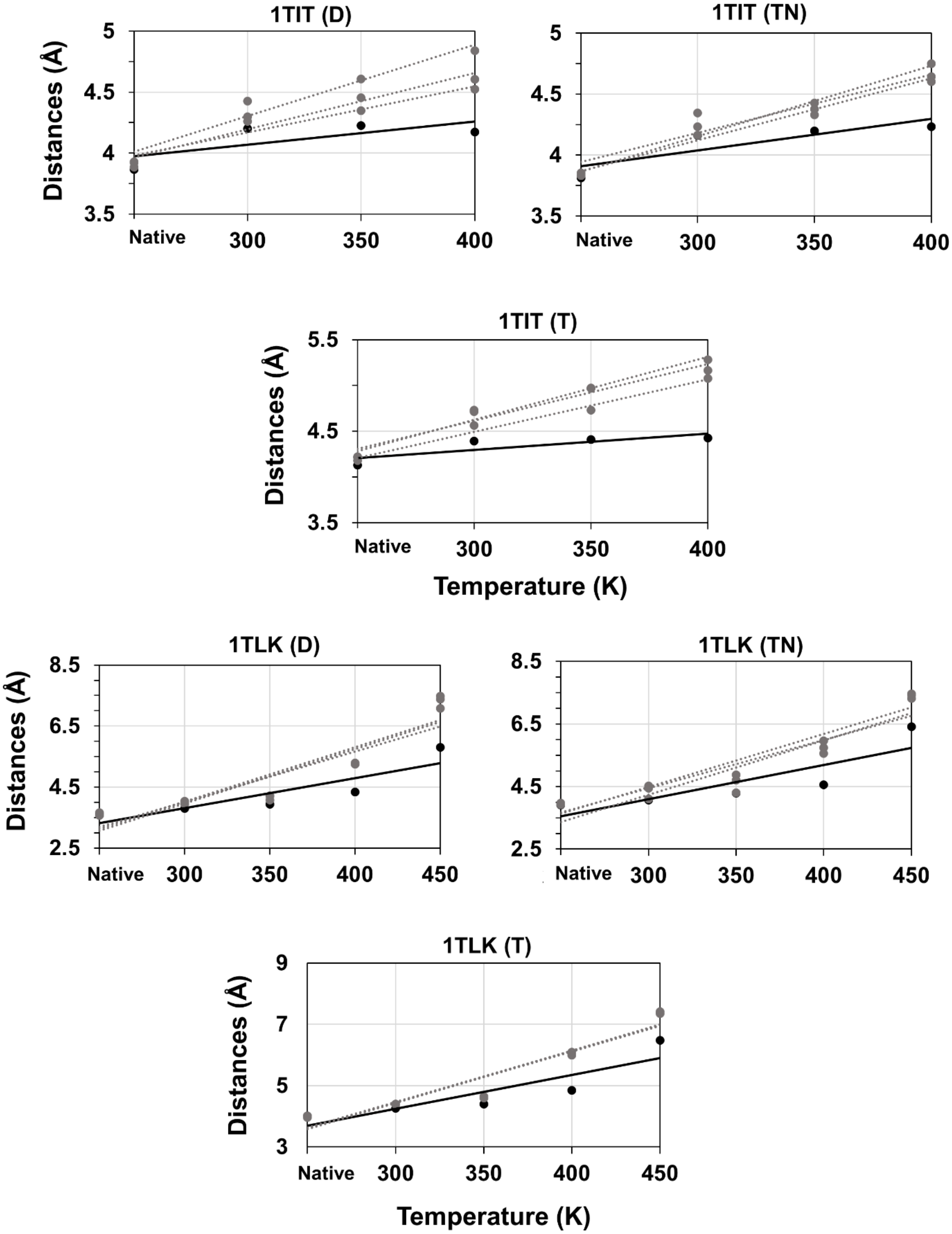


Table 2. Line's slopes of both conserved and randomly selected non-conserved contacts distances of Figure 13. R/C is the ratio of the line's slope of the random line over those of the conserved line.

PDB Code	C	R1	R 2	R3	R1/C	R2/C	R3/C	Avg.R /C	Avg. (Avg.R /C)
1E3Y (D)	0.0064	0.0123	0.0153	0.018	1.9218	2.3906	2.8125	2.3750	1.8263
1E3Y (TN)	0.0067	0.0105	0.0118	0.0129	1.5671	1.7611	1.9253	1.7512	
1E3Y (T)	0.0086	0.011	0.012	0.0119	1.2790	1.3953	1.3837	1.3527	
1DGN (D)	0.0163	0.0196	0.0235	0.0277	1.2025	1.4417	1.6994	1.4479	1.3968
1DGN (TN)	0.0079	0.0113	0.0113	0.0124	1.4304	1.4304	1.5696	1.4768	
1DGN (T)	0.0079	0.0094	0.0103	0.0103	1.1899	1.3038	1.3038	1.2658	
1A1W (D)	0.0081	0.0222	0.0231	0.0292	2.7407	2.8519	3.6049	3.0658	2.4656
1A1W (TN)	0.0095	0.0167	0.0175	0.02	1.7579	1.8421	2.1053	1.9018	
1A1W (T)	0.008	0.0192	0.0196	0.0195	2.4	2.45	2.4375	2.4292	
1RIS (D)	0.0023	0.0104	0.0104	0.0143	4.5217	4.5217	6.2174	5.0869	3.4544
1RIS (TN)	0.0041	0.0087	0.0098	0.0124	2.1220	2.3902	3.0244	2.5122	
1RIS (T)	0.0041	0.0112	0.0108	0.012	2.7317	2.6341	2.9268	2.7642	
2ACY (D)	0.0026	0.0093	0.0091	0.0096	3.5769	3.5	3.6923	3.5897	4.5368
2ACY (TN)	0.0015	0.0128	0.0104	0.0086	8.5333	6.9333	5.7333	7.0666	
2ACY (T)	0.0029	0.0083	0.0076	0.0098	2.8621	2.6207	3.3793	2.9540	
2AW0 (D)	0.0028	0.0058	0.0066	0.007	2.0714	2.3571	2.5	2.3095	2.1545
2AW0 (TN)	0.003	0.0068	0.0067	0.0076	2.2667	2.2667	2.5333	2.3556	
2AW0 (T)	0.0043	0.0077	0.0078	0.0077	1.7907	1.8140	1.7907	1.7985	
1TEN (D)	0.002	0.0146	0.0144	0.0214	7.3	7.2	10.7	8.4000	7.5216
1TEN (TN)	0.0024	0.0199	0.0188	0.0221	8.2917	7.8333	9.2083	8.4444	
1TEN (T)	0.0031	0.0146	0.0191	0.0195	4.7097	6.1613	6.2903	5.7204	
1TIT (D)	0.0019	0.0038	0.0046	0.0058	2	2.4211	3.0526	2.4912	2.6618
1TIT (TN)	0.0026	0.0048	0.0051	0.0058	1.8462	1.9615	2.2308	2.0128	
1TIT (T)	0.0018	0.0057	0.0062	0.0069	3.1667	3.4444	3.8333	3.4815	
1TLK (D)	0.0098	0.0165	0.0178	0.0179	1.6837	1.8163	1.8265	1.7755	1.6010
1TLK (TN)	0.0109	0.0154	0.017	0.0174	1.4128	1.5596	1.5963	1.5229	
1TLK (T)	0.0111	0.0167	0.0167	0.0167	1.5045	1.5045	1.5045	1.5045	

RMSF Evaluation

In the second contact analysis, the averaged contact distance root-mean-square-fluctuation (RMSF) is graphed vs. temperature for the selected conserved and non-conserved sets (Fig. 14). As we expected from the RMSD results, the RMSF of 1A1W, 1DGN, 2AW0 and 1TIT increase gradually from 300K to 400K. There is a smaller increase in the contacts fluctuation of 2ACY, 1TEN and 1TLK at 300-350K and larger increase at 400-450K. The RMSF of 1E3Y and 1RIS slightly change at 300-400K and increases at the higher temperatures. The results of this test confirm that the conserved contacts withstand more thermal shaking and do not fluctuate in distance as much as non-conserved contacts. Therefore, the conserved contacts demonstrate a more significant role in protein's structural stability.

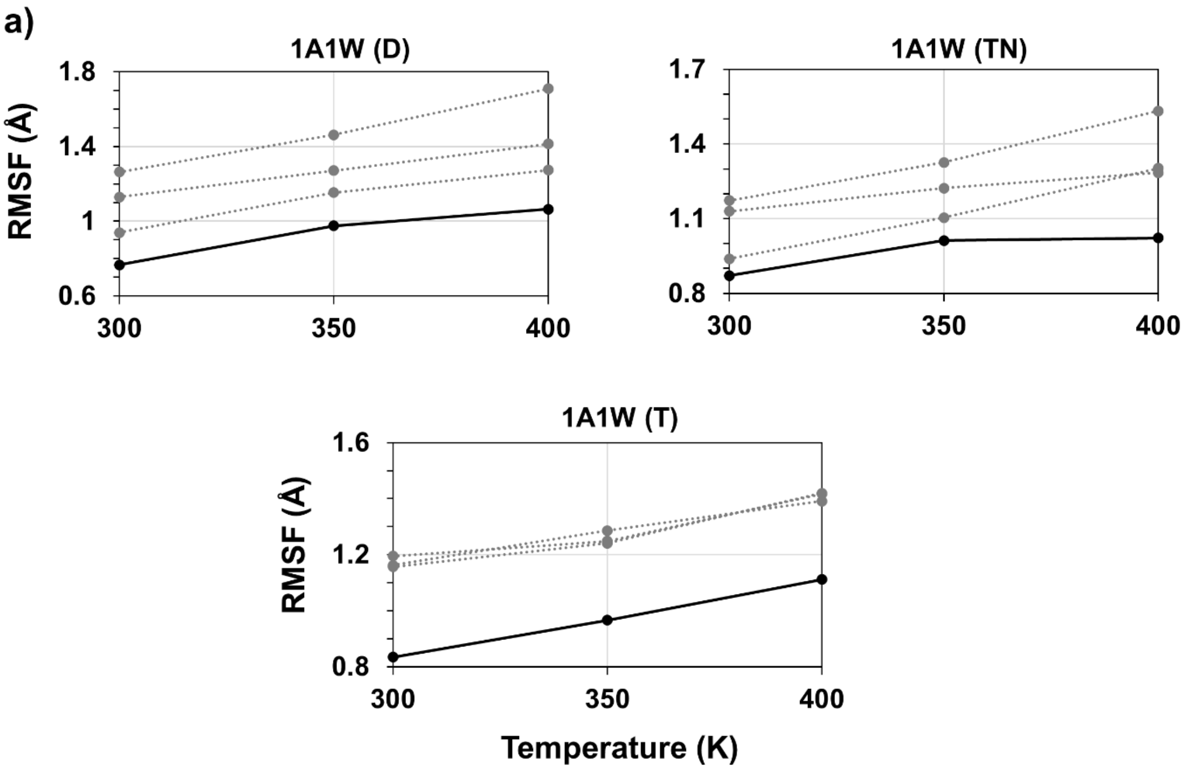


Fig. 14. Average contact distance RMSF vs. temperature. Solid and dashed lines display conserved and randomly selected non-conserved contacts, respectively. (a) Death domains: 1A1W, 1DGN, 1E3Y (b) α/β-plaits: 1RIS, 2ACY, 2AW0. (c) Immunoglobulins: 1TEN, 1TIT, 1TLK. D, T and TN are abbreviations of the terms Direct, Toggle and Toggle without acidic and basic residues.

Figure 14 Continued

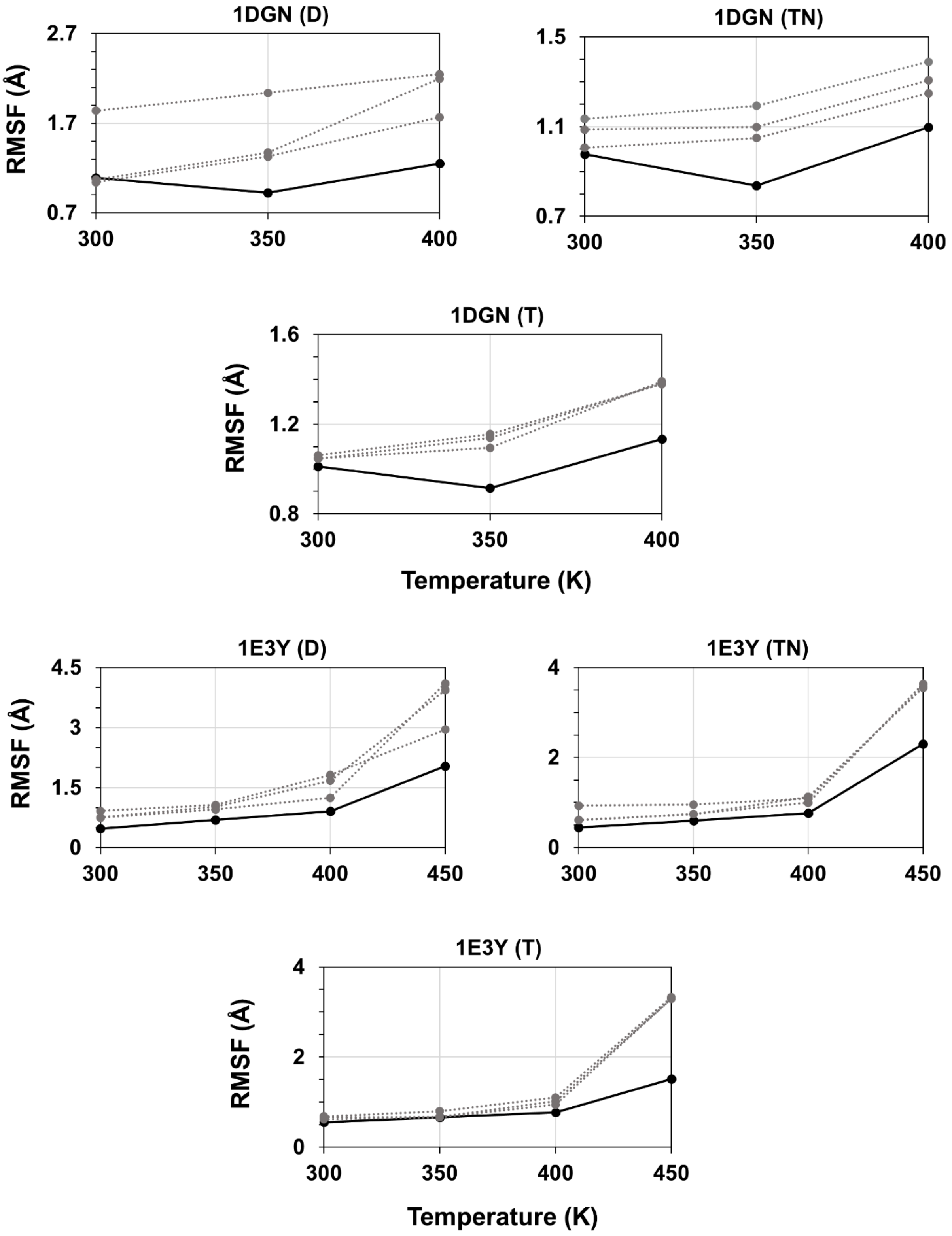


Figure 14 Continued

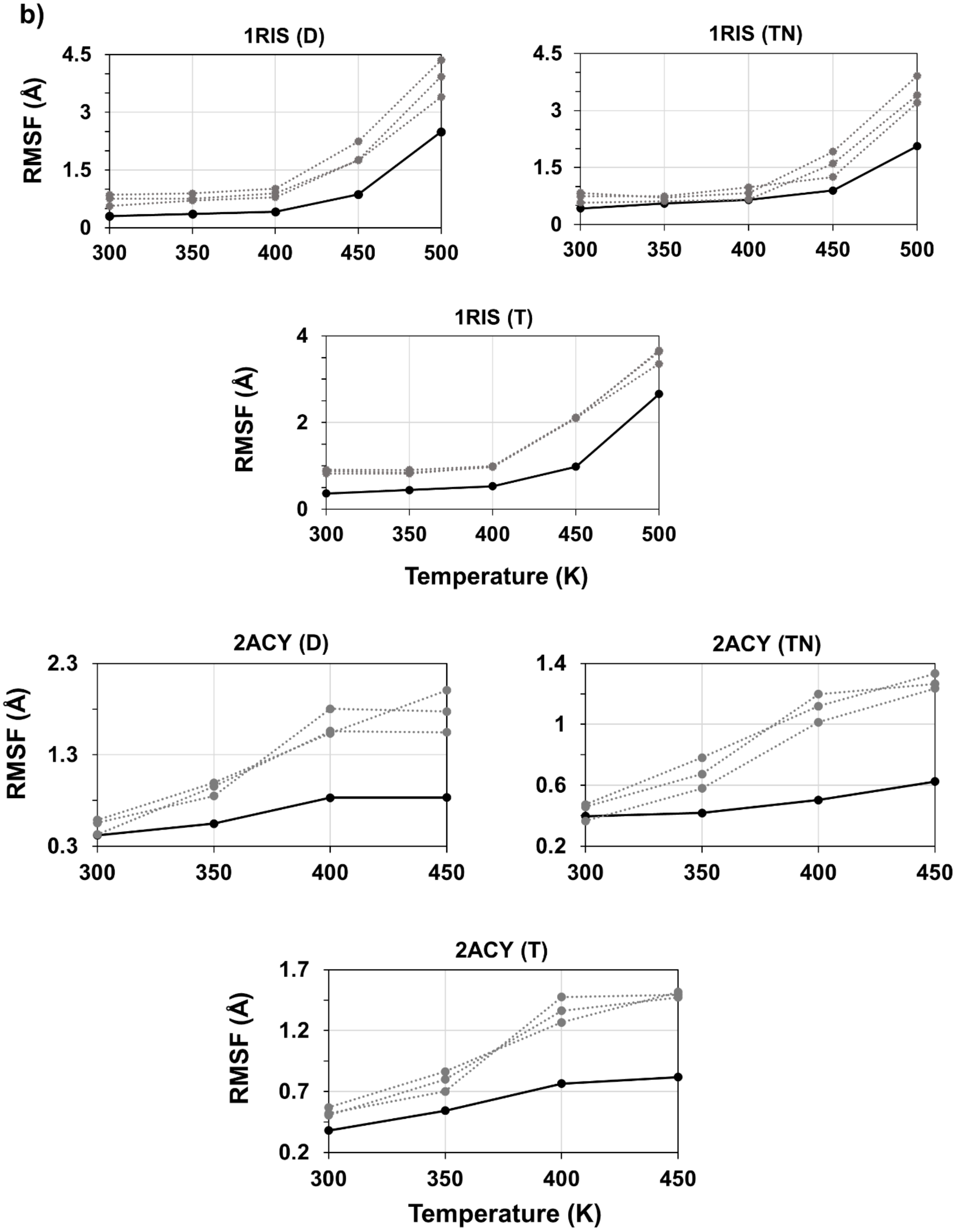


Figure 14 Continued

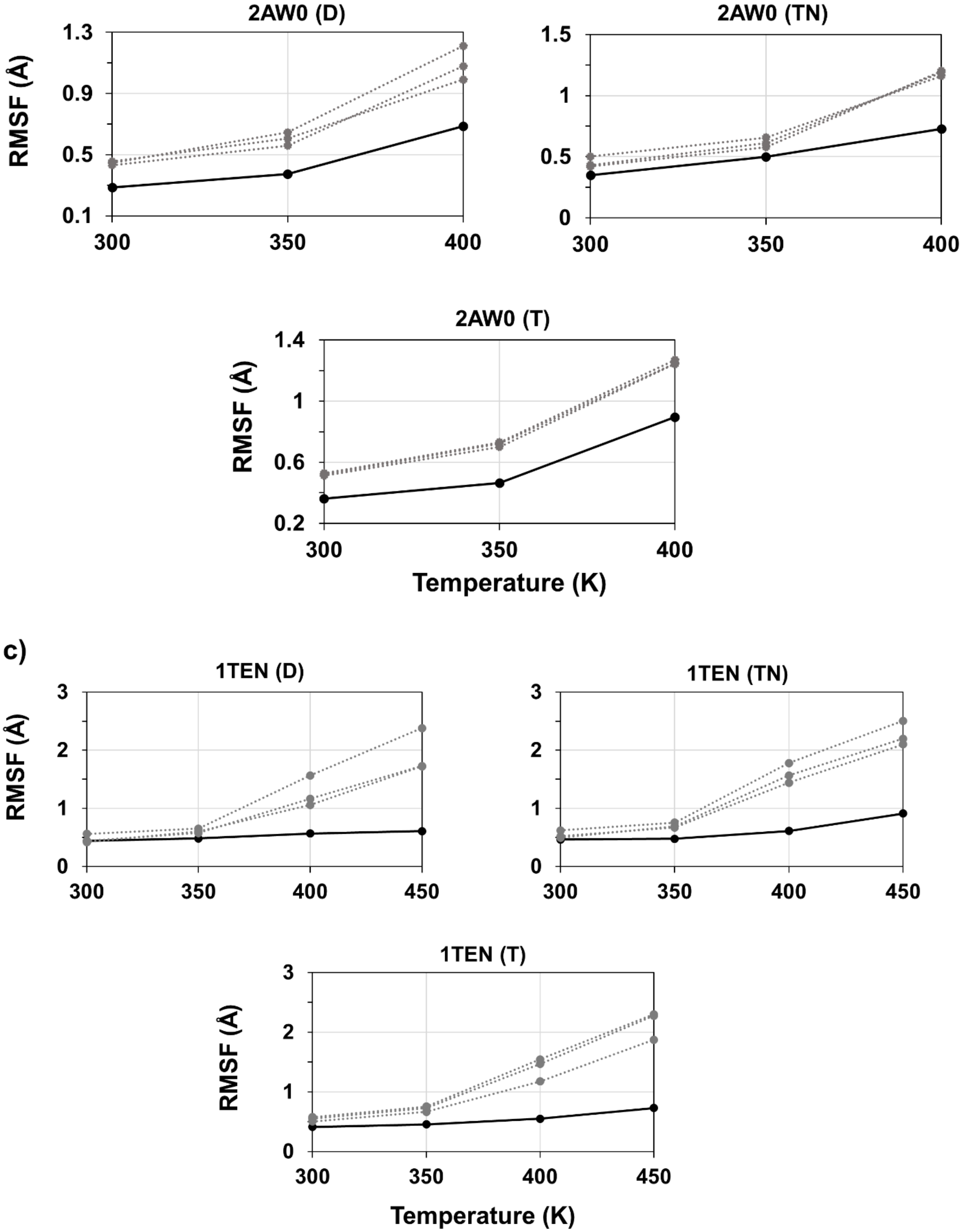
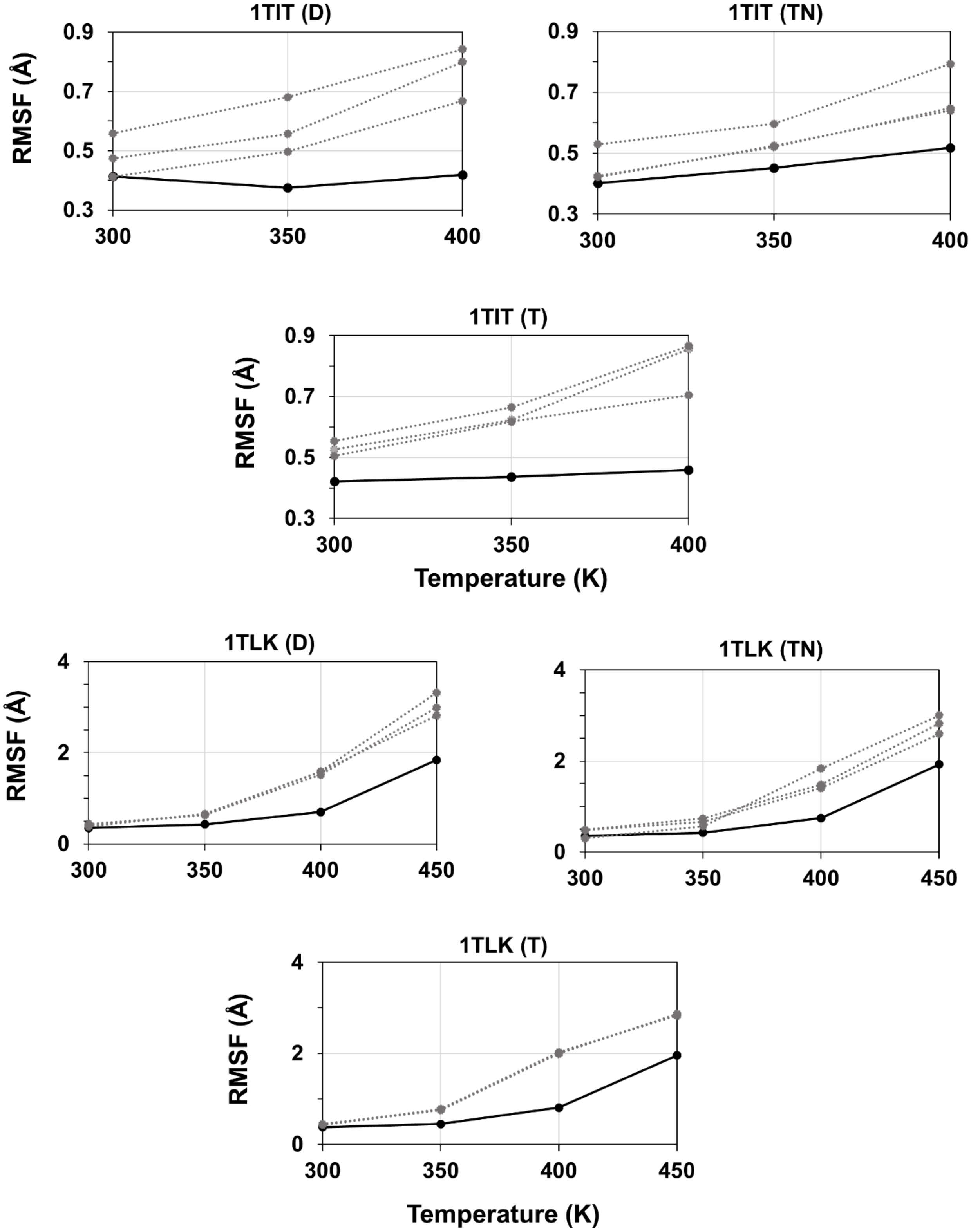


Figure 14 Continued



Calculating the Fraction of the Remaining Contacts as Proteins Unfold

In the third test, the fraction of the contacts remaining was calculated for the conserved, and randomly selected non-conserved contacts during the unfolding process (Fig. 15). As the protein unfolds, the number of the contacts remaining decreases with the simulation time. As shown in Figure 15, the number of the conserved contacts remaining are greater than those involving non-conserved contacts. Furthermore, in several cases the differences are most evident in the early steps of unfolding as in the case of 1RIS, 2AW0 and 1TLK. These results indicate that the conserved contacts break apart on a later timescale than the non-conserved contacts as the protein unfolds.

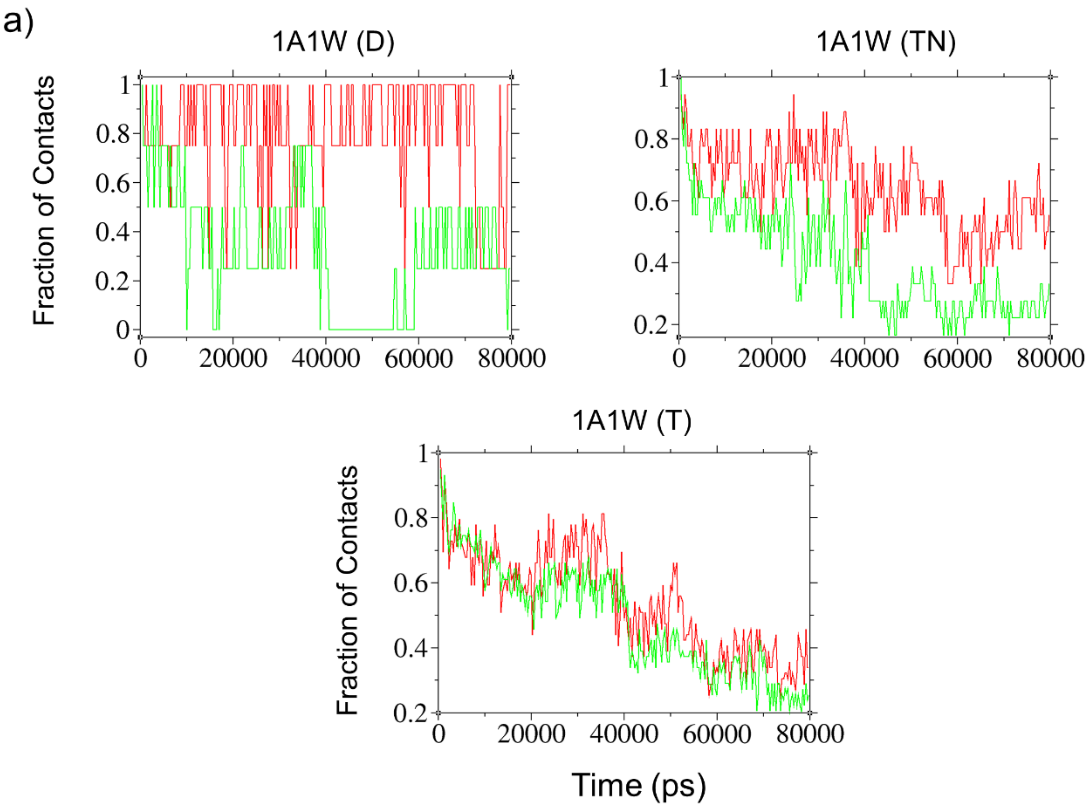


Fig. 15. Fraction of contacts remaining as the protein unfolds. Conserved and randomly selected non-conserved contacts are shown in red and green, respectively. (a) Death domains: 1A1W, 1DGN, 1E3Y (b) α/β-plaits: 1RIS, 2ACY, 2AW0. (c) Immunoglobulins: 1TEN, 1TIT, 1TLK. D, T and TN are abbreviations of the terms Direct, Toggle and Toggle without acidic and basic residues.

Figure 15 Continued

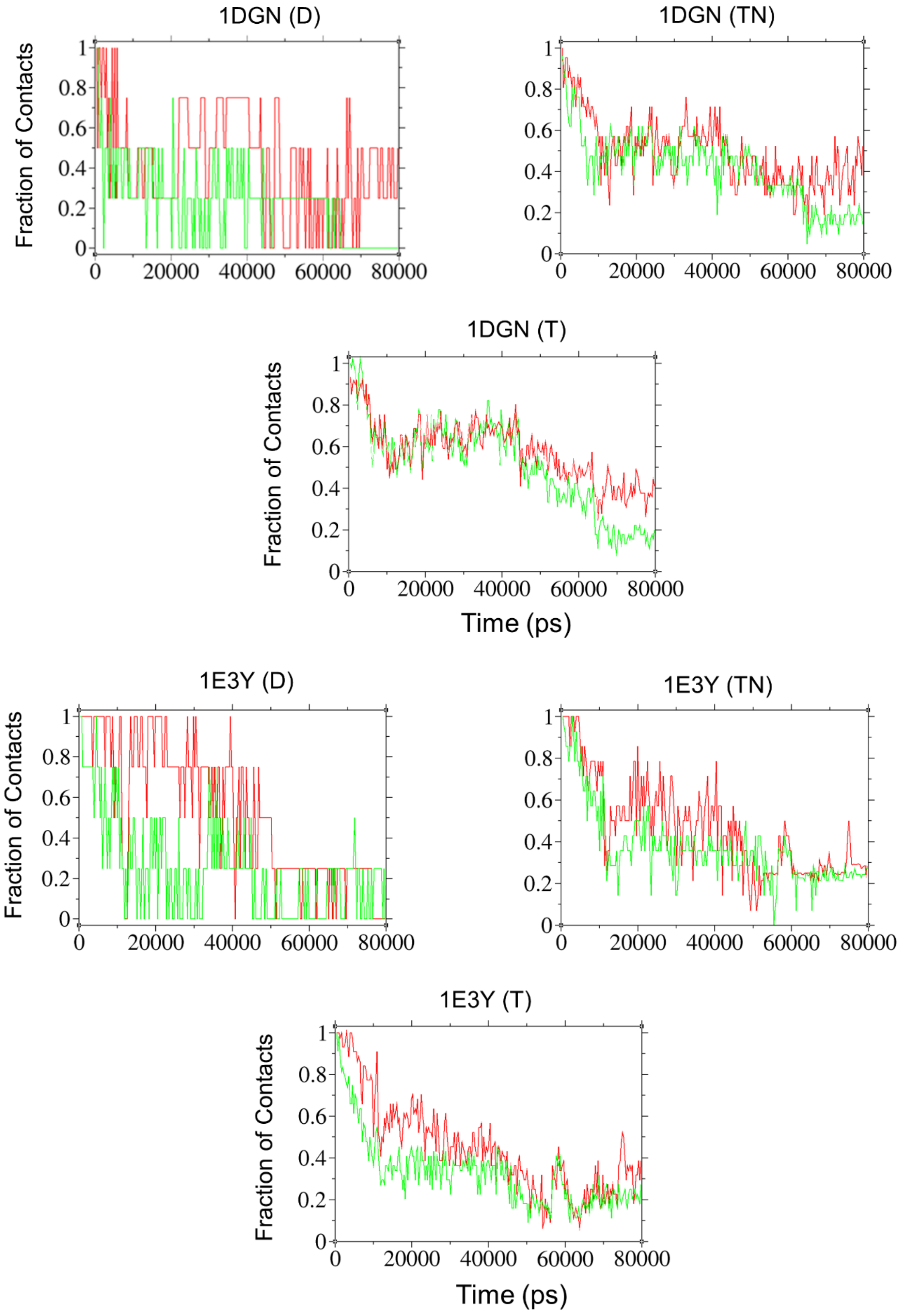


Figure 15 Continued

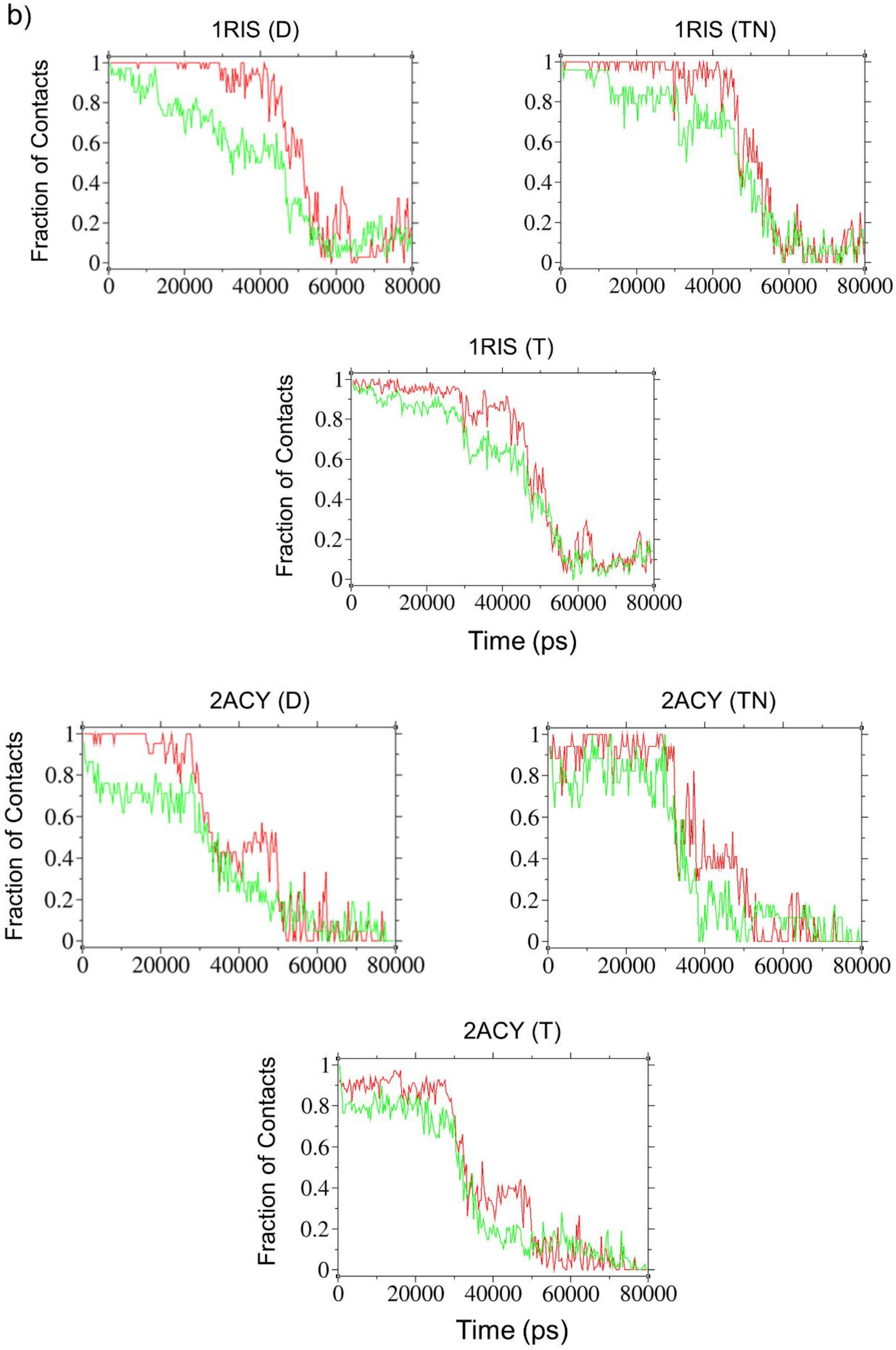


Figure 15 Continued

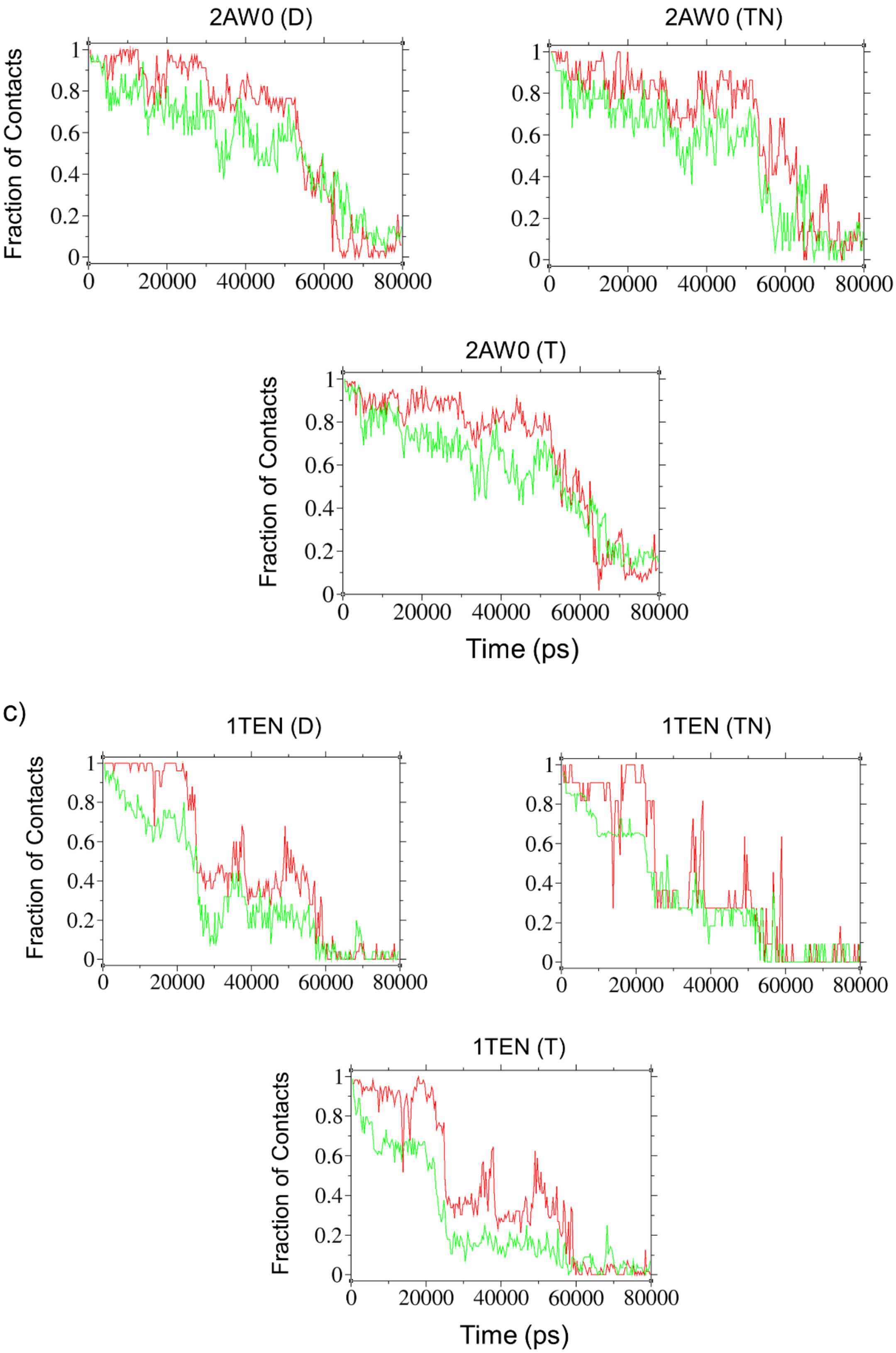
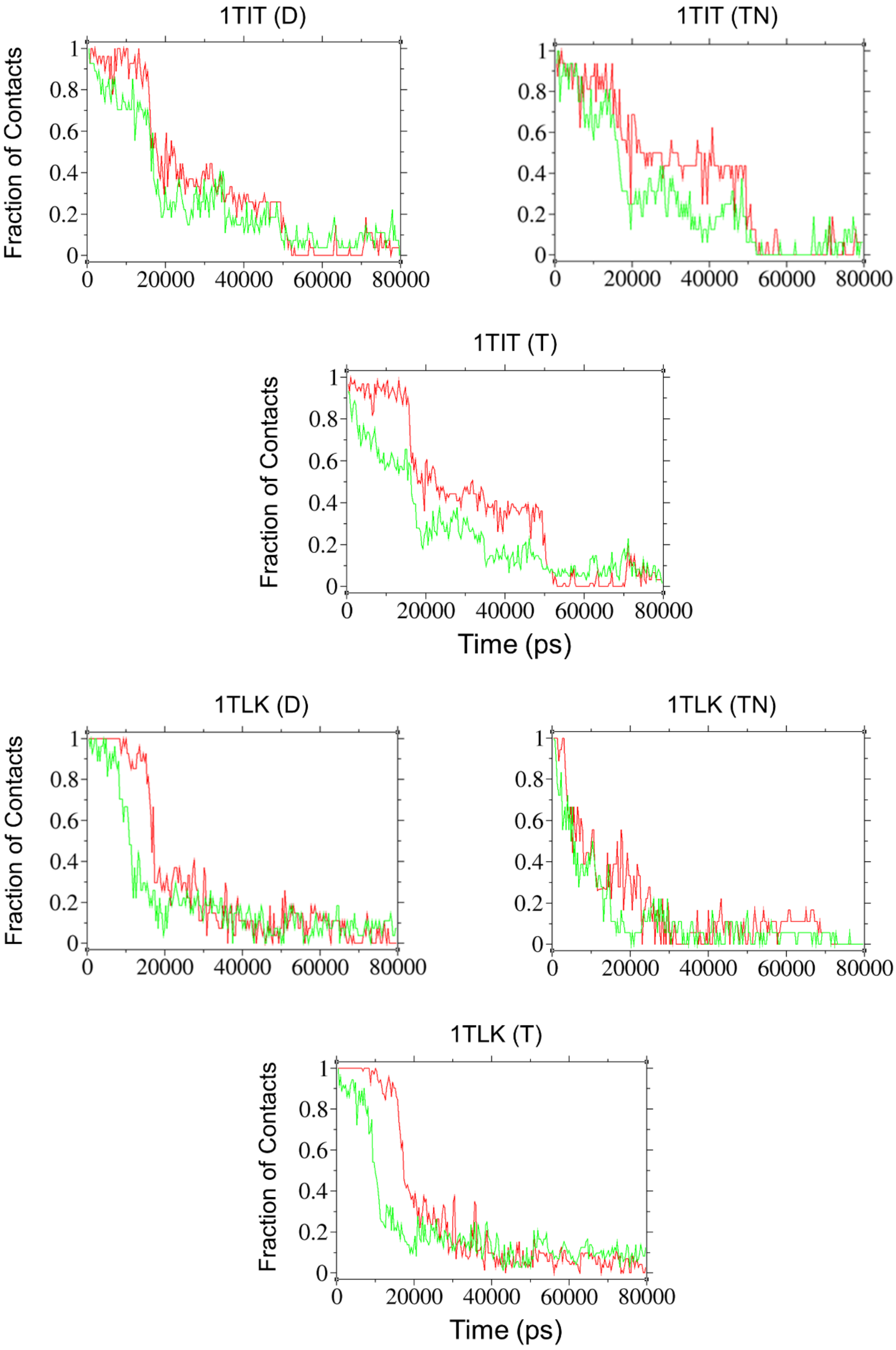


Figure 15 Continued



The Fraction of the conserved contacts remaining is also compared with the entire non-conserved and native contacts as shown in Figure 16.

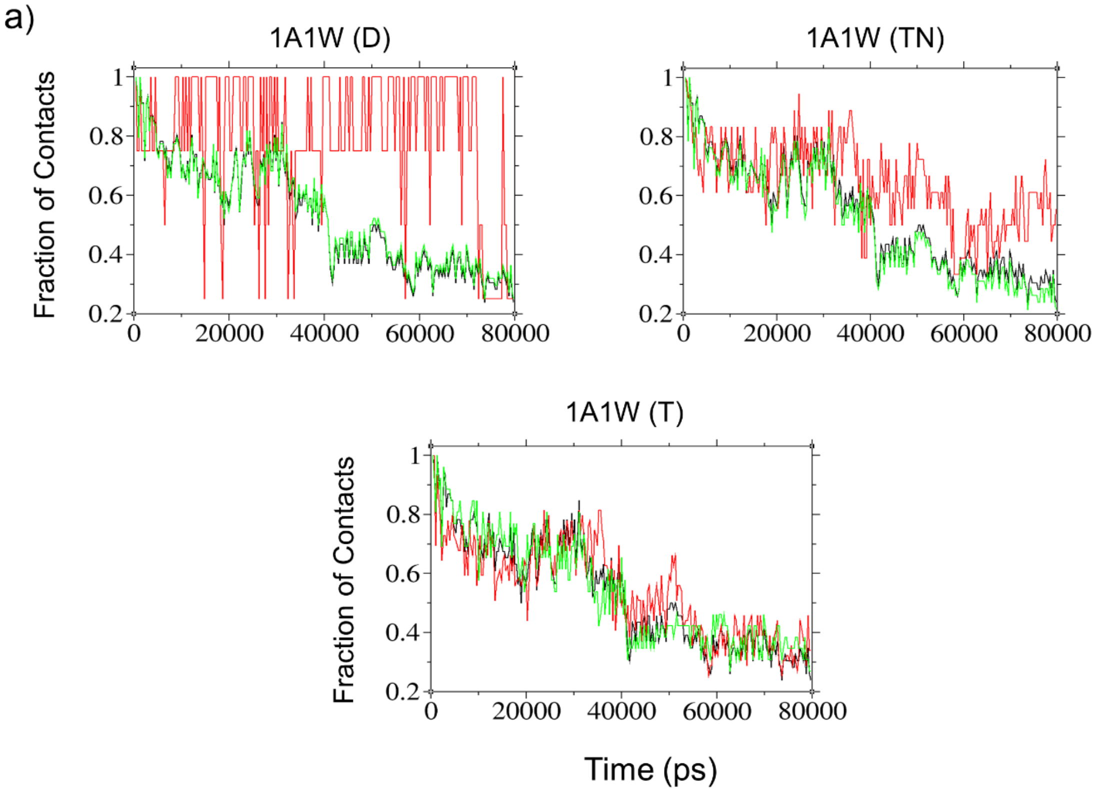


Fig. 16. Fraction of contacts remaining as the protein unfolds. The set of conserved contacts is compared with all non-conserved contacts and native contacts. Conserved, non-conserved and native contacts are shown in red, green and black respectively. (a) Death domains: 1A1W, 1DGN, 1E3Y (b) α/β plaits: 1RIS, 2ACY, 2AW0. (c) Immunoglobulins: 1TEN, 1TIT, 1TLK. D, T and TN are abbreviation of Direct, Toggle and Toggle without acidic and basic residues.

Figure 16 Continued

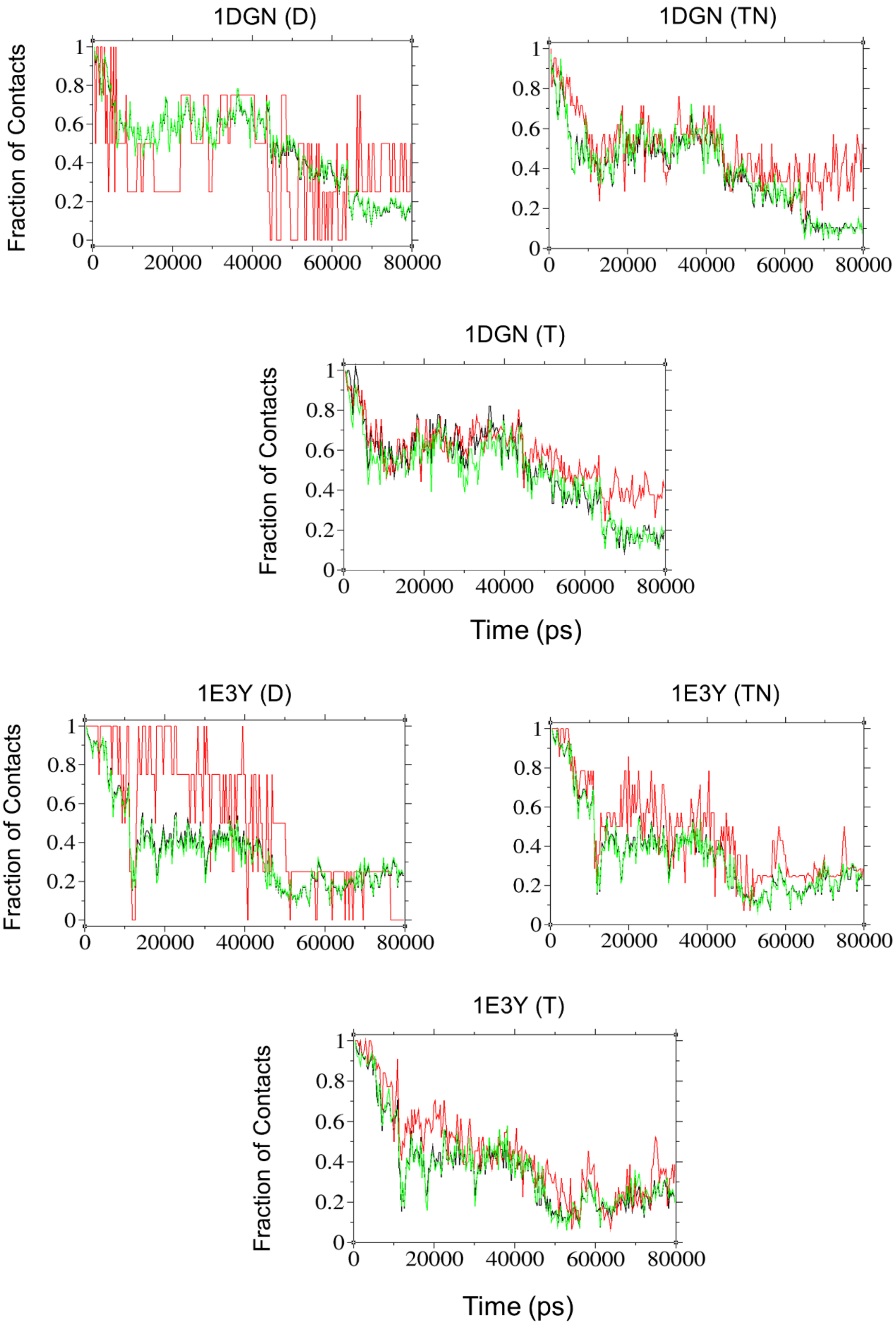


Figure 16 Continued

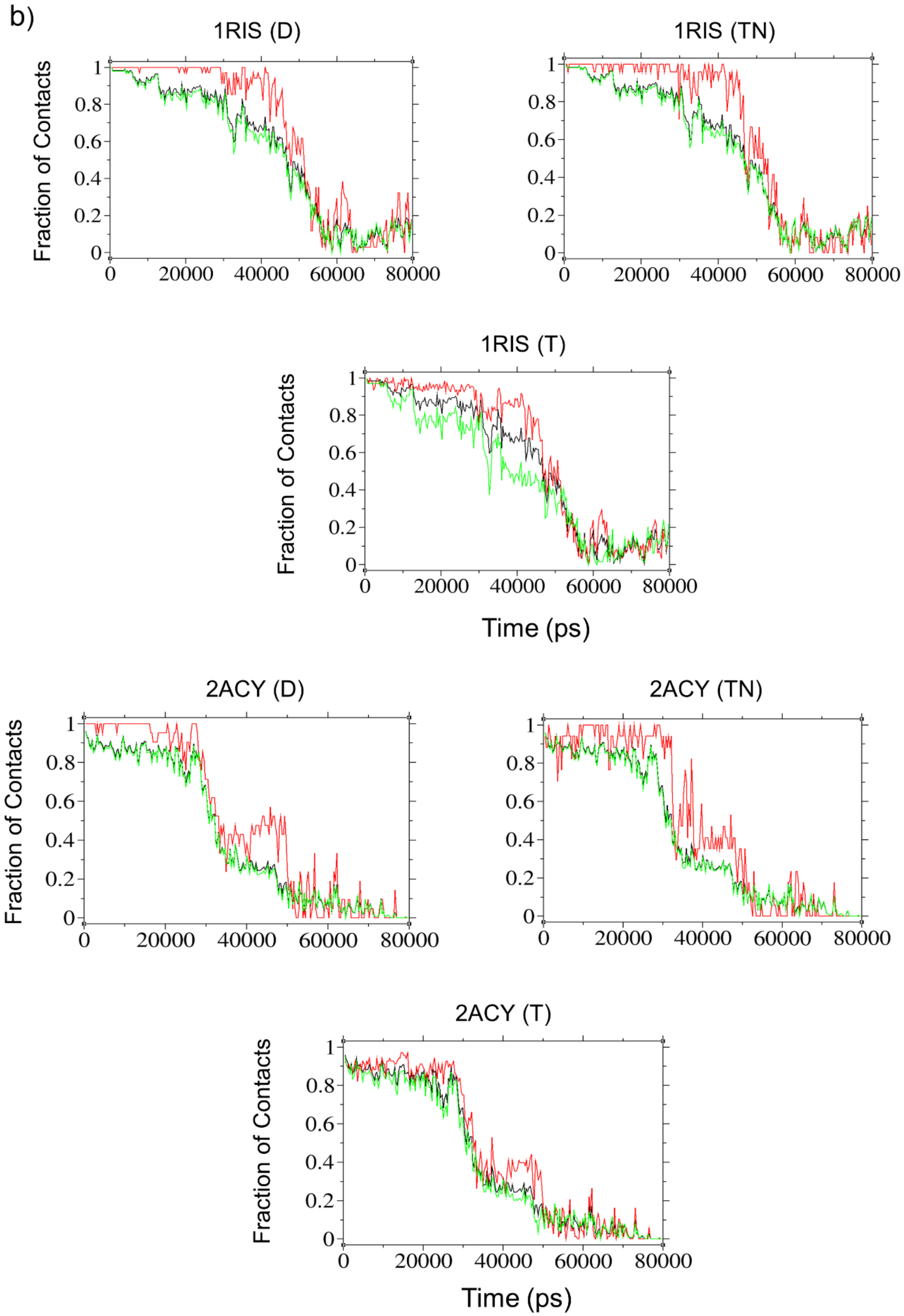


Figure 16 Continued

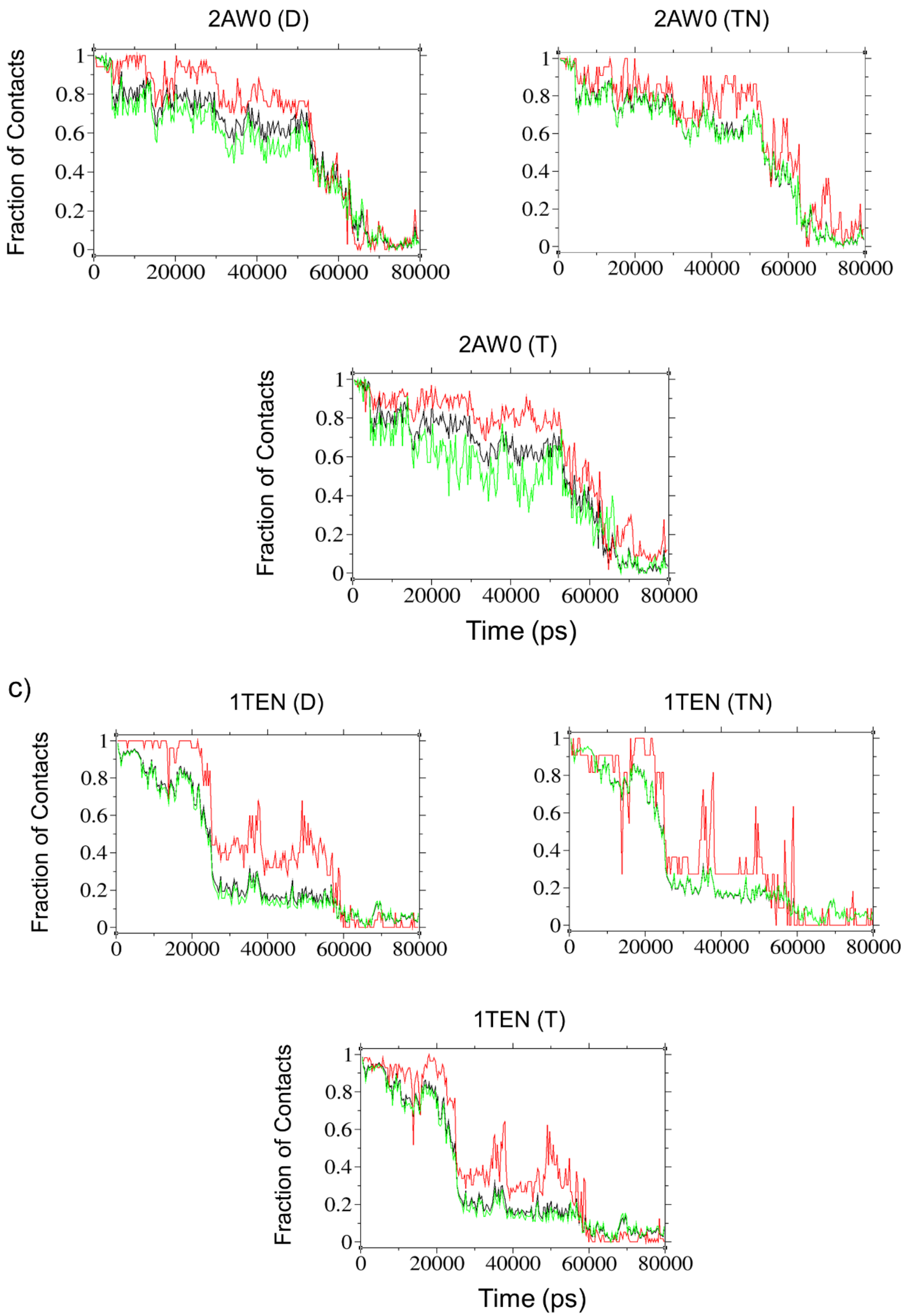
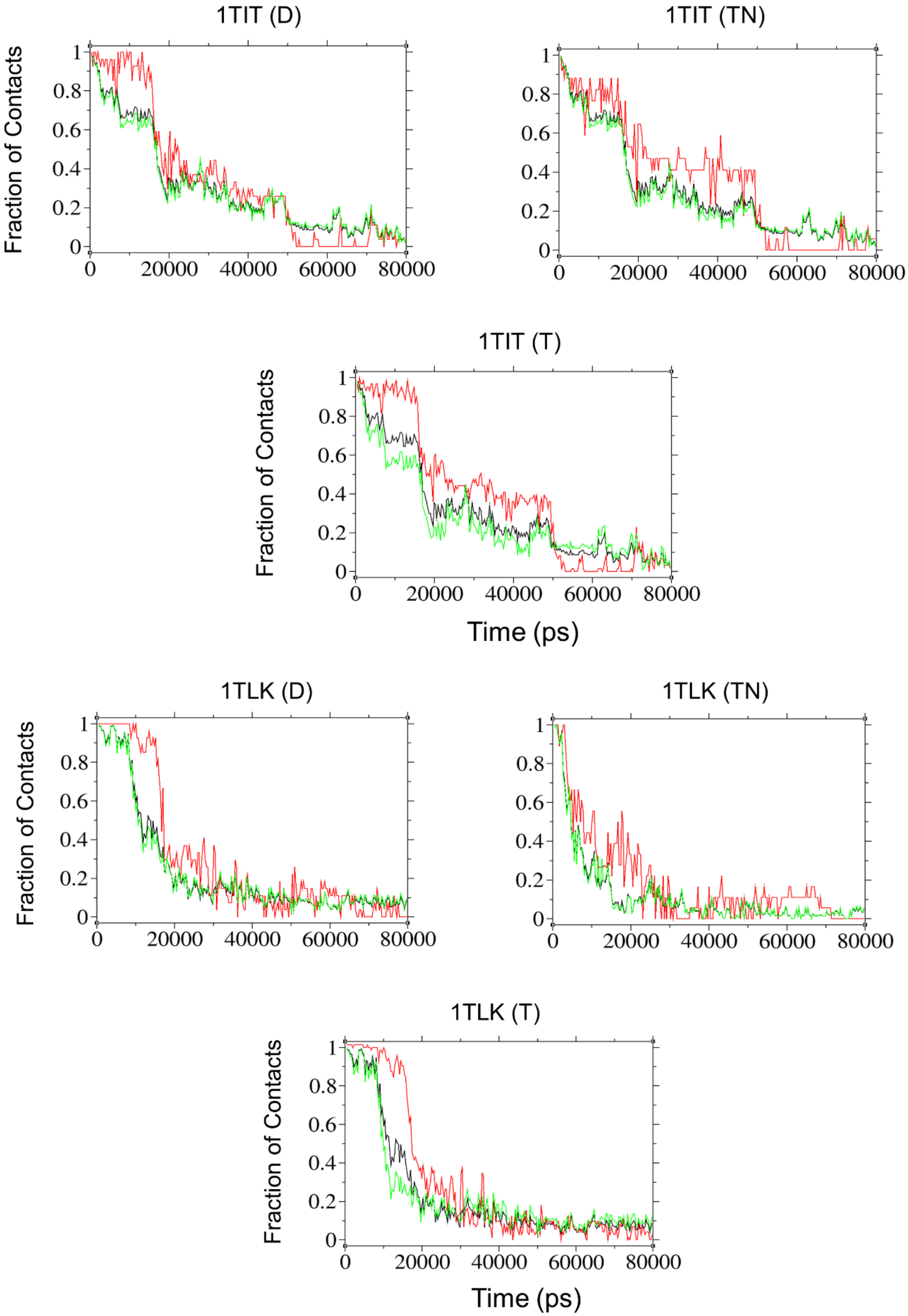


Figure 16 Continued



Correlation to Experimental Studies

Some of our protein models have been also studied experimentally. These studies calculated the Phi-value (ϕ) after mutating the specific residues of a protein. Phi-values are a measure derived from mutational studies that probe the contribution of individual amino acids in the structure of the transition state using kinetic and thermodynamic calculations [145]. Amino acides with significant ϕ-values play an important role in proper folding of a protein in the transition state [146]. Table 3, shows the residues with significant ϕ-values and bolds those that are part of the conserved contact network. Many residues with high ϕ-values that are identified as significant amino acids in protein folding and stability, demonstrate the same result in our computational study by being part of the evolutionary conserved contact network. Moderate to high Phi-values are denoted based on the following criteria: ϕ-values between 0.3-1.0 at ϕ ^{0M} and $\phi^{midpoint}$ for 1RIS [147]. ϕ-values between 0.3-1.0 at ϕ^{1M} for both 1TIT [148] and 1E3Y [149]. ϕ-values between 0.3-1.0 for 1TEN [150] and 2ACY [151]. Residues 160 of 1E3Y, 67 of 1RIS, 42, 45, and 94 of 2ACY, 851, 865, 867 of 1TEN and 2, 41, 47 of 1TIT have significant ϕ-values but they are not part of the long-range conserved network. These residues may help folding by stabilizing the secondary structures and so they are not part of the long-range network. Table 4 demonstrate all amino acid residues in the conserved networks and those that are studied experimentally. Most residues of conserved network that are studied experimentally indicated the significant ϕ-values. There are some residues in the conserved network that are studied in the lab, but they don’t show the high ϕ-values. These residues may play a more significant role in folding the proteins after the transition state.

Table 3. Correlation between residues that are part of the Toggle conserved contact networks and experimental ϕ-values. Residues in bold indicate amino acids that are common in both studies.

Protein (PDB Code)	Residues with significant ϕ-values (Experimental Studies)
1E3Y	**101, 104, 112, 115, 140, 141, 144**, 160, **161, 162, 165, 173**
1RIS	**4, 6, 8, 9, 26, 30, 60, 65**, 67, **75**
2ACY	**11, 13, 30**, 42, 45, **47, 51, 54, 64,** 94
1TEN	**821, 835, 837, 849**, 851, **858, 860, 863**, 865, 867, 869, **871, 873**
1TIT	2, **19, 21, 23, 25, 30, 32, 36**, 41, 47, 49, **56, 58, 60, 71, 73, 75**

Table 4. Illustration of residues that are part of the Toggle conserved contact networks. Underlined residues indicate amino acids that are studied by experimental method. Residues in bold demonstrated significant ϕ-values.

1E3Y	1RIS	2ACY	1TIT	1TEN
A 99	R 2	I 7	E 17	D 816
A 100	R 3	S 8	T 18	T 818
F 101	**Y 4**	V 9	**A 19**	A 819
N 102	E 5	D 10	H 20	L 820
V 103	**V 6**	**Y 11**	**F 21**	**I 821**
I 104	N 7	E 12	E 22	T 822
C 105	**I 8**	**I 13**	**I 23**	W 823
D 106	**V 9**	F 14	E 24	F 824
N 107	L 10	G 15	**L 25**	K 825
G 109	N 11	K 16	**V 30**	P 826
W 112	P 12	V 17	H 31	D 831
R 113	I 25	Y 25	**G 32**	G 832
L 115	**I 26**	T 26	Q 33	I 833
A 116	R 28	A 28	W 34	E 834
R 140	A 29	E 29	K 35	**L 835**
V 141	**L 30**	**G 30**	**L 36**	T 836
R 142	E 31	K 31	**I 49**	**Y 837**
E 143	Y 33	L 33	K 55	G 838
S 144	A 35	G 34	**H 56**	I 839
L 145	R 36	L 35	I 57	K 840
R 146	V 37	V 36	**L 58**	**I 849**
I 147	E 38	G 37	I 59	E 855
W 148	K 39	W 38	**L 60**	N 856
V 158	V 40	V 39	H 61	Q 857
L 161	E 41	Q 40	N 62	**Y 858**
V 162	E 42	T 46	G 69	S 859
G 163	L 43	**V 47**	E 70	**I 860**
A 164	G 44	Q 48	**V 71**	G 861
L 165	G 58	G 49	S 72	**L 863**
S 167	Y 59	**L 51**	**F 73**	E 868
C 168	**F 60**	**P 54**	Q 74	**Y 869**
M 170	L 61	M 61	**A 75**	E 870
N 171	W 62	**W 64**	A 76	**V 871**
L 172	Y 63	L 65	A 81	S 872
V 173	Q 64	D 76	A 82	**L 873**
L 176	**V 65**	R 77		I 874
V 177	E 66	A 78		S 875
Q 187	**L 75**	S 79		R 876
N 188	E 78			E 887
	L 79			T 888
	R 86			
	R 87			
	V 88			
	M 89			

CHAPTER III

SIGNIFICANCE OF EVOLUTIONARY CONSERVED LONG-RANGE INTERACTIONS IN GOVERNING THE TOPOLOGY AND FOLDING IN PROTEINS

INTRODUCTION

Decades of research have yet to comprehensively elucidate all the underlying mechanisms in protein folding and universally predict these structures. The details governing native-state formation of proteins from their linear amino acid sequence is one the most challenging areas remaining at the forefront of the scientific community. Strides towards understanding the protein folding process and topological determination of proteins have been made through numerous *in vitro* and *in silico* studies [49, 97]; however, convincing evidence for the intricate details of protein folding remains desired, including the elusiveness of proteins sharing a common topology, yet lacking in shared sequence. Thus, factors other than amino acid sequence and secondary structure must be considered for enabling these proteins to form the same topology. The evolutionary conserved long-range interactions within these proteins are considered as the critical determinants in governing their common topology throughout this investigation. Long-range interactions which are non-local, are defined as interactions between residues that are close in 3D space but distant in the primary structure.

The folding of a polypeptide chain into a unique 3D is guided by long- and short-range interactions (also known as contacts) along the chain. Several investigations have contributed to understanding the role of these interactions in the protein's structures [9-13, 15, 17, 19]. Long range interactions not only demonstrated a significant role in the structural stability of proteins, but also indicated an essential role in protein folding [16, 17, 21, 22, 152]. Formation of the long-range interactions in the early stage of folding will govern the protein topology and will provide a productive folding [22-24, 137-140]. Within all types of long-range interactions, the hydrophobic interactions indicated a major role in protein folding, especially when they form in the core of the protein structure [130-136]. The focus of this study was showing the significance of evolutionary conserved long-range interactions especially the hydrophobic interactions in protein folding. The role of the selected interactions in the formation of protein topology and forming the native network was tested using bioinformatics, macromolecular simulations and network science (Fig. 17). Here, conserved networks were elucidated in three superfamilies (the death domains, α/β-plaits and immunoglobulins) using a novel bioinformatics approach (Appendix A). These superfamilies share a common Greek-key topology (Fig. 10) yet differ in secondary structure composition, function and sequence identity. 3D networks were then constructed for three members of each superfamily and their role in forming the gross native-like topology and the formation of the consensus network was tested by two different computational methods: simulated annealing and the formation of a giant cluster, respectively. The results of these studies indicate that the evolutionary conserved contacts are the significant determinants of their shared topology and are critical to the protein folding process.

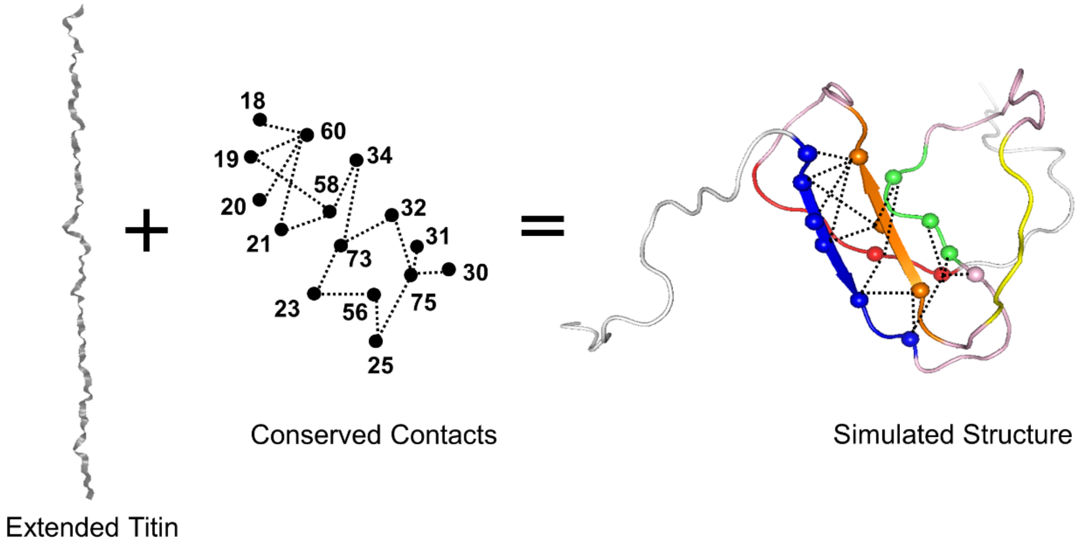


Fig. 17. The schematic representation of titin folding into its native state with the application of the conserved contacts. The simulated structure shows the attainment of the gross native-like Greek-key topology.

MATERIALS AND METHODS

Model System of Proteins

The main methodology for elucidating the conserved networks can be found in Appendix 1. For the SA and Giant Cluster studies, the same three sets of proteins were used in the MD studies in chapter two are further studied in here. They are: Death domains [human death domain of the FAS-associated death domain -1E3Y [87]; human death effector domain - 1A1W [88]; human iceberg - 1DGN [89]], α/β plaits [human fourth metal-binding domain of the Menkes copper-transporting ATPase - 2AW0 [90]; *Thermus thermophilus* ribosomal S6 - 1RIS [91]; bovine acylphosphatase - 2ACY [92]] and Immunoglobulins [titin - human titin - 1TIT [93]; turkey telokin - 1TLK [94]; human tenascin - 1TEN [95]].

Overview of Simulated Annealing Procedure

To determine the role of the conserved contact network in building the protein topology, simulated annealing was performed using CNS 1.3 [64]. This program suite is commonly used for macromolecular structure determination by X-ray crystallography and solution-state nuclear magnetic resonance (NMR) spectroscopy. Simulated annealing is so powerful that it can convert an extended protein structure into a well-defined 3D structure through the application of Nuclear Overhauser Effect (NOE) distance restraints [62]. In this study the evolutionary conserved network of long-range interactions (herein referred to as contacts) (not including acidic and basic residues) was applied as physical restraints to generate the proteins 3D fold from a linear primary structure. The percentage of the number of contacts that were entered as constraints in simulation are shown in Table 5. The same process was repeated for three different sets of equivalent randomly selected non-conserved contacts (not including acidic and basic residues) as a control. Figure 18 depicts a schematic representation of the simulated annealing process using CNS for a set of conserved contacts and an equivalent set of non-conserved contacts. This schematic demonstrates the more important role of the conserved contacts in folding and formation of the protein topology compared to the non-conserved contacts.

Table 5. The % of contacts applied as physical restraints in the SA simulations are shown as the percentage of the total number of the contacts. These percentages are the same for both the conserved and non-conserved sets in each protein.

Protein **(PDB code)**	**% Contacts**
1A1W	5.2%
1DGN	5.8%
1E3Y	3.1%
1RIS	6.5%
2ACY	5.2%
2AW0	5.2%
1TEN	6.5%
1TIT	4.8%
1TLK	5.4%

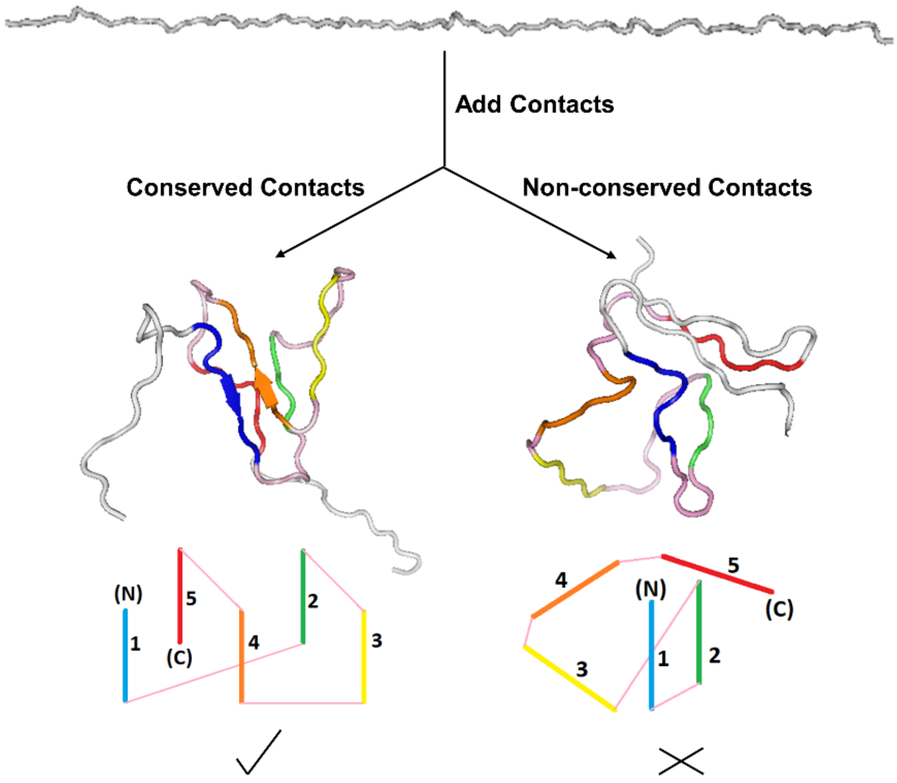


Fig. 18 . The schematic of folding of titin (PDB code: 1tit) using CNS. The structures show that conserved contacts can direct the formation of the Greek-key topology from the linear primary sequence compared to randomly selected non-conserved contacts. This is illustrated by the correct topology obtained on the left with a check mark.

Simulated Annealing Protocol

Our optimized simulated annealing protocol consists of an initial 50-step minimization followed by 10000 high temperature steps at 50000 K. The system was cooled to 250 K over 10000 steps followed by a final 200 step minimization. Each simulation is directed to produce the 10 lowest energy simulated structures. As a control, the simulation was repeated for three proteins, applying different seeds for random number generator to get different initial velocities. We discovered in the course of this research that CNS doesn’t work well in our system when there are too many pairwise constraints between the same residues are involved in the simulation (data not shown). Thus, a minimalist set of atoms involving Cα and Cβ atoms were used to reduce complexity.

RMSD and TM-Score Evaluation

To quantify the topological similarities between the simulated structures and the native form, the RMSD and the TM-score were calculated for each superimposed structure, after aligning the simulated structures with the native one. RMSD is a fast and easily calculated metric of protein structural similarity [153]. Since the RMSD measures the distances between all amino acids pairs equally, a small number of local structural change could result in a high RMSD, even when the overall topologies of the compared structures are similar. Additionally, RMSD is not only determined by the overall goodness to fit but also depending on the proteins length [154]. The TM-score overcomes these limitations by using a variant of the Levitt–Gerstein (LG) metric [155] providing a length independent measurement and limits the impact of divergent pairs of atoms in superimposed structures.

$$RMSD = \sqrt{\frac{1}{L} \sum_{i=1}^{L} d_i^z} \qquad (16)$$

$$LG = \frac{1}{L} \sum_{i=1}^{L} \frac{1}{1+\left(\frac{d_i}{d_0}\right)^2} \qquad (17)$$

$$TM\text{-}Score = Max \left[\frac{1}{L} \sum_{i=1}^{L} \frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right] \qquad (18)$$

In the above formulas, L is the proteins length, d_i is the distance between the ith matched Cα atom and d_0 is a scaling factor to normalize the matches. For small proteins the optimal value of d_0 is 4.5 Å. The LG-based metric gives a value between 0 and 1 where 1 is an exact match. The maximum value of LG that can be obtained by superposition is the TM-score [153, 156].

Calculating the Formation of a Giant Cluster

To further test the role of conserved interactions in topology and folding, network science approaches were utilized. Computational network studies offer important and novel avenue for analyzing complex protein structures [157]. Each protein can be constructed as a network where the amino acids are nodes and long-range contacts are edges. The long-range interaction network is found for each model protein using Contact [158] and the DegLr program [65] written in the Greene laboratory at Old Dominion University [65].

In the formation of a giant cluster as a network is forming, those links which connect the greatest number of nodes to produce the largest cluster are more significant to the network structure. In this study, we develop and apply a Giant Cluster Method test. It is performed by manually adding one contact at a time to generate a non-fragmented cluster. The Pajek program is utilized to provide a visual of the networks [159]. The conserved contacts compete with the non-conserved contacts to make a giant cluster. In this study all conserved contacts including those involving acidic and basic residues (R,E,D,K) are considered. For each protein, two equivalent sets of conserved and non-conserved contacts were selected. These sets were tested separately by randomly adding one contact at a time to see which could generate a giant non-fragmented network of the protein by using a fewer number of the contacts (Fig.19).

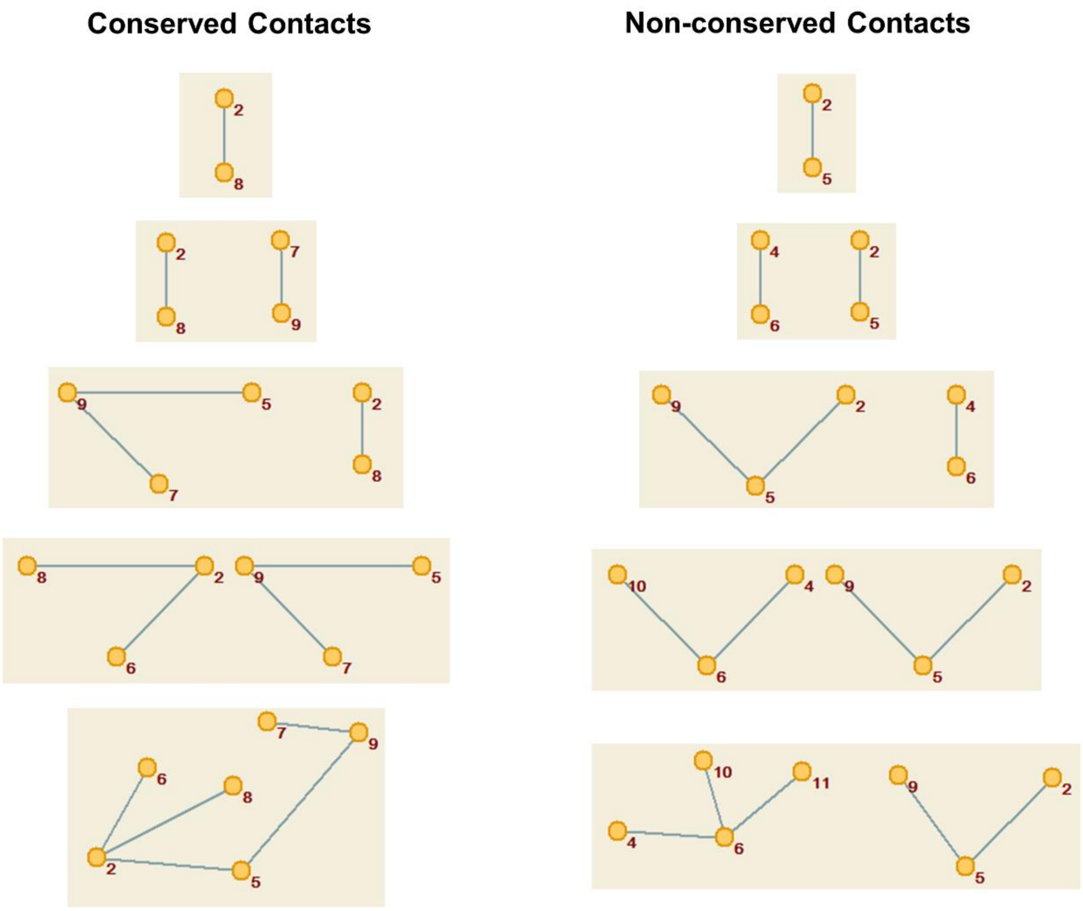


Fig. 19. The schematic representation of Giant Cluster Method. Two equivalent sets of conserved and non-conserved contacts were selected. These sets were tested separately by randomly adding one contact at the time to see which one can make a giant non-fragmented network of the protein by using a fewer number of the contacts.

RESULTS AND DISCUSSION

Simulated Annealing

The Crystallography & NMR System (CNS) [63], was employed to generate the 3D protein structures from the reduced atom set by applying the conserved contacts (not including acidic and basic residues) (Fig. 20). This process is also repeated for three equivalent sets of randomly selected non-conserved contacts as a control. Figure 21 illustrates an example of simulated structures generated with the application of non-conserved contacts. Table 6 organizes the results comprehensively, indicating that 50–100% of the simulated structures fold into their correct gross native-like topology when the conserved contacts are applied. Replication of this process with three different sets of randomly selected non-conserved contacts, indicated that most structures would at best generate 10% of the gross native-like topology, and that only in one instance did 20% of the structures achieve a gross native-like topology. These results indicate the importance of the conserved contacts in guiding the formation of the topology and we extrapolate this finding to suggest that they are key to the generating the native structure during the early stages of the protein folding process. As a control, the simulation was repeated for three proteins, applying different seeds for random number generator to get different initial velocities (Fig. 22).

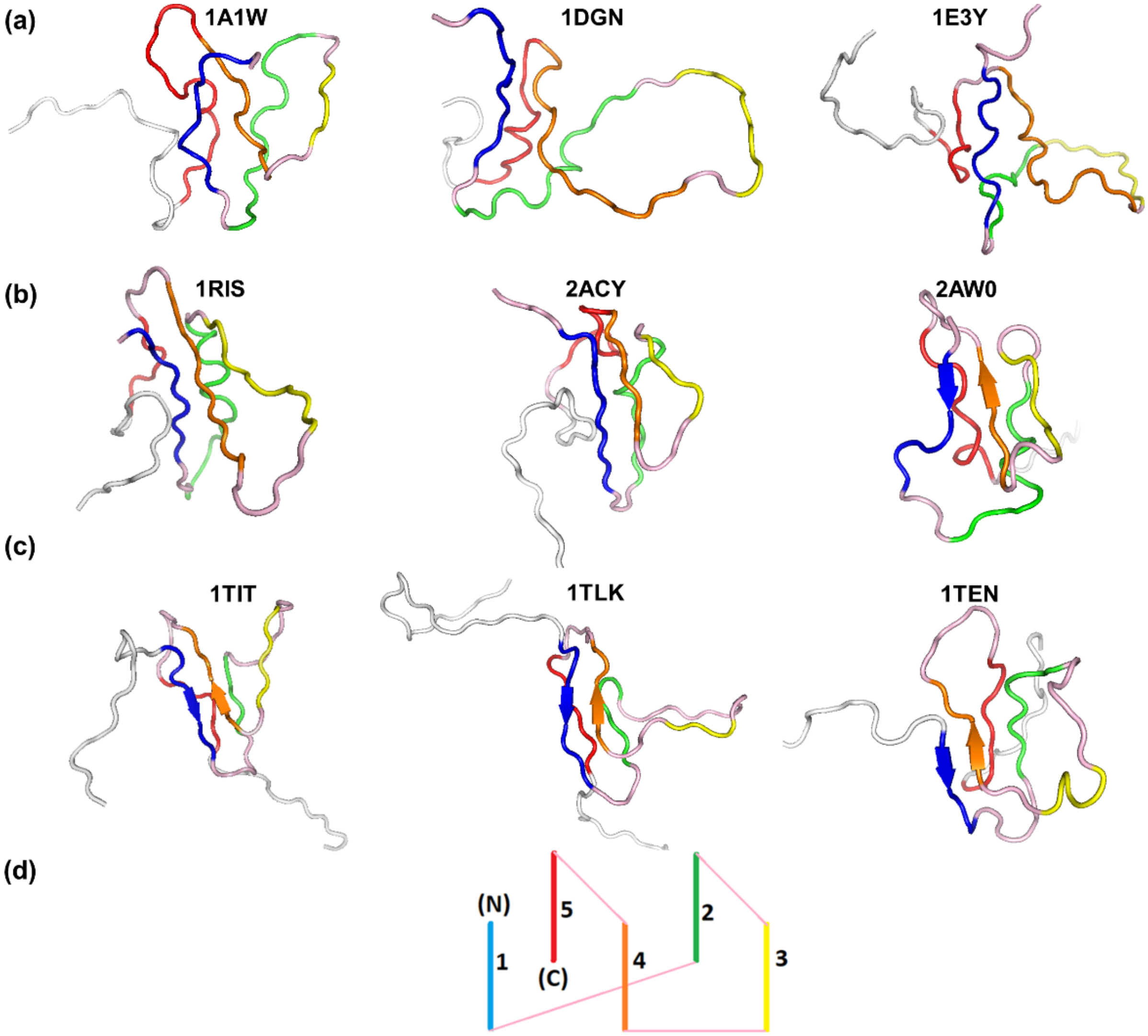


Fig. 20. Illustrations of nine simulated structures according to their superfamilies: (a) Death-domains, (b) α/β plaits, (c) Immunoglobulins (d) Greek-key topology schematic with the secondary elements color-coded. The locations of the N- and C-termini are specified These structures were generated with CNS by applying the conserved contacts as the physical constraints. These structures are color-coded according to their synonymous regions in the shared Greek-key topology. The core of the fold is comprised of elements 1, 2, 4 and 5.

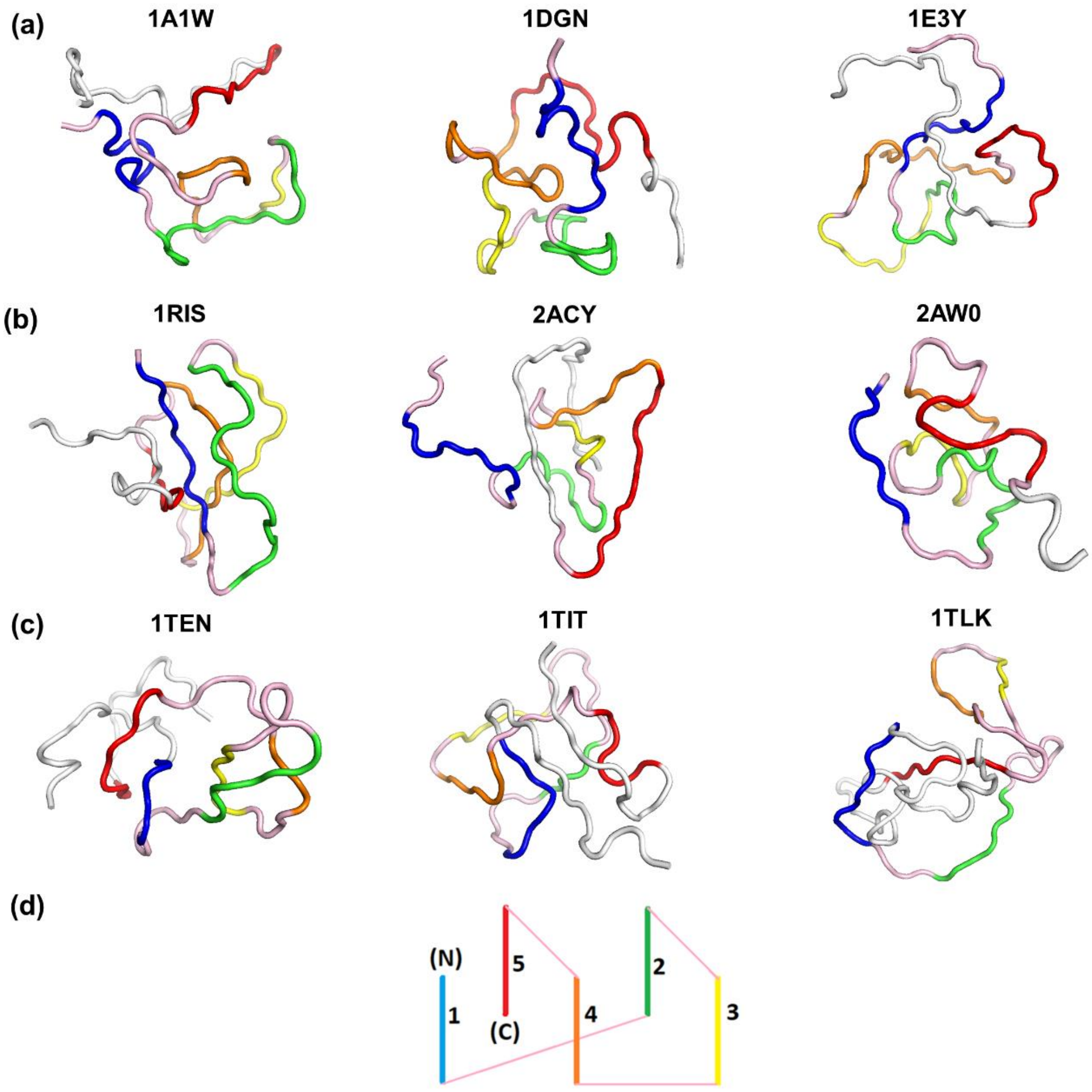


Fig. 21. Illustrations of nine selected misfolded structures grouped according to their superfamilies: (a) Death-domains. (b) α/β-plaits. (c) Immunoglobulins. These structures were generated with SA procedures in CNS by applying the non-conserved interactions as the physical constraints (described in methods). These structures are color-coded according to their synonymous regions of their shared Greek-key topology. The color-coded schematic of Greek-key topology is shown at the bottom.

Table 6. The percentage of the simulated structures that have the gross native-like Greek-key topology. The simulations were conducted with conserved and three groups of randomly selected sets of equivalent numbers of contacts.

Protein (PDB Code)	Conserved Contacts	Non-conserved Contacts (Set 1)	Non-conserved Contacts (Set 2)	Non-conserved Contacts (Set 3)
1A1W	80%	10%	10%	0%
1DGN	100%	10%	10%	0%
1E3Y	60%	10%	0%	0%
1RIS	70%	10%	10%	20%
2ACY	70%	10%	10%	0%
2AW0	70%	10%	0%	10%
1TEN	50%	0%	0%	10%
1TIT	60%	10%	0%	10%
1TLK	60%	0%	0%	0%

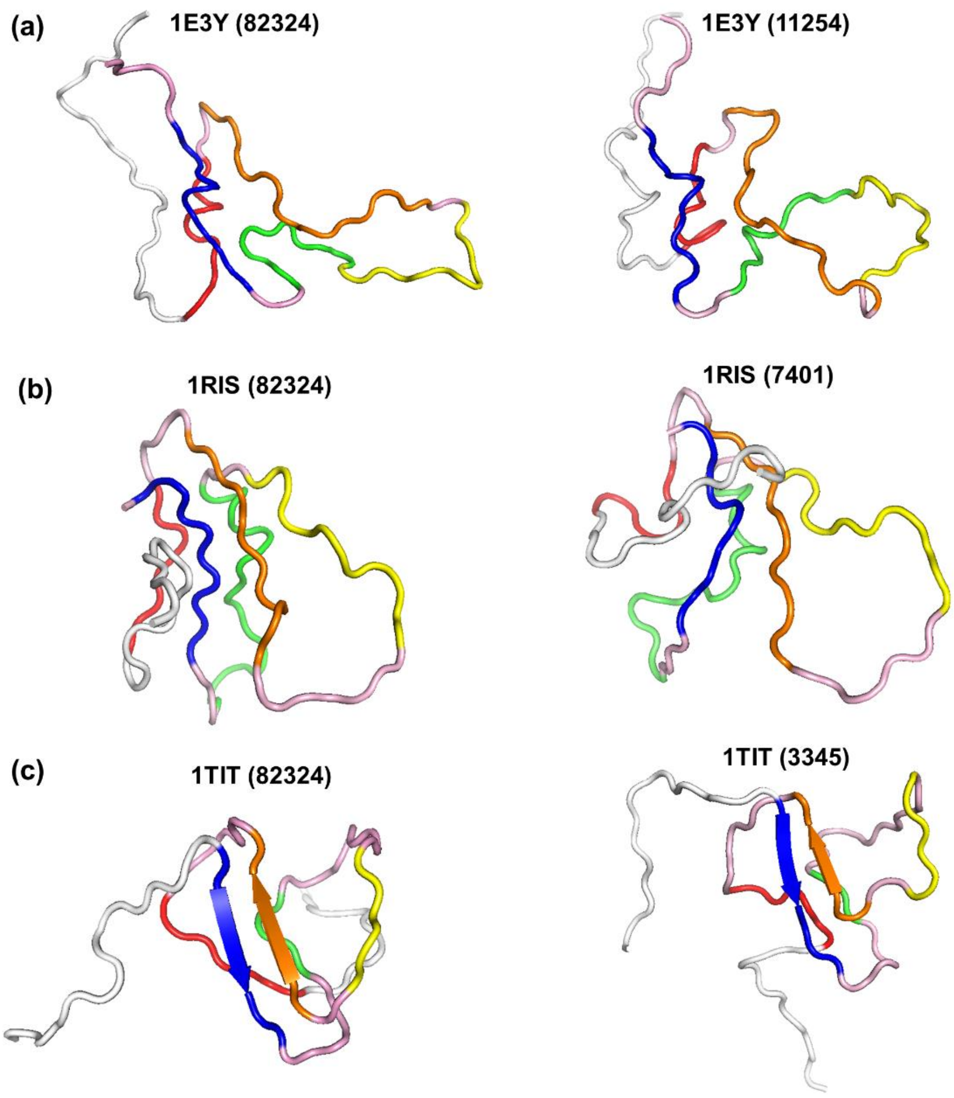


Fig. 22. Applying different seed for random number generator to get different initial velocities. Seed numbers are shown in parenthesis for each protein. 82324 is the default seed number in CNS. These simulated structures are generated using CNS when the conserved interactions are utilized as constraints. One protein is shown as an example for each superfamily, (a) Death domains. (b) α/β-plaits. (c) Immunoglobulins. As shown here, the simulated structures that are made by different seed numbers are similar in each protein and they demonstrate the Greek-key topology.

RMSD and TM-Score

To quantify the structural comparisons between the simulated structures and the native structure in each protein model, the RMSD and TM-scores were calculated for each simulated structure using the TM-align program [153]. The averaged RMSD of 10 simulated structures in each set of conserved and non-conserved contacts was calculated for each protein. In Figure 23, one RMSD value was calculated for a set of conserved contacts and three RMSD values are shown for three different sets of non-conserved contacts. The limitations of the RMSD calculations, described in methods, required the additional calculation of TM-scores for all simulated structures using both conserved and non-conserved contact sets as shown in Figure 24. The combined results indicated that the RMSD calculations have a smaller difference between conserved and non-conserved sets. While the TM-scores were significantly greater for the simulated structures made by the conserved contacts when compared with structures generated by non-conserved contacts. Thus, the closeness of the simulated structures made by the conserved contacts to the native structure, suggests the significance of theses contacts in building the protein native topology. Interestingly, the average TM-value of the structures generated by conserved contacts is even higher, when those simulated structures without the Greek-key topology were eliminated from the list (Fig. 25).

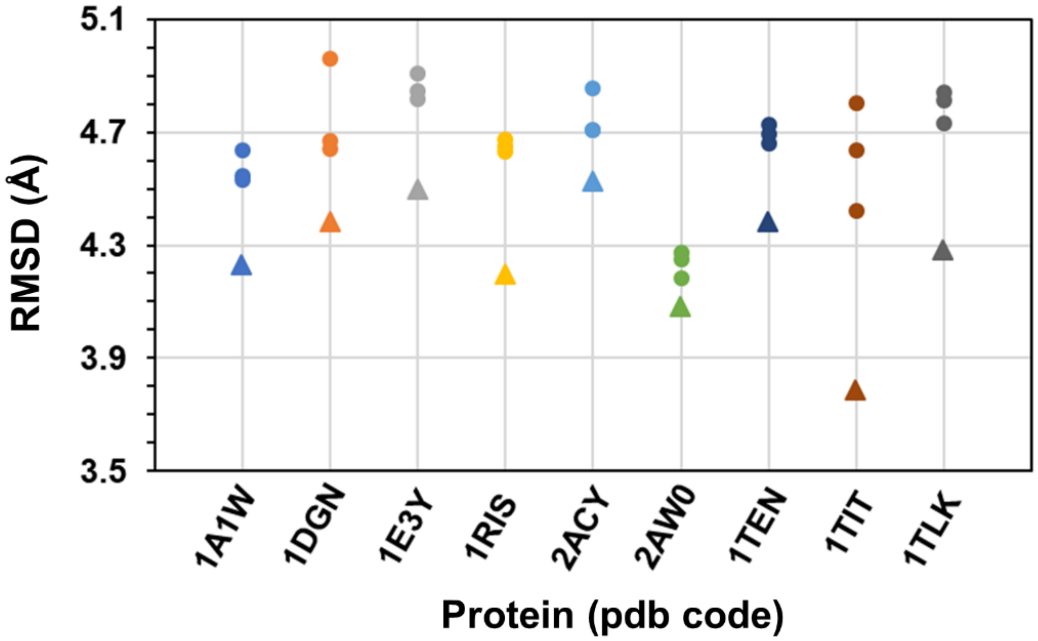


Fig. 23. RMSD of simulated structures. Structures that are made by conserved and non-conserved contacts are shown as ▲ and ●, respectively, including one set of the conserved contacts and three different sets of the randomly selected non-conserved contacts. Each point is representative of the average RMSD for 10 simulated structures in each set of the conserved and non-conserved contacts. The range of the standard deviation for the conserved and non-conserved contacts is 0.15 - 0.57 and 0.22 - 0.47, respectively. The graph was generated with Excel (Office 365).

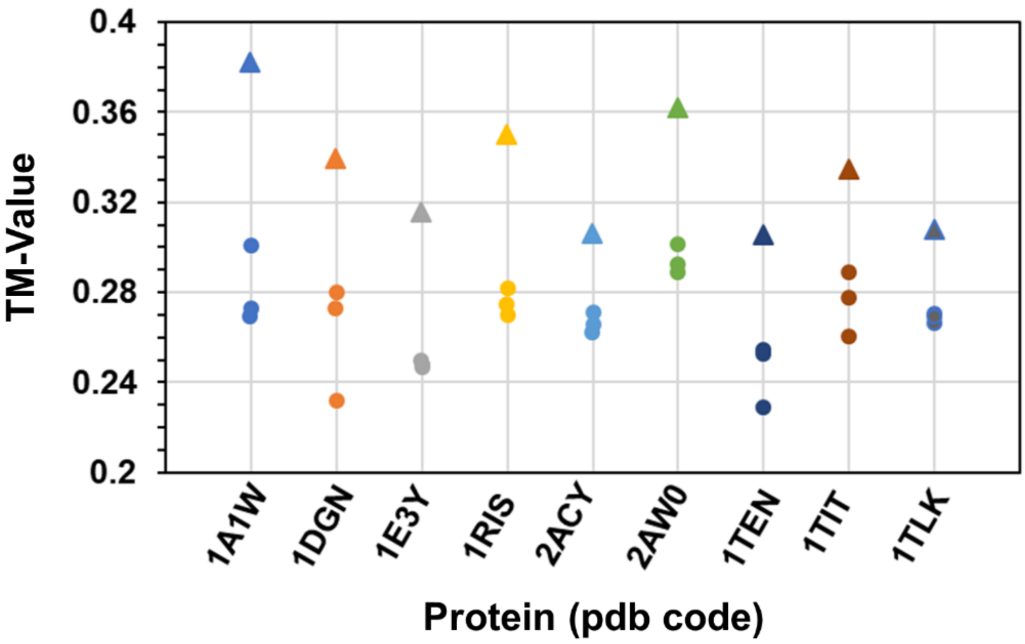


Fig. 24. TM-score of simulated structures. Structures that are made by conserved and non-conserved contacts are shown in ▲and ●, respectively, including one set of the conserved contacts and three different sets of the randomly selected non-conserved contacts. Each point is representative of the averaged TM-score for 10 simulated structures in each set of conserved and non-conserved contacts. The range of the standard deviation for the conserved and non-conserved sets is 0.02 - 0.07 and 0.012- 0.05, respectively. The graph was generated with Excel (Office 365).

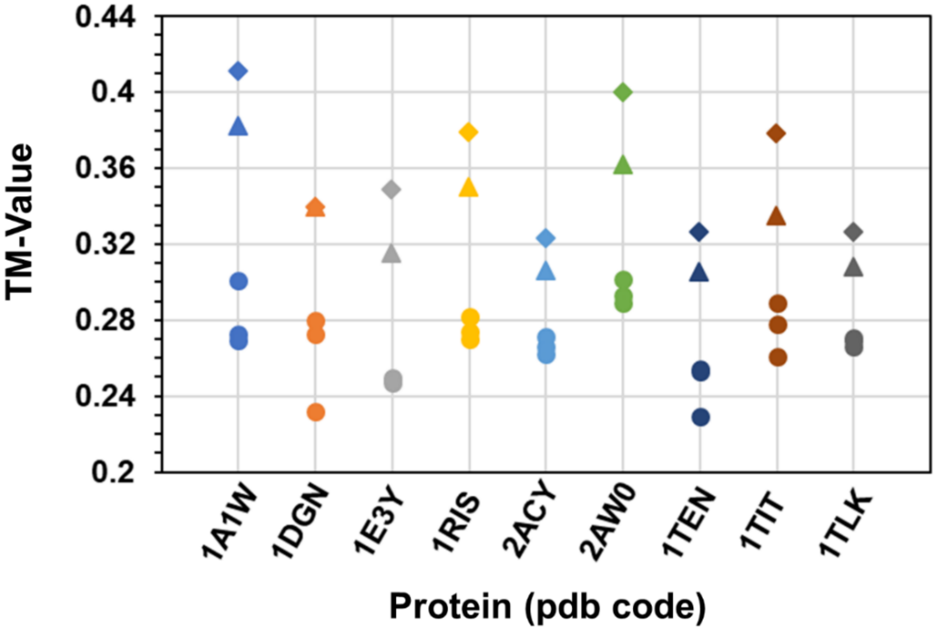


Fig. 25. TM-value of simulated structures. Structures that are made by conserved and non-conserved contacts are shown in ▲ and ●, respectively, including one set of the conserved contacts and three different sets of the randomly selected non-conserved contacts. Each point is representative of the averaged TM-value for 10 simulated structures in each set of conserved and non-conserved contacts. The ♦ represents the average TM-value of only native-like simulated structures generated by conserved contacts. The range of the standard deviation for the conserved and non-conserved contacts is 0.021 - 0.024 and 0.0114- 0.05 respectively.

Giant Cluster

The application of network science enabled us to monitor the formation of a giant cluster which we propose parallels features of protein folding. In this study we generate giant clusters through the directed addition of contacts. Figure 26 graphs the fraction of added contacts against the number of fragments. The results indicate that the conserved contacts make a giant cluster by using a fewer number of atom-to-atom contacts compared to the non-conserved contacts. These results demonstrate the significance of the conserved contacts in building the protein network structure network. Since the conserved contacts make a giant cluster at a faster rate than the non-conserved contacts within the protein, one might deduce that within the Greek-key topology these contacts would ideally be the first ones to form, serving as a scaffold or nucleus to correctly generate the rest of the structure.

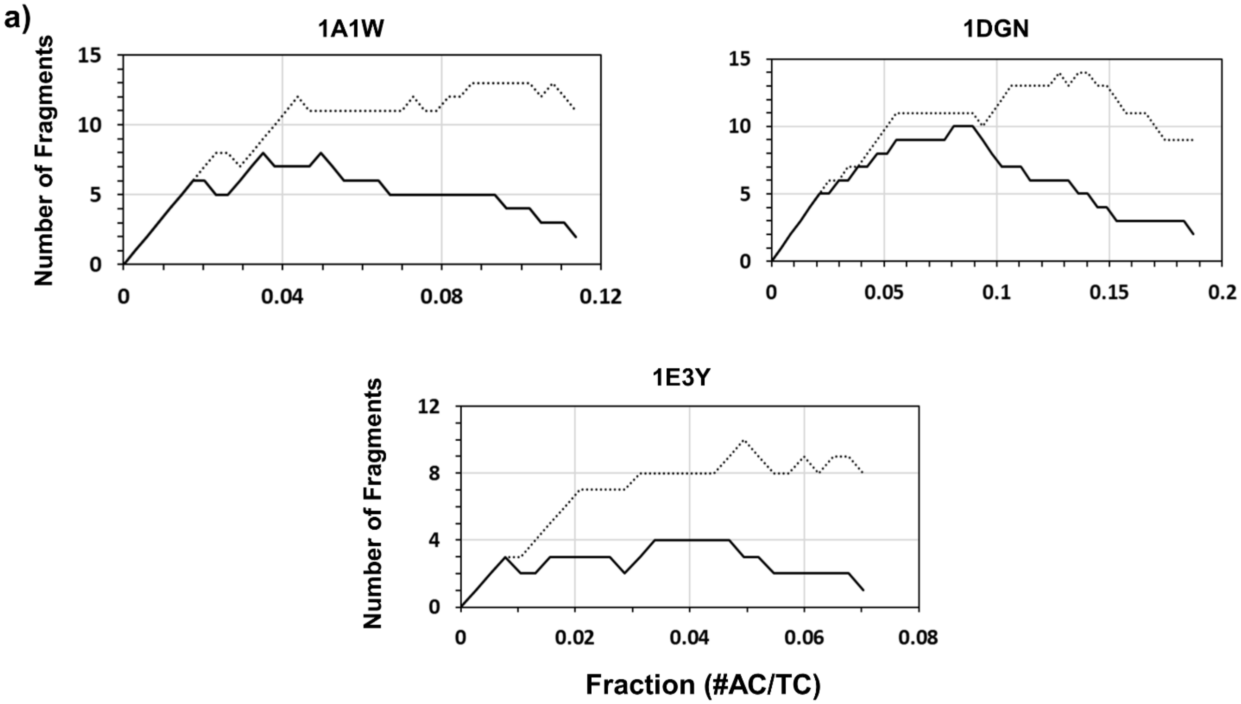
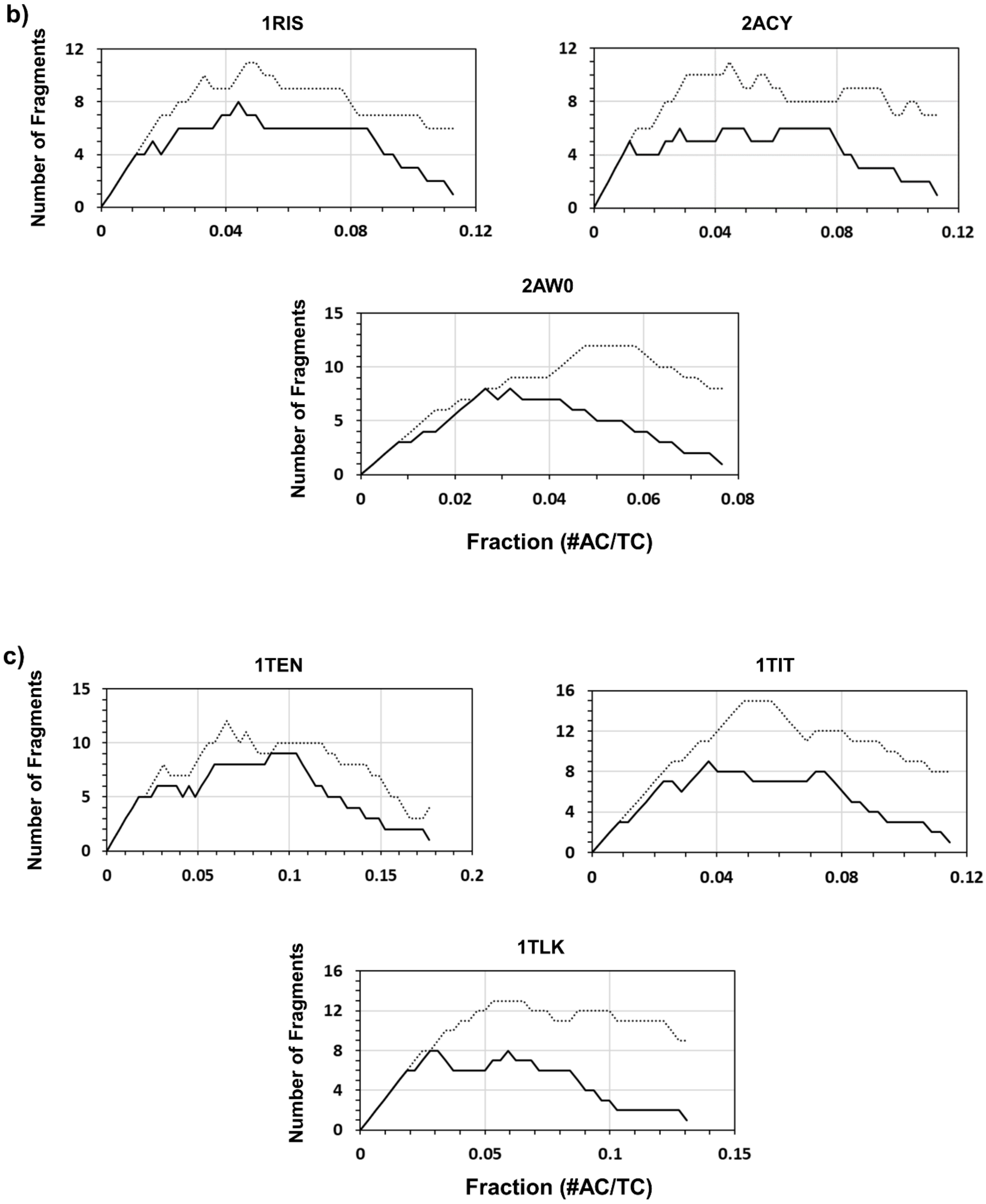


Fig. 26. Formation of a giant network cluster for each protein using long-range contacts. (a) Death domains: 1A1W, 1DGN, 1E3Y (b) α/β plaits: 1RIS, 2ACY, 2AW0. (c) Immunoglobulins: 1TEN, 1TIT, 1TLK. The solid and dotted lines represent the conserved and non-conserved contacts, respectively. AC and TC are abbreviation of added contacts and total contacts, respectively. The graph was generated with Excel (Microsoft Office 365).

Figure 26 Continued



Reported Experimental Studies

There are experimental folding studies involving five of the model proteins that ideally can be used for comparison with the computational SA results. In these studies, mutations were performed on specific residues and Phi-values (ϕ) were calculated from folding and stability studies. Residues that show significant Phi-values are proposed to play an important role in guiding the protein structure throughout the folding process by forming early in the transition-state [146]. Table 7 shows the residues with significant ϕ-values in five proteins. Moderate to high Phi-values are denoted based on the following criteria: ϕ-values between 0.3-1.0 at ϕ^{0M} and $\phi^{midpoint}$ for 1RIS [147] . ϕ-values between 0.3-1.0 at ϕ^{1M} for both 1TIT [148] and 1E3Y [149]. ϕ-values between 0.3-1.0 for 1TEN [150] and 2ACY [151]. The amino acids, in bold, are also part of the conserved contact network from this work. Interestingly, there is reasonably good correspondence between the computational and reported experimental results. 40-80% of the residues with high Phi-values are in the conserved networks.

Table 7. Correlation between residues that are part of the conserved contact network in the SA simulations and residues with medium to high ϕ-values are shown. Residues in bold indicate amino acids that are part of the conserved contact networks.

Protein (PDB Code)	Residues with significant ϕ-values (Experimental Studies)
1E3Y	101, 104, **112, 115,** 140, 141, **144**, 160, **161, 162, 165**, 173
1RIS	**4, 6, 8, 9, 26, 30,** 60, **65**, 67, **75**
2ACY	**11, 13, 30**, 42, 45, 47, **51**, 54, 64, 94
1TEN	**821, 835, 837**, 849, 851, **858, 860**, 863, 865, 867, 869, 871, **873**
1TIT	2, **19, 21, 23, 25, 30, 32**, 36, 41, 47, 49, **56, 58, 60,** 71, **73, 75**

CHAPTER IV

NETWORK CONNECTIVITY, CENTRALITY AND FRAGMENTATION IN THE GREEK-KEY PROTEIN TOPOLOGY

INTRODUCTION

While the amino acid sequence alone can encode the 3D structure of small simple systems, the intricacies reside in how the long-range interactions govern this process. Advances can be gained from comparative studies and particularly by studying groups of proteins that share common structures but are very divergent in sequence. Since some proteins with different sequences, secondary structures and functions share the same topology, it has been proposed that there are a consensus set of determinants encoding each unique topology [96]. This hypothesis highlights the role of the non-covalent long-range interactions in the protein folding process. As observed in other studies, long-range interactions are proposed to play a significant role in formation and stabilization of the overall three-dimensional form of a protein [15-18]. Whereas, local interactions primarily dictate secondary structure [17].

The identification and topological role of a specific subset of the long-range interaction network between selected amino acids can be investigated using two approaches. The first is a comparative analysis of divergent proteins and the second utilizes network principles.

The content in this chapter is reprinted with permissions from "Haratipour Z, Aldabagh H, Li Y, Greene LH. Network Connectivity, Centrality and Fragmentation in the Greek-Key Protein Topology. The Protein Journal. 2019;38:497-505".

Thus, for our present study, nine proteins from three groups that share the same Greek-key topology (Fig. 10) but differ in sequence, secondary structure composition and biological function are selected to serve as our model system. They come from the following three superfamilies: the death domains, α/β-plaits, and immunoglobulins, which are classified as all-α, mixed α/β, and all-β, respectively. The structures of proteins in these superfamilies can be visualized as two layers of secondary elements, packed together via a central four-element motif (Fig. 27). This enables us to significantly enhance diversity in order to conduct an exploration to elucidate potential determinants of topology.

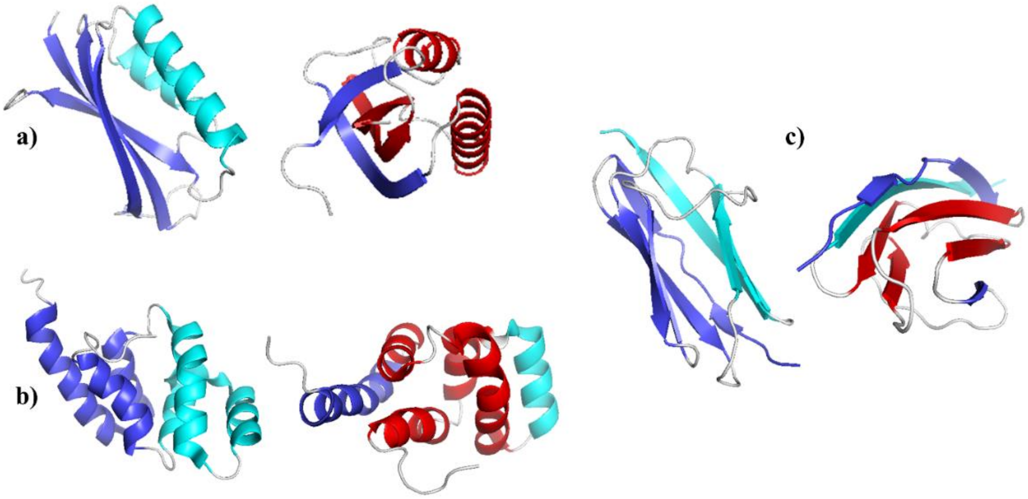


Fig. 27. Similar structural packing of (a) 1RIS, (b) 1E3Y and (c) 1TIT. Proteins are selected from three different superfamilies: α/β-plaits, death domains and immunoglobulins respectively. All proteins consist of two bundles that are shown in blue and cyan. These two units are packed together via a central four-element motif that is shown in red. This central motif consists of two pairs of secondary structural elements, one from each unit arranged in a Greek-key topology. Figure reproduced from [160].

The second involves the application of betweenness centrality and diameter to analyze the nature of the long-range interaction network within each protein. In the context of proteins as network systems, residues with high betweeness centrality (BC) scores are considered to govern the network [161]. We extrapolate this to mean that they may play an important role in the formation and stability of the network. Two additional measures, fragmentation and diameter, are applied to further test the role of BC in our networks. Here, the robustness of network integrity under directed and random attack enables us to monitor the contribution of different residues. Thus, they should have a more or less important role in the network stability.

In this work, we show the importance of these specific geographical regions in the 3D structure of select proteins using network principles. This approach offers a unique and rigorous methodology to interrogate structures from an interdisciplinary perspective. It also advances an earlier, more limited study involving a subset of Greek-key proteins which facilitates a deeper appreciation for this highly populated and very versatile fold [96].

MATERIALS AND METHODS

Protein Networks

Each protein can be constructed as a network where the amino acids are nodes and long-range contacts are edges. The long-range interaction network is found for each model protein using Contact [158] and the DegLr program [65] written in the Greene lab. In this exploratory research investigation, a long-range contact occurs when there exists at least a pair of heavy atoms, each from two residues separated by at least seven other amino acids in sequence, within a certain cutoff distance from each other. The cutoff distance for this research investigation is 7Å for both death domains and α/β-plaits, but only 6Å for immunoglobulins (Ig) since they are potentially composed of more tightly packed β-sheet secondary structures. For completion, we also constructed networks with 5Å cutoffs . Distance cutoffs in other published studies range for example from 4-8Å although these calculations vary in the use of atom types from Cα to all heavy atoms or the use of spheres with a defined radius [15, 16, 65, 67, 162-166].

Betweeness Centrality and Closeness Centrality

To identify the residues that play the most important role in controlling or maintaining the protein network structure, betweenness centrality (BC) is calculated for all selected proteins. BC is a measure of total number of shortest paths between all possible pairs of nodes (*i*,*j*) that pass through node (*m*), *i.e.*,

$$BC = \sum_{i \neq j} \frac{(i,m,j)}{(i,j)} \qquad (19)$$

The ratio of (*i*,*m*,*j*)/(*i*,*j*) demonstrates how significant the role of node m is in connections between *i* and *j* [161]. Nodes with high BC play a crucial role in the network connectivity and centrality and they are proposed to control the network. BC can be applied to diverse systems to include proteins [96, 167]. The protein long-range interaction network is analyzed using the Pajek Large Network Analysis Program, Version 4.08 [168] to calculate the betweenness centrality for all amino acids. The BC data was also analyzed to determine the mean and two standards of deviation from the mean (≥ 2SD) for statistical analysis using Sigma Plot, Version 14.

A second statistical analysis involves the calculation of Z-scores. The Z-score calculates the number of standard deviations below or above the mean for each BC value. The calculation of the Z-score was done using the program Excel (Office 365). The basic Z-score formula for a BC sample is Z = (X – μ) / σ, where X is the BC value, μ is the mean, σ is the standard deviation. Residues with a Z-score ≥ 2 are considered to have high BC. Calculations smaller angstrom cutoffs (5Å) can also be found in for comparison to the 6Å and 7Å cutoffs used in this work.

Closeness centrality (CC) is an interesting measure which may facilitate further understanding the BC values. Here we can distinguish which amino acids are closer to all the other amino acids within the network [169]. Using the same cutoffs as BC (6Å, Igs and 7Å α/β-plaits and death domains), we generated CC for all amino acids using the Pajek Large Network Analysis Program.

Fragmentation

Fragmentation is performed to investigate the role of the residues with high BC value in the robustness of the long-range interaction networks in the select proteins. We designed a program in Python called *fragmentation* which measures the size of the largest cluster (S) shown as a fraction of nodes of the cluster with respect to the total system size, when a fraction (*f*) of the nodes are removed randomly in an attack mode [170]. The program removes nodes from the original network one at the time and calculates *S* and *f* after each removal. We consider two scenarios. In the first, randomly selected residues are removed from the original network (S = 1.0) one at the time until the main network completely collapsed (S = 0.0). In the second, residues with high BC are removed from the original network followed by the random removal of the other nodes until S = 0.0 (Fig. 28). When one node is removed, the new network might be still fully connected, or disconnected into network fragments. If the network is fragmented, random removal of the nodes is applied to the largest network fragment.

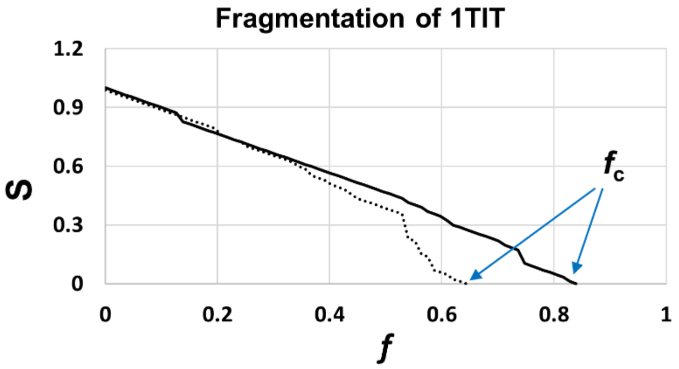


Fig. 28. Network fragmentation of 1TIT under random failures (solid line) and attacks (dash line). The fragmentation under random failures and attacks is shown for 1TIT as an example. This test was performed for all nine protein several times. The size of the largest cluster (S) as a function of the fraction of removed nodes (f). f_c is the fraction of the removed nodes when S = 0. Figure reproduced from [160].

Network Diameter

The network diameter (d) which represents the interconnectedness of a network is defined as the average length of the shortest paths between any two nodes in the graph. The diameter describes how all nodes can communicate with each other in a network. The smaller diameter represents the shorter path between any two nodes and higher connectivity in the network [171]. To test the significant role of the nodes with high BC value in maintaining the protein network's connectivity, diameter is calculated for each protein. We designed a program called *diameter* which calculate the changes in the diameter of the protein networks as a function of the fraction of the removed nodes. This algorithm finds the length of the shortest paths between all possible pairs of nodes in the network and then calculates the average of these values that is called diameter. Then, it removes the set of high BC nodes from the original network one at the time from highest to lowest and recalculates the diameter after each removal. After each elimination, the algorithm checks whether the network is fragmented and then selects the largest network fragment to proceed further node removals. The same process is also performed for 1000 sets of randomly selected nodes. Finally, the average of these 1000 runs and the standard deviation, are calculated.

RESULTS AND DISCUSSION

Betweenness Centrality Calculation

To understand and compare the nature of robustness in the networks for our model group of proteins, three computational studies are conducted using network principles. In the first test, BC enables the identification of residues in accordance with network science theory that govern the network [161, 170]. BC is calculated and graphed against amino acid numbers for each protein network. As shown in Figure 29, the residues with the highest BC values, based on a calculation of two standards of deviation or more from the mean, are selected in each graph to test how they behave differently from the other residues in protein network structure. The geographical position of the select residues are determined in the 3D structure of the model proteins. The proteins are color-coded based on the schematic of the Greek-key topology as shown in Figure 30. The selection of high BC values is also further supported by a Z-score analysis (Fig. 31) [160].

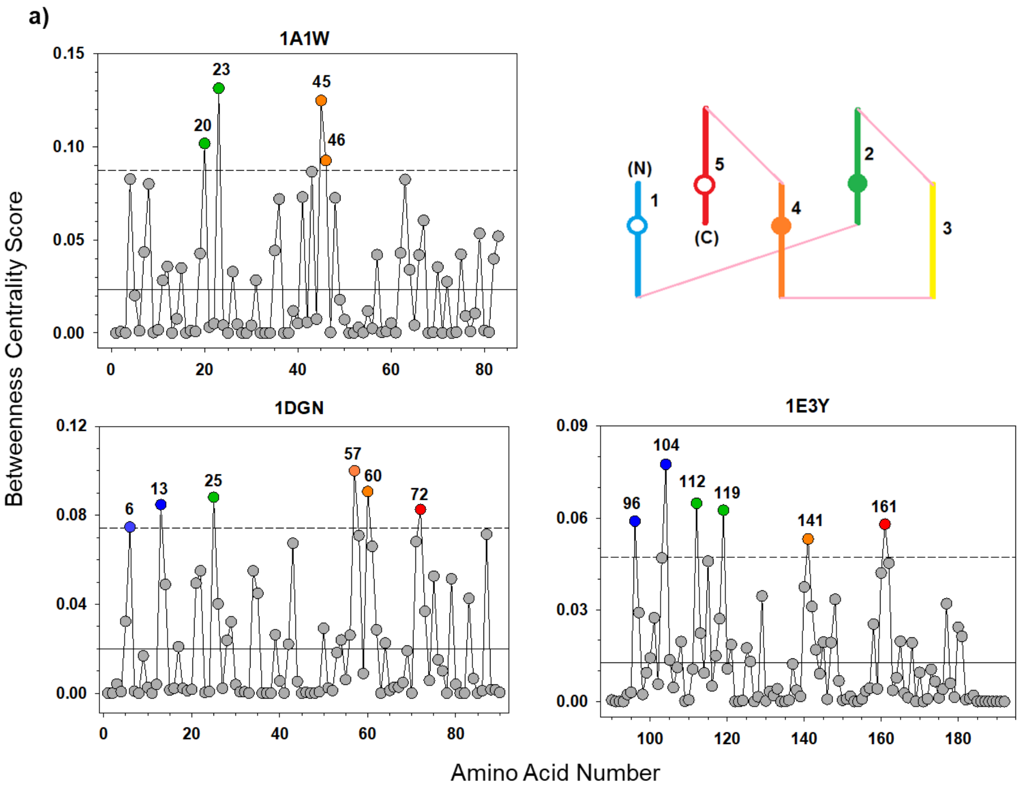
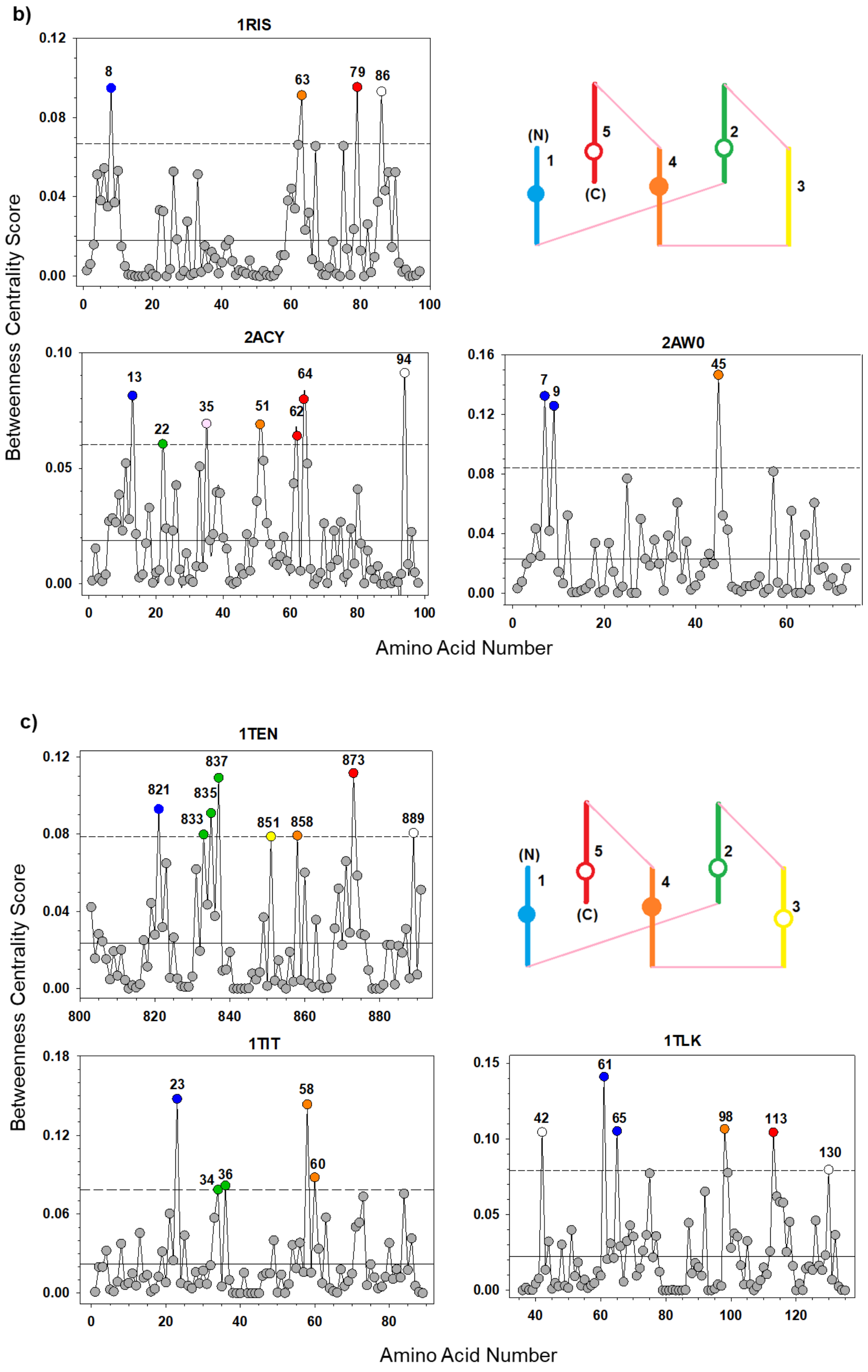


Fig. 29. Betweeness centrality versus amino acid numbers. (a) Death domains: 1A1W, 1DGN, 1E3Y, (b) α/β plaits: 1RIS, 2ACY, 2AW0, (c) Immunoglobulins: 1TEN, 1TIT, 1TLK. Residues with the BC values higher than the doubled standard deviation in each protein are shown in solid colored circles according to the color-coded schematic of Greek-key topology. High BC residues that are found in the similar geographical positions in all three proteins in each set are shown in solid circles and those that are only present in one or two proteins are shown in open circles in a generic schematic of the Greek-key topology. The solid line and the dashed line in each graph show the mean and two standards of deviation, respectively. Figure reproduced from [160].

Figure 29 Continued



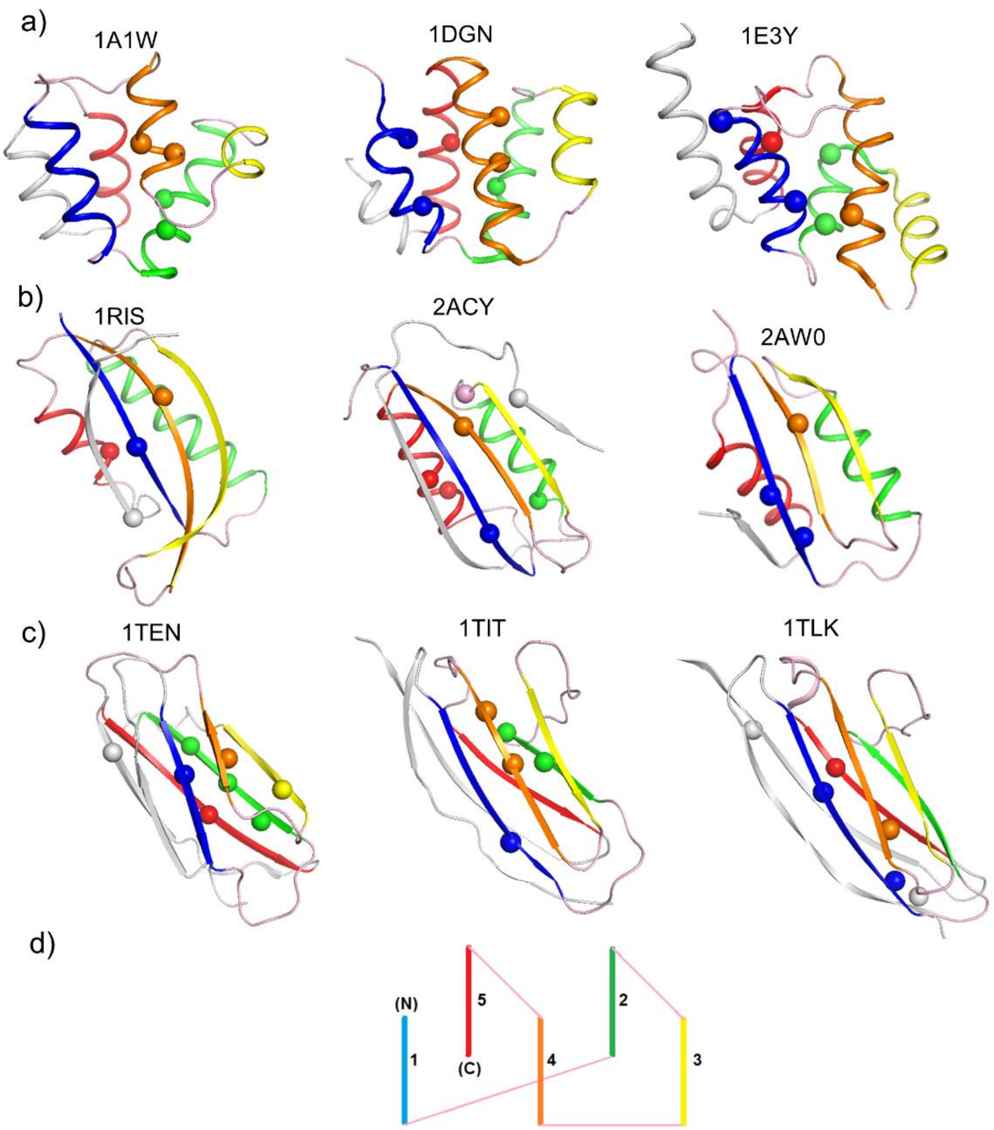


Fig. 30. Illustration of high BC residues in the 3D structure of proteins that color coded based on the schematic of the Greek-key topology. Select high BC residues are shown in spheres in all proteins. (a) Death domains: 1A1W, 1DGN, 1E3Y. (b) α/β plaits: 1RIS, 2ACY, 2AW0. (c) Immunoglobulins: 1TEN, 1TIT, 1TLK. (d) Schematic of the Greek-key topology. Common structural elements are shown in blue, green, yellow, orange and red and loops are shown in pink. Figure reproduced from [160].

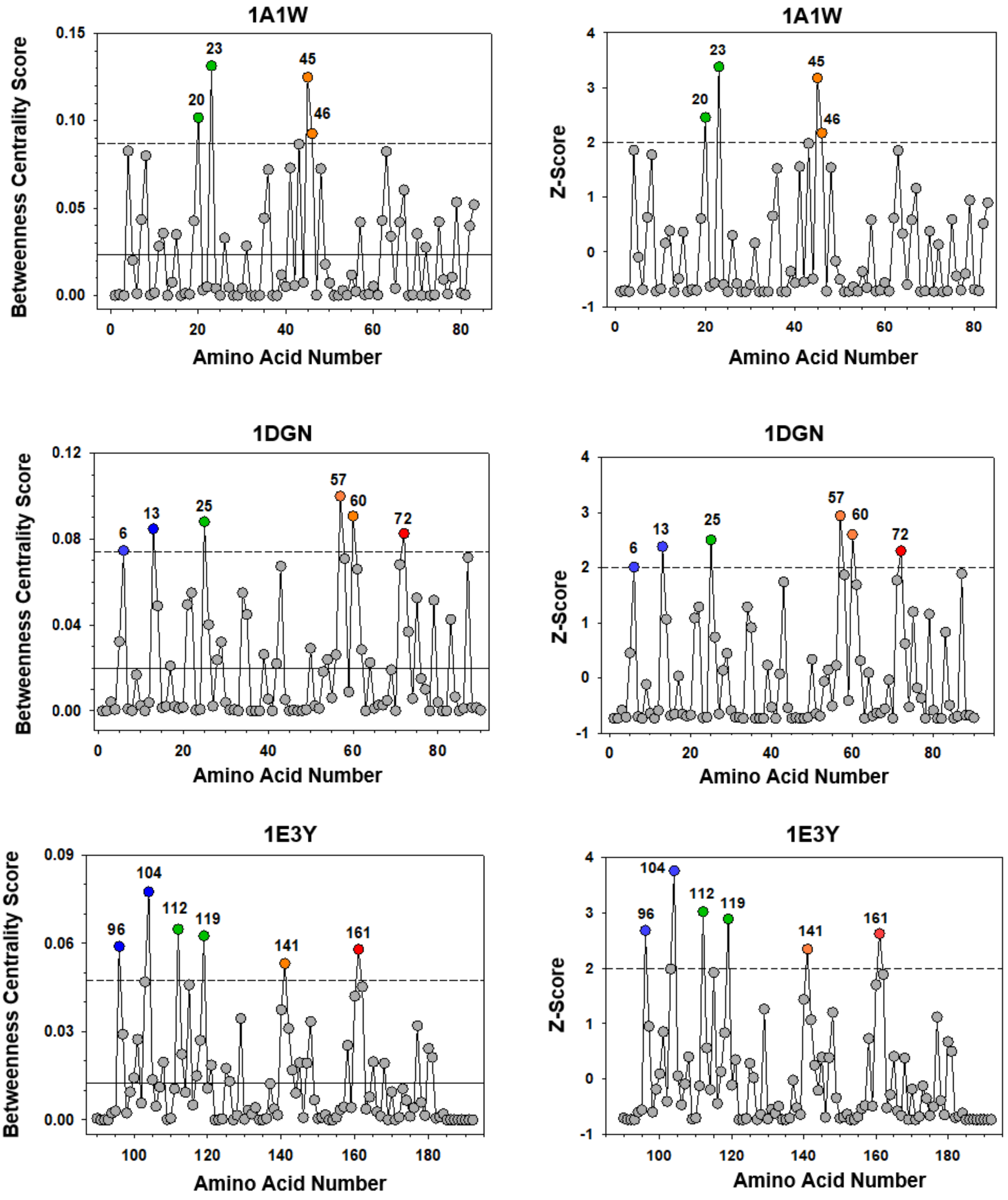


Fig. 31. Comparison between two standards of deviation from the mean and Z-scores. In all. proteins those residues with BC values ≥ 2SD score also Z-score ≥ 2, except residue 34 of 1TIT and residue 57 of 2AW0. However, the BC value of residue 57 is very close to the 2SD value and residue 34 has a Z-score very close to 2. Figure reproduced from [160].

Figure 31 Continued

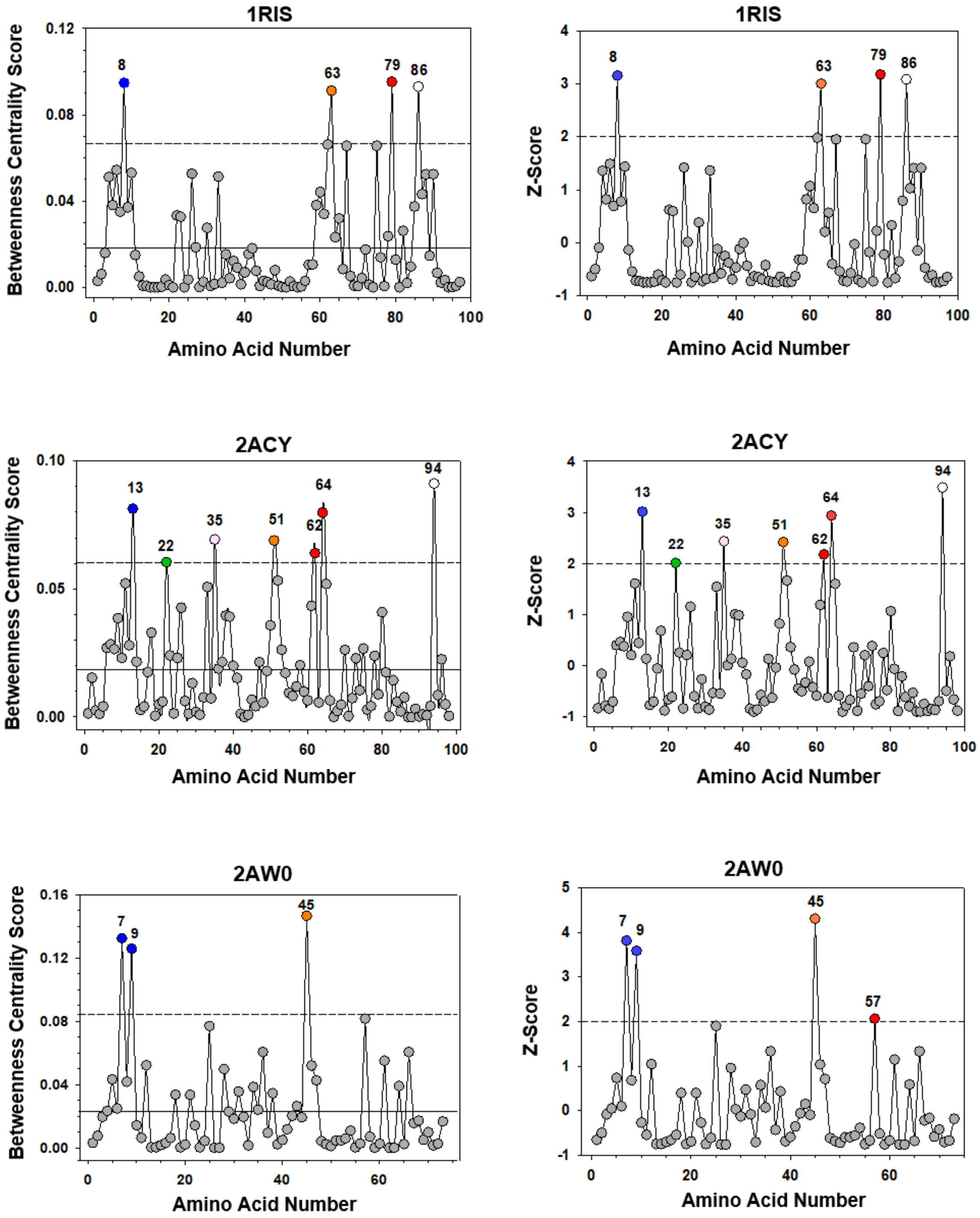
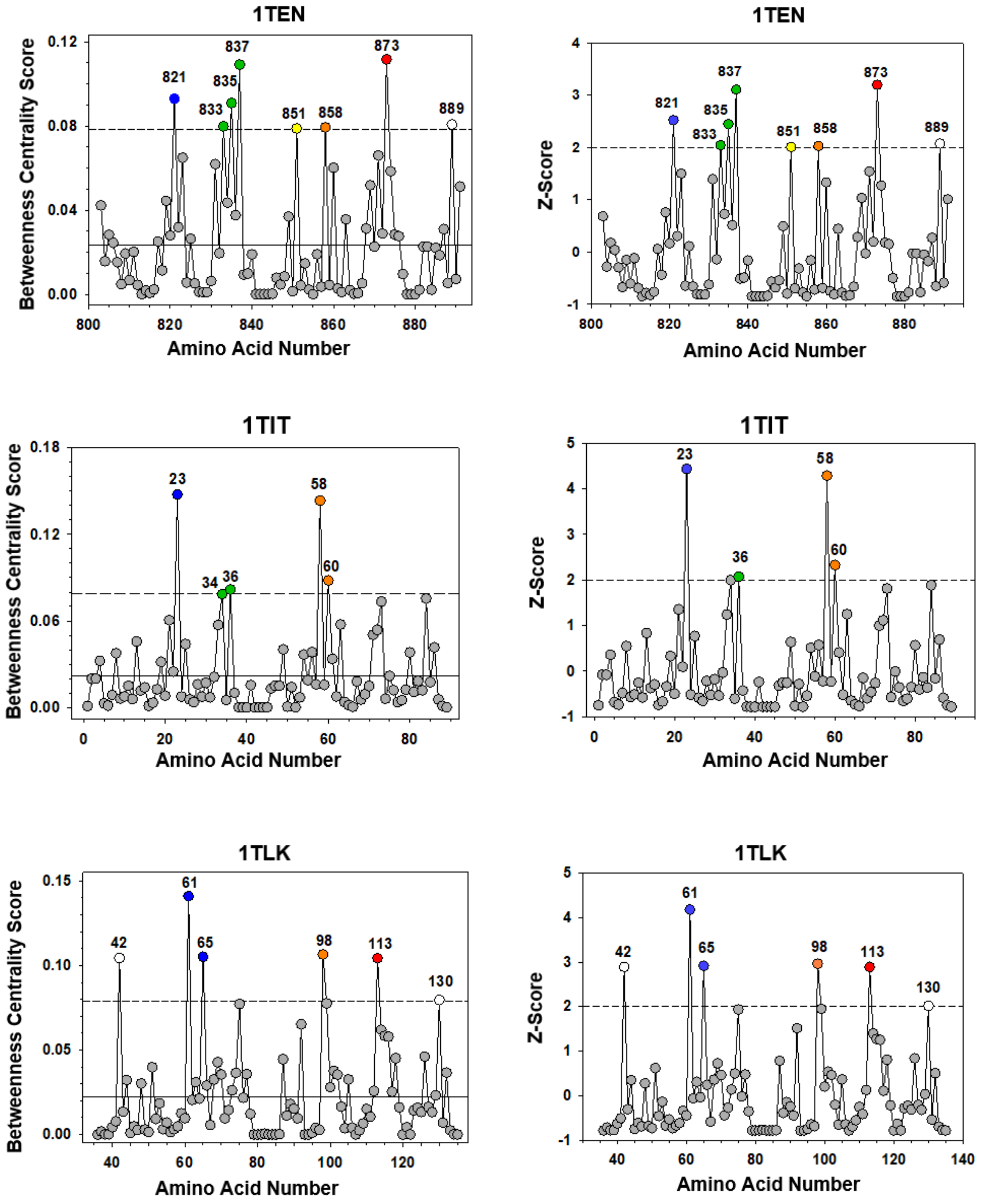


Figure 31 Continued



Fragmentation Evaluation

To test the role of the high BC residues in the protein structural stability, two computational approaches are applied. In the fragmentation test, we evaluated the integrity of the network subject to random as well as directed attacks by removing nodes with high BC values. Fragmentation is performed for each protein network and f_c which is the fraction of the removed nodes when S = 0 is recorded for both the directed attack and the random removal. This test is performed several times for each protein and the average of the f_c of all runs are provided and graphed (Fig. 32) [160].

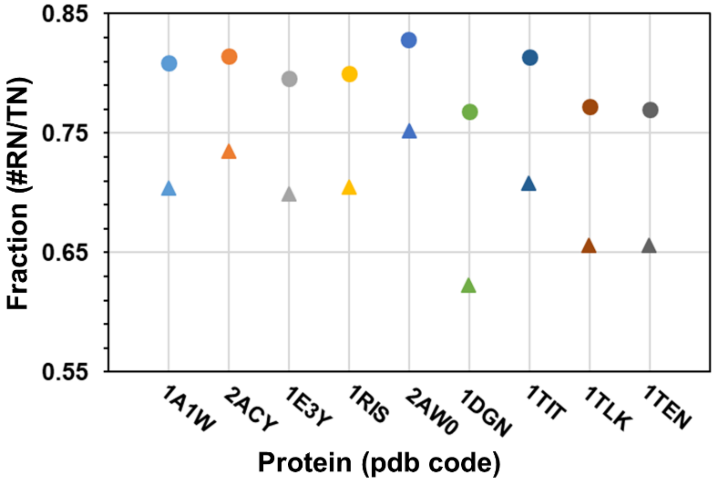


Fig. 32. Network fragmentation under random failures (●) and directed attacks (▲). These fractions are recorded when the size of the main network is zero (S = 0) and it is completely collapsed. RN and TN are abbreviation of removed nodes and total nodes, respectively. The standard deviation of 5 runs of randomly selected nodes in each protein demonstrated a very small range (2.8x 10^{-3} – 7.6 x 10^{-2}). Figure reproduced from [160].

Diameter Evaluation

In the second test, we calculated changes in the network diameter after elimination of two different sets of nodes: high BC nodes and randomly selected nodes. The diameter calculation can feasibly assess the impacts of node and links with respect to the ability of traversing a network. To have statistically more accurate result, 1000 different sets of randomly selected residues were run for each protein (Fig. 33) [160].

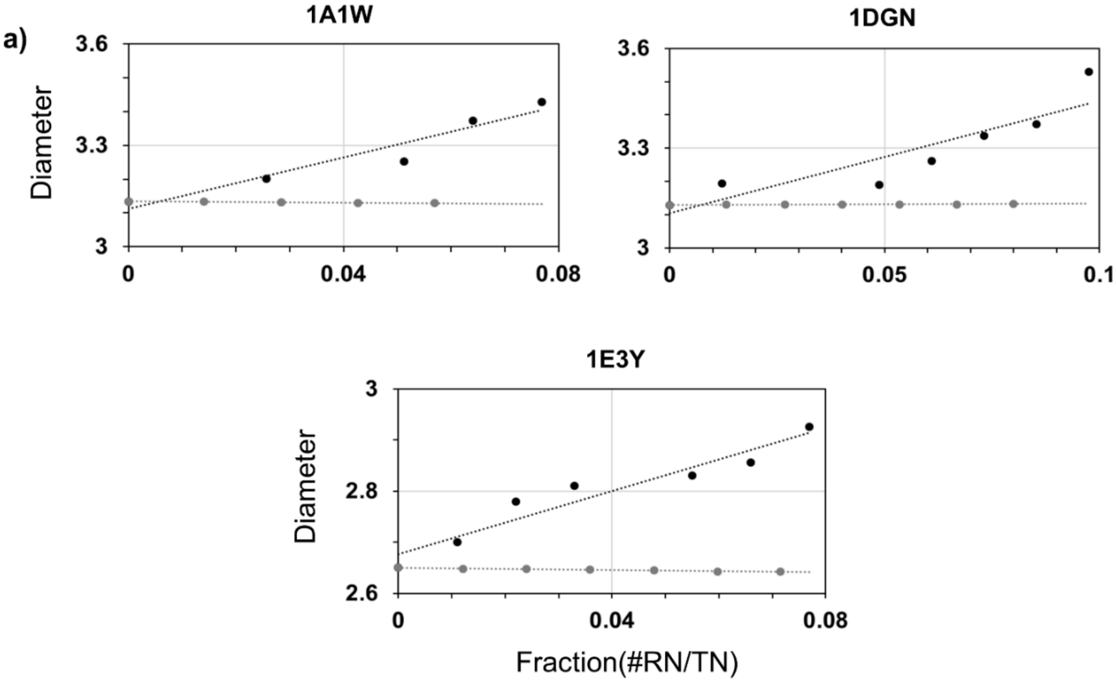
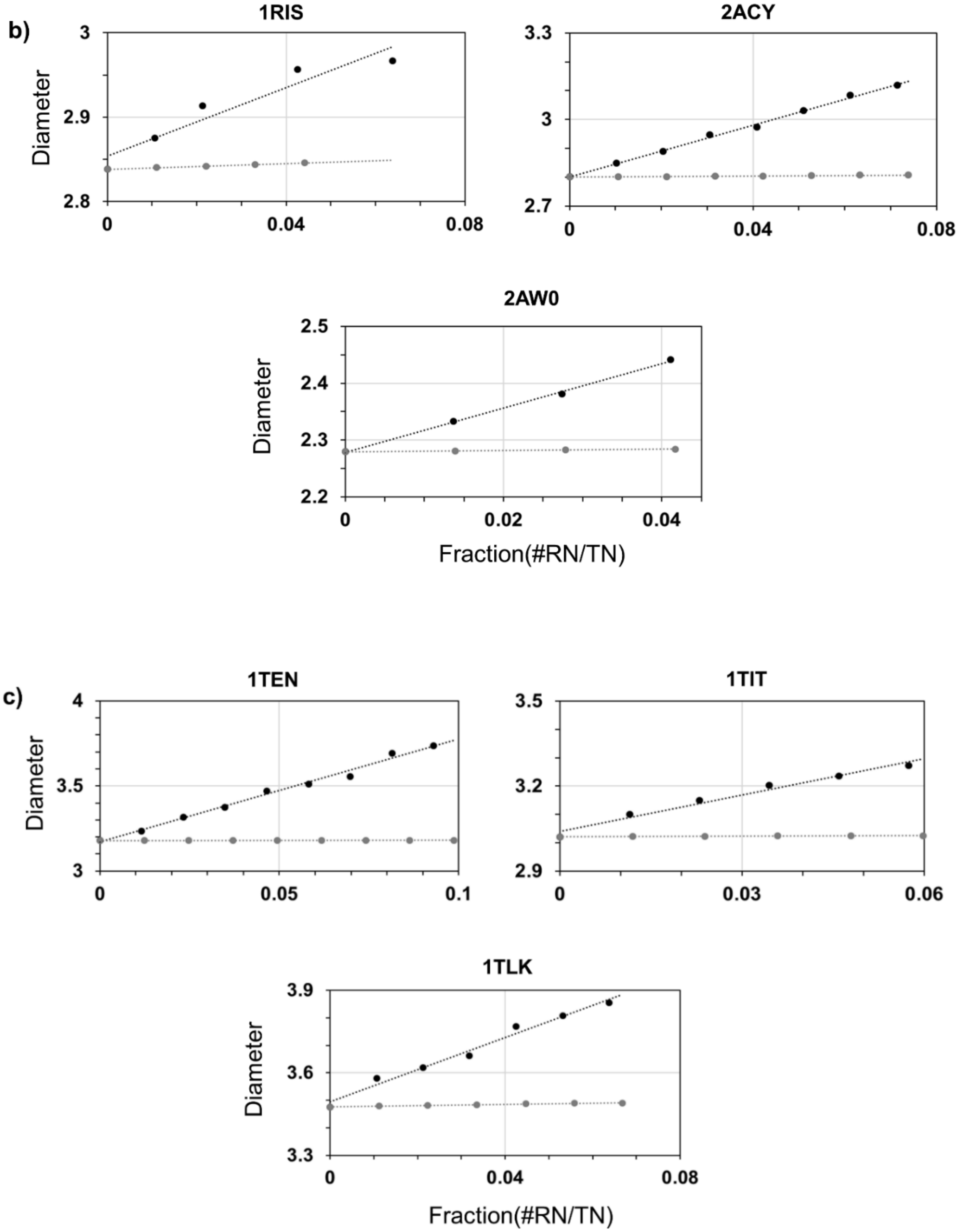


Fig. 33. Changes in the diameter of the protein networks as a function of the fraction of the removed nodes. Black circles represent nodes with high BC scores. Gray circles show the average of 1000 runs of randomly selected nodes. The standard deviation of 1000 runs of demonstrated a very small range ($1.3x10^{-4}$ - $3.7x10^{-3}$). (a) Death domains: 1A1W, 1DGN, 1E3Y. b) α/β-plaits: 1RIS, 2ACY, 2AW0. (c) Immunoglobulins: 1TEN, 1TIT, 1TLK. RN and TN are abbreviation of removed nodes and total nodes, respectively. Figure reproduced from [160].

Figure 33 Continued



The Application of Network Science to the Study of Protein Structures

Computational network approaches can offer important new avenues to analyze complex protein structures [157, 165]. The aim of this study was to apply different measures used in network science to elucidate specific residues that may play a role in the structural organization and stability of the Greek-key topology. Towards this end we selected a test set of Greek-key proteins with different sequences, secondary structure and functions for this purpose. The application of BC values was conducted to look for common geographical positions in all model proteins which may suggest the importance of these residues and regions in governing and stabilizing the Greek-key topology. Most of the residues found with the highest BC values are in the central core of the Greek-key topology at elements colored blue, green, orange, and red (Fig. 30). Among the residues with the highest BC scores, in all proteins but one, (1A1W), they are positioned on element 1 (blue) and 4 (orange) which is consistent with an earlier more restrictive study of three Greek-key proteins [96]. However, there is some variation in the other elements. Three proteins do not have residues with high BC on element 2, three do not have high BC residues on element 5, and eight are missing high BC residues on element 3. Even if one or two of the core elements in a protein model did not show the residues with the highest BC value (≥ 2SD) , it demonstrates residues with the BC value very close to the 2SD delineation, like; residues 4 and 8 in element 1 and 63 in element 5 of 1A1W, residues 25 in element 2 and 57 in element 5 of 2AW0, residue 73 in element 5 of 1TIT, residue 75 in element 2 of 1TLK [160].

Calculated long-range interactions involving the high BC residues (≥ 2SD) is shown in Table 8. Seven out of nine proteins have contacts between elements 1-4. The remaining pairs of elements have more variation between proteins. It is, however, interesting to consider that these contacts may interact early when the Greek-key topology forms [96]. Thus, they can potentially act as a scaffold to facilitate rapidly and correctly making the native structure as proposed by the 'levels of separation' model [121].

Table 8. Calculated long-range contacts within selected high BC residues in element 1, 2, 4, 5. Table reproduced from [160].

Protein	Element 1-4	Element 1-5	Element 4-5	Element 1-2	Element 2-4	Element 2-5
1E3Y	Ile104-Val141	Asp96-Leu161 Ile104-Leu161		Ile104-Trp112	Trp112-Val141 Leu119-Val141	Leu119-Leu161
1A1W					Leu20-Leu45 Leu20-Phe46 Leu23-Leu45 Leu23-Phe46	
1DGN	Phe13-Leu57	Leu6-Phe72 Phe13-Phe72	Leu57-Phe72	Phe13-Leu25	Leu25-Leu57	Leu25-Phe72
1RIS	Ile8-Tyr63	Ile8-Leu79				
2AW0	Ile7-Val45 Ile9-Val45					
2ACY		Ile13-Gln62	Leu51-Gln62	Ile13-Phe22		Phe22-Trp64
1TIT	Ile23-Leu58			Ile23-Trp34	Trp34-Leu58 Trp34-Leu60 Leu36-Leu58 Leu36-Leu60	
1TEN	Ile821-Tyr858	Ile821-Leu873		Ile821-Tyr835	Leu835-Tyr858 Tyr837-Tyr858	Ile833-Leu873 Tyr835-Leu873
1TLK	Phe61-Cys98 Val65-Cys98	Phe61-Tyr113				

The colors correspond to the secondary elements in the Greek-key schematic.

In an examination of the high BC residues that are present in the 6 or 7 Å graphs, in comparison to 5Å cutoff graphs, there are some common features (Fig. 34). In all but 1A1W, the proteins have high BC residues on elements 1 (blue) and 4 (orange). Although interestingly, residues 4 and 8 in the 7Å cutoff graph are very close to the 2SD delineation. Alternatively, it may be that element 2 (green) plays an analogous role in the topology to element 1 (blue) enabling this change to be substitutive in nature. In general, as we reduce the cutoff distance, the number of the contacts and the size of the network changes which can have some effect on the network properties. In this instance we look for trends to provide insight [160].

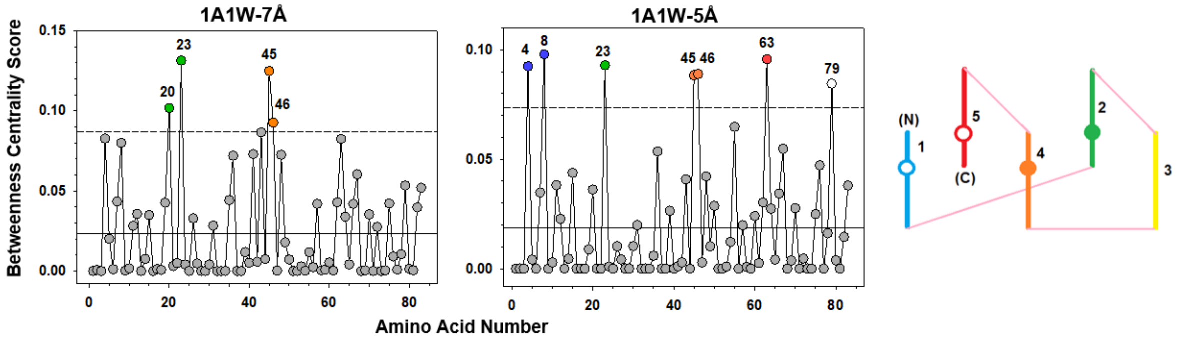


Fig. 34. Betweenness centrality scores at 5Å and 6 or 7Å cutoff distances. Residues with the high-BC values higher than the two standard deviation or more cutoff in each protein are shown in solid colored circles according to the color-coded schematic of Greek-key topology. In the schematic of the Greek-key topology on the right side of the graphs, solid circles show a high BC residue that is present in both cutoffs. Open circles show the high BC residue that is only present in one of the cutoffs. The solid line and the dashed line in each graph show the mean and two standards of deviation, respectively. Figure reproduced from [160].

Figure 34 Continued

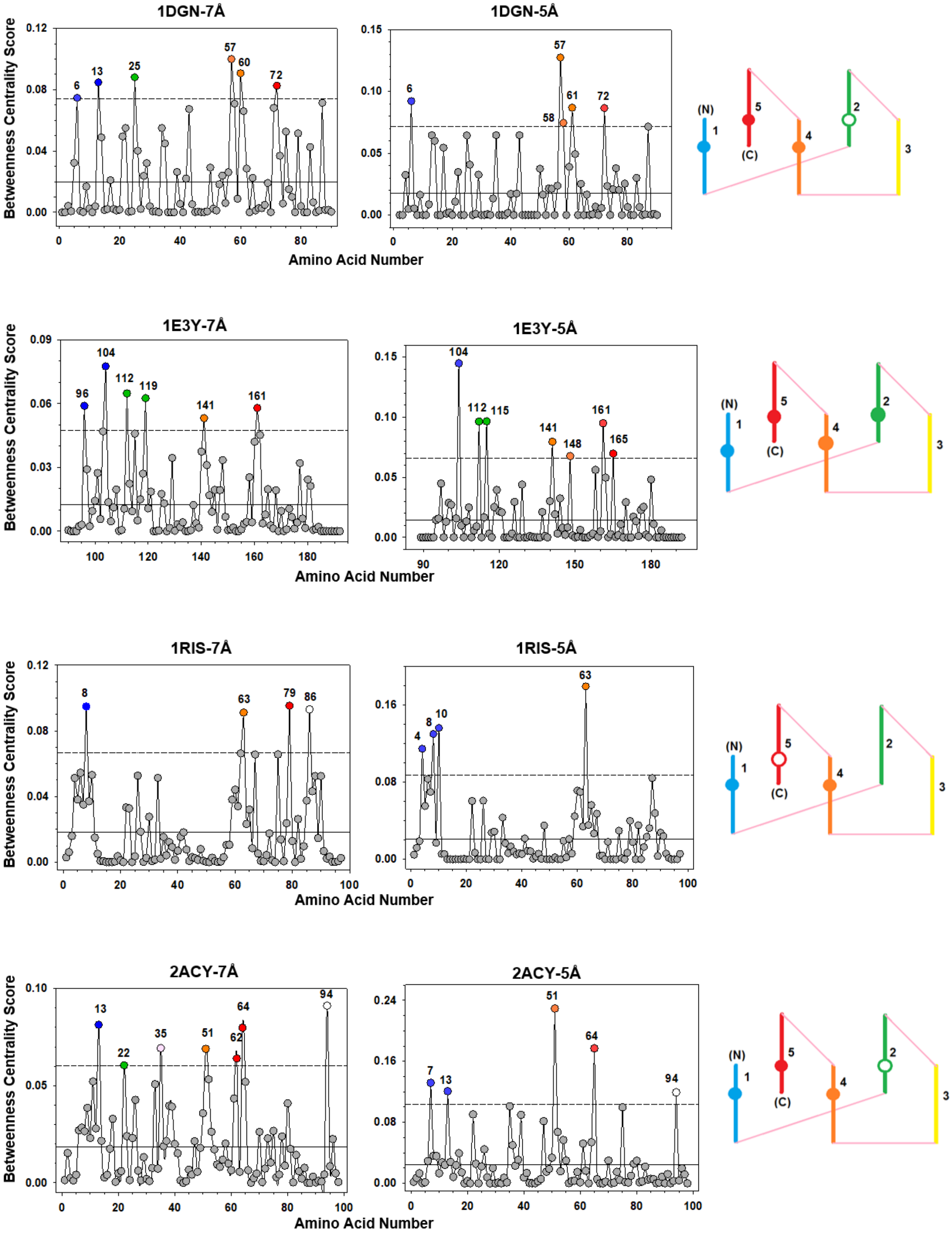
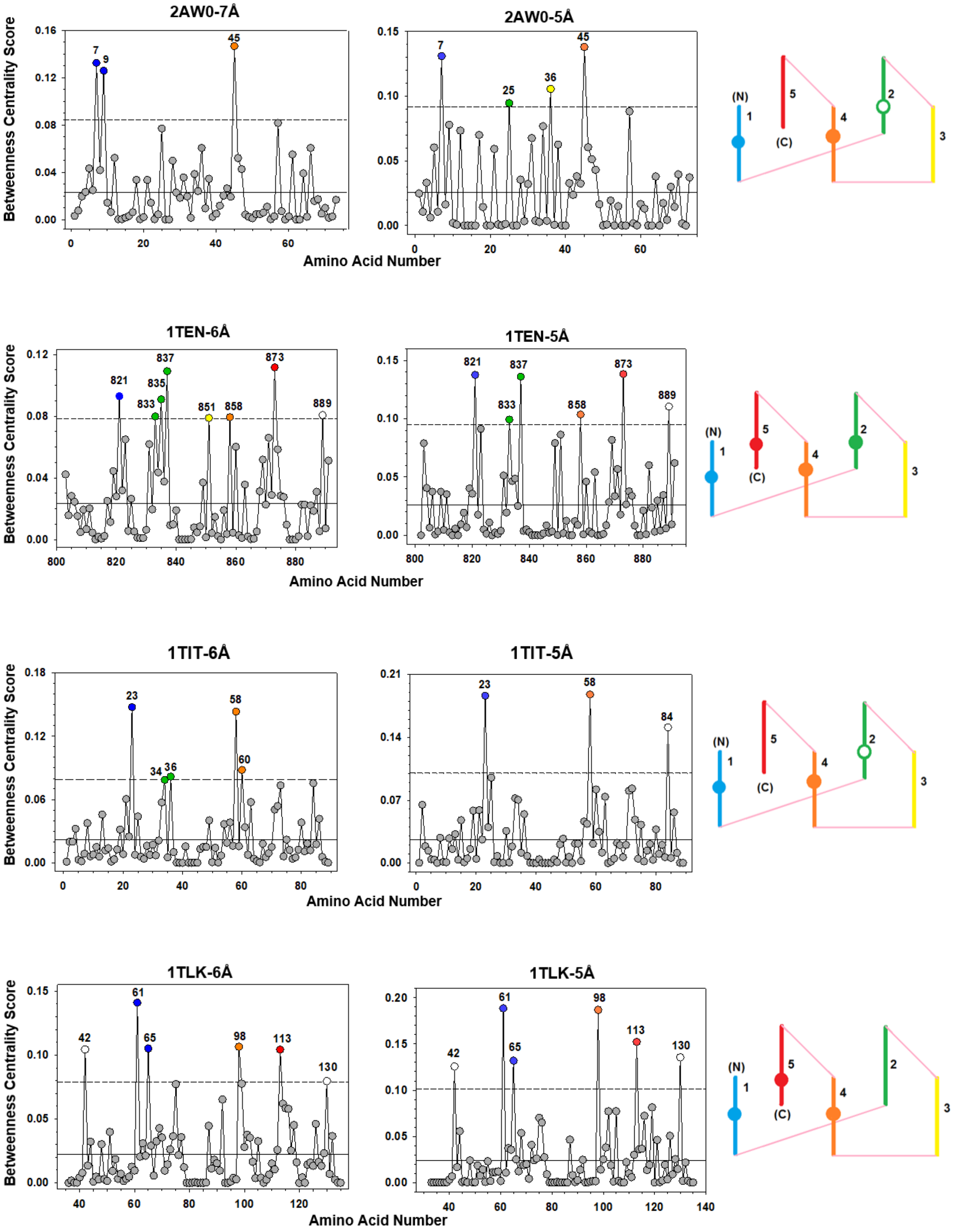


Figure 34 Continued



A further analysis of BC residues involved applying the CC measure to the long-range interaction networks (Fig. 35). The results indicate that overwhelmingly, the residues with high CC have high BC scores. Therefore, part of being strategically connected in terms of BC appears to be ‘closeness’ to all other amino acids in the network [160].

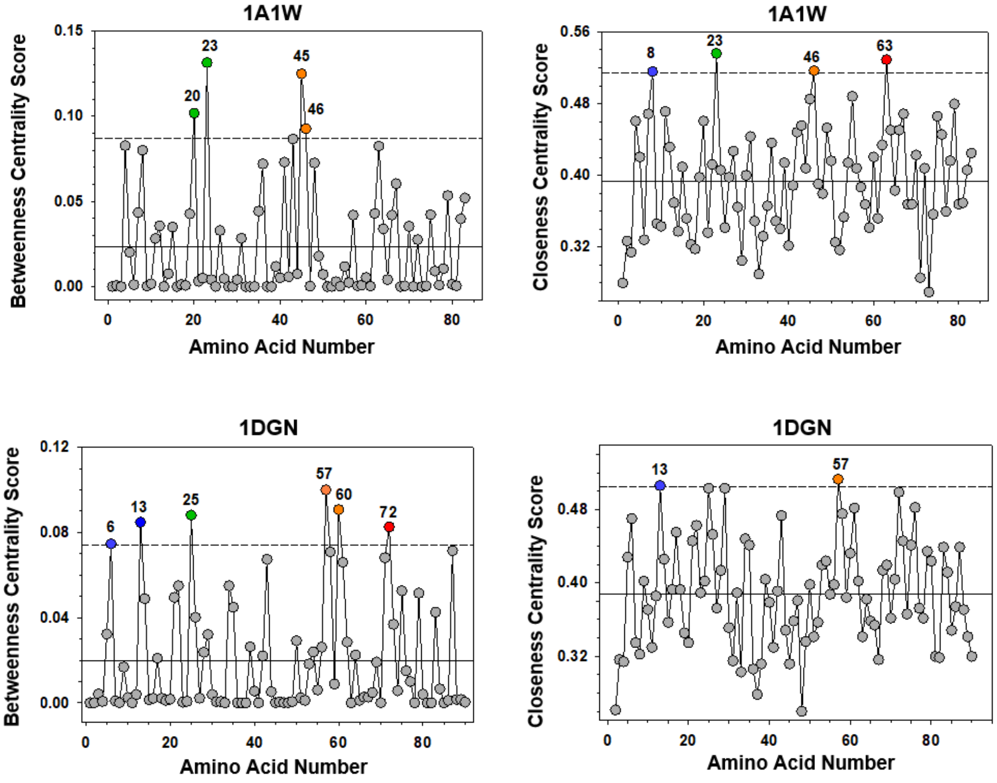


Fig. 35. Comparison of betweenness centrality (BC) and closeness centrality (CC) scores at 6Å cutoff for immunoglobulins and 7Å cutoff distance for α/β plaits and death domains. Residues with the BC and CC values higher than the doubled standard deviation value in each protein are shown in solid colored circles according to the color-coded schematic of Greek-key topology. The solid line and the dashed line in each graph show the mean and two standards of deviation, respectively. Figure reproduced from [160].

Figure 35 Continued

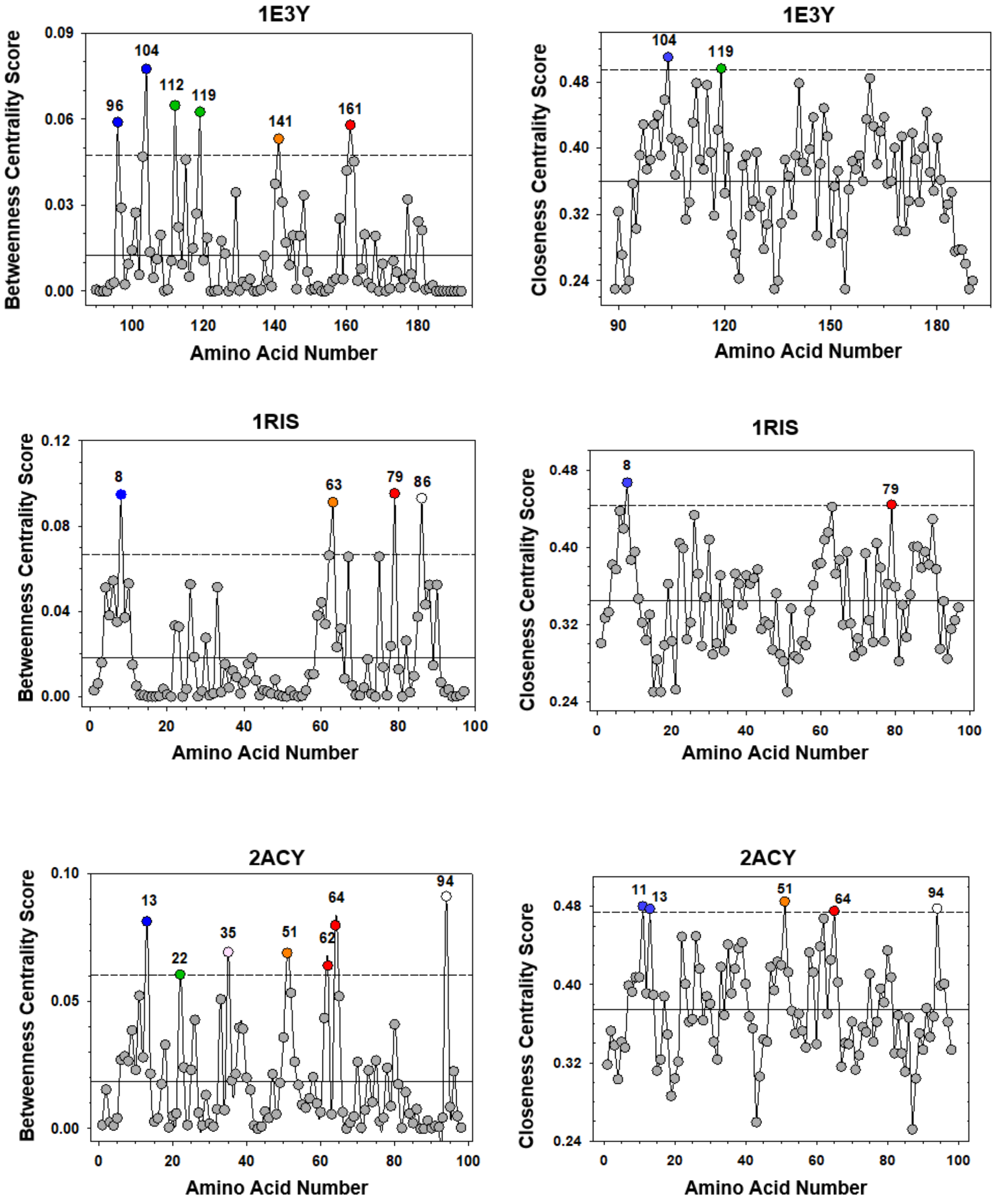


Figure 35 Continued

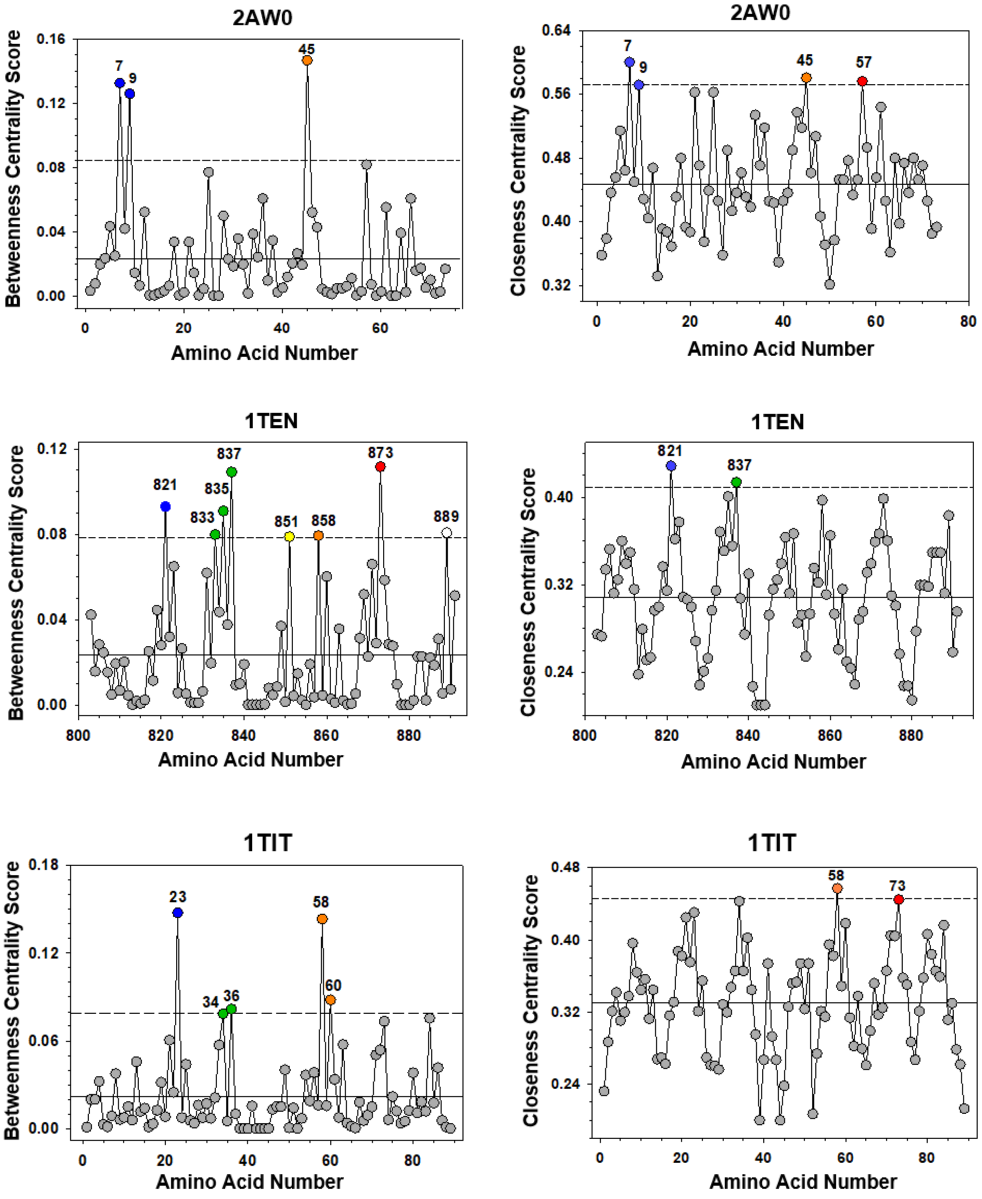
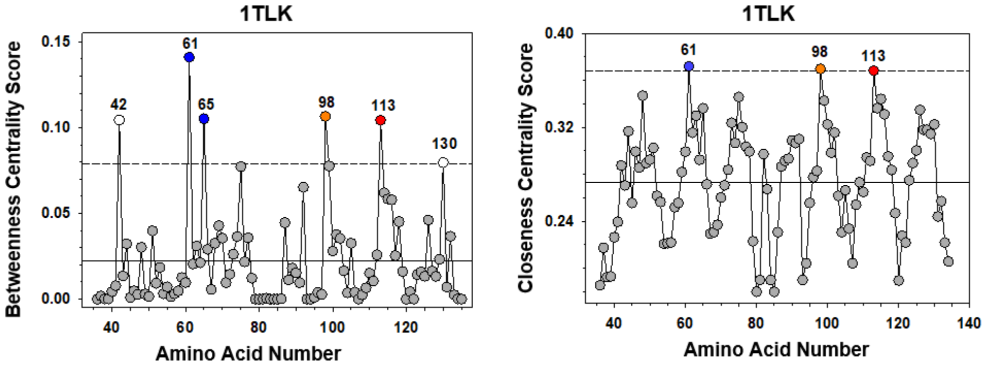


Figure 35 Continued



The results of fragmentation and diameter methods also indicate the importance of the residues with high BC value in the network connectivity, centrality and fragmentation of protein networks. In the fragmentation test shown in Figure 32, in all proteins the main network completely collapsed (S = 0.0) after randomly removing 76-82% of total number of the residues. But this range is reduced to 62-73%, when the high BC nodes are removed in advance. These results not only suggest a similar robustness in all protein networks but also show the preferred role of the residues with high BC value in holding the network together. Based on the result of the diameter test shown in Figure 33, there is a sharper increase in the diameter after removing nodes with high BC values compared to the randomly selected nodes. These observed results further reveal the importance of these residues in the connectivity and lethality of the protein network [160].

Reported Experimental Results

There are also experimental studies on five of the model proteins that can be used to compare to the network results. In these studies, mutations are performed on the specific residues and Phi-value (ϕ) are calculated from folding experiments. Phi-values are a measure derived from mutational studies that probe the contribution of individual amino acids in the structure of the transition state using kinetic and thermodynamic calculations [145]. Residues that show significant Phi-values are considered to play an important role in guiding the protein structure and folding process by forming early in the transition-state [146]. As shown in Table 9, some of the high BC residues that demonstrated a significant role in governing the Greek-key topology in our computational network studies, are also confirmed by the experimental method. Thus, there may be common roles for high BC residues relating to the stability of the transition-state and folding or they relate only to one of these biophysical functions.

Table 9. Correlation between residues with high BC values and experimental ϕ-values. Residues in bold indicate amino acids that are common with the BC studies. Table reproduced from [160].

Protein	Residues with high BC values (Computational Studies)	Residues with significant ϕ-values (Experimental Studies)
1E3Y	96*, **104**, **112**, 119, **141**, **161**	101, **104, 112,** 115, 140, **141,** 144, 160**, 161,** 162, 165, 173
1RIS	**8**, 63, 79, 86*	4, 6, **8**, 9, 26, 30, 60, 65, 67, 75
2ACY	**13**, 22, 35*, **51**, 62*, **64**, **94**	11, **13**, 30, 42, 45, 47, **51**, 54, **64**, **94**
1TIT	**23**, 34*, **36**, **58**, **60**	2, 19, 21, **23**, 25, 30, 32, **36**, 41, 47, 49, 56, **58, 60**, 71, 73, 75
1TEN	**821**, 833, **835**, **837**, **851**, **858**, **873**, 889	**821, 835, 837**, 849, **851**, **858**, 860, 863, 865, 867, 869, 871, **873**

Moderate to high Phi-values are denoted based on the following criteria: ϕ-values between 0.3-1.0 at ϕ^{0M} and $\phi^{midpoint}$ for 1RIS [147]. ϕ-values between 0.3-1.0 at ϕ^{1M} for both 1TIT [148] and 1E3Y [172] ϕ-values between 0.3-1.0 for 1TEN [150] and 2ACY [151]. Residues with high BC that have not been experimentally studied are denoted with *.

CHAPTER V

CONCLUSION AND FUTURE WORK

The research in this dissertation provides a comprehensive investigation into the determinants of structural stability and topology for the Greek-key proteins. These proteins are an ideal model system because they vary in sequence, secondary structure and function but share a common form. A group of nine Greek-key proteins selected from three different superfamilies, were studied by different computational approaches.

As discussed in chapter 2, selected proteins were subjected to high temperature conditions using MD simulation, to test the behavior and stability of the conserved long-range contacts in comparison to non-conserved contacts. During the unfolding simulations, the conserved contacts demonstrated more persistence in comparison to the nonconserved contacts in the protein. Under high temperature conditions, the conserved contacts are found to be more resistant to breaking and showed less fluctuation than the other native contacts. These results suggest an important role for the conserved contacts in the inherent structural stability of our model proteins. This also infers that they are important for topological determination.

Following the stability studies our work focused on investigating the determinants of topology and folding. The two major computational investigations reveal that fundamental determinants of protein topology consist of evolutionarily conserved long-range interaction networks. More specifically, through the application of SA simulations using CNS the significant role of the conserved contacts at the onset of folding was shown. The network study involving the formation of a giant network cluster supports the idea that the conserved contacts are more important than non-conserved contacts in governing and building the protein network structure.

The purpose of the last study was to examine the role of the high BC residues in the structural determination of our protein models. Residues with the highest betweenness centrality values proposed to control the protein network are generally found on the same elements in the selected protein models. These geographical positions are the central four-element motif of the Greek-key topology that have been considered as the core of the topology with element 1 and 4 potentially most crucial. Even if one or two of the core elements in a protein model did not show the residues with the highest BC value (≥ 2SD), it demonstrates residues with the BC value very close to the 2SD delineation. The importance of the high BC residues in the connectivity and lethality of the protein networks were tested by different network measures: fragmentation and diameter tests. The results demonstrate the significance of the high BC residues in the specific geographical positions and may possibly guide the Greek-key topology in these proteins. The importance of some of these specific residues were also confirmed with the experimental analysis.

In summary, it can be suggested that a conserved network of long-range interactions play a significant role in building and stabilizing the protein structure. This conserved network indicated to be a main determinant of a common Greek-key topology in proteins that differ in sequences and secondary structures. It should also be mentioned that the specific geographical positions of residues can make them an important determinant in controlling and maintaining the protein network.

To further examine the role of the conserved interactions network in the protein's structural stability, other analysis can be performed on the unfolded MD trajectories. The unfolded trajectories of each protein can be clustered and analyzed to see at which point the protein will completely lose its topology. This analysis will demonstrate how similar is the stability of the Greek-key topology in the nine different protein. Also, pairwise residue interaction energies and energy correlations from protein MD simulation trajectories can be generated and analyzed using gRINN (get Residue Interaction eNergies and Networks). This analysis would highlight those interactions that show more stability during the unfolding process. The importance of the selected residues and interactions in protein folding, can be further tested by NMR spectroscopy. The selected residues can be labeled and monitored during the folding or unfolding process using NMR. Also, another interesting study can be performed to test the role of the high BC residues in protein folding. The selected residues can be mutated on the extended form of the protein and then folded by simulated annealing method. The RMSD and TM_score of the simulated structures made by the mutated type would be compared with those made by the wild type. This analysis can be performed for all the high BC residues in each protein one at the time to see which one can affect the protein folding the most and play a more significant role in protein folding.

REFERENCES

1. Nelson, D. L., Lehninger, A. L. & Cox, M. M. (2008). Lehninger principles of biochemistry. Macmillan.
2. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH–a hierarchic classification of protein domain structures. Structure 5, 1093-1109.
3. Greene, L. H., Lewis, T. E., Addou, S., Cuff, A., Dallman, T., Dibley, M., Redfern, O., Pearl, F., Nambudiry, R., Reid, A., Sillitoe, I., Yeats, C., Thornton, J. M. & Orengo, C. A. (2006). The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res. 35, D291-D297.
4. Andreeva, A., Howorth, D., Chandonia, J., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. (2007). Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. 36, D419-D425.
5. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536-540.
6. Voet, D., Voet, J. G. & Pratt, C. W. (2016). Fundamentals of biochemistry: life at the molecular level. John Wiley & Sons.
7. Csaba, G., Birzele, F. & Zimmer, R. (2009). Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. BMC Struct. Biol. 9, 23.
8. Sillitoe, I., Dawson, N., Thornton, J. & Orengo, C. (2015). The history of the CATH structural classification of protein domains. Biochimie 119, 209-217.
9. Makhatadze, G. I. (2017). Linking computation and experiments to study the role of charge–charge interactions in protein folding and stability. Phys. Biol. 14, 013002.
10. Durell, S. R. & Ben-Naim, A. (2017). Hydrophobic-hydrophilic forces in protein folding. Biopolymers 107, e23020.
11. Ben-Naim, A. (2011). The rise and fall of the hydrophobic effect in protein folding and protein-protein association, and molecular recognition. Open J. Biophys. 1, 1.
12. Tanford, C. & Reynolds, J. (2003). Nature's robots: a history of proteins. OUP Oxford.
13. Pace, C. N., Scholtz, J. M. & Grimsley, G. R. (2014). Forces stabilizing proteins. FEBS Lett. 588, 2177-2184.
14. Obalinsky, T. R. (2006). Protein folding: New research. Nova Publishers.
15. Selvaraj, S. & Michael Gromiha, M. (2003). Role of Hydrophobic Clusters and Long-Range Contact Networks in the Folding of (α/β)8 Barrel Proteins. Biophys. J. 84, 1919-1925.
16. Gromiha, M. M. & Selvaraj, S. (1999). Importance of long-range interactions in protein folding. Biophys. Chem. 77, 49-68.
17. Sengupta, D. & Kundu, S. (2012). Role of long- and short-range hydrophobic, hydrophilic and charged residues contact network in protein’s structural organization. BMC bioinformatics 13, 142.
18. Cregut, D., Civera, C., Macias, M. J., Wallon, G. & Serrano, L. (1999). A tale of two secondary structure elements: when a β-hairpin becomes an α-helix. J. Mol. Biol. 292, 389-401.
19. Tanaka, S. & Scheraga, H. A. (1975). Model of protein folding: inclusion of short-, medium-, and long-range interactions. Proc. Natl. Acad. Sci. 72, 3802-3806.
20. Go, N. & Taketomi, H. (1978). Respective roles of short-and long-range interactions in protein folding. Proc. Natl. Acad. Sci. 75, 559-563.
21. Kumarevel, T. S., Gromiha, M. M., Selvaraj, S., Gayatri, K. & Kumar, P. K. R. (2002). Influence of medium- and long-range interactions in different folding types of globular proteins. Biophys. Chem. 99, 189-198.
22. Thakur, A. K., Meng, W. & Gierasch, L. M. (2018). Local and non-local topological information in the denatured state ensemble of a β-barrel protein. Protein Sci. 27, 2062-2072.

23. Meng, W., Lyle, N., Luan, B., Raleigh, D. P. & Pappu, R. V. (2013). Experiments and simulations show how long-range contacts can form in expanded unfolded proteins with negligible secondary structure. Proc. Natl. Acad. Sci. 110, 2123-2128.

24. Meng, W., Luan, B., Lyle, N., Pappu, R. V. & Raleigh, D. P. (2013). The denatured state ensemble contains significant local and long-range structure under native conditions: analysis of the N-terminal domain of ribosomal protein L9. Biochemistry 52, 2662-2671.

25. Plaxco, K. W., Simons, K. T. & Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. J. Mol. Biol. 277, 985-994.

26. Ouyang, Z. & Liang, J. (2008). Predicting protein folding rates from geometric contact and amino acid sequence. Protein Sci. 17, 1256-1263.

27. Gromiha, M. M. & Selvaraj, S. (2001). Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. J. Mol. Biol. 310, 27-32.

28. Istomin, A. Y., Jacobs, D. J. & Livesay, D. R. (2007). On the role of structural class of a protein with two-state folding kinetics in determining correlations between its size, topology, and folding rate. Protein Sci. 16, 2564-2569.

29. Harihar, B. & Selvaraj, S. (2009). Refinement of the long-range order parameter in predicting folding rates of two-state proteins. Biopolymers 91, 928-935.

30. McKee, T. & McKee, J. R. (2003). Biochemistry: the molecular basis of life. McGraw-Hill New York.

31. Nölting, B. (2005). Protein folding kinetics: biophysical methods. Springer Science & Business Media.

32. Fersht, A. (1999). Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding. Macmillan.

33. Fersht, A. R. (1995). Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. Proc. Natl. Acad. Sci. 92, 10869-10873.

34. Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. J. Mol. Biol. 254, 260-288.

35. Fersht, A. R., Matouschek, A. & Serrano, L. (1992). The folding of an enzyme: I. Theory of protein engineering analysis of stability and pathway of protein folding. J. Mol. Biol. 224, 771-782.

36. Pandit, A. D., Jha, A., Freed, K. F. & Sosnick, T. R. (2006). Small proteins fold through transition states with native-like topologies. J. Mol. Biol. 361, 755-770.

37. Fersht, A. R. & Daggett, V. (2002). Protein folding and unfolding at atomic resolution. Cell 108, 573-582.

38. Wolynes, P. G. (2015). Evolution, energy landscapes and the paradoxes of protein folding. Biochimie 119, 218-230.

39. Brooks, C. L., Gruebele, M., Onuchic, J. N. & Wolynes, P. G. (1998). Chemical physics of protein folding. Proc. Natl. Acad. Sci. 95, 11037-11038.

40. Lehninger, A. L., Nelson, D. L., Cox, M. M. & Cox, M. M. (2005). Lehninger principles of biochemistry. Macmillan.

41. Thommen, M., Holtkamp, W. & Rodnina, M. V. (2017). Co-translational protein folding: progress and methods. Curr. Opin. Struct. Biol. 42, 83-89.

42. Guinn, E. J., Tian, P., Shin, M., Best, R. B. & Marqusee, S. (2018). A small single-domain protein folds through the same pathway on and off the ribosome. Proc. Natl. Acad. Sci. 115, 12206-12211.

43. Tian, P., Steward, A., Kudva, R., Su, T., Shilling, P. J., Nickson, A. A., Hollins, J. J., Beckmann, R., von Heijne, G., Clarke, J. & Best, R. B. (2018). Folding pathway of an Ig domain is conserved on and off the ribosome. Proc. Natl. Acad. Sci. 115, E11284-e11293.

44. Nilsson, O. B., Hedman, R., Marino, J., Wickles, S., Bischoff, L., Johansson, M., Muller-Lucks, A., Trovato, F., Puglisi, J. D., O'Brien, E. P., Beckmann, R. & von Heijne, G. (2015). Cotranslational Protein Folding inside the Ribosome Exit Tunnel. Cell Rep. 12, 1533-1540.

45. Holtkamp, W., Kokic, G., Jager, M., Mittelstaet, J., Komar, A. A. & Rodnina, M. V. (2015). Cotranslational protein folding on the ribosome monitored in real time. Science 350, 1104-1107.

46. Waudby, C. A., Dobson, C. M. & Christodoulou, J. (2019). Nature and Regulation of Protein Folding on the Ribosome. Trends Biochem. Sci. 44, 914-926.

47. Motojima, F. (2015). How do chaperonins fold protein? Biophysics (Nagoya-shi) 11, 93-102.

48. Munoz, V. & Cerminara, M. (2016). When fast is better: protein folding fundamentals and mechanisms from ultrafast approaches. Biochem. J. 473, 2545-2559.

49. Dill, K. A. & MacCallum, J. L. (2012). The Protein-Folding Problem, 50 Years On. Science 338, 1042-1046.

50. Bouvier, B., Zakrzewska, K. & Lavery, R. (2011). Protein–DNA recognition triggered by a DNA conformational switch. Angew. Chem. Int. Ed. 50, 6516-6518.

51. Prisner, T., Marko, A. & Sigurdsson, S. T. (2015). Conformational dynamics of nucleic acid molecules studied by PELDOR spectroscopy with rigid spin labels. J. Magn. Reson. 252, 187-198.

52. Shinoda, T., Arai, K., Shigematsu-Iida, M., Ishikura, Y., Tanaka, S., Yamada, T., Kimber, M. S., Pai, E. F., Fushinobu, S. & Taguchi, H. (2005). Distinct conformation-mediated functions of an active site loop in the catalytic reactions of NAD-dependent D-lactate dehydrogenase and formate dehydrogenase. J. Biol. Chem. 280, 17068-17075.

53. Nussinov, R. & Tsai, C. (2015). Allostery without a conformational change? Revisiting the paradigm. Curr. Opin. Struct. Biol. 30, 17-24.

54. Bringas, M., Petruk, A. A., Estrin, D. A., Capece, L. & Martí, M. A. (2017). Tertiary and quaternary structural basis of oxygen affinity in human hemoglobin as revealed by multiscale simulations. Sci. Rep. 7, 10926.

55. Gaudin, Y. (2016). Protein conformational changes, from molecules to living organisms: an interplay between biology and physics. Oxford University Press 102, 353.

56. Klepeis, J. L., Lindorff-Larsen, K., Dror, R. O. & Shaw, D. E. (2009). Long-timescale molecular dynamics simulations of protein structure and function. Curr. Opin. Struct. Biol. 19, 120-127.

57. Hospital, A., Goñi, J. R., Orozco, M. & Gelpí, J. L. (2015). Molecular dynamics simulations: advances and applications. Adv. Appl. Bioinform. Chem. 8, 37.

58. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. a. & Karplus, M. (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. J. Comput. Chem. 4, 187-217.

59. Brooks, B. R., Brooks III, C. L., Mackerell Jr, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C. & Boresch, S. (2009). CHARMM: the biomolecular simulation program. J. Comput. Chem. 30, 1545-1614.

60. Shaw, D. E., Deneroff, M. M., Dror, R. O., Kuskin, J. S., Larson, R. H., Salmon, J. K., Young, C., Batson, B., Bowers, K. J. & Chao, J. C. (2008). Anton, a special-purpose machine for molecular dynamics simulation. Commun. ACM 51, 91-97.

61. Shaw, D. E., Grossman, J., Bank, J. A., Batson, B., Butts, J. A., Chao, J. C., Deneroff, M. M., Dror, R. O., Even, A. & Fenton, C. H. (2014). Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. IEEE Press 41-53.

62. Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M. & Pannu, N. S. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr. D Biol. Crystallogr. 54, 905-921.

63. Brunger, A. T. (2007). Version 1.2 of the Crystallography and NMR system. Nat. Protoc. 2, 2728.

64. Brünger, A. T. (1992). X-PLOR: version 3.1: a system for x-ray crystallography and NMR. Yale University Press.
65. Greene, L. H. & Higman, V. A. (2003). Uncovering network systems within protein structures. J. Mol. Biol. 334, 781-791.
66. Barabási, A.-L. (2016). Network science. Cambridge University Press.
67. Di Paola, L. & Giuliani, A. (2015). Protein contact network topology: a natural language for allostery. Curr. Opin. Struct. Biol. 31, 43-48.
68. Vishveshwara, S., Ghosh, A. & Hansia, P. (2009). Intra and inter-molecular communications through protein structure network. Curr. Protein Pept. Sci. 10, 146-160.
69. Costanzi, S. (2016). Topological Analyses of Protein-Ligand Binding: A Network Approach. Curr. Protein Pept. Sci. 17, 37-40.
70. Bhattacharyya, M., Ghosh, S. & Vishveshwara, S. (2016). Protein structure and function: looking through the network of side-chain interactions. Curr. Protein Pept. Sci. 17, 4-25.
71. Gromiha, M. M. (2009). Multiple contact network is a key determinant to protein folding rates. J. Chem. Inf. Model. 49, 1130-1135.
72. Dokholyan, N. V., Li, L., Ding, F. & Shakhnovich, E. I. (2002). Topological determinants of protein folding. Proc. Natl. Acad. Sci. 99, 8637-8641.
73. Böde, C., Kovács, I. A., Szalay, M. S., Palotai, R., Korcsmáros, T. & Csermely, P. (2007). Network analysis of protein dynamics. FEBS Lett. 581, 2776-2782.
74. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. (2019). Critical assessment of methods of protein structure prediction (CASP)—Round XIII. Proteins: Struc. Func. and Bioinform. 87, 1011-1020.
75. Guzenko, D., Lafita, A., Monastyrskyy, B., Kryshtafovych, A. & Duarte, J. M. (2019). Assessment of protein assembly prediction in CASP13. Proteins: Struc. Func. and Bioinform. 87, 1190-1199.
76. Garnier, J. (1990). Protein structure prediction. Biochimie 72, 513-524.
77. Lee, J., Freddolino, P. L. & Zhang, Y. (2017). Ab initio protein structure prediction: From protein structure to function with bioinformatics. Springer. p. 3-35.
78. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. (2016). Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. Proteins: Struc. Func. and Bioinform. 84, 4-14.
79. Abriata, L. A., Tamò, G. E. & Dal Peraro, M. (2019). A further leap of improvement in tertiary structure prediction in CASP13 prompts new routes for future assessments. Proteins: Struc. Func. and Bioinform. 87, 1100-1112.
80. Cheng, J., Tegge, A. N. & Baldi, P. (2008). Machine learning methods for protein structure prediction. IEEE Rev. Biomed. Eng. 1, 41-49.
81. Wardah, W., Khan, M. G., Sharma, A. & Rashid, M. A. (2019). Protein secondary structure prediction using neural networks and deep learning: a review. Comput. Biol. Chem.
82. Miyazawa, S. (2018). Prediction of Structures and Interactions from Genome Information. In: Nakamura H, Kleywegt G, Burley SK&Markley JL, editors. Integrative Structural Biology with Hybrid Methods,: Springer Singapore. p. 123-152.
83. Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S. & Weiner, P. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. J. Am. Chem. Soc. 106, 765-784.
84. Ołdziej, S., Czaplewski, C., Liwo, A., Chinchio, M., Nanias, M., Vila, J., Khalili, M., Arnautova, Y., Jagielska, A. & Makowski, M. o. (2005). Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: Assessment in two blind tests. Proc. Natl. Acad. Sci. 102, 7547-7552.
85. Das, R., Qian, B., Raman, S., Vernon, R., Thompson, J., Bradley, P., Khare, S., Tyka, M. D., Bhat, D. & Chivian, D. (2007). Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@ home. Proteins: Struc. Func. and Bioinform. 69, 118-128.

86. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J. & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. Nat. Methods 12, 7.
87. Berglund, H., Olerenshaw, D., Sankar, A., Federwisch, M., McDonald, N. Q. & Driscoll, P. C. (2000). The three-dimensional solution structure and dynamic properties of the human FADD death domain. J. Mol. Biol. 302, 171-188.
88. Eberstadt, M., Huang, B., Chen, Z., Meadows, R. P., Ng, S.-C., Zheng, L., Lenardo, M. J. & Fesik, S. W. (1998). NMR structure and mutagenesis of the FADD (Mort1) death-effector domain. Nature 392, 941-945.
89. Humke, E. W., Shriver, S. K., Starovasnik, M. A., Fairbrother, W. J. & Dixit, V. M. (2000). ICEBERG: A Novel Inhibitor of Interleukin-1β Generation. Cell 103, 99-111.
90. Gitschier, J., Moffat, B., Reilly, D., Wood, W. I. & Fairbrother, W. J. (1998). Solution structure of the fourth metal-binding domain from the Menkes copper-transporting ATPase. Nat. Struct. Biol. 5, 47-54.
91. Lindahl, M., Svensson, L. A., Liljas, A., Sedelnikova, S. E., Eliseikina, I. A., Fomenkova, N. P., Nevskaya, N., Nikonov, S. V., Garber, M. B., Muranova, T. A. & et al. (1994). Crystal structure of the ribosomal protein S6 from Thermus thermophilus. EMBO J. 13, 1249-1254.
92. Thunnissen, M. M., Taddei, N., Liguri, G., Ramponi, G. & Nordlund, P. (1997). Crystal structure of common type acylphosphatase from bovine testis. Structure 5, 69-79.
93. Improta, S., Politou, A. S. & Pastore, A. (1996). Immunoglobulin-like modules from titin I-band: extensible components of muscle elasticity. Structure 4, 323-337.
94. Holden, H. M., Ito, M., Hartshorne, D. J. & Rayment, I. (1992). X-ray structure determination of telokin, the C-terminal domain of myosin light chain kinase, at 2.8 Å resolution. J. Mol. Biol. 227, 840-851.
95. Leahy, D. J., Hendrickson, W. A., Aukhil, I. & Erickson, H. P. (1992). Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. Science 258, 987-991.
96. Higman, V. A. & Greene, L. H. (2006). Elucidation of conserved long-range interaction networks in proteins and their significance in determining protein topology. Physica A 368, 595-606.
97. Finkelstein, A. (2018). 50+ years of protein folding. Biochemistry (Moscow) 83, S3-S18.
98. Deller, M. C., Kong, L. & Rupp, B. (2016). Protein stability: a crystallographer's perspective. Acta Crystallogr. F Struct. Biol. Commun. 72, 72-95.
99. Pace, C. N., Fu, H., Fryar, K. L., Landua, J., Trevino, S. R., Shirley, B. A., Hendricks, M. M., Iimura, S., Gajiwala, K. & Scholtz, J. M. (2011). Contribution of hydrophobic interactions to protein stability. J. Mol. Biol. 408, 514-528.
100. Pace, C. N., Fu, H., Lee Fryar, K., Landua, J., Trevino, S. R., Schell, D., Thurlkill, R. L., Imura, S., Scholtz, J. M. & Gajiwala, K. (2014). Contribution of hydrogen bonds to protein stability. Protein Sci. 23, 652-661.
101. Somero, G. N. (2004). Adaptation of enzymes to temperature: searching for basic "strategies". Comp. Biochem. Physiol. B Biochem. Mol. Biol. 139, 321-333.
102. Fields, P. A., Dong, Y., Meng, X. & Somero, G. N. (2015). Adaptations of protein structure and function to temperature: there is more than one way to 'skin a cat'. J. Exp. Biol. 218, 1801-1811.
103. Dill, K. A. (1990). Dominant forces in protein folding. Biochemistry 29, 7133-7155.
104. Gough, J. D. & Lees, W. J. (2005). Effects of redox buffer properties on the folding of a disulfide-containing protein: dependence upon pH, thiol pKa, and thiol concentration. J. Biotechnol. 115, 279-290.
105. Kumar, S., Tsai, C.-J. & Nussinov, R. (2000). Factors enhancing protein thermostability. Protein Eng. 13, 179-191.
106. Araya, C. L., Fowler, D. M., Chen, W., Muniez, I., Kelly, J. W. & Fields, S. (2012). A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. Proc. Natl. Acad. Sci. 109, 16858-16863.
107. Bieri, M., Kwan, A. H., Mobli, M., King, G. F., Mackay, J. P. & Gooley, P. R. (2011). Macromolecular NMR spectroscopy for the non-spectroscopist: beyond macromolecular solution structure determination. FEBS J. 278, 704-715.

108. Krishnan, V. & Rupp, B. (2001). Macromolecular structure determination: comparison of X-ray crystallography and NMR spectroscopy. e LS.

109. Magliery, T. J. (2015). Protein stability: computation, sequence statistics, and new experimental methods. Curr. Opin. Struct. Biol. 33, 161-168.

110. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. (2006). Protein stability promotes evolvability. Proc. Natl. Acad. Sci. 103, 5869-5874.

111. Freddolino, P. L., Harrison, C. B., Liu, Y. & Schulten, K. (2010). Challenges in protein-folding simulations. Nat. Phys. 6, 751.

112. Samish, I., MacDermaid, C. M., Perez-Aguilar, J. M. & Saven, J. G. (2011). Theoretical and computational protein design. Annu. Rev. Phys. Chem. 62, 129-149.

113. Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T. B., Montelione, G. T. & Baker, D. (2012). Principles for designing ideal protein structures. Nature 491, 222.

114. Yin, S., Ding, F. & Dokholyan, N. V. (2007). Eris: an automated estimator of protein stability. Nat. Methods 4, 466.

115. Potapov, V., Cohen, M. & Schreiber, G. (2009). Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. Protein Eng. Des. Sel. 22, 553-560.

116. Scheraga, H. A., Khalili, M. & Liwo, A. (2007). Protein-folding dynamics: overview of molecular simulation techniques. Annu. Rev. Phys. Chem. 58, 57-83.

117. Geer, M. A. & Fitzgerald, M. C. (2014). Energetics-based methods for protein folding and stability measurements. Annu. Rev. Anal. Chem. 7, 209-228.

118. Magliery, T. J. & Regan, L. (2004). Combinatorial approaches to protein stability and structure. Eur. J. Biochem. 271, 1595-1608.

119. Magliery, T. J., Lavinder, J. J. & Sullivan, B. J. (2011). Protein stability by number: high-throughput and statistical approaches to one of protein science's most difficult problems. Curr. Opin. Struct. Biol. 15, 443-451.

120. Tzul, F. O., Schweiker, K. L. & Makhatadze, G. I. (2015). Modulation of folding energy landscape by charge–charge interactions: Linking experiments with computational modeling. Proc. Natl. Acad. Sci. 112, E259-E266.

121. Greene, L. H. & Grant, T. M. (2012). Protein folding by 'levels of separation': A hypothesis. FEBS Lett. 586, 962-966.

122. Gromiha, M. M. & Selvaraj, S. (2004). Inter-residue interactions in protein folding and stability. Prog. Biophys. Mol. Biol. 86, 235-277.

123. Bagler, G. & Sinha, S. (2007). Assortative mixing in protein contact networks and protein folding kinetics. Bioinformatics 23, 1760-1767.

124. Grimsley, G. R., Shaw, K. L., Fee, L. R., Alston, R. W., Huyghues-Despointes, B. M., Thurlkill, R. L., Scholtz, J. M. & Pace, C. N. (1999). Increasing protein stability by altering long-range coulombic interactions. Protein Sci. 8, 1843-1849.

125. Tripathi, S., Garcìa, A. E. & Makhatadze, G. I. (2015). Alterations of nonconserved residues affect protein stability and folding dynamics through charge–charge interactions. J. Phys. Chem. B 119, 13103-13112.

126. Loladze, V. V., Ibarra-Molero, B., Sanchez-Ruiz, J. M. & Makhatadze, G. I. (1999). Engineering a thermostable protein via optimization of charge− charge interactions on the protein surface. Biochemistry 38, 16419-16423.

127. Gribenko, A. V. & Makhatadze, G. I. (2007). Role of the charge–charge interactions in defining stability and halophilicity of the CspB proteins. J. Mol. Biol. 366, 842-856.

128. Strickler, S. S., Gribenko, A. V., Gribenko, A. V., Keiffer, T. R., Tomlinson, J., Reihle, T., Loladze, V. V. & Makhatadze, G. I. (2006). Protein stability and surface electrostatics: a charged relationship. Biochemistry 45, 2761-2766.

129. Nick Pace, C., Alston, R. W. & Shaw, K. L. (2000). Charge–charge interactions influence the denatured state ensemble and contribute to protein stability. Protein Sci. 9, 1395-1398.
130. Gromiha, M. M., Pathak, M. C., Saraboji, K., Ortlund, E. A. & Gaucher, E. A. (2013). Hydrophobic environment is a key factor for the stability of thermophilic proteins. Proteins: Struc. Func. and Bioinform. 81, 715-721.
131. Takano, K., Yamagata, Y. & Yutani, K. (1998). A general rule for the relationship between hydrophobic effect and conformational stability of a protein: stability and structure of a series of hydrophobic mutants of human lysozyme. J. Mol. Biol. 280, 749-761.
132. Kellis Jr, J. T., Nyberg, K. & Fersht, A. R. (1989). Energetics of complementary side chain packing in a protein hydrophobic core. Biochemistry 28, 4914-4922.
133. Zhou, R., Huang, X., Margulis, C. J. & Berne, B. J. (2004). Hydrophobic Collapse in Multidomain Protein Folding. Science 305, 1605-1609.
134. Kalinowska, B., Banach, M., Wisniowski, Z., Konieczny, L. & Roterman, I. (2017). Is the hydrophobic core a universal structural element in proteins? J. Mol. Biol. 23, 205.
135. Ventura, S., Vega, M. C., Lacroix, E., Angrand, I., Spagnolo, L. & Serrano, L. (2002). Conformational strain in the hydrophobic core and its implications for protein folding and design. Nat. Struct. Mol. Biol. 9, 485.
136. Munson, M., Balasubramanian, S., Fleming, K. G., Nagi, A. D., O'Brien, R., Sturtevant, J. M. & Regan, L. (1996). What makes a protein a protein? Hydrophobic core designs that specify stability and structural properties. Protein Sci. 5, 1584-1593.
137. Klein-Seetharaman, J., Oikawa, M., Grimshaw, S. B., Wirmer, J., Duchardt, E., Ueda, T., Imoto, T., Smith, L. J., Dobson, C. M. & Schwalbe, H. (2002). Long-range interactions within a nonnative protein. Science 295, 1719-1722.
138. Felitsky, D. J., Lietzow, M. A., Dyson, H. J. & Wright, P. E. (2008). Modeling transient collapsed states of an unfolded protein to provide insights into early folding events. Proc. Natl. Acad. Sci. 105, 6278-6283.
139. Marsh, J. A. & Forman-Kay, J. D. (2009). Structure and disorder in an unfolded state under nondenaturing conditions from ensemble models consistent with a large number of experimental restraints. J. Mol. Biol. 391, 359-374.
140. Bertoncini, C. W., Jung, Y.-S., Fernandez, C. O., Hoyer, W., Griesinger, C., Jovin, T. M. & Zweckstetter, M. (2005). Release of long-range tertiary interactions potentiates aggregation of natively unstructured α-synuclein. Proc. Natl. Acad. Sci. 102, 1430-1435.
141. Childers, M. C. & Daggett, V. (2017). Insights from molecular dynamics simulations for computational protein design. Mol. Syst. Des. Eng. 2, 9-33.
142. Kufareva, I. & Abagyan, R. (2011). Methods of protein structure comparison. Homology Modeling,: Springer. p. 231-257.
143. Vaught, A. (1996). Graphing with Gnuplot and Xmgr: two graphing packages available under linux. Linux J. 1996, 7.
144. Tama, F., Gadea, F. X., Marques, O. & Sanejouand, Y. H. (2000). Building-block approach for determining low-frequency normal modes of macromolecules. Proteins: Struc. Func. and Bioinform. 41, 1-7.
145. Fersht, A. R. (1999). Structure and Mechanism in Protein Science. New York: WH Freeman and Company.
146. Fersht, A. R. & Sato, S. (2004). Φ-value analysis and the nature of protein-folding transition states. Proc. Natl. Acad. Sci. 101, 7976-7981.
147. Otzen, D. E. & Oliveberg, M. (2002). Conformational plasticity in folding of the split β-α-β protein S6: evidence for burst-phase disruption of the native state. J. Mol. Biol. 317, 613-627.
148. Fowler, S. B. & Clarke, J. (2001). Mapping the Folding Pathway of an Immunoglobulin Domain: Structural Detail from Phi Value Analysis and Movement of the Transition State. Structure 9, 355-366.

149. Steward, A., McDowell, G. S. & Clarke, J. (2009). Topology is the principal determinant in the folding of a complex all-alpha Greek key death domain from human FADD. J. Mol. Biol. 389, 425-437.
150. Cota, E., Steward, A., Fowler, S. B. & Clarke, J. (2001). The folding nucleus of a fibronectin type III domain is composed of core residues of the immunoglobulin-like fold. J. Mol. Biol. 305, 1185-1194.
151. Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M. & Dobson, C. M. (1999). Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. Nat. Struct. Mol. Biol. 6, 1005-1009.
152. Diana, D., De Rosa, L., Palmieri, M., Russomanno, A., Russo, L., La Rosa, C., Milardi, D., Colombo, G., D'Andrea, L. D. & Fattorusso, R. (2015). Long range Trp-Trp interaction initiates the folding pathway of a pro-angiogenic beta-hairpin peptide. Sci. Rep. 5, 16651.
153. Zhang, Y. & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33, 2302-2309.
154. Zhang, Y. & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. Proteins: Struc. Func. and Bioinform. 57, 702-710.
155. Gerstein, M. & Levitt, M. (1998). Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. Protein Sci. 7, 445-456.
156. Hung, L.-H. & Samudrala, R. (2012). Accelerated protein structure comparison using TM-score-GPU. Bioinformatics 28, 2191-2192.
157. Greene, L. H. (2012). Protein structure networks. Brief. Funct. Genomics 11, 469-478.
158. Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G. & McCoy, A. (2011). Overview of the CCP4 suite and current developments. Acta Crystallogr. D Biol. Crystallogr. 67, 235-242.
159. De Nooy, W., Mrvar, A. & Batagelj, V. (2018). Exploratory social network analysis with Pajek. Cambridge University Press.
160. Haratipour, Z., Aldabagh, H., Li, Y. & Greene, L. H. (2019). Network Connectivity, Centrality and Fragmentation in the Greek-Key Protein Topology. Protein J. 1-9.
161. Dorogovtsev, S. N. & Mendes, J. F. F. (2003). Evolution of networks: From biological nets to the internet and WWW. Oxford: Oxford University Press.
162. Miller, E. J., Fischer, K. F. & Marqusee, S. (2002). Experimental evaluation of topological parameters determining protein-folding rates. Proc. Natl. Acad. Sci. 99, 10359-10363.
163. Li, J., Wang, J. & Wang, W. (2008). Identifying folding nucleus based on residue contact networks of proteins. Proteins: Struc. Func. and Bioinform. 71, 1899-1907.
164. Aftabuddin, M. & Kundu, S. (2007). Hydrophobic, hydrophilic, and charged amino acid networks within protein. Biophys. J. 93, 225-231.
165. Khor, S. (2018). Folding with a protein's native shortcut network. Proteins: Struc. Func. and Bioinform. 86, 924-934.
166. Chen, C., Li, L. & Xiao, Y. (2006). Identification of key residues in proteins by using their physical characters. Phys. Rev. E Stat. Nonlin. Soft. Matter. Phys. 73, 041926.
167. Vendruscolo, M., Dokholyan, N. V., Paci, E. & Karplus, M. (2002). Small-world view of the amino acids that play a key role in protein folding. Phys. Rev. E Stat. Nonlin. Soft. Matter. Phys. 65, 061910.
168. Batagelj, V. & Mrvar, A. (2001). Pajek—analysis and visualization of large networks. International Symposium on Graph Drawing: Springer. p. 477-478.
169. de Nooy, W., Mrvar, A. & Batageli, V. (2011). Exploratory Social Network Analysis. 2nd Ed ed: Cambridge Univesity Press.
170. Crucitti, P., Latora, V., Marchiori, M. & Rapisarda, A. (2004). Error and attack tolerance of complex networks. Physica A 340, 388-394.
171. Takes, F. W. & Kosters, W. A. (2011). Determining the diameter of small world networks. Proceedings of the 20th ACM international conference on Information and knowledge management, p. 1191-1196.

172. Steward, A., McDowell, G. S. & Clarke, J. (2009). Topology is the Principal Determinant in the Folding of a Complex All-alpha Greek Key Death Domain from Human FADD. J Mol Biol 389, 425-437.
173. Shindyalov, I. N. & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. 11, 739-747.
174. Collaborative, C. P. (1994). The CCP4 suite: programs for protein crystallography. Acta Crystallogr. D Biol. Crystallogr. 50, 760.

APPENDIX A

DETERMINATION OF EVOLUTIONARY CONSERVED INTERACTIONS

(studies designed and conducted by Lesley Greene with the computational programming assistance of Joshua Pothen)

28 proteins were selected for the network studies. Nine proteins with PDB codes 1Q5Y:A, 1RIS, 1UOS:A, 2ACY, 1URN:A, 1B7F:A, 1GH8:A, 1RKJ:A, and 2AW0 were selected to study the α/β plait superfamily. Similarly, ten proteins with PDB codes 1TIT, 1WIT, 2VAA:B, 3CD4, 1CQK:A, 1TEN, 1G84:A, 1HE7:A, 1TLK and 1HNG:A were chosen for the immunoglobulin superfamily, and nine proteins with PDB codes 1E3Y:A, 1DDF, 1D2Z:A, 1N3K:A, 1C15:A, 1UCP:A, 1DGN:A, 1A1W, and 3CRD were chosen for the death domain superfamily. Proteins within each superfamily were then structurally aligned. The combinatorial extension-Monte Carlo program [173] was used to generate the structure-based sequence alignment, which was then edited by hand to better align the residues based on a visual inspection of side chain orientation. These proteins were selected to maximize sequence and functional diversity within each superfamily. The average % sequence identity for the α/β-plaits, Igs and death domains is 9.5, 10 and 11.7%, respectively. The average R.M.S.D for the α/β-plaits, Igs and death domains is 3.1Å, 2.9Å and 3.1Å, respectively. These calculations were conducted with the combinatorial extension program using the online server [173] .

All heavy atom contact files were generated for each protein using the program Contact (CCP4) [158, 174]. For the α/β-plait and immunoglobulin superfamilies, only interactions of 6Å or less were calculated since the proteins are mainly composed of β-sheets which pack close together or α-helices and a β-sheet which are also viewed to pack close in space. For the death domains, all interactions within 7Å were calculated, since the structures contain six α-helices which occupy a large volume of space and appear to increase the diameter of the protein in comparison to the other two structures. From these contact files all pairs of residues involved in long-range interactions, where the interacting residues were separated by at least nine other residues in the primary structure, were extracted for subsequent studies. For qualification, nine residue separation means there are nine residues between the two making contact. In other words, the computer program will consider a contact between residues 1 and 10, but not between residues 1 and 9. The use of a minimum of nine residue separation helps further ensure that local interactions within the same β-strand or α-helix is avoided based on the periodicity of side chain orientation in the β-stand and the number of residues per turn in a conventional α-helix.

The evolutionary conserved interaction networks within these proteins were identified with the computer programming assistance of Joshua Pothen in accordance with the following procedure. A computer program was developed to generate the conserved network by combining the structure-based sequence alignment and calculated long-range interactions in three methodes. In the Direct method (D), the initial conserved network was generated by examining the protein residues in each line of the alignment. The algorithm then determines if these residues form contacts with other equivalently aligned residues using the calculated long-range interactions. If so, then the contacts amongst the proteins are said to be structurally and directly equivalent and thus considered part of the conserved network. A Toggle Method (T) was then applied, which takes into account the potential small variations in secondary structure stabilization and naturally occurring structural variability [96]. When a residue is checked for making a contact with another residue, the program examines what secondary structural element that second residue is a part of in the protein. For instance, if it is part of a β-sheet, then because of the alternating positions of side chains within β-strands, the first residue is checked for potential contacts with amino acids that are two residues behind and ahead of that second residue. If either of these contacts are made, they are considered structurally equivalent to the interaction between the first residue and the second residue in other proteins. Similarly, if the second residue is part of a turn, it is checked for residues that are one residue behind and ahead of it, and if it is part of an α-helix, it is also checked for residues that are one, three and four residues behind and ahead of it.

In the third method called Toggle with no acidic or basic residues (TN), any contact within the conserved network was removed if arginine, lysine, aspartic acid and/or glutamic acid (R,E,D,K) were present in one or more of the positions. This refinement allows the conserved network to contain mainly nonpolar interactions.

APPENDIX B

PROTOCOL OF CHARMM

The processor to perform molecular dynamic simulation using CHARMM is as follows:

1. Download the protein pdb-ID from the protein data bank and transfer it to ODU turing cluster where the CHARMM is located.
2. Perform minimization to minimize the potential energy of the system to the lowest possible point. Edit *protein_gen.inp* file and run it.

Note: Since most of the input files have the “go to” statement and it will crash running with CHARMM, we need to run them with the stream file. So, insert the name of the input file into the stream file and run it with the following command:

```
Charmmq    protein_stream.inp   1   protein_gen.out
```

As a result, we will have *protein_min.crd* that shows the coordinate of all the atoms after minimization.

3. Perform solvation by placing the protein in the specific water box based on protein’s size and shape. Edit *protein_gens.inp* file and run it with the stream file as follow;

```
Charmmq    protein_stream.inp   1   protein_gens.out
```

As a result, we will have *protein_solve.crd* that shows the coordinate of all the atoms after adding a box of water.

4. Perform neutralization of the system by randomly replacing some of the water molecules with ions. Edit *protein_neutral.inp* file and run it with the stream file as follow:

```
Charmmq    protein_stream.inp   1   protein_neutral.out
```

5. Perform equilibration and then MD simulation at several different temperatures. The highest temperature should be enough to unfold protein. Edit *protein_e_temp.inp* file and run it with stream file as follow:

   ```
   Charmmq_mpi    protein_stream.inp   32   protein_e_temp.out
   ```

6. Run MD simulation. Edit *protein_d_temp.inp* file and run it with stream file as follow:

   ```
   Charmmq_mpi    protein_stream.inp   32   protein_d_temp.out
   ```

Note: 32 is the number of the processors that can be changed. When we have more than one processor, the command "*charmmq*" change to "*charmmq_mpi*".

7. Perform merging step to merge all unfolding trajectories as one file. Edit *ptotein_merge.inp* file and run it with stream file as follow:

   ```
   Charmmq    protein_stream.inp   1  protein_merge_temp.out
   ```

   This step should be performed for each temperature separately.

<u>The processor to analyze the MD trajectories using CHARMM is as follows:</u>

1. Calculate the RMSD for all unfolded trajectories. Edit *protein_rmsd.inp* and run it with a stream file as follow:

   ```
   Charmmq    protein_stream.inp   1  protein_rmsd_temp.out
   ```

   As a result we will have *protein_temp.rmsd* that can be opened as a graph with visualizing program *xmgr*.

2. Calculate the residue-residue distances for all the residues as the during the unfolding process as the temperature increase. Edit *protein_traj_dmat.inp* and run it with the following command:

   ```
   Charmmq   protein_traj_dmat.inp   1   protein_traj_dmat.out
   ```

As a result, we will have *protein_temp_traj.dmat* that shows the residue-residue distances.

3. Calculate the atom-atom distances for all the atoms during the unfolding process as the temperature increase. Edit *protein_traj_atom.inp* and run it with the following command:

   ```
   Charmmq   protein_traj_atom.inp   1   protein_traj_atom.out
   ```

   As a result, we will have *protein_temp_traj_atom.dmat* that shows the atom-atom distances.

4. Calculate the distance between two atoms of two residues for the selected contacts during the unfolding process as the temperature increase. Edit *atom_min_dist3.f* file and run it with the following command:

   ```
   F95 atom_min_dist3.f
   ./a.out
   ```

   As a result, we will have *a.out* that shows the distance between two atoms of two residues for the selected contacts.

5. Find the most persistent contacts as protein unfold. To make a *.dmatp* file for a selected contact, run *dmatzero.awk* using following command:

   ```
   Awk  -f   dmatzero.awk   protein_temp.dmat > protein_temp.dmatp
   ```

   Make a *.dmatp* file also for the native contacts by the following command:

   ```
   Awk  -f  dmatproj_lr.awk   protein_native.dmat > protein_native.dmatp
   ```

   Convert zero to one for the selected contacts in the *protein_temp.dmatp* file, edit *protein_cfracvstime.inp* and then run the following command ;

   ```
   Charmmq   protein_stream.inp   1   protein_cfracvstime.out
   ```

   As a result, we will have *.traj* file that can be opened as a graph with *xmgr* program.

APPENDIX C

PROTOCOL OF CNS

The processor to perform simulated annealing using CNS is as follows:

1. Go to the Turing cluster at ODU and type the following command to access to CNS:

   Enable_lmod

   Module load cns

   cns_web

2. Make a file of amino acids sequence and name it *protein_strat.seq* to use it as an input file. For example:

```
MET ASP PRO PHE LEU VAL LEU LEU HIS SER VAL SER SER
SER LEU SER SER SER GLU LEU THR GLU LEU LYS TYR LEU
CYS LEU GLY ARG VAL GLY LYS ARG LYS LEU GLU ARG VAL
GLN SER GLY LEU ASP LEU PHE SER MET LEU LEU GLU GLN
ASN ASP LEU GLU PRO GLY HIS THR GLU LEU LEU ARG GLU
LEU LEU ALA SER LEU ARG ARG HIS ASP LEU LEU ARG ARG
VAL ASP ASP PHE GLU LEU GLU HIS HIS HIS HIS HIS HIS
```

3. In the CNSsolve web under input files (general), Open and edit *generate_seq.inp* and then run it with the following command:

   cns_solve <generate_sequence.inp> generate_seq.out

   As a result, we will have *generate_seq.mtf* that is protein topology file.

4. In the CNSsolve web under input files (NMR), Open and edit *generate_extended.inp* and then run it with the following command:

   cns_solve <generate_extended.inp> generate_extended.out

   As a result, we will have *protein_extended.pdb* that is the extended structure (linear form) of protein.

5. In the CNSsolve web under input files (NMR), Open and edit *anneal.inp* and then run it with the following command:

```
cns_solve <anneal.inp> anneal.out
```

As a result, we will have simulated structures of the protein.

APPENDIX D

PROTOCOL OF FRAGMENTATION AND DIAMETER PROGRAMS

Fragmentation Program

The *fragmentation* program is written in Python language. The program designed to attack the protein network and remove the high BC residues in advance and then continue to remove the randomly selected residues until the network completely collapse. This program can also be performed without a set of high BC residues to do fragmentation completely randomly.

<u>The protocol for the Fragmentation program is as follows:</u>

1. Prepare a set of long-range contacts and a set of high BC residues in a text format as input files. The input files should be placed in the same drive as the program. For example:

Long-range contacts

```
A  B
2  28
4  47
7  56
12  34
12  85
 .
```

High BC residues

```
 5
14
28
71
.
.
```

2. Insert the name of the input files in the script at sections shown in bold.

```
test_file=open("Long-range contacts.net","ab+")
num_lines = file_len("Long-range contacts.net.net")
text_in_line=test_file.readlines()
test_file.close()
HighBC_file=open("HighBC_residues.net","ab+")
HighBC_lines=HighBC_file.readlines()
l=len(HighBC_lines)
HighBC_file.close()
Output= open("Output_fragmentation.txt","w+")
```

3. Run the script.
4. The result will appear as a text file called *Output_fragmentation* in the same drive as the program located. The output consists of the size of the largest cluster (S) shown as a fraction of nodes of the cluster with respect to the total system size, when a fraction (*f*) of the nodes are removed randomly in an attack mode.

Diameter Program

The *Diameter* program is written in Python language. The program is designed to remove a set of high BC residues from the original protein network, and then start over and remove 1000 different sets of the randomly selected residues from the original protein network. The network diameter is calculated after each removal.

<u>The protocol for the Diameter program is as follows:</u>

1. Prepare a set of long-range contacts and a set of high BC residues in a text format as input files. The input files should be placed in the same drive as the program. For exmaple:

Long-range contacts

```
A  B
2  28
4  47
7  56
12  34
12  85
 .
 .
```

High BC residues

```
 5
14
28
71
.
.
```

2. Insert the name of the input files in the script at sections shown in bold.

```
test_file=open("long-range_contacts.net","ab+")
num_lines = file_len("long-range_contacts.net.net")
text_in_line=test_file.readlines()
test_file.close()
HighBC_file=open("HighBC_residuce.net","ab+")
HighBC_lines=HighBC_file.readlines()
Output= open("Output_diameter.txt","w+")
```

3. Run the script.
4. The result will appear as a text file called *Output_diameter* in the same drive as the program located. The output consists of the diameter (average of the shortest paths) of the protein network after removing each node.

APPENDIX E

LIST OF THE CONSERVED CONTACTS

These are the results from the study conducted and described in Appendix A.

1A1W-Toggle				1A1W-TN		1A1W-Direct	
L 5	L 43	L 20	S 41	L 7	F 46	L 49	L 63
L 5	F 46	L 20	G 42	L 8	F 46	L 63	V 79
V 6	L 43	L 20	L 45	L 8	L 63	R 64	V 79
L 7	F 46	T 21	S 41	H 9	F 46	R 64	D 80
L 7	L 75	E 22	L 66	V 11	L 63		
L 7	R 78	L 23	L 45	S 12	F 46		
L 8	G 42	L 23	L 63	L 20	L 45		
L 8	L 43	L 23	L 66	L 23	L 45		
L 8	F 46	L 23	L 67	L 23	L 63		
L 8	L 63	K 24	L 45	L 23	L 67		
H 9	L 43	K 24	L 66	C 27	L 49		
H 9	F 46	C 27	L 45	C 27	L 63		
S 10	L 43	C 27	L 49	L 45	L 63		
S 10	L 75	C 27	L 63	F 46	L 63		
S 10	R 78	C 27	L 66	L 49	L 63		
V 11	L 43	L 36	M 48	T 60	V 79		
V 11	L 63	E 37	M 48	L 63	V 79		
V 11	L 75	R 38	M 48	L 67	V 79		
V 11	R 78	L 45	L 63				
S 12	G 42	F 46	L 63				
S 12	L 43	H 59	F 82				
S 12	F 46	T 60	V 79				
S 12	L 75	T 60	D 80				
S 14	L 75	T 60	F 82				
S 14	R 78	E 61	V 79				
L 15	G 42	E 61	V 79				
L 15	L 67	L 63	V 79				
L 15	L 75	R 64	V 79				
S 17	S 41	R 64	D 80				
E 19	L 67	L 67	V 79				

1DGN-Toggle				1DGN-TN		1DGN-Direct	
K 9	I 58	N 23	L 57	F 13	F 72	V 61	F 72
K 9	L 83	A 24	H 75	I 14	I 58	F 72	M 87
K 9	K 86	L 25	L 57	T 21	L 57	I 73	M 87
R 10	I 58	L 25	V 61	T 21	L 76	I 73	G 88
R 11	I 58	L 25	F 72	I 22	L 57		
I 12	L 83	L 25	H 75	N 23	L 57		
I 12	K 86	L 25	L 76	L 25	L 57		
F 13	R 55	L 26	L 57	L 25	V 61		
F 13	A 54	D 27	L 57	L 25	F 72		
F 13	F 72	L 29	L 57	L 25	L 76		
F 13	L 83	L 29	V 61	L 26	L 57		
I 14	A 54	L 29	F 72	L 29	V 61		
I 14	R 55	L 29	H 75	L 29	F 72		
I 14	I 58	D 39	L 60	L 57	F 72		
I 14	L 83	M 40	L 60	I 58	F 72		
H 15	A 54	N 41	L 60	V 61	F 72		
H 15	L 83	R 44	L 60	T 62	F 72		
S 16	A 54	L 57	F 72	F 72	M 87		
S 16	L 83	I 58	F 72	I 73	M 87		
V 17	A 54	T 62	F 72	L 76	M 87		
V 17	L 76	F 72	M 87	C 77	M 87		
V 17	L 83	I 73	M 87				
G 18	L 83	I 73	G 88				
A 19	K 53	I 73	H 90				
T 21	L 57	K 74	M 87				
T 21	H 75	K 74	H 90				
T 21	L 76	L 76	M 87				
I 22	K 53	C 77	M 87				
I 22	A 54	C 77	G 88				
I 22	L 57	C 77	H 90				
I 22	L 76						

1E3Y-Toggle				1E3Y-TN		1E3Y-Direct	
A 99	L 161	W 112	V 141	A 99	L 161	W 148	L 161
A 99	L 176	W 112	S 144	A 100	L 145	L 161	V 177
A 100	L 145	W 112	L 145	V 103	L 161	V 162	V 177
F 101	V 141	R 113	R 140	C 105	L 145	V 162	Q 178
F 101	R 142	R 113	S 144	W 112	S 144		
F 101	L 145	L 115	R 140	L 115	W 148		
N 102	R 142	L 115	V 141	L 115	L 161		
N 102	L 145	L 115	S 144	L 115	L 165		
N 102	L 172	L 115	W 148	A 116	S 144		
N 102	L 176	L 115	L 161	L 145	L 161		
V 103	V 141	L 115	A 164	W 148	L 161		
V 103	R 146	L 115	L 165	V 158	V 177		
V 103	L 161	L 115	C 168	L 161	V 177		
V 103	N 171	A 116	S 144	V 162	V 177		
V 103	V 173	A 116	L 145				
V 103	L 176	K 125	S 144				
I 104	V 141	K 125	I 147				
I 104	R 146	L 145	L 161				
I 104	N 171	L 145	S 167				
I 104	V 173	W 148	L 161				
C 105	V 141	V 158	N 171				
C 105	R 142	V 158	V 177				
C 105	E 143	V 158	Q 187				
C 105	L 145	L 161	M 170				
D 106	V 141	L 161	V 177				
D 106	L 145	L 161	V 177				
N 107	N 171	L 161	Q 178				
N 107	V 173	V 162	V 177				
G 109	V 141	V 162	Q 178				
G 109	E 143	V 162	N 188				
D 111	V 141	G 163	M 170				
D 111	L 161	G 163	V 177				
D 111	L 165	G 163	Q 178				

1RIS-Toggle						1RIS-TN		1RIS-Direct	
R 2	V 65	V 9	G 58	E 31	L 75	Y 4	Y 63	Y 4	Y 63
R 3	Q 64	V 9	Y 59	N 32	L 75	Y 4	V 65	Y 4	Q 64
R 3	V 65	V 9	F 60	Y 33	L 75	V 6	Y 63	Y 4	V 65
Y 4	W 62	V 9	L 61	Y 33	E 78	V 6	V 65	E 5	W 62
Y 4	Y 63	V 9	R 86	Y 33	L 79	V 6	V 88	E 5	Y 63
Y 4	Q 64	V 9	R 87	A 35	V 65	N 7	Y 63	V 6	L 61
Y 4	V 65	V 9	V 88	A 35	E 66	N 7	V 88	V 6	W 62
E 5	W 62	V 9	M 89	R 36	Q 64	I 8	I 26	V 6	Y 63
E 5	Y 63	L 10	I 26	R 36	V 65	I 8	Y 63	V 6	M 89
E 5	Q 64	L 10	G 58	R 36	E 66	I 8	V 88	N 7	F 60
E 5	V 65	L 10	Y 59	V 37	Y 63	V 9	V 88	N 7	L 61
E 5	M 89	L 10	F 60	V 37	Q 64	L 10	I 26	N 7	R 87
V 6	L 61	L 10	L 61	V 37	V 65	I 26	Y 63	N 7	V 88
V 6	W 62	L 10	R 86	V 37	E 66	I 26	L 75	N 7	M 89
V 6	Y 63	L 10	R 87	E 38	W 62	A 29	L 75	I 8	Y 59
V 6	Q 64	N 11	G 58	E 38	Y 63	L 30	Y 63	I 8	F 60
V 6	V 65	N 11	Y 59	E 38	Q 64	L 30	V 65	I 8	R 86
V 6	V 88	N 11	R 86	E 38	V 65	L 30	L 75	I 8	R 87
V 6	M 89	N 11	R 87	E 38	E 66	N 32	L 75	V 9	Y 59
N 7	F 60	P 12	G 58	K 39	W 62	Y 33	L 75	V 9	R 86
N 7	L 61	P 12	Y 59	K 39	Y 63	A 35	V 65	V 9	R 87
N 7	W 62	P 12	R 86	K 39	Q 64	V 37	Y 63	A 29	L 79
N 7	Y 63	I 25	E 78	K 39	V 65	V 37	V 65	L 30	Y 63
N 7	R 87	I 25	L 79	V 40	L 61	V 40	Y 63	V 37	Y 63
N 7	V 88	I 26	Y 63	V 40	W 62			V 37	Q 64
N 7	M 89	I 26	L 75	V 40	Y 63			V 37	V 65
I 8	I 26	I 26	L 79	V 40	Q 64			E 38	Y 63
I 8	G 58	R 28	L 75	E 41	W 62			E 38	Q 64
I 8	Y 59	R 28	E 78	E 41	Y 63			K 39	W 62
I 8	F 60	A 29	L 75	E 41	Q 64			K 39	Y 63
I 8	L 61	A 29	E 78	E 42	L 61			K 39	Q 64
I 8	W 62	A 29	L 79	E 42	W 62			V 40	L 61
I 8	Y 63	L 30	Y 63	L 43	W 62			V 40	W 62
I 8	R 86	L 30	V 65	L 43	Y 63			E 41	W 62
I 8	R 87	L 30	L 75	G 44	L 61				
I 8	V 88	L 30	E 78						
I 8	M 89	L 30	L 79						

2ACY-Toggle						2ACY-TN		2ACY-Direct	
I 7	L 51	I 13	Q 48	L 33	W 64	I 7	L 51	V 9	L 51
I 7	Q 52	I 13	G 49	G 34	Q 52	I 7	G 53	V 9	Q 52
I 7	G 53	I 13	D 76	G 34	G 53	S 8	L 51	V 9	G 53
S 8	Q 50	I 13	R 77	G 34	P 54	V 9	L 51	D 10	Q 50
S 8	L 51	I 13	A 78	L 35	L 51	V 9	G 53	D 10	L 51
S 8	Q 52	I 13	S 79	L 35	Q 52	Y 11	L 51	Y 11	G 49
S 8	G 53	F 14	T 46	L 35	G 53	Y 11	A 78	Y 11	Q 50
V 9	G 49	F 14	V 47	L 35	P 54	I 13	T 26	Y 11	L 51
V 9	Q 50	F 14	D 76	V 36	Q 50	I 13	A 78	Y 11	S 79
V 9	L 51	F 14	R 77	V 36	L 51	F 14	A 78	E 12	Q 48
V 9	Q 52	F 14	A 78	V 36	Q 52	Y 25	M 61	E 12	G 49
V 9	G 53	F 14	S 79	V 36	G 53	T 26	L 51	E 12	R 77
V 9	S 79	G 15	T 46	G 37	G 49	T 26	M 61	E 12	A 78
D 10	Q 48	G 15	D 76	G 37	Q 50	A 28	M 61	E 12	S 79
D 10	G 49	G 15	R 77	G 37	L 51	G 30	L 51	I 13	V 47
D 10	Q 50	K 16	D 76	G 37	Q 52	G 30	M 61	I 13	Q 48
D 10	L 51	V 17	T 46	W 38	Q 50	L 33	M 61	I 13	D 76
D 10	S 79	V 17	V 47	W 38	L 51	G 34	G 53	I 13	R 77
Y 11	V 47	V 17	D 76	W 38	Q 52	L 35	L 51	F 14	V 47
Y 11	Q 48	V 17	R 77	V 39	G 49	V 36	L 51	F 14	D 76
Y 11	G 49	Y 25	M 61	V 39	Q 50	V 36	G 53	F 14	R 77
Y 11	Q 50	Y 25	W 64	Q 40	Q 50	G 37	L 51	E 29	L 65
Y 11	L 51	T 26	L 51			W 38	L 51	G 30	L 51
Y 11	A 78	T 26	M 61					V 36	L 51
Y 11	S 79	T 26	L 65					V 36	Q 52
E 12	T 46	A 28	M 61					V 36	G 53
E 12	V 47	A 28	W 64					G 37	L 51
E 12	Q 48	E 29	M 61					G 37	Q 52
E 12	G 49	E 29	W 64					W 38	Q 50
E 12	R 77	E 29	L 65					W 38	L 51
E 12	A 78	G 30	L 51					W 38	Q 52
E 12	S 79	G 30	M 61					V 39	G 49
I 13	T 26	G 30	W 64					V 39	Q 50
I 13	T 46	K 31	M 61					Q 40	Q 50
I 13	V 47	L 33	M 61						

2AW0-Toggle						2AW0-TN		2AW0-Direct	
Q 3	V 45	I 9	N 40	S 26	V 45	T 5	V 45	E 4	V 45
Q 3	E 46	I 9	S 41	S 26	L 57	V 6	V 45	E 4	E 46
Q 3	Y 47	I 9	N 42	K 28	L 57	V 6	Y 47	E 4	Y 47
E 4	T 44	I 9	G 43	K 28	A 60	I 7	V 45	T 5	T 44
E 4	V 45	I 9	T 44	K 28	I 61	I 7	A 68	T 5	V 45
E 4	E 46	I 9	V 45	P 29	Y 47	N 8	A 68	V 6	G 43
E 4	Y 47	I 9	F 66	P 29	D 48	I 9	A 68	V 6	T 44
T 5	G 43	I 9	D 67	P 29	L 57	M 12	I 21	V 6	V 45
T 5	T 44	I 9	A 68	G 30	E 46	I 21	V 45	V 6	T 69
T 5	V 45	I 9	T 69	G 30	Y 47	I 25	V 45	I 7	N 42
T 5	E 46	D 10	N 40	G 30	D 48	I 25	L 57	I 7	G 43
T 5	Y 47	D 10	S 41	V 31	V 45	S 26	V 45	I 7	D 67
V 6	N 42	D 10	N 42	V 31	E 46	S 26	L 57	I 7	A 68
V 6	G 43	D 10	F 66	V 31	Y 47	P 29	Y 47	I 7	T 69
V 6	T 44	D 10	D 67	V 31	D 48	P 29	L 57	N 8	S 41
V 6	V 45	D 10	A 68	K 32	V 45	V 31	V 45	N 8	N 42
V 6	Y 47	D 10	T 69	K 32	E 46	V 31	Y 47	N 8	F 66
V 6	T 69	G 11	S 41	K 32	Y 47	S 33	V 45	N 8	D 67
I 7	N 40	G 11	F 66	K 32	D 48	S 33	Y 47	I 9	S 41
I 7	S 41	G 11	D 67	S 33	T 44	I 34	V 45	I 9	F 66
I 7	N 42	M 12	I 21	S 33	V 45	V 36	V 45	I 9	D 67
I 7	G 43	M 12	N 40	S 33	E 46			V 24	I 61
I 7	T 44	M 12	S 41	S 33	Y 47			I 25	V 45
I 7	V 45	M 12	N 42	I 34	G 43			V 31	V 45
I 7	D 67	M 12	G 43	I 34	T 44			V 31	E 46
I 7	A 68	M 12	F 66	I 34	V 45			V 31	Y 47
I 7	T 69	M 12	D 67	I 34	E 46			K 32	V 45
N 8	N 40	I 21	V 45	R 35	T 44			K 32	E 46
N 8	S 41	I 21	I 61	R 35	V 45			S 33	T 44
N 8	N 42	V 24	A 60	R 35	E 46			S 33	V 45
N 8	G 43	V 24	I 61	V 36	V 45			S 33	E 46
N 8	F 66	I 25	V 45					I 34	G 43
N 8	D 67	I 25	L 57					I 34	T 44
N 8	A 68	I 25	A 60					R 35	T 44
N 8	T 69	I 25	I 61						

1TEN-Toggle				1TEN-TN		1TEN-Direct	
D 816	L 863	I 833	L 873	T 817	I 860	T 817	L 863
T 817	I 860	I 833	I 874	T 818	I 860	A 819	I 860
T 817	G 861	I 833	S 875	A 819	Y 858	L 820	S 859
T 817	L 863	I 833	R 876	A 819	I 860	L 820	I 860
T 818	S 859	E 834	S 872	I 821	N 856	I 821	Y 858
T 818	I 860	E 834	L 873	I 821	Y 858	T 822	Q 857
T 818	G 861	E 834	I 874	I 821	I 860	T 822	Y 858
T 818	L 863	E 834	S 875	I 821	L 873	W 823	N 856
A 819	Y 858	L 835	I 849	W 823	N 856	W 823	L 873
A 819	S 859	L 835	Y 858	W 823	Y 858	I 833	S 875
A 819	I 860	L 835	V 871	W 823	L 873	I 833	R 876
A 819	G 861	L 835	S 872	F 824	N 856	E 834	I 874
A 819	L 863	L 835	L 873	P 826	S 875	E 834	S 875
L 820	Q 857	L 835	I 874	G 832	S 875	L 835	I 874
L 820	Y 858	T 836	E 870	I 833	L 873	T 836	S 872
L 820	S 859	T 836	V 871	I 833	S 875	T 836	L 873
L 820	I 860	T 836	S 872	L 835	Y 858	T 836	I 874
I 821	N 856	T 836	L 873	L 835	L 873	Y 837	I 849
I 821	Q 857	T 836	I 874	Y 837	Y 858	Y 837	Y 858
I 821	Y 858	Y 837	I 849			Y 837	S 872
I 821	S 859	Y 837	Y 858			G 838	E 870
I 821	I 860	Y 837	Y 869			G 838	V 871
I 821	L 873	Y 837	E 870			I 839	Y 869
T 822	N 856	Y 837	V 871			I 839	E 870
T 822	Q 857	Y 837	S 872			E 870	T 888
T 822	Y 858	G 838	E 870			V 871	E 887
W 823	E 855	G 838	V 871			V 871	T 888
W 823	N 856	I 839	Y 869				
W 823	Q 857	I 839	E 870				
W 823	Y 858	K 840	E 870				
W 823	L 873	E 868	T 888				
F 824	N 856	Y 869	E 887				
K 825	N 856	Y 869	T 888				
P 826	S 875	E 870	T 888				
D 831	R 876	V 871	E 887				
G 832	I 874	V 871	T 888				
G 832	S 875						

1TIT-Toggle				1TIT-TN		1TIT-Direct	
E 17	N 62	H 31	Q 74	T 18	L 60	E 17	N 62
T 18	I 59	H 31	A 75	A 19	L 58	A 19	L 60
T 18	L 60	G 32	F 73	A 19	L 60	H 20	I 59
T 18	H 61	G 32	Q 74	H 20	L 60	H 20	L 60
T 18	N 62	G 32	A 75	F 21	H 56	F 21	L 58
A 19	L 58	G 32	A 76	F 21	L 58	E 22	I 57
A 19	I 59	Q 33	S 72	F 21	L 60	E 22	L 58
A 19	L 60	Q 33	F 73	I 23	H 56	I 23	H 56
A 19	H 61	Q 33	Q 74	I 23	F 73	I 23	F 73
H 20	I 57	Q 33	A 75	L 25	H 56	V 30	A 75
H 20	L 58	W 34	I 49	L 25	A 75	V 30	A 76
H 20	I 59	W 34	L 58	V 30	A 75	H 31	Q 74
H 20	L 60	W 34	V 71	H 31	A 75	H 31	A 75
F 21	H 56	W 34	S 72	G 32	F 73	G 32	Q 74
F 21	I 57	W 34	F 73	G 32	A 75	Q 33	S 72
F 21	L 58	W 34	Q 74	W 34	L 58	Q 33	F 73
F 21	I 59	K 35	E 70	W 34	F 73	Q 33	Q 74
F 21	L 60	K 35	V 71			W 34	I 49
E 22	H 56	K 35	S 72			W 34	L 58
E 22	I 57	K 35	F 73			W 34	S 72
E 22	L 58	K 35	Q 74			K 35	E 70
I 23	K 55	L 36	G 69			K 35	V 71
I 23	H 56	L 36	E 70			L 36	G 69
I 23	I 57	L 36	V 71			L 36	E 70
I 23	F 73	L 36	S 72			E 70	A 82
E 24	H 56	G 69	A 82			V 71	A 81
L 25	H 56	E 70	A 82			V 71	A 82
L 25	A 75	V 71	A 81				
V 30	A 75	V 71	A 82				
V 30	A 76	S 72	A 82				
H 31	F 73						

1TLK-Toggle				1TLK-TN		1TLK-Direct	
G 56	S 103	P 77	V 132	S 57	I 102	S 57	V 105
G 56	V 105	E 78	K 133	A 58	I 102	A 59	I 102
S 57	I 102	E 79	A 134	A 59	L 100	R 60	T 101
S 57	S 103	V 80	C 135	A 59	I 102	R 60	I 102
S 57	V 105	V 81	K 136	F 61	C 98	F 61	L 100
A 58	T 101	V 82	A 137	F 61	L 100	D 62	S 99
A 58	I 102	V 83	V 138	F 61	I 102	D 62	L 100
A 58	S 103	M 84	T 139	C 63	C 98	C 63	C 98
A 58	V 104	M 85	C 140	C 63	C 115	C 63	C 115
A 59	L 105	M 86	K 141	V 65	C 98	P 71	A 117
A 59	T 106	M 87	A 142	V 65	C 115	P 71	V 118
A 59	I 107	W 88	I 143	V 65	A 117	E 72	K 116
A 59	S 108	W 89	L 144	P 71	A 117	E 72	A 117
R 60	S 109	W 90	Y 145	V 73	C 115	V 73	K 116
R 60	L 110	W 91	T 146	V 73	A 117	M 74	T 114
R 60	T 111	W 92	C 147	M 74	A 117	M 74	C 115
R 60	I 112	W 93	K 148	W 75	L 100	M 74	K 116
F 61	C 113	F 94	K 149	W 75	C 115	W 75	I 90
F 61	S 114	F 95	Y 150			W 75	L 100
F 61	L 115	F 96	T 151			W 75	T 114
F 61	T 116	F 97	C 152			F 76	K 112
F 62	I 117	F 98	K 153			F 76	Y 113
D 63	C 118	K 99	A 154			K 77	A 111
D 64	S 119	K 100	K 155			K 77	K 112
D 65	L 120	K 101	Y 156			K 112	T 127
C 66	N 121	K 102	T 157			Y 113	C 126
C 67	C 122	D 103	K 158			Y 113	T 127
C 68	S 123	D 104	Y 159				
C 69	C 124	D 105	T 160				
K 70	C 125	D 106	K 161				
V 71	N 126	K 107	T 162				
V 72	C 127	Y 108	C 163				
V 73	C 128	Y 109	T 164				
V 74	A 129	T 110	T 165				
P 75	V 130	C 111	C 166				
P 76	A 131	C 112	T 167				

VITA

ZEINAB HARATIPOUR
zhara001@odu.edu
Department of Chemistry and Biochemistry
Old Dominion University
Norfolk, VA 23529

EDUCATION

Old Dominion University	**Norfolk, VA**
PhD, Chemistry	May 2020
Old Dominion University	**Norfolk, VA**
Master of Science, Chemistry	May 2018
Islamic Azad University	**Mashhad, Iran**
Master of Science, Physical Chemistry	May 2011
Islamic Azad University	**Mashhad, Iran**
Bachelor of Science, Chemistry	July 2007

PUBLICATIONS AND PRESENTATIONS

Z. Haratipour, H. Aldabagh, Y. Li, L.H. Greene, *Network Connectivity, Centrality and Fragmentation in the Greek-Key Protein Topology*. The protein journal, 2019, 38(5), 497-505.

Z. Haratipour, L.H. Greene, *Significance of evolutionary conserved long-range interactions in governing the topology and folding in proteins*. (to be submitted)

Z. Haratipour, J. Poutsma, L.H. Greene, *Role of the Conserved Long-Range Interaction Networks in the Structural Stability of Proteins.* (to be submitted)

M. R. Housaindokht, M. R. Bozorgmehr, H. Eshtiagh Hosseini, R. Jalal , A. Asoodeh, M. Saberi, **Z.Haratipour**, H. Monhemi, *Structural properties of the truncated and wild types of Taka-amylase: A molecular dynamics simulation and docking study,* Journal of Molecular Catalysis B: Enzymatic, 2013, vol. 95, 36– 40.

Z. Haratipour, J. Poutsma, L.H. Greene, *Conserved long-range interaction networks, energetics and the determinants of protein topology,* the 257th ACS National Meeting, Orlando, FL, March 32- April 4, 2019. (Oral presentation)

Z. Haratipour, L.H. Greene, *Characterization of Long-Range Interaction Networks in Proteins using Key Network Principles,* Graduate Achievement Day at Old Dominion University, Norfolk, VA, March 2018. (Poster presentation)

Z. Haratipour, L.H. Greene, *Characterization of long-range interaction networks in proteins using key network principles*, the 255th ACS National Meeting, New Orleans, LA, March 18-22, 2018. (Oral presentation)

Z. Haratipour, L.H. Greene, *The significance of the conserved long-ranged interaction networks in determining protein topology,* The 95th Virginia Academy of Sciences Annual Meeting, 16-18 May 2017, Virginia Commonwealth University, Richmond, VA. (Oral presentation)