

Old Dominion University

## ODU Digital Commons

---

Chemistry & Biochemistry Theses & Dissertations

Chemistry & Biochemistry

---

Summer 2012

# Identification of Persistent Long Range Interactions in G<sub>A</sub>95 and G<sub>B</sub>95 Through Thermal Unfolding Simulations

Milen Redai Tesfamariam  
*Old Dominion University*

Follow this and additional works at: [https://digitalcommons.odu.edu/chemistry\\_etds](https://digitalcommons.odu.edu/chemistry_etds)



Part of the [Amino Acids, Peptides, and Proteins Commons](#), [Bacteriology Commons](#), [Biochemistry Commons](#), [Computational Chemistry Commons](#), and the [Molecular Biology Commons](#)

---

### Recommended Citation

Tesfamariam, Milen R.. "Identification of Persistent Long Range Interactions in G<sub>A</sub>95 and G<sub>B</sub>95 Through Thermal Unfolding Simulations" (2012). Master of Science (MS), Thesis, Chemistry & Biochemistry, Old Dominion University, DOI: 10.25777/wqg9-ew61  
[https://digitalcommons.odu.edu/chemistry\\_etds/147](https://digitalcommons.odu.edu/chemistry_etds/147)

This Thesis is brought to you for free and open access by the Chemistry & Biochemistry at ODU Digital Commons. It has been accepted for inclusion in Chemistry & Biochemistry Theses & Dissertations by an authorized administrator of ODU Digital Commons. For more information, please contact [digitalcommons@odu.edu](mailto:digitalcommons@odu.edu).

**IDENTIFICATION OF PERSISTENT LONG RANGE INTERACTIONS IN G<sub>A</sub>95  
AND G<sub>B</sub>95 THROUGH THERMAL UNFOLDING SIMULATIONS**

by

Milen Redai Tesfamariam  
B.S. September 2007, University of Asmara, Eritrea

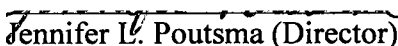
A Thesis Submitted to the Faculty of  
Old Dominion University in Partial Fulfillment of the  
Requirements for the Degree of

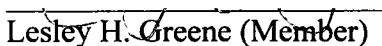
MASTER OF SCIENCE

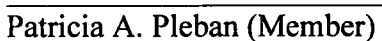
CHEMISTRY

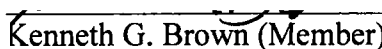
OLD DOMINION UNIVERSITY  
August 2012

Approved by:

  
Jennifer L. Poutsma (Director)

  
Lesley H. Greene (Member)

  
Patricia A. Pleban (Member)

  
Kenneth G. Brown (Member)

## ABSTRACT

### IDENTIFICATION OF PERSISTENT LONG RANGE INTERACTIONS IN G<sub>A</sub>95 AND G<sub>B</sub>95 THROUGH THERMAL UNFOLDING SIMULATIONS

Milen Redai Tesfamariam  
Old Dominion University, 2012  
Director: Dr. Jennifer L. Poutsma

For over five decades, different experiments have been performed to research how proteins attain their native three dimensional structures. However, the folding problem continues to be a puzzle in modern science. The design of two proteins that have maximal sequence identity but different folds and functions is one method that is being used to study the relationship between protein structure and amino acid sequence. In particular, mutant proteins of *Streptococcus* protein G, G<sub>A</sub> and G<sub>B</sub>, have 95% sequence identity and a 3 $\alpha$  helix fold and 4 $\beta$ / $\alpha$  fold, respectively. Molecular dynamics simulations of G<sub>A</sub>95 and G<sub>B</sub>95 at high temperatures were used to unfold the proteins and observe how the long range interactions between the amino acids change during the unfolding process. Comparison of the persistent interactions with the locations of the non-identical residues will provide further insight into how these amino acids encode the protein fold.

Three independent simulations of each protein were performed at 550 K. For each trajectory, the long range contact distances versus time were calculated. The most important long range interactions in maintaining the 3 $\alpha$ -helical fold of G<sub>A</sub>95 are  $\alpha$ 1/  $\alpha$ 2 and  $\alpha$ 2/ $\alpha$ 3 interactions, which include Ala16-Ile30, Ile17-Ile30, Leu20-Ile30, Tyr29-Leu45, Tyr29-Ile49, and Ile33-Val42. Four of these interactions have one of the non-identical residues, Leu20, Ile30, and Leu45. Residues 20 and 30 are found to be more important than residue 45 in G<sub>A</sub>95 because they form a strong long range interaction

between  $\alpha 1$  and  $\alpha 2$  helices, which may explain their stability during the unfolding simulation.

In  $G_B95$ , interactions between  $\beta 1/\beta 2$ ,  $\beta 3/\beta 4$ ,  $\beta 1/\alpha$ , and  $\alpha/\beta 4$  are the most important ones in determining the  $4\beta+\alpha$  fold. These include Thr1-Ala20, Tyr3-Ala20, Thr44-Thr53, Tyr45-Phe52, Val42-Thr55, Gly41-Thr55, Leu5-Phe30, Tyr3-Ala26, and Phe30-Phe52. Five interactions have one of the non-identical residues. Of all these interactions, the most important interaction is Tyr45-Phe52, which was observed to have a significant number of contacts at the end of the simulation. Hence, residue 45 is the most important non-identical residue in  $G_B95$ .

Copyright, 2012, by Milen Redai Tesfamariam, All Rights Reserved.

*This thesis is dedicated to my father and mother who loved and supported me to be where I am now.*

## ACKNOWLEDGMENTS

Thanks to my advisor, Dr. Jennifer Poutsma, who has guided this research and helped in every way during these graduate studies. Her patience and instruction using CHARMM were well received. Thanks to Dr. Lesley Greene who focused my research on the computational aspects of protein folding. To Old Dominion University, Department of Chemistry, thanks for providing me teaching assistantship. Thanks to Valerie DeCosta, Graduate Program Assistant, for her excellent support. Thanks to Alicia Herr, Department Manager, for her careful supervision of the teaching schedule. Tammy Subotich, Stockroom Manager, without whom the labs could not run. Thanks to Janice Moore, Administrative Assistant, for myriad activities.

Great and lifelong thanks to my parents, Redai and Emuna, who have always been by my side, giving me advice and support. Their encouragement during tough times has ensured my success. I also thank my brother, Bereket, who always checked how I was doing and cheered me on.

Most importantly, I thank God for giving me this opportunity to learn and for giving me the strength to succeed. How great are your works, O Lord!

## TABLE OF CONTENTS

	Page
LIST OF TABLES .....	viii
LIST OF FIGURES .....	ix
 Chapter	
I. INTRODUCTION.....	1
PROTEIN STRUCTURE .....	1
STABILITY OF A PROTEIN STRUCTURE .....	4
PROTEIN FOLDING .....	5
PROTEIN FOLDING PATHWAY .....	6
PREDICTION OF THE TERTIARY STRUCTURE FROM AMINO ACID SEQUENCE .....	10
COMPUTATIONAL AND EXPERIMENTAL STUDIES USING $G_A$ AND $G_B$ MUTANTS.....	15
THERMAL UNFOLDING SIMULATION OF $G_A^{95}$ AND $G_B^{95}$ .....	21
II. METHODOLOGY.....	23
CHARMM: MOLECULAR DYNAMICS SIMULATION PROGRAM.....	23
MOLECULAR DYNAMICS SIMULATIONS .....	28
INITIALIZATION.....	30
EQUILIBRATION .....	32
PRODUCTION DYNAMICS .....	32
ANALYSIS OF THE MOLECULAR DYNAMICS SIMULATIONS.....	33
III. RESULTS AND DISCUSSION.....	36
IV. CONCLUSION .....	73
REFERENCES.....	77
APPENDIX.....	79
VITA.....	85



## LIST OF TABLES

Table	Page
1. Distances for the Long Range Interactions in 2kdl300.....	45
2. Distances for the Long Range Interactions in 2kdm300.....	45
3. Total Number of Contacts <10Å in 2kdl550A .....	50
4. Total Number of Contacts <10Å in 2kdl550B .....	52
5. Total Number of Contacts <10Å in 2kdl550C1 .....	54
6. Total Number of Contacts <10Å in 2kdl550C2 .....	55
7. Average Number of Contacts <10Å in 2kdl550A, 2kdl550B and 2kdl550C1 ....	58
8. Average Number of Contacts <10Å for each 2 ns Interval of 2kdl550A (0-10ns), 2kdl550B (0-10ns) and 2kdl550C2 (2-12ns) .....	60
9. Total Number of Contacts <10Å in 2kdm550A .....	63
10. Total Number of Contacts <10Å in 2kdm550B .....	64
11. Total Number of Contacts <10Å in 2kdm550C .....	66
12. Average Number of Contacts <10Å for each 2 ns Interval of 2kdm550A (0-10ns), 2kdm550B (0-10ns) and 2kdm550C (4-14ns).....	69

## LIST OF FIGURES

Figure	Page
1. Structure of a protein.....	2
2. Energy landscape diagram for protein folding .....	7
3. Models of protein folding.....	9
4. The structures and amino acid sequences of Staphylococcal protein G proteins, with Protein Data Bank (PDB) code of PSD-1 for G <sub>A</sub> and GB1 for G <sub>B</sub> .....	11
5. Sequence alignment for designed proteins of G <sub>A</sub> (top) and G <sub>B</sub> (bottom) .....	13
6. Backbone topology of G <sub>A</sub> 95 and G <sub>B</sub> 95 .....	14
7. Folding pathways of G <sub>A</sub> 88 (above) and G <sub>B</sub> 88 (below) from the reverse unfolding simulation process.....	16
8. C $\alpha$ RMSD vs. time graphs of trajectories at 298 k and 348 k for NMR and homology models with (A) 88% sequence identities and (B) 95% sequence identities.....	19
9. (a) Cluster analysis on trajectories. b) NMR structures of G <sub>B</sub> 88 and G <sub>A</sub> 88, with PDB codes of 2JWU and 2JWS, respectively .....	20
10. Proper dihedral and improper dihedral .....	26
11. Lennard-Jones potential for van der Waals interactions .....	27
12. Solvation of GA95 and GB95 in truncated octahedron water box.....	31
13. Alignment of 300K minimized average structure (light gray) of (A) 2kdl300 and (B) 2kdm300, with minimized NMR structure (black) in VMD .....	37
14. RMSD vs. time graph of 2kdl300 (black = all amino acids, Dark gray = residues 8-51) and 2kdm300 (light gray).....	37
15. RMSDs of samples of (A) 2kdl550 and (B) 2kdm550 as a function of time, relative to RMSDs of 2kdl300 and 2kdm300 .....	39

16.	RGYRs of samples of (A) 2kdl550 and (B) 2kdm550 as a function of time, relative to RGYRs of 2kdl300 and 2kdm300 .....	40
17.	Snapshots of unfolding trajectory of 2kdl550A (top) and 2kdm550A (below), at ~ 6 ns, 7 ns and 10 ns .....	41
18.	G <sub>A</sub> 95 long range distance vs. time graphs of (A) Leu20-Ile30, (B) Thr25-Ile49, and (C) Leu32-Leu45 .....	47
19.	G <sub>B</sub> 95 long range distance vs. time graphs of (A) Gly41-Thr55, (B) Leu5-Phe30, and (C) Tyr45-Phe52.....	48
20.	Average structure of G <sub>A</sub> 95 showing the important long range interactions.....	62
21.	Two views of the average structure of G <sub>B</sub> 95 showing the important long range interactions.....	70

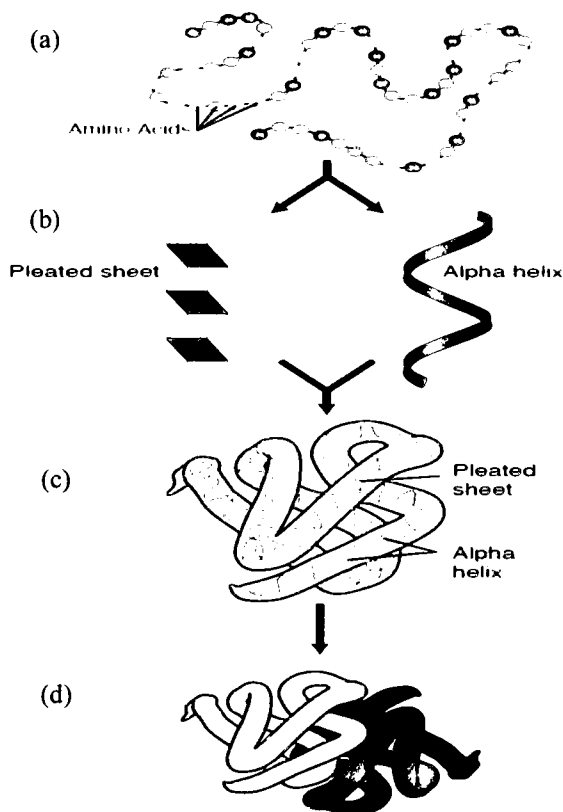
## **CHAPTER I**

### **INTRODUCTION**

#### **Protein Structure**

Proteins are biomolecules, mainly composed of amino acids. They are diverse in nature and are found in abundance in cells. More than 50% of the dry weight of cells is made up of proteins, which indicates that they play a vital role in cellular structures and functions. Proteins have a wide range of functions and can be classified according to their molecular function into enzymes, regulatory proteins, transport proteins, structural proteins, contractile and motile proteins, storage proteins, and protective proteins.<sup>1</sup> In order for proteins to function, they must be correctly folded into their native three dimensional structures. Therefore, failure to fold correctly or to remain correctly folded will lead to malfunctions in a living system and hence to different diseases such as Alzheimer's and Parkinson's diseases, type II diabetes, cystic fibrosis, and some types of cancers.<sup>2</sup>

Proteins have complex structures. Their structure can be described in terms of four levels of organization.<sup>1</sup> The amino acid sequence of polypeptide chain(s) is known as the primary structure. The residues in the primary structure are joined together by peptide bonds. An example of a primary structure is shown in Figure 1a. The primary structure determines the characteristics of a protein. Proteins contain at least 40 residues, with the majority of the polypeptides consisting of 100 to 1000 residues.<sup>3</sup> With the 20 amino acids available, there are a large number of different possible amino acid sequences, and hence, a variety of proteins with different properties.<sup>3</sup>



**Figure 1.** Structure of a protein. (a) Primary structure; (b) Secondary structure:  $\beta$ -pleated sheet and  $\alpha$ -helix; (c) Tertiary structure; and (d) quaternary structure.<sup>4</sup>

Secondary structure is one of the higher levels of protein structure, in which the peptide chains arrange themselves into regular folding patterns such as helices, sheets, and turns through hydrogen bonding interactions.<sup>1, 3</sup> The  $\alpha$ -helix is a right-handed secondary structure (Figure 1b). It has 3.6 residues per turn and an average length of 12 residues, which means that an  $\alpha$ -helix contains about 3 helical turns. In the  $\alpha$ -helix, the hydrogen bonds are formed between the amide hydrogen and carbonyl oxygen of the  $n$  and  $n+4$  backbone residues, respectively.<sup>3</sup>

$\beta$  sheets are another type of secondary structure, which is formed by hydrogen bonds between the polypeptide backbones. They are different from  $\alpha$ -helices in that the hydrogen bonding is between neighboring polypeptide chains rather than within one polypeptide chain as in an  $\alpha$ -helix (Figure 1b). There are two types of  $\beta$  sheets: parallel and anti-parallel. In a parallel  $\beta$  sheet, the hydrogen-bonded polypeptide chains are lined up in the same direction, i.e. all the adjacent chains are lined up from N terminal to C terminal; whereas in anti-parallel  $\beta$  sheet, the polypeptide chains are in the opposite direction, i.e., a polypeptide chain with N  $\rightarrow$  C terminal is lined up with an adjacent C  $\rightarrow$  N terminal chain.  $\beta$  sheets can contain from 2 to 22 polypeptide strands, with an average of 6 strands.<sup>3</sup> Each strand is made up of up to 15 residues, with an average number of 6 residues.<sup>3</sup>

Tertiary structure is the functional form of a protein that is formed when secondary structures are bent and folded to form a more compact three-dimensional shape. (Figure 1c).<sup>1,5</sup> Quaternary structure is the highest level of protein structure, which is formed when separate polypeptide chains with a characteristic tertiary structure interact and assume a specific geometry.<sup>1, 3</sup> The interacting polypeptide chains are known as subunits and they can be identical or non-identical (Figure 1d). The formation of quaternary structure is important because it provides stability to proteins that are unstable by themselves, can form an active enzyme by bringing the different catalytic sites of subunits together, and allows subunits with different binding affinities to a ligand to work together cooperatively.<sup>1</sup>

## Stability of a Protein Structure

The primary structure of proteins is formed by covalent peptide bonds whereas the higher levels of protein structures: secondary, tertiary and quaternary structures are formed by non-covalent interactions which are weak, but play a major role in stabilizing the protein structures. The non-covalent interactions include hydrophobic interactions, van der Waals forces, hydrogen bonds, electrostatic interactions, and disulfide bonds.<sup>1</sup>

Non-polar side chains of amino acids minimize their contacts with polar solvents such as water by aggregating in the core of the protein. The aggregation of the non-polar side chains is known as the hydrophobic effect and this interaction is the major determinant of the native structure of a protein and the protein folding pathway.<sup>3</sup> Hydrogen bonds can be formed between peptide backbones, surface side chains, and between surface side chains and water. The stabilization energy contributed by a hydrogen bond is small. However, due to the large number of hydrogen bonds in a protein, hydrogen bonds play a significant role in stabilizing the protein structure.<sup>1</sup> Electrostatic interactions or salt bridges are formed between charged residues, which are mainly found on the protein surface.<sup>3</sup> In addition to the charged residues, the N-terminal and C-terminal can contribute to this interaction.<sup>1</sup> Charged residues are able to interact with water when the protein is unfolded. Electrostatic interactions between charged residues are formed when the protein folds and these new interactions result in the loss of entropy from the side chains and the loss of solvation free energy from breaking the hydrogen bonds with water. As a result, even though electrostatic interactions are strong, they contribute little to the stability of the protein structure.<sup>3</sup> Van der Waals forces are attractive and repulsive forces which are formed between adjacent non-bonded atoms.<sup>1</sup>

The repulsive force arises when the interaction between electrons of adjacent atoms is strong.<sup>6</sup> The attractive forces are mainly due to dipole-induced dipole interactions that results from fluctuations in the electron distribution of adjacent atoms. Van der Waals forces are important protein interactions and the attractive portion, known as the London dispersion force, is a major contributor to the hydrophobic effect.<sup>1</sup> Disulfide bonds are formed from the oxidation of thiol groups of nearby cysteine residues. These bonds are somewhat important in stabilizing the protein structures and can help to “lock in” a particular backbone conformation during folding.<sup>3</sup>

## **Protein Folding**

Protein folding is the process in which proteins adopt their native three-dimensional structure.<sup>7</sup> For over five decades, different studies have been carried out to learn how proteins attain their native state. However, the folding problem continues to be a puzzle in modern science and different researches are still in progress.<sup>8</sup> The protein folding problem has two parts: (1) the prediction of the tertiary structure of proteins from the amino acid sequence and (2) the description of protein folding pathways or the mechanism of protein three-dimensional structure formation.<sup>9,10</sup> Solving the protein folding problem will help understand the underlying mechanisms of how proteins fold. This information will provide further insight into how proteins misfold, how they cause different diseases, and hence, how to treat these diseases. In addition, knowledge of how amino acids dictate a tertiary structure will enable us to predict the structure directly from an amino acid sequence and avoid the difficulty of obtaining an experimental structure. The design of sequences that will encode new protein folds will also be



possible.<sup>2</sup> If protein structures can be predicted, then we will be able to determine the functions of proteins and design structure-based drugs.<sup>11</sup>

### **Protein Folding Pathway**

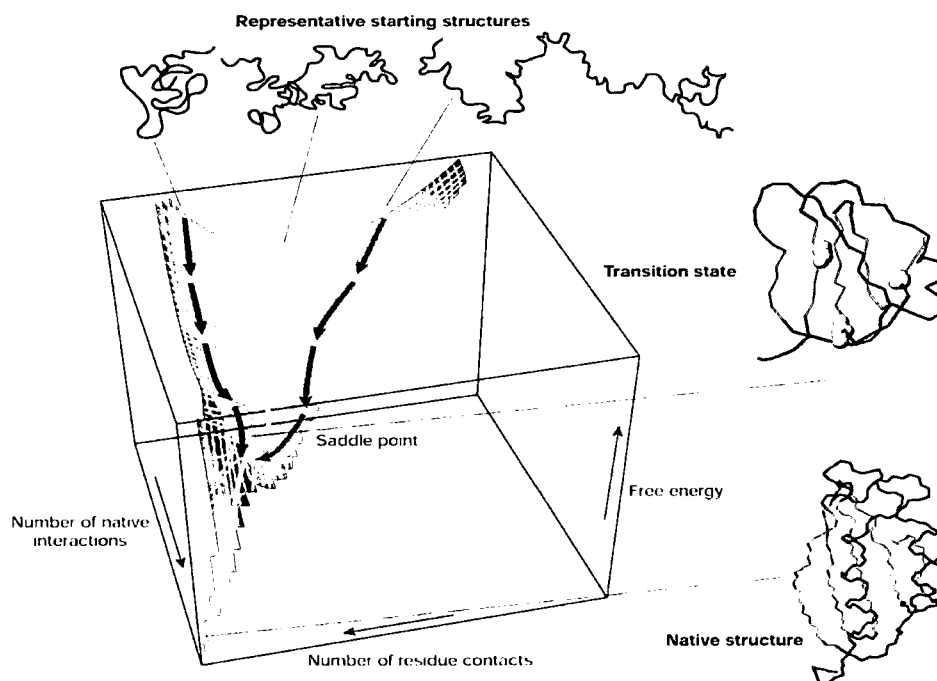
In 1957, Christian Anfinsen demonstrated that denatured ribonuclease A (RNase A), a single chain protein with 124 residues, can spontaneously fold back to the native structure upon removal of the denaturing agents. This experiment proved that a protein's primary structure contains all the information required for a protein to fold to its native three-dimensional structure.<sup>3</sup> Christian Anfinsen won a Noble prize for his work.<sup>12</sup> However, how exactly the amino acids dictate the stable functional three-dimensional structure is still a mystery.

In 1968, Cyrus Levinthal suggested that unfolded polypeptide chains can have a very large number of possible conformations and therefore on a biological time scale, it is impossible for a protein to attain its correct stable conformation by random sampling of all the possible conformations.<sup>13</sup> It only takes seconds or less for proteins to fold to their native three dimensional structures.<sup>14</sup> This argument is known as the Levinthal paradox and can be explained by using a protein with 100 amino acids. Assuming that there are only two conformational possibilities for each amino acid, the total number of possible conformations for the protein will be  $2^{100}$ . If it takes  $10^{-13}$  sec for the protein to test each possible conformation and find the stable conformation, it will take:

$$(10^{-13} \text{ sec}) (2^{100}) = 1.27 \times 10^{17} \text{ sec} = 4 \times 10^9 \text{ years (approximate age of the earth)}^1$$

The Levinthal paradox has led to the concept of a protein folding pathway, which can be depicted as an energy landscape diagram from the unfolded state of the protein to

the native or folded state (Figure 2).<sup>13</sup> The energy landscape resembles a funnel in that a wide top becomes narrower as it approaches the bottom. The wide part of the landscape on the top represents the different possible denatured structures of a protein, which are characterized by high free energy. The middle of the energy landscape, which contains a saddle point, corresponds to the transition state. The transition state is a pivotal region with an energy barrier that all molecules must pass through to fold to their native states.<sup>2</sup>



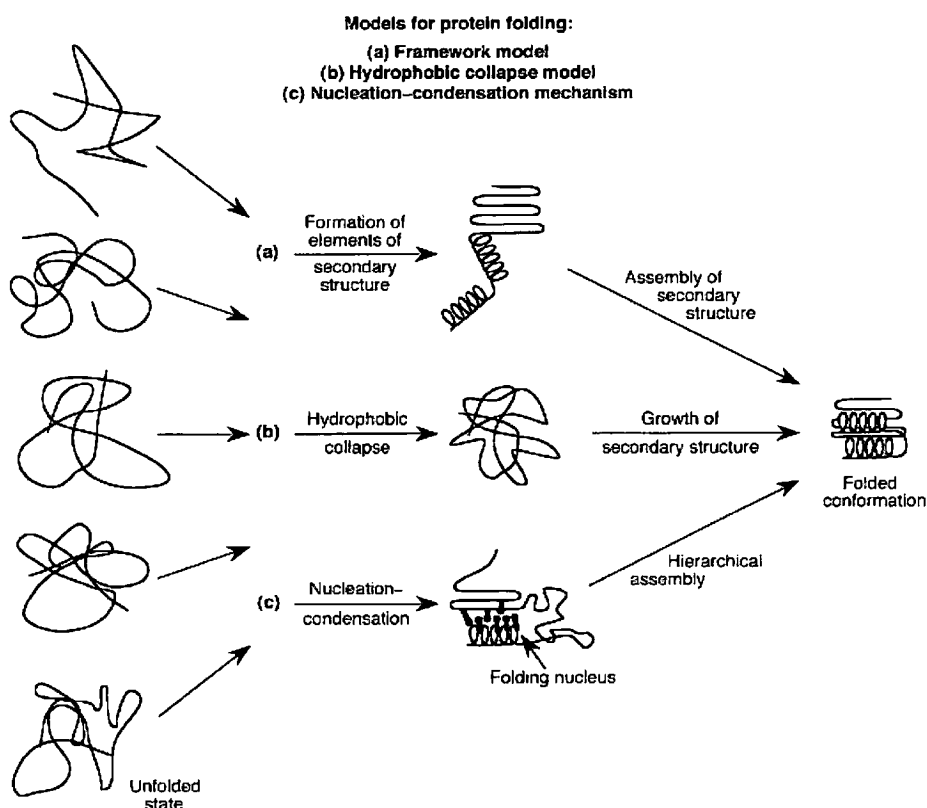
**Figure 2.** Energy landscape diagram for protein folding. The structures which are superimposed on the transition state are different possible conformations of the transition state. The balls in the transition state represent the residues from which the native structure is established.<sup>2</sup>

In the energy landscape diagram, the transition state is shown by an ensemble of structures to indicate that the formation of the transition state involves simultaneous formation and breaking of non-covalent interactions which result in the formation of structures with close, but different, conformations and similar energy.<sup>15</sup> The narrow part of the energy landscape represents a decrease in free energy and the number of possible conformations, as the contacts between residues increase leading toward the native state.<sup>2</sup>

A variety of parallel paths can exist as a protein folds and usually includes the formation of several transiently populated species known as intermediates, which are partially folded structures.<sup>8,16</sup> There are different proposed models of protein folding (Figure 3). One of the proposed models is known as Framework model (Figure 3a). In the Framework model, as a protein folds, intermediates containing secondary structures form first.<sup>17</sup> The  $\alpha$ -helices can form in less than 1  $\mu$ s, and less frequently, the  $\beta$ -hairpins can also form as an isolated secondary structure. Since, the secondary structures by themselves are not stable; they will assemble into stable tertiary structure through hydrophobic and long range interactions.<sup>8,18</sup>

The second model of folding is known as hydrophobic collapse (Figure 3b). In the hydrophobic collapse model, the protein first collapses through hydrophobic effect. Then stable secondary structures start to form in the collapsed state.<sup>17</sup> The collapsed intermediate state that mostly consists of secondary structures is known as a molten globule.<sup>3</sup> The third model of protein folding is called nucleation-condensation mechanism which involves the formation of folding nucleus in the transition state (Figure 3c).<sup>17,19</sup> The folding nucleus contain a key set of interactions between residues

from which the native structure is established.<sup>8</sup> The efficiency of folding and the folding pathway are dependent upon the amino acid sequence of a protein and considering the large number of proteins with different amino acid sequences, it is difficult to find a general folding pathway. In addition, whereas large proteins with more than ~100 residues have significantly populated intermediates during folding, small proteins with 60-100 residues can exhibit two state folding, without an intermediate.<sup>2</sup>



**Figure 3.** Models of protein folding. (a) Framework model; (b) Hydrophobic collapse model; and (c) Nucleation-condensation mechanism.<sup>17</sup>

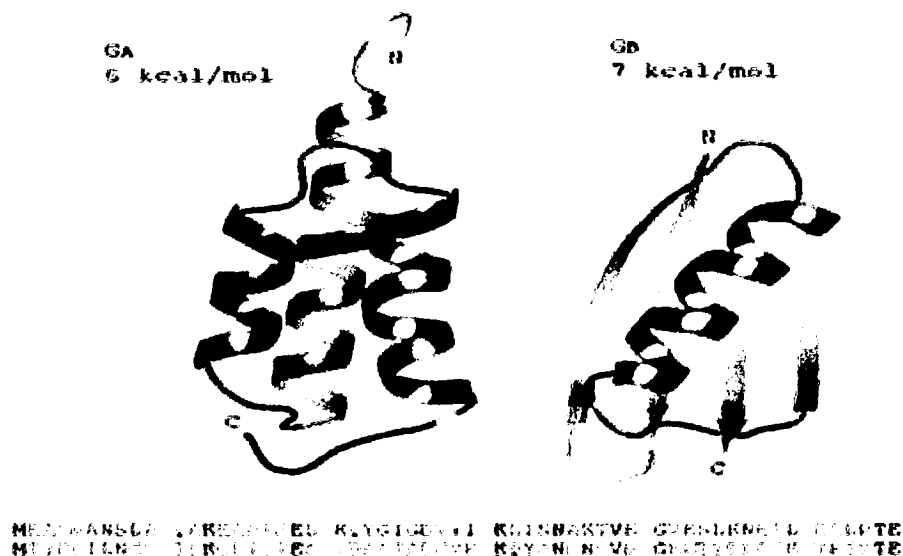
## **Prediction of the Tertiary Structure from Amino Acid Sequence**

Understanding how an amino acid sequence determines protein structure remains a very challenging problem.<sup>20</sup> This difficulty arises from the fact that the degree of contribution of an amino acid to a particular protein fold varies for different proteins. Mutations of some amino acids in a protein can have an insignificant effect on the protein stability, whereas mutations of some other amino acids in the same protein result in complete unfolding. Secondly, a polypeptide chain has a large number of possible conformations and this makes it difficult to calculate the most preferred conformation and study the relationship between the amino acid sequence and the most preferred structure.<sup>21</sup>

The most common structure prediction method from amino acid sequence is known as comparative or homology modeling. In this modeling method, the sequence of a protein with an unknown structure, called the target, is aligned with the sequence of a protein with a known structure, called the template. If the target and the template have detectable similarity, then the three dimensional structure of the target can be predicted from the structure of the template. Comparative modeling is based on the fact that three dimensional structures are more conserved than sequences, i.e., proteins with large difference in sequence can have similar structures. More accurate structure prediction can be attained if the sequence identities between the target and the template are higher.<sup>22</sup>

A new approach to study the relationship between protein structure and amino acid sequence is the design of two proteins that have maximal sequence identity but different folds, and hence functions. The non-identities between the two proteins would

then be responsible for coding the folds.<sup>21</sup> Alexander et al. used protein G as a starting point, which is a multi-domain cell wall protein from *Streptococcus*.<sup>20</sup> Protein G contains two types of domains which bind to serum proteins in the blood: the G<sub>A</sub> domain that binds to human serum albumin (HSA) and the G<sub>B</sub> domain that binds to the constant (Fc) region of immunoglobulin G (IgG).<sup>23,24</sup> The *Streptococcus* bacteria use protein G to bind to serum proteins in the host, which hides the bacteria from detection by the host. The G<sub>A</sub> and G<sub>B</sub> domains both have 56 amino acids.<sup>20</sup> In the G<sub>A</sub> domain, amino acids 1-8 and 54-56 are not ordered in the Nuclear Magnetic Resonance (NMR) structures, whereas all 56 amino acids in the G<sub>B</sub> domain are well ordered.<sup>25</sup> The G<sub>A</sub> and G<sub>B</sub> domains share only 16% sequence identity and have different folds: a 3- $\alpha$  helical structure and an  $\alpha/\beta$  fold, respectively (Figure 4).<sup>20</sup>

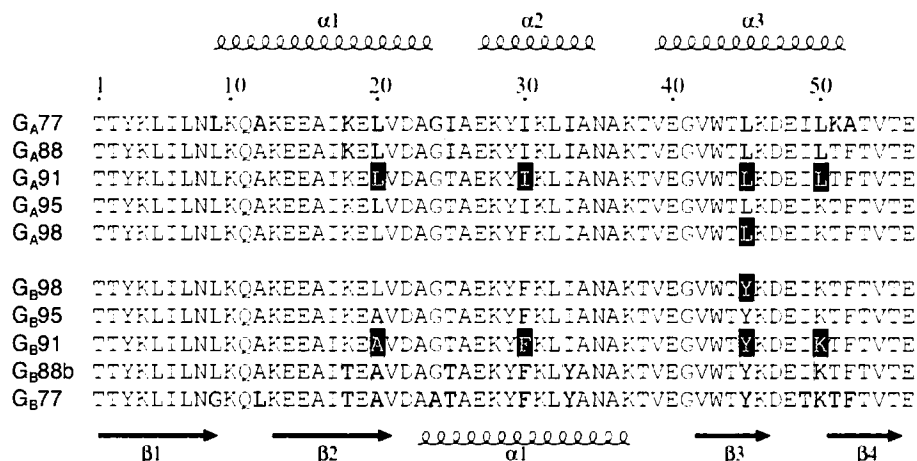


**Figure 4.** The structures and amino acid sequences of Streptococcal protein G proteins, with Protein Data Bank (PDB) code of PSD-1 for G<sub>A</sub> and GB1 for G<sub>B</sub>. The identical residues are in bold.<sup>20</sup>

The  $G_A$  and  $G_B$  domains were mutated in such a way that the binding sites of HSA and IgG were encoded in both proteins, so that the HSA-binding site is functional in the 3- $\alpha$  fold and latent in the  $\alpha/\beta$  fold, but the IgG-binding site is functional in the  $\alpha/\beta$  fold and latent in the 3- $\alpha$  fold. This initial mutation resulted in the design of two proteins that have 30% sequence identity and 40 non-identical amino acids, called  $G_{A30}$  and  $G_{B30}$ . Further mutations resulted in the design of proteins with 77% ( $G_{A77}$  and  $G_{B77}$ , with 13 non-identical amino acids), and 88% ( $G_{A88}$  and  $G_{B88}$ , with only 7 non-identical amino acids) sequence identities.<sup>20</sup> Starting from the parent  $G_A$  protein, a total of 24 mutations were made to design  $G_{A88}$ , and 17 mutations were made to the parent  $G_B$  protein to design  $G_{B88}$ .<sup>21</sup> The designed proteins maintained the folds and functions of their parent proteins. Studies using thermal denaturation showed that as the sequence identity increased, the stabilities of the designed proteins decreased. Moreover, the stability of the  $G_B$  mutants relative to the  $\Delta G_{\text{unfolding}}$  of the parent was less than that of the  $G_A$  mutants. The  $\Delta G_{\text{unfolding}}$  of  $G_{A88}$  was  $\approx 4$  kcal/mol, whereas the  $\Delta G_{\text{unfolding}}$  of  $G_{B88}$  was  $\approx 2$  kcal/mol.<sup>20</sup>

Increasing the amino acid identities in both proteins helps minimize the number of amino acids which are responsible for a specific fold and thus makes the study of the relationship between sequence and protein fold easier. Interestingly, it was possible to attain proteins with even higher sequence identities of 91% ( $G_{A91}$  and  $G_{B91}$ ), 95% ( $G_{A95}$  and  $G_{B95}$ ), and 98% ( $G_{A98}$  and  $G_{B98}$ ) (Figure 5). These proteins have the same structures and functions as their wild type proteins but their stabilities were further compromised with  $\Delta G_{\text{unfolding}} \approx 3$  kcal/mol. The  $G_{A98}$  and  $G_{B98}$  are very unstable and,

therefore, difficult to study experimentally.<sup>25</sup> Thus, any future studies on these proteins would be more feasible for G<sub>A</sub>95 and G<sub>B</sub>95 than for the 98% identity proteins.

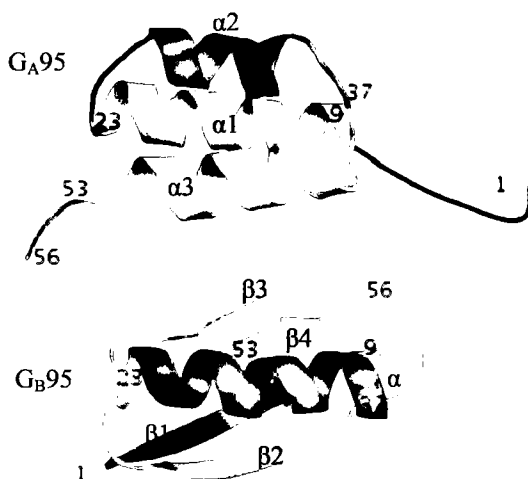


**Figure 5.** Sequence alignment for designed proteins of G<sub>A</sub> (top) and G<sub>B</sub> (bottom). The non-identical residues for each are highlighted. The secondary structures at the top and bottom of the alignment are for G<sub>A</sub>95 and G<sub>B</sub>95, respectively.<sup>25</sup>

The G<sub>A</sub>95 and G<sub>B</sub>95 proteins differ only by three residues at positions 20, 30, and 45 in the amino acid sequence. Residue 20 is leucine in G<sub>A</sub>95 and alanine in G<sub>B</sub>95; residue 30 is isoleucine in G<sub>A</sub>95 and phenylalanine in G<sub>B</sub>95; and residue 45 is leucine in G<sub>A</sub>95 and tyrosine in G<sub>B</sub>95. As with their parent proteins, G<sub>A</sub>95 and G<sub>B</sub>95 have a 3-α fold and a 4β + α fold, respectively. The first 8 amino acids (1-8) in G<sub>A</sub>95 are not ordered in its NMR structure, whereas in G<sub>B</sub>95, these amino acids form the first β strand. Amino acids 9-23 in G<sub>A</sub>95 form the first α helix; while they form the turn between the first and second β strands, the second β strand, and the turn between the second β strand and central helix in G<sub>B</sub>95. The only similarity between G<sub>A</sub>95 and G<sub>B</sub>95 is that the amino



acids 27-33 in G<sub>A</sub>95 form the second  $\alpha$  helix and amino acids 24-37 in G<sub>B</sub>95 form the central helix. The third  $\alpha$  helix in G<sub>A</sub>95 is formed from amino acids 39-51. The amino acids 39-51 in G<sub>B</sub>95 form the turn between the central helix and the third  $\beta$  strand, the third  $\beta$  strand, and the first part of the fourth strand. The remaining part of the fourth  $\beta$  strand in G<sub>B</sub>95 is formed from amino acids 52-56, whereas the amino acids 52-56 in G<sub>A</sub>95 are not ordered (Figure 6).



**Figure 6.** Backbone topology of G<sub>A</sub>95 and G<sub>B</sub>95. The residues 1, 9, 23, 37, 53, and 56 are shown to compare the locations of the different secondary structures in both proteins.<sup>25</sup>

NMR studies of the hydrophobic interactions in G<sub>A</sub>95 and G<sub>B</sub>95 show that in G<sub>A</sub>95, Leu20 and Ile30 are within the hydrophobic core; whereas in G<sub>B</sub>95, all three non-identical residues, Ala20, Phe30, and Tyr45 are found in the hydrophobic core. Moreover, in G<sub>A</sub>95, the residues that form a tight hydrophobic network are Ala16, Leu20, Ile30, Ile33, and Ile49. In G<sub>B</sub>95, the Tyr3, Leu5, Phe30, and Phe52 network

forms about 50% of the hydrophobic core. The hydrophobic core involves the residues in the  $\beta 1$  and  $\beta 4$  strands, which form the tails in the 3- $\alpha$  fold. Of the three non-identical residues in  $G_A95$  and  $G_B95$ , residue 45 is an important determinant for the formation of the hydrophobic core and thus for the switch between the 3- $\alpha$  fold and  $4\beta + \alpha$  fold. Residue Tyr45 in  $G_B95$  forms strong contacts with Phe52 and Asp47 and these contacts are thought to stabilize the  $\beta 3/\beta 4$  hairpin turn. In  $G_A95$ , mutation of L45Y and A52F destabilizes the 3- $\alpha$  fold by -1.5 kcal/mol, but Tyr45 and Phe52 in  $G_B95$  increases the stability of the  $4\beta + \alpha$  fold by +2 kcal/mol.  $G_A95$  maintains a 3- $\alpha$  fold even when residue 30 is mutated to the  $G_B95$  residue (F30L). Complimentarily,  $G_B95$  maintains a  $4\beta + \alpha$  fold if residue 20 is mutated to the  $G_A95$  residue (A20L).<sup>25</sup>

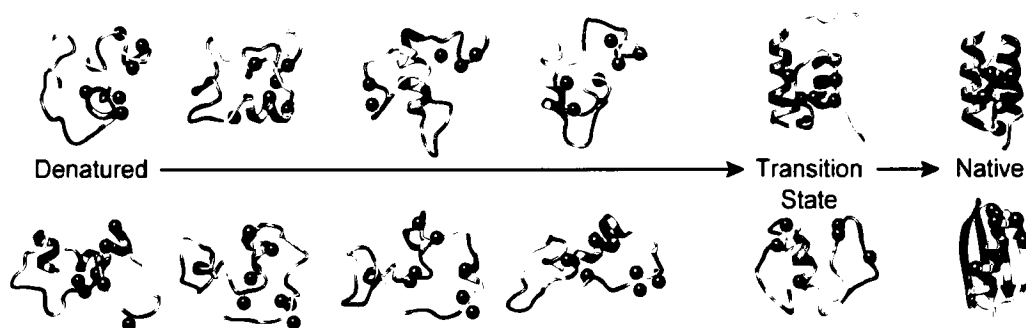
### Computational and Experimental Studies Using $G_A$ and $G_B$ Mutants

To date, three molecular dynamics studies have been published on the  $G_A$  and  $G_B$  mutant proteins.<sup>26,27,28</sup> Morrone et al. research the folding mechanisms of  $G_A88$  and  $G_B88$  using experiments involving pH changes and molecular dynamics simulations. Urea induced equilibrium and kinetic experiments at different pH ranges were performed on  $G_A88$  and  $G_B88$ .<sup>26</sup> Both the equilibrium and kinetic experiments indicate that  $G_A88$  and  $G_B88$  follow a two-state folding mechanism and that  $G_A88$  is more stable than  $G_B88$ . However, it was suggested that the denatured  $G_B88$  has residual structure.

Room temperature and unfolding molecular dynamic simulations, using *in lucem* molecular mechanics (*ilmm*) software, were performed to confirm experimental results. It was observed that the central  $\alpha$ -helix in  $G_B88$  was maintained throughout the

unfolding process, that the structure  $\beta 1/\beta 2$  was observed more often than  $\beta 3/\beta 4$  and that both are separated from each other.

Morrone et al. also studied the reverse of the unfolding simulation process to see how the  $G_{A88}$  and  $G_{B88}$  attain their  $\alpha$ -helical and  $\alpha/\beta$  structures (Figure 7). In  $G_{A88}$ , the denatured state contained some residual structures with different main chain interactions, which become compact and form helical structures in the transition state (TS). The native structure of  $G_{B88}$  can be detected in its denatured state because it has two hairpin regions that are separated by a central helix. The central helix in  $G_{B88}$  is more stable than that of  $G_{A88}$ , which may constrain the  $G_{B88}$  sequence from folding to the  $3\alpha$  helix structure. Hydrogen bonds of Asp47 and Glu48 with Thr1 favored the formation of the  $3\alpha$  fold in the denatured state of  $G_{A88}$ ; whereas in  $G_{B88}$  hydrogen bonds between Asp47-Lys50 and Asp47-Tyr45 favor the formation of  $\beta 3/\beta 4$  hairpin in the denatured state of  $G_{B88}$ .



**Figure 7.** Folding pathways of  $G_{A88}$  (above) and  $G_{B88}$  (below) from the reverse unfolding simulation process. The balls indicate the non-identical residues in  $G_{A88}$  and  $G_{B88}$ .<sup>26</sup>

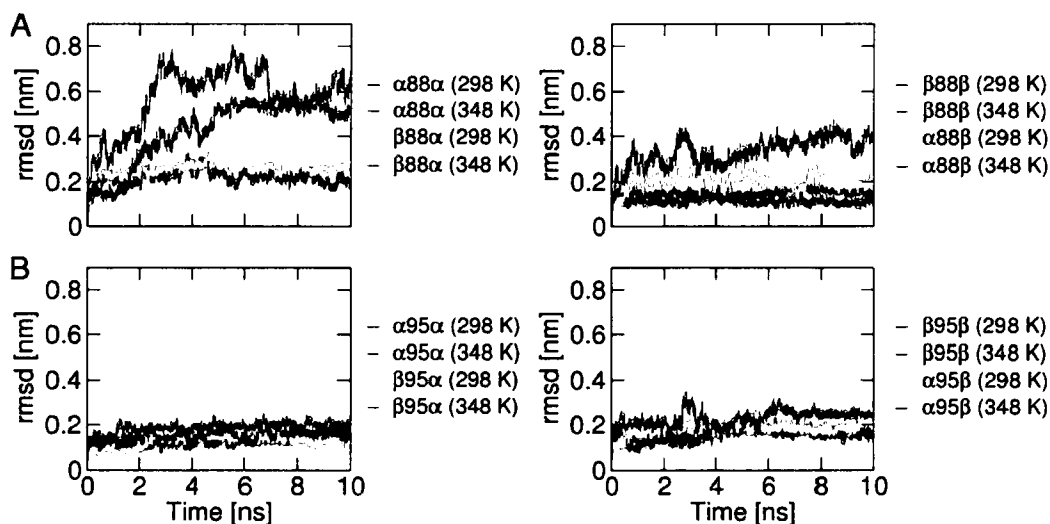
Consequently, the formation of a stable  $\beta 3/\beta 4$  hairpin favors the formation of the  $4\beta+\alpha$  fold in  $G_B95$ . These observations indicate that whether the proteins attain an  $\alpha$  or  $\beta/\alpha$  fold is determined early in the folding pathway; and it can be detected from their residual structures and long range interactions in the denatured state. Only a few residues determine the native structures of the proteins and these residues work together to form stabilized-long range interactions. Formation of some or all of these contacts creates a folding nucleus in the denatured state.  $G_A88$  and  $G_B88$  have different nuclei according to their amino acid sequences and this leads to the different folds of these proteins.<sup>26</sup>

Allison et al. used computer modeling to study why  $G_A$  and  $G_B$  mutants that have high amino acid sequence identities of 88 and 95% have different structural preferences, but were unable to come up with any concrete conclusions.<sup>27</sup> NMR structures were studied alongside homology models which were created by fitting sequences that fold into one type of structure onto the other structure. For example, the  $G_A88$  sequence was fitted onto the  $4\beta+\alpha$  fold of  $G_B88$  to form a homology model  $\alpha 88\beta$ ; whereas  $\alpha 88\alpha$  is just the NMR structure of  $G_A88$ . The similarity of the homology structures with the NMR template structures was verified by calculating the Nuclear Overhauser Effect (NOE) distances and comparing the results with experimental data.<sup>27,29</sup> 95% of  $\beta 88\beta R$  (reduced, or back bone) NOE data was satisfied by  $\alpha 88\beta$ , but only 87% of  $\alpha 88\alpha R$  NOE data was satisfied by  $\beta 88\alpha$ . The results indicate that the alpha sequence is well-matched with the  $4\beta+\alpha$  structure, whereas the beta sequence was less compatible with the  $3-\alpha$  fold.<sup>27</sup> This result agrees with the CASP8 (Critical Assessment of Techniques for Protein Structure Prediction) competition results for  $G_A95$  and  $G_B95$ , where most web servers predicted the  $4\beta+\alpha$  fold, for both  $G_A95$  and  $G_B95$  sequences.<sup>30</sup>

Molecular dynamics simulations were performed on the NMR and homology models using the GROMOS program and force field at 298 K and 348 K for 10 ns. The stabilities of the structures from the trajectory were determined using root mean square deviations (RMSD) of the C $\alpha$  atoms (Figure 8). Residues 10-50 of the 3- $\alpha$  structures and all the residues in 4 $\beta$ + $\alpha$  structures were used to calculate the RMSD. The RMSD results show that the 3- $\alpha$  structures with 88% sequence identity were less stable than that of the 4 $\beta$ + $\alpha$  structures. Both  $\alpha$ 88 $\alpha$  and  $\beta$ 88 $\alpha$  had RMSDs greater than 0.5 nm at 348 K. Comparing the secondary structures in  $\alpha$ 88 $\alpha$  and  $\beta$ 88 $\alpha$ , the 3  $\alpha$  helices in  $\alpha$ 88 $\alpha$  are very stable and the fluctuation in the RMSD is caused by changes in the three dimensional arrangement of the helices. In the  $\beta$ 88 $\alpha$  structure, the ends of the helices were disordered and the first  $\alpha$ -helix was not stable during the simulation process. The  $\beta$ 88 $\beta$  structures were stable at 298 K and 348 K, but  $\alpha$ 88 $\beta$  had slightly higher RMSDs at both temperatures which indicate that the  $\alpha$ 88 $\beta$  is less stable than  $\beta$ 88 $\beta$  structure. The deviations in the RMSDs of  $\alpha$ 88 $\beta$  could be due to the disruption of the strong interactions among residues 45, 47, and 52 that form the  $\beta$ 3/ $\beta$ 4 hairpin and are reduced by the presence of Leu45 instead of Tyr45.

Surprisingly, structures with 95% sequence identity had smaller RMSDs, at both 298 K and 348 K. The stability of the structures with 95 % sequence identity contradicts the experimental results where the stabilities of the designed G<sub>A</sub> and G<sub>B</sub> proteins decrease with increasing sequence identity. In addition, from the molecular dynamics simulation, Allison et al. were unable to differentiate the structural preferences of the proteins with 95% sequence identity because of their similar RMSDs. The stability of the sequence-structure combinations with 88 and 95% sequence identities were further

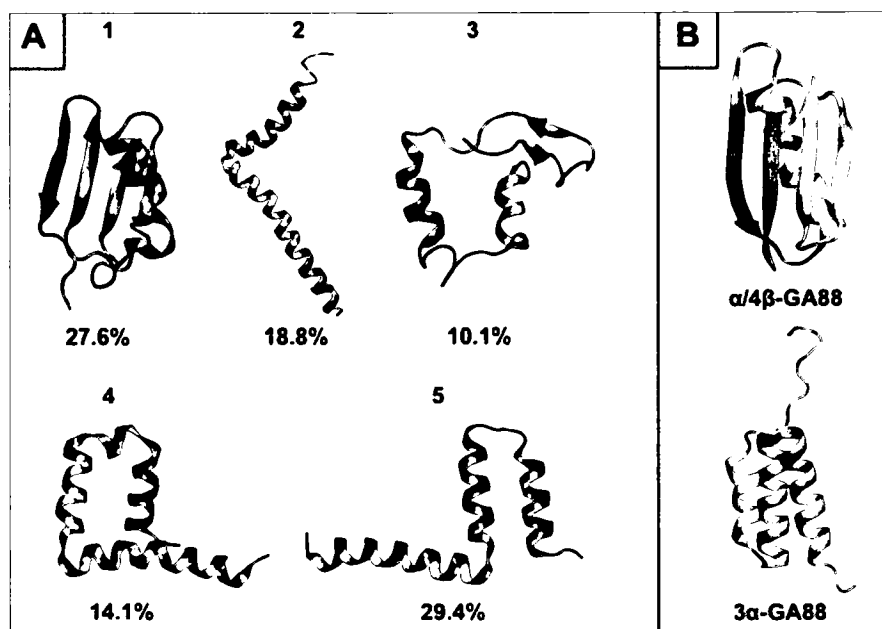
studied by calculating their intraprotein potential energies through different methods. However, the results were contradictory and thus, could not contribute to conclusions as to which structure is more preferred.<sup>27</sup>



**Figure 8.** C $\alpha$  RMSD vs. time graphs of trajectories at 298 K and 348 K for NMR and homology models with (A) 88% sequence identities and (B) 95% sequence identities.<sup>27</sup>

Lazim et al. performed a replica exchange molecular dynamics (REMD) simulation to study the folding and unfolding mechanism of the 3- $\alpha$  fold and 4 $\beta$ + $\alpha$  fold.<sup>28</sup> The 4 $\beta$ + $\alpha$  fold of G<sub>B</sub>88 was used as a template structure on which the primary structure of G<sub>A</sub>88 was aligned. The G<sub>A</sub>88 will thus have the 4 $\beta$ + $\alpha$  fold as its non-native structure and this enables one to see the conformational variations that occur when the non-native 4 $\beta$ + $\alpha$  fold of G<sub>A</sub>88 folds to its native 3- $\alpha$  fold. REMD was used because it allows more efficient sampling of conformational space than regular molecular dynamic simulations by using different temperatures at one time.

During the simulations, large variations in the  $C\alpha$ -RMSDs with respect to the  $G_{A88}$  NMR structure were observed, with the lowest value being 4.34 Å at 270 K. To further investigate the conformational variations, cluster analysis was performed on the trajectory at 270 K. Analysis on five clusters show that during the 75 ns simulation process, more than half of the trajectory consisted of  $\alpha$ -helical structures of the 3- $\alpha$  fold, which confirms the variation in the secondary structure from  $\beta$ -strands to  $\alpha$ -helix (Figure 9).



**Figure 9.** (A) Cluster analysis on trajectories. The percentages indicate the occurrence of the different structures during the 75 ns simulation time. (B) NMR structures of  $G_{B88}$  and  $G_{A88}$ , with PDB codes of 2JWU and 2JWS, respectively.<sup>28</sup>

The simulated proteins at lower temperatures were observed to have a higher propensity to form the 3 $\alpha$ -fold. During the simulation, the first eight residues, which are

disordered in the NMR structure, added on to the first helix and the first four residues in the third helix formed a random coil. In addition, the 3- $\alpha$  fold was not as compact as the NMR structure of G<sub>A</sub>88, because the implicit solvent was unable to completely account for the entropic cost of desolvation of the hydrophobic core. Nevertheless, the three helices were present in the lowest RMSD structures. The folding of the three  $\alpha$  helices: H1 (residues 9-23), H2 (residues 27-34) and H3 (residues 39-51) during the trajectories at 270 K and 304 K were further studied by calculating their C $\alpha$ -RMSD relative to the three helical domains from the NMR structure of G<sub>A</sub>88. The C $\alpha$ -RMSD of H1 and H3 decrease through time indicating their folding. The C $\alpha$ -RMSD of H2 was approximately constant which indicates the conservation of this  $\alpha$  helix from the 4 $\beta$ + $\alpha$  fold to the 3 $\alpha$ -fold during the simulation process. Overall, the unfolding pathway consisted of the  $\beta$ -sheet separating into the  $\beta$ 1/ $\beta$ 2 and  $\beta$ 3/ $\beta$ 4 hairpins, which were separated by the central helix. These hairpins then formed the first and third helices, while at the same time, the hydrophobic core of the 4 $\beta$ + $\alpha$  unpacked to form the 3- $\alpha$  fold.<sup>28</sup>

### **Thermal Unfolding Simulation of G<sub>A</sub>95 and G<sub>B</sub>95**

Computational experiments were performed to determine the relationship between the amino acid sequences and the tertiary structures of G<sub>A</sub>95 and G<sub>B</sub>95. The study mainly focuses on how the three non-identical residues dictate the different folds in these proteins sharing high sequence identity. G<sub>A</sub>95 and G<sub>B</sub>95 were thermally unfolded using the molecular dynamics simulation method and the CHARMM program. The purpose of unfolding the G<sub>A</sub>95 and G<sub>B</sub>95 was to see how the long range interactions between the amino acids change during the simulation process. Long range interactions



are defined as contacts that are between 7 or more amino acids in the primary structure and are within or at 6.5 Å distance in the tertiary structure. The long range interactions that are persistent or long-lasting during the unfolding simulation are important in determining the protein folds. Persistent long range interactions were examined to determine if they consisted of the three non-identical residues.

## CHAPTER II

### METHODOLOGY

#### **CHARMM: Molecular Dynamics Simulation Program**

CHARMM (Chemistry at HARvard Molecular Mechanics) is a molecular simulation and modeling program that is used for theoretical investigation of the structures, dynamics, and energies of biological macromolecular systems such as proteins, nucleic acids, carbohydrates, lipids and small molecules such as ligands in solution, vacuum, and crystal environments.<sup>6, 31</sup> The main use of CHARMM and other simulation and modeling programs is to obtain information about a molecular system that is difficult to determine experimentally.<sup>31</sup> CHARMM was first designed in the late 1970s in the laboratory of Professor Martin Karplus at Harvard University and its efficiency and applicability have been successfully developed over many years.<sup>31,32</sup>

The CHARMM program is based on empirical potential energy functions or force fields which are calibrated to experimental results such as structural data obtained from X-ray crystallography and NMR, dynamic data obtained from spectroscopy and thermodynamic data.<sup>32</sup> The program contains various analysis facilities which are used to compare structures, evaluate energies, calculate time series and correlation functions.<sup>6</sup> The potential energy is computed in CHARMM from the PSF (Protein Structure File), which contains complete information about the composition and connectivity of the molecular system of interest and the Cartesian coordinates, which are the atomic positions of the molecular system and are usually obtained from X-ray crystal or NMR

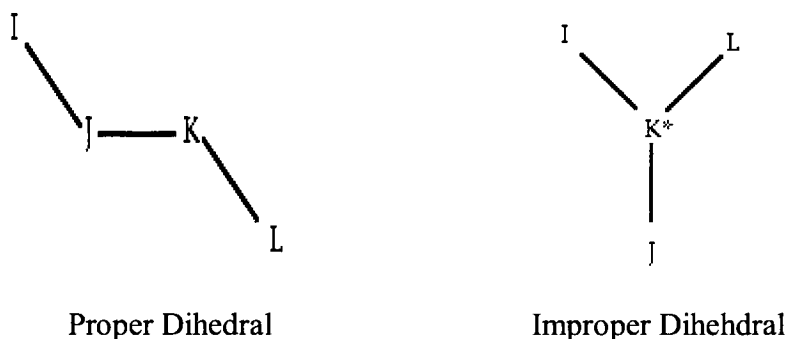
structures from the Protein Data Bank.<sup>32,33</sup> In the case of proteins, the composition consists of the primary sequence.

The CHARMM potential energy function,  $U_R$ , is expressed as the sum of internal or bonded energy terms, and the sum of external or non-bonded energy terms (eq 1).<sup>6,31</sup> The internal energy terms include bond (b), angle ( $\theta$ ), Urey-Bradley (UB, S), dihedral angle ( $\varphi$ ), improper angle ( $\omega$ ), and the back bone torsional correction (CMAP,  $\varphi, \psi$ ). In the potential energy equation, the parameters  $K_b$ ,  $K_\theta$ ,  $K_{UB}$ ,  $K_\varphi$ , and  $K_\omega$  are force constants; and the terms  $b_0$ ,  $\theta_0$ ,  $S_0$ ,  $\varphi_0$ , and  $\omega_0$  are the respective equilibrium values.<sup>31</sup> All the internal energy terms, except the dihedral angle, are harmonic meaning the energy is calculated as a function of deviation from the equilibrium values.<sup>6,31</sup>

$$\begin{aligned}
 U_R = & \underbrace{K_b (b - b_0)^2}_{\text{bonds}} + \underbrace{K_\theta (\theta - \theta_0)^2}_{\text{angles}} + \underbrace{K_{UB} (S - S_0)^2}_{\text{Urey-Bradley}} \\
 & + \underbrace{K_\varphi [1 + \cos n\varphi - \delta]}_{\text{dihedrals}} + \underbrace{K_\omega (\omega - \omega_0)^2}_{\text{impropers}} \quad (1) \\
 & + \underbrace{\epsilon_{ij}^{min} \frac{R_{ij}^{min}}{r_{ij}}^{12} - 2 \frac{R_{ij}^{min}}{r_{ij}}^6}_{\text{non-bonded pairs}} + \frac{q_i q_j}{4\pi\epsilon_0\epsilon r_{ij}} \\
 & + \underbrace{U_{CMAP}(\varphi, \psi)}_{\text{residues}}
 \end{aligned}$$

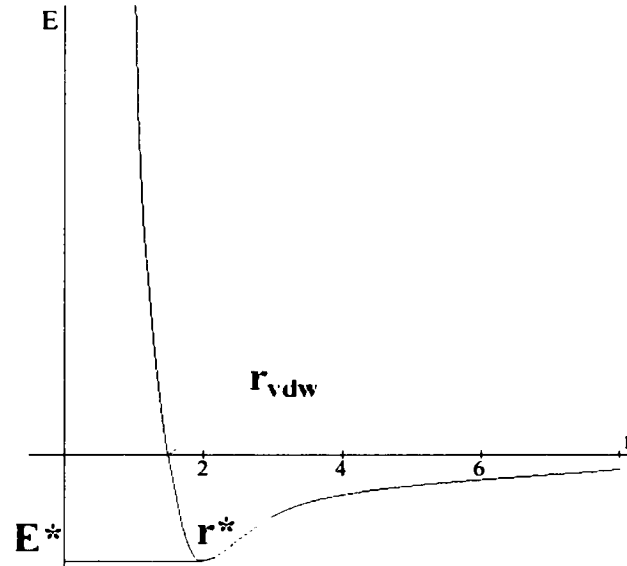
The bond potential represents the energy of a covalent bond between two atoms, which is calculated as the difference between the actual bond (b) length and the equilibrium bond length ( $b_0$ ).  $K_b$  is the force constant that determines the bond strength. The angle potential is a function of displacement of a bond angle between three consecutive atoms ( $\theta$ ) from its equilibrium value,  $\theta_0$ .<sup>6</sup> For three atoms separated by two

bonds,  $A - B - C$ , the Urey-Bradley potential is a function of the change in the distance between A and C (S) from its equilibrium value  $S_0$ . The Urey-Bradley potential is added in the CHRAMM force fields only in special cases and is used to restrain the motions of the bonds involved in the angle  $A - B - C$ .<sup>31</sup> The dihedral angle and the improper angle potentials are torsional terms that represent the steric rotational barriers in atoms separated by three covalent bonds (Figure 10).<sup>6, 33</sup> The dihedral angle potential is defined by four consecutive atoms  $I - J - K - L$  and is used to constrain rotation around a bond.<sup>6</sup> The dihedral angle potential is periodic with a force constant of  $K_\phi$ , where the  $\phi$  is the dihedral angle between the planes  $I, J, K$  and  $J, K, L$ ;  $\delta$  is the angle at which the potential is at its minimum; and  $n$  is the multiplicity of the dihedral angle as it is rotated  $360^\circ$ .<sup>31</sup> The improper angle potential applies for atoms  $I, J, L$  bonded to a central atom  $K$  and is used mainly to maintain the planarity of an atom (e.g. carbonyl carbon) and chirality of a tetrahedral atom (e.g.  $C_\alpha$  in proteins) in a molecular structure.<sup>31, 33</sup>  $K_\omega$  is the force constant for the improper angle potential, where the  $\omega$  is the angle between the planes  $I, J, K$  and  $J, K, L$  and  $\omega_0$  is the equilibrium improper dihedral angle.<sup>31</sup> CMAP is a torsional correction term used for protein backbones with energies significantly far away from the minimum energy or equilibrium values. The CMAP increases the accuracy of force fields and hence gives more accurate information about the dynamics of the protein.<sup>34</sup>



**Figure 10.** Proper dihedral and improper dihedral.<sup>6</sup>

The external energy term is the interactions between non-bonded atoms or atoms separated by 3 or more covalent bonds, and is expressed by the sum of van der Waals and electrostatic interactions.<sup>6</sup> The Lennard-Jones (LJ) term is used to express the repulsive and attractive forces involved in the van der Waals interaction (Figure 11). In the LJ term,  $\epsilon_{ij}^{min}$  is the energy between atoms  $i$  and  $j$  separated by distance,  $r_{ij}$ ; and  $R_{ij}^{min}$  is the distance at which the LJ term is at its minimum.<sup>31</sup> The attractive interaction occurs when  $r_{ij}$  is at a longer distance. At shorter distances, the repulsive interaction becomes dominant.  $r_{vdw}$  is the distance at which the LJ becomes zero and the attraction and repulsion forces are equal.<sup>6</sup> The attractive and repulsive potential are represented by  $-r_{ij}^{-6}$  and  $-r_{ij}^{-12}$ , where the power numbers indicate the optimum distance at which the attractive and repulsive interactions occur, respectively. The electrostatic potential is represented by the Coulomb equation, where  $qi$  and  $qj$  are the charges of atoms  $i$  and  $j$ , which are separated by a distance  $r_{ij}$ .  $\epsilon$  is the dielectric constant for the medium, relative to the permittivity of vacuum  $\epsilon_0$ .<sup>31</sup>



**Figure 11.** Lennard-Jones potential for van der Waals interactions. The well depth represents the LJ minimum energy  $E^*$  ( $\epsilon_{ij}^{min}$ ) and is located at  $r^*$  ( $R_{ij}^{min}$ ).<sup>6</sup>

When non-bonded terms are calculated, all possible pairs of atoms should be evaluated. Hence, calculating the non-bonded terms is the most time consuming part of the molecular dynamics simulation. In order to reduce the computation time, only those interactions that are within or at a cutoff distance are kept in the non-bonded list.<sup>6</sup> However, the cutoff distance has the disadvantage of creating discontinuity in the energy function at the cutoff distance and significantly affect the computational results.<sup>32</sup> The discontinuity in the energy function is caused by atom pairs that are close to the cutoff boundary. The atom pairs can be within the cutoff distance at one time step and contribute to the potential energy. In another time step, they can be out of the cutoff distance, even with a very limited movement of these atoms, and not contribute to the potential energy. Switching and shifting functions are used to avoid this discontinuity.<sup>33</sup>

The switching function applies a second cutoff distance. The potential energy is calculated without modification for distances within the first cutoff distance; and is gradually switched to zero between the first and last cutoff distance. The shifting function modifies the entire potential energy surface and shifts the potential at the cutoff distance to zero.<sup>6</sup> The long range electrostatic interactions are important, and ignoring them by applying the cutoff distance can result in reduced accuracy. Ewald summation is a method used to calculate the long range electrostatic interactions in molecular systems with periodic boundary conditions.<sup>35</sup>

### **Molecular Dynamics Simulations**

Molecular dynamics simulation is a computational method used to study time dependent behavior of molecular systems such as the conformational changes of proteins and nucleic acids. It is also used in the determination of structures from x-ray crystallography and NMR experiments, and in drug design. Molecular dynamics simulation is based on Newton's second law of motion,  $F = ma$ , where  $F$  is the force exerted on a particle with mass  $m$  and acceleration  $a$ . Integrating the equation of motion yields the change in the positions, velocities and accelerations of the particles of the system versus time (eq 2), known as the trajectory. The trajectory is started from the initial positions of all the atoms, the initial distribution of velocities, and acceleration. The initial positions are usually obtained from x-ray crystal or NMR structures and are used in the calculation of the potential energy. The gradient of the potential energy function is the force  $F$  and is used to determine the acceleration. The initial velocities are

randomly selected from a Maxwell-Boltzmann distribution at a relatively low temperature.<sup>6</sup>

There are different methods for integrating the equation of motion. The most commonly used integration method is known as the Verlet algorithm. In the Verlet algorithm, positions and accelerations at time  $t$  and positions at time  $t - \delta t$  are used to calculate new positions at time  $t + \delta t$ . The Verlet algorithm (eq 4) is derived from the following two equations (eq 2, 3).

$$r(t + \delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 \quad (2)$$

$$r(t - \delta t) = r(t) - v(t)\delta t + \frac{1}{2}a(t)\delta t^2 \quad (3)$$

$$r(t + \delta t) = 2r(t) - r(t - \delta t) + a(t)\delta t^2 \quad (4)$$

The Verlet algorithm is straightforward but it does not use velocities. Another method of integration, in which the velocities are explicitly calculated, is known as the Verlet leap-frog algorithm. In this method, the velocities are first calculated at time  $t + 1/2\delta t$  (eq 5), and these are used to calculate the positions,  $r$ , at time  $t + \delta t$  (eq 6).<sup>36</sup> Thus, “the velocities leap over the positions, and then the positions leap over the velocities”.<sup>6</sup> The velocity at time  $t$  is calculated by eq 7.<sup>36</sup> The Verlet leap-frog algorithm was used in this study.

$$v(t + \frac{1}{2}\delta t) = v(t - \frac{1}{2}\delta t) + a(t)\delta t \quad (5)$$

$$r(t + \delta t) = r(t) + v(t + \frac{1}{2}\delta t)\delta t \quad (6)$$



$$v(t) = \frac{1}{2} v(t - \frac{1}{2}\delta t) + \frac{1}{2} v(t + \frac{1}{2}\delta t) \quad (7)$$

The detailed procedure for running the molecular dynamics simulations is explained below.

## Initialization

NMR structures of G<sub>A</sub>95 and G<sub>B</sub>95 from the Protein data bank, with PDB codes of 2kdl and 2kdm, respectively, were used as the initial structures. The molecular dynamic simulations were performed in the CHARMM (CHARMM27 force field), under microcanonical conditions, i.e., constant number of atoms (N), volume (V), and energy (E).<sup>6, 31</sup> Before starting the molecular dynamics simulation, energy minimization of the NMR structures is necessary to remove any strong van der Waals interactions that might cause local structural distortion and result in an unstable simulation. Each structure was minimized with constraints.

After minimization, the structures were placed into a pre-equilibrated water box. The protein structure is placed into the center of the water box and any water molecules that overlap the protein are removed.<sup>6</sup> The water model that was used in this study was TIP3P and the shape of the water box was a truncated octahedron (Figure 12).<sup>37</sup> The TIP3P is a simple model in which the interaction between the three atoms of water is represented by van der Waals with point charges at their center. The model is designed to mimic the real water molecule.<sup>33</sup> The number of TIP3P water molecules that were used for G<sub>A</sub>95 and G<sub>B</sub>95 were slightly different; 6484 for G<sub>A</sub>95 and 6492 for G<sub>B</sub>95. A

van der Waals switching function between 6 and 8 Å, an electrostatic shifting function cutoff of 8 Å and a 10 Å cutoff distance for the non-bonded list were used.



**Figure 12.** Solvation of G<sub>A</sub>95 and G<sub>B</sub>95 in truncated octahedron water box. Pictures generated using the Visual Molecular Dynamics (VMD) program.<sup>38</sup>

Periodic boundary conditions were applied to prevent the water molecules from diffusing away from the protein and to limit the number of water molecules needed, which reduces the simulation time. The periodic boundary condition is created by replicating the water box in all directions. During the simulation, if a water molecule leaves the right side of the primary water box, then its image enters the left side. This way, the periodic boundary condition insures that the protein is solvated throughout the simulation time.<sup>36</sup> The water molecules were minimized to readjust the water molecule to the presence of the minimized protein structure.<sup>6</sup>

## Equilibration

The equilibration step involves assigning random velocities to each atom of the solvated structures at 60 K. The proteins were simulated by increasing the temperature by 50 K every 0.2 ps until the desired temperature was reached. The G<sub>A</sub>95 and G<sub>B</sub>95 were simulated at different temperatures: 300 K, 350 K, 400 K, 450 K, 500 K, and 550 K; for 10000 time steps. Each step took 0.002 ps and the intervals were run 5 times giving a total equilibration time of 100 ps. In the first 20 ps, all atoms were restrained to their positions in the corresponding energy-minimized NMR structures. After the first equilibration, the constraints were removed and the structures were further simulated for 80 ps to make sure that a stable structure was attained at the desired temperature. During the equilibration, a switching function was used for van der Waals interactions with a cutoff distance between 8 and 11 Å. Long range Electrostatic interactions with a cutoff distance of 11 Å were included and the Ewald summation was used to add on electrostatic interactions that were longer than 11 Å. All bond lengths involving hydrogen were constrained using the SHAKE algorithm.<sup>37</sup>

## Production Dynamics

Production dynamics is the final step of simulation, in which the equilibrated structures are simulated further for the desired time length. During this production phase, the trajectory was saved for later analysis. Thermal unfolding simulations at different temperature were run in intervals of 10000 time steps. Each step took 0.002 ps and the intervals were run 500 times, giving a total simulation time of 10 ns. The non-bonded

terms were treated using the same conditions as the equilibration dynamics. The SHAKE algorithm was also used to constrain the bonds involving hydrogen.

### Analysis of the Molecular Dynamics Simulations

The coordinates that are saved during the molecular dynamics simulations were used for analysis.<sup>6</sup> The average structure was obtained by averaging the geometries from every 100 steps of the simulation. The average structure was minimized to remove any artifacts from the averaging procedure. The trajectory of each production dynamics was visualized in Visual Molecular Dynamics (VMD). Time dependent properties such as the root mean square deviation (RMSD), radius of gyration (RGYR) and long range contact distances were compared using the XMGRase graphics program.

RMSD is the average distance between atoms in two conformations of the same molecule, calculated after superimposing the two conformations.

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N (r_i^{(j)} - r_i^{(k)})^2} \quad (8)$$

In eq 8,  $r_i^{(j)}$  is the coordinate of the  $i$ -th atom in conformation ( $j$ ) and  $r_i^{(k)}$  is the coordinate of the  $i$ -th atom in conformation ( $k$ ).  $r_i^{(j)} - r_i^{(k)}$  is the distance for atom  $i$  between conformations ( $j$ ) and ( $k$ ).  $N$  is the number of atoms in the molecule. For a more precise calculation of the RMSD, all atoms in the conformations or selected atoms should be optimally aligned. RMSD was used to compare structures, where one of the conformations is used as the reference structure.<sup>33</sup> Two conformations of a molecule are

considered to be similar if their RMSD is  $< 3.5 \text{ \AA}$ . In molecular dynamics simulations, the structure under study is said to be stable and folded if its RMSD, relative to the reference structure, is  $< 3.5 \text{ \AA}$  during the entire trajectory. In this study, the G<sub>A</sub>95 and G<sub>B</sub>95 NMR structures were used as the reference structure in all RMSD calculations. The C<sub>α</sub> RMSD between the room temperature average structures and the NMR structures were determined using only the C<sub>α</sub> backbone atoms. A C<sub>α</sub> RMSD versus time trajectory was calculated by determining the RMSDs between the geometries at certain time steps and the NMR structure. The RMSD trajectory for the room temperature simulation was compared to those for the unfolding simulations. C<sub>α</sub> RMSDs were calculated in CHARMM and graphically visualized in XMGRace.

RGYR is the average distance of the atoms from the center of mass of the molecule  $r_{com}$  (eq 9). RGYR is a way to estimate the size of a molecule. The smaller the RGYR, the more compact the molecule.<sup>33</sup>

$$RGYR = \sqrt{\frac{1}{N} \sum_{i=1}^N r_i - r_{com}}^2 \quad (9)$$

CHARMM was used to calculate the RGYR of the simulated structures versus time. The room temperature graph was compared to those for the unfolding simulations using XMGRace. The RGYRs were also compared with the calculated RMSDs.

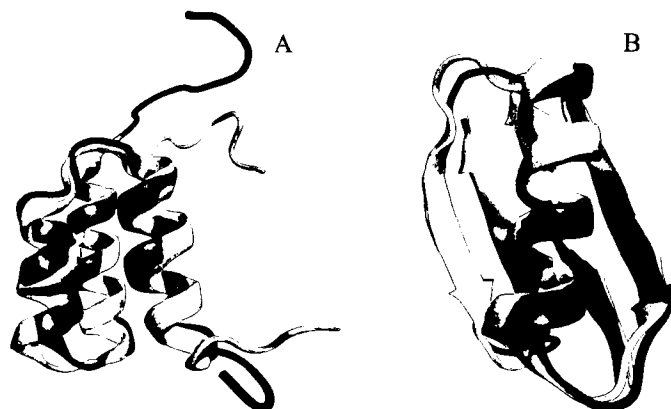
The most important part of the analysis was determining the long range interactions; these are very vital in determining the three dimensional structures of the proteins as they are involved in the nucleation process during protein folding.<sup>26, 39</sup> Long

range interactions were first determined using 9 or more amino acids in the primary structure. However, the list of long range interactions in G<sub>B</sub>95 did not include an interaction between Tyr45 and Phe52, a key interaction in stabilizing the 4 $\beta$ + $\alpha$  fold.<sup>25</sup> When 7 or more amino acids were used as a cutoff, the Ty45r-Phe52 interaction was included in the list. Choosing a 7 or more amino acid separation was important, because it is large enough to preclude any side chain interactions within a helix structure. Thus, most long range interactions will be in the tertiary structure rather than the secondary structure. The dmat option in the CHARMM was used to calculate the long range interactions in the room temperature average structures by setting the cutoff distance as 6.5 Å. The dmat calculates the distance between the centers of mass of the residues. The output from the dmat option provides the list of contacts that are within or at 6.5 Å. Then, an awk program was used to obtain only the list of long range interactions that are 7 or more residues apart. The long range contacts in the room temperature and thermally unfolded trajectories were calculated using CHARMM and an in-house fortran program written by Dr. Jennifer Poutsma. The coordinates of each long range contact was determined at every ps. This fortran program uses the atom to atom distances to calculate an average distance between the residues. The results were visualized using XMGRase.

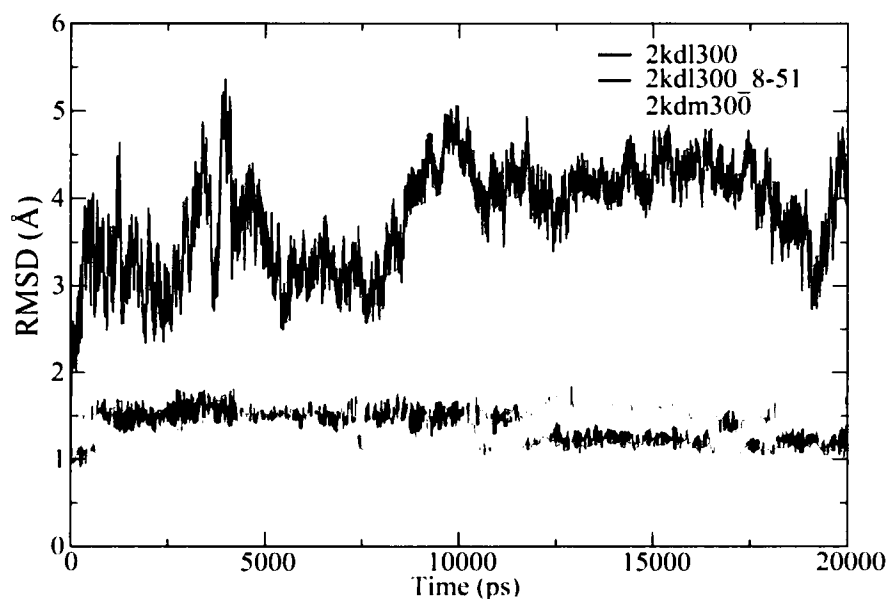
## CHAPTER III

### RESULTS AND DISCUSSION

The codes for the G<sub>A</sub>95 and G<sub>B</sub>95 300 K simulations are 2kdl300 and 2kdm300, respectively; where 2kdl and 2kdm are PDB codes of G<sub>A</sub>95 and G<sub>B</sub>95 and the 300 refers to the temperature in Kelvin. The average structures of G<sub>A</sub>95 and G<sub>B</sub>95, which were calculated from the room temperature 20 ns simulations, were aligned with the minimized NMR structures to calculate their RMSDs (Figure 13). The C<sub>α</sub>-RMSD of 2kdl300 was calculated two different ways: 1) the entire amino acid sequence was used and 2) the disordered residues from the NMR structures, 1- 8 and 52-56, were excluded. The all amino acid C<sub>α</sub>-RMSD (4.2 Å) was higher than the shortened sequence RMSD (1.4 Å). In the RMSD versus time plot (Figure 14), the all amino acid RMSD graph was very high and fluctuating, as opposed to the graph for the shortened sequence, which has only small fluctuations and was almost flat. This result indicated that the tails were disordered, but the 3  $\alpha$ -helices were stable and folded during the simulation, which was in agreement with experiment. The RMSD between the 2kdm300 average structure and NMR structure was 1.3 Å and the RMSD versus time graph was flat with little fluctuation which indicates this structure was stable and folded (Figure 14). Both 2kdl300 and 2kdm300 average structures were used as reference structures for comparison to the thermally unfolded structures.



**Figure 13.** Alignment of 300K minimized average structure (light gray) of (A) 2kd1300 and (B) 2kdm300, with minimized NMR structure (black) in VMD.

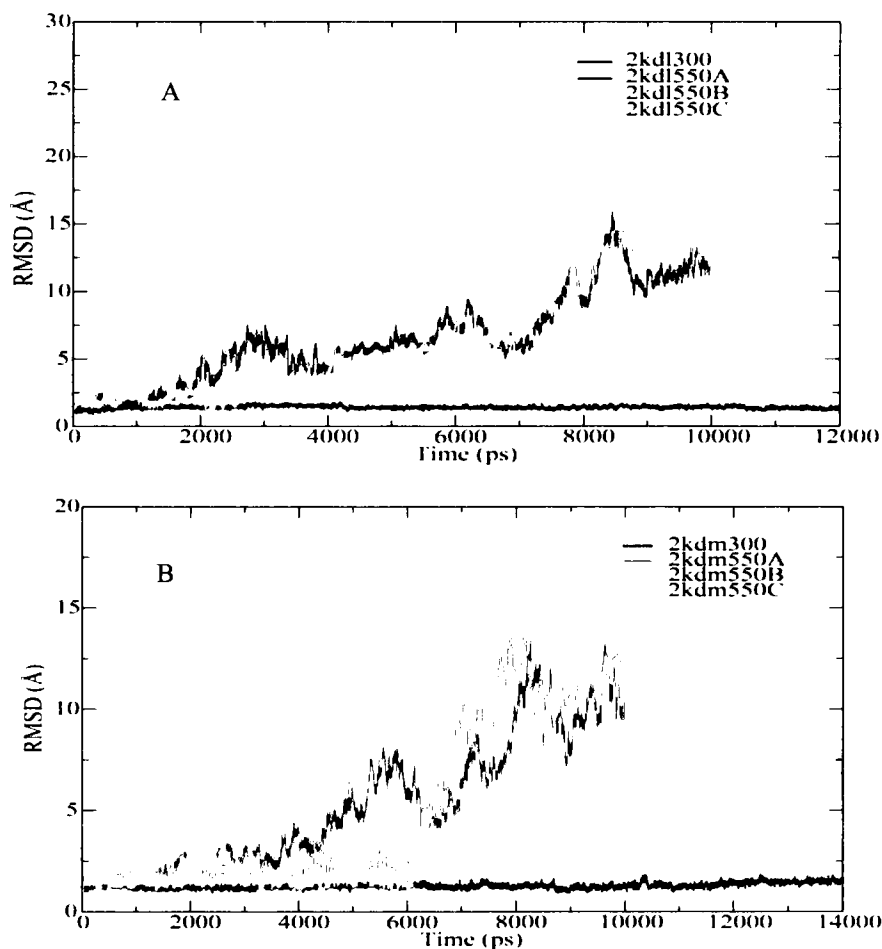


**Figure 14.** RMSD vs. time graph of 2kd1300 (black = all amino acids, dark gray = residues 8-51) and 2kdm300 (light gray).

Unfolding simulations were performed by first increasing the temperature to its final value, 350 K, 400 K, 450 K, 500 K, and 550 K, in 50 K increments every 0.2 ps. Dynamics were then run on the system for an additional 10ns. RMSDs of the proteins

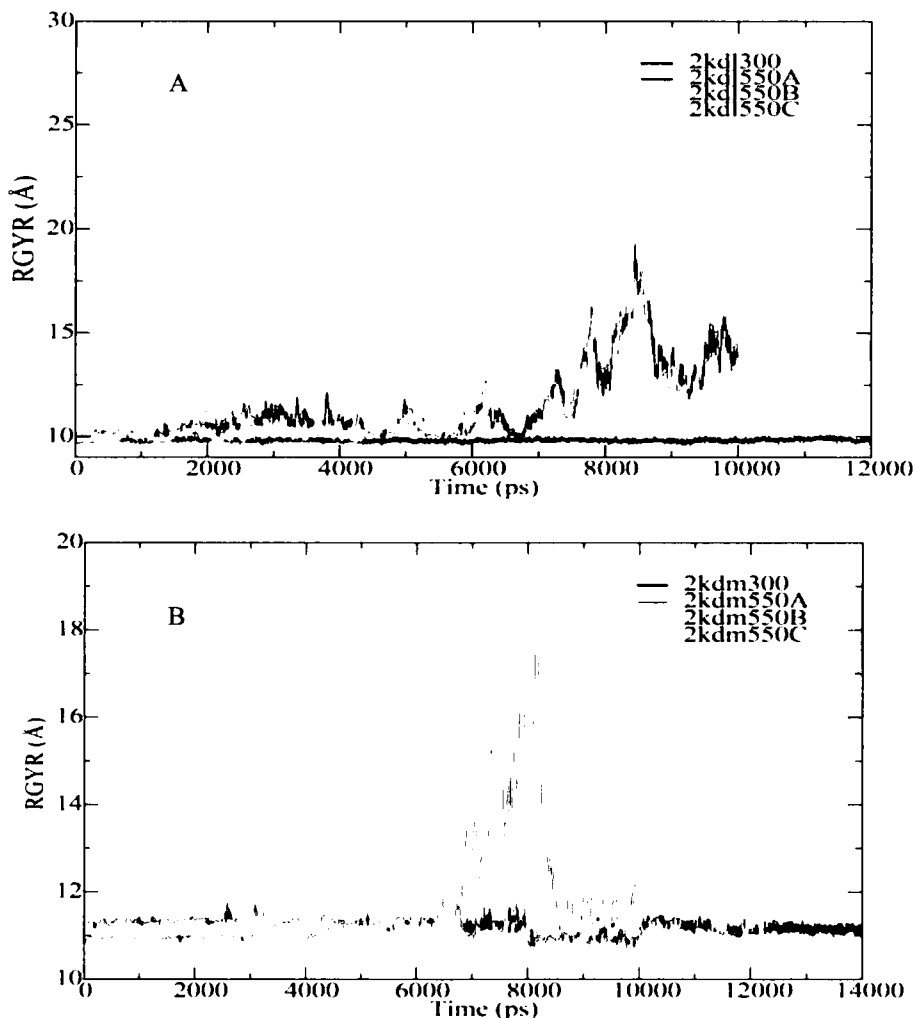


were calculated and compared with the RMSDs of 2kd1300 and 2kdm300. As with 2kd1300, only residues 8-51 were included in the RMSD calculation. The RMSD results show that  $G_A95$  and  $G_B95$  only unfolded at 550 K. Thus, all future unfolding simulations were run at 550K for 10 ns. The codes for the  $G_A95$  and  $G_B95$  550 K simulations are 2kd1550 and 2kdm550, respectively. Three independent unfolding simulations were performed for each protein by randomly choosing the initial velocities. The codes for the three samples of  $G_A95$  are 2kd1550A, 2kd1550B, and 2kd1550C. The three samples of  $G_B95$  are also coded as 2kdm550A, 2kdm550B, and 2kdm550C.  $C\alpha$ -RMSDs of all the 2kd1550 and 2kdm550 samples were  $> 10 \text{ \AA}$  at some point during the simulation (Figure15). The higher RMSDs indicated that the proteins are unfolded and this assessment was confirmed by visualization of the trajectories using VMD (see below). Note that while a small amount of secondary or tertiary structure may remain, most of it has been significantly disrupted.



**Figure 15.** RMSDs of samples of (A) 2kdl550 and (B) 2kdm550 as a function of time, relative to RMSDs of 2kdl300 and 2kdm300.

RGYR versus time of the 2kdl550 and 2kdm550 samples were also calculated and compared with the RGYR of the room temperature structures (Figure 16). All the unfolded proteins, except for the first sample of 2kdm550, had higher RGYR values than that of 2kdl300 and 2kdm300, indicating that they were less compact than the reference structures. This result is not surprising as the size of a protein was expected to increase as it unfolded.

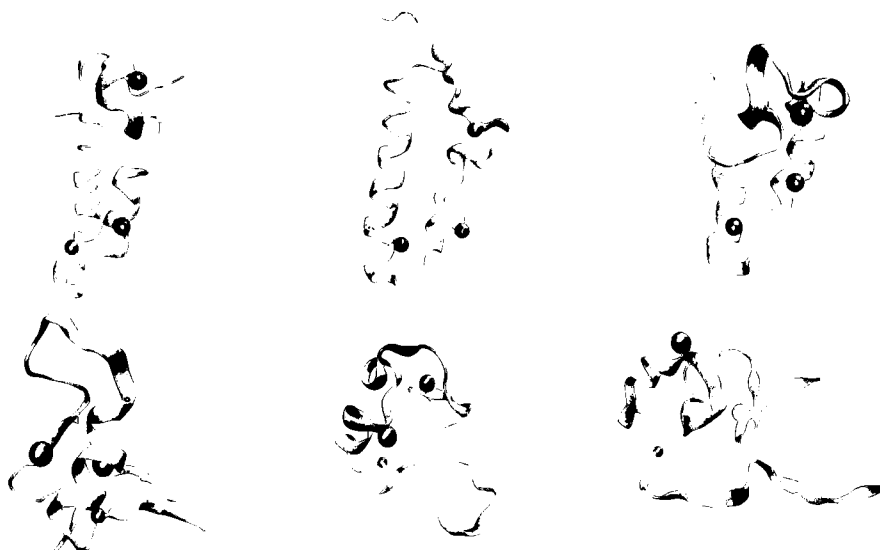


**Figure 16.** RGYRs of samples of (A) 2kdl550 and (B) 2kdm550 as a function of time, relative to RGYRs of 2kdl300 and 2kdm300.

The RMSD of each sample was compared with its RGYR. In 2kdl550, the RMSDs of all the samples became smaller near the end of each simulation, but the concomitant reduction in the RGYR was much larger. The lower RGYR value indicated that the protein became more compact at the end of the simulation. For 2kdm550, 2kdm550A was even more compact than its reference structure, but it still had a high RMSD. The 2kdm550B was also very compact at the end of the simulation, but still had

a high RMSD. These results indicated that most of the unfolded structures were compact, meaning that they lacked native contacts and were more like molten globules than folded structures. Also, the increase in the RGYR (or RMSD) followed by a decrease suggested—the increase in size was necessary for unfolding, followed by the protein collapse to a molten globular structure.

The trajectories from the unfolding simulations of 2kd1550 and 2kdm550 were visualized in VMD. Snapshots were taken for the first trajectories of 2kd1550 and 2kdm550 at  $\sim 6$  ns, 7 ns and 10 ns to see the general behavior of the proteins as they unfolded (Figure 17). At  $\sim 6$  ns, G<sub>A</sub>95 started to unfold at its second and third helices. There was slight unfolding at the  $\alpha 1/\alpha 2$  turn and the section of  $\alpha 1$  connected to the  $\alpha 1/\alpha 2$  turn, but the  $\alpha 1$  helix returned back to its stable conformation at  $\sim 7$  ns. At  $\sim 10$  ns, the



**Figure 17.** Snapshots of unfolding trajectory of 2kd1550A (top) and 2kdm550A (below), at  $\sim 6$  ns, 7 ns and 10 ns. The balls show the three non-identical amino acids in G<sub>A</sub>95 and G<sub>B</sub>95.

$\alpha 2$  helix partially refolded. The  $\alpha 2$  never completely unfolded during the simulation process. The trajectory indicated that only the third helix completely unfolded, but that interactions between the helices were broken. The  $\alpha 3$  helix moved away from  $\alpha 2$  helix, but the  $\alpha 2$  helix remained close to  $\alpha 1$  helix throughout the simulation except when it started to unfold at  $\sim 6$  ns. Trajectories in the second and third samples of 2kd1550 were also compared and some similarities and differences were observed. In the second sample, the  $\alpha 3$  helix started to unfold at  $\sim 6$  ns; then the  $\alpha 2$  helix and the section of  $\alpha 1$  connected to the  $\alpha 1/\alpha 2$  turn unfolded at  $\sim 7$  ns. At  $\sim 10$  ns, the three helices were unfolded with partial refolding of the  $\alpha 2$  helix. It was observed that the three helices moved away from each other during unfolding. The third sample was simulated for 12 ns and like the first sample, its  $\alpha 2$  and  $\alpha 3$  helices started to unfold at around 6 ns. At  $\sim 7$  ns, the  $\alpha 1$  helix also unfolded but the  $\alpha 2$  helix started to partially refold. At  $\sim 10$  ns, the  $\alpha 1$  and  $\alpha 3$  helices were unfolded but the  $\alpha 2$  helix was partially folded between 7 and 10 ns. The  $\alpha 2$  helix was close to  $\alpha 1$  but the  $\alpha 3$  was far from  $\alpha 2$ . After 10 ns, the helices were completely unfolded.

Surprisingly, residues 1-8 in the tail added onto the first helix in all unfolding simulations. This behavior was also seen in the dynamics simulation of  $G_{A88}$  by Lazim et al.<sup>28</sup> In addition, pictures of the trajectory from unfolding simulations of  $G_{A88}$  by Morrone et al also show the same behavior of the N-terminal.<sup>26</sup> All three studies used different force fields; hence the observation is not an artifact of the force field. Even more interestingly, the extension of the first helix was not seen in the simulation of  $G_{A95}$  or  $G_{A88}$  at room temperature.<sup>28</sup> However, it is unclear how the behavior of the tail fits in the unfolding pathway or if it is an important factor.

The unfolding of G<sub>B</sub>95 started at the turns between  $\alpha/\beta$ 2 and  $\alpha/\beta$ 3. At around 7 ns, the central helix was mostly unfolded and the  $\beta$ 1/ $\beta$ 2 and  $\beta$ 3/ $\beta$ 4 hairpins had moved away from the central helix but the strands were still in contact, though the  $\beta$ 1/ $\beta$ 4 distance had increased. At ~10 ns, the protein had completely unfolded, but the  $\beta$ 3/ $\beta$ 4 hairpin still remained. Some differences were observed for the other runs of G<sub>B</sub>95. In the second sample, the  $\beta$ 1/ $\beta$ 2 and  $\beta$ 3/ $\beta$ 4 hairpins started to move away from the central helix; however, unlike the first sample, the  $\beta$ 1/ $\beta$ 2 hairpin started to come apart at ~7 ns. Moreover, at ~10 ns, the  $\alpha$ -helix was partially recovered. In the third sample, which was simulated for 14 ns, the  $\beta$ 1/ $\beta$ 2 and  $\beta$ 3/ $\beta$ 4 hairpins were at their original positions until ~10 ns, when they started to move apart. The  $\beta$ 1/ $\beta$ 2 hairpin started breaking after 10 ns, followed by the separation of the  $\beta$ 3/ $\beta$ 4 hairpin. The  $\alpha$ -helix was stable until ~10 ns and then was only partially unfolded. This observation agrees with Lazim et al.'s result that the folding pathway of G<sub>B</sub>88 involved the separation of  $\beta$ 1/ $\beta$ 2 and  $\beta$ 3/ $\beta$ 4 hairpins by the central helix; and the  $\alpha$ -helix is partially conserved even after unfolding.<sup>28</sup> Morrone et al. also observed that the central  $\alpha$ -helix in G<sub>B</sub>88 was maintained throughout the unfolding process and that the  $\beta$ 1/ $\beta$ 2 and  $\beta$ 3/ $\beta$ 4 hairpins were separated from each other.<sup>26</sup>

The long range interactions were determined from the minimized NMR structures of 2kdl and 2kdm. The total number of long range interactions in G<sub>A</sub>95 was 18. G<sub>B</sub>95 had 41 long range interactions. Most of the long range contacts in G<sub>B</sub>95 contained similar residues, because, in a  $\beta$ -sheet structure, long range interactions could be formed between strands; whereas in a  $\alpha$ -helix structure, the long range interactions were only between helices. Only 16 of the contacts in G<sub>B</sub>95 were taken for analysis, by choosing those contacts involving the non-identical residues and the shortest contact for

a set of interactions between similar residues.  $G_{A95}$  and  $G_{B95}$  had only one common long range interaction between residue 16 and 30, i.e., Ala-Ile in  $G_{A95}$  and Ala-Phe in  $G_{B95}$ . The long range interactions versus time during the unfolding simulations of 2kdl550 and 2kdm550 were calculated and compared. The contacts were qualitatively analyzed from the distance vs. time graphs in XMGRase.

The minimum, maximum, and average distances of the long range interactions were determined for the 2kdl300 and 2kdm300 simulations. A cutoff distance of 6.5 Å was used to choose the long range interactions, however, most of the long range interactions had an average contact distance of  $> 6.5$  Å. This discrepancy could have been because the initial sets of long range interactions were determined from the minimized NMR structures. The average of the maximum distances was used as the cutoff point within which the long range interactions in the unfolded proteins were considered to be in contact. The average of the maximum distances for 2kdl300 was 10.30 Å, but the average of the maximum distances for 2kdm300 was 7.70 Å (Table 1 and 2). A cutoff point of 10 Å was taken in order to compare the two proteins under the same conditions.

**Table 1. Distances for the Long Range Interactions in 2kdl300**

Contacts	Avg. Contact (Å)	Min Contact (Å)	Max Contact (Å)
Ile6-Val39	7.08	5.22	10.03
Leu9-Val39	6.51	5.37	9.83
Ala12-Val39	6.65	5.37	10.23
Ala12-Val42	7.18	5.54	10.44
Ala12-Trp43	5.47	7.46	12.25
Ala16-Ile30	9.00	7.58	13.38
Ala16-Ile33	7.32	5.82	10.53
Ala16-Lys46	6.36	5.62	7.91
Ile17-Ile30	6.75	5.59	10.39
Glu19-Lys46	6.99	5.87	8.85
Leu20-Ile30	6.83	5.74	9.59
Thr25-Ile49	8.04	5.90	11.50
Ala26-Ile49	9.15	7.07	12.24
Tyr29-Leu45	9.51	6.78	12.28
Tyr29-Ile49	6.37	5.34	8.23
Leu32-Leu45	7.29	5.68	10.29
Ile33-Val42	5.78	4.96	7.88
Ile33-Leu45	6.78	5.42	9.47
			Avg. = 10.30

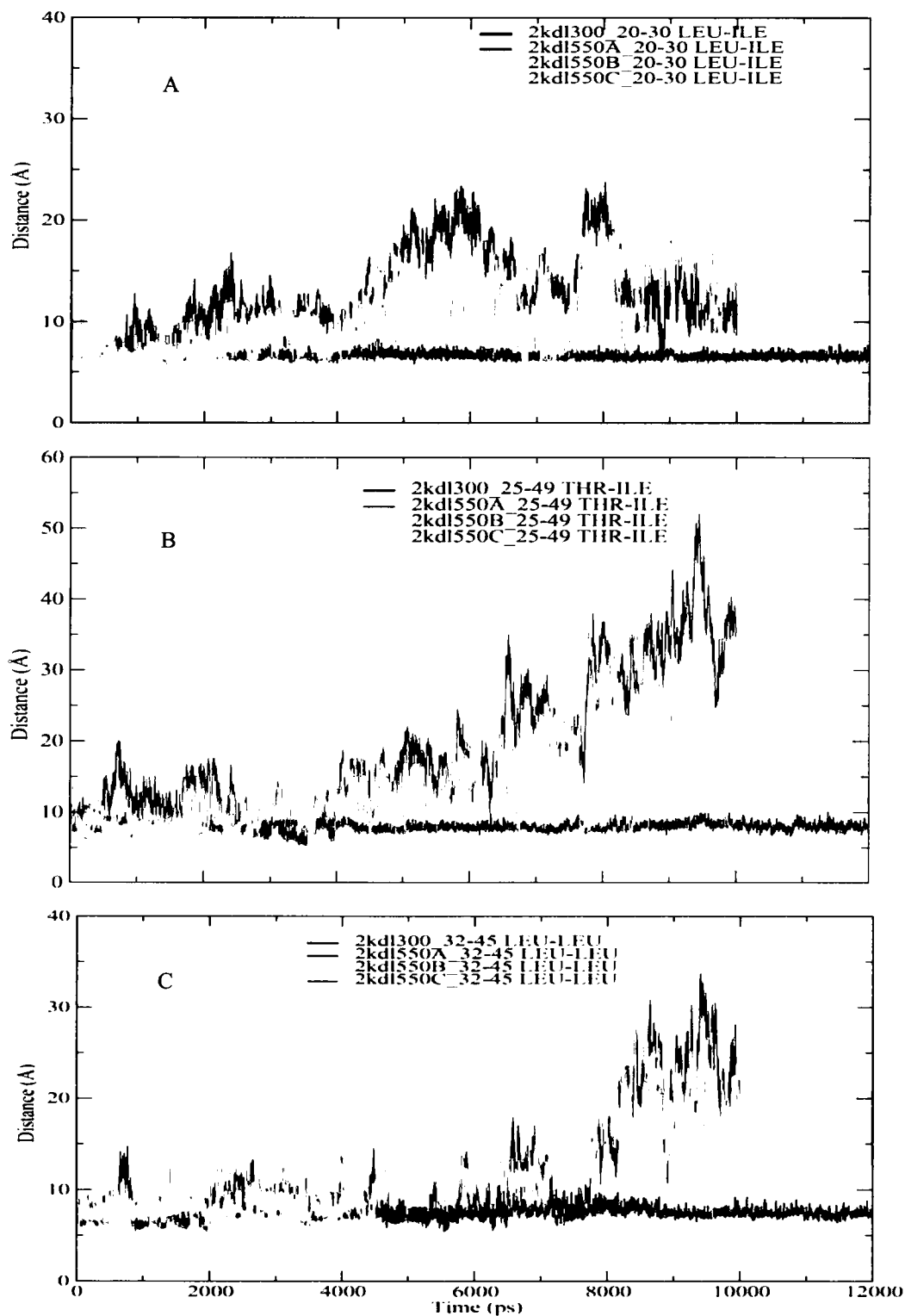
**Table 2. Distances for the Long Range Interactions in 2kdm300**

Contacts	Avg. Contact (Å)	Min Contact (Å)	Max Contact (Å)
Thr1-Ala20	5.36	4.62	10.91
Tyr3-Ala20	6.23	5.37	7.84
Tyr3-Ala26	6.66	6.03	8.01
Lys4-Ala16	6.03	5.35	7.17
Lys4-Thr51	6.10	5.48	7.15
Leu5-Phe30	5.77	5.28	6.44
Ile6-Thr53	5.41	4.93	6.27
Leu7-Glu14	5.75	5.03	6.77
Leu7-Val54	5.70	5.00	7.02
Ala16-Phe30	6.99	5.94	8.86
Lys18-Tyr29	6.89	5.98	8.92
Phe30-Phe52	7.14	6.40	8.45
Gly41-Thr55	6.39	5.45	9.21
Val42-Thr55	5.12	4.50	7.53
Thr44-Thr53	4.96	4.48	5.85
Tyr45-Phe52	6.10	5.53	6.87
			Avg. = 7.70

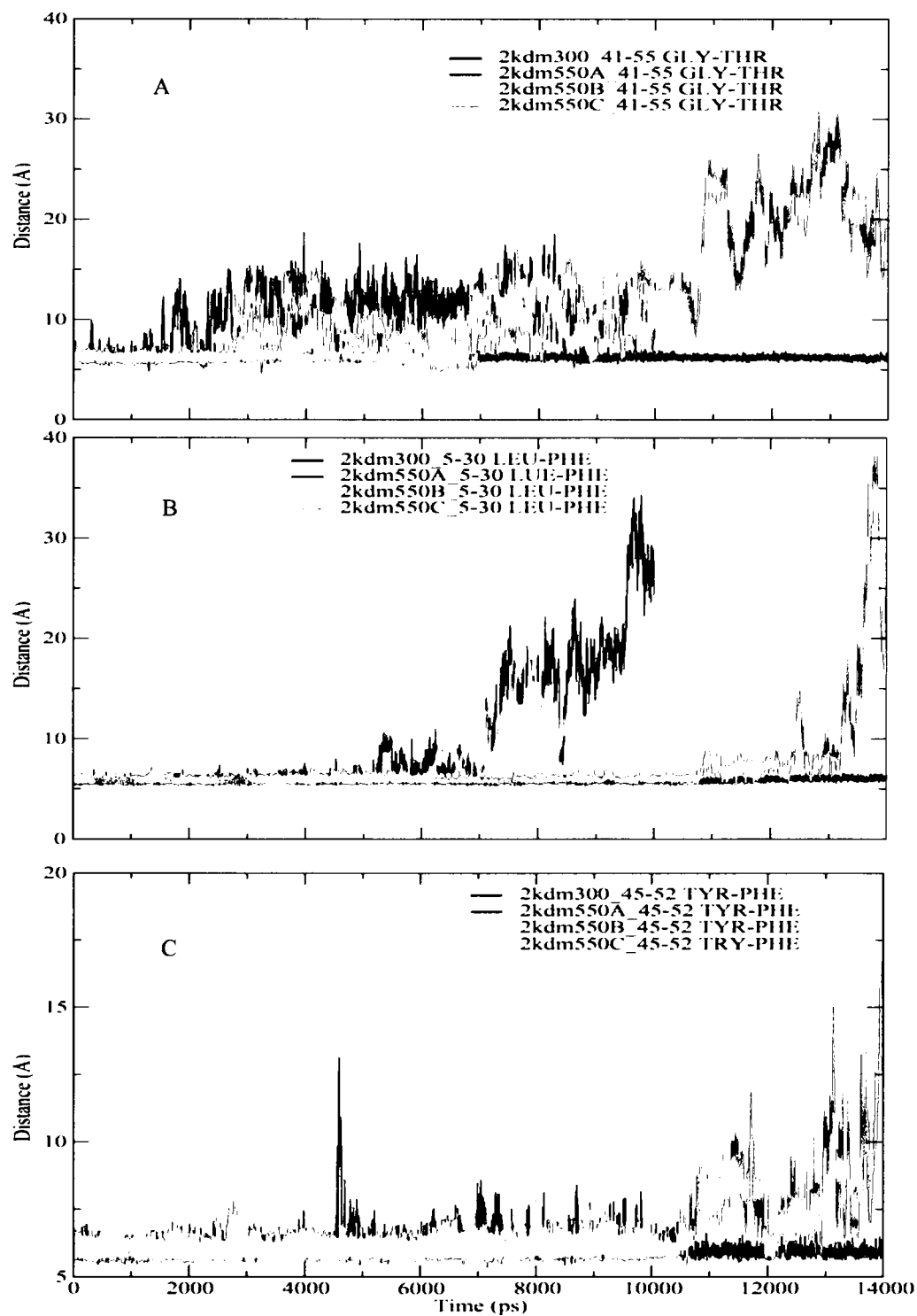


The contacts that were within 10 Å distance during the simulation were counted, by dividing the simulation time into 5 intervals (0-2 ns, 2-4 ns, 4-6 ns, 6-8 ns, and 8-10 ns). One snapshot per picosecond was used for this analysis; thus, the maximum number of contacts that a long range interaction could have at each interval was 2000. The long range interactions were classified as persistent (see definition in next paragraph) or not persistent. Only three graphs from each protein that represent the general behavior of the long range interactions are shown here (Figures 18, 19). Contacts Leu20-Ile30 in G<sub>A</sub>95 and Gly41-Thr55 in G<sub>B</sub>95, Figures 18A and 19A are an example of interactions that are persistent during the simulation but with fluctuating contact distances in the middle. Contact Thr25-Ile49 in G<sub>A</sub>95 is an example of an interaction that was not persistent during the unfolding simulation (Figures 18B). Interactions Leu32-Leu45 in G<sub>A</sub>95 and Leu5-Phe30 in G<sub>B</sub>95, Figures 18C and 19B, are interactions that are persistent but broke at the end of the simulation. Contact Tyr45-Phe52 in G<sub>B</sub>95 is an example of an interaction that is persistent with slight fluctuations in the distance (Figure 19C).

The long range interactions were classified as not persistent if the amino acids involved separated quickly during the simulation and the number of contacts in the last two or three intervals, usually 6-10 ns, were close or equal to zero. The long range interactions that remained for a significant portion of the simulation were classified as persistent interactions and could be separated into two different types: (1) consistent (C), those with a similar number of contacts in each interval or where the number of contacts decreased through time; (2) non-consistent (NC), those with an inconsistent number of contacts that reached zero or a minimum number of contacts in the middle of the simulation, but had contacts at the end of the simulation.



**Figure 18.**  $G_{A95}$  long range distance vs. time graphs of (A) Leu20-Ile30, (B) Thr25-Ile49, and (C) Leu32-Leu45.



**Figure 19.** G<sub>B</sub>95 long range distance vs. time graphs of (A) Gly41-Thr55, (B) Leu5-Phe30, and (C) Tyr45-Phe52.

The persistent contacts were further classified according to the interactions between the secondary structures. Although most of the contacts which are important in keeping the tertiary structure of G<sub>A</sub>95 were between the helices, there were some interactions between the disordered tails or turns and an  $\alpha$  helix. The long range interactions were classified into interactions between (1)  $\alpha$ 1 and  $\alpha$ 2, (2)  $\alpha$ 2 and  $\alpha$ 3, (3)  $\alpha$ 1 and  $\alpha$ 3, (4)  $\alpha$ 3 and the N-terminal tail, and (5)  $\alpha$ 1/ $\alpha$ 2 turn and  $\alpha$ 3. In G<sub>B</sub>95, many of the long range contacts were between the different  $\beta$  strands, as expected, and between  $\beta$ -strands and the  $\alpha$ -helix. The persistent contacts were classified into interactions between (1)  $\beta$ 1 and  $\beta$ 2, (2)  $\beta$ 3 and  $\beta$ 4, (3)  $\beta$ 1 and  $\beta$ 4, (4)  $\beta$ 1 and  $\alpha$ , (5)  $\alpha$  and  $\beta$ 4, (6)  $\beta$ 1 and  $\beta$ 3, (7)  $\beta$ 2 and  $\alpha$ , (8)  $\alpha$ / $\beta$ 3 turn and  $\beta$ 4. These classifications were used in the tables below to indicate the type of the long range interactions between secondary structures (SS). The persistent contacts in G<sub>A</sub>95 and G<sub>B</sub>95 were further analyzed in terms of the three non-identical (NI) residues.

Tables 3-6 show the total number of times the long range interactions were in contact, i.e., in which their distances was less than 10Å, for the three simulations of G<sub>A</sub>95 at 550K. The contacts have been grouped first, by whether or not they are persistent and then, by the type of secondary structure interaction. The first column of the tables shows the residue numbers and the name of the amino acids that form the long range interaction. The last two columns are used to indicate (1) the type of secondary structure interaction and if the interaction involves one of the non-identical residues (NI); and (2) if the long-lasting interactions are consistent (C) or non-consistent (NC). The type of secondary structure contact is specified by the number given and refers to the numbers given in the previous paragraph. The same classification method was used

for the G<sub>B</sub>95 Tables 9-11.

In 2kd1550A (Table 3), the first six contacts were those that were not persistent and the rest were persistent interactions. Not surprisingly, the contacts that were not persistent involved helix 3 which completely unfolded during the trajectory. All the long range contacts involving the non-identical residues remained during the simulation to some extent. Interactions between residues Ala16-Ile30, Ile17-Ile30, and Leu20-Ile30 were persistent but had an inconsistent number of contacts over the simulation. These contacts were between the  $\alpha$ 1 and  $\alpha$ 2 helices and contained the non-identical residues 20

**Table 3. Total Number of Contacts < 10 Å in 2kd1550A**

Contacts	0 - 10 ns	0 - 2 ns	2 - 4 ns	4 - 6 ns	6 - 8 ns	8 - 10 ns	SS/NI <sup>a</sup>	NC/C <sup>b</sup>
Ala12-Val42	4239	1281	1392	1506	60	0	3	
Ala12-Trp43	5151	1953	1336	1830	32	0	3	
Ala16-Lys46	2993	1427	983	583	0	0	3	
Glu19-Lys46	4563	1999	1931	633	0	0	3	
Thr25-Ile49	1886	441	1437	7	1	0	4	
Ala26-Ile49	1847	799	1048	0	0	0	4	
Ala16-Ile30	2641	579	685	238	1075	64	1, NI	NC
Ala16-Ile33	2445	1306	327	10	802	0	1,	NC
Ile17-Ile30	1837	801	646	19	180	191	1, NI	NC
Leu20-Ile30	2690	1682	586	142	0	280	1, NI	NC
Tyr29-Leu45	4322	1087	1928	686	621	0	2, NI	NC
Tyr29-Ile49	3424	1883	952	347	242	0	2	C
Leu32-Leu45	6189	1845	1630	1641	1073	0	2, NI	C
Ile33-Val42	6701	1996	1864	1821	1020	0	2	C
Ile33-Leu45	3517	1822	514	312	869	0	2, NI	NC
Leu9-Val39	5004	1507	1507	1846	144	0	3	NC
Ala12-Val39	4426	1745	1519	1021	141	0	3	C
Ile6-Val39	4355	1294	1341	1528	192	0	4	NC

<sup>a</sup> Secondary structure contacts, see text for explanation of numbers. NI indicates contact contains non-identical residue.

<sup>b</sup> C = consistent, NC = non-consistent.

and 30. Additional persistent interactions that contained the non-identical residue 45 were Tyr29-Leu45, Leu32-Leu45 and Ile33-Leu45. These contacts were between the  $\alpha 2$  and  $\alpha 3$  helices. Only the Leu32-Leu45 interaction had a consistent number of contacts. Other contacts between  $\alpha 1$  and  $\alpha 2$  and between  $\alpha 2$  and  $\alpha 3$  and not involving residues 20, 30, or 45 were also found to remain for a large portion of the simulation: Ala16-Ile33, Tyr29-Ile49, and Ile33-Val42. Even though there was a constant increase in RMSD and RGYR of 2kd1550A, the number of contacts in the long range interactions did not necessarily show the same behavior. Many of the interactions had an inconsistent number of contacts during the simulation.

In 2kd1550B (Table 4), the first six contacts in the table were not persistent. However, these were a different set of contacts than those observed for 2kd1550A and this could be due to the complete unfolding of the  $\alpha 2$  helix, as observed in the trajectory of 2kd1550B. The rest of the long range interactions were persistent throughout the simulation process. Interestingly, the long range interaction Leu32-Leu45, which had a large number of contacts throughout 2kd1550A was among those that broke early in 2kd1550B. The  $\alpha 1$ - $\alpha 2$  interactions Ala16-Ile30, Ile17-Ile30, and Leu20-Ile30, which consist of non-identical residues 20 and 30, were also persistent during this simulation. However, all three contacts were inconsistent in 2kd1550A, but contact Leu20-Ile30 was consistent in 2kd1550B. This interaction was unique because it was formed between two non-identical residues. Again, as with 2kd1550A, the long range interactions Tyr29-Leu45 and Ile33-Leu45 (contacts between  $\alpha 2$  and  $\alpha 3$  and containing the non-identical residue 45) were persistent. Similarly, Ile33-Val42, which was formed between  $\alpha 2$  and  $\alpha 3$  and did not involve a non-identical residue, was also persistent. Interactions between

**Table 4. Total Number of Contacts < 10Å in 2kd1550B**

Contacts	0 - 10 ns	0 - 2 ns	2 - 4 ns	4 - 6 ns	6 - 8 ns	8 - 10 ns	SS/NI <sup>a</sup>	NC/C <sup>b</sup>
Ile6-Val39	245	245	0	0	0	0	4	
Ala12-Val39	2695	1882	397	323	93	0	3	
Thr25-Ile49	2991	1666	1146	125	54	0	4	
Ala26-Ile49	2693	1603	1053	0	7	30	2	
Tyr29-Ile49	4218	2000	1898	217	103	0	2	
Leu32-Leu45	3645	1708	1361	478	98	0	2, NI	
Ala16-Ile30	2469	1579	541	147	202	0	1, NI	NC
Ala16-Ile33	4277	1910	741	1128	498	0	1	NC
Ile17-Ile30	5156	1937	1782	655	780	2	1, NI	NC
Leu20-Ile30	6173	1959	1786	1140	1139	149	1, NI	C
Tyr29-Leu45	6052	1477	1854	1921	800	0	2, NI	NC
Ile33-Val42	5367	1977	1977	351	583	479	2	NC
Ile33-Leu45	3914	1954	1524	199	235	2	2, NI	NC
Leu9-Val39	2650	1946	127	320	51	206	3	NC
Ala12-Val42	4269	1951	397	1234	662	25	3	NC
Ala12-Trp43	2856	1996	664	30	177	19	3	NC
Ala16-Lys46	3891	2000	1623	57	211	0	3	NC
Glu19-Lys46	1834	1278	0	3	19	534	3	NC

<sup>a</sup> Secondary structure contacts, see text for explanation of numbers. NI indicates contact contains non-identical residue.

<sup>b</sup> C = consistent, NC = non-consistent.

Leu9-Val39 and Glu19-Lys46 were among those which were formed between  $\alpha 1$  and  $\alpha 3$  helices. The two contacts did not contain non-identical amino acids but had significant number of contacts at the end of the unfolding simulation. Comparing the number of contacts with RMSD and RGYR of 2kd1550B, the numbers were not in agreement with the RMSD and RGYR because only a few interactions had a consistent decrease in the number of contacts as expected with increasing RMSD and RGYR. However, most of the persistent interactions had an inconsistent number of contacts over time.

The 2kd1550C simulation was run for a total of 12 ns. In sharp contrast to the

2kdl550A and 2kdl550B trajectories to the 10 ns simulation, only two interactions had fewer than 1000 contacts during the (2-4 ns) interval and many were close to 2000 contacts. This implied that during simulation C the protein began to unfold at a later time than in the previous two simulations. To make sure that the protein was completely unfolded, the simulation was run for an additional 2ns, giving a total time of 12 ns. The interactions were classified into two categories: interactions between 0-10 ns (2kdl550C1), and interactions between 2-12 ns (2kdl550C2). The data for 2kdl550C1 (Table 5), indicated that six interactions were not persistent. Unlike 2kdl550A and 2kdl550B, these interactions included contacts between  $\alpha 1$  and  $\alpha 2$ ; and this was in agreement with the observation in its trajectory that the  $\alpha 1$  helix was unfolded during the simulation. One of these interactions, Ile33-Leu45, contained the non-identical residue 45 (persistent in A, but not persistent in B). Moreover, the contact Ile33-Val42, which was persistent in 2kdl550A and 2kdl550B, had a very small number of contacts in the 8-10 ns interval, but was not persistent.

Like in 2kdl550A and 2kdl550B, the interactions between Ala16-Ile30, Ile17-Ile30, and Leu20-Ile30 were among those that were persistent. In 2kdl550C1, the interactions between  $\alpha 2$  and  $\alpha 3$  were persistent with a significant number of contacts in the 8-10 ns interval. These interactions are between Tyr29-Leu45 and Leu32-Leu45, which contain the non-identical residue Leu45; and between Tyr29-Ile49, which had the largest number of contacts during the last interval. The interaction between Glu19-Lys46, which is located between the  $\alpha 1$  and  $\alpha 3$  helices, had a significant number of contacts until the last interval, as was seen in the 2kdl550B simulation. Interactions between Thr25-Ile49 and Ala26-Ile49 occur between the  $\alpha 1/\alpha 2$  turn and  $\alpha 3$  helix. These



interactions had a significant number of contacts only in the third simulation.

**Table 5. Total Number of Contacts < 10Å in 2kd1550C1**

Contacts	0 - 10 ns	0 - 2 ns	2 - 4 ns	4 - 6 ns	6 - 8 ns	8-10 ns	SS/NI <sup>a</sup>	NC/C <sup>b</sup>
Ala12-Val42	5508	1996	1864	1621	27	0	3	
Ala12-Trp43	4733	1998	1948	787	0	0	3	
Ala16-Ile33	4149	2000	1918	205	26	0	1	
Ala16-Lys46	5074	2000	1842	1232	0	0	3	
Ile33-Val42	4737	2000	1981	728	3	25	2	
Ile33-Leu45	4683	1990	1493	1181	19	0	2, NI	
Ala16-Ile30	4104	1759	1188	820	296	41	1, NI	NC
Ile17-Ile30	6696	1993	1985	1875	843	0	1, NI	NC
Leu20-Ile30	4014	2000	1701	242	0	71	1, NI	NC
Tyr29-Leu45	3998	1198	1773	920	7	100	2, NI	NC
Tyr29-Ile49	7305	1966	1509	1756	375	1699	2	NC
Leu32-Leu45	3964	1890	1583	368	0	123	2, NI	NC
Leu9-Val39	6037	1969	1915	1414	739	0	3	NC
Ala12-Val39	5703	1924	1993	1300	486	0	3	NC
Glu19-Lys46	4944	2000	1800	991	14	139	3	NC
Ile6-Val39	4526	1432	1765	1002	326	1	4	NC
Thr25-Ile49	4014	1927	716	761	314	296	5	NC
Ala26-Ile49	4558	1686	458	1028	388	998	5	NC

<sup>a</sup> Secondary structure contacts, see text for explanation of numbers. NI indicates contact contains non-identical residue.

<sup>b</sup> C = consistent, NC = non-consistent.

Upon changing the simulation timescale to 2-12 ns (2kd1550C2, Table 6), eleven interactions, instead of 6, were considered short-lived. This sample has the largest number of contacts that were not persistent. This result was most likely due to the longer simulation time. The trajectory for 2kd1550C showed that at ~ 10-12 ns, the protein was completely unfolded. In fact, the only interactions with contacts in the 10-12 ns interval were Leu20-Ile30, Tyr29-Ile49, Thr25-Ile49, and Ala26-Ile49. The interactions that had zero contacts in the 8-10 ns and 10-12 ns intervals included the non-identical residue-

containing interactions: Ala16-Ile30 (has 41 contacts in 8-10 ns interval), Ile17-Ile30, and Ile33-Leu45. In all the other samples, Ala16-Ile30 and Ile17-Ile30 were persistent.

**Table 6. Total Number of Contacts < 10Å in 2kdI550C2**

Contacts	2 - 12 ns	2 - 4 ns	4 - 6 ns	6 - 8 ns	8 -10 ns	10 -12 ns	SS/NI <sup>a</sup>	NC/C <sup>b</sup>
Ile6-Val39	4526	1765	1002	326	1	0	4	
Leu9-Val39	6037	1915	1414	739	0	0	3	
Ala12-Val39	5703	1993	1300	486	0	0	3	
Ala12-Val42	5508	1864	1621	27	0	0	3	
Ala12-Trp43	4733	1948	787	0	0	0	3	
Ala16-Ile30	4104	1188	820	296	41	0	1, NI	
Ala16-Ile33	4149	1918	205	26	0	0	1	
Ala16-Lys46	5074	1842	1232	0	0	0	3	
Ile17-Ile30	6696	1985	1875	843	0	0	1, NI	
Ile33-Val42	4756	1981	728	3	25	19	2	
Ile33-Leu45	4683	1493	1181	19	0	0	2, NI	
Leu20-Ile30	4049	1701	242	0	71	35	1, NI	NC
Tyr29-Leu45	3998	1773	920	7	100	0	2, NI	NC
Leu32-Leu45	3964	1583	368	0	123	0	2, NI	NC
Tyr29-Ile49	7495	1509	1756	375	1699	190	2	NC
Glu19-Lys46	4941	1800	991	14	139	0	3	NC
Thr25-Ile49	4035	716	761	314	296	21	5	NC
Ala26-Ile49	4589	458	1028	388	998	31	5	NC

<sup>a</sup> Secondary structure contacts, see text for explanation of numbers. NI indicates contact contains non-identical residue.

<sup>b</sup> C = consistent, NC = non-consistent.

The interaction Leu20-Ile30 that contains the two non-identical residues had contacts until the end of the simulation. However, the number of contacts was not as significant as for the other samples which, again, could be due to the longer simulation times. The interactions Tyr29-Leu45 and Leu32-Leu45 which were between the  $\alpha 2$  and  $\alpha 3$  helices and had a significant number of contacts in the 2kdI550C1, had no contacts in the 10-12 ns interval in 2kdI550C2. However, Tyr29-Ile49, an  $\alpha 2$ - $\alpha 3$  interaction had contacts at

the 8-10 ns interval in the 2kd1550C1, and still had contacts in 10-12 ns interval of 2kd1550C2. As in 2kd1550C1, Tyr29-Ile49 has the largest number of contacts in the last interval. The interaction Glu19-Lys46 which was between the  $\alpha 1$  and  $\alpha 3$  helices had a significant number of contacts until the 8-10 ns interval in 2kd1550C1, but has no contacts in the 10-12 ns interval of the 2kd1550C2. Again, these results are in agreement with the observation that at  $\sim 10$ -12 ns, all the  $\alpha$  helices were completely unfolded. The interactions Thr25-Ile49 and Ala26-Ile49, which are between the  $\alpha 1/\alpha 2$  turn and the  $\alpha 3$  helix, remained during the whole 12 ns simulation time. The number of contacts versus time compared with RMSD and RGYR of 2kd1550C, showed that the number of contacts for most of the interactions decreased with time as expected from the RMSD and RGYR results. But, none of the long-lasting interactions had a consistent number of contacts versus time.

To summarize, the three samples of G<sub>A</sub>95 which were simulated at 550 K show some similar trends. In all cases, the long range interactions which are between  $\alpha 1$  and  $\alpha 2$ : Ala16-Ile30, Ile17-Ile30, and Leu20-Ile30 were persistent during the simulations. All of these interactions involve the non-identical residues 20Leu and/or 30Ile. The only exception was that Ala16-Ile30 and Ile17-Ile30 interactions were not persistent in the 12 ns simulation of 2kd1550C2, which, as stated earlier, could have been due to the longer simulation time. The long range interactions between  $\alpha 2$  and  $\alpha 3$  helices: Tyr29-Leu45, Leu32-Leu45 and Ile33-Leu45, which contain the non-identical residue Leu45, were persistent in the 2kd1550A simulation. However, the Leu32-Leu45 interaction, which had a consistent number of contacts in 2kd1550A, was among those which were not persistent in 2kd1550B, 2kd1550C1, and 2kd1550C2. In addition, the interaction Ile33-

Leu45 was not persistent in 2kdl550C.

Looking at those interactions that do not contain the non-identical residues, the Ile33-Val42 interaction between  $\alpha 2$  and  $\alpha 3$  helices was an important contact during the simulations. This long range interaction had a consistent number of contacts in 2kdl550A and was persistent in the 2kdl550B. Even though the 2kdl550C simulation had only a small number of contacts for this interaction, the contact never completely disappeared. In addition, Tyr29-Ile49 interaction between  $\alpha 2$  and  $\alpha 3$  helices had a significant number of contacts in 2kdl550C1 and 2kdl550C2. Although the number of contacts was small, Tyr29-Ile49 was also persistent in 2kdl550A and 2kdl550B. The Glu19-Lys46 interaction between the  $\alpha 1$  and  $\alpha 3$  helices was persistent in 2kdl550B, 2kdl550C1 and 2kdl550C2; but not persistent in 2kdl550A. In addition, this interaction had a significant number of contacts during the last interval of the 2kdl550C1 simulation. The long range interaction Leu9-Val39, also between the  $\alpha 1$  and  $\alpha 3$  helices, was persistent in all runs, except in 2kdl550C2, with a significant number of contacts in the last interval of 2kdl550B. The  $\alpha 1/\alpha 2$  turn -  $\alpha 3$  helix interactions between Thr25-Ile49 and Ala26-Ile49 were persistent only in 2kdl550C1 and 2kdl550C2.

To get a quantitative comparison of the three simulations, the average number of contacts was calculated for the 2kdl550A, 2kdl550B and 2kdl550C1/2kdl550C2 runs (Tables 7 and 8). The persistent interactions in the average values of 2kdl550A, 2kdl550B and 2kdl550C1 (Table 7) were classified into two tiers: (1) those interactions that had >100 contacts in the 8-10 ns interval; and (2) interactions that had > 300 contacts in the 6-8 ns interval and 20 – 99 contacts in the 8-10 ns interval. Tier 1 includes interactions between (1) the  $\alpha 1$  and  $\alpha 2$  helices (Leu20-Ile30); (2) the  $\alpha 2$  and  $\alpha 3$

helices-(Tyr29-Ile49 and Ile33-Val42); (3) the  $\alpha 1$  and  $\alpha 3$  helices (Glu19-Lys46); and (5) the  $\alpha 1/\alpha 2$  turn and  $\alpha 3$  helix (Ala26-Ile49). Persistent interactions in tier 2 are those interactions between (1) the  $\alpha 1$  and  $\alpha 2$  helices (Ala16-Ile30 and Ile17-Ile30); (2) the  $\alpha 2$  and  $\alpha 3$  helices (Tyr29-Leu45 and Leu32-Leu45); and (3) the  $\alpha 1$  and  $\alpha 3$  helices (Leu9-Val39).

**Table 7. Average Number of Contacts < 10Å in 2kd1550A, 2kd1550B and 2kd1550C1**

Contacts	0 -10 ns	0 - 2 ns	2 - 4 ns	4 - 6 ns	6 - 8 ns	8 – 10 ns
Ile6-Val39	3042	990	1035	843	173	0
Leu6-Val39	4564	1807	1183	1193	311	69
Ala12-Val39	4275	1850	1303	881	240	0
Ala12-Val42	4672	1743	1218	1454	250	8
Ala12-Trp43	4247	1982	1316	882	70	6
Ala16-Ile30	3071	1306	805	402	524	35
Ala16-Ile33	3624	1739	995	448	442	0
Ala16-Lys46	3986	1809	1483	624	70	0
Ile17-Ile30	4563	1577	1471	850	601	64
Glu19-Lys46	3780	1759	1244	542	11	224
Leu20-Ile30	4292	1880	1358	508	380	167
Thr25-Ile49	2964	1345	1100	298	123	99
Ala26-Ile49	3033	1363	853	343	132	343
Tyr29-Leu45	4791	1254	1852	1176	476	33
Tyr29-Ile49	4982	1950	1453	773	240	566
Leu32-Leu45	4599	1814	1525	829	390	41
Ile33-Val42	5602	1991	1941	967	535	168
Ile33-Leu45	4038	1922	1177	564	374	1

From the above contacts, those that were not persistent in one or more trajectories were considered less significant. For example, interaction Glu19-Lys46 was not persistent in 2kd1550A, yet had a significant number of contacts in the 8-10 ns interval of the 2kd1550B simulation and a very small number of contacts from 2-8 ns. Interaction Ala26-Ile49 had a significant number of contacts in 2kd1550C1, but was not persistent in 2kd1550A and 2kd1550B. Ile33-Val42 was persistent in the 2kd1550A and

2kd1550B, but not persistent in 2kd1550C1. The interaction Tyr29-Ile49 was not persistent in 2kd1550B, but had a significant number of contacts (~100) in the 8-10 ns interval. Thus, the important interactions in tier 1 were Leu20-Ile30 and Tyr29-Ile49. The Ile33-Val42 interaction was still considered important (but less so) because even though it was not persistent in 2kd1550C1, there were still contacts present at the end of the simulation. In tier 2, the most important interactions were Leu9-Val39, Ala16-Ile30, Ile17-Ile30, Tyr29-Leu45 and Leu32-Leu45. The interaction Leu32-Leu45 was similar to the Tyr29-Ile49 interaction because it also was not persistent in 2kd1550B, but had a significant number of contacts (~100) in the fourth interval.

The long-lasting interactions in the average values of the 2kd1550A, 2kd1550B and 2kd1550C2 simulations (Table 8) were classified into three tiers: (1) those that had >100 contacts in the last interval; (2) interactions that had > 300 contacts in the 6-8 ns interval and 20 – 99 contacts in the 8-10 ns interval; and (3) those that were classified as tier 2 in the average values calculated using the 2kd1550C1 simulation, and had > 300 contacts in the fourth interval and zero contacts in the last interval for the 2kd1550C2 averages. Tier 1 included interactions between (1) the  $\alpha 1$  and  $\alpha 2$  helices: Leu20-Ile30; (2) the  $\alpha 2$  and  $\alpha 3$  helices: Ile33-Val42; and (3) the  $\alpha 1$  and  $\alpha 3$  helices: Glu19-Lys46. Tier 2 included interactions between (1) the  $\alpha 1$  and  $\alpha 2$  helices: Ala16-Ile30, Ile17-Ile30; and (5) the  $\alpha 1/\alpha 2$  turn and the  $\alpha 3$  helix: Ala26-Ile49, Tyr29-Ile49. Tier 3 included those interactions between (2) the  $\alpha 2$  and  $\alpha 3$  helices: Tyr29-Leu45 and Leu32-Leu45.

**Table 8. Average Number of Contacts < 10Å for each 2 ns Interval of 2kd1550A (0-10ns), 2kd1550B (0-10ns) and 2kd1550C2 (2-12ns)**

Contacts	Total	1 <sup>a</sup>	2 <sup>b</sup>	3 <sup>c</sup>	4 <sup>d</sup>	5 <sup>e</sup>
Ile6-Val39	2565	1101	781	618	64	0
Leu9-Val39	3907	1789	1016	968	65	69
Ala12-Val39	3633	1873	1072	610	78	0
Ala12-Val42	4007	1699	1137	922	241	8
Ala12-Trp43	3581	1966	929	620	70	6
Ala16-Ile30	2485	1115	682	227	439	21
Ala16-Ile33	2957	1711	424	388	433	0
Ala16-Lys46	3319	1756	1279	213	70	0
Ile17-Ile30	3899	1574	1434	506	320	64
Glu19-Lys46	3114	1692	974	217	53	178
Leu20-Ile30	3637	1781	871	427	403	155
Thr25-Ile49	2328	941	1115	149	117	7
Ala26-Ile49	2481	953	1043	129	335	20
Tyr29-Leu45	4391	1446	1567	871	507	0
Tyr29-Ile49	4390	1797	1535	313	681	63
Leu32-Leu45	3969	1712	1120	706	431	0
Ile33-Val42	4941	1985	1523	725	543	166
Ile33-Leu45	3375	1756	1073	177	368	1

<sup>a</sup> 0-2 ns for 2kd1550A and 2kd1550B, 2-4 ns for 2kd1550C2.

<sup>b</sup> 2-4 ns for 2kd1550A and 2kd1550B, 4-6 ns for 2kd1550C2.

<sup>c</sup> 4-6 ns for 2kd1550A and 2kd1550B, 6-8 ns for 2kd1550C2.

<sup>d</sup> 6-8 ns for 2kd1550A and 2kd1550B, 8-10 ns for 2kd1550C2.

<sup>e</sup> 8-10 ns for 2kd1550A and 2kd1550B, 10-12 ns for 2kd1550C2.

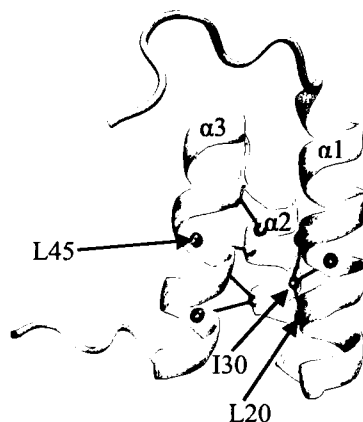
Again, those interactions that were persistent in all trajectories were considered significant. The only interactions that were excluded from the tiers were Glu19-Lys46 and Ala26-Ile49. The Glu19-Lys46 interaction in tier 1 is not persistent in 2kd1550A. While it is persistent in 2kd1550B, the number of contacts during the 2-8 ns intervals was not significant. Moreover, this interaction had a significant number of contacts only in the 8-10 ns time range of 2kd1550C but not in the 6-8 ns and 10-12 ns intervals. The Ala26-Ile49 interaction in tier 2 had a significant number of contacts in 2kd1550C2, but was not persistent in 2kd1550A and 2kd1550B. The most important interactions were

Leu20-Ile30 and Ile33-Val42 in tier 1; Ala16-Ile30, Ile17-Ile30, and Tyr29-Ile49 in tier 2; and Tyr29-Leu45 and Leu32-Leu45 in tier. Even though Leu32-Leu45 and Tyr29-Ile49 were not persistent in 2kd1550B, they had a significant number of contacts (~100) in the fourth interval. In addition, in 6-10 ns run of 2kd1550C2, Tyr29-Leu45 and Leu32-Leu45 had a similar number of contacts.

The long range interactions that were important in determining the tertiary structure of  $G_{A95}$  were  $\alpha 1/\alpha 2$  interactions: Ala16-Ile30, Ile17-Ile30, Leu20-Ile30; and  $\alpha 2/\alpha 3$  helices interactions: Tyr29-Leu45, Tyr29-Ile49, Leu32-Leu45, and Ile33-Val42. All of these important interactions were hydrophobic. Moreover, the interactions Leu9-Val36 and Glu19-Lys46 could have also contributed, but to a lesser extent. Contact Leu9-Val39 was persistent in 2kd1550A, 2kd1550B, and 2kd1550C1, but not in 2kd1550C2. The Glu19-Lys46 interaction, which could be an electrostatic interaction, was not persistent in 2kd1550A. However, it was persistent in 2kdm550B; with small number of contacts in 2-8 ns run. Alexander et al. have experimentally determined that residues Ala16, Leu20, Ile30, Ile33, and Ile49 form a tight hydrophobic core in  $G_{A95}$ .<sup>25</sup> In this study, these residues form the most important long range interactions that determine the  $3\alpha$  fold of  $G_{A95}$ . From the results it is clear that even though most of the long range interactions involve residues 20, 30 and 45, those that do not are also important. In the trajectory of the 2kd1550 samples, the  $\alpha 1$  and  $\alpha 2$  helices were more stable than  $\alpha 3$ , most likely due to the stabilizing effect of the 20-30 interaction between the two non-identical residues. Therefore, residues Leu20 and Ala30 can be more important than Leu45 in determining the  $3\alpha$  fold of  $G_{A95}$ . Furthermore, Leu20 can be the most important residue in  $G_{A95}$  because according to Alexander et al.,  $G_{A95}$  is



unfolded when Leu20 is mutated.<sup>25</sup> Figure 20 shows how these interactions hold the 3 $\alpha$ -fold together. It is very interesting to see that residue Tyr29 forms long range interactions with Leu45 and Ile49. In addition, its neighboring residue Ile30 forms long range interaction with Ala16, Ile17, and Leu20. This shows how the long range interactions are interconnected to determine the structure of G<sub>A</sub>95.



**Figure 20.** Average structure of G<sub>A</sub>95 showing the important long range interactions.

In 2kdm550A (Table 9), all the long range interactions, except Ala16- Phe30, were persistent during the simulation process. Ala16- Phe30 is the only long range interaction that was formed in both G<sub>A</sub>95 and G<sub>B</sub>95 and was one of the important interactions for folding in G<sub>A</sub>95. Most of the interactions had a significant number of contacts until the end of the simulation despite the simulation RMSD being over 10 Å for the last part of the simulation. All the long range interactions except Gly41-Thr51 and Tyr45-Phe52 had a consistent number of contacts. In addition, the interactions Thr1-Ala20, Tyr3-Ala20, Lys4-Ala16, Leu7-Glu14, Gly41-Thr55, Val42-Thr55, Thr44-Thr53, Tyr45-Phe52, and Phe30-Phe52 had a large number of contacts during the last

**Table 9. Total Number of Contacts < 10Å in 2kdm550A**

Contacts	0 - 10 ns	0 - 2 ns	2 - 4 ns	4 - 6 ns	6 - 8 ns	8 - 10 ns	SS/NI <sup>a</sup>	NC/C <sup>b</sup>
Ala16-Phe30	5441	2000	1998	1435	8	0	7, NI	
Thr1-Ala20	8684	1997	1998	1972	1505	1212	1, NI	C
Tyr3-Ala20	9133	2000	2000	2000	1921	1212	1, NI	C
Lys4-Ala16	7068	2000	2000	1593	813	662	1	C
Leu7-Glu14	7768	2000	2000	1579	1402	787	1	C
Gly41-Thr55	3806	1821	688	333	125	839	2	NC
Val42-Thr55	7133	1989	1260	1209	1198	1477	2	C
Thr44-Thr53	10000	2000	2000	2000	2000	2000	2	C
Tyr45-Phe52	9974	2000	2000	1974	2000	2000	2, NI	NC
Ile6-Thr53	7001	2000	2000	1999	990	12	3	C
Leu7-Val54	7143	2000	1964	1841	1095	243	3	C
Tyr3-Ala26	7262	2000	2000	1981	1060	221	4	C
Leu5-Phe30	7216	2000	2000	1985	1159	72	4, NI	C
Phe30-Phe52	7933	2000	2000	1995	1407	531	5, NI	C
Lys4-Thr51	6987	2000	1997	1984	805	201	6	C
Lys18-Tyr29	6796	1999	1999	1755	1043	0	7	C

<sup>a</sup> Secondary structure contacts, see text for explanation of numbers. NI indicates contact contains non-identical residue.

<sup>b</sup> C = consistent, NC = non-consistent.

interval of the simulation, four of which contain a non-identical residue. Thr1-Ala20, Tyr3-Ala20, Lys4-Ala16, and Leu7-Glu14 form the  $\beta$ 1/ $\beta$ 2 hairpin. Gly41-Thr55, Tyr45-Phe52, and Thr44-Thr53 form the  $\beta$ 3/ $\beta$ 4 hairpin. Those interactions that were in the  $\beta$ 3/ $\beta$ 4 hairpin, Thr44-Thr53 and Tyr45-Phe52, had the maximum possible number of contacts at the end of the simulation, which indicated that they were not separated when the protein was unfolded and this was also observed in the unfolding trajectory. Interaction Phe30-Phe52 is formed between the  $\alpha$  helix and  $\beta$ 4 and has the non-identical residue 30. The persistent interactions did not include the  $\beta$ 1-  $\beta$ 4 interaction which means that the  $\beta$ -sheet had broken up into two  $\beta$ -hairpin turns during the unfolding simulation. This agrees with the unfolding trajectory that the  $\beta$ 1/ $\beta$ 2 and  $\beta$ 3/ $\beta$ 4 hairpins

were further away during the simulation. The number of contacts in 2kdm550A was compared with the RMSD and RGYR versus time graphs. Even though the RMSD showed that the protein was unfolded, most of the long range interactions were in contact even at the last interval of the simulation. 2kdm550A was the only protein which was more compact, i.e., had less RGYR relative to the reference structure, which could be due to the fact that most of the long range interactions had significant number of contacts at the end of the simulation.

In 2kdm550B (Table 10), all of the interactions were persistent. All interactions except Thr1-Ala20, Ala16-Phe30, and Lys19-Tyr29, had a consistent number of

**Table 10. Total Number of Contacts < 10 Å in 2kdm550B**

Contacts	0 - 10 ns	0 - 2 ns	2 - 4 ns	4 - 6 ns	6 - 8 ns	8 - 10 ns	SS/NI <sup>a</sup>	NC/C <sup>b</sup>
Thr1-Ala20	6430	2000	1976	1996	458	0	1, NI	NC
Tyr3-Ala20	7073	2000	2000	2000	1073	0	1, NI	C
Lys4-Ala16	6956	2000	2000	2000	956	0	1	C
Leu7-Glu14	7003	2000	2000	1998	1005	0	1	C
Gly41-Thr55	7692	2000	2000	1993	859	840	2	C
Val42-Thr55	8619	2000	2000	2000	1435	1184	2	C
Thr44-Thr53	10000	2000	2000	2000	2000	2000	2	C
Tyr45-Phe52	10000	2000	2000	2000	2000	2000	2, NI	C
Ile6-Thr53	6656	2000	2000	2000	656	0	3	C
Leu7-Val54	6447	2000	2000	1997	450	0	3	C
Tyr3-Ala26	6492	2000	2000	2000	492	0	4	C
Leu5-Phe30	8001	2000	2000	2000	1197	804	4, NI	C
Phe30-Phe52	6927	2000	2000	2000	791	136	5, NI	C
Lys41-Thr51	6885	2000	2000	2000	878	7	6	C
Ala16-Phe30	7205	2000	2000	1999	565	641	7, NI	NC
Lys18-Tyr29	7293	1989	1975	1991	1177	161	7	NC

<sup>a</sup> Secondary structure contacts, see text for explanation of numbers. NI indicates contact contains non-identical residue.

<sup>b</sup> C = consistent, NC = non-consistent.

contacts. The long range interactions with a significant number of contacts at the last interval were between  $\beta 3$  and  $\beta 4$  (2): Gly41-Thr55, Val42-Thr55, Thr44-Thr53, and Tyr45-Phe52;  $\beta 1$  and  $\alpha$  (4): Leu5-Phe30; and  $\beta 2$  and  $\alpha$  (7): Ala16-Phe30. From the above long range interactions, Tyr45-Phe52, Leu5-Phe30, and Ala16-Phe30 are those that have one of the non-identical amino acids. None of these interactions were between  $\beta 1$  and  $\beta 4$ . Moreover,

it is interesting to see that, as with 2kdm55A, the interactions between  $\beta 3$  and  $\beta 4$ , Thr44-Thr53 and Tyr45-Phe52 were in contact throughout the simulation; though, the  $\beta 1/\beta 2$  hairpin turn was broken during the last interval of this simulation. The trajectory of 2kdm550B showed that the  $\beta 1/\beta 2$  and  $\beta 3/\beta 4$  hairpins were far apart as in the trajectory of 2kdm550A. The  $\beta 3/\beta 4$  hairpin did not separate throughout the simulation. In addition, the central helix was partially refolded at the end of the simulation and this can explain why the interactions between  $\beta 1$  and  $\alpha$ , Leu5-Phe30 and  $\beta 2$  and  $\alpha$ , Ala16-Phe30 were persistent interactions with significant number of contacts at the end of the simulation. Since most of the contacts were consistently decreasing with time, the results were mostly in agreement with the RMSD and RGYR versus time graphs.

The thermal unfolding simulation of 2kdm550C was run for 14 ns. When it was simulated for 10 ns, almost all of the long range interactions had the maximum number of contacts even at the end of the simulation. This result means that the protein had only started to unfold, which was also indicated by the slight increase in RMSD. When simulated for 14 ns, all the long range interactions except Gly41-Thr55 were considered persistent (Table 11). Half of the persistent interactions were consistent. In addition,

**Table 11. Total Number of Contacts < 10Å in 2kdm550C**

Contacts	1 <sup>a</sup>	2 <sup>b</sup>	3 <sup>c</sup>	4 <sup>d</sup>	5 <sup>e</sup>	6 <sup>f</sup>	7 <sup>g</sup>	8 <sup>h</sup>	SS/NI <sup>i</sup>	NC/C <sup>j</sup>
Gly41-Thr55	7438	2000	1369	1469	1778	754	68	0	2	
Thr1-Ala20	13010	1989	2000	1938	1999	2000	1725	1359	1, NI	NC
Tyr3-Ala20	13220	2000	2000	2000	2000	2000	2000	1220	1, NI	C
Lys4-Ala16	12369	2000	2000	2000	2000	2000	1986	383	1	C
Leu7-Glu14	10879	2000	2000	1999	2000	1993	871	16	1	NC
Val42-Thr55	9456	2000	1859	1887	1942	1238	530	0	2	NC
Thr44-Thr53	11529	2000	2000	2000	2000	2000	1267	262	2	C
Tyr45-Phe52	13454	2000	2000	2000	2000	2000	1941	1513	2, NI	C
Ile6-Thr53	12719	2000	2000	2000	2000	2000	1638	1081	3	C
Leu7-Val54	11006	2000	1928	1961	2000	1916	1157	44	3	NC
Tyr3-Ala26	13371	2000	2000	2000	1999	2000	2000	1372	4	NC
Leu5-Phe30	13091	2000	2000	2000	2000	2000	2000	1091	4, NI	C
Phe30-Phe52	12725	2000	2000	2000	2000	2000	1533	1192	5, NI	C
Lys4-Thr51	11651	2000	2000	2000	2000	2000	1525	126	6	C
Ala16-Phe30	10878	1999	2000	2000	1998	1999	882	0	7, NI	NC
Lys18-Tyr29	11983	1969	1998	2000	1974	1998	1976	68	7	NC

<sup>a</sup> 0-14 ns, <sup>b</sup> 0-2 ns, <sup>c</sup> 2-4 ns, <sup>d</sup> 4-6 ns, <sup>e</sup> 6-8 ns, <sup>f</sup> 8-10 ns, <sup>g</sup> 10-12 ns, <sup>h</sup> 12-14 ns.

<sup>i</sup> Secondary structure contacts, see text for explanation of numbers. NI indicates contact contains non-identical residue.

<sup>j</sup> C = consistent, NC = non-consistent.

many interactions had a significant number of contacts (~1000) during the last interval (12-14 ns) of the simulation. These were the long range interactions between  $\beta$ 1 and  $\beta$ 2 (1): Thr1-Ala20 and Tyr3-Ala20;  $\beta$ 3 and  $\beta$ 4 (2): Tyr45-Phe52;  $\beta$ 1 and  $\beta$ 4 (3): Ile6-Thr53;  $\beta$ 1 and  $\alpha$  (4): Tyr3-Ala26, Leu5-Phe30; and  $\alpha$  and  $\beta$ 4 (5): Phe30-Phe52. From the above interactions, Thr1-Ala20, Tyr3-Ala20, Tyr45-Phe52, Leu5-Phe30 and Phe30-Phe52 have one of the non-identical residues. The Tyr45-Phe52 interaction which was observed to have the maximum possible number of contacts at the end of the simulation in kdm550A and 2kdm550B had the highest number of contacts of all the contacts during the kdm550C 12-14 ns interval. The Thr44-Thr53 interaction had a much smaller number of contacts than for 2kdm550A and 2kdm550B, which could be the result of the

extended time. The number of contacts in 2kdm550C was significantly high until the 10-12 ns interval and this was in agreement with the small RMSD and RGYR graphs. At the 12-14 ns interval, the protein unfolded with high RMSD and RGYR, but most interactions had significant contacts during this interval. The trajectory of 2kdm550C also showed that the unfolding started at around 10 ns, in agreement with the large number of contacts for the last interval of the simulation. It was also interesting to see that a  $\beta$ 1- $\beta$ 4 interaction, Ile6-Thr53, was in the persistent interactions list, which could be due to the unfolding of the protein at the last interval. Moreover, interactions between  $\beta$ 1 and  $\alpha$ , Tyr3-Ala26, Leu5-Phe20; and  $\alpha$  and  $\beta$ 4, Phe30-Phe52 were probably persistent interactions because according to the trajectory, the  $\alpha$  helix was only partially unfolded.

The three samples of G<sub>B</sub>95 which were simulated at 550 K had similarity in that most of the contacts remained much longer than for G<sub>A</sub>95. Most likely, G<sub>B</sub>95 seems to unravel within a short period of time, giving the  $\beta$ 3/  $\beta$ 4 hairpin; whereas G<sub>A</sub>95 seems to unfold more slowly. In all the samples, the long range interactions between  $\beta$ 1 and  $\beta$ 2 (1): Tyr3-Ala20;  $\beta$ 3 and  $\beta$ 4 (2): Thr44-Thr53, Tyr45-Phe52;  $\beta$ 1 and  $\alpha$  (4): Leu5-Phe30; and  $\beta$ 2 and  $\alpha$  (7): Lys18-Tyr29 remained for a significant period of time. Tyr3-Ala20, Tyr45-Phe52, and Leu5-Phe30 involved one of the non-identical residues. Even though the interactions Thr44-Thr53 and Lys18-Tyr29 did not have the non-identical residues, they were still important interactions in G<sub>B</sub>95. The long range interaction Thr1-Ala20 in the  $\beta$ 1/ $\beta$ 2 hairpin was persistent in 2kdm550A and 2kdm550C, and it has the non-identical residue 20Ala. The long range interaction Phe30-Phe52 between  $\alpha$  and  $\beta$ 4 was also persistent in 2kdm550A and 2kdm550C, and has the non-identical residue 30Phe. Moreover, long range interactions that did not involve one of the non-identical residues

were also observed in at least two of the samples. These were Leu7-Glu14, Val42-Thr55, Leu7-Val54, and Tyr3-Ala26.

The average number of contacts in the three samples of 2kdm550 was calculated. In the 2kdm550C simulation, the first two intervals (0-2 ns, 2-4 ns) were excluded, and the interval 4-6 ns was taken as the first interval for the calculation. The average results showed that all of the long range interactions were present for long periods of time during the simulation process (Table 12). In addition, all except Ala16-Phe30 and Gly41-Thr55 had a large number of contacts until the fourth interval (6-8 ns). Note that the only identical interaction between G<sub>A</sub>95 and G<sub>B</sub>95, 16-30, was one of the interactions that was most important in determining the 3 $\alpha$  fold of G<sub>A</sub>95. A persistent  $\beta$ 2- $\alpha$  interaction between 16-30 was observed only in 2kdm550B and thus, this contact was not important in determining the structure of G<sub>B</sub>95. Residue 30, which is Ile in G<sub>A</sub>95 and Phe in G<sub>B</sub>95, is part of the central helix in both proteins. The reason why 16-30 was important only in G<sub>A</sub>95 might be due to the difference in the structures of Ile and Phe. Though both are hydrophobic, the structure of Ile is branched and bulky, whereas, the benzene ring of Phe is flat. Those interactions with a significant number of contacts at the last interval were considered the most important ones and include interactions between  $\beta$ 1 and  $\beta$ 2 (1) Thr1-Ala20, Tyr3-Ala20;  $\beta$ 3 and  $\beta$ 4 (2) Thr44-Thr53, Tyr45-Phe52, Val42-Thr55, Gly41-Thr55;  $\beta$ 1 and  $\alpha$  (3) Leu5-Phe30, Tyr3-Ala26; and  $\alpha$  and  $\beta$ 4 (4) Phe30-Phe52. All of these interactions, except those containing Thr, form hydrophobic contacts. Alexander et al. have found that approximately 50% of the hydrophobic interactions in G<sub>B</sub>95 are formed from residues Tyr3, Leu5, Phe30, and Phe52.<sup>25</sup> All of these residues were also found in the most important long range

interactions. Figure 21 shows how these long range interactions are interconnected in G<sub>B</sub>95. The interactions that involved the non-identical residues were  $\beta$ 1- $\beta$ 2 interactions: Thr1-Ala20, Tyr3-Ala20;  $\beta$ 3- $\beta$ 4 interaction: Tyr45-Phe52;  $\beta$ 1- $\alpha$  interaction: Leu5-Phe30;  $\alpha$ - $\beta$ 4 interaction: Phe30-Phe52. It was interesting to see that none of these important interactions were between  $\beta$ 1 and  $\beta$ 4, which indicated that these strands were probably the easiest part of the protein to unfold.

**Table 12. Average Number of Contacts < 10Å for each 2 ns Interval of 2kdm550A (0-10ns), 2kdm550B (0-10ns) and 2kdm550C (4-14ns)**

Contacts	Total	1 <sup>a</sup>	2 <sup>b</sup>	3 <sup>c</sup>	4 <sup>d</sup>	5 <sup>e</sup>
Thr1-Ala20	8045	1978	1991	1989	1229	857
Tyr3-Ala20	8475	2000	2000	2000	1665	811
Tyr3-Ala26	7708	2000	2000	1994	1184	531
Lys4-Ala16	7464	2000	2000	1864	1252	348
Lys4-Thr51	7174	2000	1999	1995	1069	111
Leu5-Phe30	8103	2000	2000	1995	1452	656
Ile6-Thr53	7459	2000	2000	2000	1095	364
Leu7-Glu14	7217	2000	2000	1857	1093	268
Leu7-Val54	6889	1987	1988	1918	901	96
Ala16-Phe30	6508	2000	1999	1811	485	214
Lys18-Tyr29	7368	1996	1983	1915	1399	76
Phe30-Phe52	7862	2000	2000	1998	1244	620
Gly41-Thr55	5189	1763	1489	1027	351	560
Val42-Thr55	7116	1959	1734	1482	1054	887
Thr44-Thr53	9176	2000	2000	2000	1756	1421
Tyr45-Phe52	9809	2000	2000	1991	1980	1838

<sup>a</sup> 0-2 ns for 2kdm550A and 2kdm550B, 4-6 ns for 2kdm550C.

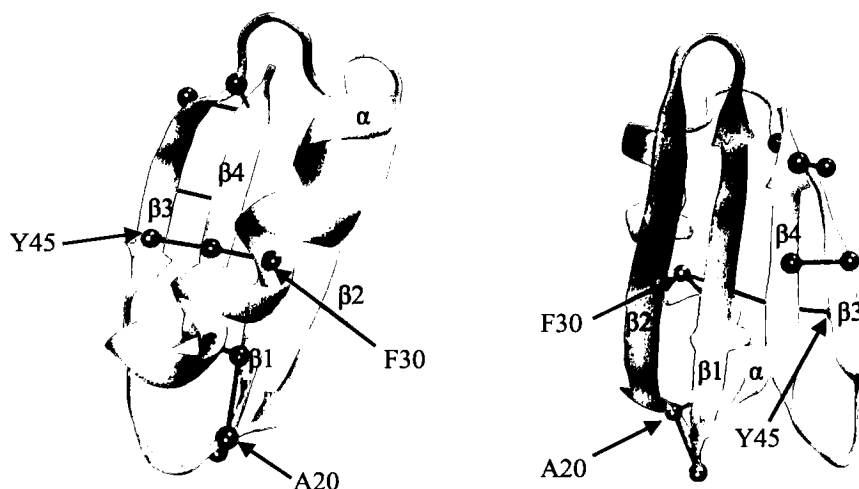
<sup>b</sup> 2-4 ns for 2kdm550A and 2kdm550B, 6-8 ns for 2kdm550C.

<sup>c</sup> 4-6 ns for 2kdm550A and 2kdm550B, 8-10 ns for 2kdm550C.

<sup>d</sup> 6-8 ns for 2kdm550A and 2kdm550B, 10-12 ns for 2kdm550C.

<sup>e</sup> 8-10 ns for 2kdm550A and 2kdm550B, 12-14 ns for 2kdm550C.





**Figure 21.** Two views of the average structure of  $G_B95$  showing the important long range interactions.

This study showed that Tyr45-Phe52 is a very critical interaction for stabilizing the  $4\beta+\alpha$  fold of  $G_B95$  as this was the only persistent interaction in all three simulations. This is in agreement with Alexander et al.'s observation that the residues Tyr45 and Phe52 increased the stability of the  $4\beta + \alpha$  fold by +2 kcal/mol.<sup>25</sup> Residue 45 is one of the non-identical residues and these results demonstrate how very important this amino acid is in determining the three dimensional structure of  $G_B95$ . In this study, the interaction Thr44-Thr53 also had a similar number of contacts as Tyr45-Phe52. The interactions Tyr45-Phe52, Thr44-Thr53, and also Val42-Thr55 and Gly41-Thr55; which were shown to be important from the averaging data, can explain why the  $\beta3/\beta4$  hairpin did not separate during the thermal unfolding simulations of 2kdm550A and 2kdm550B and only partially unfolded in the 2kdm550C simulation. Two of the other important interactions as determined by averaging were Thr1-Ala20 and Tyr3-Ala20 and they explain why the  $\beta1/\beta2$  hairpin remained for most of the simulation before separating.

These interactions have the non-identical residue Ala20. The interactions between  $\alpha$  and  $\beta$ 1: Leu5-Phe30; and  $\alpha$  and  $\beta$ 4: Phe30-Phe30, which also had significant number of contacts at the end of the simulation, contained the non-identical residue Phe30. The central helix in 2kdm550B and 2kdm550C was partially stable during the simulation and this could be due to the persistent interactions between  $\beta$ 1 and  $\alpha$  (Leu5-Phe30) and  $\alpha$  and  $\beta$ 4 (Phe30-Phe52).

The fact that Thr44-Thr53, Val42-Thr55, Gly41-Thr55, and Tyr3-Ala26 did not involve a non-identical residue but were persistent indicated that the long range interactions with a non-identical residue work together with those without a non-identical residue to form the tertiary structure of G<sub>B</sub>95. It was interesting to see that residue Phe52 formed long range interactions with both Tyr45 and Phe30, all of which contain a benzene ring. This set of contacts shows the interconnectivity of the long range interactions. Moreover, residues Tyr45 and Phe30 may be more important in determining the structure of G<sub>B</sub>95 than that of Ala20 because both Tyr45 and Phe30 had long range interactions with residue Tyr52, and Tyr45-Phe52 was the most persistent interaction, and Phe30 also had a long range interaction with residue Leu5; creating a mini-network. Ala20 was also part of a mini-network involving residues Thr1, Tyr3 and Ala26. However, neither of the other non-identical residues were part of this network. This observation agrees with the results of Alexander et al.'s experiment that G<sub>B</sub>95 maintains the 4 $\beta$ + $\alpha$  fold even when residue 20 is mutated.<sup>25</sup> The unfolding simulation of 2kdm550 involved the breaking of the  $\beta$ 1- $\beta$ 4 interactions, which was shown by the two hairpins separating from the central helix and was usually the starting point of the protein unfolding. This is supported by the persistent interactions that only one contact

in 2kdm550C (Ile6-Thr53) was between  $\beta$ 1- $\beta$ 4, but 2kdm550A and 2kdm550B did not have persistent interactions between  $\beta$ 1 and  $\beta$ 4.

## CHAPTER IV

### CONCLUSION

In this study, thermal unfolding simulations have been performed on G<sub>A</sub>95 and G<sub>B</sub>95 at different temperatures. Both proteins unfolded at 550 K and three independent simulations were run at this temperature. The unfolding simulations were used to study how the long range interactions behave as the proteins unfold. Those long range interactions that are persistent or long-lasting were considered the most important ones in determining the tertiary structures of G<sub>A</sub>95 and G<sub>B</sub>95. The long range interactions were also compared to see if they contain any of the non-identical residues that distinguish the sequences of G<sub>A</sub>95 and G<sub>B</sub>95.

The average number of contacts for each long range interaction of G<sub>A</sub>95 was calculated for the 10ns simulations. In addition, averages were also calculated using only the 2-12 ns interval of the third trajectory. The persistent interactions were classified into tiers according to the number of contacts that they have in the last two intervals. Results show that the persistent interactions that are important in determining the tertiary structure of G<sub>A</sub>95 are  $\alpha 1/\alpha 2$  interactions: Ala16-Ile30, Ile17-Ile30, Leu20-Ile30; and  $\alpha 2/\alpha 3$  interactions: Tyr29-Leu45, Tyr29-Ile49, Leu32-Leu45, and Ile33-Val42. Moreover,  $\alpha 1/\alpha 3$  interactions: Leu9-Val36 and Glu19-Lys46 may also contribute to a lesser extent. The interactions, Ala16-Ile30, Ile17-Ile30, Leu20-Ile30, Tyr29-Leu45 and Leu32-Leu45 have one of the non-identical residues, Leu20, Ile30, and Leu45. However, the interactions, Tyr29-Ile49 and Ile33-Val42, do not involve the non-identical residues. It can be concluded that not only are the interactions that have the non-identical residues

important, but also those without non-identical residues. All of the interactions are needed to maintain the 3 $\alpha$ -fold of G<sub>A</sub>95.

The most reliable long-lasting interaction is formed between Leu20 and Ile30, which indicates that these residues may be more important than residue Leu45 in determining the fold of G<sub>A</sub>95. Leu20-Ile30 is between  $\alpha$ 1 and  $\alpha$ 2 helices and can help explain why these helices were more stable than  $\alpha$ 3 during the unfolding simulations. Hence, Leu20-Ile30 can be the main interaction that prevents G<sub>A</sub>95 from assuming a 4 $\beta$ + $\alpha$  fold. From these non-identical residues, Leu20 can be the most important residue in G<sub>A</sub>95.

The average of number of contacts for each long range interaction of G<sub>B</sub>95 was calculated by using 2kdm550A, 2kdm550B, and 2kdm550C. The 2kdm550C was simulated for an extended period of 14 ns and the number of contacts from 4-14 ns was used for the calculation. The persistent interactions, hence the most important ones, were those with a significant number of contacts in the last interval of the simulation. These interactions were between  $\beta$ 1 and  $\beta$ 2 (1) Thr1-Ala20, Tyr3-Ala20;  $\beta$ 3 and  $\beta$ 4 (2) Thr44-Thr53, Tyr45-Phe52, Val42-Thr55, Gly41-Thr55;  $\beta$ 1 and  $\alpha$  (3) Leu5-Phe30, Tyr3-Ala26; and  $\alpha$  and  $\beta$ 4 (4) Phe30-Phe52. In addition to the interactions that involved one of the non-identical residues, important interactions also included those without non-identical residues such as Thr44-Thr53, Val42-Thr55, Gly41-Thr55, and Tyr3-Ala26. Those interactions between  $\beta$ 3 and  $\beta$ 4, particularly Thr44-Thr53 and Tyr45-Phe52, had a higher number of contacts. This explains why the  $\beta$ 3/ $\beta$ 4 hairpin did not separate during the simulation. One of the interactions in the  $\beta$ 3/ $\beta$ 4 hairpin, with the highest number of contacts and the only interaction that was persistent in all three samples, Tyr45-Phe52,

involved the non-identical residue Tyr45. Thus, this contact was a very important interaction and stabilized the  $4\beta+\alpha$  fold of G<sub>B</sub>95. Consequently, residue Tyr45 was probably the most important residue in determining the tertiary structure of G<sub>B</sub>95. Thr44-Thr53, Val42-Thr55, and Gly41-Thr55 which did not contain any non-identical residues, were also important interactions in the  $\beta 3/\beta 4$  hairpin. Thr1-Ala20 and Tyr3-Ala20 involved the non-identical residue 20 and presumably contributed to the stability of the  $\beta 1/\beta 2$  hairpin. This structure remained for almost the whole simulation before separating during the unfolding process. The other significant long range interactions, Leu5-Phe30 and Phe30-Phe52, contained the non-identical residue Leu30. Residue 30 appeared to be more important than 20 because it was part of a network of long-lasting interactions involving 45-52, 30-52 and 5-30.

The thermal unfolding simulations used in this study provided an increased understanding of how the long range interactions, with and/or without non-identical residues, determined the tertiary structures of G<sub>A</sub>95 and G<sub>B</sub>95. Moreover, Morrone et al. observed that G<sub>A</sub>88 and G<sub>B</sub>88 formed and stabilized different long range interactions in the denatured state and thus, different folding nuclei that dictated the tertiary structure were formed for each protein.<sup>26</sup> The G<sub>A</sub>95 and G<sub>B</sub>95 simulations identified the interactions that most likely contributed to the folding nuclei that form in the denatured state of these proteins. A very different set of interactions was located for each protein, supporting the theory that it is the long range interactions in the denatured state that led to the different structures for these two proteins. Even though it was possible to make a conclusion as to which interactions are the most important ones, additional information should be obtained and the conclusions further verified by running more simulations of

the proteins and for longer time periods. Breaking down residue-residue interactions into atom-atom interactions and studying the exact geometry of interactions and changes during unfolding would also be useful.

## REFERENCES

- (1) Garrett, R. H.; Grisham, C. M. *Biochemistry*; 4th ed.; Mary Finch: Boston, 2010.
- (2) Dobson, C. M. *Nature* **2003**, *426*, 884-890.
- (3) Voet, D.; Voet, J. G.; Pratt, C. W. *Fundamentals of Biochemistry*; 2nd ed.; John Wiley & Sons, Inc.: Hoboken, 2006.
- (4) National Institutes of Health. National Human Genome Research Institute. "Talking Glossary of Genetic Terms". <http://www.genome.gov/glossary/?id=169> (accessed August 4, 2012).
- (5) Matthews, C. R. *Curr. Opin. Struct. Biol.* **1991**, *1*, 28-35.
- (6) Stote, R.; Dejaegere, A.; Kuznetsov, D.; Falquet, L. *Theory of Molecular Dynamics Simulation Tutorial*; 1999.
- (7) Weissman, J. S. *Chem. Biol.* **1995**, *2*, 255-260.
- (8) Fedjukina, D. V.; Cavagnero, S. *Annu. Rev. Biophys.* **2011**, *40*, 337-59.
- (9) Travaglini-Allocatelli, C.; Ivarsson, Y.; Jemth, P.; Gianni, S. *Curr. Opin. Struct. Biol.* **2009**, *19*, 3-7.
- (10) Buchner, G. S.; Murphy, R. D.; Buchete, N.; Kubelka, J. *Biochim. Biophys. Acta* **2011**, *1814*, 1001-1020.
- (11) Summa, C. M.; Levitt, M. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 3177-3182.
- (12) Anfinsen, C. B.; Haber, E.; Sela, M.; White, F. H., Jr. *Proc. Natl. Acad. Sci. USA* **1961**, *47*, 1309-14.
- (13) Lu, D.; Liu, Z. *Annu. Rep. Prog. Chem.* **2010**, *106*, 259-273.
- (14) Zwanzig, R.; Szabo, A.; Bagchi, B. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 20-22.
- (15) Jackson, S. E. *Folding Des.* **1998**, *3*, 81-91.
- (16) Sosnick, T. R.; Barrick, D. *Curr. Opin. Struct. Biol.* **2011**, *21*, 12-24.
- (17) Nolting, B.; Andert, K. *Proteins: Struct., Funct., Bioinf.* **2000**, *41*, 288-98.
- (18) Daggett, V.; Fersht, A. R. *Trends Biochem. Sci.* **2003**, *28*, 18-25.
- (19) Fersht, A. R. *Curr. Opin. Struct. Biol.* **1997**, *7*, 3-9.
- (20) Alexander, P. A.; He, Y.; Chen, Y.; Orban, J.; Bryan, P. N. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 11963-11968.
- (21) He, Y.; Chen, Y.; Alexander, P.; Bryan, P. N.; Orban, J. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 14412-14417.
- (22) Marti-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sanchez, R.; Melo, F.; Sali, A. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 291-325.
- (23) Falkenberg, C.; Bjorck, L.; Akerstrom, B. *Biochemistry* **1992**, *31*, 1451-7.
- (24) Myhre, E. B.; Kronvall, G. *Infect. Immun.* **1977**, *17*, 475-82.
- (25) Alexander, P. A.; He, Y.; Chen, Y.; Orban, J.; Bryan, P. N. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 21149-21154.
- (26) Morrone, A.; McCully, M. E.; Bryan, P. N.; Brunori, M.; Daggett, V.; Gianni, S.; Travaglini-Allocatelli, C. *J. Biol. Chem.* **2011**, *286* (5), 3863-3872.
- (27) Allison, J. R.; Bergeler, M.; Hansen, N.; van Gunsteren, W. F. *Biochemistry* **2011**, *50* (50), 10965-10973.
- (28) Lazim, R.; Mei, Y.; Zhang, D. *J. Mol. Mod.* **2012**, *18* (3), 1087-1095.
- (29) Williamson, M. P. *Annu. Rep. NMR Spectrosc.* **2009**, *65*, 77-109.
- (30) Horst, J.; Samudrala, R. *F1000 Biol. Rep.* **2009**, *1*, 69.



- (31) Brooks, B. R.; Brooks Iii, C. L.; Mackerell Jr, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoseck, M.; Im, W.; Kuczera, K.; Lazaridis, T. *J. Comp. Chem.* **2009**, *30* (10), 1545-1614.
- (32) Schleif, R. *Analysis of Protein Structure and Function: A Beginner's Guide to CHARMM*; 2006.
- (33) Becker, O. M.; Karplus, M. *A Guide to Biomolecular Simulations*; Springer: The Netherlands, 2006; Vol. 4.
- (34) Buck, M.; Bouguet-Bonnet, S.; Pastor, R. W.; MacKerell, A. D. *Biophys. J.* **2006**, *90*, L36-L38.
- (35) York, D. M.; Wlodawer, A.; Pedersen, L. G.; Darden, T. A. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 8715-8718.
- (36) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press Inc.: New York, 1987.
- (37) Radkiewicz, J. L.; Brooks, C. L. *J. Am. Chem. Soc.* **1999**, *122* (2), 225-231.
- (38) Humphrey, W.; Dalke, A.; Schulten, K. *J. Mol. Graph.* **1996**, *14*, 33-8, 27-8.
- (39) Higman, V. A.; Greene, L. H. *Phys. A* **2006**, *368* (2), 595-606.

## APPENDIX

### Figure 1

Dear Sirs,

I am a Chemistry Master's student at Old Dominion University, Norfolk VA. I would like to request permission to use the following web page for my thesis:

<http://www.genome.gov>

Can I use figure from the protein section? If so, do I need letter of permission from you and how should the pictures be referenced? Can I get author's name and year of publication?

I would really appreciate your fast reply.

Thank you in advance.

Milen Tesfamariam  
Chemistry MS Student  
Old Dominion University

**NHGRI Webmaster (NIH/NHGRI)** [nhgriwebmaster@mail.nih.gov](mailto:nhgriwebmaster@mail.nih.gov)

Here is the exact location of the image on our site, in color and high-res if you need it:

<http://www.genome.gov/glossary/index.cfm?id=169>

Yes, you may use it. All gov't images are in the public domain. But here is some information on how to cite it. It is always common courtesy to link back to us when possible:

<http://www.genome.gov/glossary/index.cfm?p=howtocite> (where it says? id=0 please change the "0" to the id of the actual link, in the case above it's "169").

Let me know if you have any other questions.

-Webmaster

**Figure 2**

**NATURE PUBLISHING GROUP LICENSE  
TERMS AND CONDITIONS**

Aug 06, 2012

This is a License Agreement between Milen Tesfamariam ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

**All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.**

License Number	2963291117453
License date	Aug 06, 2012
Licensed content publisher	Nature Publishing Group
Licensed content publication	Nature
Licensed content title	Protein folding and misfolding
Licensed content author	Christopher M. Dobson
Licensed content date	Dec 18, 2003
Volume number	426
Issue number	6968
Type of Use	reuse in a thesis/dissertation
Requestor type	non-commercial (non-profit)
Format	electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Figures	Figure 1
Author of this NPG article	no
Your reference number	
Title of your thesis / dissertation	Identification of Persistent Long Range Interactions in GA95 and GB95 through Thermal Unfolding Simulations
Expected completion date	Aug 2012
Estimated size (number of pages)	80
Total	0.00 USD

**Figure 3**

**JOHN WILEY AND SONS LICENSE  
TERMS AND CONDITIONS**

Aug 06, 2012

This is a License Agreement between Milen Tesfamariam ("You") and John Wiley and Sons ("John Wiley and Sons") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by John Wiley and Sons, and the payment terms and conditions.

**All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.**

License Number	2960351383428
License date	Aug 01, 2012
Licensed content publisher	John Wiley and Sons
Licensed content publication	Proteins: Structure, Function and Bioinformatics
Licensed content title	Mechanism of protein folding
Licensed content author	Bengt Nölting, Karl Andert
Licensed content date	Sep 18, 2000
Start page	288
End page	298
Type of use	Dissertation/Thesis
Requestor type	University/Academic
Format	Electronic
Portion	Figure/table
Number of figures/tables	1
Number of extracts	
Original Wiley figure/table number(s)	Figure 1
Will you be translating?	No
Order reference number	
Total	0.00 USD

**Figures 5 and 6**

Dear Sirs,

I am a Chemistry Master's student at Old Dominion University, Norfolk VA. I would like to request permission to use the following journal for my thesis:

Alexander, P. A.; He, Y.; Chen, Y.; Orban, J.; Bryan, P. N., A Minimal Sequence Code for Switching Protein Structure and Function. *PNAS* **2009**, *106*, 21149-21154.

Can I use figures from the journal? If so, do I need letter of permission from you and how should the pictures be referenced?

I would really appreciate your fast reply.

Thank you in advance.

Milen Tesfamariam  
Chemistry MS Student  
Old Dominion University

**PNAS Permissions** [PNASPermissions@nas.edu](mailto:PNASPermissions@nas.edu)

Dear Milen Tesfamariam,

Permission is granted for your use of the figures (1-6 of the referenced article) as described in your message below. Please cite the full journal references.

Please let us know if you have any questions.

Thank you!

Best regards,  
Kelly Gerrity for  
Diane Sullenberger  
Executive Editor  
PNAS

**Figure 8**

**Title:** Current Computer Modeling Cannot Explain Why Two Highly Similar Sequences Fold into Different Structures

**Author:** Jane R. Allison, Maike Bergeler, Niels Hansen, and Wilfred F. van Gunsteren

**Publication:** Biochemistry

**Publisher:** American Chemical Society

**Date:** Dec 1, 2011

**Copyright** © 2011, American Chemical Society

**PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE**

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

Permission is granted for your request in both print and electronic formats, and translations. If figures and/or tables were requested, they may be adapted or used in part.

Please print this page for your records and send a copy of it to your publisher/graduate school.

Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.

One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

**Figure 9****SPRINGER LICENSE  
TERMS AND CONDITIONS**

Aug 06, 2012

This is a License Agreement between Milen Tesfamariam ("You") and Springer ("Springer") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Springer, and the payment terms and conditions.

**All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.**

License Number	2960350629050
License date	Aug 01, 2012
Licensed content publisher	Springer
Licensed content publication	Journal of Molecular Modeling
Licensed content title	Replica exchange molecular dynamics simulation of structure variation from $\alpha/4\beta$ -fold to $3\alpha$ -fold protein
Licensed content author	Raudah Lazim
Licensed content date	Jan 1, 2011
Volume number	18
Issue number	3
Type of Use	Thesis/Dissertation
Portion	Figures
Author of this Springer article	No
Order reference number	
Title of your thesis / dissertation	IDENTIFICATION OF PERSISTENT LONG RANGE INTERACTIONS IN GA95 AND GB95 THROUGH THERMAL UNFOLDING SIMULATIONS
Expected completion date	Aug 2012
Estimated size(pages)	76
Total	0.00 USD

## VITA

Milen Redai Tesfamariam

### Education

- Old Dominion University  
Department of Chemistry and Biochemistry  
Norfolk, VA 23529  
M.S. in Chemistry, August 2012
- University of Asmara  
College of Health Sciences  
Asmara, Eritrea  
B.S. in Pharmacy, September 2007

### Publication

- Tesfamariam M. R.; Poutsma J. L. Identification of Persistent Long Range Interactions in G<sub>A</sub>95 and G<sub>B</sub>95 through Thermal Unfolding Simulations. (2012, in preparation)

### Teaching Experience

- Introductory Chemistry Lab (CHEM 106 and CHEM 108)
- Foundation of Chemistry Lab (CHEM 115 and CHEM 122)