A STUDY OF BARSTAR FOLDING EVENTS USING BOUNDARY VALUE SIMULATIONS

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Master of Science

by

Jacob Morris Yunger

January 2007

ABSTRACT

This study revolves around a computational algorithm called SDEL (Stochastic Difference Equation in Length) that generates approximate protein folding trajectories on the atomically detailed resolution scale. The protein studied is Barstar- a barnase inhibitor. Because of the protein's interesting structure (four alpha helices, three beta strands) and relatively small size (89 residues), Barstar is an optimal choice for running complete folding trajectories on a computer. 12 pathways were generated with SDEL, starting from a structurally wide selection of unfolded conformations, yet all ending with the native configuration. We tracked hydrogen bonds, dihedral angles, native and non-native contacts, and energetic along these folding pathways. The resulting trajectories show: 1) Barstar follows the Hydrophobic Collapse folding scenario, 2) native α-helices begin forming earlier in the trajectory than the β-sheets, 3) particular residues maintain a propensity for helical structure in their unfolded state, and 4) specific non-native contacts persist during the folding trajectory. Strong correlations were found between the SDEL pathways and data from NMR, CD, and other experimental studies.

BIOGRAPHICAL SKETCH

Jacob (Yaacov) Yunger was born in downtown Toronto, Canada on February 26, 1980. In May of 1997, he graduated the Irving Zucker College of Hamilton, Ontario – an unsupervised boarding high school. He then spent a year studying ancient texts in Aramaic across the road from a socialist community in the State of Israel. Jacob returned to North America in 1998, spending four years at Yeshiva University located in Spanish Harlem, double majoring in physics and philosophy with a minor in mathematics. Deferring graduate school for a year, Jacob returned to the Middle East as a resident counselor and mentor for unsupervised students studying ancient texts in Aramaic. Finally arriving in Ithaca, Jacob began a graduate program of Physics at Cornell University in August of 2003 – attending his first co-ed class since kindergarten. On August 15, 2004, Jacob married Mera Bender of Maryland after convincing her he was in Ithaca studying to become a rabbi. After three and a half years at Cornell, a wiser, humbler, and less-haired Jacob graduated with an MS in physics. He “looks forward” to a life of trying to find ways of making use of this degree.

“Rabbi Masya ben Charash said: Initiate a greeting to every person; and be a tail to the lions rather than a head to the foxes.” – Pirkei Avot (4:15)

To Mera

תְּנוּ לָהּ מִפְּרִי יָדֶיהָ. וִיהַלְלוּהָ בַּשְּׁעָרִים מַעֲשֶׂיהָ׃

ACKNOWLEDGMENTS

During this roller coaster ride, there was help along the way.

I would like to thank the Cornell Physics staff for always being friendly, supportive, and helpful. Special appreciation goes for the TA funding I received for nine semesters.

Thanks to Paul, Peter, Sourish, Chris, Xe, Svet, Jan and Bruno for paving the road while keeping it real.

A big thank-you to Jim Alexander, for all his encouragement early in my career. An equally sized thank you to all those who discouraged me - your negativity only made me try harder to prove you wrong.

Scientists aught to be people of great minds and great character. Such role models I had valuable private time with include Paul Ginsparg, Don Holcomb, Raphael Littauer and David Mermin. I learned more from our short conversations than from all my courses combined.

Thanks to my Special Committee: Don Hartill, Tomas Arias and Sol Gruner.

Thank you, Ron Elber, for giving me a computer, a project, and help with my first serious scientific endeavor. Thanks for the quick email responses, and for the opportunity.

My love for science, physics, and education grew exponentially with my time spent with Phil Krasicky. Thank you for teaching the teachers, exciting the students, and helping me foster a career path in physics education.

Thanks to my parents for only wanting me to be happy. Thanks to God for humbling me without letting go.

Big love to my wife, Mera. She sacrificed by coming to Ithaca for three years, and was my raft while I was sinking. I look forward to keeping afloat together.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF GRAPHS

CHAPTER 1

INTRODUCTION

Proteins

Proteins are biologically significant macromolecules. Every protein has its own specific function within an organism such as chemical reaction catalysis, cargo transport, and other important biological functions. Each molecule is made up of a chain (or multiple chains) consisting of a linear amino acid sequence, which is joined together with covalent peptide bonds. The peptide pieces contains a *backbone* consisting of a succession of two carbons and one nitrogen, and one of twenty *sidechain* groups, which are covalently attached to the first carbon (knows as the C_α). The sidechains can be as simple as one hydrogen (Glycine), or more complex loops (e.g. Tryptophan). In this way, the makeup of the protein is defined solely by the linear code of the amino acid sequence.[1] While the protein is synthesized as a linear macromolecule, the protein will not function as it should until taking a unique three-dimensional compact shape. The unfolded protein conformations are often called *denatured* configurations. The unique three-dimensional folded structure that allows it to be functional is called the *native* structure.

Protein structure has four-fold classification. The linear sequence of amino acid coding is known as the protein's *primary structure*. The torsion angle between the N and C_α atoms, and the C_α and C atoms in the peptide, are called the ϕ and ψ dihedral angles, respectively. Local internal structure within the native conformation is classified as *secondary structure*. These structures are due mostly to hydrogen bonding between backbone parts. The two most commonly defined secondary structure forms are α-helices and β-sheets. In protein α-helices, there are 3.6 peptide

pieces – or, *residues* - per turn. The hydrogen bonds needed for this conformation come from the contact between the backbone carbonyl (C-O) of residue n and the amino (N-H) of reside n+4. The β-sheets are h-bonded strands of parallel or anti-parallel slightly coiled stretches with only two residues per turn. One of the dogmas in the protein folding community is that secondary structure - and eventually tertiary and quaternary structure - can be predicted from the amino acid sequence alone. Some secondary structure prediction algorithms have a 70-80% success rate where the failures are attributed to such things as tertiary interactions that fix secondary structure elements.[2]

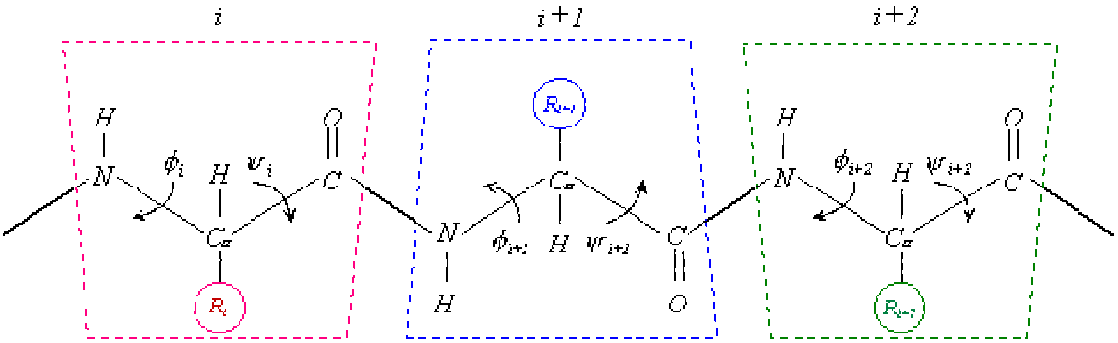


Figure 1. ϕ and ψ angles in the protein backbone. Courtesy of: http://hpcio.cit.nih.gov/protein/Foldin14.gif

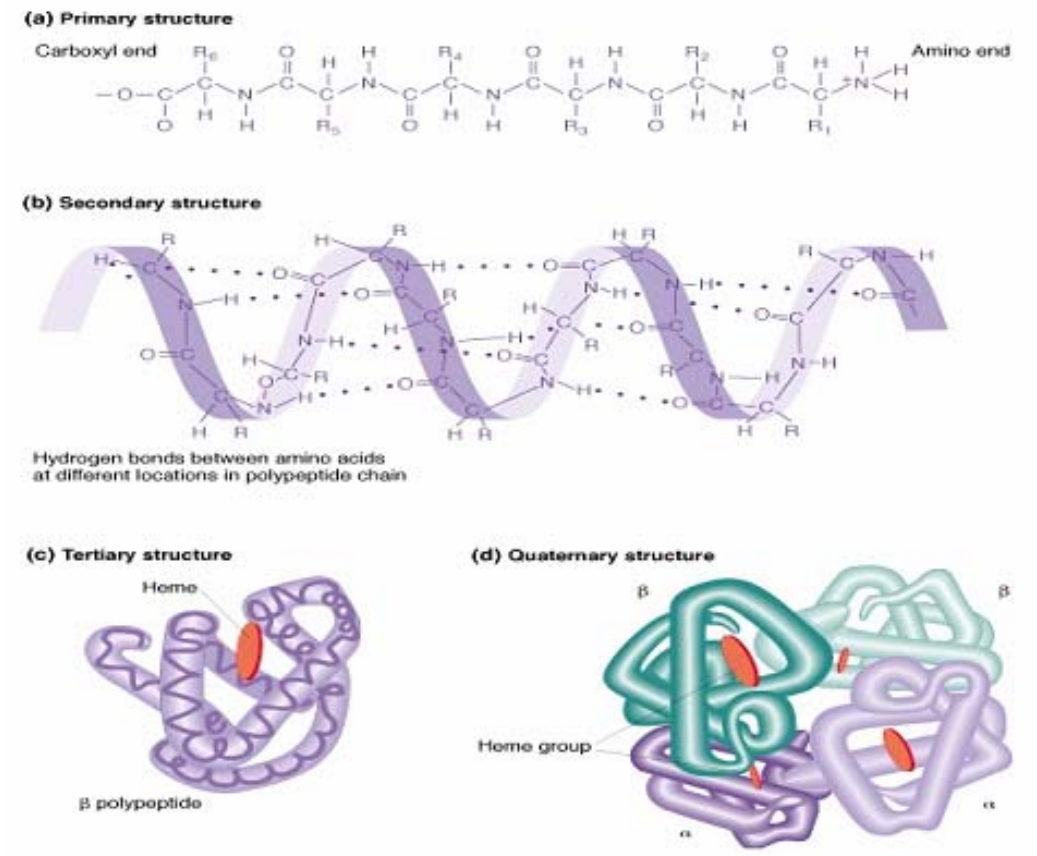


Figure 2. Four levels of protein structure. Courtesy of: http://www.bio.miami.edu/dana/104/proteinstructure.jpg

Tertiary structure is the packing together of secondary structure elements. Native tertiary structure implies the functionality of the protein - cavities can store cargos while arms can provide mobility to motor proteins - but one native structure can perform more than one task. Larger proteins can contain more than one polypeptide chain, joined together with a variety of bonding interactions including hydrogen bonding, salt bridges, and disulfide bonds. The association of two or more peptide chains into a multi-subunit structure defines the *quaternary structure*. The subunits can sometimes act cooperatively during their initial native formation and have slight conformational changes during chemical processes.[3]

The energy landscape of a protein determines its conformations. The protein folding community asserts that the native conformation of the protein is the energy landscape's global minimum. Natural selection, then, codified which sequences would lead to stable proteins necessary for biological functions.[4] On the other hand, folding is robust in the sense that changes in environment (pH, temperature, denaturant, etc) can still allow the protein to complete the folding process.[5]

Protein Folding

The process whereby a protein arranges itself into a unique three-dimensional structure is called *folding*. When all relevant biological processes are performing correctly, the proteins fold towards the native conformation. When proteins do not fold, or fold incorrectly, they can be responsible for prion related illnesses like Mad Cow disease, and amyloid related illnesses like Alzheimer's, by aggregating into insoluble plaques. Following the assumption that all the information for the tertiary structure is coded within the primary sequence, interference to proper folding must come through either sequence mutations during protein generation or external environmental deterrents. Investigations into the folding mechanism and pathways can pinpoint possible locations along the folding trajectory where misfolds can take place.

A naïve assumption would be that protein follow random paths upon folding, taking many random configurations to locate the native structure. This process would require multiple mistries, including many cases of time-consuming unfolding so that the protein can backtrack and try a different route. This slow process is quantified by the famous Levinthal's Paradox.[6] Say a particular protein is 100 amino acids in length. A highly underestimated count would give each amino acid two possible

conformations- the ‘right’ and ‘wrong’ ones (in reference to the native structure). If this were the case, this protein would need to visit as many as 2^{100} possible conformations to arrive at the native conformation. If one allows for 1 picosecond between configurational attempts, it would take as long as 10^{18} seconds – 10^{10} years! – to find the native state. Beyond the obvious biological problem of waiting the age of the universe for a single protein to fold, we already know that most proteins fold on the microsecond to second time scale. Unable to let folding try random configurations, nature must have built into the sequence of proteins another aspect that would always bias the folding towards the native state.

A proposal to resolve this ‘paradox’s is that the energy landscape of the protein is funnel shaped, where the width of the funnel is proportional to the conformational entropy at a given energy. The unique native state is located at the point at the bottom of the funnel. This funnel shape biases the folding pathways towards the native state, whereby minimizing the random conformational guesses the protein must take to progress.[7] The search difficulties for the folding pathway due to landscape roughness are partially simplified by this reduction of the entropy due to collapse.

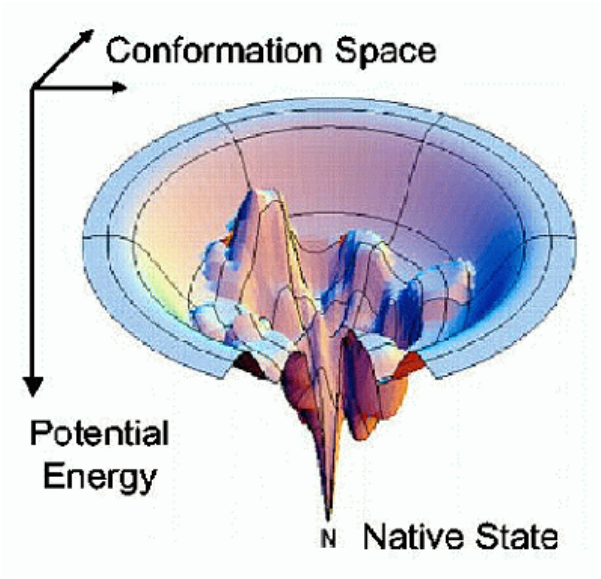


Figure 3. Protein Energy Landscape Funnel. Courtesy of: http://parasol.tamu.edu/groups/amatogroup/foldingserver/FAQ_Technique.php

Studies have been devoted to the understanding the collapsed form of the protein, referred to as it *molten globular* state - an intermediate between the extended chain of the unfolded form and the native conformation. Specific attention has been paid to quantifying the degree of collapse for this globular molecule.[8] There is no consensus yet as to the amount of secondary structure needed to define this globular state. The theory of Nucleation, for example, claims that these globular structures contain specific parts that act as a framework to guide folding into that structure. The folding rate can thereby controlled (by evolution or engineering) by placing specific amino acids into those positions that make the collapsed form stronger (or weaker, depending on your folding rate requirements).[9]

There are two main theories that describe global folding events. One scenario places secondary structure formation in the early folding events. The alternative approach places the initial collapse of the chain due to hydrophobic forces before secondary structure formation.[10,11] Experimental studies have seen both theories in effect, with larger occurrences of the former theory.[12,13]

This Study

The protein examined in this work is Barstar. This macromolecule is an inhibitor of the extracellular endoribonuclease barnase, and is found in the bacteria *Bacillus amyloiquefaciens*. Failure to express active Barstar is fatal to the bacteria. The Barstar-Barnase complex makes one of the tightest protein-protein contacts, making the pair an optimal study for protein-protein interactions.[14] Barstar alone serves as a model protein for folding studies. It is a single-unit 89-mer, containing four α-helices and a three stranded parallel β-sheet. The four Barstar helices span the following residues: Ser14-Ala25, Asn33-Gly43, Gln55-Thr63, Glu68-Gly81. The parallel strands span: Lys1-Asn6, Leu-49-Arg54, Asp83-Ser89. The majority of Barstar's residues are involved in secondary structure, and there is one extended loop between the first two helices which is used for binding with barnase.[15] Barstar contains two cysteine residues that tend to aggregate, whereby impeding crystallization of its wild-type. For this reason, a C40A and C82A mutant of the protein is often prepared for experimental use. Studies have shown that comparing solution structures of wild-type Barstar with crystal structure of the mutant type exemplify only subtle changes upon binding to barnase.[16]

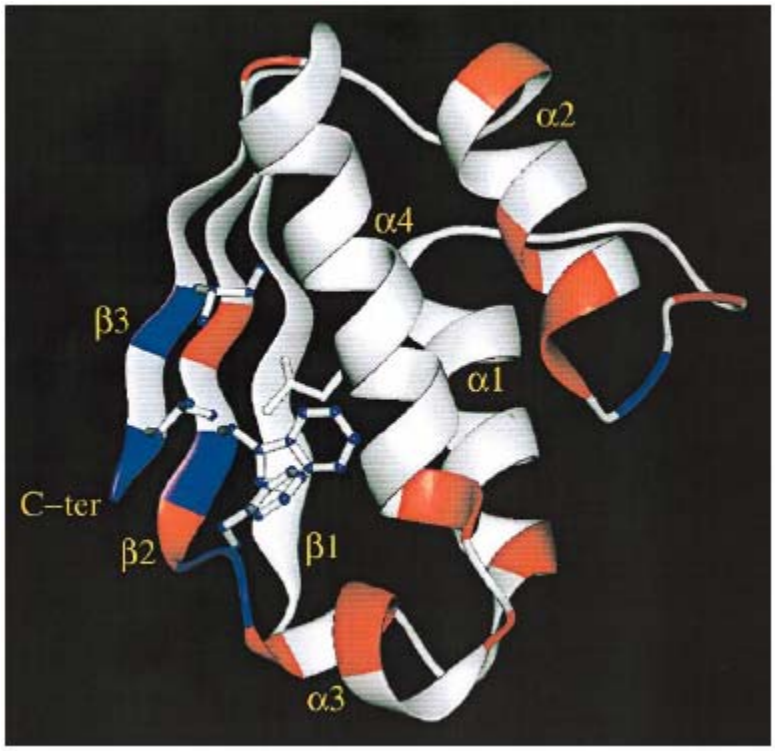


Figure 4. Native Barstar with labeled secondary structure motifs. Courtesy of: Wong, K.-B., Fersht, A.R., Freund, S.M.V. (1997) *J. Mol. Biol.,* **13**, 506

Early stopped-flow fluorescence studies depict multiple pathways and multiple early transient intermediates for Barstar folding, which are dependant on denaturant concentration.[17] One experimentally observed structural intermediate is a native-like conformation that has been shown to also be capable of inhibiting barnase activity.[18] Other reports show that there exist intermediates that are 'fast-refolding' unfolded states that differ from 'slow-refolding' unfolded states by *cis* and *trans* conformations, respectively, of the Tyr47-Pro48 bond.[19] The energy barrier to reach the native state with a *cis* Tyr-Pro bond is high enough to trap the *trans* intermediate for a significant length of time.[20]

Experimental studies have shown that cold-denatured Barstar is not completely in a random coil configuration; rather it has a preference for local structure.

Specifically, regions Ser12-Lys21 (part of the first helix), Try29-Glu46 (part of the second helix) and Leu51-Phe56 (part of the second strand) have residual structure in helical (ϕ,ψ) space while Barstar is unfolded. The former two regions are native-like, while the latter region is non-native when compared with the native β-sheet structure. These non-random pieces found in unfolded Barstar are considered to be potential initiation sites for protein folding; during folding, the sampling of conformational space is non-random for these regions, which can help overcome the Levinthal paradox by reducing the volume of conformational search.[21]

Other experiments, studying urea-unfolded Barstar, also find residual structure within the unfolded conformations. For example, five residues in Helix 1 show propensity towards populating helical regions in (ϕ,ψ) space. In contrast, most residues located between those involved in secondary structure are shown to have strong propensity towards populating beta-regions in (ϕ,ψ) space while the protein is unfolded. The regions corresponding to β-strands 1 and 2 indicate quick conformational averaging between regions in (ϕ,ψ) space, implying a lack of preferentiality. Strand 3, on the other hand, shows preference for non-native helical structure in the unfolded state.[22]

It has been thoroughly shown that as Barstar folds, it performs cooperative rapid hydrophobic collapse into a partially organized compact state, which then converts more slowly to the native state.[23] Through studies of folding pieces of Barstar that would natively hold secondary structure, it has been shown that hydrophobic collapse is a *necessary* precursor for the structure formation. Furthermore, the collapsed nucleus is centered on helix 1, which is almost completely formed at this stage in the folding.[24]

There have been many simulations done on the Barstar-barnase complex, such as Brownian dynamics[25]. To this author's knowledge there are no studies that compute a complete folding trajectory for Barstar on a computer. The computational approach taken by this work is to compute room temperature atomically detailed folding trajectories that do not assume a reaction coordinate or equilibrium state. To expedite the calculations, we modeled water solvation by a continuum – called the Generalized-Born model (GB).[26] This method has the obvious limits where structured water molecules would make a significant contribution to folding kinetics.

The goal for this study is to investigate the order of folding events for the entire folding trajectory of simulated Barstar, using the SDEL algorithm. Special focus will be on the order of folding events, to discern if, for Barstar, secondary structure forms during or after the hydrophobic collapse. In addition, we hope to see small-scale structural changes over the fold, such as order of secondary structure formation, and the structural propensity of residues.

CHAPTER 2

THE ALGORITHM

Computer calculations and simulations have become a powerful tool for the complex protein system. Ideally, all computer simulations of protein folding would be at the atomistic level, including the complete details of the solvent environment. However, the time step for molecular simulations can be as short as the femtosecond range, while overall general protein folding times can last from microseconds to seconds, which would leave such calculations expensive and impractical. In this vein, many simulations cut the folding time by either focusing in on specific parts,[27] or running high temperature unfolding runs.[28] Keeping to folding the entire chain, many minimalist models have been proposed[29] which included on-lattice[30] and off-lattice models.[31] As always, experiment is used to discern which parts of the model are artifacts and which belong to idealized or real proteins. Early potentials for these proteins involve the binary hydrophobic/polar possibilities (HP model[32]), with an assortment of potential strengths between constituents. Later, potentials involving the full variety of amino acids, or ones taken from a statistical distribution, were used. Some currently popular force fields include AMBER,[33] CHARMM,[34] OPLS,[35] and TIP3P for water solvation.[36]

Typical trajectory calculations either solve an initial boundary value problem (e.g. deterministic Newton's equations), or by using stochastic approaches (e.g. Langevin Equation). The Stochastic Difference Equation (SDE) algorithm[37] was devised to approximate classical trajectories for long time scale dynamics, based on the optimization of the action between two known end points.

The classical action parameterized by length has the form:

$$S = \int_{X_i}^{X_f} \sqrt{2(E - U(X))}\, dl \qquad (1)$$

X_i and X_f are the (mass weighted) coordinates for the initial unfolded and final folded conformations, respectively. E is the total energy, U is the potential energy, and dl is a (mass weighted) length element in the path. As a consequence to the *principle of least action*, folding trajectories with these fixed boundary points and total energy that produce a stationary S are considered to be optimal. In a discretized form, the above equation becomes:

$$S = \sum_{X_i}^{X_f} \sqrt{2(E - U(X))}\, \Delta l_{i,i+1} \qquad (2)$$

This step size $\Delta l_{i,\, i+1}$ is the distance between structures of the folding trajectory:

$$\Delta l_{i,\, i+1} = |X_i - X_{i+1}|$$

The goal is to use a sufficiently small step $\Delta l_{i,\, i+1}$ (within the usual computational price constraints) to get a reasonable approximation to the classical action. The optimization process will eventually lead to a succession of optimal structures that makes S stationary. The set of coordinates for these optimal structures are in fact slices along the trajectory as a function of the length index.

The *principle of least action* states:

$$\delta S / \delta X = 0 \qquad (3)$$

This statement implies that the first order variation of the action is minimal.[38] Therefore, to actually find the stationary solution, we optimize the norm of the gradient of the action:

$$T = \sum_i (\delta \mathrm{S}/\delta \mathrm{X_i})^2 + \lambda(\Delta l_{i,i+1} - <\Delta l>)^2 \qquad (4)$$

The second term on the right hand side is a penalty function, (where λ is an empirical constant) that ensures that our intermediate structures are equally spaced along the path:

$$(<\Delta l>) = \frac{1}{N+1}\sum_i \Delta l_{i,i+1} \qquad (5)$$

CHAPTER 3

COMPUTATIONAL METHOD

The main tool for our simulations was the molecular dynamics package MOIL. The most current publicly available version can be found at: http://cbsu.tc.cornell.edu/software/moil/moil.html. MOIL is a molecular modeling computer package, with the capabilities to run energy calculations, structure/path minimization, molecular dynamics, SDEL and other calculations.

Because in SDEL the action of a pathway between two conformations is minimized, the program's input can be either initial and final structures, or a pre-determined pathway to be optimized. For our calculations we chose the latter option. The structure of native Barstar was taken from the Protein Data Bank (1BTA), and no structural modifications were made. To obtain our collection of unfolded structures for the SDEL starting structures, we ran a high temperature (400K) MD run for 20,000 steps (1 femtosecond per step). We turned off electrostatic forces during the run, essentially eliminating hydrogen bonding and overall electrostatic attraction. After 20,000 steps, Barstar was completely unfolded into an almost linear chain. A set of 50 structures was sampled every 400 steps from the unfolding trajectory. This set of unfolded structures and the native structure of Barstar became the respective boundaries.

In principle, we could have started the SDEL calculations by generating initial guesses from these points. However, such calculations could be especially difficult since the term $\sqrt{(E-U)}$ can become imaginary if the initial paths include structures with potential energies higher than E. It was therefore useful to precede the SDEL calculation with a calculation of the minimum energy path. We used the SPW (Self-

Penalty Walk)[39] functional to compute minimum energy paths. The functional was optimized with conjugate gradient algorithm for 2000 steps. A Generalized Born[40] term was added to the calculations to mimic solvation with water molecules. As part of the SPW optimization, we added more intermediate structures within the pathways to ensure that the RMS between sequential structures was on the order of 0.4-0.8Å. For an analysis of how unique these paths are relative to each other, see Appendix II.

We decided to choose from this collection of paths only those where the RMS between initial and final structures was then limited to be no less than 5.5 Å; this eliminated paths whose 'most unfolded' structure was too structurally similar to the native conformation. We chose 12 final paths for SDEL, whose initially structures spanned the variety of unfolded configurations, from almost collapsed to completely linear. The RMS for the unfolded-to-native structures for these paths ranged from 5.8 to 17.2 Å. The number of structures per path ranged from 19 to 52, for a total of approximately 400 structures.

As seen in equation (2), SDEL requires a value for the total energy as an input. This value can be estimated from canonical equilibrium argument. We estimated to be –4000 kcal/mol, based on the average potential energy of a short simulation which was added to the average thermal kinetic energy 3N/2 KT. The path functional - equation (4) - was then optimized using a simulated annealing algorithm for the whole coordinates for each of the 12 paths. The annealing was done in cycles of 1000 steps, for a total of 5000 steps. The RMS between the paths from SPW and their SDEL counterparts were on the order of 0.2 Å.

The unfolding trajectory run was done on a Dell Latitude D600 laptop, and lasted about 48 hours. All SDEL calculations were run on the LINUX cluster at

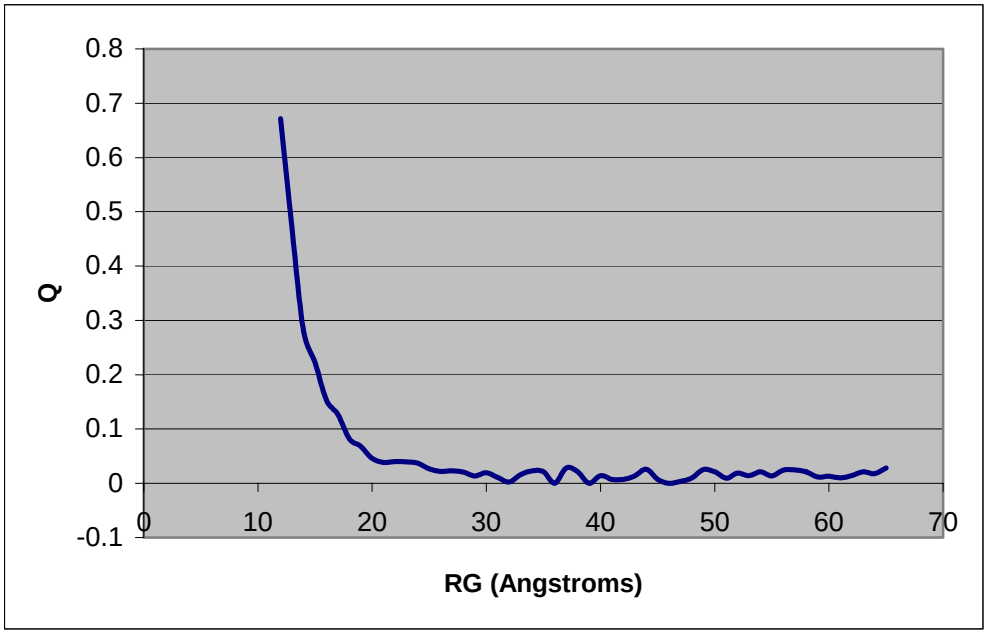
Cornell University's Computer Science Department. The number of CPUs for each path varied with a maximum of 47. The run time for the SDEL calculations for a single trajectory took no more than 24 hours.

CHAPTER 4

RESULTS

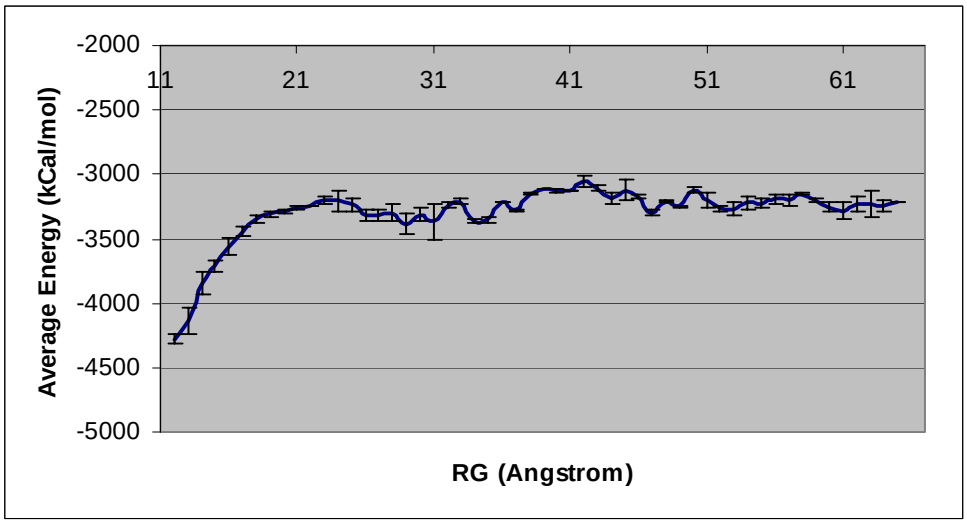
Once all 12 pathways were calculated, a choice of a progress variable was needed for proper data analysis. Using a time progress variable is an obvious first guess; however, SDEL minimizes the action functional over length and the length step is too large to allow interpolations to properly relate length to time. Choosing structure index as a progress variable would also not be useful, since the 12 paths do not have the same number of structures. In some cases, paths that begin with a structure of a larger Radius of Gyration (RG) have less intermediate structures than paths that begin with a structure of smaller RG. A third, and popular, option is to use Q (the fraction of contacts that are native) as the progress variable. Using Q often fails, though, when the rate of increase in Q is very far from being monotonic in folding time (or structure index). For Barstar, the Q for the trajectories remains close to zero until the molten globule state, which is most of the length of our trajectories.

In the end, we chose the protein's Radius of Gyration as our variable to track the reaction progress. Barstar's RG decreases consistently over the entire trajectory, with a *slower* decrease once the collapse occurs and internal structure begins to form. Graph 1 tracks the protein's Q as a function of RG. (To read folding trajectory graphs as a function of RG, one must follow the curve from right to left – from higher to lower radius of gyration). Q remains close to 0 for almost the entire path length, reiterating the lack of Q's utility for a progress variable.



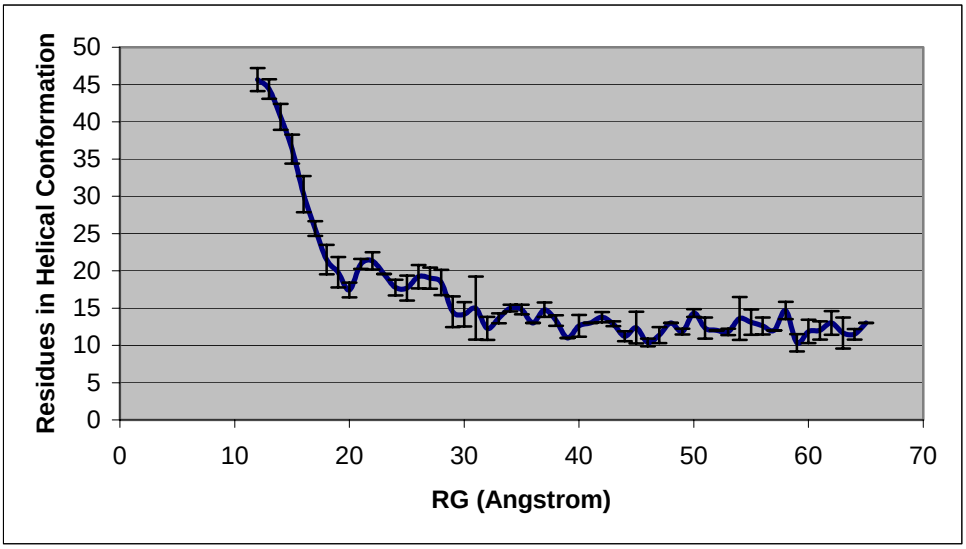
Graph 1. Q (averaged over the 12 paths) as a function of RG.

We also followed the energy of the protein over the entire fold. As can be seen from Graph 2, during the rapid collapse of the protein from a linear extension (RG of 65Å) until an RG of approximately 20Å the protein retains a relatively constant potential energy. The protein first collapses since the stretched configuration in entropically extremely unlikely, and the structure folds to a more probable self-avoiding configuration. From our graph and error-bars there does not seem to be an energetic barrier upon entry into the molten globular stage. But, this is our first clear indication that Barstar has two lengths scales for the folding path, with the end of the collapse occurring at an RG of around 20Å.

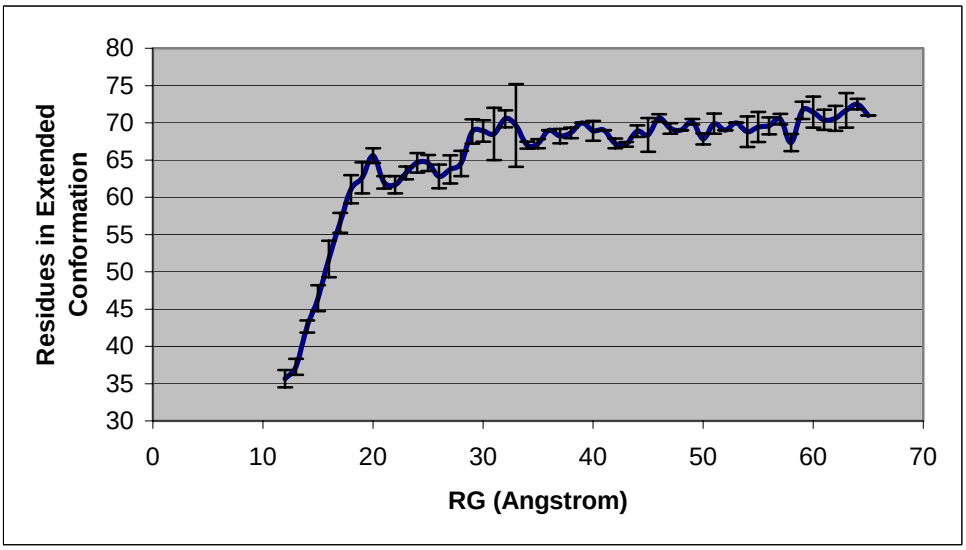


Graph 2. Average energy of the structures binned by RG.

In general, the residues within a protein remain in an extended conformation until residual structure begins to form. However, some residues can show propensity to *remain* in an extended conformation even during later folding stages, and some may show preference for helical conformations during early folding stages. To track this, one can follow the dihedral (ϕ,ψ) angles of all the residues as the folding progresses. Keeping with the broadest definition, we considered all (-ϕ,-ψ) as a helical residue and (-ϕ,+ψ) as an extended (beta) residue. Graphs 3 and 4, respectively, show the general count for all residues that take helical or extended dihedral values over the entire pathway. As expected, there is a sharp transition after an RG of about 20Å, where the number of residues in helical form dramatically increases (indicating the building of α-helical structure), and the number of residues in extended form dramatically decreases



Graph 3. Average count of residues taking helical dihedral angle values over the entire pathway.



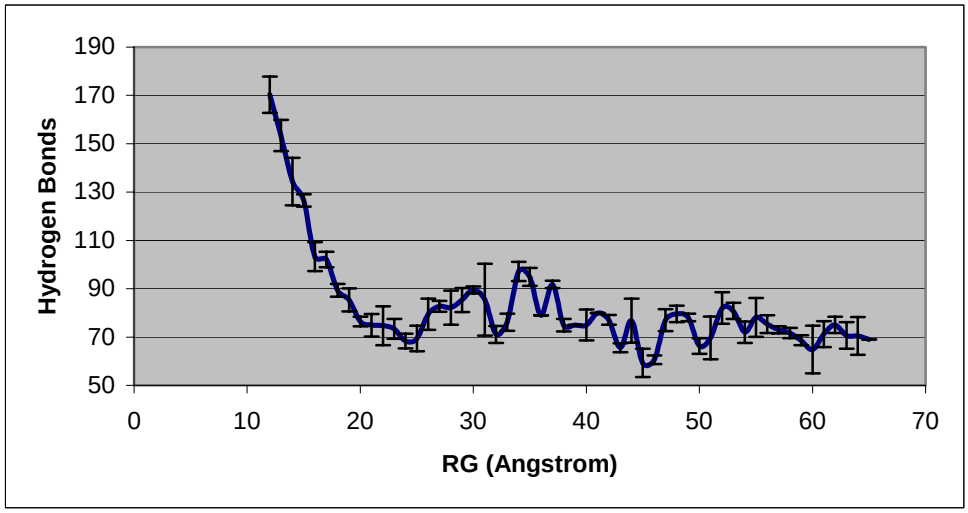
Graph 4. Average count of residues taking extended dihedral angle values over the entire pathway.

There are two interesting features of these graphs worth noting. First: one would normally expect a random-coil conformation of a protein *not* to take any

specific formation, even if this specific formation contained a heavy weight of residues in extended form. Since we started most of our simulation runs with structure that were linear or almost linear, the count for residues in extended conformation is the *largest* at the beginning of the trajectories. This number decreases and the protein folds, heading towards the final minimum value when the only residues left in extended conformation are those involved in forming the native β-sheets.

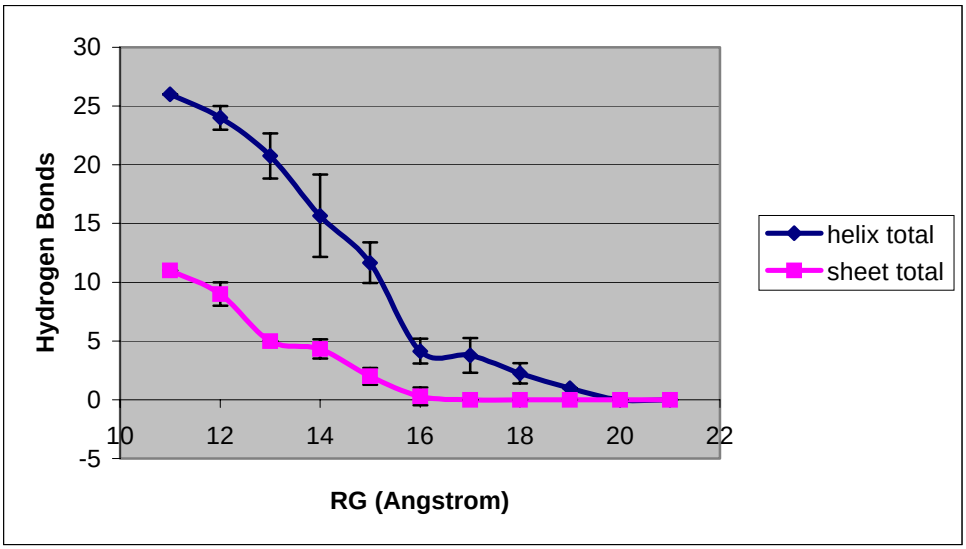
Second: In both Graph 3 and Graph 4, there seems to be an intermediary stage within the folding. Even within our error bars, the count for residues with helical conformation takes a slight jump up and for extended conformation take a jump down, at an RG of about 30Å. This would imply that for an intermediate stage the hydrophobic collapse is not yet done, yet residual structure is already beginning to form. These helical switches occur most densely in the low 20s and 70s residue range, or Helices 1 and 4.

To take yet another perspective, we also tracked the total number of hydrogen bonds between atoms (for backbone and sidechains) both involved and not involved in secondary structure formation. The results are shown in Graph 5. The h-bonds count seems erratic during the earlier folding events, implying that bonds are transiently being formed and often broken. The erratic count slows after an RG of 30Å, and the count sharply increases after the expected end of the collapse. *A priori*, one can claim that the final increase in the number of h-bonds corresponds to the building of more permanent secondary structure bonds.



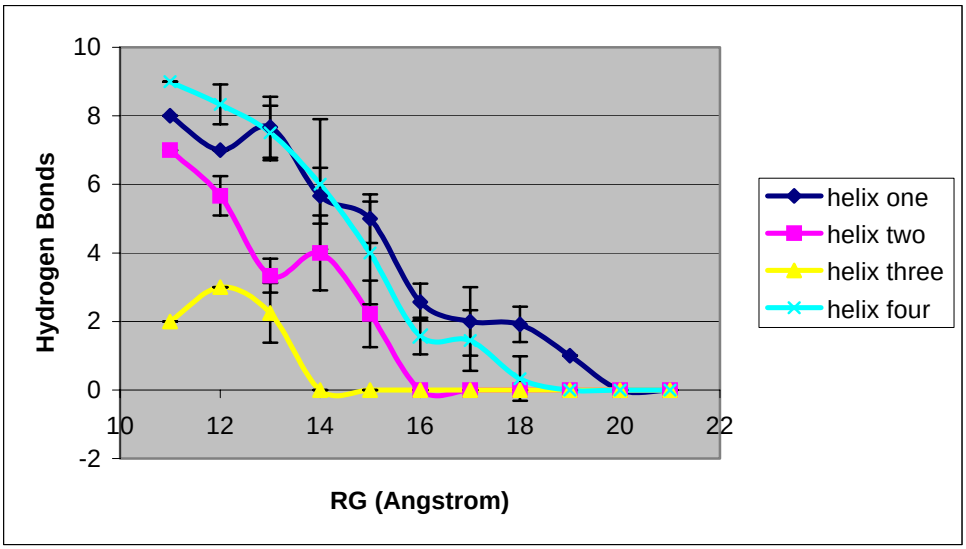
Graph 5. Average count for hydrogen bonds between all atoms for the entire pathway.

To focus in further, we also looked at the succession of hydrogen bond formation between residues that contain only native secondary structures. These results are shown in Graphs 6, 7, and 8. Graph 6 depicts h-bond formation for the collective set of helical residues and the collective set of sheet residues. Before the folding Barstar reaches an RG of 20Å, no native secondary structure forms – again points to the beginning of structure formation once the collapse stage has ended. In the range of folding RG between 20Å to 16Å, only the residues involved in native helical structure begin to lock into place, while the sheet residues are still structurally dormant. Only after an RG reaches 16Å do the sheets begin to form, at around the same rate as the helical formation.

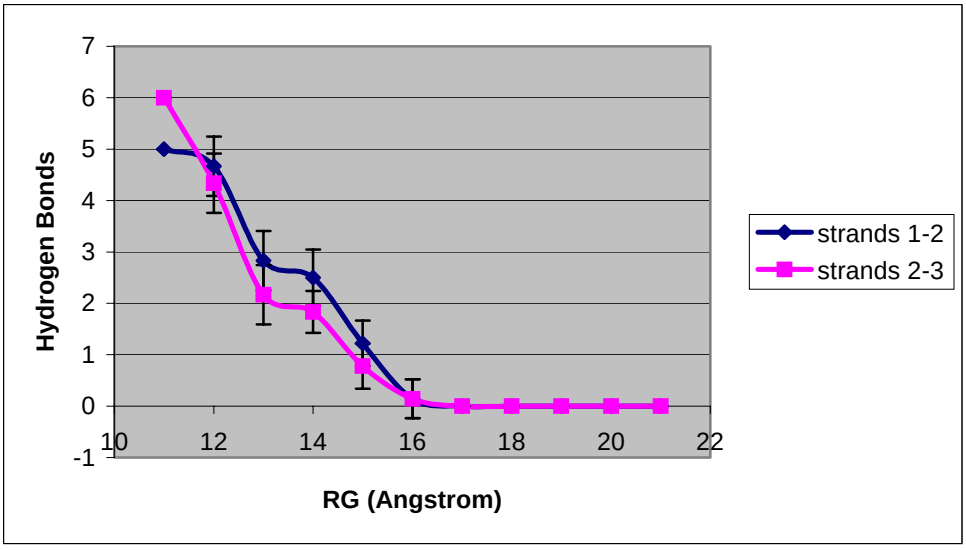


Graph 6. Hydrogen bond formation for residues with native secondary structure.

In Graph 7 we plotted the hydrogen bond formation count by specific helices. As expected, no native structure forms before an RG of 20Å. What is most interesting in this plot is the succession of native structure initiation. The order from our calculations is: Helix 1, Helix 4, Helix 2, and Helix 3. There is a significant lapse of 6Å of folding between the initiation of Helix 1 and Helix 3. In contrast, the results shown in Graph 8 depict the h-bonds between β-strands 1 and 2 and between β-strands 2 and 3 forming around the same RG (16Å), and their rate of growth is comparable.



Graph 7. Hydrogen bonds within single native helices.

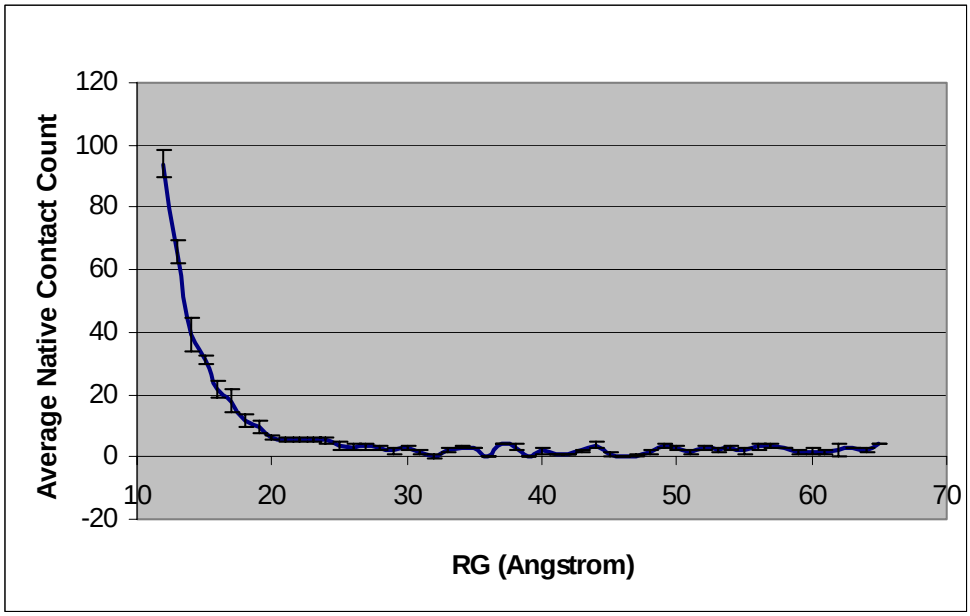


Graph 8. Hydrogen bonds between single native β-strands.

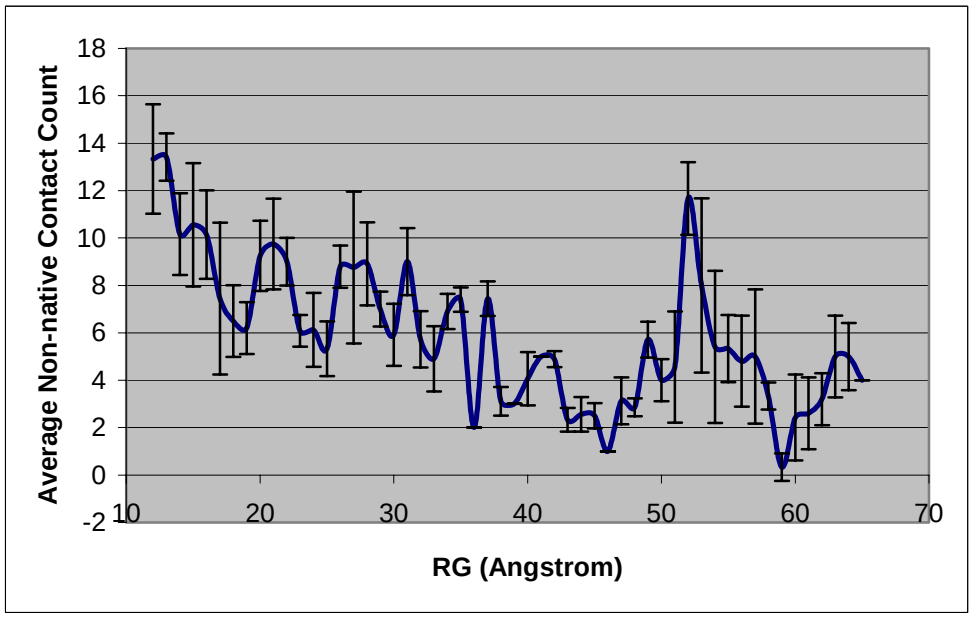
To quantify the overall folded tertiary structure, we calculated the number of native and non-native contacts within the protein over the folding pathway. We defined a contact as the event when the geometric centers of two residues are within

6Å of each other. In Graph 9, we track the native contacts over the folding trajectory. The shape of this curve is not surprisingly similar to Graph 1; native secondary and tertiary structure only begins once the protein has collapsed to an RG of around 20Å. Graph 10, however, is much more interesting. (The data-point for native Barstar – where the count drops to zero – is not shown.) In this graph we follow the non-native contacts over the entire folding pathway. Two elements of this graph are worthy of note. First, the erratic nature of this graph implies that non-native contacts are continually being formed and broken as the folding occurs. This suggests trials the protein attempts before directing itself forwards towards the native conformation.

Second, the RG 'area' for native and non-native contacts has significant overlap, and the number of non-native contacts tends to *increase* as the folding progresses. (The native structure – where the count drops to zero – is not shown). This is counter-intuitive since one would expect this number to remain constant as these contacts are periodically made and broken. Indeed, none of these non-native contacts remain constant throughout the trajectory. However, there is a high reoccurrence of the non-native contacts between residue n and n+3 for those involved in Helices 1, 2 and 4, Sheet 2 and the loop connecting Helix 1 and 2. In one trajectory, residues 47 and 50 occur in non-native contact for 30% of the structures.



Graph 9. Count of native contacts formed during the folding pathway.



Graph 10. Count of non-native contacts formed during the folding pathway.

The following (Figure 5) are visual representations of Barstar over the pathway. The radius of gyration for these structures, respectively, is: 66 Å, 59 Å, 26 Å, 20 Å, 17 Å,

15 Å, and 12 Å (native Barstar). The fourth structure is considered to be the transition structure where the collapse stage has ended and secondary structure begins to form.

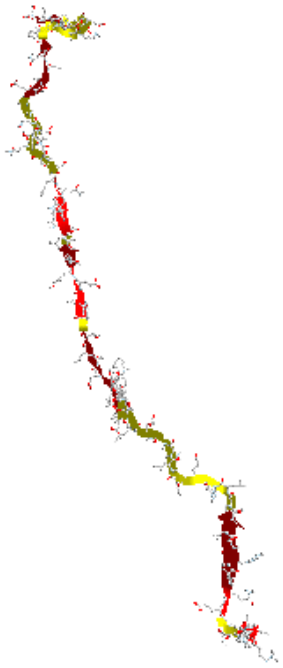


Figure 5. Seven structures along the folding pathway

Figure 5 (Continued)

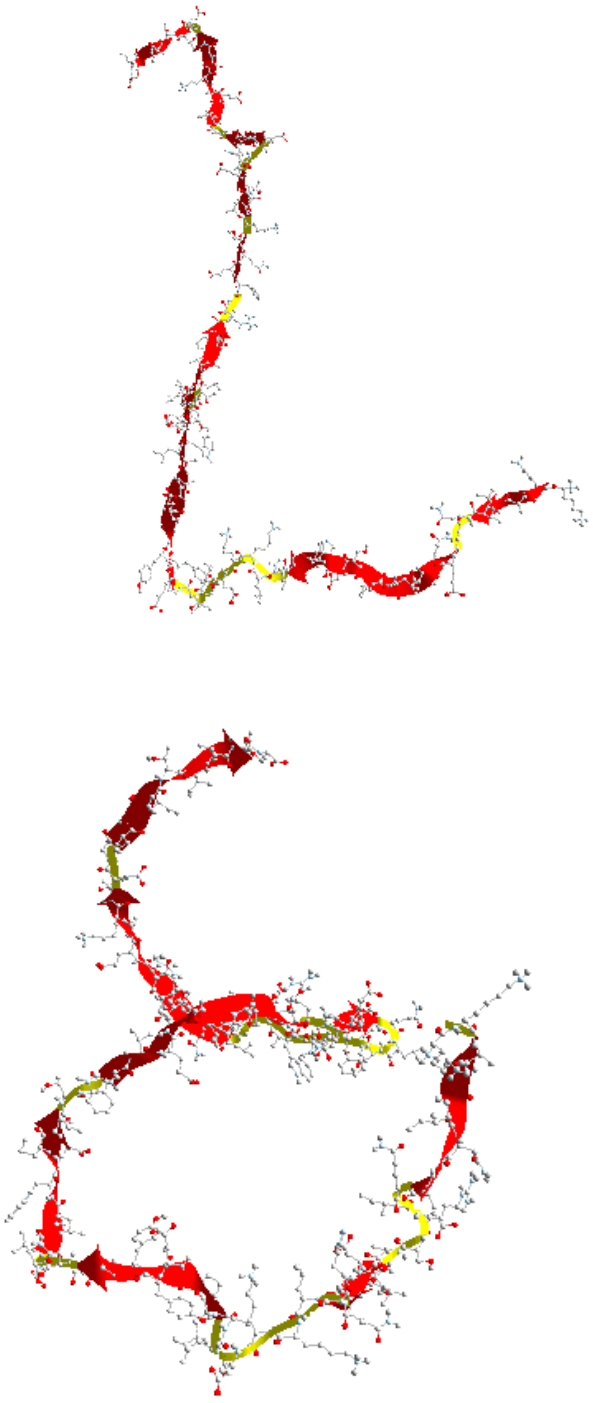


Figure 5 (Continued)

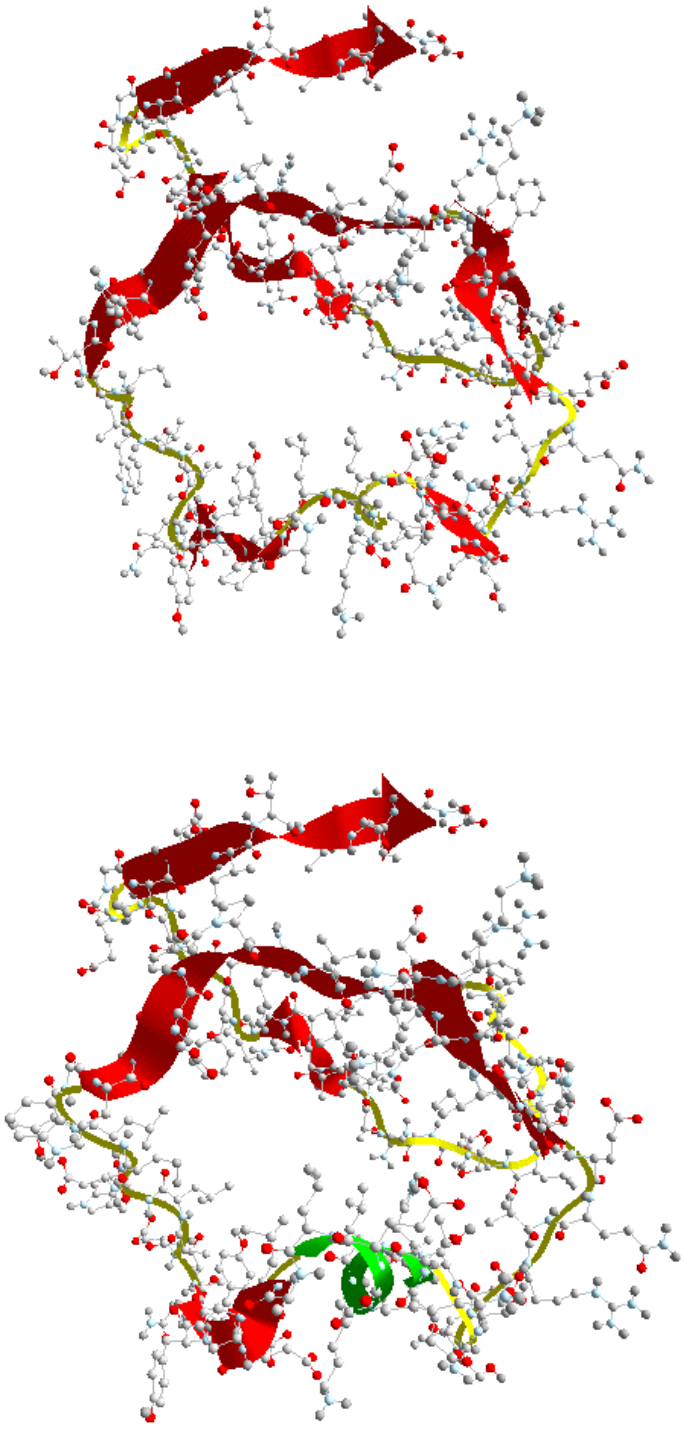
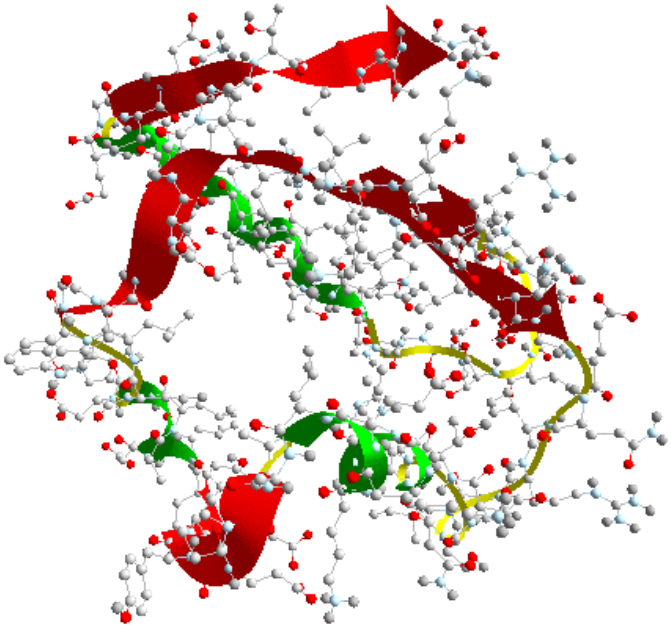
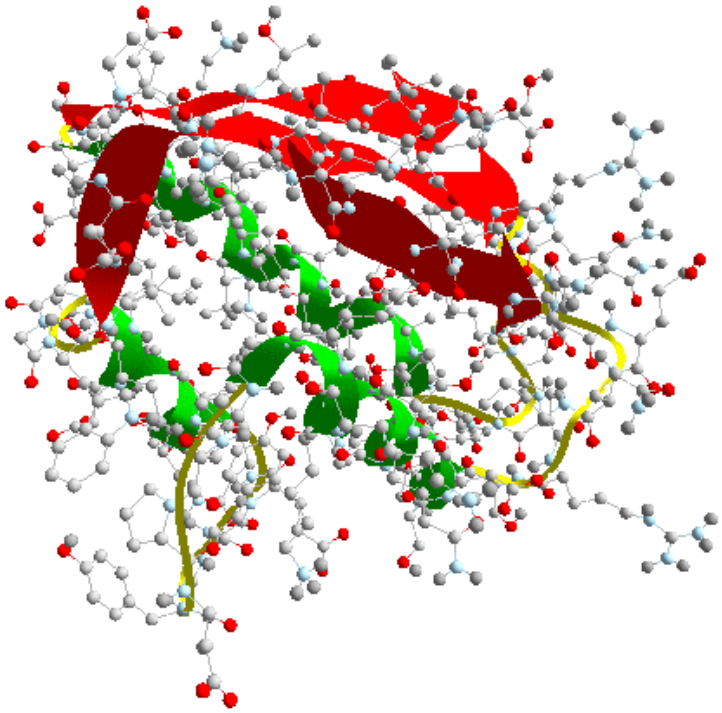


Figure 5 (Continued)





CHAPTER 5

DISCUSSION

It has been thoroughly shown by experiment that as Barstar folds, it performs cooperative rapid hydrophobic collapse into a partially organized compact state, which then coverts more slowly to the native state. According to a CD experiment by Agashe et. al. (1995),[41] the Barstar chain rapidly collapsed (within 4ms) to a globular form without any optically active secondary (or tertiary) structure. As seen from virtually all the graphs in the preceding section, our simulated Barstar has two folding length scales with a transition at a radius of gyration of around 20Å. Before this transition, the molecule is undergoing pure collapse, as the RG heads from the extended 65Å to 20 Å. Graphs 8-10 add that during this collapsing stage no secondary structure is being formed.

Using NMR techniques, Bhavish et. al. (2004)[42] discovered that in urea-unfolded Barstar, the majority of residues display properties indicative of propensities toward extended conformations. This is true for the simulated Barstar as well. Graph 4 depicts the high concentration of residues located in the extended (beta) conformation, for the majority of the pathway, (however this could just be an artifact from the fact we started some of our trajectories with extended chains that are entropically unlikely). This is not the case for a pure random-coil configuration, where there is no preference, even for extended configurations. These NMR experiments also observed that Helix 1 contained five (out of 11) residues that strongly preferred a helical conformation. The conclusion of those authors was that Helix 1 might be considered an initiation site for folding. Graph 7 clearly shows that Helix 1 is the first to begin

forming in our simulation, out of all possible secondary structure locations. Furthermore, the authors say the regions corresponding to β-strands 1 and 2 indicate quick conformational averaging between regions in (ϕ,ψ) space, implying a lack of preferentiality. This could explain why – as seen in Graph 6 – our β-sheets begin folding much later that the initiation of helical forming.

In Graph 10, the non-native contacts within the simulated Barstar appear to transiently form and break over the early collapse stage. Also, when comparing to Graph 9, we can see that there is significant overlap in the region where native contacts drive the folding and where non-native contacts drive the folding. Even after the transition point of an RG of 20Å when the number of native contact quickly climbs, the count for non-native contacts *continues* to climb. In fact, there is high reoccurrence of specific non-native contacts between residue n and n+3 for Helices 1, 2 and 4, Sheet 2 and the loop connecting Helix 1 and 2. In one trajectory, residues 47 and 50 occur in non-native contact in 30% of the structures. It would seem that these non-native contacts are an important part of secondary structure formation. While it is not clear from our Graph 2 where the energy barriers are for Barstar folding, the non-native contacts may possibly act continually throughout the folding process to facilitate crossing the barriers shown in experimental papers. It can be proposed that the erratic yet consistent formation of the observed specific non-native contacts acts as a built-in annealing technique for the protein to overcome barriers during folding. In this sense, these contacts could be intentional alignments instead of random configurational attempts.

CHAPTER 6

CONCLUSIONS

This study has successfully found strong correlations between experimental data for folding Barstar and trajectories found using the SDEL computational protocol. As with experiment, our folding trajectories depicted two length scales for Barstar folding, where the protein undergoes a collapse stage prior to forming secondary structure. During the simulated protein's collapsing stage, residues involved in secondary structure formation – particularly within Helix 1 – held a propensity for helical conformation. Once collapsed, the order of secondary structure formation follows closely to the succession seen from experimental studies.

There are two caveats to the above successes. First, the same experimental studies that conclude that Barstar's Helix 1 retains a strong helical propensity in the protein's unfolded state, find similar characteristics within Helix 2. Our pathways have Helix 2 forming later in the trajectory, after Helix 4. Second, the number of intermediate structures within our pathway is significant, yet the step size between structures is large and the trajectories, therefore, can only be treated as approximate.

SDEL's strong correlation with experiment bolsters the assertion that it is a powerful computational tool for approximating molecular dynamical pathways. These pathways can be folding trajectories, motor protein action strokes, or the dynamics of prion and amyloid fibrils. Future steps on this particular project would include introducing more intermediate structures within the simulated trajectories – especially within the latter folding stage – whereby enriching the information garnered from the paths. An interesting sub-project – not studied here – would be to investigate the bonds and contacts between sidechains alone. Finally, to complete this study with a

look at the rates for each folding stage, one would need to reintroduce a time scale to the trajectories. This can be accomplished by the Milestoning technique[43]currently available for use in the MOIL package.

APPENDIX I

MORE ON SDEL

Structural Orientation

For the SDEL algorithm to work properly, it factors out from individual structures in Cartesian space along the trajectory their overall translations and rotations. This is accomplished by subjecting the minimization of the target function in equation (4) to the linear constraints:

$$\sum_j x_{ij} = 0 \qquad (6)$$

where x_{ij} are the mass weighted Cartesian coordinate vectors of atom j in structure i, and

$$\sum_j (x_{ij} \times x_{ij}^0) = 0 \qquad (7)$$

where x_{ij}^0 are the coordinates of the initial structure (X_i) that is used to define the "laboratory" reference frame.

More details on the simulated annealing optimization of SDEL can be found in: Cárdenas, A., Elber, R., (2003). *Biophysical Journal*, **85**, 2919-2939.

APPENDIX II

COMPARING PATHWAYS

Contact Metric

The SPW algorithm (mentioned above) produces least energy pathways between two molecular conformations. A logical next step would be to check the uniqueness of the paths. While the initial structures for the paths vary greatly, there is no clear intuition as to where these trajectories may converge along the folding landscape.

To study this, a *contact metric* was devised to compare pathways. In essence, the metric measures the number and constancy of specific internal contacts that two paths share along their respective folding trajectory. The greater similarity of the contacts would imply a degree of non-uniqueness between the paths.

We first define a way of counting contacts:

$c_k^a = 0$ if not in contact

$c_k^a = 1$ if in contact (12)

where *k* is a specific contact pair and *a* is the path index. Averaging over the path:

$$n_k^a(t^a) = \frac{1}{N}\sum_t c_k^a \quad (13)$$

where *t* is the structure index, and *N* is the total number of structures in path a. The metric to compare two paths (*a* and *b*) is defined as:

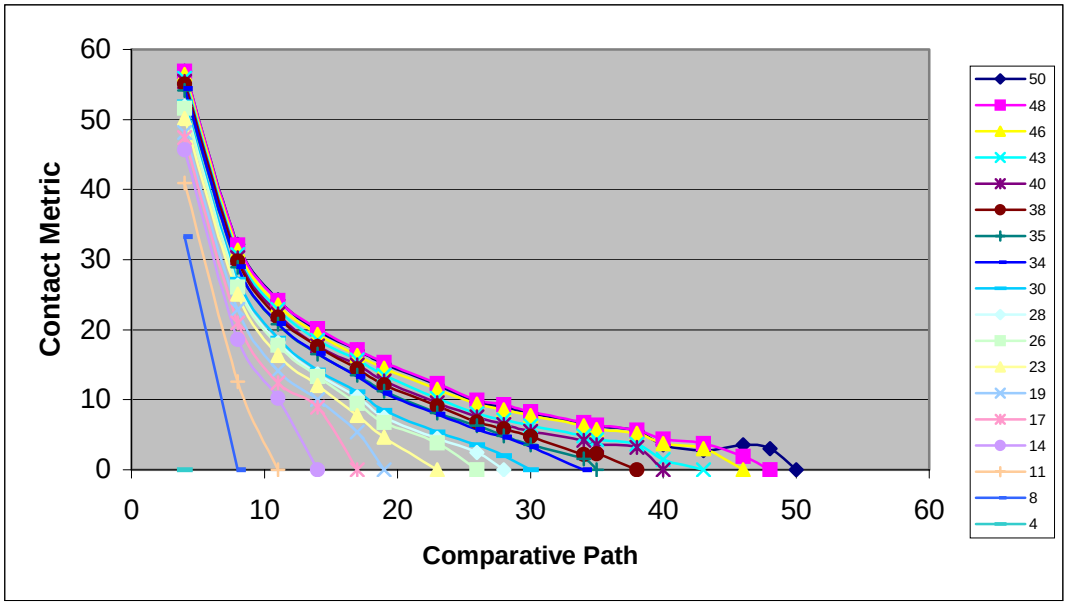
$$m_{a,b}(t) = \sum_k | n_k^a(t^a) - n_k^b(t^b) | \quad (14)$$

A smaller $m_{a,b}(t)$ implies that the paths are more similar. When the metric equals zero, the paths are identical.

This contact metric was calculated for 20 paths resulting from the SPW algorithm.

Results

Graph 11 depicts the contact metric count for each of the twenty paths as related to all other paths.



Graph 11. Contact Metric for each path related to all other paths

This graph shows the strong dissimilarity of paths that begin from very different structures. For example: Paths 46, 48 and 50 all converge with each other rather quickly, however they converge with paths 26, 28 and 30 only further down in their trajectories towards paths 4 and 8. More exactly: paths 46, 48 and 50 become *as similar* to paths 4 and 8 as are paths 26, 28 and 30 are to paths 4 and 8. In contrast, paths 34 and 35 begin with very similar structures (RMS of 0.6Å). These paths

converge very quickly, i.e. are equally as similar to all subsequent paths. In the end, it is safe to conclude that our selection of paths spans a wide enough selection of paths to make our path selections independent.

REFERENCES

[1] Nelson, D. L., Cox, M.M. (2005).*Lehninger Principles of Biochemistry*. W.H. Freeman and Company, NY

[2] Rost, B., Sander, C., (1993). Proc. Natl. Acad. Sci. USA, **90**, 7558-7562

[3] Nelson, D.L., Cox, M.M., ibid.

[4] Onuchic, J.N., Luthey-Schulten, Z., Wolynes, P.G. (1997). Annu. Rev. Phys. Chem. **48**, 545-600

[5] Onuchic, J.N., Nymeyer, H., Garcia, A. E., Chahine, J., Socci, N.D. (2000). Adv. Prot. Chem. **53**, 88

[6] Levinthal, C., (1968). J. Chem. Phys. **65**, 44-45.

[7] Onuchic, J.N, Wolynes, P.G., Luthey-Schulten, Z., Socci, N.D. (1995) Proc Natl Acad Sci USA, **92**, 3626-3630

[8] Chan, H.S., Dill, K.A. (1991). *Annu. Rev. Biophys. Chem.* **20**, 447-490

[9] Mirny, L. Abkevich, V., Shakhnovoch, E. (1998). *Proc. Natl. Acad. Sci. USA.* **95**, 4976-4871

[10] Englander, S.W., Sosnick, T.R., Mayne, L.C., Shtilerman, M., Qi, P.X., Bai, Y. (1998). Acc. Chem. Res. **31**, 737-744

[11] Fersht, A.R., Matouschek, A., Serrano, L. (1992). J. Mol. Biol. **224**, 771-782.

[12] Honig, B., Cohen, F.E., (1996). Fold. Des. **1**, R17-R20.

[13] Fersht, A.R. (1997).Curr. Opin. Sruct. Biol., **7**, 3-9

[14]Hartley, R.W., (1989). *TIBS*, **14**, 450-454 .

[15] Lubienski, M.J, Bycroft, M., Freund, S.M.V. & Fersht, A.R. (1994). *Biochemistry*, **33**, 8866-8877.

[16] Buckle, A.M., Schrieber, G. & Fersht, A. R. (1994). *Biochemistry*, **33**, 8878-8889.

[17] Shastry, M.C.R., Udgaonkar, J.B. (1995). *Journal of Molecular Biology*, **247**, 1013-1027.

[18] Shastry, M.C.R., Agashe, V.R. & Udgaonkar, J.B. (1994). *Protein Science*, **3**, 1409-1417.

[19] Schreiber, G., Fersht, A.R. (1993). *Biochemistry*, **32**, 11195-11203.

[20] Killick, T.R., Freund, S.M.V., Fersht, A.R. (1999). *Protein Science*, **8**, 1286-1291.

[21] Wong, K-B, Freund, S.M.V., Fersht, A.R. (1996). *Journal of Molecular Biology*, **259**, 805-818.

[22] Bhavesh, N.S., Juneja, J., Udgaonkar, J.B., Hosur, R.V. (2004). *Protein Science*, **13**, 3085-3091.

[23] Agashe, V.R., Shastry, M. C. R., Udgaonkar, J.B. (1995). *Nature*, **377**, 754-757.

[24] Nolting, B., Golbik, R., Neira, J.L., Soler-Gonzales, A.S., Schreiver, G., Fersht, A.R. (1997). *Biochemistry*, **94**, 826-830.

[25] Spaar, A., Dammer, C., Gabdoulline, R.R., Wade, R.C.& Helms, V. (2006). Bio*physical Journal*, **90**, 1913-1924.

[26] Tsu, V., Case, D.A., (2000). *Biopolymers*, **56**, 275-291.

[27] Hirst, J.D., Brooks, C.L. (1995). *Biochemistry*. **34**, 7614-7621

[28] Hunenberger, P.H., Mark, A.E., van Gunsteren, W.F. (1995). *Biochemistry*. **31**, 7745-7748

[29] Covell, D.G., Jernigan, R.L. (1990). *Biochemistry*. **29**, 3287-3284

[30] Lau, K.F., Dill, K.A. (1989). *Macromolecules*. **22**, 3986

[31] Hao, M-H, Scheraga, H.A. (1994). *J. Phys. Chem.* **98**, 4940-4948

[32] Chan, H.S., Dill, K.A. (1998). *Proteins: Structure, Function and Genetics* **30**, 6388-6392

[33] Cornell et. al. (1995).

[34] MacKerell, A.D., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J,D., et. al. (1998). *J. Phys. Chem. B.* **102**, 3586-3616

[35] Jorgensen, W.L., Maxwell, D.S., Tirado-Rives, J. (1996). *J. Am. Chem. Soc.* **118**, 11225-11236

[36] Jorgensen, W.L., Chadrasekhar, J., Madura, J.D., Impey, R.W., Klein, M.L. (1993). *J. Chem. Phys.* **79**, 926-935

[37] Elber, R., Ghosh, A., Cerdenas, A. (2002). *Journal of Chemical Physics*, **35**, 396-403.

[38] Goldstein, H., Poole, C., Safko, J. *Classical Mechanics, 3rd edition*. (2002). Addison Wesley, CA.

[39] Czerminski, R., Elber, R. (1990). *Int J Quantum Chemistry*, **24**, 167-186

[40] D.T., Eisenberg; L. Wesson (1992). *Protein Sci.*, **1**, 227235.

[41] Agashe, V.R., Shastry, M. C. R., Udgaonkar, J.B. (1995). *Nature*, **377**, 754-757.

[42] Bhavesh, N.S., Juneja, J., Udgaonkar, J.B., Hosur, R.V. (2004). *Protein Science*, **13**, 3085-3091.

[43] Faradjian, A.K., Elber, R. (2004). *Journal of Chemical Physics*. **120**, 10880.