Data Analysis of Villin Headpiece Subdomain Folding Simulations.

Jagannath Krishnan

Department of Computer Science,
Stanford University
March 2002

Abstract

The Folding@Home project simulates protein folding in atomistic detail using a distributed computing approach. My project explores the folding of the villin headpiece by analyzing data generated by the Pande group at Stanford, using the Folding@Home approach. A total of 226.752 microseconds of folding time were simulated and 21 independent folding trajectories were observed. Starting from an extended state a relaxation to an unfolded state is observed which is accompanied by the occurrence of at least one alpha helix. Plateaus in the energy landscape are also observed in nearly all the folded trajectories till the final energy barrier is crossed by which time the hydrophobic core is formed. Helices are observed to form, break and reform, some being more stable than others. Some of the folded structures were seen to have at least 2 of the 3 helices, and rmsd < 3.0 when compared with the experimentally obtained structure (1VII.pdb).

The computationally 'impossible' task of simulating the folding of proteins in atomistic detail is being made possible by the use of a radically new distributed computing approach in the folding@home project. Tens of thousands of computers collaborate to generate megabytes of data as the story of each atom is recorded during the folding process. This project seeks to understand the process of protein folding by analyzing the vast amount of data generated while simulating the folding of the villin headpiece.

Introduction

Protein folding has been called one of the greatest scientific challenges of our time [1]. The limitation of present day experimental methods in determining the structures of proteins has fueled a great deal of interest in trying to predict the structure of a given protein using computational methods. Ab initio structure prediction is the Holy Grail of this field because it empowers the biologist with the ability to predict the structure of even completely new proteins unlike homology or threading based approaches. Protein folding studies the folding trajectory of protein molecules from expanded to native conformation. Understanding the process of folding will not only enable us to predict structure but will also permit us to synthesize new molecules that fold into desired shapes and has implications in drug design.

Until recently ab initio structure prediction has been dismissed as a pipe-dream. This is partly because of the difficulty in modeling the folding process in software but more because the computational prowess to carry out the simulations simply does not exist yet. A recent simulation of the villin headpiece that was reported simulated the events in 1 microsecond of folding time[2]. This project used a massively parallel supercomputer to speed up the process of simulation. This approach has its limitations as the amount of communication between the processors is excessive and limits the speedup that can be achieved. The time simulated using this approach falls short of the folding time of even very small molecules.

The Folding@Home project attacks the problem using a distributed computing approach. It is based on the insight that given a very large number of molecules that are simulated, each with slightly different forces from the solvent, the probability that a few of the molecules will overcome the energy barrier to reach the native conformation is very high. Small protein molecules are thought to have no intermediate stages and cross the energy barrier just once. For larger molecules with multiple energy barriers the fact that one of the trajectories has crossed a barrier is used to restart all the trajectories from that state. Thus individual simulations don't need to communicate with each other and given M processors there is a factor of M speedup in the simulation [3]. The simulations are distributed among tens of thousands of processors worldwide and communicate with a central server that gathers data.
The aim of my project is to analyze the simulation data obtained for the villin headpiece and to learn about its folding pathways. Most of the data is present is a MySQL database and has been analyzed using Perl scripts I wrote that connect to the database and generate data which is then examined using Mathematica. The data describes various properties of the molecule at each nano second as the molecule goes from an unfolded to a folded state.

In the next subsection the villin headpiece is introduced. Section 2 presents the overall characteristics of the folding process. Section 3 singles out a representative folded trajectory and analyzes it in detail. Section 4 analyzes presents an ensemble level analysis of the trajectories. Section 5 summarizes the results. References are listed in section 6 and acknowledgements follow in section 7.

Villin Headpiece Subdomain

The Folding@Home project has simulated the folding of the villin headpiece subdomain, which is a 36-residue fast folding protein. It is the C terminal domain of the villin actin binding protein. Being a fast folding protein it has been studied both experimentally and by simulation although previously reported simulations are not as long as the one reported in this paper. The molecule being studied has 3 alpha helices joined by short turns (pdbcode 1vII). The experimentally determined structure is shown in figure 1. This figure also shows the four Phe residues, three of which form the hydrophobic core of this molecule.

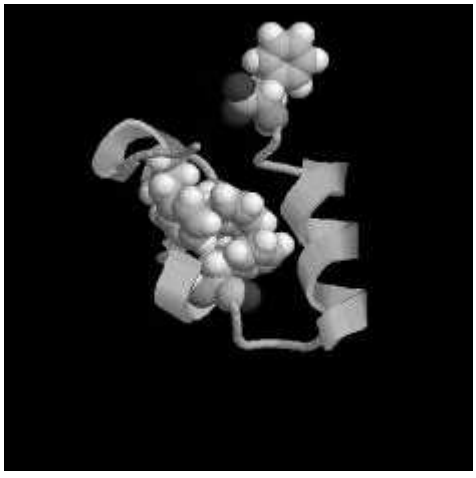

Figure 1. Structure of the villin headpiece subdomain. Note the alpha helices at each terminal and also the single turn helix in the middle. Three of the four Phenylalanine's come together to form the core.

2.Characteristics of the folding process

This section elaborates upon the method used to simulate the folding and introduces terminology that is used in describing the results. It also describes the overall manner in which the protein folding data can be viewed.

As mentioned in section 1, the basic idea behind Folding@Home is to start numerous simulations in tandem. The forces on the molecule are modeled in atomistic detail and are recalculated on the order of femtoseconds (e-12 seconds). The coordinates of all atoms and properties of the molecule like the radius of gyration, the solvent accessible surface area(SASA), the rmsd from 1VII.pdb, energy etc. are all recorded in the database after each nanosecond. This snapshot at each nanosecond is referred to as a frame. A complete simulation run which has a record at each nanosecond is called a trajectory since it describes the properties of a molecule over time as it undergoes conformational changes. The solvent molecules play a key role in protein folding as they collide with the atoms of the protein causing perturbations that constantly change the forces on the protein's atoms. The Generalized Born implicit solvent model is used to simulate the effect of the solvent molecules in this folding simulation. Thus for each trajectory (or simulation run) the protein is modeled in its extended conformation. Slightly different forces due to the solvent are applied on the protein and the simulation is started. These differences in forces cause the trajectories to fold in different ways.

On analyzing the data it is observed that all the trajectories start out in the extended state (E) and in a few frames collapse to a more relaxed state with lower energy. Early appearance of secondary structure is also noticed in this state as at least one alpha helix can be seen to have formed by this time. This state will be referred to as the unfolded state U in the rest of this paper. The state allows us to study the various conformations that the molecule could be in between the extended state and the folded state and provides important information about the intermediate stages of the folding process. On an average with the passage of time the total potential energy of the molecule decreases and formation of the hydrophobic core is seen as well as formation of additional secondary structures. The relative order in which these events occur is described in a later section. In this analysis we distinguish between the composite data set that describes all the trajectories and the folded set which describes only those trajectories that successfully folded.

3. Analysis of a representative folding trajectory

One major difference between the experimental and computational techniques of studying protein folding is that the experimental techniques study the properties of only groups of molecules. Due to this the structures that are observed experimentally are rough averages of quantities like bond lengths and bond angles over many molecules. The simulation method described in the previous section provides a window into the world of this molecule. The data for any given trajectory reveals that bonds are continually being formed and broken; helices form, disappear and reform. As the molecule undergoes conformational changes per frame its other observable properties like the radius of gyration and its rmsd from the 'native' (experimental) structure vary. Since the molecule is never at rest it is difficult to define just what parameters qualify a particular conformation to be a part of the unfolded or folded states. For the purpose of this analysis the folded set is defined as those trajectories that have at least one frame in which the rmsd of the molecule is below a threshold value of 3 and 10 or more residues are part of a helix. By these standards the simulations resulted in 21 complete and independent folding trajectories of folded final structures.

In the remainder of this section we look at one member of this set and study the way in which it folded.

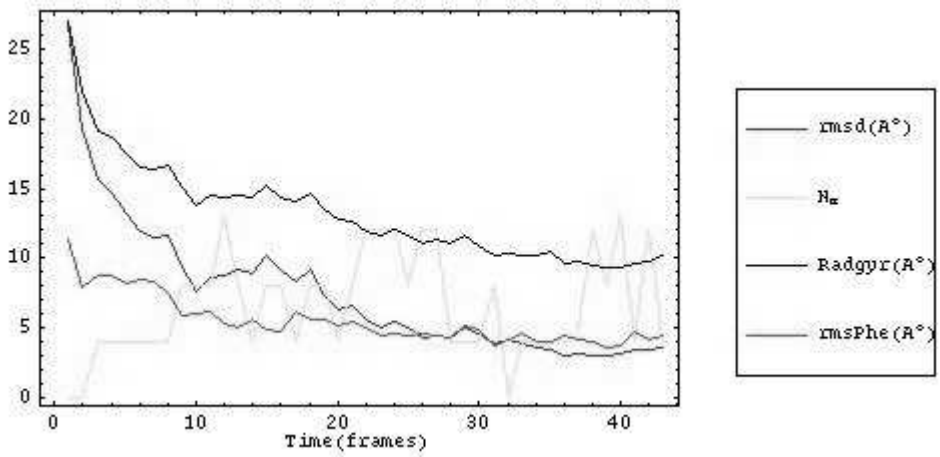

Figure 2. A summary of various values plotted against time. These include overall rmsd from the 1VII structure, number of helical residues, radius of gyration of the molecule, rmsd of the Phe core.

In figure 2 above we see how various characteristics of the molecule vary over time. As can be seen from the overall rmsd of the molecule from the 1VII structure the protein starts in an extended state. Between frames 1 and 10 there is a rapid collapse of the molecule to a globular state. Between frames 10 and 19 we see that there is a quiet stage in the folding process where the molecule is in some intermediate conformation that is still very different from the folded conformation. This quiet time has also been observed by Kollman et al. in their simulation of villin headpiece folding [2]. Finally at frame 19 we see that the molecule collapses yet again and comes closer to the native state in which it remains. The radius of gyration of the molecule(blue), which is a measure of its size, follows exactly the same pattern and we see that in the end it reaches around 10 Angstroms. The 1VII structure has a radius of gyration of 9.6 Angstroms. Videos of some similar folding trajectories created by other members of the Pande group have shown that one possibility for why the intermediate stage is seen for this long quiet stage is that the C-terminal Phenylalanine, being hydrophobic, folds into the core along with the other 3 Phe groups. This is fairly stable and it takes a significant perturbation to push this Phe group out so that the rest of the core and the remaining helices can form. This could not be identified in the analysis done by me so far.

Turning our attention to the formation of the helices we find that the first helix is formed in the second frame itself. Thereafter there is a constant fluctuation in the number of residues that are helical. By the 20^{th} frame when the rmsd starts to come close to the native structure an increase in the average number of helical residues can be seen. In this particular trajectory it is because of the formation of a second helix between residues 15-18. At frame 32 we find that both helices are lost at the same time. This has also been seen in other trajectories where the middle helix and the C-terminal helix were present. There are trajectories where this is not the case too. Further analysis is required before the significance, if any, of this observation can be understood.

As can be seen from the size of the Phe core (pink), the Phe groups in the core come together very early in the process, by frame 9. The radius of gyration of the Phe core in the native state is 5.6 Angstroms. This occurs at the same time as the formation of the second helix. In contrast the helices are formed before the radius of gyration reaches it native value and this indicates that the formation of the secondary structures preceded the complete collapse of the molecule. This can also be seen from the plot of the solvent accessible surface area (SASA) versus time in figure 3 below. The SASA reaches its native values long after the helices are formed. Also plotted is the total potential energy against time. We see from these figures that it is the rapid hydrophobic collapse that drives the folding events that result in the formation of the intermediates.

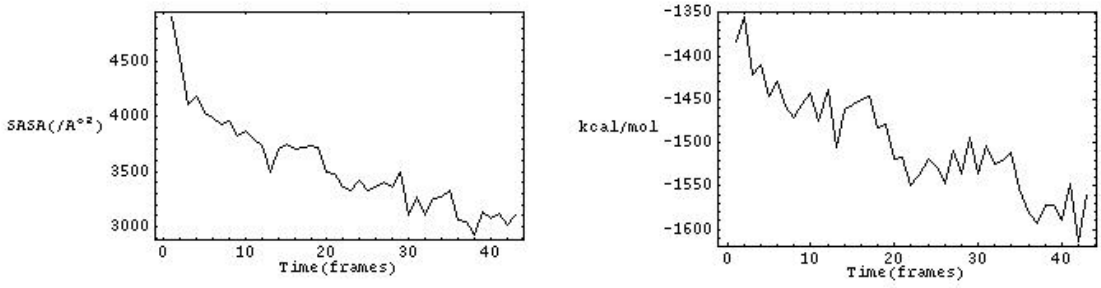

Figure 3. Solvent accessible surface area and total potential energy against time.

Note however that there is wide variation in the order of these events among members of the folded set and these results merely serve to illustrate the kinds of events seen in the various folding trajectories. In the next section we examine ensemble wide trends to make these findings more concrete.

4. *Ensemble* ***level analysis***

On an ensemble level, the three states - extended, unfolded and folded - are seen in the composite data set. There is a great variation in the intermediates seen in these trajectories. Other than small helices mainly hydrogen bonded turns and bends were found in these intermediates. Folded structures display a good number of alpha helical residues and a hydrophobic core. The non-helical regions were again bends and turns. A smattering of isolated beta bridges was also seen. Three of the folded structures also had extended strands, though they were restricted to 4 residues. These strands and bridges were not seen as precursors to the helices. Visual analysis suggests that these were merely random events. Alpha helices were frequently converted to PI helices and vice-versa. When studying the order of formation of the helices in the folded trajectories it was found that in a majority of the trajectories (14 of 21) the N terminal helix was formed first. In these cases the alpha helix in the middle usually was second to form, although not always. The alpha helix in the middle was the first to form in 6 of the 21 folded trajectories. In only one case did the N terminal helix form first, and that too along with the C terminal helix. It must be realized that the C terminal is far from the residues that form the core, although the helix in the middle is close to the hydrophobic core. Thus the C terminal region may be more free to form the helix as long as it is not in contact with the core, whereas the residues in the other helix may have to wait for all the forces that cause the hydrophobic collapse to stabilize before these helices can be formed.

Figure 4 plots the statistical energy of the composite set as a function of the radius of gyration and the RMSD from the 1VII structure. The statistical energy is computed as the log of the fraction of the total population. Since the composite set is not in equilibrium we cannot rely on the energy boundaries quantitatively but they do serve to illustrate the energy boundaries of various events in the folding process. The figure clearly shows a single descent in energy towards the folded state, which has low radius of gyration and rmsd.

Figure 5 has a similar graph but this time for the folded set only. The extended, unfolded and folded regions can be identified in this plot too at the top, middle and bottom regions of the contour respectively.

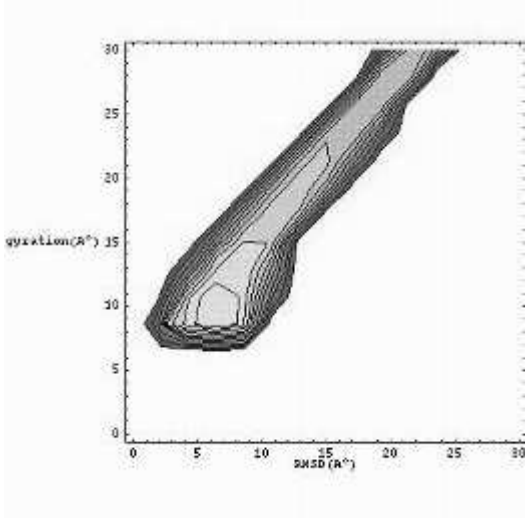

Figure 4. Statistical Energy of the composite data set as a function of radius of gyration and RMSD from the 1VII structure. Values range from 1.09 (red) to 10.9 (yellow).

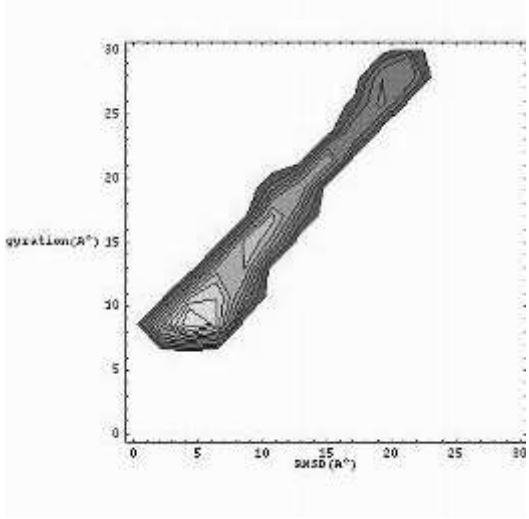

Figure 5. Statistical Energy of the folded data set as a function of radius of gyration and RMSD from the 1VII structure. Values range from 0.69 (red) to 5.56 (yellow).

In figure 6 we plot the statistical energy as a function of the number of helical residues and the radius of gyration of the hydrophobic core. This figure shows that a high number of helical residues are only observed when the residues forming the hydrophobic core have collapsed. Thus hydrophobic collapse seems to drive the formation of secondary structure. However note the region where there are less than 5-6 helical residues. In this region we see that helices exist regardless of formation of the core. This suggests that some helices can be formed even before the hydrophobic collapse. This is in contrast to results obtained in an earlier simulation of the Folding@Home project where a beta-hairpin folding was simulated. This can be attributed again to the fact that the C terminal residues seem independent of the core and the helix formation in that region does not depend on any other interactions as long as the core doesn't bond with some residues in this region and preclude formation of the helix. This theory needs to be verified by much more rigorous analyses, of course. The plot does suggest however that the order of hydrophobic collapse and secondary structure formation may differ form protein to protein and may differ on the layout of the sequence (e.g. location of the core, location of residues that will form the secondary structures)

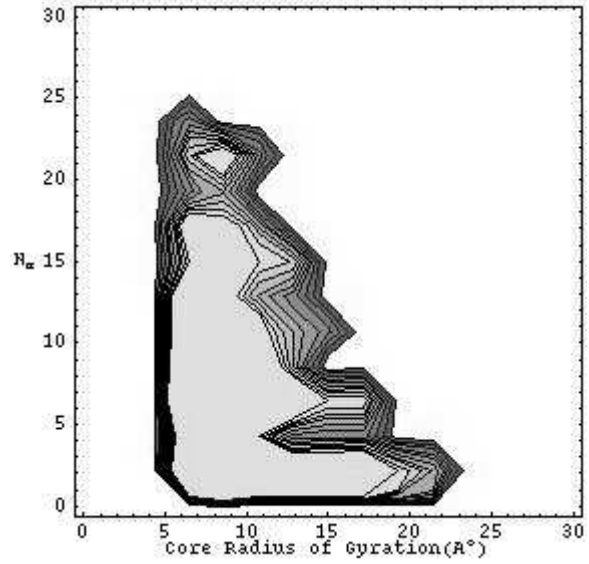

Figure 5. Statistical Energy of the folded data set as a function of the number of helical residues and radius of gyration of the core from the 1VII structure. Values range from 0.69 (red) to 4.41 (yellow).

To study this even further the times at which the rmsd of the Phe core and the number of helices (not helical residues) meet certain thresholds is given in the figures below which analyze the folded set. If the hydrophobic collapse were to occur before the formation of the desired number of the secondary structure then most of the points on the plot would lie below the 45 degree line. Most points would likewise lie above the 45 degree line if secondary structure was formed before the tertiary structure (hydrophobic core). The definition of secondary structure formation is varied from formation of 1 helix to a stringent condition – formation of all 3 helices. The definition of the formation of tertiary structure is also varied from the mild – RMSD of core from the native structure < 5.0 to stringent – rmsd < 3.0. As can be seen from these plots (figure 7a through c) the first helix is formed before the hydrophobic collapse, even if we take a very lenient view of the collapse (rmsd of Phe core < 5.0). There is no clear winner (figure 7b) among whether 2 helices are formed first or whether rmsd falls below 5.0. If both secondary and tertiary are given medium definitions (2 helices should be formed and rmsd of core < 4.0) then secondary structure seems to be formed first once again. This is significant as 2 of the helices of this protein are small and an rmsd of 4.0 is not very strict. Finally if we define formation of secondary structure as the formation of all the helices we see that when the rmsd threshold is 4 Angstroms tertiary structure seems to appear after secondary structure, and when the rmsd threshold is 3.0 there is no clear winner.
Based on these observations it seems that formation of the secondary structure often precedes formation of tertiary structure. The fact that there are points on either side of the diagonal in each graph reminds us that this is not a rule but a trend. This is contrary to what has been observed in other molecules so far. Again this can be attributed to the fact that the helix at the C terminal can form fairly independently. Unlike the beta hairpin simulation[4] in which the formation of the hairpin is dependent on formation of the core, this is not the case for the villin molecule. Another reason that the core formation is delayed maybe the interaction of the C terminal Phe group that becomes a part of the core and has to be ejected before the rest of the hydrophobic collapse can occur. This needs to be verified by further analysis of the data.

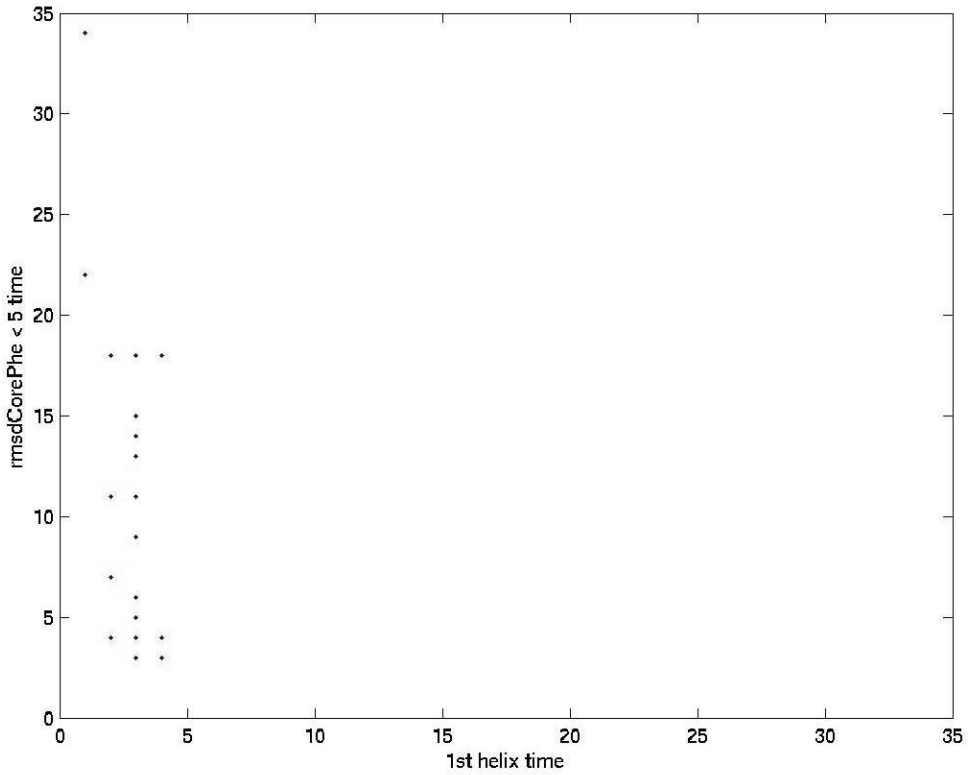

Figure 7a. Time at which the RMSD of the Phe core with the 1VII structure falls below 5.0 is plotted on the y axis while time at which the first helix is formed is plotted on the x axis.

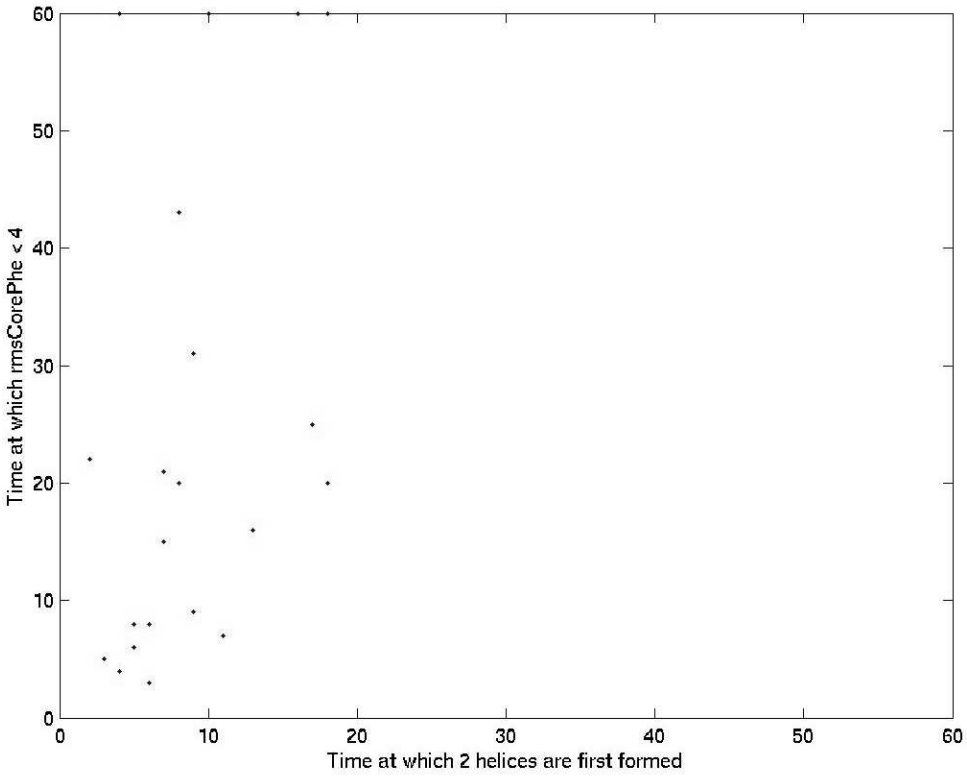

Figure 7b. Time at which the RMSD of the Phe core with the 1VII structure falls below 4.0 is plotted on the y axis while time at which the first two helices are formed is plotted on the x axis.

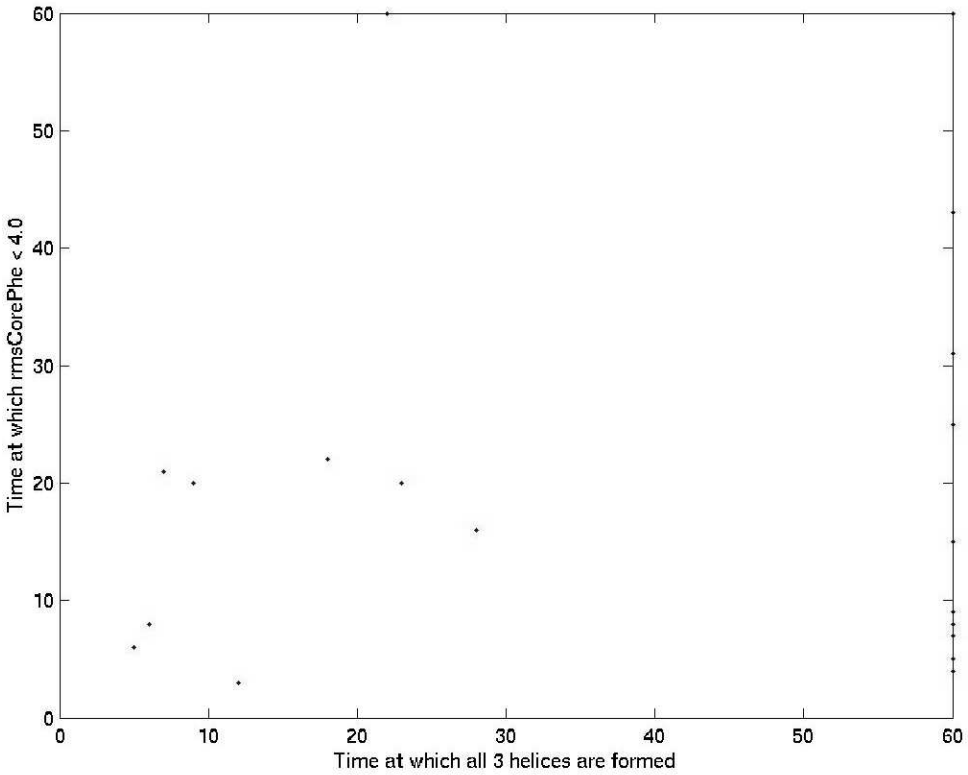

Figure 7c. Time at which the RMSD of the Phe core with the 1VII structure falls below 4.0 is plotted on the y axis while time at which the first three helices are formed is plotted on the x axis.

Note: The figures could not be compressed without loss of detail.

5. Conclusion

Based on the analysis above it can be concluded that (i) There are no well defined intermediates on the folding pathway (ii) The folding process is driven by the hydrophobic collapse at least in the initial and very final stages (iii) Formation of secondary structure usually precedes hydrophobic collapse in the villin headpiece (iv) Helices keep forming and breaking and are not as stable as the core tertiary structure. (v) Secondary structure elements do not appear all at once. In the case of the villin headpiece a preference is shown for forming secondary structure from the C-terminal to the N-terminal.

The study brings to light the fantastic detail in which the folding process can be described using the ensemble dynamics approach can be used to study protein. There are still many questions that are unanswered and a systematic mining of all the data can yield even more insights about the folding process.

6. Acknowledgements

I am very grateful to Dr Vijay Pande and Bojan Zagrovic for their guidance in this project, and to Dr. Douglas Brutlag for encouraging me to take it up.

7.References

[1] August 28, 1999, National Public Radio's Science Friday program (http://search.npr.org/cf/cmn/cmnpd01fm.cfm?PrgDate=08/28/1998&PrgID=5" TARGET ="NEW")

[2]Yong Duan & Peter A. Kollman, "Pathways to a Protein Folding Intermediate Observed in a 1-Microsecond Simulation in Aqueous Solution," *Science* **282**, 740-44 (1998).

[3]Shirts, M. R. & Pande, V. S. (2001), “Mathematical Analysis of Coupled Parallel Simulations” Phy. Rev. ser. B, 57, 13985-13988

[4] B. Zagrovic, E. Sorin & V. Pande (2001) “Beta-Hairpin Folding Simulations in Atomistic Detail Using an Implicit Solvent Model”, JMB 313 151-169