The Early Events of Protein Folding: Simulations of Polyalanine Folding into an Alpha-Helix

Thesis by

Ruth Ann Bertsch

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

1998
(Submitted August 5, 1997)

Acknowledgments

I thank the Department of Energy (Biocatalysis by Design DE-FG36-93CH10581) for providing support for this research and the National Institutes of Health for a National Research Service Award for the author. Additional support for the Materials Simulation Center (MSC) facilities came from the National Science Foundation Grand Challenge Application Grant (ASC 92-17368 and CHE 95-12279), Durip (DAAG55-97-1-0140), Chevron Petroleum Technology Co., Asahi Chemical, Aramco, Owens-Corning, Exxon, Asahi Glass, Nippon Steel, Hercules, Avery Dennison, BP Chemical, and the Beckman Institute. Some calculations were carried out at the Illinois National Center for Supercomputing Applications (NCSA), funded by the National Science Foundation (NSF).

I would like to thank my primary advisor, Sunney I. Chan, for supporting me financially and academically throughout my six years at Caltech. Since the beginning, he has trusted my scientific abilities and allowed me to pursue my own path. For example, he let me spend over six months in the library, something virtually unprecedented at a graduate research institution designed to produce papers. However, this time made this work possible.

I would like to thank William A. Goddard, III, for providing me the tools, people, and resources with which I could learn what I had to learn to do the research I wanted to do. Also, I appreciate the fabulous scientific discussions we have had. Both Bill and Nagarajan Vaidehi have made science fun for me.

Nagarajan Vaidehi, Darryl L. Willick, and Timothy M. McPhillips have been invaluable. They have been willing to debug my technical problems, and they did not stop until the problems were solved. Thanks to the help of Darryl and Tim, I now consider programming an exciting tool that works for me and a toy that delights me. Vaidehi and Tim have also provided excellent scientific advice.

I am in debt to M. Susan Melnik for teaching me and helping me extensively with $\LaTeX2_\epsilon$ and to Gary Holt for $\LaTeX2_\epsilon$ and postscript debugging. Neither stopped helping me until my problems were solved.

Last, I would like to thank the Chan, Goddard, and Rees groups for scientific support and friendship. In particular, Brian Schultz, Ron S. Rock, Gary Mines, and Silvia Cavagnero have been invaluable.

Abstract

The kinetics of α-helix formation in polyalanine and polyglycine eicosamers (20-mers) were examined using the Newton-Euler Inverse Mass Operator (NEIMO) method (Jain *et al.* (1993) *J. Comp. Phys. 106*: 258–268), a new type of torsional coordinate molecular dynamics (MD). One hundred fifty-five (155) different MD experiments were carried out on extended $(\text{Ala})_{20}$ under identical conditions for 0.5 ns each, and 129 of the simulations (83%) formed a persistent α-helix. In contrast, the extended state of $(\text{Gly})_{20}$ only formed a right-handed α-helix in two of the 20 MD experiments (10%), and these helices were not as long or as persistent as those of polyalanine. This is consistent with the helix propensities of the natural amino acids.

The analysis of all 155 simulations show helix formation to be a competition between the rates of

(a) forming local hydrogen bonds (i.e., hydrogen bonds between any residue i and its $i+2$, $i+3$, $i+4$, or $i+5$th neighbor) and

(b) forming nonlocal hydrogen bonds (HBs) between residues widely separated in sequence.

Local HBs grow rapidly into an α-helix; but, nonlocal HBs usually retard helix formation by "trapping" the polymer in irregular, "balled-up" structures. Most trajectories formed some nonlocal HBs, sometimes as many as eight. But, for $(\text{Ala})_{20}$, most of these eventually rearranged to form local HBs that lead to α-helices. A simple kinetic model describes the rate of converting nonlocal HBs into α-helices.

Torsional coordinate MD speeds folding by eliminating bond and angle degrees of freedom and reducing dynamical friction. Thus, the observed times of 80 to 500 ps are likely to be lower bounds on real rates. However, we believe the sequential

steps observed here mirror those of real systems. When compensating for the effect of dynamic friction, the half live for α-helix formation of $(\text{Ala})_{20}$ is estimated to be 209 ps.

Chapters 2 and 3 describe two trajectories of $(\text{Ala})_{20}$ folding into an α-helix. Different types of analyses are used to understand the process of formation and simplify the megabytes of information available in each trajectory. Chapter 2 illustrates a trajectory that forms an α-helix fast, whereas Chapter 3 describes a trajectory where helix formation was retarded by nonlocal HBs.

These simulations attempt to elucidate the early events of protein folding. As elaborated in Chapter 1, the early events may be vital to controlling folding yield and the folding/aggregation partition.

Contents

List of Figures

List of Tables

List of Abbreviations

$(\text{Ala})_{20}$	an eicosamer (20-mer) of polyalanine. See Appendix D.
CD	circular dichroism
$(\text{Gly})_{20}$	an eicosamer (20-mer) of polyglycine. See Appendix D.
HB	hydrogen bond. The abbreviation is used as a noun, an adjective, and a verb.
$i, i+n$ HB	a hydrogen bond between the carbonyl oxygen of residue i and the amide proton of residue $i+n$. This is different from a $i, i-n$ HB which links the carbonyl oxygen of residue $i, i+n$ to the amide proton of residue i.
MD	molecular dynamics (For a definition, see Appendix A.)
NEIMO	Newton Euler Inverse Mass Operator method. See Chapter 6, Section 6.3 and Jain *et al.* (1993).
nonlocal HB	Essentially all nonhelical, non-$(i, i+2)$ HBs. Rigorously, all HBs between the carbonyl oxygen of residue i and the amide proton of any residue in the following ranges: $i+6$ to $i+20$ or $i-2$ to $i-20$.
PDB	Protein Data Bank (Bernstein *et al.*, 1977)
ϕ, ψ	Two backbone dihedral, or torsional, angles of a protein. See Appendix D.

Chapter 1 The Protein Folding Problem

Abstract

These simulations attempt to elucidate the early events of protein folding. Understanding the early events of protein could help structure prediction. More important, the early events may be vital to controlling folding yield and the folding/aggregation partition. Proteins do not always fold in high yield. However, for any protein should be possible to mutate one or two residues, improve the folding/aggregation partition, yet not alter the stability or function of the native state significantly. A major cause of protein misfolding is aggregation, and its features are described.

The view that proteins fold via pathways arose to resolve the Levinthal Paradox. New explanations include a "funnel" model which allows for an almost infinite array of parallel folding pathways and a model by Debe & Goddard (1997) whose recalculations of the size of conformational space indicate that a protein may be able to randomly search all low and medium energy conformations and still fold in less than a second.

The order of events in protein folding are described by the diffusion-collision, framework, hydrophobic collapse model, and a new paradigm.

Although these early events of protein folding are vital to current research, they are very difficult to monitor experimentally because folding is too fast for most existing techniques. In contrast, molecular dynamics, until recently, has not been able to extend its simulation times long enough to simulate the entire folding process. Fortunately, this thesis illustrates progress in that area.

1.1 Why is understanding protein folding important?

Understanding how proteins fold will be one of the most important accomplishments in twenty-first century science for two reasons. First, understanding the process by which the randomly oriented chain becomes the more ordered, functional molecule might help scientists predict the final folded form given the amino acid sequence (Karplus & Weaver, 1976). Structure prediction is currently at a rudimentary level. The mechanism of folding may become the key to predicting native structure.

The more immediate benefit to studying folding is that kinetics often determines the final structure of the protein, i.e., whether the molecules fold to their native state or whether they aggregate into nonfunctional conglomerates. In other words, folding does not always work in either *in vitro* or *in vivo* expression systems. The failure of particular proteins to fold correctly is often an insurmountable obstacle in biological research. Isolating proteins which fold inefficiently can be a prohibitively costly industrial process (Georgiou & De Bernardez-Clark, 1991). Understanding the mechanism of folding could help scientists improve its efficiency.

1.1.1 Structure prediction

Understanding the process of how a protein arrives at its final form could eventually help structure prediction. Current attempts to predict protein structure, given the amino acid sequence, are not reliable for the majority of proteins. To further evaluate progress in this field, in 1994 and 1996 about 70 research groups were given previously-unknown protein sequences and asked to predict their three-dimensional structures. The organizers of this informal contest withheld the experimentally solved structures from the groups. The results of the contest were discussed at the first and second meetings on the Critical Assessment of Techniques for Protein Structure Prediction (CASP and CASP2). Authors have summarized the successes and limitations of structure prediction in three categories: comparative modeling using homologous proteins, threading onto existing structures, and *ab initio* (force field) predictions.[1]

First, comparative modeling, using the structures of homologous proteins, is highly successful at predicting the tertiary structure of a protein if it has over 70% sequence homology with a family of structurally well-characterized proteins. In fact, several public domain and commercial software packages are available for this. The results are used to aid X-ray crystallographers in molecular replacement and to help experimentalists visualize their protein and then design new experiments. However, for proteins with less than 30% sequence homology to structurally characterized proteins, the results are unreliable and do not improve with energy minimization techniques. In addition, if the initial sequence alignment is incorrect, the predicted structure is guaranteed to be wrong (Mosimann *et al.*, 1995).

Either threading or *ab initio* procedures are used when the unknown protein has no detectable sequence homology to proteins with known structures. Threading methods assume the unknown sequence folds into some topology already present in the Protein Data Bank (PDB). CASP showed that although current threading methods are capable of selecting the correct fold from a data base of structures, the methods are not yet reliable.

In contrast to threading techniques, *ab initio* methods do not presume the structure of the unknown protein is similar to anything in the PDB. This makes *ab initio* methods the only methods that could potentially identify an entirely new topology. In practice, however, they do worse than any other method at predicting tertiary structure. *Ab initio* methods generally either employ a force field to simulate a low energy structure or they attempt to interpret and use information from multiply aligned sequences of homologous proteins of unknown structure (Dunbrack *et al.*, 1997).

The algorithm that was heralded in national newspapers in 1995, LINUS, is an example of an *ab initio* program (Srinivasan & Rose, 1995). Although the authors

[1] See, for example, the special issue of *PROTEINS: Structure, Function, and Genetics* that was entirely devoted to the 1994 contest (23(3), 1995).

report, "LINUS effectively determines the secondary and supersecondary structure of five [small] proteins," they admit, "extensive atomic detail is beyond" their scope (Srinivasan & Rose, 1995). Their submissions to CASP and CASP2 did not stand out significantly.

The CASP and CASP2 judges underscored that a major problem that keeps the field from advancing more quickly is the lack of an accurate, reproducible method for evaluating the quality of a particular prediction. Root-mean-square deviations (RMSD) of various atoms or residues are helpful, but they are dominated by a few large errors in dihedral angles, which are then propagated by the internal coordinate system (Srinivasan & Rose, 1995). One bad helix can make the RMSD of an excellent prediction look bad (Su, 1997). Srinivasan and Rose prefer difference dihedral angle plots that compare, per dihedral angle of each residue, the difference between the experimental and predicted structures (1995). Not only are more analytic parameters necessary, but a better vocabulary would be enormously helpful for describing fits of topologies and structural similarity (Madej *et al.*, 1995).

In contrast to three-dimensional prediction, it is easier to evaluate the prediction of the secondary structure of proteins. Estimating whether a portion of sequence will fold into an α-helix, β-pleated sheet, or a turn or a loop is not more than 72–75% accurate (Frishman & Argos, 1997). Algorithms that achieve this accuracy use both information from the local interactions among neighboring residues and from the known structures of proteins with homologous sequences. A popular example of a program with 72% accuracy is the publically available software PHD (Rost & Sander, 1995). It uses homology information in a neural network to predict both secondary structure and solvent accessibility (Rost & Sander, 1995).

Interestingly, there appears to be an upper limit of accuracy for secondary structure prediction algorithms that use only information from local interactions and no information from nonlocal interactions or homologous proteins. If neither homology modeling nor nonlocal techniques supplements secondary structure prediction algorithms, they would only achieve 64% accuracy (Jaenicke, 1991; Branden & Tooze, 1991). (In contrast, random guessing would be 33% accurate (Branden & Tooze, 1991.) For example, Chou and Fasman statistical predictions average 50% accuracy, while the stereochemical methods of Lim average 56% (Branden & Tooze, 1991). Neural nets using only local information predict secondary structure with up to 64% accuracy (Qian & Sejnowski, 1988). Because these methods fail to predict secondary structure 100% correctly, "no method based solely on local information is likely to produce significantly better results for non-homologous proteins" (Qian & Sejnowski, 1988). The failure of algorithms based only on local information implies that local interactions among residues are significant, but they do not entirely determine the final state of the protein.

In fact, folding to the native state must depend on nonlocal forces because the collapse is a cooperative process in which events at one end of the polypeptide depend on and influence events at the other end. It has already been shown that adding nonlocal information helps improve secondary structure prediction. For example, structural information from homologous proteins offer nonlocal information. But, without employing homology modeling, Frishman and Argos (1996) used secondary structural information to predict the position of nonlocal, hydrogen bonds between residues on neighboring β-strands. These predictions about nonlocal interactions improved the secondary structure predictions to 68% accuracy; the authors claim the accuracy would improve 5–7% if homology alignments were included.

Structure prediction will improve with a better understanding of nonlocal forces. Studying the kinetic process of protein folding will probably elucidate these nonlocal forces.

1.1.2 Efficient folding

Protein folding will have tremendous applications to structure prediction once we thoroughly comprehend the process. However, studying it could immediately yield solutions to problems which are critical now. Folding an overexpressed protein is often an insurmountable obstacle in industrial and laboratory protein syntheses (Georgiou & De Bernardez-Clark, 1991). Theoretically, a molecule will eventually adopt the conformation of its thermodynamic minimum, given the proper conditions.[2] However, the free energy of stabilization of the native state of a protein is very small (Jaenicke, 1991), and many local, thermodynamic minima exist. We might expect that the protein could become trapped in one of these minima, preventing it from folding on a reasonable time scale.[3]

Needless to say, folding conditions are not always practical or attainable industrially. Nor do *E. coli* or yeast cells always provide optimal conditions for the assembly of foreign or over-expressed proteins. In fact, nature does not even fold each naturally-expressed polymer strand perfectly. For example, optimal growth of the tailspike protein of the *Salmonella* phage P22 has a yield of less than 50% *in vivo.* "During secretion, misfolded, misassembled, and unassembled polypeptides are retained in the

[2] For example, even rubisco, an 800 kDa, 12–14 subunit protein famous for requiring chaperones to fold, can fold correctly, unassisted, if at 70 nM and 15°C conditions (Viitanen *et al.*, 1990; Goloubinoff *et al.*, 1991). Besides concentration and temperature, other variables that can alter the folding yield for a particular protein include ionic strength, pH, and the presence of sugars, surfactants (Wetlaufer & Xie, 1995) counterions, cofactors, and chaperones. In addition, slowly dropping unfolded protein into a refolding sample has enhanced folding yield (Fischer *et al.*, 1992).

[3] Two classes of possible examples of proteins getting trapped in semi-stable intermediate states are particular proteases and influenza hemagglutinin. Refolding α-lytic protease gets trapped in a semistable, intermediate state until the protease can interact with its pro sequence (Baker *et al.*, 1992a). In the presence of the pro region, the protease rapidly refolds to its native state but remains associated with the pro region until another protease degrades the pro region and activates α-lytic protease. α-lytic protease is unusual because it requires an extra polypeptide almost as long as itself to fold to its stable, active, native state. (The pro region has 166 amino acids, and the protease 198 (Baker *et al.*, 1992b.)) This feature of α-lytic protease may have evolved to protect it from other, active proteases when it is in a vulnerable (i.e., partially folded) state (Agard, 1997).

Other proteases also get trapped in metastable states. The subtilisins E and BPN$'$ also require their pro-region to fold from meta-stable states to their native states (Zhu *et al.*, 1989; Eder *et al.*, 1993). Carboxypeptidase Y has a similar story (Winther & Sørensen, 1991). Active plasminogen activator inhibitor-1 (PAI-1) converts with a half life of 1 hour to a more stable, inactive form. The metastable, active state is regenerated after denaturing and refolding the stable state (Banzon & Kelly, 1992).

Influenza hemagglutinin folds at neutral pH to one configuration, but converts to another at low pH. The low pH form is more stable than the high pH form. The low pH structure is more resistant to denaturation, even at high pH, and the conversion is irreversible (Baker & Agard, 1994; White, 1993). In addition, when hemagglutinin is expressed in *E. coli* without the receptor-binding chain, it folds into the low-pH structure (Chen *et al.*, 1995). Thus, hemagglutinin can fold to a quasi-stable form in the presence of the receptor binding chain at neutral pH.

ER [endoplasmic reticulum] and specifically degraded" (Jaenicke, 1991). Nonetheless, many biochemists still believe that any protein possessing native covalent bonds will always fold to its native form, given the proper set of conditions (Lorimer, personal communication).[4]

1.1.3 Controlling folding yield by selective mutation: The amino acid sequence influences the kinetic outcome

Some amino acid residues are vital not for stabilizing the final state but for bringing the protein to the final state. The classic example of how individual residues in a sequence can dictate folding yields is the temperature sensitive folding (*tsf*) and suppressor (*su*) mutations of the tailspike protein of the *Salmonella* phage P22 (Yu & King, 1988). The protein has mutants which cannot fold at elevated temperatures where the wild type can. However, these mutants, once folded at lower temperatures, are biologically active and as stable as the native, wild type protein. Counter mutations can suppress the effects of these *tsf* mutants. Applying these suppressor mutations to the wild type protein increases its folding efficiency above normal (Mitraki *et al.*, 1991). Folding mutants have been isolated in other proteins too. D-Lactate dehydrogenase has *tsf* mutants which are stable at elevated temperatures and are biologically active (Truong *et al.*, 1991). Interleukin-1β, a monomeric, single domain protein which has one mutant (K97V) which is at least as stable than the wild type. This mutant folds less efficiently than the wild type both *in vivo* and *in vitro* (Wetzel & Chrunyk, 1993). Clearly, some amino acids are important not for structure or function but for determining the folding pathway.

[4] Of course, the example in footnote 3 of the proteases and hemagglutinin is an exception. Both α-lytic protease and influenza hemagglutinin need other atoms (a pro region and extra protons or the absence of the receptor-binding chain, respectively) to fold to their most stable states; once there, these proteins discard the extra atoms. Thus, these other atoms can be thought of as folding "catalysts," if the term "catalyst" is interpreted loosely.

1.2 Mechanisms of protein folding

1.2.1 Resolutions of the Levinthal Paradox

The Levinthal Paradox

For decades it was believed that proteins must fold via pathways rather than by randomly sampling conformational space until arriving at the global minimum (Levinthal, 1968). This assumption is based on calculations of Levinthal, Bloomfield, and Wetlaufer estimated in the late 1960s (Wetlaufer, 1973; Levinthal, 1969). In their model, a 50-residue protein, sampling ten conformations per residue at a rate of 100 residue conformations per 0.1 picosecond, would take 3×10^{27} years to fold (Wetlaufer, 1973), longer than the age of the universe, which is only in the billions (i.e., probably $\sim 10 \times 10^{9}$) of years. Thus, there are too many conformational states for a protein to search and find the global minimum on a biological time scale. This is the "Levinthal Paradox" (Dill & Chan, 1997).

There are several ways to resolve the Levinthal Paradox: First, one can postulate that proteins fold on an energy landscape that directs the protein through the plethora of nonproductive conformations to the native state. Two models describe this landscape either as a tunnel or a funnel. In the folding "pathways" model, most molecules in the ensemble metamorphose through the same succession of structural intermediates that lead to the native state. The funnel model supposes the protein arrives at the native state via any of a virtually infinite number of pathways, similar to the number of paths a drop of rain can take from any summit in a mountain range to the delta of a single river.

The second way, but not necessarily independent way, to resolve the Levinthal Paradox is to assume the original calculations were gross overestimates of the number of conformations available to a protein. Both compaction and secondary structure formation probably limit the size of conformational space. Recent research suggests the number of accessible configurations is small enough to search randomly and still fold correctly on a biological scale. Different calculations offer either an exponential dependence on the number of residues in the protein or a power dependence for the number of accessible configurations (Dill & Chan, 1997; Debe & Goddard, 1997). Recalculating the number of accessible states and biasing the energy landscape are not necessarily incompatible models, depending on who you talk to.

Folding pathways

For decades most scientists have postulated that proteins fold along directed pathways to avoid exhaustive conformational searches. This pathway was thought to be a well-defined trajectory consisting of small, finite numbers of obligatory structures leading to the native state (Dill & Chan, 1997). This model of folding leads to searches for intermediates, the key structures that help proteins avoid hopeless searches through all of conformational space (Dill & Chan, 1997).

The pathway theory has problems. First, there cannot be a single folding pathway for any particular sequence because proteins successfully fold to the native state starting from many different initial conditions. For example, proline isomerization is not necessary for the entire folding population of a protein containing a proline residue. Such a protein must have at least two folding pathways because the fraction of denatured polypeptides with the wrong peptidyl-proline bond will need to isomerize but the fraction with the native bond will not. The work of Radford, Dobson, and coworkers on hen egg white lysozyme indicate there have to be multiple, parallel folding pathways. The group discovered that different subpopulations of folding hen egg white lysozyme fold their α and β-domains at different times and in different sequences (Radford *et al.*, 1992; Miranker *et al.*, 1993). Wright *et al.* argue that multiple folding pathways are vital for evolutionary adaptation because mutations to residues critical to a single pathway should not prevent the protein from folding by other means (Wright *et al.*, 1988).

The folding funnel

The more recent view of the pathway model and resolution of the Levinthal Paradox is to ignore them both. Dill and Chan (1997) assert to us, "Thermodynamic texts are full of examples of systems having nearly Avogadro's number of microscopic degrees of freedom that nevertheless reach stable states on observable time scales." Dill and Chan (1997) believe that the Levinthal paradox is an artifact of describing the energy landscape for folding as a flat energy surface with a single, narrow well for the native state.

Instead, Dill claims recent research supports an energy surface shaped like a funnel (Dill, 1987; Wolynes *et al.*, 1995), as in Figure 1.1. The y axis represents changes in free energy, and, the lateral area of the funnel (represented in the x domain) is proportional to the configurational entropy of the protein. As the protein drops in free energy, it folds and compacts. As it compacts, entropic barriers are erected, and they limit the conformations the molecule can sample (Dill, 1990).

The funnel allows the protein to avoid exhaustively searching each configuration and allows it to fall more-or-less downhill towards the native state. The fall may be smooth, like a kitchen funnel, or bumpy, like the passage down a mountain range full of slopes, valleys, moguls, and passes. Furthermore, the fall will take many different courses because the proteins will start from many different positions on the funnel. This is analogous to water drops draining from a mountain range (Dill & Chan, 1997).

This model, also called the "landscape" model, implies there are many different denatured states and that there is no universal, intermediate structure or transition state through which 100% of the population passes (Dill & Chan, 1997).

The degree of ruggedness of the landscape determines the quantity of mountain valleys in which the protein can become trapped in nonnative structures. In essence, the degree of ruggedness determines how sharply the landscape is biased towards native-like, low-energy states instead of merely towards low energy states.

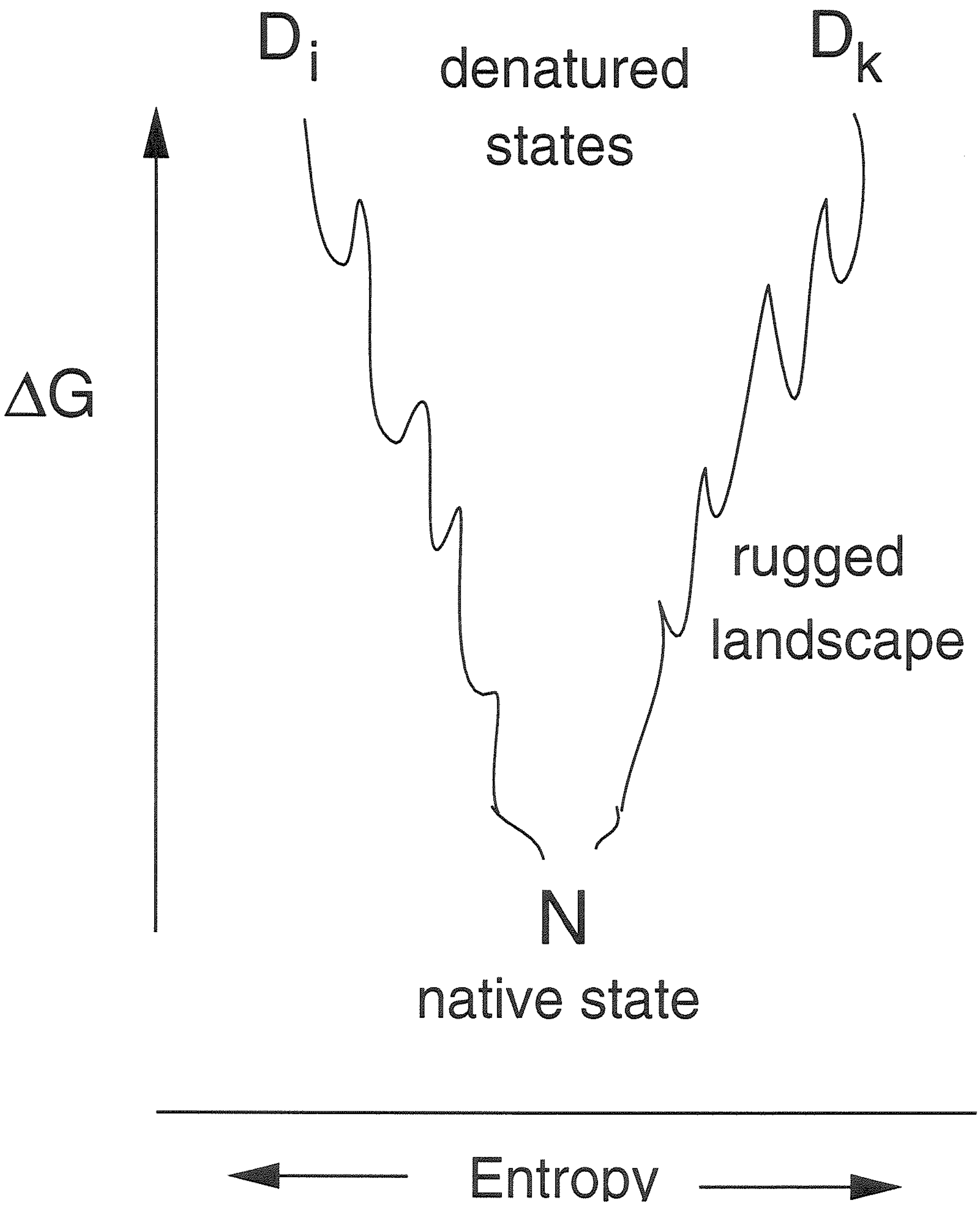


Figure 1.1: Diagram of the folding funnel, a model of the energy landscape for folding of a globular protein. The y axis plots internal free energy of the protein, and the x axis roughly represents the size of conformational space. The lateral area of the funnel is proportional to the configurational entropy of the protein, i.e., the number of structures accessible to it. As the protein drops in free energy, it folds and compacts. As it compacts, entropic barriers are erected that limit the conformations the molecule can sample. For an artistic drawing in perspective, see Figure 4, Dill & Chan, 1997.

Criticisms of the folding funnel

Critics of the folding funnel ask how an energy landscape can be biased enough to shrink the number of conformations from more than 10^{24} conformations (for a 50-mer, assuming 3^n conformations per n residues) to a number feasible for biological folding, e.g., to 10^6 or 10^{12} conformations. Such critics liken this folding funnel to a folding tornado because it must be capable of finding the native conformation in more than a mole of nonnative structures (Goddard, 1997).

Calculations explain that constraining the volume of a protein can significantly reduce the configurational space accessible to a protein. For example, Dill calculates that a 50-residue protein only has an upper limit of 3×10^{11} configurations to sample (Dill, 1985). These calculations are described in the next section, "Volume compaction and an exponential dependence," p. 12. Debe and Goddard (1997) independently performed simulations that further limit configurational space, and this work is discussed in Section "Volume compaction and a power dependence," p. 13. However, neither group finds evidence that the folding states are "channeled" or directed into native-like states.

Volume compaction and an exponential dependence

Dill drastically reduces the size of the configurational space of a folding protein. With this smaller size, he concludes that a protein can fold within experimental time scales to conformations at or near the global free energy minimum via a "biased reversible search" (Dill, 1985).

Dill's recalculations assume that proteins are heteropolymers whose hydrophobic residues want to bury their atoms away from the solvent and whose polar residues belong at the surface. The ratio of polar to nonpolar residues appears to have limits because globular proteins must dissolve in polar water but remain compact, with a hydrophobic core for cooperative stability. The model estimates that "an upper bound on the number of conformations in the globular state is" $(1.7)^n$, where n is the number of amino acid residues in the protein. For a 50-residue protein, the upper bound would be 3×10^{11} configurations. "The number of conformations of relatively low free energy is significantly smaller than this" (Dill, 1985).

Volume compaction and a power dependence

More recent computer simulations further lower the estimate of the number of conformations available to a folding protein and conclude the number has a power, not exponential, dependence on the number of residues (Debe *et al.*, 1997). For a given length of protein, Debe and Goddard generated structures with residues in one of six dihedral configurations, and polymer growth was biased towards low energy structures. The only energy term was a 12-6 van der Waals potential with a well minimum at 5.5 Å. Thus, no atoms overlapped, and yet the structures were biased towards compact configurations. Structures were generated until the simulated ensemble included at least one structure that was "similar" to each of about 20 test structures of the particular polymer length from the PDB. "Similar" meant that the simulated structure had a similar topology to the experimental structure and that the root mean squared deviation of the α-carbons (CRMS) was less than $0.05(n) + 3.00$ Å, e.g., less than 5 Å for a 50-mer.

Although the choice of (ϕ, ψ) angles and the van der Waals potential may have bias the construction of the ensemble to native-like structures, Debe is confident the ensembles represent denatured and partially folded states well (Debe, 1997). Approximately 30% of the structures are approximately as compact as native globular proteins, with the remaining 70% of the ensemble less compact. For example, one of the structures in the ensemble of 65-residue proteins fits an NMR-determined structure of a proteolytic fragment of bacterial rhodopsin. The structure of the protein fragment is two helices at right angles in solution, and it hardly resembles a compact, globular protein.

The small number of conformations this algorithm had to sample before spanning all the topologies in the PDB indicates that a protein has only a few topologically

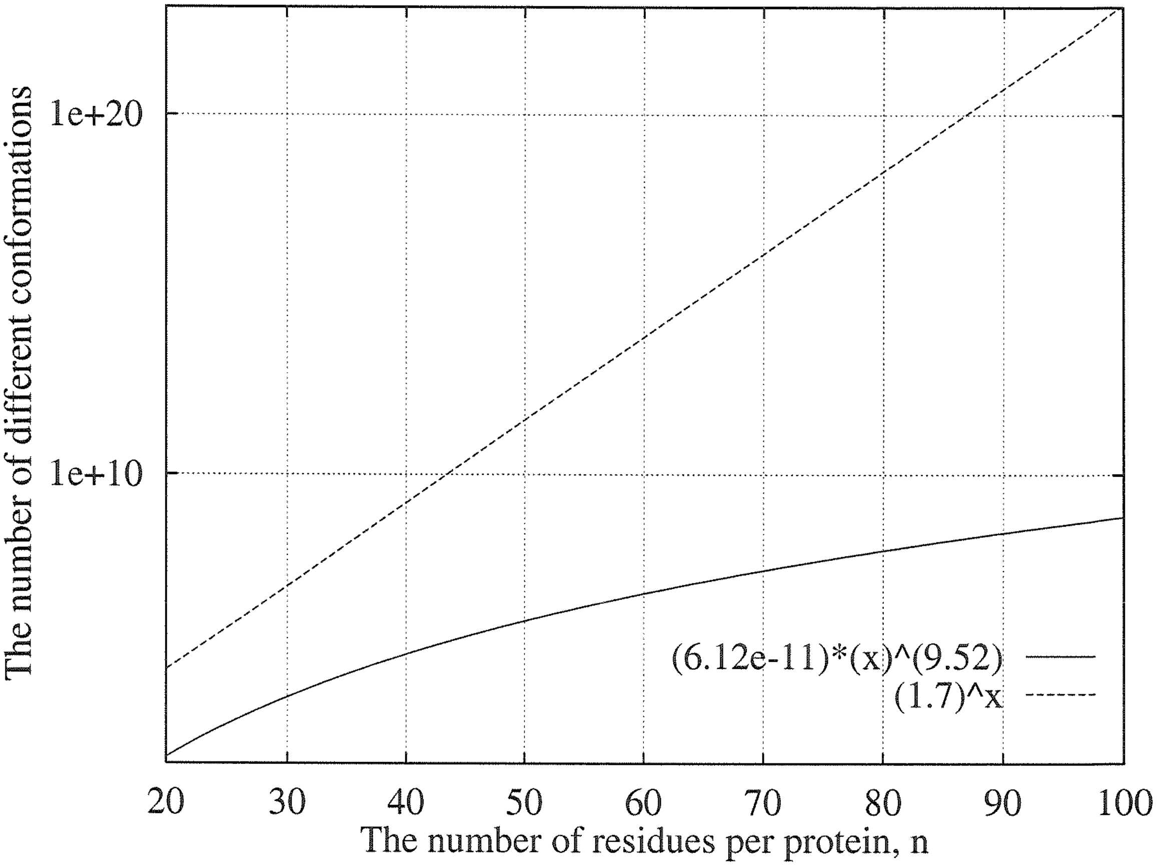


Figure 1.2: Comparison of a power and an exponential dependence of the size of conformational space on the number of residues in a protein. The function $(1.7)^x$ is the relationship Dill proposed in 1985. The function $(6.12 \times 10^{-11}) \times (n)^{9.52}$ is proposed by Debe and Goddard (1997). The y axis is logarithmic with base 10.

distinct structures that are not high in energy. For example, a 50-residue protein only has $\sim 10^6$ (9.1×10^5) such conformations (Debe & Goddard, 1997). For a 100-residue protein, this number is estimated to be 6.7×10^8. Perhaps more important, the number of topologically distinct, low to medium energy conformations increased by a power of the number of residues in the protein—$(6.12 \times 10^{-11})n^{9.52})$ where n is the number of residues—instead of by an exponential of the number of residues, e.g., by 10^n as in the Wetlaufer calculations (1973) or by 1.7^n as in the Dill calculations (1985). The power dependence is a substantial improvement for large proteins, as Figure 1.2 illustrates.

With only a million structures for a 50-mer to sample, it can explore the entire space in 10 μs, assuming a polymer can sample one conformation every 10 ps. This is well within experimental time scales!

Comparison to the funnel model. These simulations explain one feature of the folding funnel model but do not support the "channeling" aspect. By including a van der Waals potential with a minimum depth at 5.5 Å, Debe and Goddard compact the proteins. One may argue that only 30% of the states have native-like compactness. Nonetheless, the more diffuse states are still compact enough to have CRMS's which are often closer to the test PDB structures than the compact, generated configurations are. Also, some of the diffuse structures from the PDB are actually stable, structured polypeptides. For example, the proteolytic fragment of bacteriorhodopsin that forms two helices at right angles was "fully folded" to the best of its ability in the NMR tube. To the extent that this structured fragment is "folded" and "compact," many of the conformers in the generated ensemble are also compact.

The point is that the simulated ensemble is packed enough that compaction can be considered the entropic "force" that erects barriers to the rest of conformational space. This may be the mechanism by which the folding funnel excludes conformational space.

Unlike the funnel, the Debe and Goddard model does not suppose anything "channels" or directs the polymer to the global minimum. Rather, their calculations allow the protein time to walk randomly through low and medium energy configurational space until the protein reaches the global minimum. Since these calculations do not address the relative energies of the topologically distinct states, this model does not tell us either how "rugged" the landscape is or how much it channels the molecules to the final state.

Secondary structure formation reduces conformational space

Besides compaction, another mechanism of reducing the conformational space of a folding protein is to create secondary structure early. The formation of native-like, secondary structure in portions of a protein could prevent those portions from sampling astronomical numbers of other structures. Presumably, the sooner stable secondary structure forms, the sooner the conformational space of the folding protein shrinks. Thus, an understanding of the time scale of α-helix formation will help characterize the folding landscape.

Attempts to map the energy landscape

Obviously, scientists would love to map the folding landscape for a protein. The kinetics of refolding experiments, observations of refolding intermediates, and the relative thermodynamic stabilities of different measurable states have all contributed to our understanding of the folding barriers particular proteins encounter.

Unfolding experiments. *Un*folding experiments can also help describe the energy landscape of a protein and its denatured states. For instance, unfolding experiments can capture intermediates that are not observable by folding experiments.[5] Many intermediates are invisible in refolding experiments because one can only detect fast reactions if they precede slow reactions. An ensemble of molecules loses the simultaneity necessary to monitor a quick reaction after the ensemble undergoes a slow reaction. In other words, one can only see an intermediate that occurs before the rate-limiting step of a reaction. Hence, in Figure 1.3, the folding experiment in (A) will populate state I_1 long enough to observe it since the high barrier to I_2 populates or traps I_1. However, I_2 is undetectable in this experiment. Only an unfolding experiment, as in (B) can elucidate anything similar to I_2, in this case, I_2'. The hope is that I_2 and I_2' are structurally related. To illustrate this hope, the unfolding landscape in (B) was drawn a reversal of the folding landscape in (A). The landscapes have to be at least partially different because the thermodynamic conditions are different. The final condition in unfolding experiments, the unfolded state, must be energetically

[5] For example, Cavagnero (1997) used unfolding kinetics to explain the hyperthermostability of the small iron protein rubredoxin from *Pyrocuccus furiosus*.

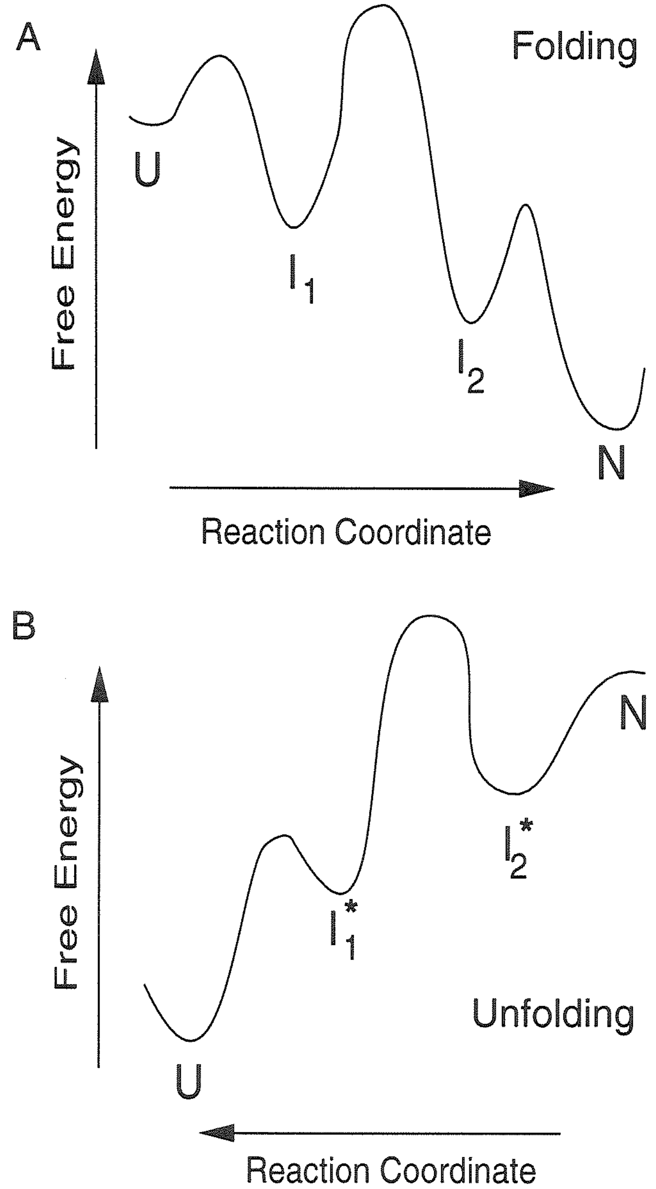


Figure 1.3: Two-dimensional slices of a folding and unfolding pathway. (A) The refolding experiment starts with the unfolded state, U on the left and folds to the native state, N on the right. The first intermediate, I_1, is observable because it gets populated because it is before a high energy barrier. Intermediate I_2 is not observable. (B) The unfolding experiment. N, on the right, is the initial state, and U, on the left, is the final state. The landscape is drawn so the protein unfolds downhill. The intermediate I_2' is experimentally observable but I_1' not. The unfolding energy diagram is drawn as a reversal of the folding diagram to illustrate experimentalists' hope that unfolding kinetics can map the folding landscape. In fact, state I_2 may or may not resemble state I_2'.

downhill from the initial, folded condition, contrary to refolding experiments.

Unfortunately, the refolding and unfolding landscapes are not necessarily related closely enough to guarantee that I_2 and I'_2 are structurally similar. Since proteins fold by multiple, parallel routes, the principle of microscopic reversibility does not apply. In other words, the unfolding path is not necessarily the refolding path, and hence, the two landscapes are not necessarily compatible.

1.2.2 The order of events in folding

Existing models of protein folding

Whatever the shape of the energy landscape, in what sequence do the segments of protein structure build the native state? Several chemists have attempted to describe mechanisms of protein folding. Karplus and Weaver describe a diffusion-collision model, where microdomains of protein adopt native-like secondary structure. These fleeting segments of secondary structure may occasionally collide and stabilize each other. Finally, they may form stable tertiary interactions (Karplus & Weaver, 1976). Englander and Baldwin believe that the contacts between amino acid side chains and the secondary structures they create mold the protein structure. In this "framework model," native secondary structure forms before tertiary structure locks into place (Kim & Baldwin, 1990). In contrast, Dill describes a protein as a heteropolymer which collapses to avoid contact with the water solvent and to maximize contacts between apolar amino acid side chains. The compaction of the polymer creates secondary structure and drives the formation of tertiary structure (Jaenicke, 1991; Chan & Dill, 1990; Chan *et al.*, 1995).

Perhaps these different models contribute the most by asking the following questions: Does secondary structure formation come before hydrophobic collapse? Is the "hydrophobic force" more influential in determining folding than the formation of secondary structure? The framework model predicts that secondary structure forms before protein collapse. The Dill model predicts they either occur simultaneously or secondary structure forms after collapse.

A new paradigm: The stages of protein folding

An alternative model is to visualize the folding of a water-soluble, single-domain protein as a continuum approximated by three states linked by two transformations. The characteristics of the states are based on experimental observations of proteins folding, most of which initiated folding in a stopped-flow by rapidly diluting out denaturant. The experimental probes were mostly circular dichroism (CD), fluorescence of ANS[6] or aromatic residues, and changes in amide proton-solvent exchange rates. Some of the following transitions may not be irreversible, according to the work of Goldberg *et al.* on hen and turkey lysozymes (Goldberg *et al.*, 1991).

A. The denatured state. Proteins start folding from the denatured state, an ensemble of conformations having larger radii of gyration than the unique, native state (Flanagan *et al.*, 1993). Denatured proteins are loose, amorphous blobs which maximize configurational entropy.

B. Condensation. In most single-domain, globular proteins, the first transformation, condensation, starts less than a millisecond after the initiation of folding by dilution from denaturants or extreme pHs or temperatures. Presumably the condensation begins with the help of initiation sites. These are local sections of polypeptide which transiently adopt native-like forms. Because they temporarily exclude much of conformational space, they serve as sites for cooperative growth of more native-like structure (Wright *et al.*, 1988).

During condensation the protein collapses, and the radius of gyration shrinks. Hydrophobic and electrostatic contacts are formed. The polymer acquires secondary structure and perhaps some tertiary contacts (Sugawara *et al.*, 1991; Mann & Matthews, 1993; Elöve *et al.*, 1992; Roder *et al.*, 1988). We

[6] 1-anilino-8-naphthalenesulfonate

expect this process could take between hundreds of microseconds to hundreds of milliseconds in single-domain, globular proteins.

During condensation and the early moments of the next stage, energy barriers to particular conformations are changing, and the protein population partitions into folding and unproductively tangling fractions. At this point, chaperones could prevent unproductively tangled molecules from aggregating irreversibly. (See Section 1.4, Chaperones.)

C. The intermediate or molten globule state. After condensation, the protein either persists as a molten globule or forms a short-lived intermediate with many of the properties of a molten globule. This intermediate moves rapidly into the next transformation, annealing (Ptitsyn & Semisotnov, 1991). Molten globules have radii of gyration slightly larger than the native protein. They have most of the secondary structure of the native state but little of its tertiary structure. Molten globules can be thought of as compact proteins having secondary structure and a fluctuating core (Kuwajima, 1989; Kuwajima *et al.*, 1989). The overall folding pattern probably resembles that of the native protein (Ptitsyn & Semisotnov, 1991).

Molten globules are sometimes observed under equilibrium conditions, either at pH extremes or in mildly denaturing conditions (Ptitsyn *et al.*, 1990). For example, α-lactalbumin forms a molten globule at moderate concentrations of denaturant and at extremely acidic or alkaline equilibrium conditions in the absence of guanidinium hydrochloride (Gdn HCl) (Kuwajima *et al.*, 1989). The intermediate is compact, as shown by a variety of methods, and possesses much secondary structure. There is no tertiary structure, as demonstrated by near-UV CD and a featureless NMR (Roder *et al.*, 1988; Elöve *et al.*, 1992; Chaffotte *et al.*, 1992). This molten globule forms within milliseconds, and its conversion to the native form is the rate-limiting step of folding (Kuwajima, 1989).

D. Annealing. The final stage of protein folding occurs when the polymer readjusts its contacts to optimize its configuration. Tertiary structure is locked into place. Annealing is often the rate-limiting step, and proteins in the process of annealing are often detected as folding intermediates. For example, the last intermediate in the major folding fraction of bovine pancreatic ribonuclease A differs from the native state by the isomerization of proline-93; and this isomerization is part of the rate-limiting step (Schmid, 1986). Similarly, the rate-limiting step in the folding of ubiquitin is the isomerization of proline-37 and/or proline-38 (Briggs & Roder, 1992).

E. The folded state. The final, folded or mature protein is in the native conformation. Folding is one outcome of a set of probabilities.

This paradigm describes how a protein folds correctly. But not all proteins fold correctly. In nature, folding a protein correctly seems to be one outcome out of a set of probabilities under a particular set of conditions. For example, renaturing octopine dehydrogenase, a monomeric homologue of lactate dehydrogenase, yields only 70% activity. The active 70% can be isolated and then denatured again. After refolding, only 70% shows activity (Jaenicke, 1988).[7] This suggests that folding octopine dehydrogenase has a 0.7 probability of success.

This paradigm tries to emphasize that folding is a cooperative, kinetic phenomenon. Protein folding/unfolding bears several hallmarks of cooperative transitions. Unfolding curves are not linear with respect to the denaturing agent; curves of percentage unfolded protein versus denaturant concentration are sigmoidal, like those of allosteric enzymes. Calorimetry demonstrates that "melting" most single-domain proteins occurs within a narrow temperature range, and the transition has a large heat capacity, greater than the heat capacity of each phase on either side of the transition.

[7] The other 30% is not aggregating intermolecularly. Instead, the two domains of octopine dehydrogenase do not always associate correctly (Jaenicke, 1988). However, octopine dehydrogenase is not a counterexample to the assertion that all proteins can find their native conformation given the proper conditions. Presumably, octopine dehydrogenase has a higher folding efficiency at low temperatures where the protein population should assume fewer conformations. There the domains might interact correctly with higher efficiency.

Despite the postulate that the native state of a protein is its global energy minimum, folding *is* under kinetic control. Folding does not always have a 100% probability of success, since the molecules sometimes fold into non-optimal structures, such as an aggregate or the high-pH form of hemagglutinin, which seems to violate the global-energy minimum hypothesis.

1.3 Aggregation: What goes wrong in protein folding? How do proteins misfold?

What do most proteins do besides fold correctly? Proteins that do not fold properly *in vitro* aggregate, in the absence of necessary cofactors or changes in chemical bonding.[8]

This is especially true of single-domain proteins.[9] Aggregation is believed to occur when solvent-exposed hydrophobic residues of incompletely folded proteins encounter nonpolar patches on other, similarly immature molecules and associate to avoid solvent contact. In other words, aggregation occurs for the same reason proteins fold: nonpolar residues avoid polar solvents. *In vitro*, aggregates often grow until they precipitate. *In vivo* bacterial expression systems, aggregates are manifested as inclusion bodies (IBs), or densely packed granules of misfolded protein which can be isolated from cell lysates. Once aggregated, as either *in vivo* inclusion bodies or *in vitro* aggregates, proteins must be dissolved by strong denaturants or detergents and then diluted under less denaturing conditions before the polymers will refold. Aggregation and IB formation are usually considered irreversible kinetic traps in the folding pathway. We assume that the factors which lead to aggregation *in vitro* are similar to those leading to inclusion body formation *in vivo*. It is assumed that the faster a molecule buries its hydrophobic residues and adopts the general form of the fully

[8] See Footnote 3, p. 6.

[9] Interestingly, the remaining 30% of octopine dehydrogenase which failed to renature did not form high-molecular-mass aggregates but small aggregates (Zettlmeissl *et al.*, 1984) or "inactive monomers with native-like secondary structure." Octopine dehydrogenase has two domains, and presumably the poorly folded molecules have incorrect inter-domain interactions (Jaenicke, 1988).

folded molecule, the less likely the molecule is to aggregate and reduce the yield of fully-folded protein.

Because we can influence the relative extents of folding and aggregation, folding must compete with aggregation. Because protein folding is a first order reaction but aggregation is second order, lowering the concentration of protein enhances folding yields *in vitro*. *In vivo*, lowering the cell growth temperature can increase yields and reduce inclusion body formation. Thus, low temperatures must hinder protein synthesis and/or aggregation more than they slow protein folding.

Aggregation is sometimes thought to be in kinetic competition with folding (Kiefhaber *et al.*, 1991) and, therefore, can be modeled mathematically as a competition against folding. See Figure 1.4. The rate law is

$$\frac{-d[U]}{dt} = k_f[U] + k_a[U]^2 \tag{1.1}$$

where $[U]$ is the concentration of unfolded protein, k_f is the first-order rate constant for folding (about 0.4 sec^{-1} for apomyoglobin), and k_a is the effective rate constant for aggregation. Kiefhaber *et al.* (1991) define k_a as

$$k_a = k_{a2} \times N \tag{1.2}$$

where k_{a2} is the intrinsic second-order aggregation rate constant and N is the mean number of monomers per aggregate. They achieve an equation relating yield of folded protein to initial concentration of denatured protein:

$$\text{Yield} = \frac{N_\infty}{U_0} = \frac{k_f}{U_0 k_a} \times \ln\left(1 + \frac{U_0 k_a}{k_f}\right), \tag{1.3}$$

N_∞ is the final concentration of folded protein. This model is consistent with experimental results on lactic dehydrogenase (Kiefhaber *et al.*, 1991).

If aggregation is a kinetic competition, the kinetics of folding should be more important in determining the folding yield than the relative stabilities of the individual

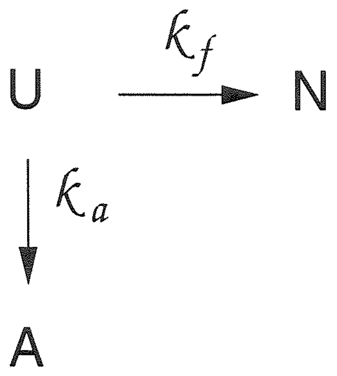


Figure 1.4: Aggregation competes with folding. This scheme illustrates a protein without kinetic intermediates. This illustrates how Kiefhaber *et al.* (1991) modeled protein folding.

states. For example, at least two proteins and their folding mutants have folding yields that are independent of the stabilities of the native states. A mutant of bovine growth hormone with eight mutations folds faster and aggregates less than the wild type hormone, yet it is as stable as the wild type once fully folded (Lehrman *et al.*, 1991). The *tsf* and *su* mutations of the P22 tailspike protein clearly alter its folding by changing the folding rates of the early monomeric intermediates. As stated before, once folded at low temperatures, the mutants appear as active biologically as the wild type and as stable to heat denaturation (Mitraki *et al.*, 1993). At these low temperatures, the *tsf* mutations do not affect folding kinetics or the aggregation/folding partition (Smith & King, 1981; Goldenberg *et al.*, 1983). Similarly, at high temperature, suppressor (*su*) mutations increase the folding yield of the *tsf* mutants by attenuating the folding retardation and decreasing aggregation (Mitraki *et al.*, 1991). But, as mentioned, the *su* mutants *do not* affect the apparent stability of the folded product (Danner & Seckler, 1993).

If aggregation is only dependent on folding kinetics, the free energies of the native state and the folding intermediates should not alter the aggregation/folding partition, as long as the folding kinetics remain constant. This may be difficult to prove, because kinetics and thermodynamics are often intertwined, even in the case of the P22 tailspike. It is true that many of the *tsf* mutants of the P22 tailspike protein, once folded, are as resistant to denaturation against heat or sodium dodecyl sulfate (a detergent) as the wild type (Mitraki *et al.*, 1993). However, these *tsf* folding mutants are actually less stable than wild type because they unfold faster after the initial unfolding phase. Their folding intermediates are less resistant to denaturation and fold more slowly than wild type intermediates. But, the mutants appear as stable as wild type because they unfold in the initial unfolding phase at the wild type rate (Danner & Seckler, 1993).

1.3.1 Mechanisms of aggregation

What properties are important in altering the folding yield of a protein? A statistical analysis of the sequences and properties of 81 different proteins expressed in *E. coli* indicates that proteins that are unlikely to form inclusion bodies have a high charge but have few turn-forming residues (asparagine, proline, glycine, and serine). The hydrophilicity of the total protein is not a good indicator of aggregation behavior (Wilkinson & Harrison, 1991).

Particular structures may be important to induce aggregation because it sometimes occurs preferentially among molecules with significant sequence homology. Although *in vivo* inclusion bodies contain hydrocarbons, glycogen, polyphosphates, and other nonprotein molecules, IBs "are highly enriched in a single protein, despite the high concentration of normal proteins in *E. coli* cytoplasm, many of which are presumably folding" simultaneously (Wetzel, 1992). *In vitro*, aggregation can occur exclusively among molecules with significant sequence homology. For example, folding P22 tailspike protein does not coaggregate with folding P22 coat protein. Thus, the folding intermediates of those P22 proteins must distinguish between folding coat and tailspike protein intermediates (Speed *et al.*, 1996). Similarly, folding tryptophanase does not coaggregate with bovine serum albumin or crude *E. coli* cell extract (London *et al.,* 1974). In contrast, folding hen egg white lysozyme coaggregates with turkey egg white lysozyme (Goldberg *et al.*, 1974), presumably because the two lysozyme intermediates have very similar structures.

Because aggregation can be protein specific, specific conformations must help induce it. For example, islet amyloid polypeptide aggregates form β-pleated sheet fibrils (Chargé *et al.,* 1995). Although aggregates of lactate dehydrogenase predictably fluoresce like the denatured protein, their ellipticities in far-UV CD spectra suggest that the aggregates have almost as much secondary structure as the native dehydrogenase (Zettlmeissl *et al.*, 1979). Phosphoglycerate kinase aggregates also have large components of β-sheet but little α-helix (Mitraki *et al.*, 1987). This does not indicate whether PGK aggregation is induced by β structures or by denatured, formerly α-helical sections.

Attempts have been made on several proteins to define the elements that cause them to aggregate. The region of bovine growth hormone that is important in determining folding yield is the third α-helix in the four helix-bundle hormone. Altering the sequence of the third α-helix can cause the resulting mutant to refold more quickly and aggregate less, yet be as resistant to denaturation as the wild type hormone (Lehrman *et al.*, 1991). Although there is a structure of the P22 tailspike protein, it is not obvious why the 32 independent, *tsf*, single amino acid substitution sites can cause the *tsf* phenotype. Of the 32 sites, 24 are exposed to solvent and about 15 are in surface turns or loops (Steinbacher *et al.*, 1994). However, it is known that the conformation of the N-terminus is not important in determining the folding and chain association pathway because some antibodies against native epitopes block productive folding while the monoclonal antibody against the N-terminus does not block folding (Speed *et al.*, 1997).

1.3.2 Time of aggregation

At what stage of folding do the polymers aggregate? When does a water-soluble protein become "committed to fold," i.e., succeed in folding enough to avoid aggregating? They do not generally aggregate in the native, fully folded form if it is easily soluble (Wetzel, 1992). Since aggregation is a second order reaction, molecules cannot aggregate faster than they can diffuse and collide. But this does not restrict the time scale much. Half-lives for second-order, diffusion-limited reactions are on the order of hundreds of nanoseconds (ns) to hundreds of microseconds (μs) for proteins at concentrations from 1 mM to 100 μM, respectively. P22 clearly aggregates either during its collapse to form the first intermediate or immediately after its formation (Mitraki *et al.*, 1991). This intermediate undergoes a first-order adjustment before it becomes susceptible to aggregation (Danner & Seckler, 1993). Similarly, we know that the monomeric protein carbonic anhydrase aggregates before reaching its second intermediate. If all of folding carbonic anhydrases are at the stage of the second intermediate or beyond, they avoid aggregating under conditions which would otherwise induce aggregation (Cleland & Wang, 1990). Turkey egg white lysozyme appears to remain capable of aggregating with denatured hen egg white lysozyme until the turkey enzyme molecules have reached the native state. This was demonstrated by initiating the folding of turkey lysozyme and then periodically injecting the solution with denatured hen lysozyme. The turkey lysozyme continued to aggregate until it had completely folded. In contrast, when turkey lysozyme was allowed to fold without supplementing it with unfolded hen lysozyme, turkey lysozyme essentially stopped aggregating in a fifth of the folding time (Goldberg *et al.*, 1991).

The facts indicate that the yield-determining steps in protein folding occur early, during the process of collapse or during the early stages of annealing when hydrophobic patches are still exposed. Proteins misfold when they stay incompletely folded, trapped in a local minimum, long enough to aggregate irreversibly. If a protein aggregates after forming specific elements of secondary structure, we can deduce that the remaining native secondary structure had problems folding.

Aggregation affects most protein researchers, yet the kinetic and thermodynamic pathways leading to it have only been studied in a few proteins. The mechanism of aggregation has been best described in the P22 tailspike protein (Speed *et al.*, 1996; Danner & Seckler, 1993; Mitraki & King, 1992; Mitraki *et al.*, 1991; Haase-Pettingell

& King, 1988; Goldenberg *et al.*, 1983; Smith & King, 1981); bovine growth hormone (Lehrman *et al.*, 1991), β-amyloid peptide (Jarrett & Lansbury, 1992), and sickle cell anemia (Zubay, 1988). Preliminary studies on the mechanism of aggregation have been published on interferon-γ (Wetzel, 1992), human interleukin-1β (Wetzel, 1992), prion protein (Kocisko *et al.*, 1995), transthyretin (Saraiva *et al.*, 1984), and islet amyloid polypeptide (Chargé *et al.*, 1995). Understanding the relationships among aggregation, folding kinetics, thermodynamics, and structure is the first step to controlling folding yield by selective, strategic mutation.

1.4 Chaperones and other folding enzymes improve folding yields and rates

The existence of chaperones illustrates that the probabilities of proper folding sometimes need to be improved. Chaperones, such as GroEL and GroES, help the unfolded polymer avoid aggregating irreversibly. The chaperones probably do not affect the thermodynamic outcome of folding (Jaenicke, 1991).

Chaperones are not traditional catalysts. They often work by binding to an unfolded protein, frequently remaining attached throughout the folding process (Jaenicke, 1991). They often require ATP or GTP to remove them from the fully folded protein. They prevent the proteins they associate with from irreversibly aggregating in inclusion bodies *in vivo* or aggregates *in vitro*. The job of chaperones is to enhance the yield of protein folding.[10] Not all of them increase the rate of folding (Jaenicke, 1991) because chaperones catalyze the folding at the step where proteins are prone to aggregate, and this need not be the rate-limiting step.

There are other protein folding catalysts, such as the peptidyl-prolyl *cis-trans* isomerases (PPIases) and protein disulfide isomerase, which also accelerate the formation

[10] GroEL *reduces* the folding yield of many mutants of barnase (Gray *et al.*, 1993) and *unfolds* wild-type barnase (Corrales & Fersht, 1995). However, barnase does not epitomize protein interactions with chaperones (Gray *et al.*, 1993). Barnase is a small, 110 amino acid residue, quickly folding, single-domain protein that may not rely on chaperones to fold *in vivo*.

of mature proteins. However, these enzymes only hasten the folding of a protein which would have folded eventually. *Cis-trans* isomerization of peptidyl-proline bonds is the rate-limiting step in the refolding of slow-folding fractions of a number of proteins, e.g., cytochrome *c* (Wetzel, 1992). In order for a PPIase to enhance the folding yield of a protein, the catalyst would have to speed the transition of the folding polymer through a step prone to aggregation.

In summary, a chaperone protects a protein from aggregating until it can fold properly. Chaperones alter the probability of successful folding, and hence, the folding yield. In contrast, the peptidyl-prolyl *cis-trans* isomerases (PPIases) and protein disulfide isomerase change the rate of folding.

1.5 Current directions of investigations

1.5.1 Laboratory experiments

The critical, early events are extremely difficult to observe experimentally. They occur within the dead-time of stopped-flows (a few milliseconds) for many single-domain, water soluble proteins (Radford, 1992; Elove, 1992). And, experiments to study proteins folding at submillisecond time scales are technically difficult. Even if successful, such experiments can supply only limited information about the kinetics of folding because the experiments rely on fast probes, such as UV absorption, fluorescence changes, and CD, that can only provide a few average properties of an entire ensemble of polymers. Furthermore, the experiments are usually specific for proteins with cold denaturations (Nölting *et al.*, 1995; Ballew *et al.*, 1996; Dyer *et al.*, 1996) or hemes whose oxidation state affects the stability of the protein (Mines *et al.*, 1996; Pascher *et al.*, 1996; Winkler & Gray, 1996). Experiments that will be more generalizable to other types of proteins are being developed. For example, Chan *et al.* (1996) are initiating folding by ultrafast mixing, while Kholodenko *et al.* (1996) and Rock *et al.* (1996) are initiating folding by photolyzing an engineered, denaturing covalent bond.

1.5.2 Computer simulations

Molecular dynamics (MD) appears to be an ideal tool for investigating early folding events and how their rates depend on amino acid sequence, solvent, and ions. After all, MD can give an atom-by-atom, picosecond-by-picosecond propagation of a kinetic event.[11] Unfortunately, it has not been practical to perform MD for times long enough either to capture the critical early events of protein folding, which might require microseconds to milliseconds, or to run enough simulations to acquire a statistically significant sample of simulations.

Researchers investigating protein folding have simplified their simulations in a variety of ways. Until the recent development of the Newton-Euler Inverse Mass Operator (NEIMO) method for MD (Jain *et al.*, 1993), it had not been possible to follow the formation of an α-helix for more than a few trajectories starting from the extended, *non*helical state.

To reduce the computational cost yet describe statistical ensembles of folding proteins, many groups use lattices to run Monte Carlo (MC) simulations of copolymers consisting of only two types of residues—polar and nonpolar. These simulations are generally good for describing the degree of cooperativity of the transition to the lowest-energy state, the dependence of cooperativity on the degree of attraction between like monomers, and how the sequence of polar and hydrophobic monomers affects the folding kinetics (Onuchic & Socci, 1995; Socci *et al.*, 1996; Chan & Dill, 1994; Mirny *et al.*, 1996). Some of the simulations are so simplified they occur on two-dimensional lattices (Miller *et al.*, 1992).

Statistical ensembles of folding molecules have not been possible with more realistic representations of whole proteins. Proteins more lifelike than copolymers are

[11] Molecular dynamics operates by defining a set of equations of motion, determining the forces and accelerations on each body in the molecule, integrating over a specific timestep, moving each body to its new position, and then repeating the cycle. See Appendix A.

sometimes simulated on high coordinate lattices (Skolnick & Kolinski, 1996). For example, to simulate the folding pathway of two large α/β-proteins, triose phosphate isomerase and the α-subunit of tryptophan synthase, Godzik *et al.* (1992) put the proteins on a lattice and used the Metropolis sampling criteria (Metropolis & Ulam, 1949) to move the atoms. They claim the model represents the C_α positions to within 2.5–3 Å deviation of the crystal structure, and their simulation succeeding in predicting properties of an experimentally observable intermediate. Lattices have biases (Gregoret & Cohen, 1991), but simulating an entire protein off-lattice is more computationally expensive.

Rather than fold an α-helix *de novo*, many studies of polyalanine α-helices study its motions at equilibrium. Both Gō and Gō (1976) and Levy and Karplus (1979) performed analytical studies of preformed α-helices at equilibrium to characterize fluctuations in backbone dihedral angles. More recently, atomistic MD simulations were performed. For example, Daggett *et al.* (1991) ran 4 ns of MD simulations of a polyalanine relaxing from the ideal α-helical configuration where all $(\phi, \psi) = (-57°, -47°)$. Daggett *et al.* were able to characterize the degree of cooperatively of the helix-coil transition to determine the free energy, enthalpy, and entropy of the helix-coil transition (1991).

To describe the kinetics of the folding process, Brooks (1996) solved rate equations based on helix-coil transition theory. He calculated half-lives of 20–70 ns, depending on the sequence. In an attempt to use MD to model part of the folding process, Pleiss and Jähnig (1992) ran MD on a kinked, α-helical polyalanine to watch it straighten itself. They did not find a high energy transition state for the straightening process. They concluded that the transition was retarded by a random search over a large landscape and not by high energy or entropic barriers.

Recently groups are starting to use nonhelical polyalanine to simulate the formation, *de novo*, of an α-helix. They often use Monte Carlo techniques (Sung, 1995 & 1994) with implicit solvent, although recently Sung has used MD to do the same (Sung & Wu, 1996). Sung experimented with the AMBER force field and found that polyalanine, starting from all ϕ and ψ bonds being 180°, formed an α-helix fastest when electrostatic forces were strong. Perhaps because electrostatic forces are long-range, they seemed to be able to guide productive helix formation (Sung, 1995).

The recent development of the Newton Euler Inverse Mass Operator (NEIMO) method for MD (Jain *et al.*, 1993) enables one (a) to use explicit atom models of protein 20-mers, (b) to follow processes at the early folding stages, and (c) to run almost 200 comparable simulations—enough to make generalizations about the process of helix formation. This is a significant achievement in helix-folding simulations. The rest of this thesis describes results of simulations of polyalanine and polyglycine folding into α-helices.

1.6 References

Agard, D. (1997) seminar at the California Institute of Technology, Pasadena, CA, February 21, 1997.

Baker, D. & Agard, D.A. (1994) Kinetics versus thermodynamics in protein folding. *Biochemistry 33*: 7505-7509.

Baker, D., Sohl, J.L., Agard, D.A. (1992a) A protein-folding reaction under kinetic control. *Nature 19*: 263-265.

Baker, D., Silen, J.L., Agard, D.A. (1992b) Protease pro region required for folding is a potent inhibitor of the mature enzyme. *PROTEINS: Structure, Function, and Genetics 12*: 339-344.

Ballew, R.M., Sabelko, J., Gruebele, M. (1996) Direct observation of fast protein folding: The initial collapse of apomyoglobin. *PNAS USA 93(12)*: 5759–5764.

Banzon, J.A. & Kelly, J.W. (1992) β-Sheet rearrangements: Serpins and beyond. *Protein Engineering 5(2)*: 113-115.

Branden, C. & Tooze, J. (1991) *Introduction to Protein Structure* (New York, Garland Publ., Inc.) pp. 251-252.

Brems, D.N., Plaisted, S.M., Kauffman, E.W., Havel, H.A. (1986) Characterization of an associated equilibrium folding intermediate of bovine growth hormone. *Biochemistry 25*: 6539-6543.

Briggs, M.S. & Roder, H. (1992) Early hydrogen-bonding events in the folding reaction of ubiquitin. *PNAS USA 89*: 2017-2021.

Brooks, C.L. (1996) Helix-coil kinetics: Folding time scales for helical peptides from a sequential kinetic model. *J. Phys. Chem. 100*: 2546–2549.

Cavagnero, Silvia (1996) Towards understanding hyperthermostability of rubredoxin from *Pryrococcus furiosus*. (Ph.D. thesis, California Institute of Technology), pp. 196–217.

Chaffotte, A. F., Guillou, Y., Goldberg, M.E. (1992) Kinetic resolution of peptide bond and side chain far-UV circular dichroism during the folding of hen egg white lysozyme. *Biochemistry 31*: 9694–9702.

Chan, H.S. & Dill, K. A. (1990) Origins of structure in globular proteins. *PNAS USA 87*: 6388-6392.

Chan, H.S. & Dill, K.A. (1994) Transition states and folding dynamics of proteins and heteropolymers. *J. Chem. Phys. 100*: 12.

Chan, H.S., Bromberg, S., Dill, K.A. (1995) Models of cooperativity in protein folding. *Philosophical Transactions of the Royal Society of London, Series B. Biol. Sciences 348*: 1323.

Chan, C.-K., Hu, Y., Takahashi, S., Rousseau, D.L., Eaton, W.A., Hofrichter, J. (1996) Submillisecond protein folding kinetics studied by ultrafast mixing. *Abstracts of the Am. Chem. Soc. 212*: 150.

Chargé, S.B.P., De Konig, E.J.P., Clark, A. (1995) Effect of pH and insulin on fibrillogenesis of islet amyloid polypeptide *in vitro*. *Biochemistry 34(44)*: 14588-14593.

Chen, J., Wharton, S.A., Weissenhorn, W., Calder, L.J., Hughson, F.M. *et al.*, (1995) A soluble domain of the membrane-anchoring chain of influenza virus hemagglutinin (HA(2)) folds in *Escherichia coli* into the low pH induced conformation. *PNAS USA 92(26)*: 12205-12209.

Cleland, J.L. & Wang, D.I.C. (1990) Refolding and aggregation of bovine carbonic anhydrase B: Quasi-elastic light scattering analysis. *Biochemistry 29*: 11072-11078.

Corrales, F.J. & Fersht, A.R. (1995) The folding of GroEL-bound barnase as a model for chaperonin-mediated protein folding. *PNAS USA 92*: 5326-5330.

Danner, M. & Seckler, R. (1993) Mechanism of phage P22 tailspike protein folding mutations. *Protein Science 2*: 1869-1881.

Daggett, V., Kollman, P.A., Kuntz, I.D. (1991) A Molecular dynamics simulation of polyalanine: An Analysis of equilibrium motions and helix-coil transitions. *Biopolymers 31*: 1115–1134.

Debe, D. (1997) personal communication at Caltech.

Debe, D.A., Chan, S.I., Goddard, W.A. (1997) Resolution of the Levinthal Paradox: Practical sampling of protein folding conformations. *in progress.*

Dill, D.A. (1990) Dominant forces in protein folding. *Biochemistry 29(31)*: 7133–7154.

Dill, D.A. (1985) Theory for the folding and stability of globular proteins. *Biochemistry 24*: 1501–1509.

Dill, K.A. (1987) "The stabilities of globular proteins." In *Protein Engineering*, ed. by Oxender, D.L. & Fox, C.F. (New York, Alan R. Liss, Inc.), pp. 187-192.

Dill, K.A. & Chan, H.S. (1997) From Levinthal to pathways to funnels. *Nature Structural Biology 4(1)*: 10–19.

Dunbrack, R.L., Gerloff, D.L., Bower, M., Chen, X., Lichtarge, O., Cohen, F.E. (1997) *Folding and Design 1*: R27–R42.

Dyer, R.B., Williams, S., Woodruff, W.H. (1996) The Earliest events in protein folding: Helix dynamics in proteins and model peptides. *Abstracts of Papers of the American Chemical Society 212(2)*: 150.

Eder, J., Rheinnecker, M., Fersht, A.R. (1993) Folding of subtilisin BPN′: Characterization of a folding intermediate. *Biochemistry 32*: 18–26.

Elöve, G.A., Chaffotte, A.F., Roder, H., Goldberg, M.E. (1992) Early steps in cytochrome c folding probed by time-resolved circular dichroism and fluorescence spectroscopy. *Biochemistry 31*: 6876–6883.

Fischer, B., Perry, B., Sumner, I., Goodenough, P. (1992) A Novel sequential procedure to enhance the renaturation of recombinant protein from *Escherichia coli* inclusion bodies. *Protein Engineering 5(6)*: 593–596.

Flanagan, J.M., Kataoka, M., Fujisawa, T., Engelman, D. (1993) Mutations can cause large changes in the conformation of a denatured protein. *Biochemistry 32*: 10359–10370.

Frishman, D. & Argos, P. (1996) Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Engineering 9(2)*: 133–142.

Frishman, D. & Argos, P. (1997) Seventy-five percent accuracy in protein secondary structure prediction. *PROTEINS: Structure, Function, and Genetics 27*: 329–335.

Georgiou, G. & De Bernardez-Clark, E., eds. (1991) *Protein Refolding* (Washington D.C., American Chemical Society), p. ix.

Gō, M. & Gō, N. (1976) Fluctuations of an α-helix. *Biopolymers 15*: 1119–1127.

Goddard, William A. III (1997) personal communication at Caltech.

Godzik, A., Skolnick, J., Kolinski, A. (1992) Simulations of the folding pathway of triose phosphate isomerase-type alpha-beta barrel proteins. *PNAS USA 89*: 2629-1633.

Goldberg, M.E., Rudolph, R., Jaenicke, R. (1991) A Kinetic study of the competition between renaturation and aggregation during the refolding of denatured-reduced egg white lysozyme. *Biochemistry 30*: 2790-2797.

Goldenberg, D.P., Smith, D.H., King, J. (1983) Genetic analysis of the folding pathway for the tail spike protein of phage P22. *PNAS USA. 80*: 7060-7064.

Goloubinoff, P., Gatenby, A.A., Lorimer, G.H. (1991) "Role of chaperonins in protein folding." In: Georgiou, George & De Bernardez-Clark, Eliana, eds. *Protein Refolding* (Washington, D.C., American Chemical Society), pp. 110–118.

Gray, T.E., Eder, J., Bycroft, M., Day, A.G., Fersht, A.R. (1993) Refolding of barnase mutants and pro-barnase in the presence and absence of GroEL. *The EMBO Journal 12(11)*: 4145-4150.

Gregoret, L.M. & Cohen, F.E. (1991) Protein folding: Effect of packing density on chain conformation. *J. Mol. Biol. 219(1)*: 109–122.

Haase-Pettingell, C.A. & King, J. (1988) Formation of aggregates from a thermolabile *in vivo* folding intermediate in P22 tailspike maturation. *J. Biol. Chem. 263(10)*: 4977-4983.

Jackson, G.S., Staniforth, R.A., Halsall, D.J., Atkinson, T., Holbrook, J.J., Clarke, A.R., Burston, S.G. (1993) Binding and hydrolysis of nucleotides in the chaperonin catalytic cycle: Implications for the mechanism of assisted protein folding. *Biochemistry 32*: 2554-2563.

Jaenicke, R. (1988) "Is there a code for protein folding?" in *Protein Structure and Protein Engineering*, eds. E.L. Winnacker & R. Huber, (Berlin, Springer-Verlag) p. 16.

Jaenicke, R. (1991) Protein folding: Local structures, subunits, and assemblies. *Biochemistry 30(13)*: 3147-3161.

Jain, A., Vaidehi, N., Rodriguez, G. (1993) A Fast recursive algorithm for molecular dynamics simulations. *J. Computational Physics 106*: 258–268.

Jarrett, J.T. & Lansbury, P.T., Jr. (1992) Amyloid fibril formation requires a chemically discriminating nucleation event: Studies of an amyloidogenic sequence from the bacterial protein *Biochemistry 31(49)*: 12345-12352.

Karplus, M. & Weaver, D.L. (1976) Protein-folding dynamics. *Nature 260*: 404-406.

Kiefhaber, T. & Baldwin, R.L. (1995) Intrinsic stability of individual α helices modulates structure and stability of the apomyoglobin molten globule form. *J. Mol. Biol. 252*: 122-132.

Kim, P.S. & Baldwin, R. L. (1990) Intermediates in the folding reactions of small proteins. Ann. Rev. *Biochemistry 59*: 631-660.

Kocisko, D.A., Priola, S.A., Raymond., G.J., Chesbro, B., Lansbury, P.T. Caughey, B. (1995) Species specificity in the cell-free conversion of prion protein to protease-resistant forms: A model for the scrapie species barrier. *PNAS 92*: 3923-3927.

Kuwajima, K. (1989) The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure. *Proteins: Structure, Function, & Genetics 6*: 87-103.

Kuwajima, K., Mitani, M., Sugai, S. (1989) Characterization of the critical state in protein folding: Effects of guanidine hydrochloride and specific Ca+2 binding on the folding kinetics of alpha-lactalbumin. *J. Mol. Biol. 206*: 547-561.

Lehrman, S.R., Tuls, J.L, Havel, H.A., Haskell, R.J., Putnam, S.D., Tomich, C.-S.C. (1991) Site-directed mutagenesis to probe protein folding: Evidence that the formation and aggregation of a bovine growth hormone folding intermediate are dissociable processes. *Biochemistry 30*: 5777-5784.

Lemer, C.M.-R., Rooman, M.J., Wodak, S.J. (1995) Protein structure prediction by threading methods: Evaluation of current techniques. *PROTEINS: Structure, Function, and Genetics 23*: 337–355.

Levinthal, C. (1968) Are there pathways for protein folding? *J. Chim. Phys. 85*: 44-45.

Levinthal, C. (1969) in *Mössbauer Spectroscopy in Biological Systems* (proceedings of a meeting held at Allerton House, Monticello, Ill.), pp. 22–24.

Levy, R.M. & Karplus, M. (1979) Vibrational approach to the dynamics of an α-helix. *Biopolymers 18*: 2465–2495.

London, J., Skrzynia, C., & Goldberg, M. (1974) Renaturation of *Escherichia coli* tryptophanase after exposure to 8 M urea. *Eur. J. Biochem. 47*: 409-415.

Lorimer, G. (1992?) personal communication.

Madej, T., Gibrat, J.-F., Bryant, S.H. (1995) Threading a database of protein cores. *PROTEIN: Structure, Function, Genetics 23*: 356–369.

Mann, C.J. & Matthews, C.R. (1993) Structure and stability of an early folding intermediate of *Escherichia coli trp* aporepressor measured by far-UV stopped-flow circular dichroism and 8-anilino-1-naphthalene sulfonate binding. *Biochemistry 32*: 5282–5290.

Metropolis, N. & Ulam, S.M. (1949) *J. Am. Stat. Asso. 44*: 247.

Miller, R., Danko, C.A., Fasolka, M.J., Balazs, A.C., Chan, H.S., Dill, K.A. (1992) Folding kinetics of proteins and copolymers. *J. Chem. Phys. 96*: 1.

Miranker, A., Robinson, C.V., Radford, S.E., Aplin, R.T., Dobson, C.M. (1993) Detection of transient protein folding populations by mass spectrometry. *Science 262*: 896–900.

Mirny, L.A., Abkevich, V., Shakhnovich, E.I. (1996) Universality and diversity of the protein folding scenarios: A comprehensive analysis with the aid of a lattice model. *Folding and Design 1*: 2.

Mitraki, A., Betton, J.-M., Desmadril, M., Yon, J.M. (1987) Quasi-irreversibility in the unfolding-refolding transition of phosphoglycerate kinase induced by guanidine hydrochloride. *Eur. J. Biochem. 163*: 29-34.

Mitraki, A., Danner, M., King, J., Seckler, R. (1993) Temperature-sensitive mutations and second-site suppressor substitutions affect folding of the P22 tailspike protein *in vitro*. *J. Biol. Chem. 268(27)*: 20071-20075.

Mitraki, A., Fane, B., Haase-Pettingell, C., Sturtevant, J., King, J. (1991) Global suppression of protein folding defects and inclusion body formation. *Science 253*: 54-58.

Mitraki, A. & King, J. (1992) Amino acid substitutions influencing intracellular protein folding pathways. *FEBS 307(1)*: 20-25.

Mosimann, S., Meleshko, R., James, M.N.G. (1995) A Critical assessment of comparative molecular modeling of tertiary structures of proteins. *PROTEINS: Structure, Function, and Genetics 23*: 301–317.

Nölting, B., Golbik, R., Fersht, A.R. (1995) Submillisecond events in protein folding. *PNAS USA 92*: 10668–10672.

Pleiss, J. & Jähnig, F. (1992) Conformational transition of an α-helix studies by molecular dynamics. *European Biophysics Journal 21(1)*: 63–70.

Ptitsyn, O.B., Pain, R.H., Semisotnov, G.V., Zerovnik, E., Razgulyaev, O.I. (1990) Evidence for a molten globule state as a general intermediate in protein folding. *FEBS Letters262(1)*: 20–24.

Ptitsyn, O.B. & Semisotnov, G.V. (1991) "The Mechanism of protein folding" in *Conformations and Forces in Protein Folding*, ed. by Dill, K.A. & Nall, B.T. (American Association for the Advancement of Science, Washington, D.C.), pp. 155-168.

Qian, N. & Sejnowski, T.J. (1988) Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol. 202*: 865-884.

Radford, S.E., Dobson, C.M., Evans, P.A. (1992) The folding of hen lysozyme involves partially structured intermediates and multiple pathways. *Nature 358*: 302–307.

Roder, H., Elöve, G.A., Englander, S.W. (1988) Structural characterization of folding intermediates in cytochrome *c* by H-exchange labeling and proton NMR. *Nature 335*: 700–704.

Rost, B. & Sander, C. (1995) Progress of 1D protein structure prediction at last. *PROTEINS: Structure, Function, and Genetics 23(3)*: 295–300.

Saraiva, M.J.M., Birken, S., Costa, P.P., Goddman, D.S. (1984) Amyloid fibril protein in familial amyloidotic polyneuropathy, Portuguese type: Definition of molecular abnormality in transthyretin (prealbumin). *J. Clin. Invest. 74*: 104-119.

Schmid, F.X., Grafl, R., Wrba, A., Beintema, J. (1986) Role of proline peptide bond isomerization in unfolding and refolding of ribonuclease. *PNAS USA 83*: 872-876.

Sharp, K. & Honig, B. (1990) Electrostatic interactions in macromolecules: Theory and applications. *Annual Review of Biophysics and Biophysical Chemistry 19*: 301–332.

Skolnick, J. & Kolinski, A. (1996) "Monte Carlo lattice dynamics and the prediction of protein folds," preprint of a chapter in a book.

Smith, D. & King, J. (1981) Temperature-sensitive mutants blocked in the folding or subunit assembly of the bateriophage P22 tail spike protein. *J. Mol. Biol. 145*: 653-676.

Socci, N.D. & Onuchic, J.N. (1995) Kinetic and thermodynamic analysis of protein-like heteropolymers: Monte Carlo histogram technique. *J. Chem. Phys. 103*: 11.

Socci, N.D., Onuchic, J.N., Wolynes, P.G. (1996) Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys. 104*: 15.

Speed, M.A., Wang, D.I.C., King, J. (1996) Specific aggregation of partially folded polypeptide chains: The molecular basis of inclusion body composition. *Nature Biotechnology 14*: 1283-1287.

Speed, M.A., Morshead, T., Wang, D.I.C., King, J. (1997) Conformation of P22 tailspike folding and aggregation intermediates probed by monoclonal antibodies. *Protein Science 6(1)*: 99-108.

Srinivasan, R. & Rose, G.D. (1995) LINUS: A Hierarchic procedure to predict the fold of a protein. *PROTEINS: Structure, Function, and Genetics 22(2)*: 81–99.

Steinbacher, S., Seckler, R., Miller, S., Steipe, B., Huber, R., Reinemer, P. (1994) Crystal structure of P22 tailspike protein: Interdigitated subunits in a thermostable trimer. *Science 265*: 383-386.

Su, A. (1997) personal communication at Caltech.

Sugawara, T., Kuwajima, K., Sugai, S. (1991) Folding of Staphylococcal nuclease A studied by equilibrium and kinetic circular dichroism spectra. *Biochemistry 30*: 2698–2706.

Sung, S.-S. & Wu, X.-W. (1996) Molecular dynamics simulations of synthetic peptide folding. *PROTEINS: Structure, Function, and Genetics 25*:202–214.

Sung, S.-S. (1994) Helix folding simulations with various initial conformations. *Biophysical Journal 66*:1796–1803.

Sung, S.-S. (1995) Constant temperature simulations of helix folding. *J. theor. Biol. 173*:398–400.

Truong, H.-T., Pratt, E.A., Rule, G.S., Hsue, P.Y., Ho, C. (1991) Inactive and temperature-sensitive folding mutants generated by tryptophan substitutions in the membrane-bound D-lactate dehydrogenase of *Escherichia coli. Biochemistry 30*: 10722-10729.

Viitanen, P.V., Lubben, T.H., Reed, J., Goloubinoff, P., O'Keefe, D.P., Lorimer, G. (1990) Chaperonin-facilitated refolding of ribulosebisphosphate carboxylase and ATP hydrolysis by chaperone 60 (GroEL) are K^+ dependent. *Biochemistry 29*: 5665-5671.

Wetlaufer, D.B. (1973) Nucleation, rapid folding, and globular intrachain regions in proteins. *PNAS USA 70(3)*: 697–701.

Wetlaufer, D.B. & Xie, Y. (1995) Control of aggregation in protein refolding: A Variety of surfactants promote renaturation of carbonic-anhydrase-II. *Protein Science 4(8)*: 1535–1543.

Wetzel, R. (1992) "Protein aggregation *in vivo*: Bacterial inclusion bodies and mammalian amyloid," in *Stability of Protein Pharmaceuticals*, ed.s Tim J. Ahera & Mark C. Manning (New York, Plenum Press), pp. 43-88.

Wetzel, R. & Chrunyk, B.A. (1993) "Mutational effects on inclusion body formation," in *Biocatalyst design for stability and specificity*, eds. by Michael E. Himmel & George Georgiou (Washington, D.C., American Chemical Society), pp. 116-125.

White, J.M. (1994) in *Receptor mediated virus entry into cells* (Wimmer, E., Ed.) Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY (in press).

Wilkinson, D.L. & Harrison, R.G. (1991) Predicting the solubility of recombinant proteins in *Escherichia coli. Bio/Technology 9*: 443-448.

Winther, J.R. & Sørensen, P. (1991) Propeptide of carboxypeptidase Y provides a chaperone-like function as well as inhibition of the enzymatic activity. *PNAS USA 88*: 9330–9334.

Wolynes, P.G., Onuchic, J.N., Thirumalai, D. (1995) Navigating the folding routes. *Science 267*: 1619–1620.

Wright, P., Dyson, J., Lerner, R. (1988) Conformation of peptide fragments of proteins in aqueous solution: Implications for initiation of protein folding. *Biochemistry (27)19*: 7167-7175.

Yu, M.-H. & King, J. (1988) Surface amino acids as sites of temperature-sensitive folding mutations in the P22 tailspike protein. *J. Biol. Chem. 263(3)*: 1424-1431.

Zettlmeissl, G., Rudolph, R., Jaenicke, R. (1979) Reconstitution of lactic dehydrogenase. Noncovalent aggregation vs Reactivation. 1. Physical properties and kinetics of aggregation. *Biochemistry 18(25)*: 5567–5571.

Zettlmeissl, G., Teschner, W., Rudolph, R., Jaenicke, R. & Gade, G. (1984) Isolation, physicochemical properties, and folding of octopine dehydrogenase from *Pecten jacobaeus. Eur. J. Biochem. 143*: 401-407.

Zhu, X., Ohta, Y., Jordan, F., Inouye, M. (1989) Pro-sequence of subtilisin can guide the refolding of denatured subtilisin in an intermolecular process. *Nature 339*: 483–484.

Zubay, Geoffrey L. (1988) *Biochemistry* (New York, Macmillan Publishing Company), 2nd ed., p. 335.

Chapter 2 Example of a Fast Helix-forming Trajectory

2.1 Outline of the sequence of structural changes during helix folding

2.1.1 The trajectory

Before discussing experiments to elucidate the requirements for helix formation, let us look at an example of a simulation of polyalanine that forms a helix quickly and simply.[1] Besides looking at the animated trajectory, we can view the process with many analysis tools that are designed to extract only the important features out of the volumes of information each animation contains.

Figure 2.1 illustrates the trajectory. This trajectory forms an α-helix within 100 ps. The animation shows the polymer starting at 0 ps in the extended state, where the backbone dihedral angles $(\phi, \psi) \approx (-180°, +180°)$. (Refer to Appendix D for a definition of the dihedral angles ϕ and ψ. Within a few picoseconds, the $(\mathrm{Ala})_{20}$ relaxes to configurations where each residue could stably hydrogen bond to its neighbor two residues away. This will be referred to as an "$i, i+2$" hydrogen bond (HB) because it binds the backbone carbonyl oxygen of the ith residue to the amide proton of the ith$+2$ residue. The polymer undulates in these configurations for about 12 ps.

From about 12–15 ps the polymer forms the first α-helical hydrogen bond, a.k.a. the first $i, i+4$ HB, or, the first HB between residues separated by three intervening residues. This HB nucleates the folding of the rest of the helix by forming one loop

[1] All figures in this chapter are derived from the trajectory labeled "pAnc-450K-ad-15A" in Appendix D.

of the α-helix. This nucleation will be described in more detail in Figure 2.8. At ~20 ps, another residue (residue 12) adds to the central helical loop, thereby lengthening the central helix.

Around 35 ps the C-terminal residues appear as if they will nucleate another helix, but the HBs they form do not propagate down towards the existing central helix. Instead, the central helix grows out to the C-terminus, starting at about 37 ps and finishing by about 42 ps. The N-terminal half of the polymer flails around until about 64 ps when residues 2 and higher adopt the α-helical conformation. The N-terminal residue does not form an α-helical HB until 99 ps. For the remaining 275 ps of the dynamics simulation, the helix wiggles slightly, presumably at equilibrium.

Figure 2.1: Figure on p. 43. A representative trajectory of polyalanine folding fast. All of the illustrations of polypeptides depict the carbon and nitrogen backbone in dark grey, carboxyl oxygens in black, and the amide protons in white. The amino termini are at the bottom of each picture. In this figure only, the 8th amide proton of each conformation is equidistant from the bottom of the figure. A. The initial conformation has $\phi, \psi = 180°$. B. At 3 ps, each residue HBs to its neighbor two residues away ($i, i+2$ HBs). Three such HBs are illustrated with dashed lines. C. At 17 ps, the first $i, i+4$, α-helical HB forms when the carbonyl oxygen of residue 8 HBs to the amide proton of residue 12. D. At 30 ps, there are two α-helical HBs. E. At 37 ps, the carboxy-terminal half has almost formed a helix; residues 16 and 19 form an $i, i+3$ HB. The amino terminal half has formed an $i, i+3$ bond from residue 2 to residue 5. F. At 44 ps, residues 8-20 are in an α-helix. G. At 72 ps, the α-helix has formed, although residue 1 is disordered. All figures in this chapter are drawn from the trajectory labeled "pAnc-450K-ad-15A" in Appendix D.

2.1.2 The sequence and location of sites of nucleation and propagation

Figure 2.2, p. 44, clearly illustrates the propagation and nucleation steps in the previous simulation. At each time point, the scroll-like figure depicts each residue as an "h," a slash ("/"), or a period ("."), depending on the (ϕ, ψ) dihedrals. "h"s represent α-helical residues, i.e., those whose dihedral angles are within a 30° radius

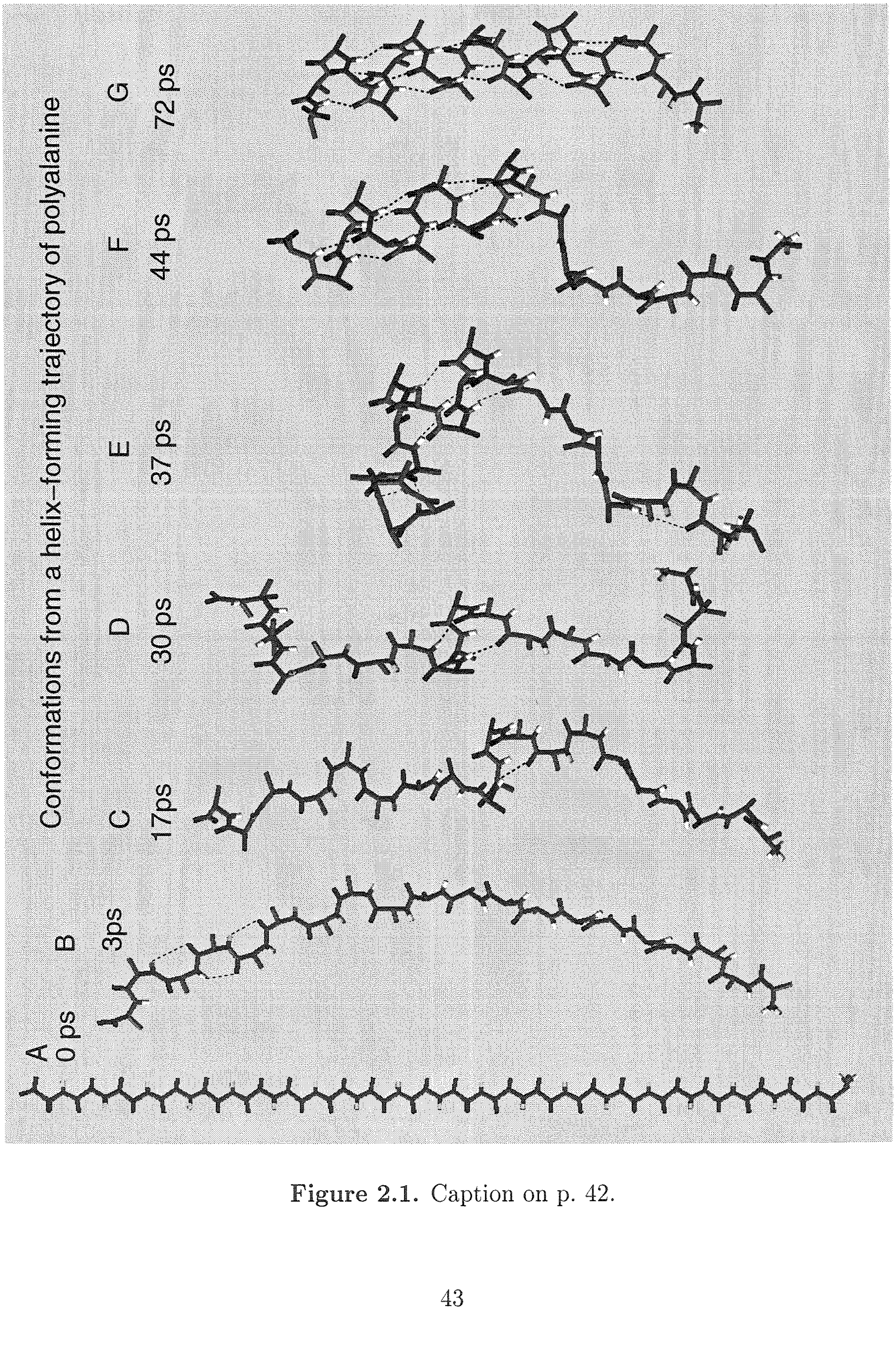

Conformations from a helix–forming trajectory of polyalanine
A 0 ps
B 3ps
C 17ps
D 30 ps
E 37 ps
F 44 ps
G 72 ps

Figure 2.1. Caption on p. 42.

of the classic α-helical dihedrals, $(\phi, \psi) = (-57°, -47°)$. The "slash" ("/") denotes a residue within a largely helical region of Ramachandran space, a region populated by residues in good quality X-ray structures (region A of PROCHECK (Laskowski *et al.*, 1993). See Figures 4.2 and 5.1). Each residue symbolized by a period (".") is in a coiled, nonhelical structure. This figure clearly illustrates nucleation at residue 9 at 12 ps. After forming an initial α-helical loop, the helical region grows out to the C-terminus and reaches it by 43 ps. Although residues near the N-terminus tried to nucleate a lasting helix starting as early as 24 ps, the N-terminus did not become an α-helix until after residue 2 formed an $i, i+4$ HB with residue 6 at 64 ps. After 64 ps, this N-terminal loop reorients and attaches itself to the base of the main helix. By 72 ps the entire helix except for residue 1 is α-helical.

Figure 2.2: pp. 44–47. Nucleation and propagation for a fast helix-forming trajectory. At each time point, each residue is symbolized as an "h," a slash ("/"), or a period ("."), depending on the (ϕ, ψ) dihedrals. "h"s represent α-helical residues, i.e., those within a 30° radius of the classic α-helical dihedral angles, $(\phi, \psi) = (-57°, -47°)$. The "/" character denotes a residue within a largely helical region of Ramachandran space, a region populated by residues in good quality X-ray structures (region A of PROCHECK (Laskowski *et al.*, 1993). See Figures 4.2 and 5.1.). Residues symbolized by a period (".") are in a coiled, nonhelical structure at all other regions of Ramachandran space.

```
# Residues classified by structure: helix or coil
# From tor file  pAnc-450K-ad-15A.tor
# from residues  2 to  19
# from      0.00 ps to   100.00 ps averaging every   10 points
# residues 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9
     1.00  . . . . . . . . . . . . . . . . . .
     2.00  . . . . . . . . . . . . . . . . . .
     3.00  . . . . . . . . . . . . . . . . . .
     4.00  . . . . . . . . . . . . . . . . . .
     5.00  . . . . . . . . . . . . . . . . . .
     6.00  . . . . . . . . . . . . . . . . . .
     7.00  . . . . . . . . . . . . . . . . . .
     8.00  . . . . . . . . . . . . . . . . . .
     9.00  . . . . . . . . . . . . . . . . . .
```

```
10.00   . . . . . . . . . . . . . . . . . .
11.00   . . . . . . . . . . . . . . . . . .
12.00   . . . . . . . h . . . . . . . . . .
13.00   . . . . . . . h / . . . . . . . . .
14.00   . . . . . . . h / . . . . . . . . .
15.00   . . . . . . . h h h . . . . . . . .
16.00   . . . . . . . h h / . . . . . . . .
17.00   . . . . . . . h / . . . . . . . . .
18.00   . . . . . . . h h / . . . . . . . .
19.00   . . . . . . . h h h / . . . . . . .
20.00   . . . . . . . h h h h . . . . . . .
21.00   . . . . . . . / h h / . . . . . . .
22.00   . . . . . . . h h h / . . . . . . .
23.00   . . . . . . . h h h h . . . . . . .
24.00   . . h / . . . h h . h . . . . . . .
25.00   . . . . . . . h h h h . . . . . . .
26.00   . . / . . . . h h h h . . . . . . .
27.00   . . h . . . . h h / . . . . . . . .
28.00   . . h . . . . h h h / . . . . . . .
29.00   . . h . . . . h / / / . . . . . . .
30.00   . . h . . . . h h / h . . . . . . .
31.00   . . h . . . . h h h h . . . . . . .
32.00   . . h . . . . h h h h . . . . . . .
33.00   / . / . . . . h h h h . . . . . . .
34.00   . . / . . . . h h / h . . . . / / .
35.00   . . h . . . . h h / / . . . . h / .
36.00   . . h . . . . h h / h . . . . h h h
37.00   . . . . . . . h h h h h . . . h h /
38.00   . . / . . . . h h h h h . . . h h .
39.00   . . h . . . . h h h h h . . . h / .
40.00   . . / . . . . h h h h h h . . h . .
41.00   . . h . . . . h h h h h h h . h . .
42.00   . . h . . . . h h h h h h h / h / /
43.00   . . / . . . . h h h h h h h h h h .
44.00   . . / . . . . h h h h h h h h h h /
45.00   . . . . . . . h h h h h h h h h h h
46.00   . . . . . . . h h h h h h h h h h h
47.00   . . . . . . / h h h h h h h h h h h
48.00   . . . . . . h h h h h h h h h h h h
49.00   . . . . . . h h h h h h h h h h h h
50.00   / h . . . . h h h h h h h h h h h h
51.00   h h . . . . h h h h h h h h h h h .
52.00   . . . . . . h h h h h h h h h h h .
```

```
53.00   . h . . . . h h h h h h h h h h h .
54.00   . h . . . . h h h h h h h h h h h .
55.00   . / . . . . h h h h h h h h h h h .
56.00   . . . . . . h h h h h h h h h h h .
57.00   . . . . . . h h h h h h . h h h h .
58.00   . . . . . . h h h h h h h h h h h .
59.00   . . . . . . h h h h h h h h h . h /
60.00   . . . . . . h h h h h h h h h h h /
61.00   . . . . . . h h h h h h h h h h h .
62.00   . . . . . . h h h h h h h h h h h .
63.00   / . . . . . h h h h h h h h h h h /
64.00   . . . h h . h h h h h h h h h h h h
65.00   . . . h . . h h h h h h h h h h h h
66.00   . . . / h . h h h h h h h h h h h h
67.00   . . . . h / h h h h h h h h h h h h
68.00   . . . . h h h h h h h h h h h h h h
69.00   . / h . h h h h h h h h h h h h h h
70.00   . h / . h h h h h h h h h h h h h h
71.00   . h h . h h h h h h h h h h h h h h
72.00   . h h h h h h h h h h h h h h h h h
73.00   . h h h h h h h h h h h h h h h h h
74.00   . h h h h h h h h h h h h h h h h h
75.00   . h h h h h h h h h h h h h h h h h
76.00   . h h h h h h h h h h h h h h h h h
77.00   . h h h h h h h h h h h h h h h h .
78.00   . h h h h h h h h h h h h h h h h h
79.00   . h h h h h h h h h h h h h h h h h
80.00   . h h h h h h h h h h h h h h h h h
81.00   . h h h h h h h h h h h h h h h h h
82.00   . h h h h h h h h h h h h h h h h h
83.00   . h h h h h h h h h h h h h h h h h
84.00   . h h h h h h h h h h h h h h h h h
85.00   . h h h h h h h h h h h h h h h h /
86.00   . h h h h h h h h h h h h h h h h .
87.00   . h h h h h h h h h h h h h h h h h
88.00   . h h h h h h h h h h h h h h h h h
89.00   . h h h h h h h h h h h h h h h h h
90.00   . h h h h h h h h h h h h h h h h h
91.00   . h h h h h h h h h h h h h h h h h
92.00   . h h h h h h h h h h h h h h h h h
93.00   . h h h h h h h h h h h h h h h h h
94.00   . h h h h h h h h h h h h h h h h h
95.00   . h h h h h h h h h h h h h h h h h
```

```
 96.00  . h h h h h h h h h h h h h h h h h
 97.00  . / h h / h h h h h h h h h h h h h
 98.00  h h h h h h h h h h h h h h h h h h
 99.00  h h h h h h h h h h h h h h h h h /
100.00  h h h h h h h h h h h h h h h h h .
```

Figure 2.3: A–C. Pages 48– 50. Stacked plots of the (ϕ, ψ) angles of the residues of $(\text{Ala})_{20}$ folding fast (pAnc-450K-ad-15A in Appendix D). A. Residues 14–19, p. 48. B. Residues 8–13, p. 49. C. Residues 2–7, p. 50. For these three figures, the y axis cycles from -180° to +180°, repeating every residue. The residues started in the extended state, and by the end of the graph at 100 ps they are in the α-helical configuration, where $(\phi, \psi) \approx (-60°, -40°)$. The horizontal lines at -52° approximate the (ϕ, ψ) angles of the α-helix. Points were averaged one point plotted per 0.50 ps simulated, i.e., one point plotted per five data points recorded. Note bene: Error bars on all graphs extend one sigma above and below the average values of the ordinate, and the error bars encompass 67% of the values the ordinates are expected to take. Periodic boundary conditions were not taken into account in calculating the error bars. So, error bars extending over 360° are inaccurate.

Figure 2.3, pp. 47–48, shows the nucleation and propagation of the α-helix in more detail than in Figure 2.2 but more clearly than in the animated trajectory. Figure 2.3 is a series of stacked plots of the (ϕ, ψ) angles of each of the residues. These plots offer the benefits of displaying the entire time evolution of the (ϕ, ψ) angles of several residues on one sheet of paper. The figure shows that residues 9–11 were the first to adopt a helical conformation. This corresponds to the nucleation at ~15ps. By 20 ps, residue 12 had also adopted α-helical dihedral angles, consistent with the trajectory. By ~40 ps, the helix had grown to residues 13–18. And by 70 ps, residues 2–8 and 19 stop flailing about, and the plots show them adopting α-helical torsions.

2.1.3 Other measures of structural rearrangement

Radius of gyration and end-to-end distance. The radius of gyration and the end-to-end distance describe gross structural changes in polymers. Figure 2.4, p. 51, graphs these quantities. From 20–25 ps, the distances drop as the N-terminus

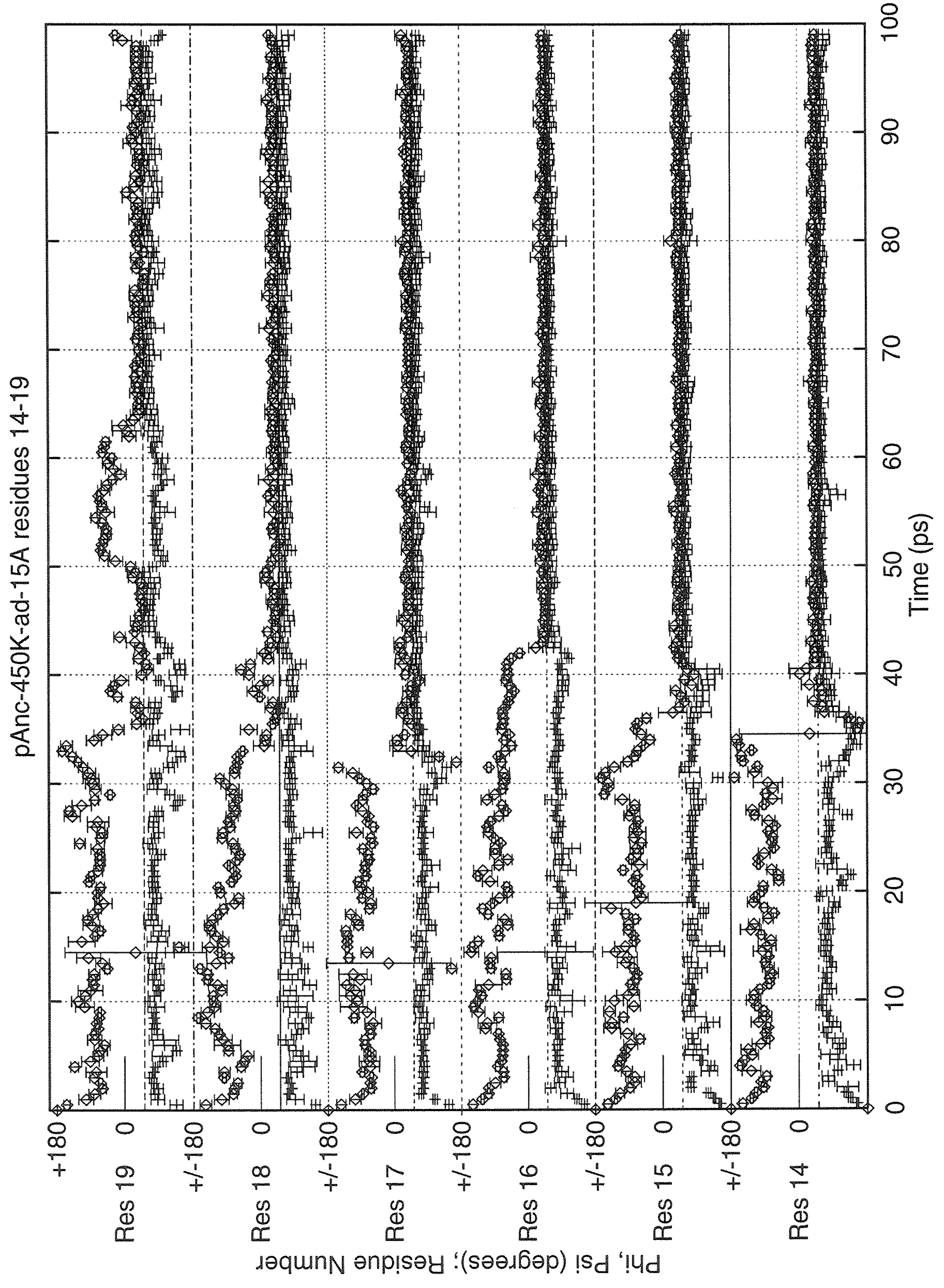


Figure 2.3A: Caption p. 47.

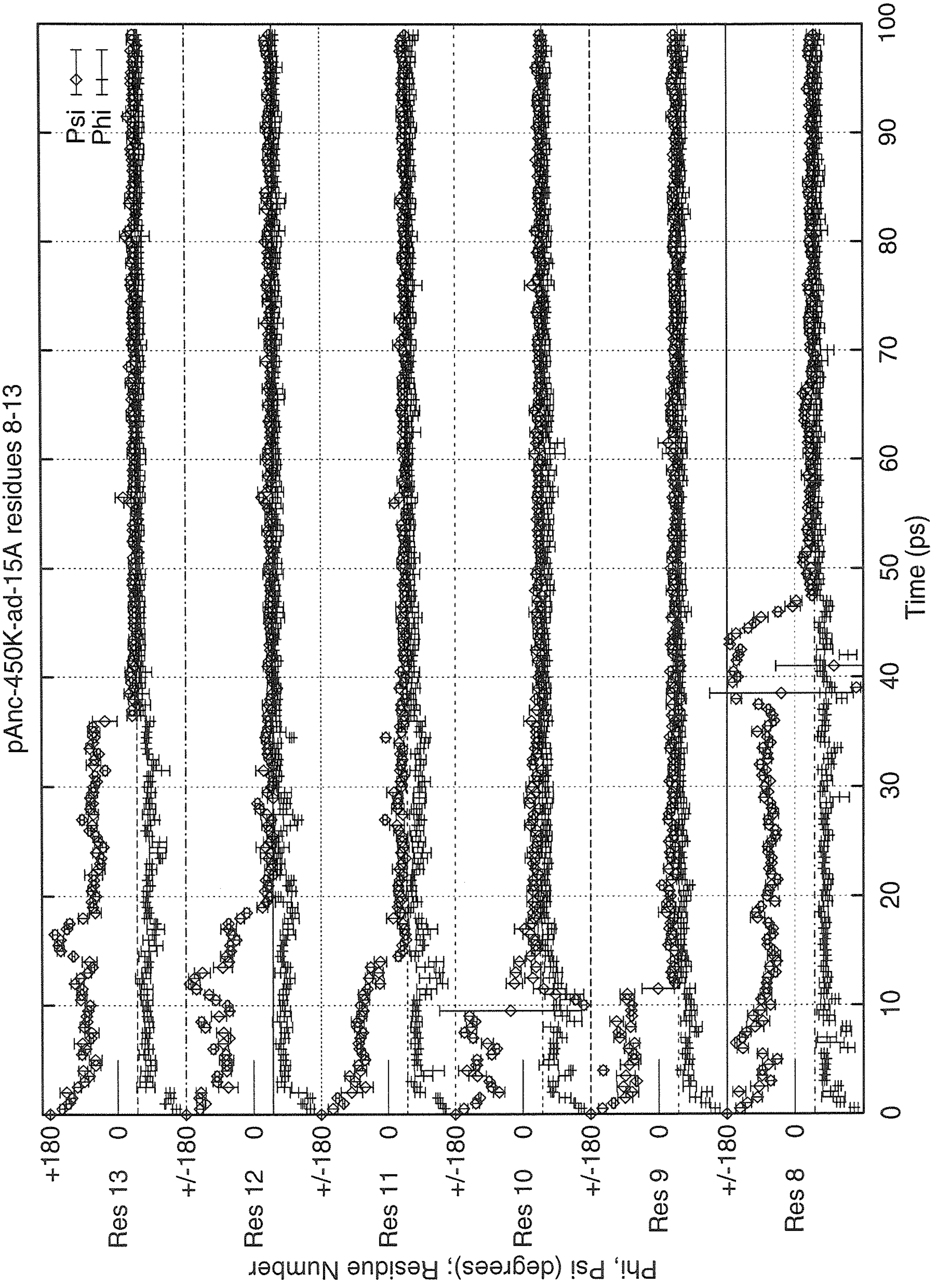


Figure 2.3B: Caption p. 47.

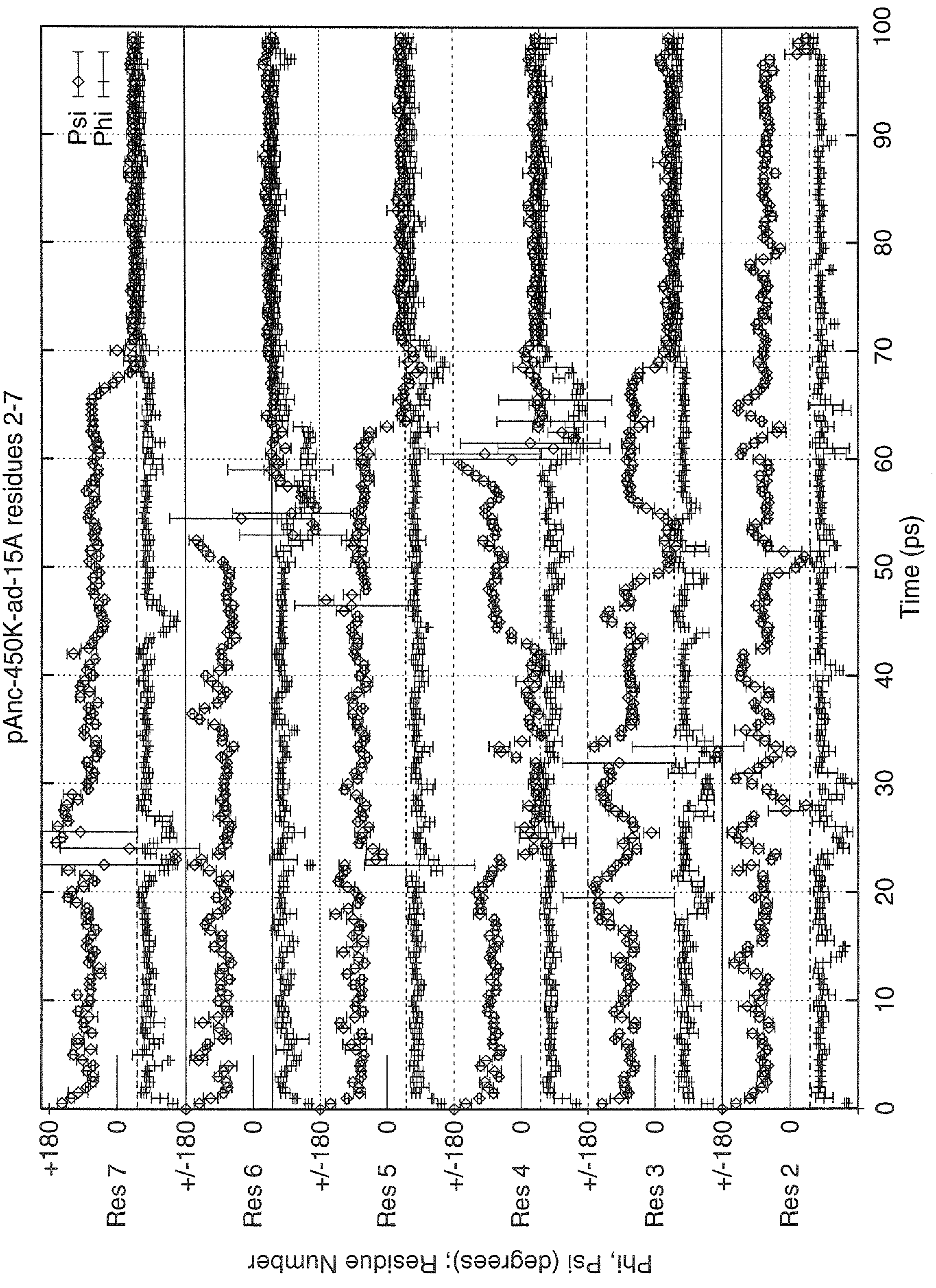


Figure 2.3C: Caption p. 47.

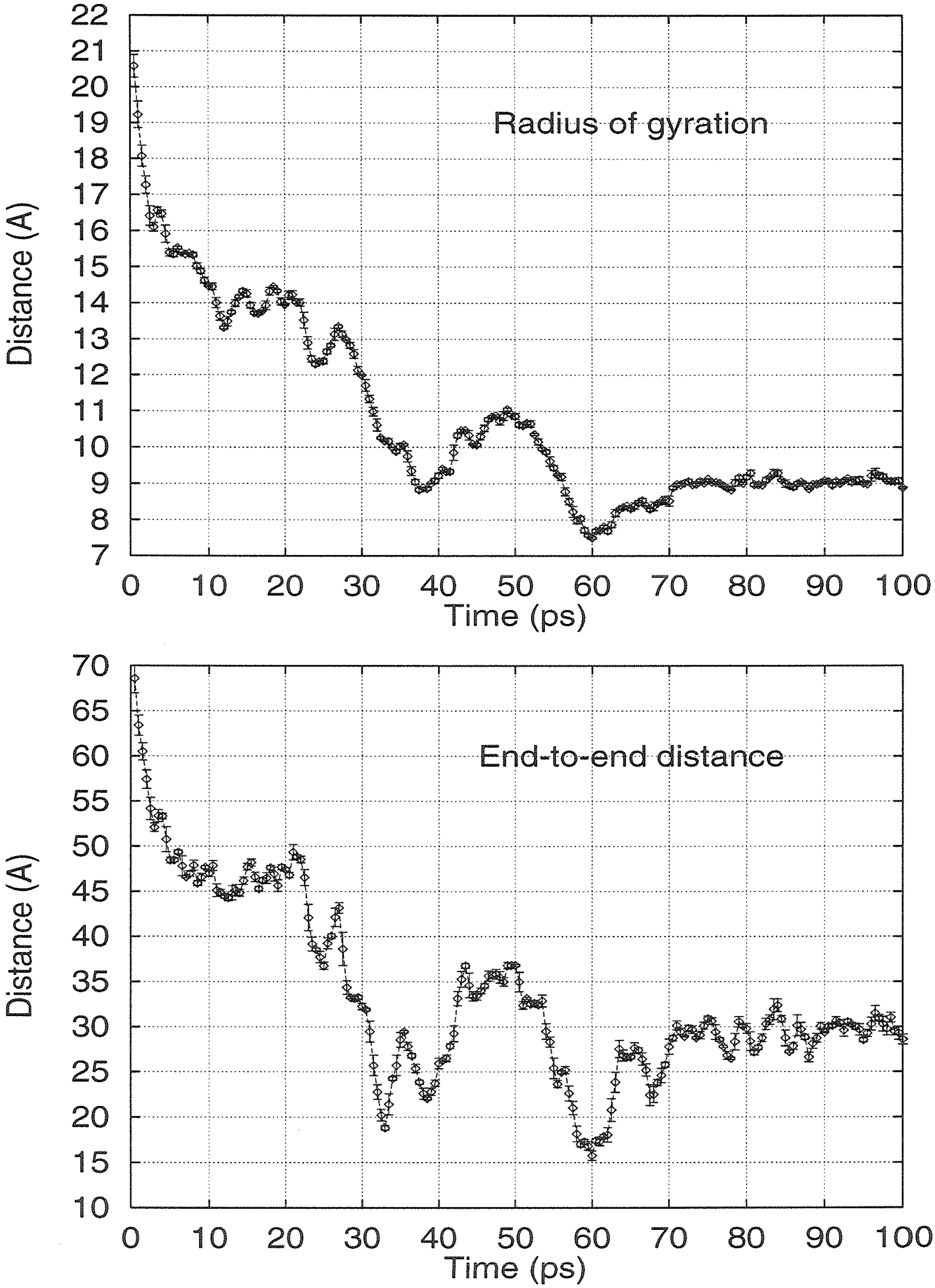


Figure 2.4: The radius of gyration and end-to-end distance during helix formation for a fast-folding trajectory.

moves from a relatively extended position to being curled up slightly, like the end of walking cane. In both plots the drop in distance from ~22 ps to 38 ps represents a gradual contraction of the polymer as both ends fold up on each other and the central helix forms. From ~38 ps to ~50 ps, the polymer visibly expands. From 50 ps to 60 ps the N-terminal half of the protein bends and pulls up close to the center of the polymer. From ~60 ps to 70 ps, the N-terminus extends, but the helix propagates, thereby attenuating the increase in distance. This attenuation is especially clear in the plot of radius of gyration.

The radius of gyration more smoothly portrays the process of helix formation than the end-to-end distance does. For example, the end-to-end distance, but not the radius of gyration, shows a dramatic drop at ~32 ps because the C-terminus moves from a relatively extended position to being curled up, like the end of a walking cane. The end-to-end distance is, by definition, biased towards the positions of only two atoms, the terminal ones, and this makes its plot jagged. In contrast, the radius of gyration averages out differences in atomic positions by squaring their displacement from a common center:

$$R_g = \sqrt{\frac{\sum_{allatoms} M_i(x_i - x_{cm})^2 + M_i(y_i - y_{cm})^2 + M_i(z_i - z_{cm})^2}{M_{total}}} \qquad (2.1)$$

where M_i is the mass of the ith atom, x_{cm} is the x coordinate of the center of mass, and M_{total} is the total mass of the polymer. The radius of gyration of the helix equals the radius of gyration of a rigid rod of that length, indicating that the residues are evenly spaced along the helix.[2]

Ramachandran plots. Another useful measure of the structure of a protein polymer is a Ramachandran plot because many protein structures, including α-helices, have characteristic backbone dihedral, (ϕ, ψ), angles. A variety of Ramachandran

[2] The change in radius of gyration was calculated using the moment of inertia of the polymer, and this matches the radius of gyration calculated assuming all masses are one. The equality is expected since the chemical and isotopic composition of each simulated monomer is constant, and the sequence is symmetric about the center.

plots are shown on pp. 54–55. First, Figure 2.5 is a series of graphs displaying the ϕ, ψ angles of every residue at a single time point from 0 ps to 100 ps. One easily sees that the residues start in the first graph at $(\phi, \psi) = (-180°, +180°)$, move in the succeeding graphs to the upper left-hand quadrant, and then coalesce into an α-helix by 100 ps.

Figure 2.5: Figure on p. 54. Ramachandran plots every 20 ps of a fast helix-forming trajectory. The (ϕ, ψ) angles of every residue at a single time point from 0 ps to 100 ps. The residues start in the first graph at $(\phi, \psi) = (-180°, +180°)$, move in the succeeding graphs to the upper left-hand quadrant, and then coalesce into an α-helix by 100 ps.

Second, plotting the trajectories each residue makes in ϕ, ψ space uncovers a local minimum. Figure 2.6 graphs the dihedral angles of residues 5, 9, 13, and 17. The dihedral angles of each residue start the simulation at (-180°, +180°). They immediately relax to a well around (-80°, +70°) containing structures with stable $i, i+2$ HBs. The polymer rattles around in this region, which is called the C7 region (Avignon *et al.*, 1969; Bystrov *et al.*, 1969), during the first 10 ps of the simulation. After ~12 ps for residue 9, ~32 ps for residue 17, ~36 ps for residue 13, ~63 ps for residue 5, the residues reoriented and jumped to the lower well at (~-60°, ~-40°) characterizing helical structures. (See Chapter 4, Section 4.2.1 and Chapter 5, Section 5.1.1, for further discussion of this local minimum.)

2.2 Why the α-helix and the C7 structure form: The energetics of helix formation

Because energetics often drives structural transitions, it is useful to determine the relative energies of the pertinent structures. Figure 2.7 graphs the energies of the different terms in the energy expression during the formation of the helix. Because the energies of the $(Ala)_{20}$ drop precipitously in the first few ps, the C7 conformation

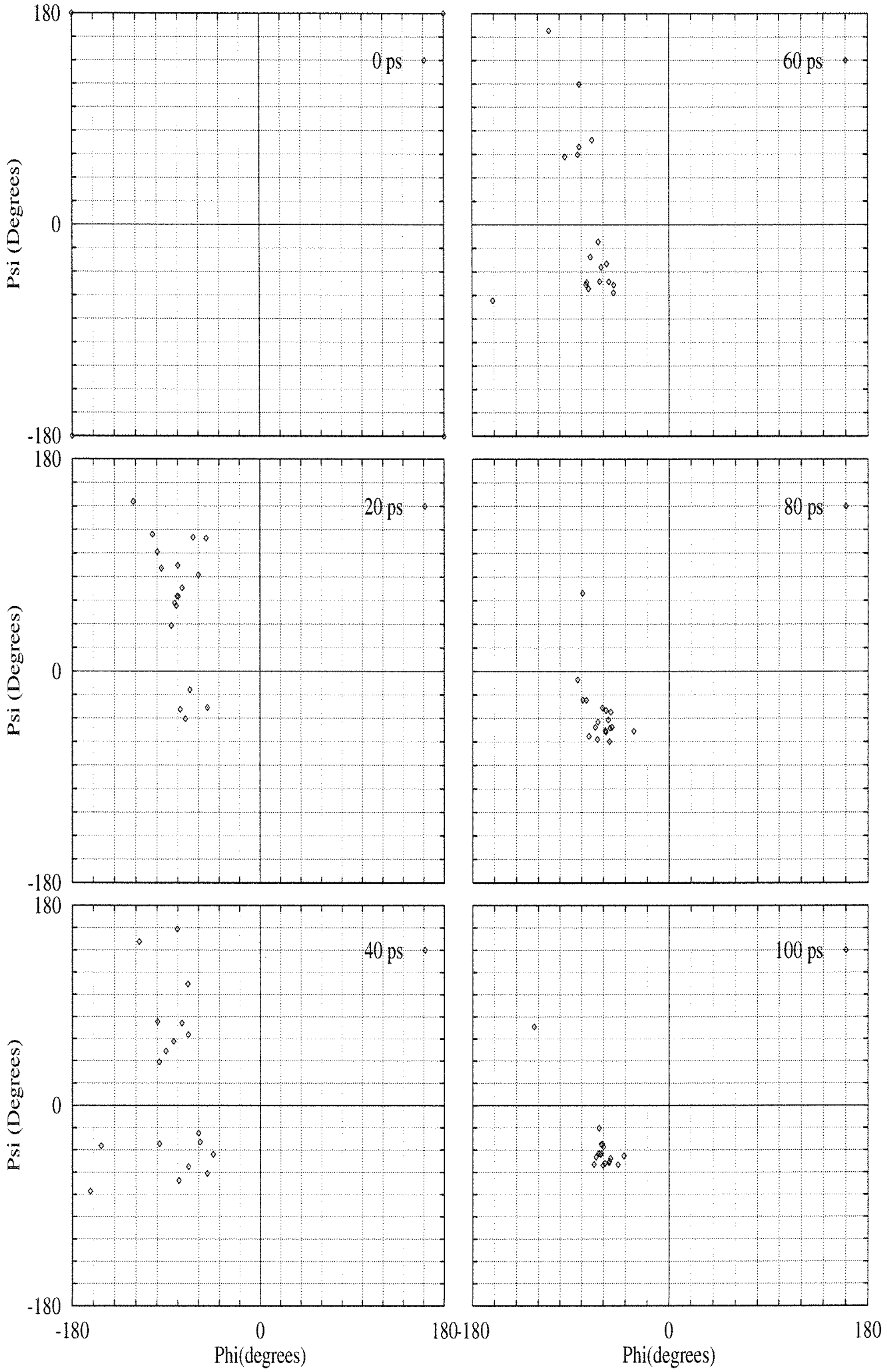


Figure 2.5: Caption on p. 53.

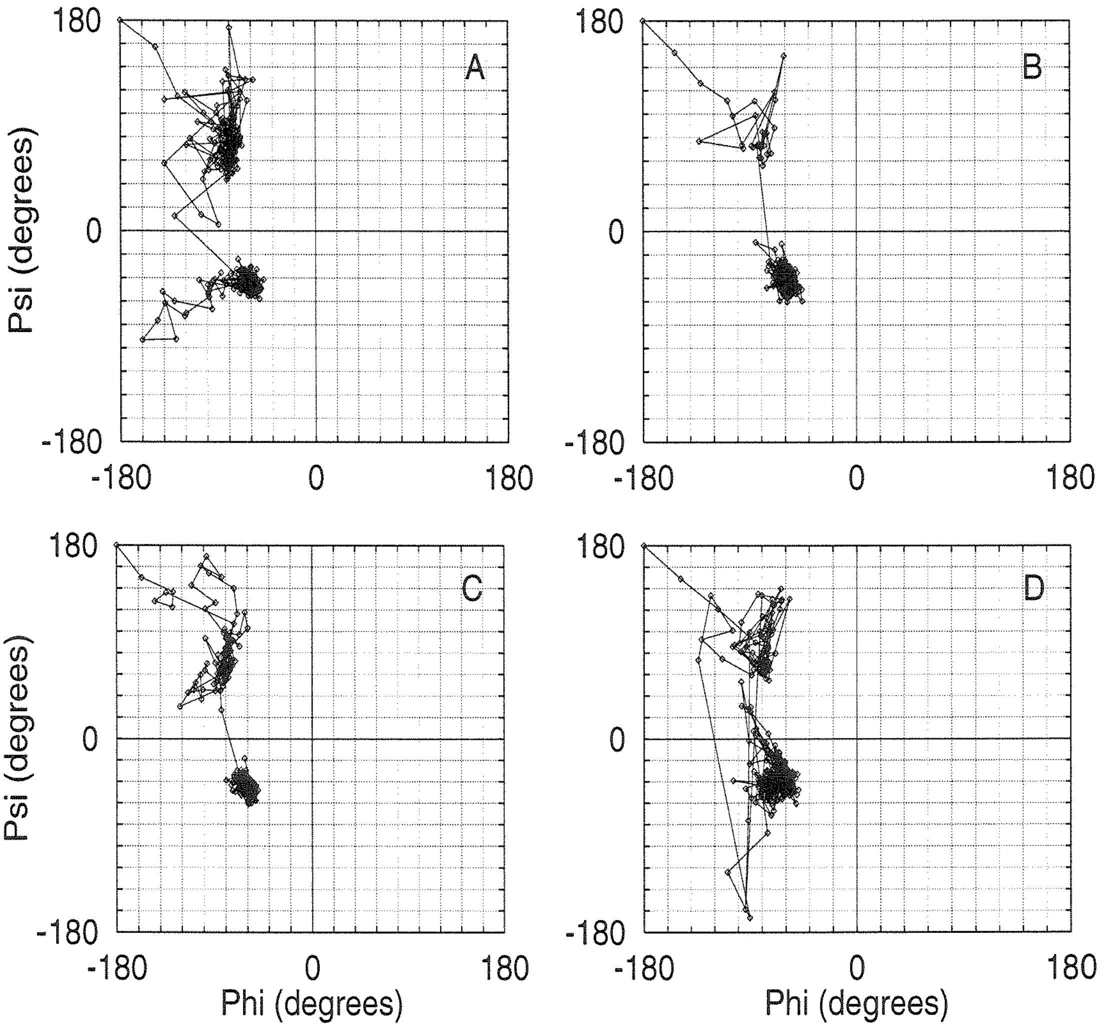


Figure 2.6: The trajectory of the ϕ, ψ dihedral angles of four residues during this fast, helix-forming trajectory. Graphs A, B, C, and D show residues 5, 9, 13, and 17, respectively. These residues clearly illustrate the existence of two energetic minima separated by a large barrier. The region at ($\sim$-60°, $\sim$-40°) characterizes α-helices. The region at ($\sim$-80°, $\sim$+70°) characterizes structures with strong, $i, i+2$ hydrogen bonds. (See Chapter 5, Section 5.1.1.)

is clearly more energetically favorable than the initial, $(\phi, \psi) \approx (-180°, +180°)$ conformation. The other two major drops occur at around 45 and 70 ps, corresponding to α-helix formation at the C-terminal third and at the N-terminal third, respectively. Thus, $(\mathrm{Ala})_{20}$ forms an α-helix because it is energetically favorable to do so.

In Figure 2.7a, the valence energy is relatively constant because the NEIMO al-

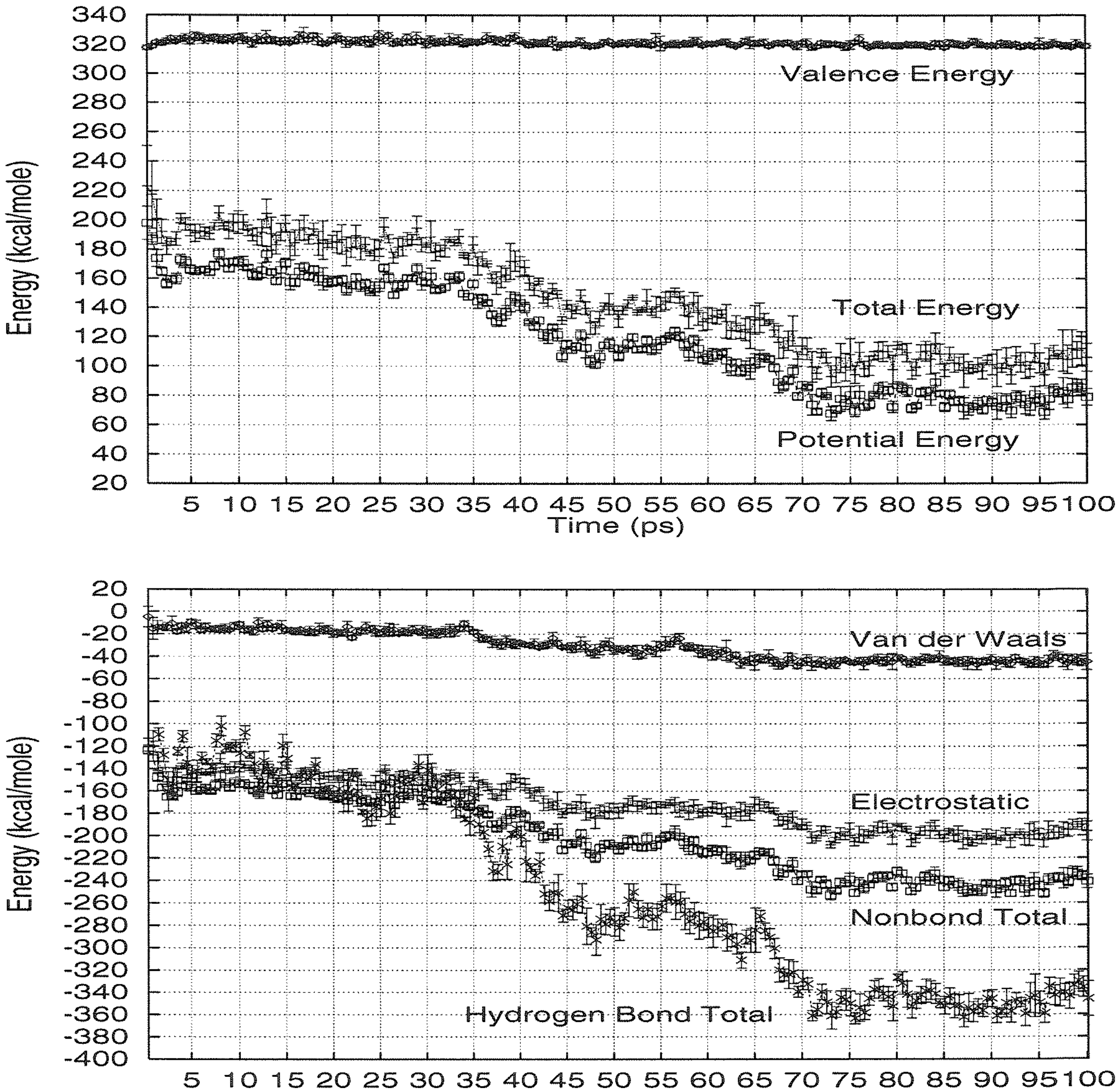


Figure 2.7: Energies of different terms in the AMBER energy expression during a fast helix-forming trajectory. Top (a): Total energy, valence, and potential energy. Bottom (b): Nonbond energy and its components van der Waals, electrostatic, and total hydrogen bonding energy. The total energy in hydrogen bonds is calculated from the hydrogen bonding term in the AMBER force field (a 12-10 potential. See Equation 2.2). Only backbone amide nitrogens and backbone carbonyl oxygens separated in sequence by at least one intervening residue were considered in the hydrogen bonding term. And, only hydrogen bonds stronger than -2 kcal/mole were totalled. Points are averaged 1 point graphed per 0.5 ps simulated or per 5 data points recorded.

gorithm (Jain *et al.*, 1993) freezes most of the components of the valence energy, i.e., NEIMO freezes bond angles and bond lengths.[3] Only the torsional component varies. The difference in total energy and the potential energy appear to differ roughly by a constant, indicating that the total energy is dominated by potential energy, not kinetic energy. In fact, the kinetic energy adopts a Boltzmann distribution about the thermal energy at 450K, as the NEIMO-Hoover algorithm dictates. The hydrogen bonded term in AMBER is a 12-10 potential applied to the proton and oxygen in each hydrogen bond (Weiner *et al.*, 1984).[4]

2.3 The events during the nucleation of a fast-folding simulation

Helix-coil transition theory hypothesizes that the rate-limiting event in helix formation is nucleation, i.e., the formation of an initial $i, i+4$ HB that should then seed the conversion of the rest of the polymer into a helix. In this thesis, I will define nucleation as the formation of the first α-helical hydrogen bond that persists long enough to become part of a contiguous chain of α-helical HBs. In this trajectory, nucleation occurs when an $i, i+2$ HB converts to an $i, i+4$ HB. As Figure 2.8 illustrates on pp. 58–61, the carboxy oxygen of residue eight is strongly hydrogen bonded to the amino nitrogen of residue 10 at 11 ps. Over the course of the next 4 ps, the carbonyl oxygen of residue 8 walks itself up to residue 12 to make the first α-helical, $i, i+4$ HB. At 12 ps it has formed a weak bond to residue 11, by 14 ps both residues 8 and 9 have $i, i+3$ HBs, and by 15 ps residue 8 is stably hydrogen bonded to residue 12.

[3] Valence energy refers to the energy in bond stretches, bends, motion along torsional angles, and inversions.

[4] In other words,

$$E_{\mathrm{HBs}} = \sum_{\mathrm{HB}\ i,j} \left(\frac{C_{ij}}{R_{ij}^{12}} - \frac{D_{ij}}{R_{ij}^{10}} \right) \tag{2.2}$$

where R_{ij} is the distance between the ith and jth atoms in the HB, and C and D are constants. The AMBER force field includes Coulombic terms for all nonneutral atoms, irregardless of whether they form HBs.

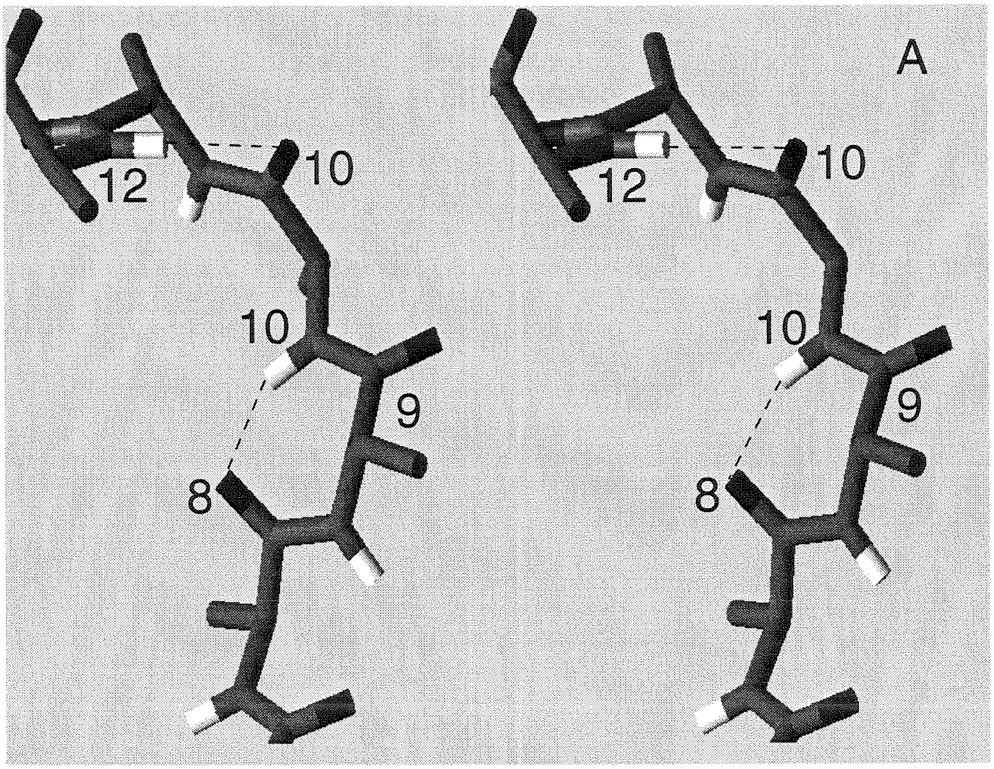


Figure 2.8: Nucleation via $i, i+2$ expansion. A. At 10.7 ps, the oxygen of residue 8 has an $i, i+2$ HB to the amide proton of residue 10, and residue 10 has an $i, i+2$ HB to residue 12. B & C on p. 60. B. At 12.4 ps, residue 8 forms an $i, i+3$ HB to residue 11. C. From 12 ps to 14 ps, residue 8 forms an $i, i+3$ HB to residue 11. By 13.8 ps (pictured), residue 9 makes an $i, i+3$ HB to residue 12. D & E on p. 61. D. At 14.2 ps, residue 12 has shifted to form HBs with both residues 8 and 9. E. Last, at 14.9 ps, the amide proton of residue 12 shifts its HB from residue 9 to residue 8, forming the first $i, i+4$ HB.

Figures 2.9 and 2.10 describe the nucleation event in more detail. They graph the hydrogen bond energies of residues 8 and 9 to residues 10, 11, and 12 and to residues 11, 12, and 13, respectively. In the picoseconds before nucleation, i.e., ~9–11 ps, the carbonyl of residue 8 forms only one HB, one to residue 10. It is ~-10 kcal/mole. At about 11 ps, the $i, i+3$ and $i, i+4$ strengthen at the cost of the $i, i+2$ bonds. By 14.0 ps, the $i, i+4$ bond is more than one standard deviation stronger than the $i, i+3$ bond. By 38 ps, the α-helical hydrogen bond is about -12 kcal/mole, according to the AMBER force field.

2.4 The types of hydrogen bonds

Figure 2.11 illustrates the most useful type of analysis for predicting whether or not a particular simulation will succeed in folding a $(\mathrm{Ala})_{20}$ into an α-helix.[5] The different types and number of HBs in this trajectory are graphed with respect to time. Initially, only $i, i+2$ HBs exist. At 12 ps the number of α-helical, $i, i+4$ HBs increases.

This simulation does not form any nonlocal HBs. Nonlocal HBs are defined as $i, i+6$ to $i, i+19$ and $i, i-2$ to $i, i-19$ HBs. Essentially, nonlocal HBs are HBs between residues separated by more than four residues. Fast helix-forming trajectories generally do not form nonlocal HBs, because they trap the polymer in nonhelical conformations, as will be shown in Chapter 4, Section 4.2.4.

[5] Although all 155 simulations on polyalanine were run starting from the same configuration of $(\mathrm{Ala})_{20}$, at the same temperature, and with the same force field and charges, each simulation was different because each $(\mathrm{Ala})_{20}$ had a different set of random initial velocities. See Chapter 6, Methods of Simulation, Section 6.6.

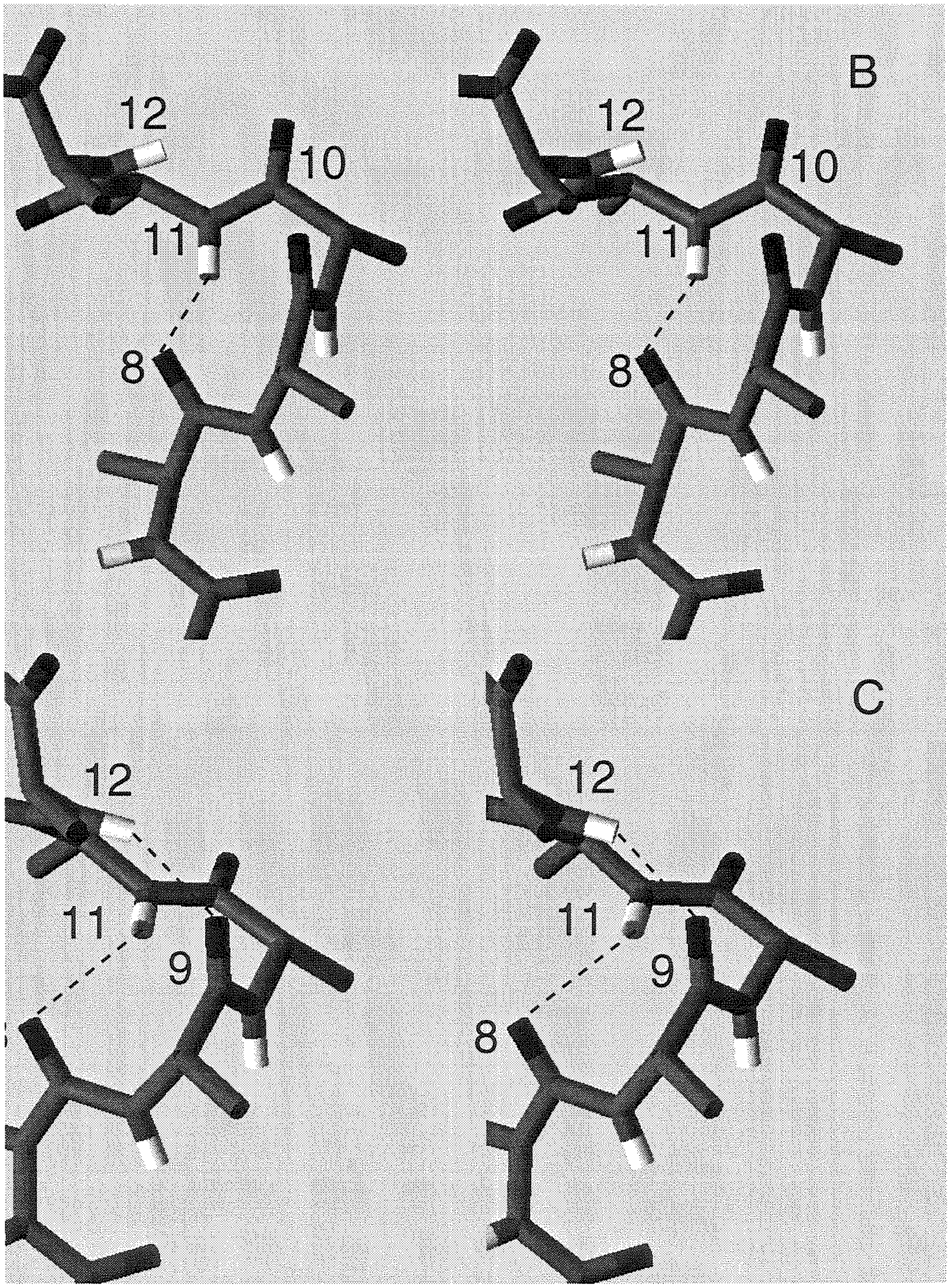


Figure 2.8B: Caption on p. 58.

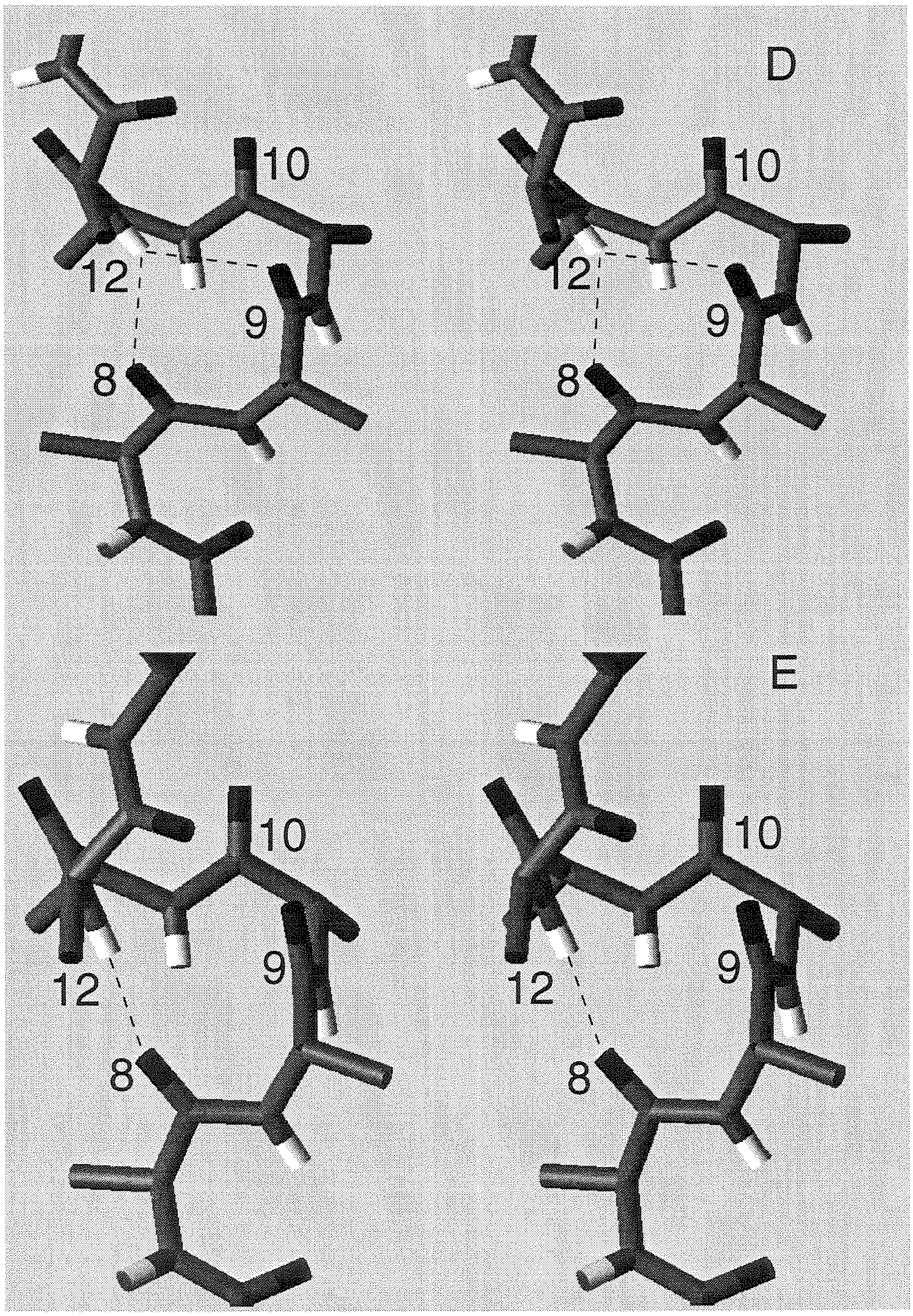


Figure 2.8C: Caption on p. 58.

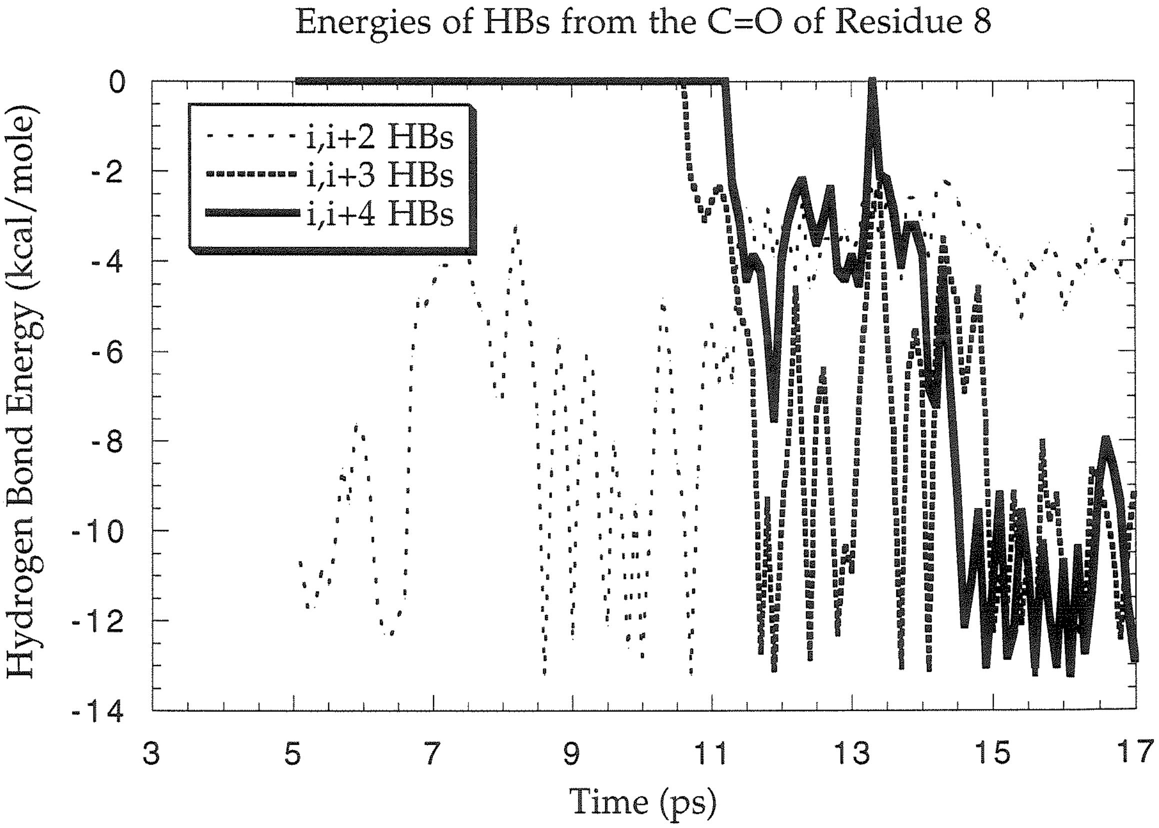


Figure 2.9: The HB energies of the carbonyl oxygen of residue 8 during nucleation. The $i, i+2$ HB is from residue 8 to 10; $i, i+3$ to 11, and the $i, i+4$ HB to residue 12.

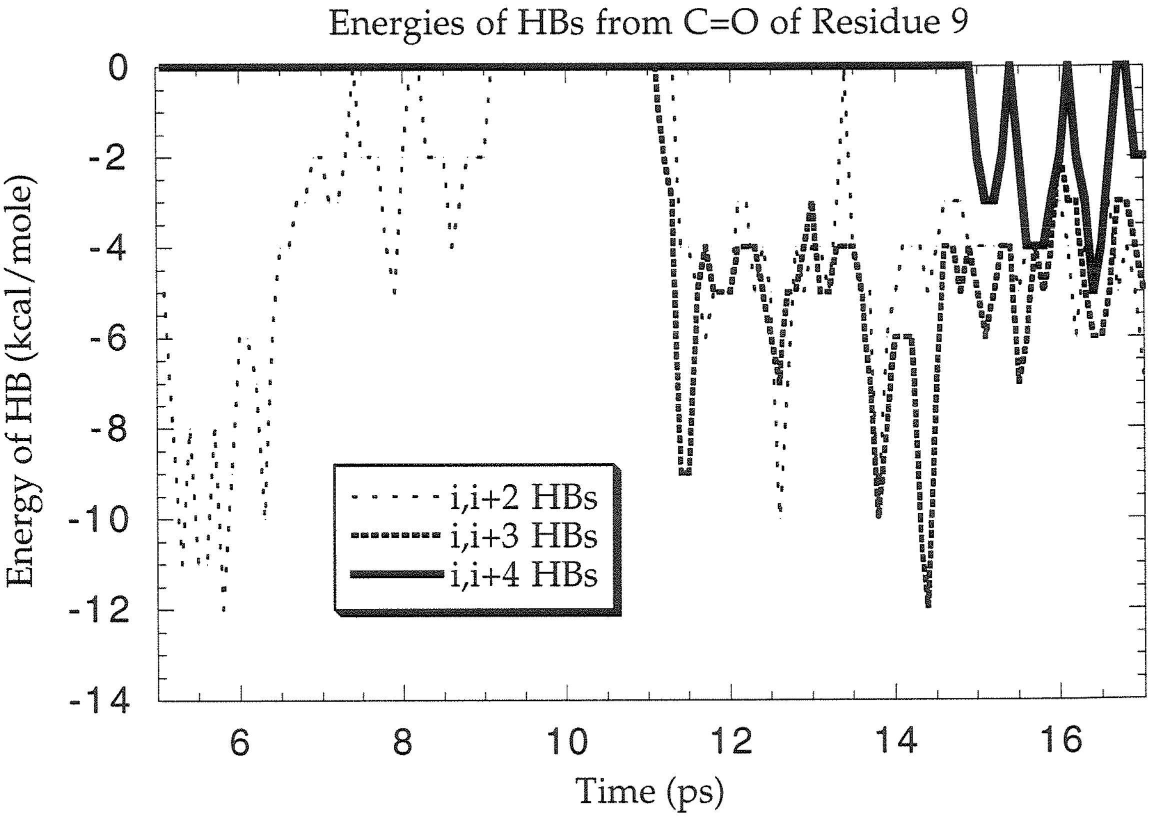


Figure 2.10: The HB energies of the carbonyl oxygen of residue 9 during nucleation. The $i, i+2$ HB is from residue 9 to 11; $i, i+3$ to 12, and the $i, i+4$ HB to residue 13.

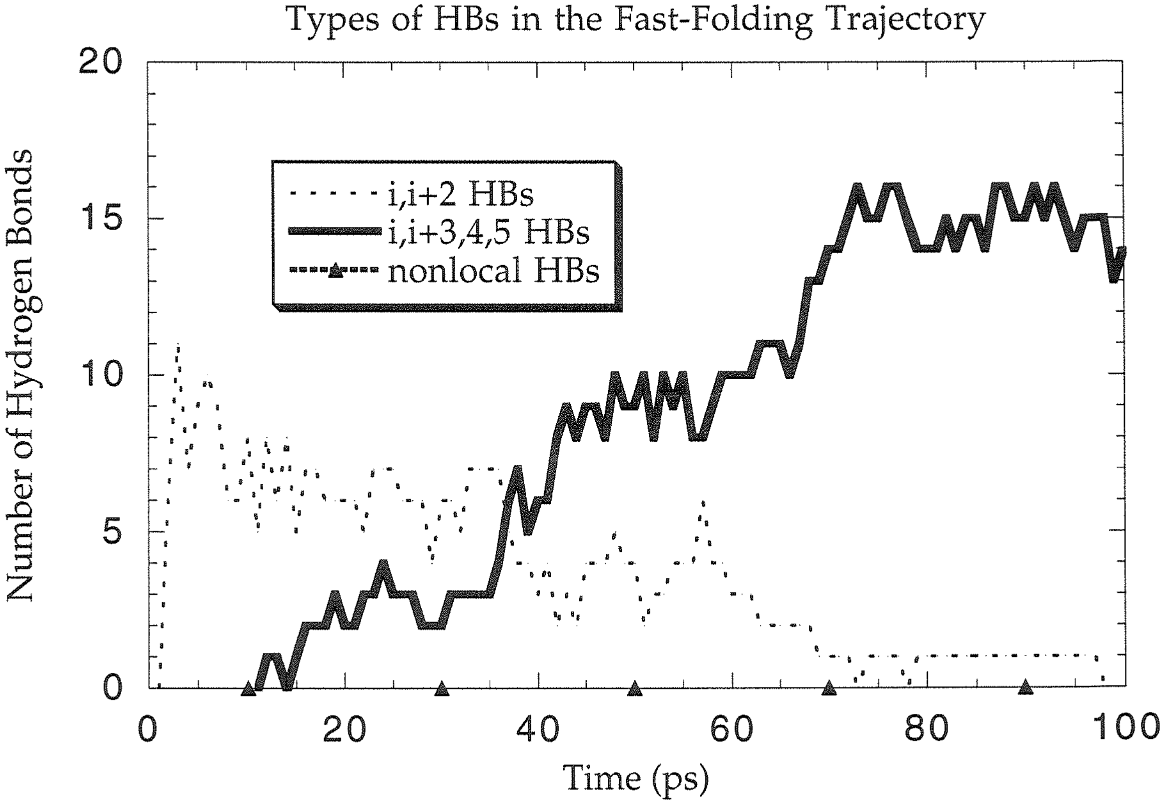


Figure 2.11: Numbers and types of HBs during a fast, helix-forming trajectory. The number of $i, i + 2$ HBs is graphed in thin, dashed lines; the number of $i, i + 4$ HBs in solid, thick, black lines; and the number of nonlocal HBs in thick, dashed lines with solid triangles superimposed. The triangles at $y = 0$ signify that the number of nonlocal HBs is 0. Nonlocal HBs are defined as $i, i + 6$ to $i, i + 19$ and $i, i - 2$ to $i, i - 19$ HBs. The total number of HBs and the total number of local HBs ($i, i + 3$ to $i, i + 5$ HBs) are equivalent in this trajectory because it has no nonlocal HBs. Points are averaged one point plotted per ten points recorded, i.e., one point plotted per 1.0 ps.

Chapter 3 Example of a Slow Helix-forming Trajectory

Not all of the 129 helix-forming simulations of polyalanine folded as quickly as the example in Chapter 2. Let us look at an example of a how helix-formation can be retarded. This chapter describes a slow helix-forming trajectory of $(\mathrm{Ala})_{20}$.[1] To compare the fast and slow helix-forming simulations, this chapter will analyze the slow helix-forming simulation with the same tools used on the fast helix-former in Chapter 2.

3.1 Outline of the sequence of structural changes during helix folding

3.1.1 The trajectory

Figure 3.1 illustrates the trajectory. As can be seen, the polypeptide did not form an α-helix quickly. Folding took 170 ps compared to the 80–100 ps for the simulation in Chapter 2. The slow helix-forming trajectory was retarded because, instead of winding up immediately into an α-helix, the polymer formed nonlocal HBs that constrained it in a nonhelical blob. Nonlocal HBs are defined here as all nonhelical, non-$i, i+2$ HBs. In more rigorous terms, nonlocal HBs are HBs between the carbonyl oxygen of residue i and the amide proton of any residue in the following ranges: $i+6$ to $i+20$ or $i-2$ to $i-20$. Nonlocal HBs are seen in Figure 3.1 from 38 to 123 ps, inclusive. By 138 ps, all of the nonlocal HBs had broken, and the polymer readily wound up into an α-helix.

[1] This is the trajectory labeled pAnc-450K-ag-15A in Appendix D.

Figure 3.1: The folding trajectory of a slowly folding polyalanine: The formation of nonlocal HBs retards helix formation. Folding is delayed for 100 ps until the polymer can break the nonlocal HBs trapping it in a globular conformation. The amino termini are at the bottom of the figure. A. The initial conformation has $(\phi, \psi) = (180°, 180°)$. B. At 2 ps, each residue hydrogen bonds to its neighbor two residues away ($i, i+2$ HBs). Three such HBs are illustrated with dashed lines. C. 18 ps. D. At 38 ps, the two halves of the polymer are stuck together by four nonlocal HBs. E. At 65 ps, the four nonlocal HBs continue to trap the polymer in a ball. However, the carboxy terminus has formed a short helical segment. F. At 123 ps, all but one of the nonlocal HBs have been replaced with local HBs. G. At 138 ps, the polymer has elongated and formed some helical segments. H. At 151 ps, residues 11-20 are in an α-helix; the amino terminal loop has not propagated yet. I. At 170 ps the helix is complete. All figures in this chapter are derived from the trajectory labeled pAnc-450K-ag-15A in Appendix D.

3.1.2 The location and sequence of nucleation and propagation

Figure 3.2 illustrates the locations and times of nucleation and propagation more clearly than snapshots of the animated trajectory can. For example, Figure 3.2 draws attention to an early section of α-helix containing two $i, i+4$ HBs at the C-terminus at 27 ps. A similarly isolated $i, i+4$ HB forms in the central portion of $(\text{Ala})_{20}$ at about 26 ps. As Figure 3.2 demonstrates, residues flit in and out of α-helical configurations for the next 100 ps. But, by 126 ps, the polymer has broken all its nonlocal HBs, and segments of residues in helical orientations can then start to propagate along the length of the polymer. According to Figure 3.2, the central helix starts consolidating into a stable region of α-helix almost immediately, and the C-terminal residues consolidate 15 ps later at ~141 ps. Thirteen picoseconds later, at ~154 ps, a stable helix forms from residues in the N-terminus that had been flickering in and out of α-helical configurations starting from 95 ps.

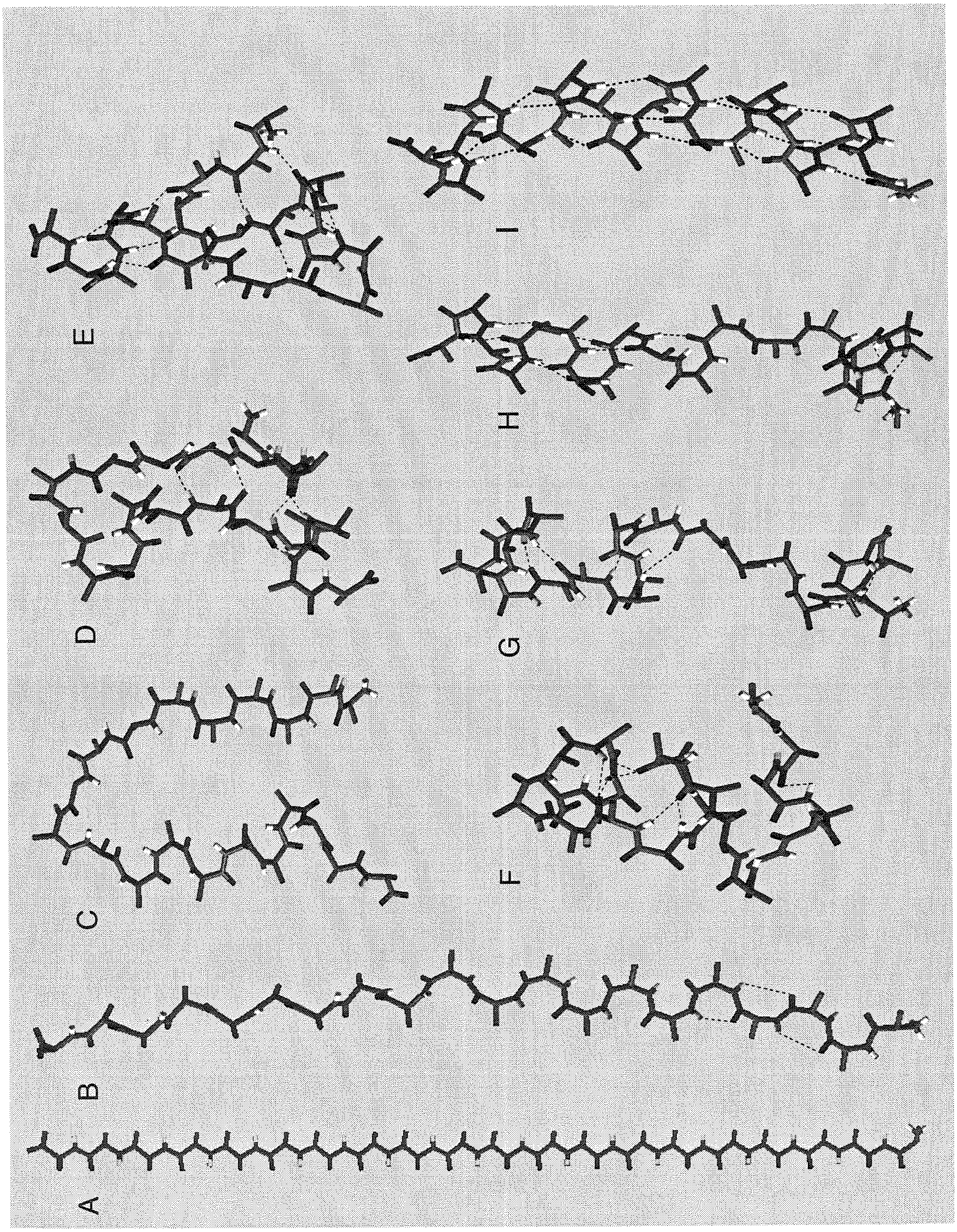


Figure 3.1. Caption p. 66.

Figure 3.2: Pages 68–72: Nucleation and propagation for a slow helix-forming trajectory. At each time point, each residue is symbolized as an "h," a slash ("/"), or a period ("."), depending on the (ϕ, ψ) dihedrals. "h"s represent α-helical residues, i.e., those within a 30° radius of the classic α-helical dihedral angles, $(\phi, \psi) = (-57°, -47°)$. The slash ("/") denotes a residue within a largely helical region of Ramachandran space, a region populated by residues in good quality X-ray structures (region A of PROCHECK (Laskowski *et al.*, 1993). See Figures 4.2 and 5.1.). Each residue symbolized by a period (".") is in a coiled, nonhelical structure at any other region of Ramachandran space.

```
# Residues classified by structure: helix or coil
# From tor file  pAnc-450K-ag-15A.tor
# from residues  2 to  18
# from      0.00 ps to    171 ps averaging every    10 points
# residues 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8
     1.00  . . . . . . . . . . . . . . . . .
     2.00  . . . . . . . . . . . . . . . . .
     3.00  . . . . . . . . . . . . . . . . .
     4.00  . . . . . . . . . . . . . . . . .
     5.00  . . . . . . . . . . . . . . . . .
     6.00  . . . . . . . . . . . . . . . . .
     7.00  . . . . . . . . . . . . . . . . .
     8.00  . . . . . . . . . . . . . . . . .
     9.00  . . . . . . . . . . . . . . . . .
    10.00  . . . . . . . . . . . . . . . . .
    11.00  . . . . . . . . . . . . . . . . .
    12.00  . . . . . . . . . . . . . . . . .
    13.00  . . . . . . . . . . . . . . . . .
    14.00  . . . . . . . . . . . . . . . . .
    15.00  . . . . . . . . . . . . . . . . .
    16.00  . . . . . . . . . . . . . . . . .
    17.00  . . . . . . . . . . . . . . . h .
    18.00  . . . . . . . . . . . . . . . . .
    19.00  . . . . . . . . . . . . . . . h /
    20.00  . . . . . . . . . . . . . . . . .
    21.00  . . . . . . . . . . / . . . . h .
    22.00  . . . . . . . . . . . . . . h h .
    23.00  . . . . . . h . . . . . . . h h h
    24.00  . . . . . . h . . . . . . . h h h
    25.00  . . . . . . h . h . . . . . h h h
    26.00  . . . . . . h . h . h . . . h h h
    27.00  . . . . . . h . h / / . . . h h .
```

```
28.00   . . . . . . h . h . / . . . h h h
29.00   . . . . . . . . h . . . . . h h h
30.00   . . . . . . h . h / / . . . h / h
31.00   h . . . . . . . h . . . . . h / h
32.00   . . . . . . . . h / / . . . h h h
33.00   h . . . . . . . h h h . . . h h h
34.00   h . . . . . . . . / h . . . h h h
35.00   . . . . . . . . . / h . . . h h h
36.00   h . . . . . . . . h h . . . / h h
37.00   h . . . . . . . . / h . . . h h h
38.00   h . . . . . . . h h h . . . h h h
39.00   h . . . . . . . . . h . . . h h h
40.00   h . . . . . . . h h h . . . h h h
41.00   h . . . . . . . . / / . . . h h h
42.00   h . . . . . . . h . / . . . h h h
43.00   h . . . . . . . h / h . . . h h h
44.00   h . . . . . h . h . / . . . h h h
45.00   . . . . . . h . h / / . . . h h h
46.00   . . . . . . h . h / h . . . h h h
47.00   . . . . . . h . h . / . . . h h h
48.00   . . . . . . h . h / / . . . h h h
49.00   . . . . . . h . h . . . . . h h h
50.00   . . . . . . h . h / h . . . h h h
51.00   . . . . . . h . h . h . . . h h h
52.00   . . . . . . h . h . h . . . h h h
53.00   . . / . . . h . h . . . . . h h .
54.00   . . / . . . h . h h . . . . h h h
55.00   . . . . . . h . h h . . . . h h h
56.00   . . . . . . h . h . / . . . h h h
57.00   . . . . . . h . . . / . . . h h h
58.00   . . . . . . . . . h h . . . h h h
59.00   . . . . . . . . / . h . . . h h h
60.00   . . / . . . . . . . h . . . h h h
61.00   . . . . . . . . . . . . . . h h h
62.00   . . h . . . . . h . . . . . . h /
63.00   . / . . . . . . . . . . . . . h /
64.00   . . . . . . / . h . . . . . . h h
65.00   . . . . . . . . h . . . . . h h h
66.00   . . . . . . h . h / . . . . h h h
67.00   . . . . . . h . h . / . . . h h h
68.00   . . . . . . . . h . . . . . h h .
69.00   . . . . . . h . h h . . . . h h .
70.00   . . . . . . h . h / . . . . h h .
```

```
 71.00   . . . . . . h . h / . . . . / / .
 72.00   . . . . . . h . h . . . . . / . .
 73.00   . . . . . . h . h / . . . . h . .
 74.00   . . . . . . . . . . h . . . h h h
 75.00   . . . . . . . . h . . . . . h h h
 76.00   . . . . . . h . h . h . . . h . .
 77.00   . . . . . . . . h h . . . . . . .
 78.00   . . . . . . h . h / . . . . . . .
 79.00   . . . . . . h . h / . . . . . . .
 80.00   . . . . . . h . h / h . . . h . .
 81.00   . . . . . . h . h / . . . . h . .
 82.00   . . h . . . h . h / . . . . h . .
 83.00   . . . . . . h . . . . . . . h . .
 84.00   . . . . . . . . h / . . . . h . .
 85.00   . . . . . . . . h h . . . . . . .
 86.00   . . . . . . . . h h . . . . . . .
 87.00   . . . . . . . . . . . . . . . . .
 88.00   . . . . . . . . . . h . . . . . .
 89.00   . . . . . . . . . . . . . . . . .
 90.00   . . . . . . . . . . . . . . . . .
 91.00   . . . . . . . . h / h . . . / . .
 92.00   . . . . . . . . h / h . . . . . .
 93.00   . . . . . . . . h / h . . . h . .
 94.00   . . . . . . . . h / h . . . h . .
 95.00   . . h . . . . . . . h . . . . . .
 96.00   . . / . . . . . . . h . . . h . .
 97.00   . . . . . . . . h h h . . . h . .
 98.00   . h . . . . . . . . h . . . h . .
 99.00   . . / . . . . . . . . . . . . . .
100.00   . . . . . . . . h / h . . . . . .
101.00   . . . . . . . . h . h . . . . . h
102.00   . h . . . . . . . . . . . . . . h
103.00   . h . . . . . . h h / . . h . . h
104.00   . . h . . . . . h h h . . h . . h
105.00   . h h . . . . . h / h . . h / . h
106.00   . . h . . . . . h / . . . h . . h
107.00   . . . . . . . . h . h . . h . . h
108.00   . . h . . . . . h / h . . h . . h
109.00   . . h . . . / . . . h . . h . . h
110.01   . . h . . . . . . . h . . h / . h
111.01   . . . . . . . . . . h . . h . . h
112.01   . h h . . . . . . . h . . h . . h
113.01   . . . . . . . . h . h . . . . . h
```

```
114.01   . . . . . . . . h / / . . h / . h
115.01   . . . . . . h . h / . . . h / . h
116.01   . . h . . . h . h / / . . h . . .
117.01   . . . . . h . . h . / . . / . . h
118.01   . . . . . . h . h . . h . h . . h
119.01   . . . . . . . . h . . h . h . . h
120.01   . . . . . . / . h . . h . h . . h
121.01   . . . . . . . . h . . h . h . . h
122.01   . . . . . . h . . . . h . h . . h
123.01   . h h . . . . . / . . h . h . . h
124.01   . h / . . . . . . . . h . h . . h
125.01   . h . . . . . . h . h h . h . . h
126.01   . h . . . . h . h / h h . h . . h
127.01   . h h . h . . . h h h h . h . . h
128.01   . h . . . / . . h h h h . h . . .
129.01   . . . . . h . . h h h h . / . . .
130.01   . . . . . h . . h h h h . h . . .
131.01   h . . . . h . . h h h h . h . . .
132.01   / . . . . h . . h h h h . h / . .
133.01   h h . h . h . . h h h h . h . . /
134.01   h . . h . h . . h h h h . h . . h
135.01   h . . h . h . . h h h h . h / . h
136.00   h . . h . h . . h h h h . . . . h
137.00   h . . h . . . . h h h h . h . . h
138.00   h . . h . . . . h h h . . h / h .
139.00   h h . h . . . . h h h . . h . . .
140.00   h . . h . . . . h h h h . h h h .
141.00   h . . h . . . . h h h h . h h h .
142.00   . / . h . . . . h h h . . h h . h
143.00   . . . h . . . . h h h h h h h h h
144.00   . . . h . . . . h h h h h h h h h
145.00   h . . h . . . . h h h h h h h h h
146.00   h / . h . . . . h h h h h h h h h
147.00   h . . h . . . . h h h h h h h h h
148.00   h . . h . . . . h h h h h h h h h
149.00   h . . h . . . . h h h h h h h h h
150.00   h . . h . . . . h h h h h h h h h
151.00   h . . h . . . . h h h h h h h h h
152.00   . . . / . . . . h h h h h h h h h
153.00   . / . . . . . . h h h h h h h h h
153.99   h . h h . . . . h h h h h h h h h
154.99   h h h h . . . . h h h h h h h h h
155.99   h h h h . . . . h h h h h h h h h
```

```
156.99  h h h h h . . . h h h h h h h h h
157.99  h h h h h . . . h h h h h h h h h
158.99  h h h h h . . . h h h h h h h h h
159.99  h h h h . . . . h h h h h h h h h
160.99  h h h h h . . . h h h h h h h h h
161.99  h h h h h . . . h h h h h h h h h
162.99  h h h h h h . . h h h h h h h h h
163.99  h h h h h . . . h h h h h h h h h
164.99  h h h h h / . . h h h h h h h h h
165.99  h h h h h h h . h h h h h h h h h
166.99  h h h h h h h h h h h h h h h h h
167.99  h h h h h h h h h h h h h h h h h
168.99  h h h h h h h h h h h h h h h h h
169.99  h h h h h h h h h h h h h h h h h
170.99  h h h h h h h h h h h h h h h h h
```

Figure 3.3: A–C. Pages 73–75. Stacked plots of the (ϕ, ψ) angles of the residues of the slowly folding $(\text{Ala})_{20}$. A. residues 14–19, p. 73. B. residues 8–13, p. 74. C. residues 2–7, p. 75. For these plots, the y axis cycles from -180° to +180°, repeating every residue. The residues started in the extended state, and by the end of the graph at 170 ps they are in the α-helical configuration, where $(\phi, \psi) \approx (-60°, -40°)$. The horizontal lines at -52° approximate the (ϕ, ψ) angles of the α-helix. Points were averaged one point plotted per 1.0 ps simulated, i.e., one point plotted per ten data points recorded. Note bene: Error bars on all graphs extend one sigma above and below the average values of the ordinate, and the error bars encompass 67% of the values the ordinates are expected to take. Periodic boundary conditions were not taken into account in calculating the error bars. So, error bars extending over 360° are not accurate.

Figure 3.3 shows nucleation and propagation in more detail than Figure 3.2. Figure 3.3 illustrates the (ϕ, ψ) angles of individual residues during the entire trajectory. This graph demonstrates that residue 10 was in an α-helical orientation from about 25 ps to the end. Similarly, residues 16–19 were in an α-helix from ~25 to ~65 ps, although their helix did not persist. Persistent helices seem to nucleate simultaneously at about 120 ps and 135 ps at residues 10–13 and 15–19, respectively.

3.1.3 Other measures of structural rearrangement

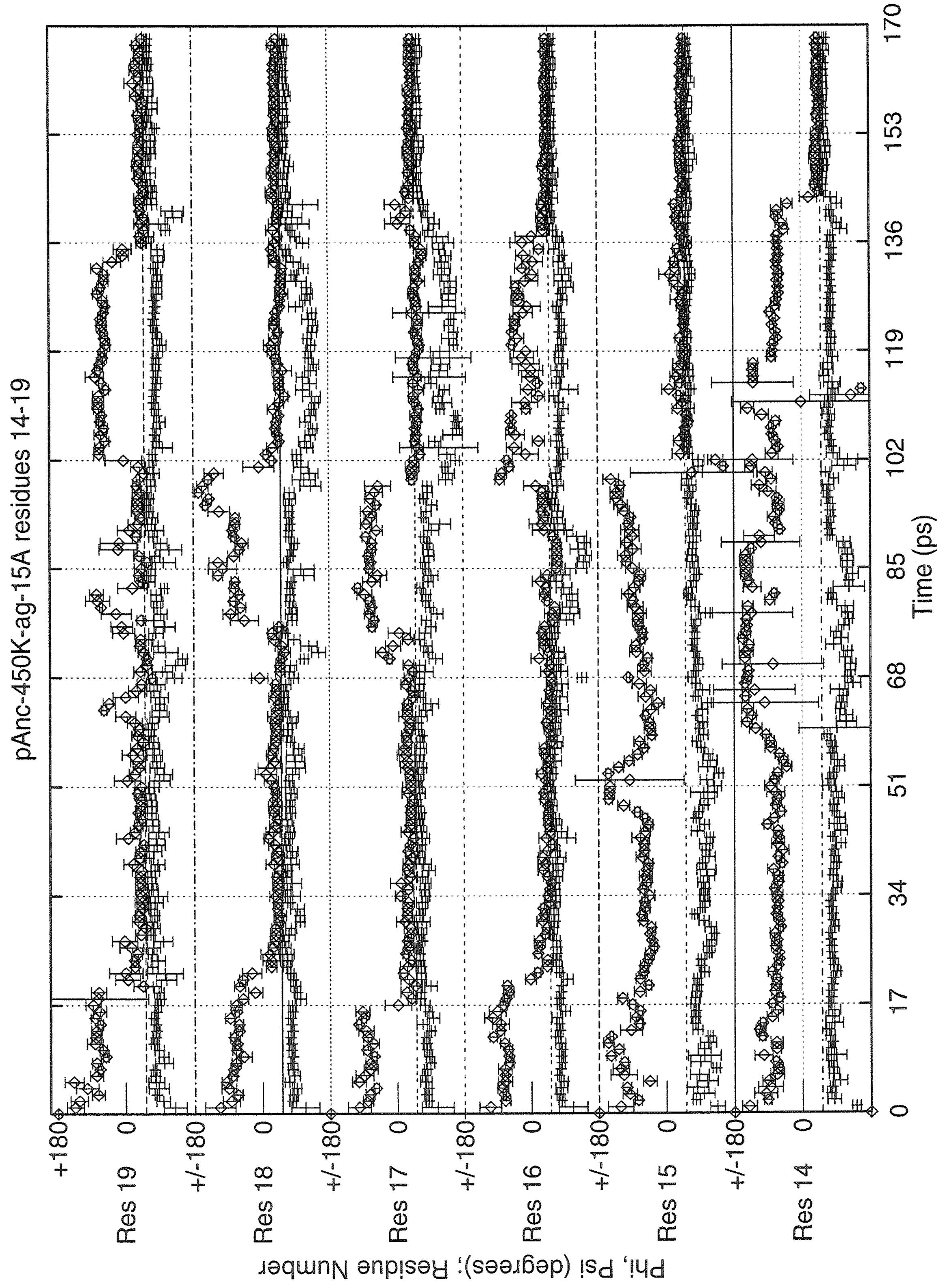


Figure 3.3A: Caption on p. 72.

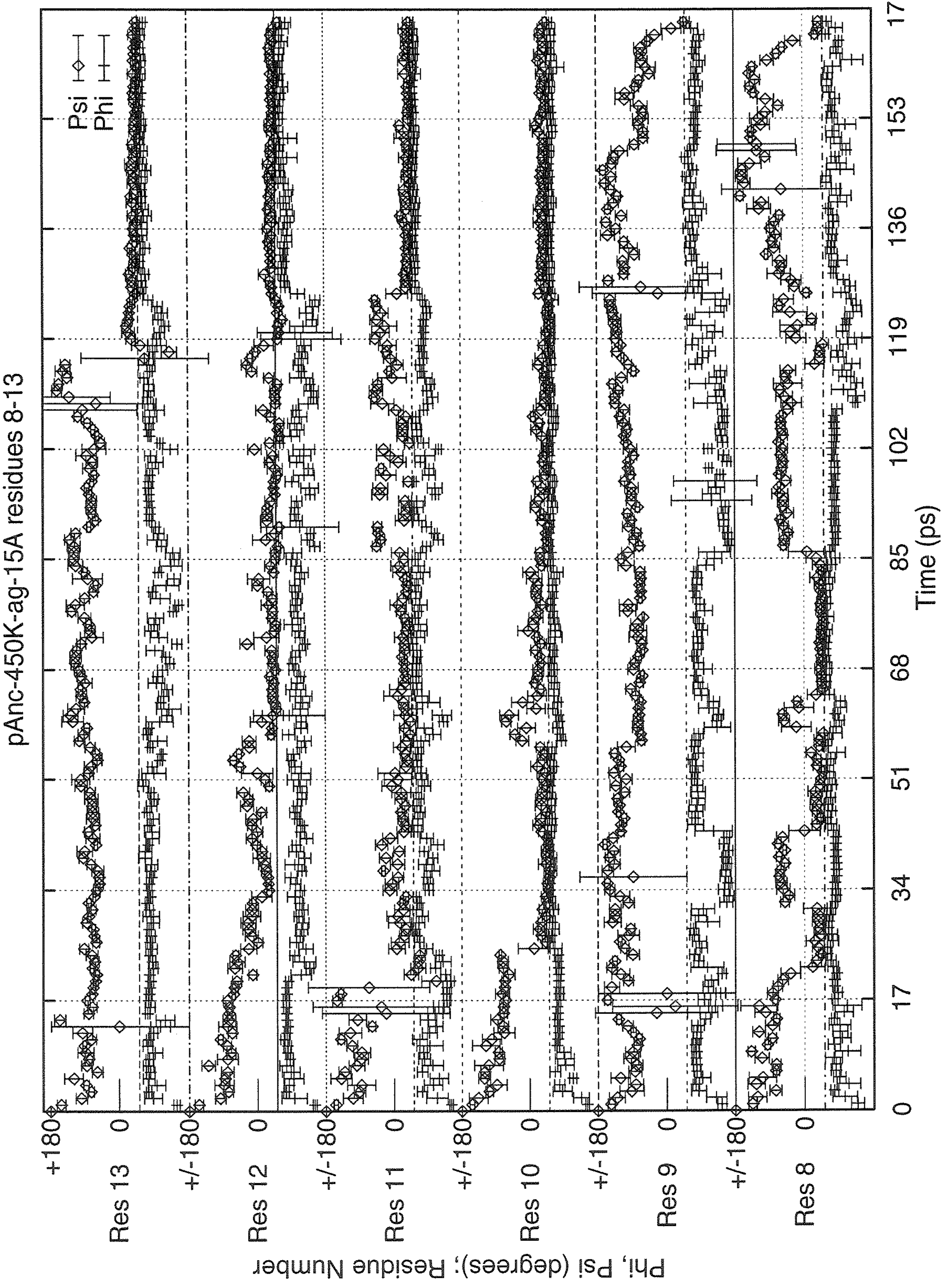


Figure 3.3B: Caption on p. 72.

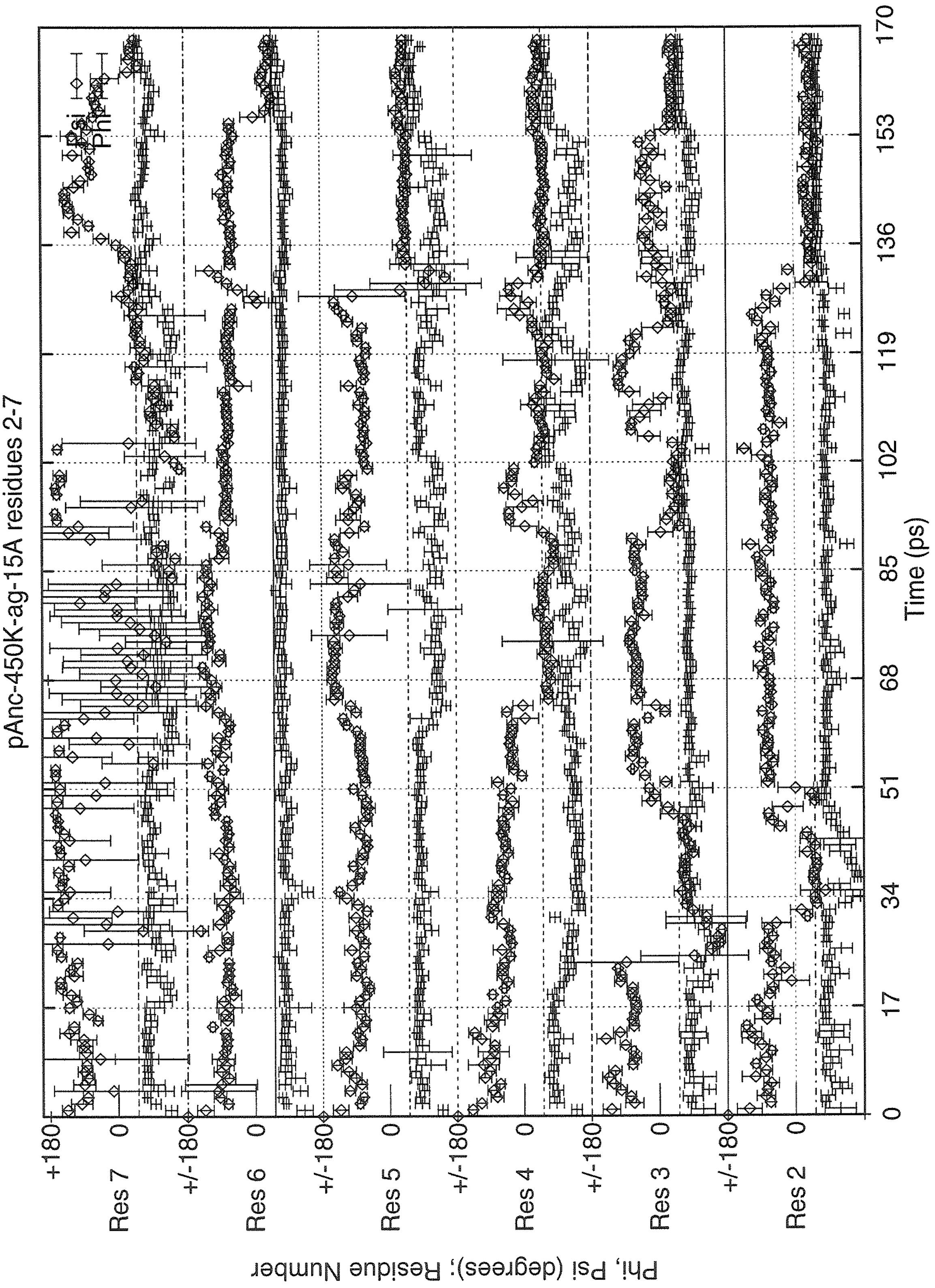


Figure 3.3C: Caption on p. 72.

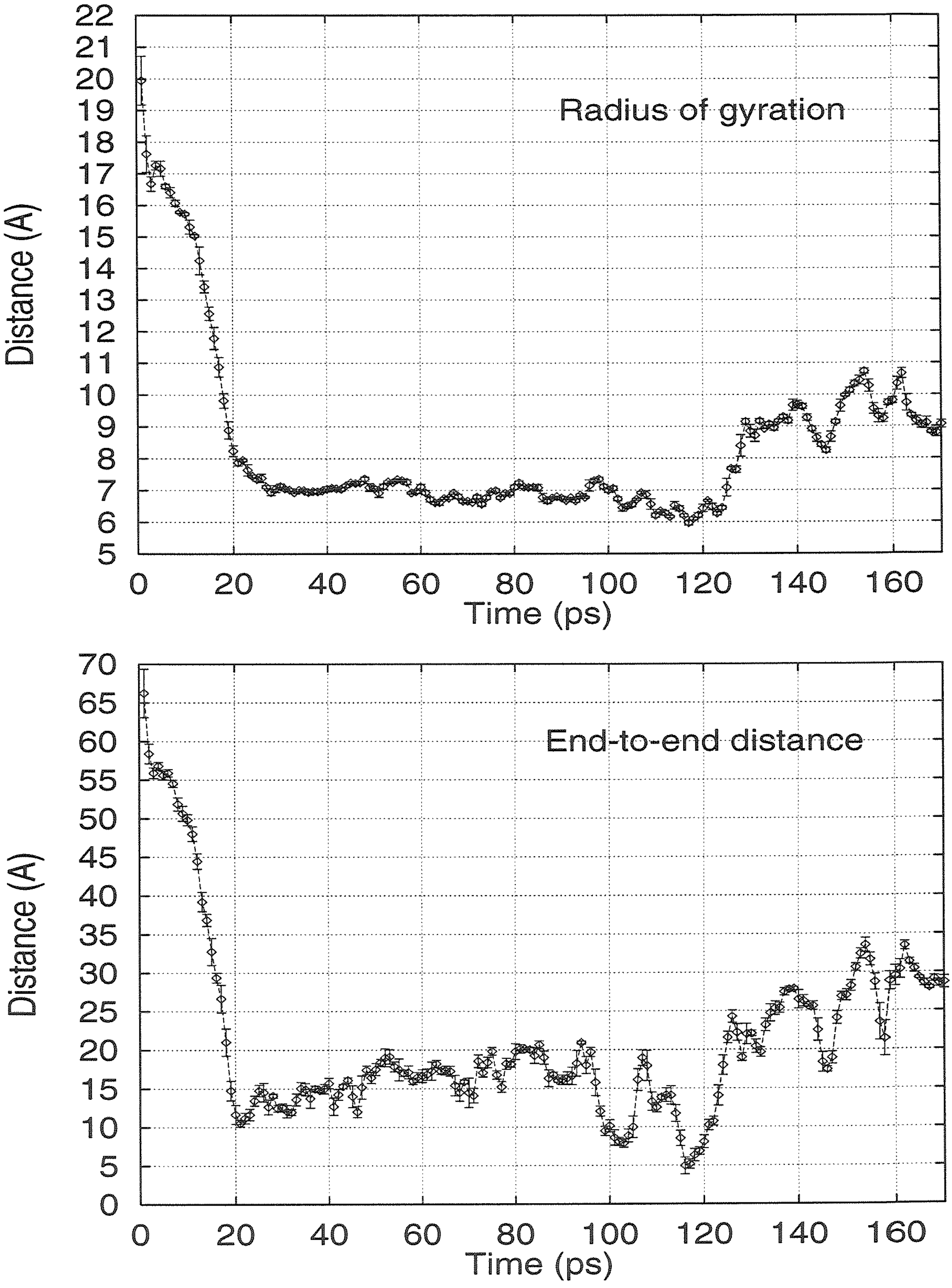


Figure 3.4: The radius of gyration and end-to-end distance during a slow helix-forming trajectory.

End-to-end distance and radius of gyration. Figure 3.4 plots the changes in the radius of gyration and end-to-end distance of the slow helix-forming trajectory. The minimum end-to-end distance at 20 ps occurs when the polymer resembles a bobby pin or narrow crochet hoop, approximated by Figure 3.1C. The end-to-end distance increases as the polymer balls up from ~30–90 ps. From ~95 ps to 125 ps, the end-to-end distance varies dramatically as the ends contract and expand in response to the gross rearrangements of the rest of the polymer. Finally, after 125 ps, the ends extend themselves, and the polymer quickly turns into an α-helix. As in Figure 2.4, the radius of gyration is relatively insensitive to changes in conformation. (See Section 2.1.3.)

Figure 3.5: p. 78. Ramachandran plots every 35 ps of the slowly folding trajectory. The (ϕ, ψ) angles of every residue at a single time point from 0 ps to 175 ps. The residues start in the first frame at $(\phi, \psi) = (-180°, +180°)$, move in the succeeding frames to the upper left-hand quadrant, and then coalesce into an α-helix by 175 ps.

Ramachandran plots. Just as Figure 2.5 illustrates for the fast-folding trajectory, Figure 3.5 shows the residues starting in the first frame at $(\phi, \psi) = (-180°, +180°)$, moving in the succeeding frames to the upper left-hand quadrant, and then coalescing into an α-helix by 100 ps.

The Ramachandran plots of the slowly folding trajectory are Figures 3.5 and 3.6. Figure 3.6 shows the trajectory spent much more time in nonhelical conformations than the fast folding trajectory. Furthermore, it is clear from this figure that the slowly folding trajectory sampled more regions of (ϕ, ψ) space than its fast folding counterpart, Figure 2.6. Residue 9 does not adopt α-helical dihedrals until just before 170 ps. This is consistent with Figure 3.1.2b.

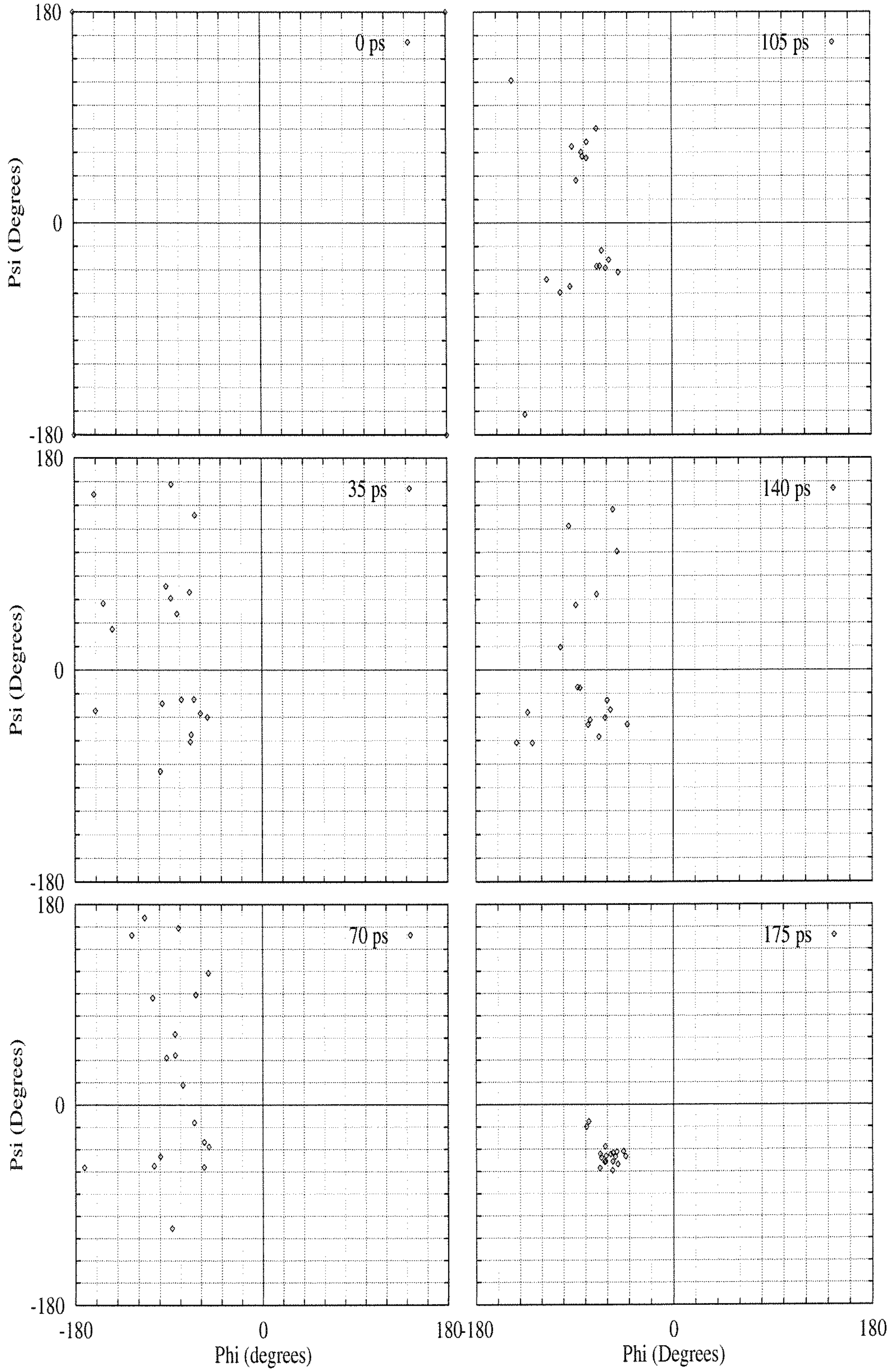


Figure 3.5: Caption on p. 77.

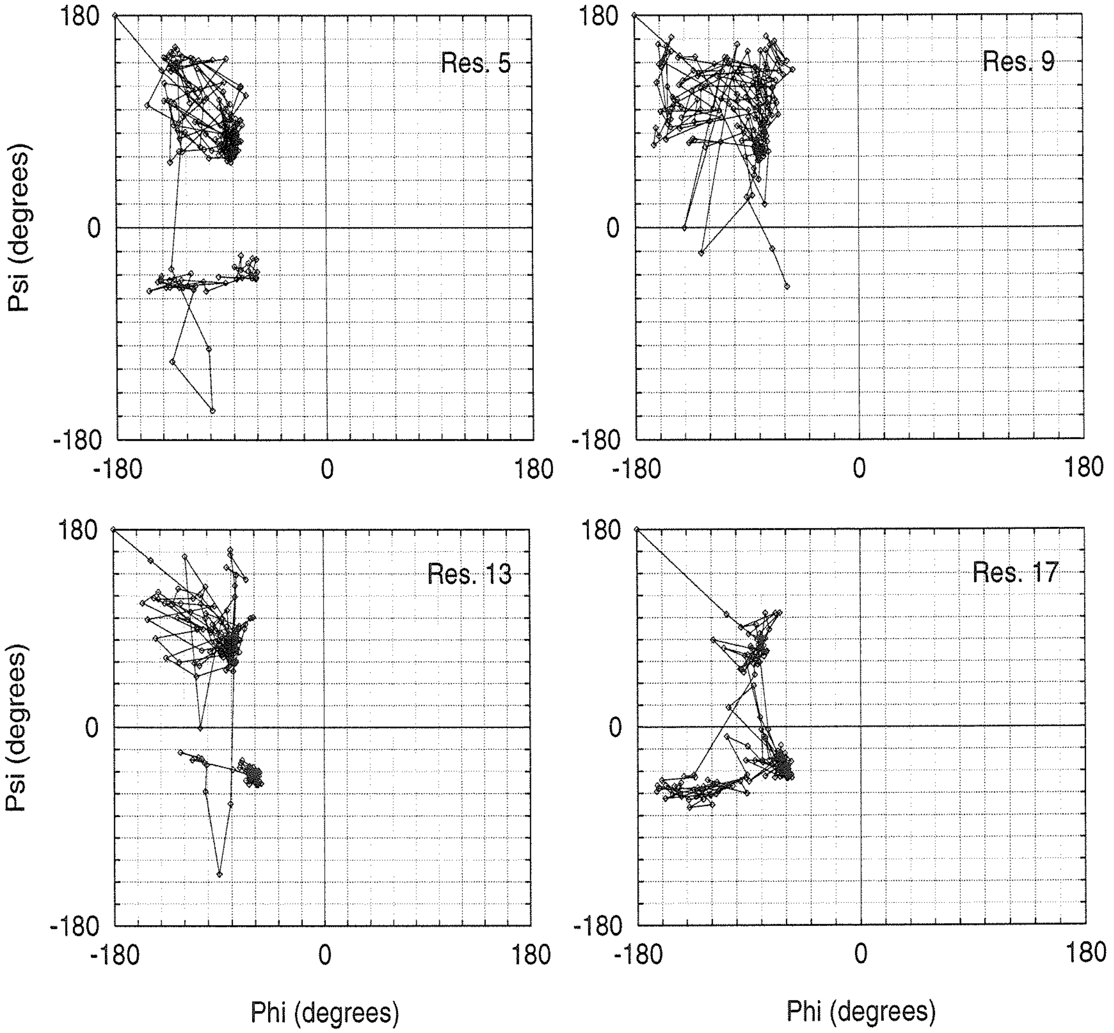


Figure 3.6: The trajectory of the ϕ, ψ dihedral angles of four residues during the slow, helix-forming trajectory. Residues 5, 9, 13, and 17 were chosen in order to compare them to the same residues in the fast-folding trajectory illustrated in Figure 2.6, p. 55. It is clear that residue 9 does not adopt α-helical dihedrals until just before 170 ps. Periodic boundary conditions at $\psi = \pm 180°$ were not taken into account when graphing the trajectories, but this only affects the appearance of the graph of residue 13.

3.2 The energetics of helix formation

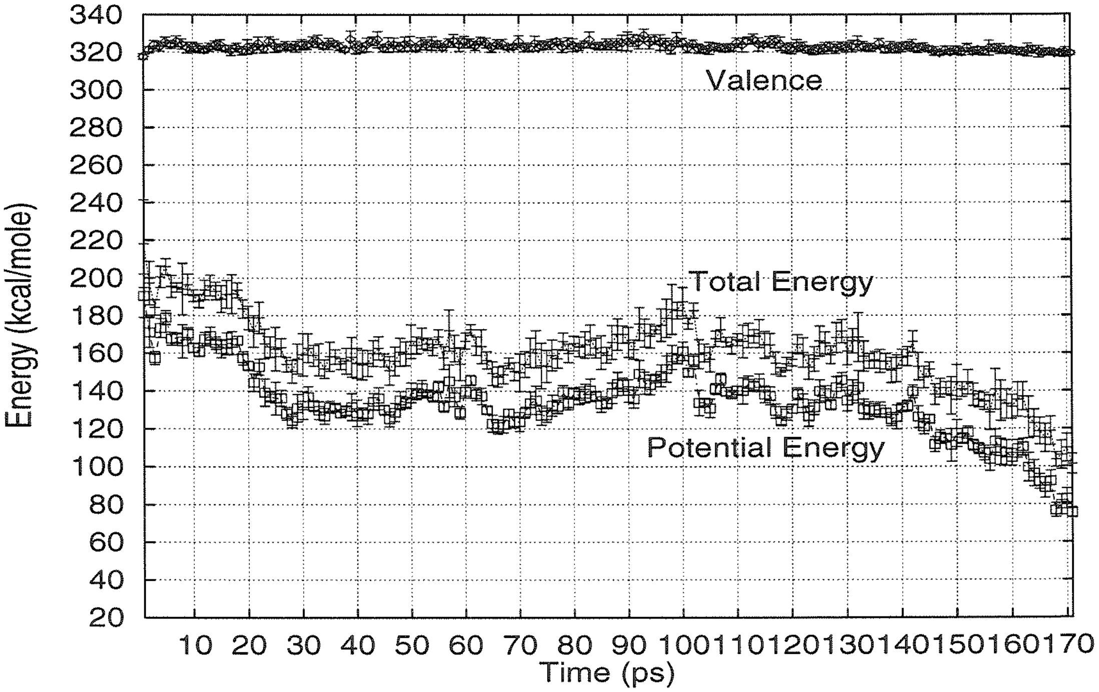
Like Figure 2.7, Figure 3.7, p. 80, graphs the energies of the different terms of the energy expression. Again, the major drops in energy occur at helix formation. The drop from 20–28 ps occurred when the nonlocal HBs pictured in Figure 3.1D, p. 66,

Figure 3.7: Figure p. 81. Energies of different terms in the AMBER energy expression during a slow, helix-forming trajectory. Top: total energy, valence, and potential energy. Bottom: Nonbond energy and its components van der Waals, electrostatic, and total hydrogen bonding energy. The total energy in hydrogen bonds is calculated from the hydrogen bonding term in the AMBER force field, a 12-10 potential (See Equation 2.2). Only backbone amide nitrogens and backbone carbonyl oxygens separated in sequence by at least one intervening residue were considered in the hydrogen bonding term. In addition, only hydrogen bonds stronger than -2 kcal/mole were totalled. Points are averaged 1 point graphed per 1.0 ps simulated or per ten data points recorded.

formed. The fluctuations in total HB energy from 90–120 ps appear to be gross structural rearrangements as the polymer breaks the tethering nonlocal HBs. The valence term remains relatively constant as a result of the NEIMO-Hoover algorithm. See Chapter 6, Section 6.3, p. 116. The difference in total and potential energy, i.e., the kinetic energy, appears constant because potential energy dominates the total energy.

3.3 The types of hydrogen bonds

The slow folding counterpart of Figure 2.11, Figure 3.8, p. 82, demonstrates that helix formation did not occur until the nonlocal HBs broke. The number of helical HBs did not increase above 4 until about 120 ps, when the number of nonlocal HBs dropped to zero. From about 25 ps to 90 ps, the polymer was balled up, trapped in nonhelical conformations. As will be shown in Chapter 4, Section 4.2.4, the best predictor of the speed of folding of $(\mathrm{Ala})_{20}$ is the maximum number of nonlocal HBs that form during a simulation.



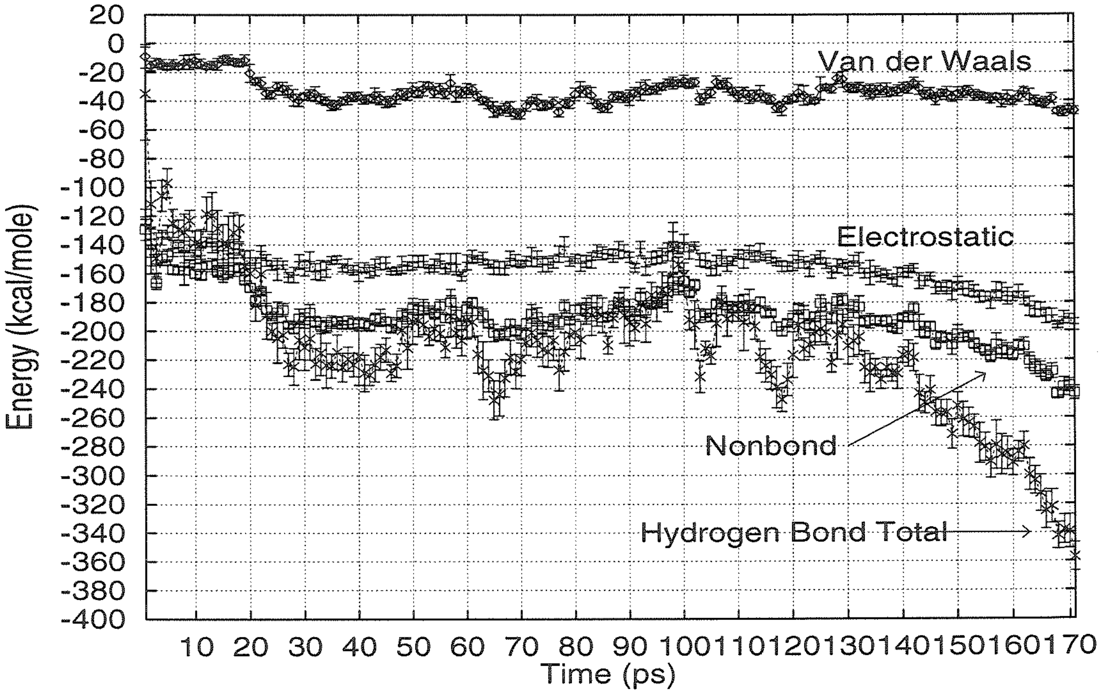


Figure 3.7: Caption p. 80.

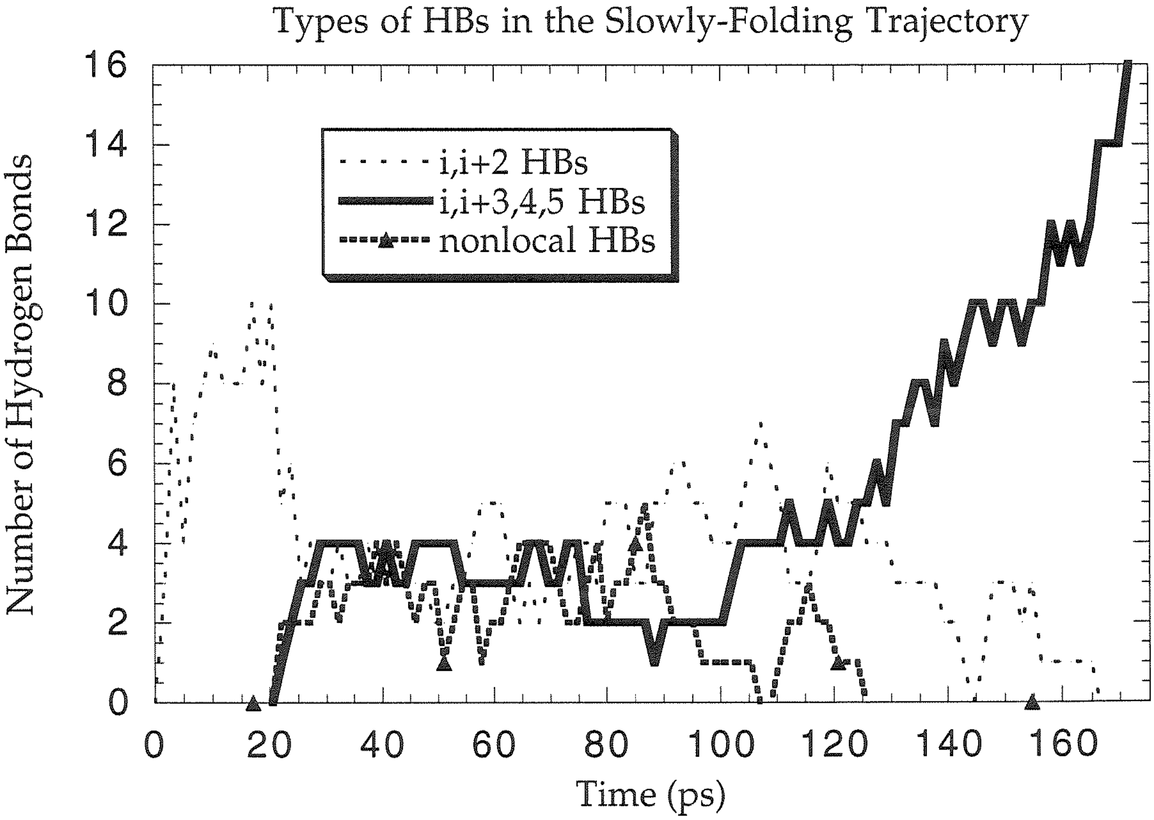


Figure 3.8: Numbers and types of HBs during a fast, helix-forming trajectory. The number of $i, i+2$ HBs is graphed in thin, dashed lines; the number of $i, i+4$ HBs in solid, thick, black lines; and the number of nonlocal HBs in thick, dashed lines with solid triangles superimposed. Nonlocal HBs are defined as $i, i+6$ to $i, i+19$ and $i, i-2$ to $i, i-19$ HBs. Points are averaged one point plotted per 17 points recorded, i.e., one point plotted per 1.7 ps.

Chapter 4 Results from all the Simulations

4.1 Trajectories and analysis routines are available

All of the polyalanine and polyglycine trajectories are available for downloading, viewing, and analysis from http://www.wag.caltech.edu/ or by anonymous ftp. Also available are all FORTRAN77 analysis programs and subroutines, including those used for reading the FORTRAN, binary data files.

4.2 Observations on polyalanine trajectories

Of the 155 simulations of extended polyalanine, 129 of them formed a helix within 500 ps. These simulations were used to identify events key to helix formation in $(\text{Ala})_{20}$.

4.2.1 Formation of i, $i+2$ hydrogen bonds, the C7 conformation

Within the first five picoseconds (ps) of every simulation, polyalanine relaxed to a conformation in which the dihedral angles of each residue were at $(\phi, \psi) \approx (-80°, +70°)$. This corresponds to a structure with strong $i, i+2$ HBs, which I will call call this region C7 in this thesis. The (~-80°, ~+70°) conformation has been called C7 because seven atoms compose the boat-shaped formed by the HB (Avignon *et al.,* 1969; Bystrov *et al.,* 1969).[1] An example of a strong $i, i+2$ HB is illustrated in Figure 4.1.

Each graph in Figure 2.6, p. 55, is the trajectory of the (ϕ, ψ) angles of one residue

[1] Refer to Footnote 1 for an explanation of why we do not call it a γ-turn.

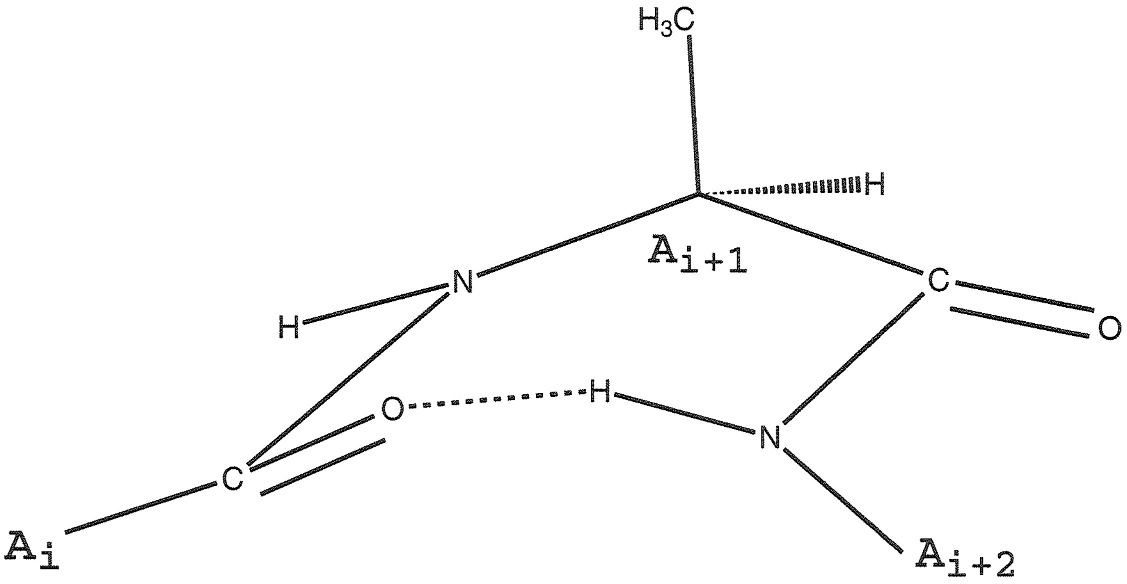


Figure 4.1: An $i, i+2$ HB in $(\mathrm{Ala})_{20}$.

during the simulation in Figure 2.1. Figure 2.6 clearly illustrates the existence of two favorable conformations: (a) the C7 conformation discussed above, and (b) the α-helix near $(\phi, \psi) = (-60°, -40°)$. Starting from the extended conformation of polyalanine, we find that all simulations went through the C7 region and 83% of the simulations continued on to form a persistent α-helix. Figure 4.2, p. 85, compares these regions to the preferred dihedral angles of proteins in the Protein Data Base (Bernstein *et al.*, 1977). See also Section 5.1.1 and Figure 5.1.

4.2.2 α-helix nucleation

Nucleation of α-helical segments occurred by several different mechanisms. In this thesis I define the nucleus of an α-helix to be the first $i, i+4$ HB that persists in the α-helical configuration until its surrounding residues also form $i, i+4$ HBs. The molecular dynamics (MD) simulations contained several patterns of α-helix nucleation.

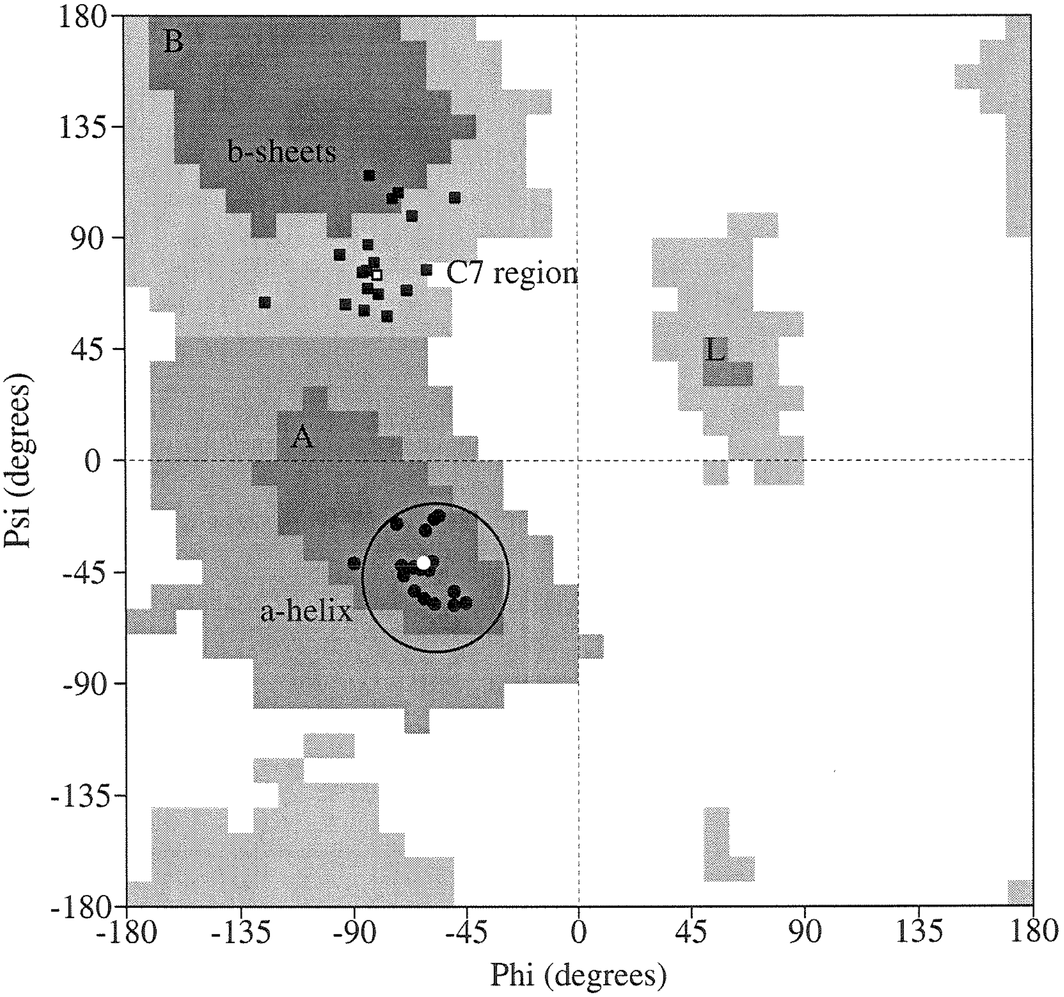


Figure 4.2: Comparison of (ϕ, ψ) angles of prominent structures in the PDB and in the simulations. The most darkly shaded regions (regions A, B, and L) are where good quality X-ray structures expect to place over 90% of their residues. Less densely shaded regions represent less favorable regions for residues. (Plot modified from PROCHECK (Laskowski *et al.*, 1993).) Briefly, the filled squares represent a polyalanine in the C7 configuration after 3 ps of simulation. (Specifically, the conformation originates from the trajectory that is labeled pAnc-450K-ad-15A in Appendix D and is depicted in Figure 2.1.) The white square is approximately at the center of the C7 region in our simulations, i.e., at $(\phi, \psi) \approx$ (-80°, +75°). The small, black, filled circles represent the polyalanine 200 ps into the same trajectory, after it had formed an α-helix. The white circle represents the average coordinates for α-helices in the PDB (Barlow & Thornton, 1988). The large circle encompasses the region considered α-helical in this study. The center is at $(\phi, \psi) = (-57°, -47°)$. Simulations were considered to have formed an α-helix if ~75% of their residues fell in the circle, i.e., within 30° of (-57°, -47°). See Chapter 6, Methods of Simulation, Section 6.7.

(a) $i, i+2$ **expansion:** Figure 2.8, p. 58, illustrates a case where the first persistent $i, i+4$ HB formed only after forming a sequence of $i, i+2$ and $i, i+3$ HBs. An $i, i+2$ HB formed first, then it lengthened its reach to become an $i, i+3$ HB, and this finally developed into an $i, i+4$ HB.

(b) $i, i+3$ **expansion:** Sometimes the nucleating $i, i+4$ HB developed from only an $i, i+3$ HB instead of from a sequence of $i, i+2$ and $i, i+3$ HBs as in (a). Sung (1995), in a Monte Carlo simulation on a hexadecamer of polyalanine in the AMBER (1984) force field at 300K, also observed a polyalanine nucleating an α-helical loop by first forming an $i, i+3$ HB and then shifting it to an $i, i+4$ HB.

(c) **Nonlocal induction:** Figure 4.3, p. 87, shows a case in which the formation of a nonlocal HB (in this case an $i, i+10$ HB) compacted the polymer, providing an opportunity for a residue at the resulting turn (in this case, residue $i-1$) to nucleate a persistent α-helix.

4.2.3 Mechanism of propagation

Propagating the nucleating α-helical HB into an α-helix did not appear to follow a specific sequence. Some simulations nucleated in the middle of the polymer, then formed a second nucleus near the amino terminus, and then formed a third one near the carboxy terminus. These nucleation sites grew independently, and then all three helices fused. Other simulations showed an α-helix forming first near the amino terminus, propagating through the amino terminal third and then through the middle third, and then fusing with an independently nucleated and propagated helical segment in the carboxy-terminal third of the polymer. At least one simulation in the group of 129 helix-forming simulations formed a helix in each possible sequence or combination of forming helices in a third of the polymer, propagating the helix through one or both of the other thirds, and/or fusing the helix with independently

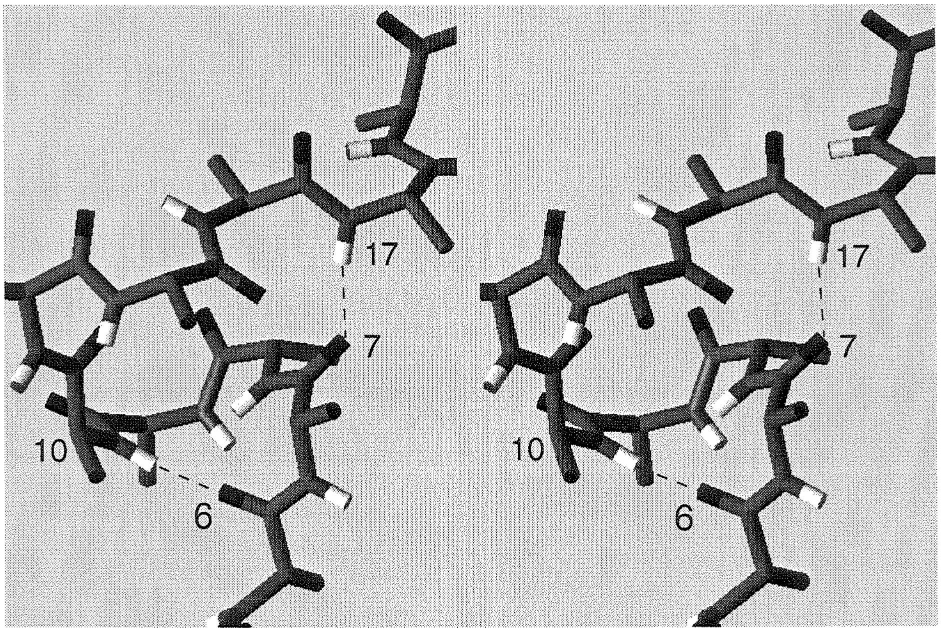


Figure 4.3: Formation of a large loop stimulates a helical loop to form. The formation of a nonlocal, $i, i+7$ HB from the carbonyl oxygen of residue 7 to the amide proton of residue 17 compacts the polymer. This provides an opportunity for the amide proton of residue 6 to form an α-helical HB with residue 10. (Trajectory pAnc-450K-al-15A at 43 ps.)

formed helices in one or both of the other thirds. See Figure 4.5, p. 89, Figure 4.4, p. 87, and Figure 4.6, p. 91, and Table 4.2.3, p. 93. There was no pattern to the method or order that the polymer thirds formed a helix.

Figure 4.4: p. 88. Helix nucleation and propagation in $(\mathrm{Ala})_{20}$. This figure illustrates a trajectory where the first nucleation is in the middle of the polymer, the helix propagates to the N-terminus, and then up to the C-terminus. A. 20.5 ps. Nucleation in the center of the polymer. The carbonyl of residue 6 has HBs to residues 8 $(i, i+2)$ and 12 $(i, i+6)$. The carbonyl of residue 8 has HBs to residues 12 and 13 $(i, i+4$ and $i, i+5)$. The N-terminus is on the left side. B. 32.8 ps. The helix is forming in the central third of the polymer. Residues 6, 7, and 9 are in $i, i+4$ HBs, extending the α-helix from residue 6 to approximately residue 13. The amino terminus is on the right side. C–E. The amino termini are at the bottom of the page. C. 54.7 ps. Residues 2–10 are in an α-helix. D. 91.3 ps. The α-helix propagates towards the C-terminal. Besides residue 11, which is out of register, most of the other residues are in $i, i+4$ HBs. E. 98.9 ps. The complete α-helix, residues 1–19. (Trajectory pAnc-450K-aw-15A.)

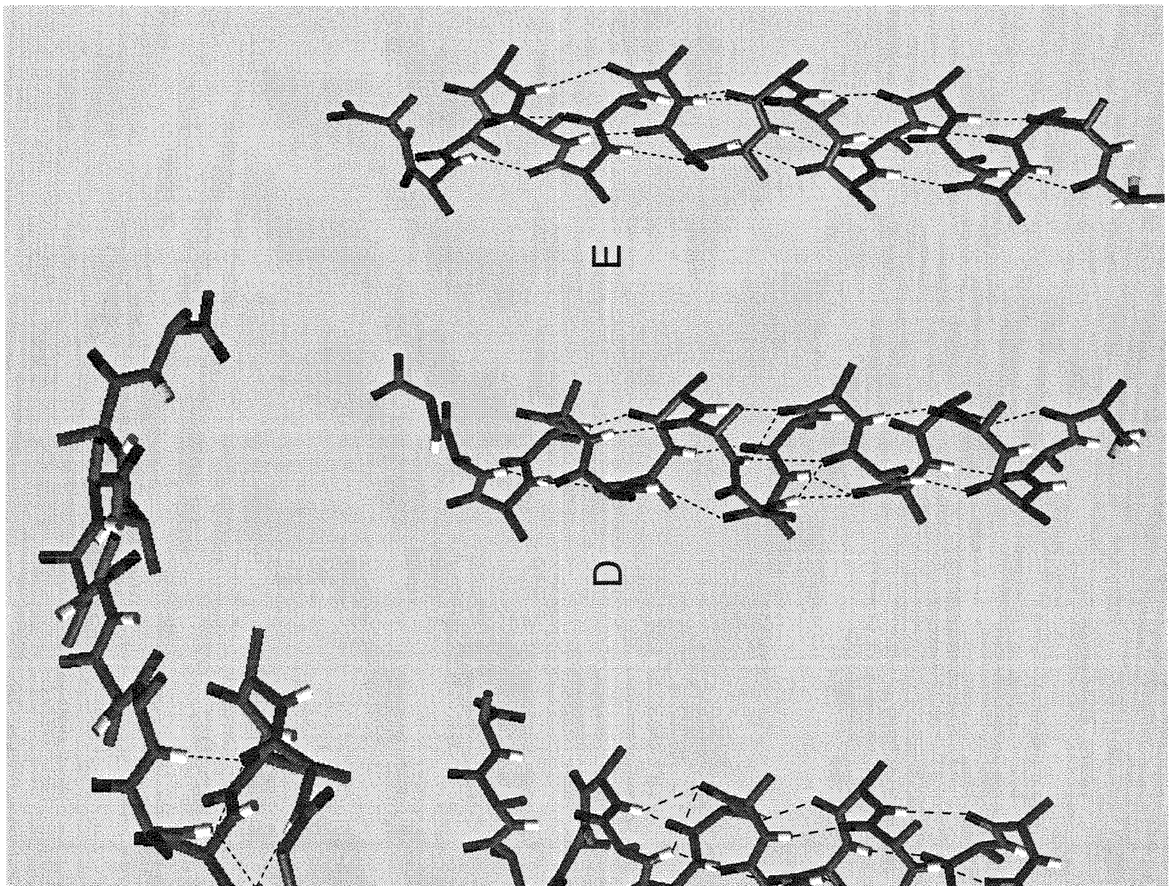


Figure 4.4: Caption on p. 87.

Figure 4.5: p. 89. Propagation and nucleation in $(\mathrm{Ala})_{20}$. This figures illustrates a trajectory where the first nucleation and propagation occurs at the N-terminus. An independent nucleation site occurs in the middle third of the polymer, and it propagates in the middle third of the polymer. Then the two α-helices fuse, and the resulting helix propagates out to the C-terminus. The N-terminus is shown in the bottom or the right side of each figure. A. 31.7 ps. Nucleation at the N-terminus. Residues 1 and 2 are hydrogen bonded in $i, i+4$ HBs to residues 5 and 6, respectively. B. 42.0 ps. The helix at the N-terminal propagates: residues 1, 2, 3, and 4 are hydrogen bonded in $i, i+4$ HBs to residues 5, 6, 7, 8, respectively. C. 51.3 ps. Residues 1–8 form an α-helix as in B. In addition, the middle section has nucleated a helix and started to propagate it. Residues 8 and 9 are in $i, i+4$ HBs to residues 12 and 13, respectively. D. 57.3 ps. The middle and N-terminal helices are fusing into one α-helix. Residues 1–15 are in a helix. Residues 1–8 are in an α-helix. The carbonyl oxygen of residue 8 is out of register and not hydrogen bonding to anything. Residues 9–15 are in a π-helix (composed of $i, i+5$ HBs). E. 59.8 ps. The helix propagates towards the C-terminus. Residues 1–12 form an α-helix. The carbonyls of residues 9 and 13 have two, bifurcating HBs as the helix propagates out to the C-terminus. F. 66.2 ps. The α-helix encompasses the entire length of the polymer. (Trajectory pAnc-450K-ar-15A.)

Helix propagation was frequently delayed either by (a) difficulties reorienting part of the polymer or by (b) nonlocal HBs. Figure 4.7, p. 94, exhibits a long stretch of α-helix that could not lengthen further because one of the terminal carbonyl oxygens of the helix had formed a stable HB to a residue far removed in sequence, preventing the helical carbonyl from hydrogen bonding to its $i+4$th neighbor.

These simulations indicate that nucleation occurs more frequently than propagation, suggesting that propagation is rate-determining. For example, some simulations had two or three nucleating events, although many only had one. Figure 4.8 depicts an example.

4.2.4 Nonlocal hydrogen bonds

As stated earlier (pp. xvi, 65), I define nonlocal HBs as all $i, i+6$ to $i, i+20$ and $i, i-3$ to $i, i-20$ HBs. We saw in Figure 4.3, p. 87, that nonlocal HBs sometimes facilitated

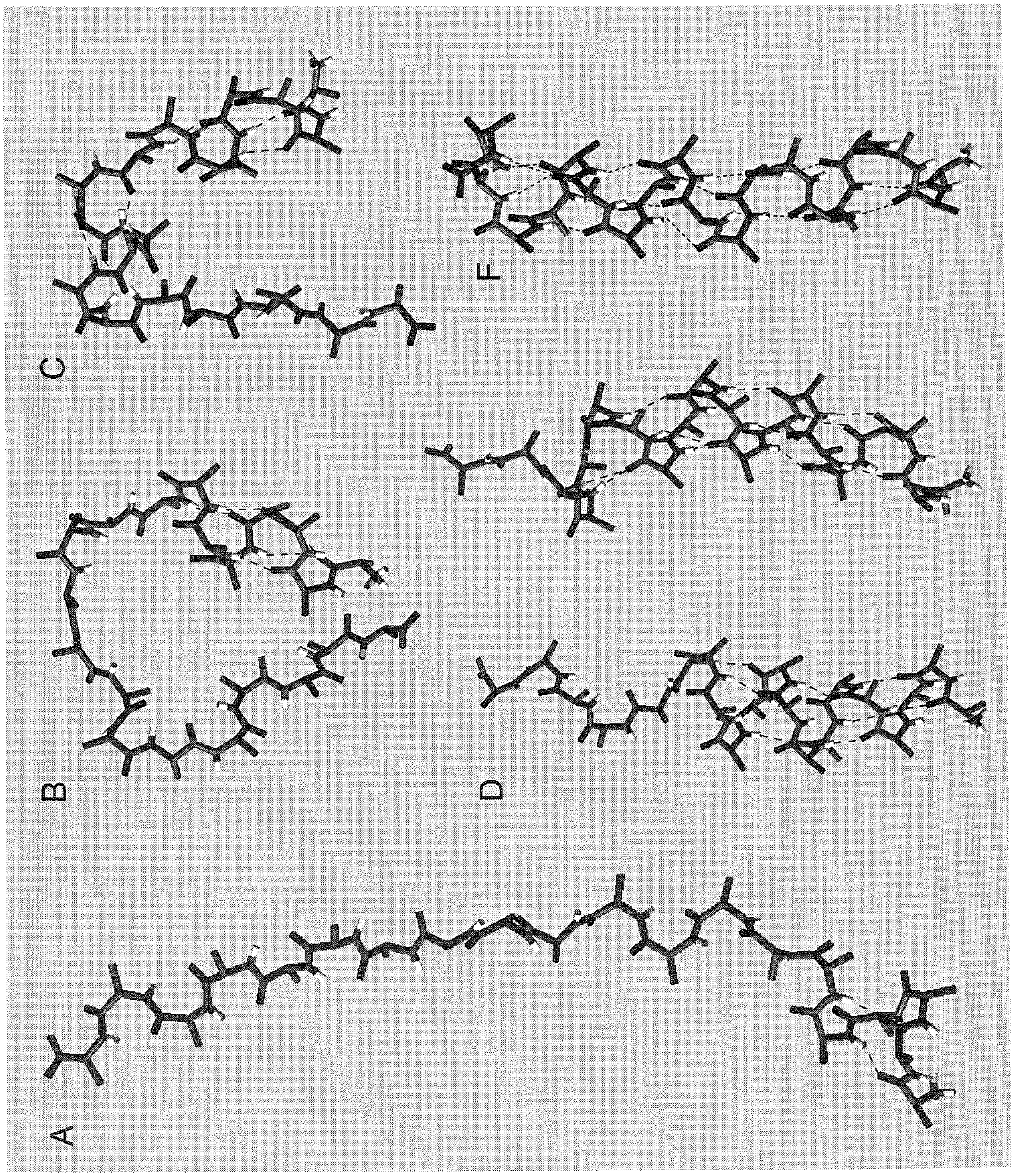


Figure 4.5: Caption on p. 89.

Figure 4.6: p. 92. Nucleation and propagation in $(\text{Ala})_{20}$. This trajectory illustrates the following order of propagation of an α-helix: (1) independent but simultaneous nucleations and propagations in the C-terminal and central thirds of the polyalanine, (2) independent helix nucleation at the N-terminus, (3) fusion of the helices in the two thirds, and (4) fusion of the long helix and the N-terminal helix. The N-terminus is the lower, left-most terminus in all diagrams. A. 37.4 ps. Independent nucleations in the C-terminal and central thirds. The carbonyl of residue 5 is hydrogen bonded to the amide nitrogen of residue 10 $(i, i+5)$, residue 7 to residue 11 $(i, i+4)$, and residue 8 to 12 $(i, i+4)$. In the C-terminal third, residue 12 is hydrogen bonded to the amide protons of residues 16 and 17 $(i, i+4$ and $5)$. B. 92.4 ps. Propagation of the helices in the C-terminal and central thirds. The middle helix has formed an α-helix from residues 5–12. The C-terminal helix has formed from residue 12–20. Residue 12 is hydrogen bonded to residues 16 and 17 $(i, i+4$ and $5)$ and residue 13 to 18 $(i, i+5)$. C. 98.6 ps. Nucleation at the N-terminus. In the central helix, the carbonyl of residue 1 is hydrogen bonded to residue 5 $(i, i+4)$. Residues 5–12 are in a helix, but residue 5 is hydrogen bonded to residue 9 in an $i, i+3$ HB. In contrast, the C-terminal helix is now fully α-helical from residues 12–20. D. 101.4 ps. Fusion of the central and C-terminal helices. Residues 5–20 are in $i, i+4$ HBs. Residue 1 is hydrogen bonded to residues 4 and 5 $(i, i+4$ and $5)$. E. 108.2 ps. The N-terminal helical loop forms a strong α-helical HB. F. 113.0 ps. The N-terminal loop has snapped onto place, making the entire polymer one α-helix. (Trajectory pAnc-450K-as-15A.)

nucleation because they forced the polymer into a loop that later tightened into turns of an α-helix. Nonetheless, in most cases nonlocal HBs impeded helix formation. Recall Figure 4.7, p. 94. Figure 3.1, p. 66, illustrates the trajectory of a polyalanine which got trapped as a ball of nonlocal HBs, delaying folding by 100 ps. In these cases, the nonhelical HBs had to break before a helix could form. In all cases that did not form a helix within the 500 ps simulation time (i.e., 17% of the total 155 runs), nonlocal HBs had trapped and tethered the $(\text{Ala})_{20}$'s in nonhelical balls.

We found that the time delay for forming a persistent α-helix could be related to the maximum number of nonlocal HBs formed during the trajectory. The delay time (τ) was defined as the elapsed time between the conformation having the largest number of nonlocal HBs and the time at which a persistent helix first formed. Figure 4.9, p. 96, shows $\ln(1/\tau)$ versus the maximum number of nonlocal HBs $(N_{MaxNLHB})$ dur-

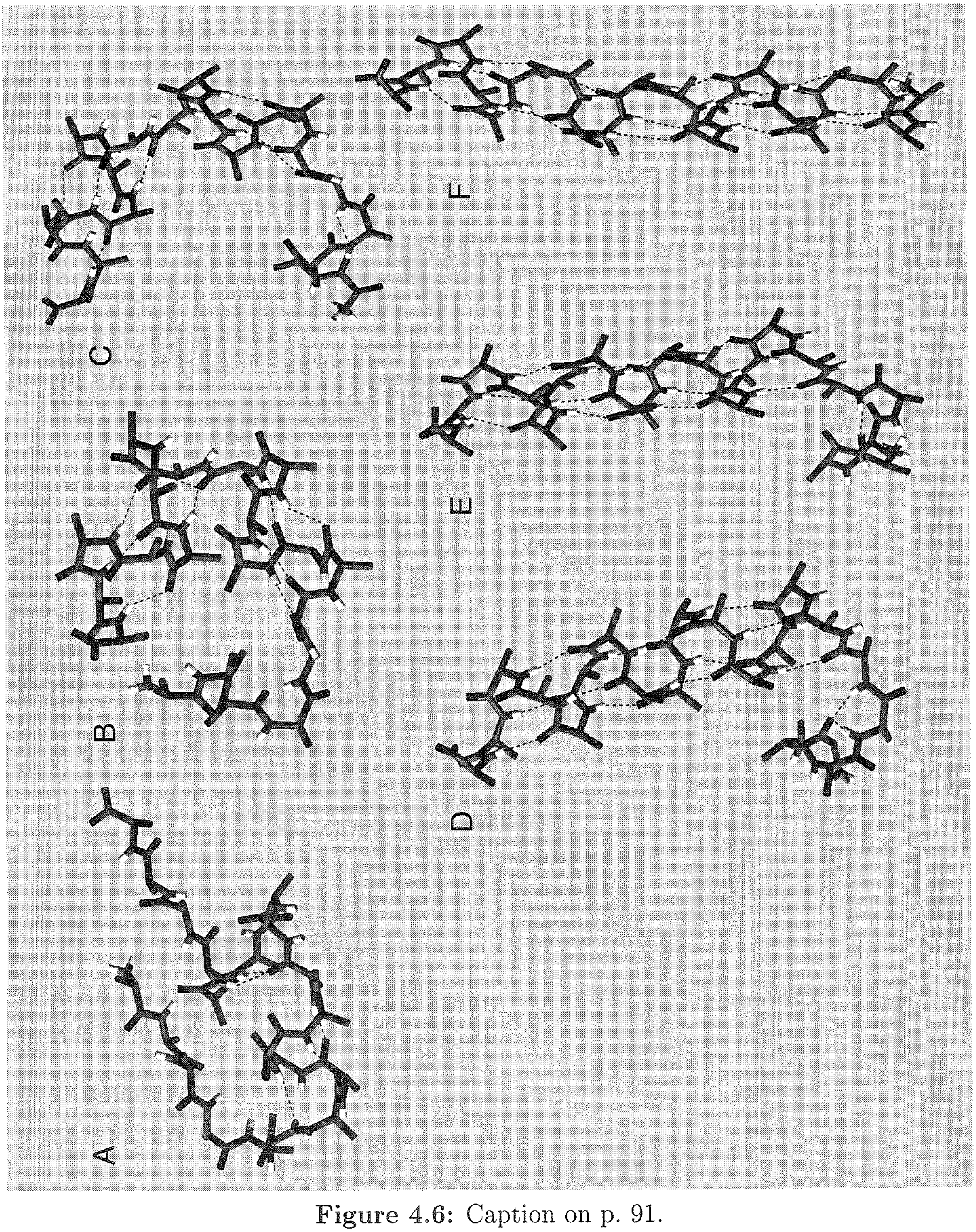


Figure 4.6: Caption on p. 91.

Table 4.1: Summary of the orders of helix propagation in the figures illustrating helix forming trajectories.

Figure	1st 1/3 to nucleate & propagate[a]	Method of joining[b]	2nd 1/3 to nucleate & propagate	Method of joining	Last 1/3 to nucleate & propagate
2.1	mid.	fuse	C-t	prop.	N-t
3.1	mid.	fuse	C-t	fuse	N-t
4.4	mid.	prop.	N-t	prop.	C-t
4.5	N-t	fuse	mid.	prop.	C-t
4.6	C-t	fuse	mid.	fuse	N-t

[a] "mid." denotes the middle third of the polyalanine. C-t and N-t denote the C-terminal and N-terminal thirds, respectively.

[b] "prop." denotes that the helical regions joined when the helix that first appeared continued adding residues, or propagating, until it reached the end of the helical section that appeared second. "Fuse" indicates two separately nucleated and propagated sections of helix that merged to form one larger section. Of course, the helix that appeared first did not fuse until after the neighboring helix had nucleated and propagated.

ing each simulation. The nearly linear relationship observed suggests that the rate constant for helix formation can be written as

$$\ln\left(\frac{1}{\tau}\right) = \left(\frac{-\Delta G_N^{\ddagger}}{k_B T}\right)$$

where k_B is the Boltzmann constant and T is the temperature. $\Delta G_N^{\ddagger}$ is the total free energy of activation for breaking nonlocal HBs and forming an α-helix. Assuming that the total free energy of activation is the sum of the free energies of activation for breaking all of the nonlocal bonds,

$$\Delta G_N^{\ddagger} = N_{MaxNLHB} * \Delta G_1^{\ddagger}$$

These simulations suggest the activation free energy for breaking a single nonlocal

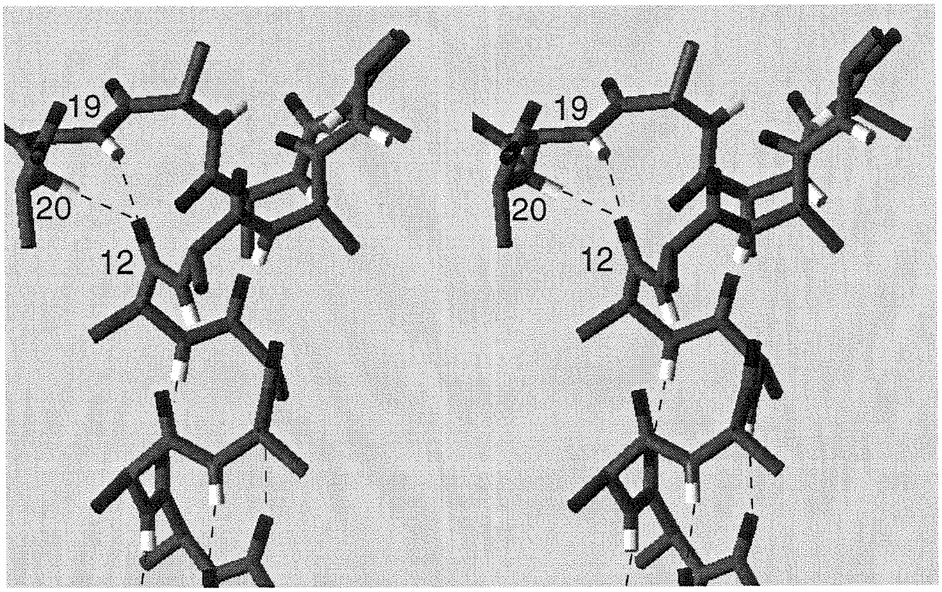


Figure 4.7: Propagation blocked by a nonlocal HB. This long stretch of α-helix (note the four $i, i+4$, HBs) cannot lengthen further because one of the terminal carbonyl oxygens of the helix (residue 12) is stably hydrogen bonded to two nonlocal protons (on residues 19 and 20). (Trajectory pAnc-450K-aw-15A at 51 ps.)

HB and forming a new, local, helix-nucleating HB is

$$\Delta G_1^\ddagger \approx 0.25 \text{ kcal/mol}$$

These simulations indicate that helix formation involves a competition between local and nonlocal HBs. Thus, factors that minimize large-scale displacements of polymer segments should favor helix formation by minimizing the formation of nonlocal HBs. This suggests there is an optimal number of residues for rapid helix formation.

4.3 Rate constant for helix formation

A rate constant for helix formation can be estimated based on the trajectories of the 155 runs. Figure 4.10 graphs the percentage of runs that formed in the given amount of time. Because only 17% of the "ensemble" had folded by the end of the simulation time, 26 of the folding times were extrapolations. For each of the 26 nonhelix-forming

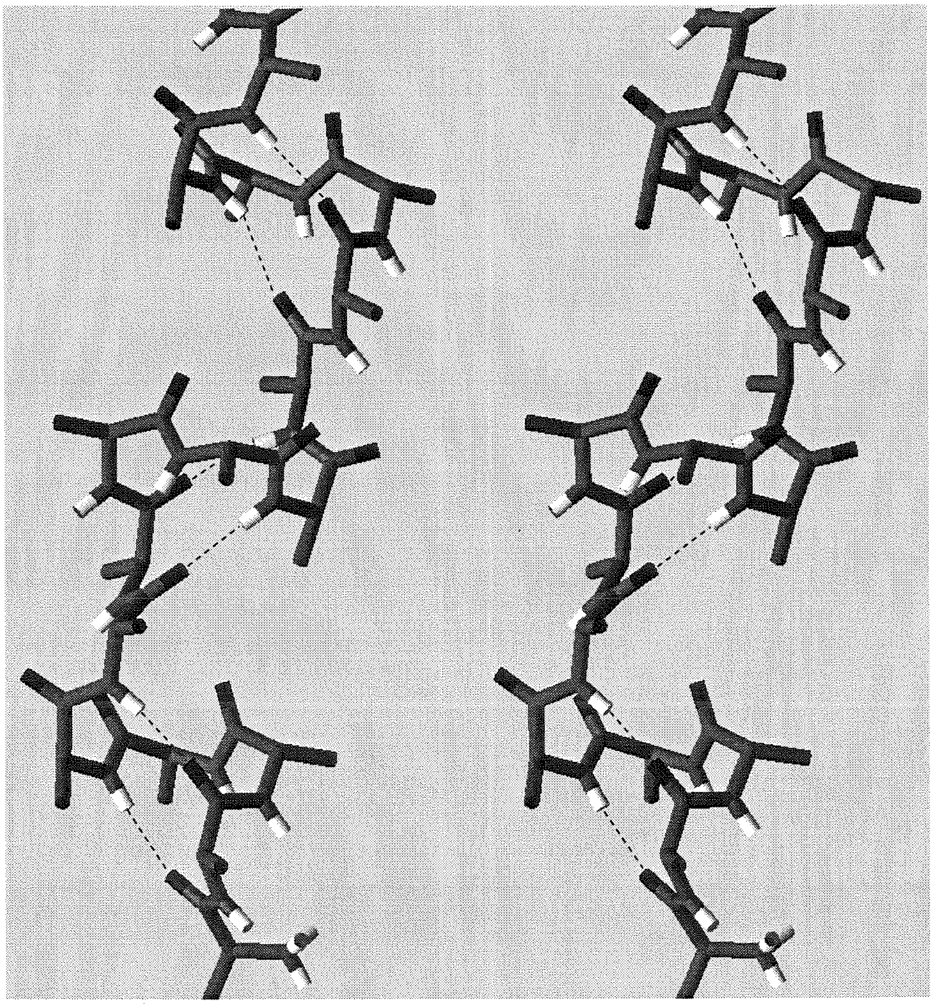


Figure 4.8: A simulation with three, apparently independent nucleation events. Notice the three growing regions of small α-helices, characterized by the marked $i, i+4$ HBs. (Trajectory pAnc-450K-ax-15A at 36 ps.)

runs, the average helix formation time was estimated from the line fitting the data in Figure 4.9 using the actual maximum number of nonlocal HBs in each nonhelix-forming trajectory. The average helix formation time computed from the line was added to the time it took each nonhelix-forming run to form the configuration with the maximum number of nonlocal hydrogen bonds. In other words, for each nonhelix-former, the extrapolated helix formation time used to graph Figure 4.10 was the sum of the time to form the structure with the most nonlocal hydrogen bonds plus the

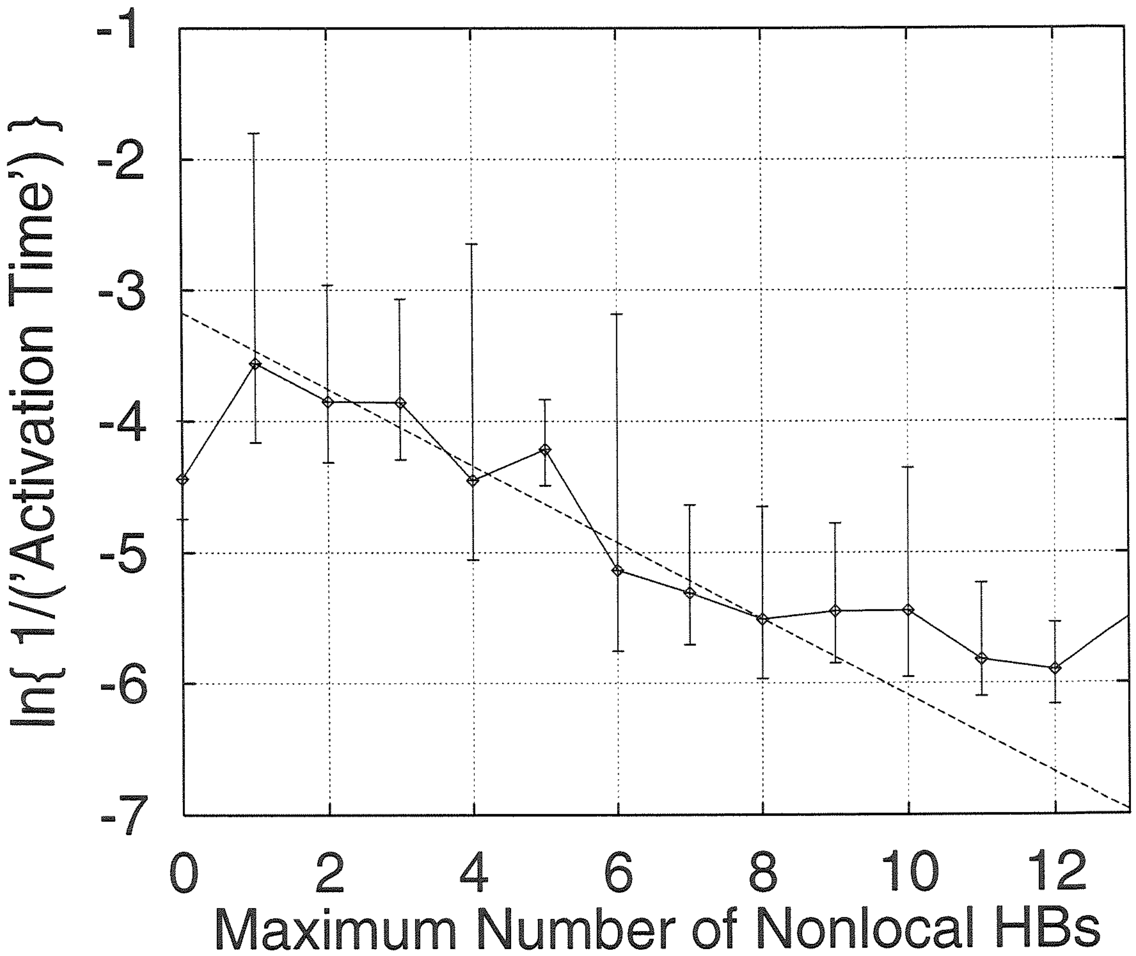


Figure 4.9: Activation free energy per nonlocal HB. The logarithm of the reciprocal of the 'Activation Time' of helix formation versus the maximum number of nonlocal HBs formed during the simulation. The activation time was defined as the time between the conformation with the largest number of nonlocal HBs and the time of helix formation. Nonlocal HBs were defined as $i, i+6$ to $i, i+20$ and $i, i-3$ to $i, i-20$ HBs. Assuming Arrhenius behavior, an activation free energy per nonlocal HB was estimated at ~0.25 kcal/mole. Error bars are ± 1 σ standard deviation, or to 68% confidence. The line drawn through points 1–7 fits $y = -(0.292)x - 3.2$.

average helix formation time calculated from

$$ln(\frac{1}{\tau}) = -(0.292)N_{MaxNLHB} - 3.2$$

The rate constant that best fit a single exponential decay was $k = 0.004779$ ps^{-1}. This yields a half-life of 209 ps for helix formation. Table 4.2 converts the rate constant

Table 4.2: Times and percentages of folding with rate constant $k = 0.004779$ ps^{-1}.

Percentage of Ensemble Folded	τ, Time of Folding
50% ($\tau_{\frac{1}{2}}$)	209 ps
68%	238 ps
95%	627 ps
99.7%	1216 ps

into percentages folded vs time.

4.4 Polyglycine does not form a helix

Excluding proline, glycine is the worst helix former among the naturally occurring amino acids (Chakrabartty *et al.*, 1994). To test if the MD distinguishes between glycine and alanine, I carried out simulations for (Gly)$_{20}$ in the same way as for (Ala)$_{20}$.

Only 10% of the 20 simulations on (Gly)$_{20}$ formed a right-handed α-helix, in contrast to 83% for polyalanine. The polyglycine helices took longer to form (an average of 400 ps versus 147 ps for polyalanine), were shorter (an average of 12 residues versus at least 18), and were much less stable than polyalanine helices. For instance, the average percentage helicity of (Gly)$_{20}$ when it had substantial helical segments was 58% (± 12%), whereas the average percentage helicity of (Ala)$_{20}$ after helix formation was 91% (± 10%). See Table 4.4. Figure 4.11, p. 101, illustrates the longest right-handed α-helix (Gly)$_{20}$ formed during the 20 simulations.

Left-handed helices formed in two of the 20 (Gly)$_{20}$ runs. The left-handed helices were also transient and unstable. Figure 4.12, p. 102, illustrates a seven residue left-handed helix that lasted for 17 ps. The other left-handed case had about five residues in a π-helix (consisting of $i, i+5$ HBs) for about 10 ps. See Table 4.4. Polyglycine seemed to favor right-handed α-helices when starting from the extended state.

Table 4.3: Comparison of α-helices in $(\mathrm{Ala})_{20}$ and $(\mathrm{Gly})_{20}$.

	Polyalanine	Polyglycine
No. of simulations	155	20
No. of complete α-helices formed	129	0
Average percentage of residues in an α-helical configuration	90% ± 10%[a]	1.9% ± 2.5%[b]
No. of α-helical segments	129/155	2/20
Average length of α-helical segment (residues)	18 res.	12 res.
Average time for RH α-helix formation	147 ps	400 ps
Average persistence time of RH α-helix	353 ps	75 ps

[a] in a helix

[b] from 100–500ps

Table 4.4: The Types of helices $(\mathrm{Gly})_{20}$ formed.

Type of helix	Frequency	Persistence time	Average No. of Residues
Right handed α-helix[2]	2/20	12 ps & 141 ps	12 residues
Left handed α-helix[3]	1/20	$\leq$ 190–207 ps	$\leq$ 6 residues
Left handed π-helix[4] (i,i+5 helix)	1/20	~136–145 ps	$\leq$ 7 residues

Polyglycine sampled all four quadrants of the (ϕ, ψ) space whereas polyalanine sampled only two (the $(-,+)$ and $(-,-)$ quadrants). This was expected, since glycine is achiral but alanine is chiral. In addition to the C7 and the α-helix regions favored by polyalanine, polyglycine also favored regions corresponding to right-handed and left-handed inversions of the C7 region and the α-helix, which we will call $\mathrm{C7}_G$ (at ($\sim$+80°, $\sim$-65°)) and α_G (at ($\sim$+60°, $\sim$+40°)). Refer to Figure 4.13, p. 103. The achirality of $(\mathrm{Gly})_{20}$ should prevent it from favoring right-handed helices over left-handed helices. Probably $(\mathrm{Gly})_{20}$ appeared to favor right-handed helices in these simulations only because 20 simulations are undoubtedly insufficient to distinguish between the probabilities of occurrence of the two unlikely events: right-handed and left-handed helix formation.

Figure 4.15, p. 105, and Figure 4.14, p. 104, illustrate the differences in the energy landscapes of polyglycine and polyalanine. Polyglycine has regions with favorable energy in both quadrants. The energy barrier between the left and right quadrants is small for $\phi = -180° = +180°$ but formidable for $\phi = 0$. In contrast, polyalanine encounters ridges of barriers in the two $\phi > 0°$ quadrants. Thus, polyalanine is not expected to sample conformations with $\phi > 0°$; and the trajectories confirm this.

Energetically polyglycine favors an α-helix just like polyalanine. However, $(\mathrm{Gly})_{20}$ has 2^{20} times as many energetically favorable conformational choices as polyalanine (i.e., the $C7_G$ and α_G regions in addition to the C7 region and the α-helix), and only one of these conformations is a right-handed α-helix. Thus entropy, not enthalpy, prevents polyglycine from forming the α-helix.

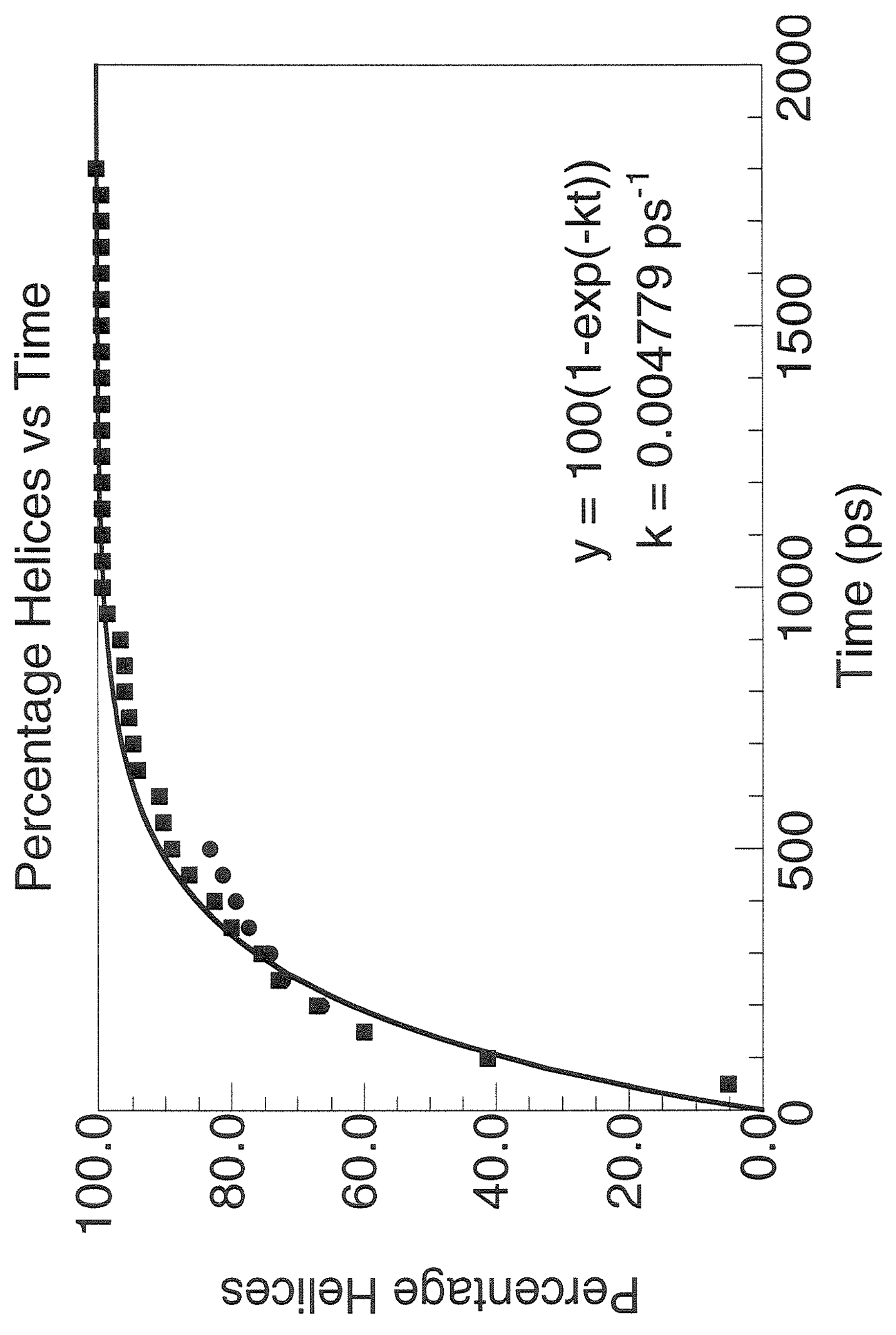


Figure 4.10: Percentage of the ensemble of 155 that formed a helix vs time. The circles represent the data from helix-forming runs, where each helix formation time is known. The squares represent the known data plus estimations of the helix formation times for the 26 runs that did not form a helix in 500 ps. The estimation procedure is described in the text on p. 94. The curve best fitting the points is $P = 100(1 - \exp(-0.004779 \times t))$, where P is the percentage folded and t is in ps.

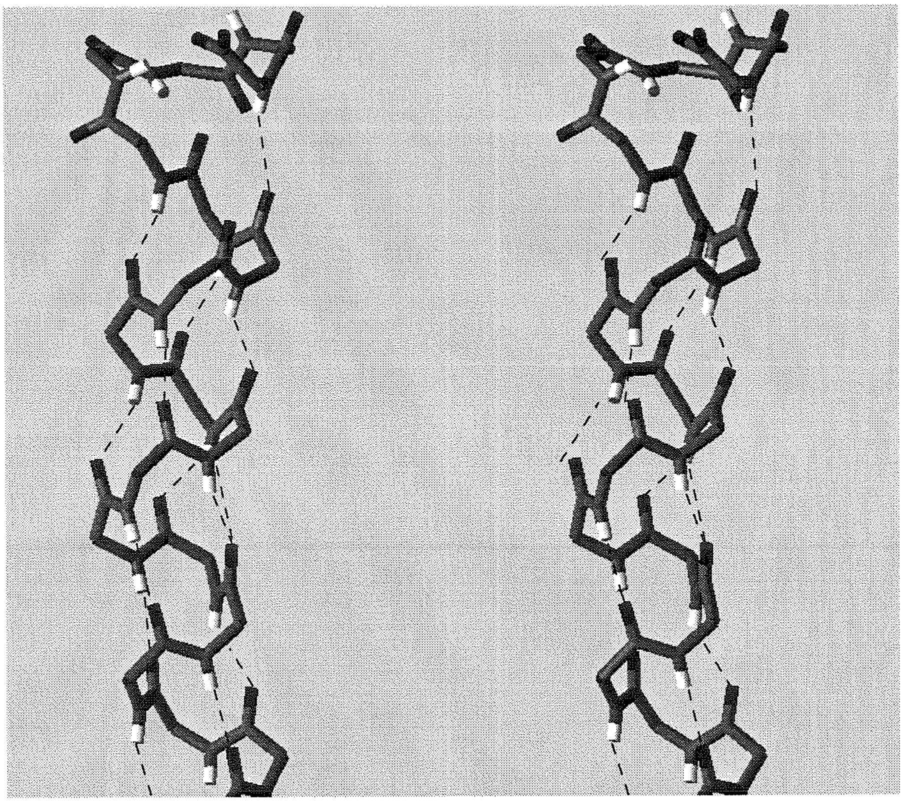


Figure 4.11: The longest right-handed α-helix the polyglycine simulations exhibited. Residues 16–20 form the beginnings of a left-handed helix. The amino terminus is at the bottom of the figure; hydrogens are white. (Trajectory pG-450K-c-15A at 410 ps.)

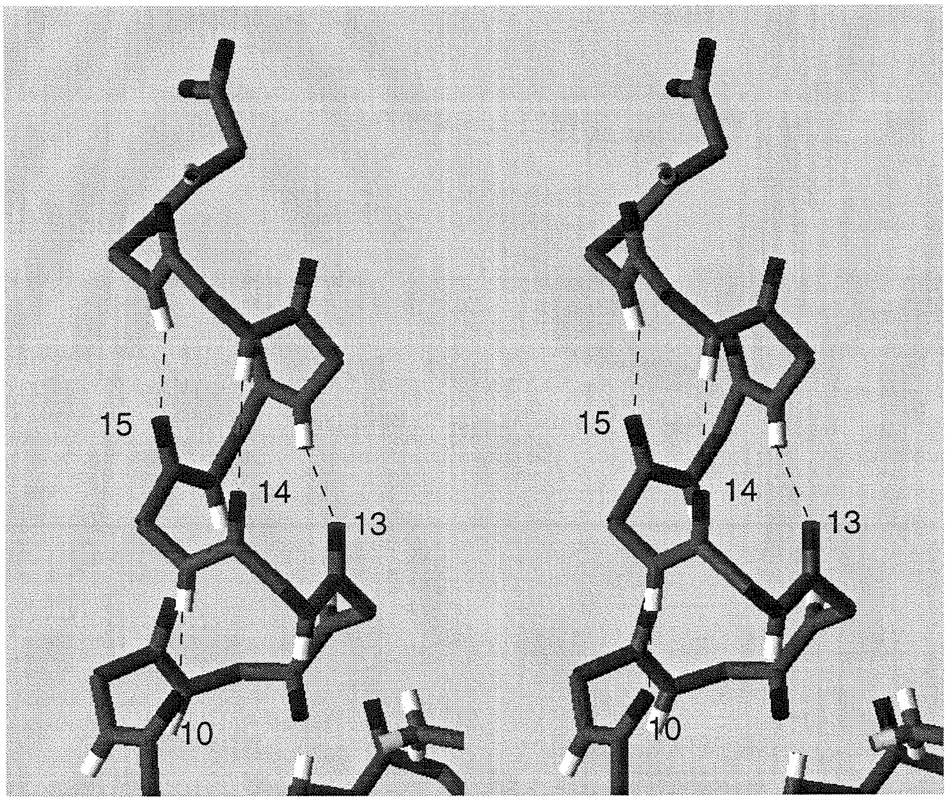


Figure 4.12: The longest left-handed α-helix the polyglycine simulations exhibited. Residues 13–18 persisted in the helix for no more than 17 ps. (Trajectory pG-450K-t-15A at 196 ps.)

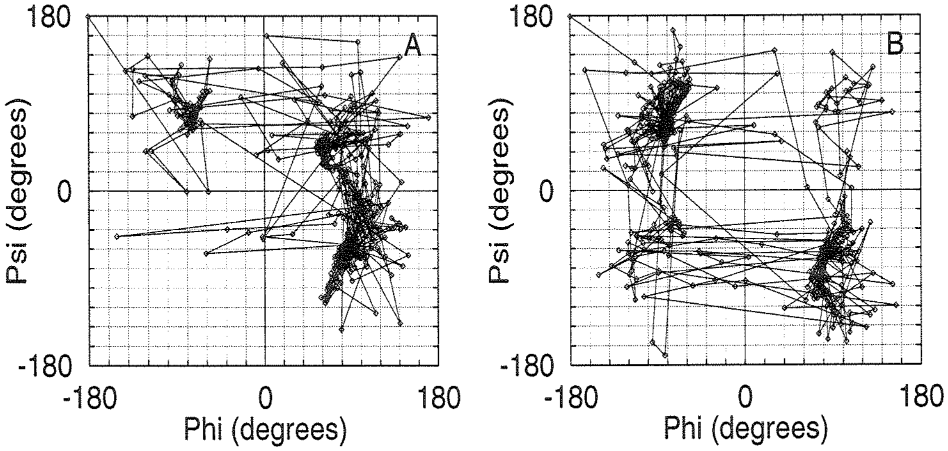


Figure 4.13: The trajectories that the dihedral angles ϕ and ψ of residue 3 (A) and residue 14 (B) took during two of the polyglycine simulations. Like Figure 2.6, this figure clearly illustrates the existence of energetic minima separated by large barriers. Both (A) and (B) exhibit the C7 region at (~-80°, ~+70°) that characterizes structures with strong, $i, i+2$ HBs. The C7$_G$ region is visible in (A and B) at (~+80°, ~-65°), and the α_G region are visible at (~+60°, ~+40°) in B. The area characterizing α-helices, at (~-60°, ~-40°), is not well populated by polyglycine. The plot does not make use of periodic boundary conditions to connect the dots to the closest torsion. (Figure A from pG-450K-u-15A and Figure B from pG-450K-h-15A.)

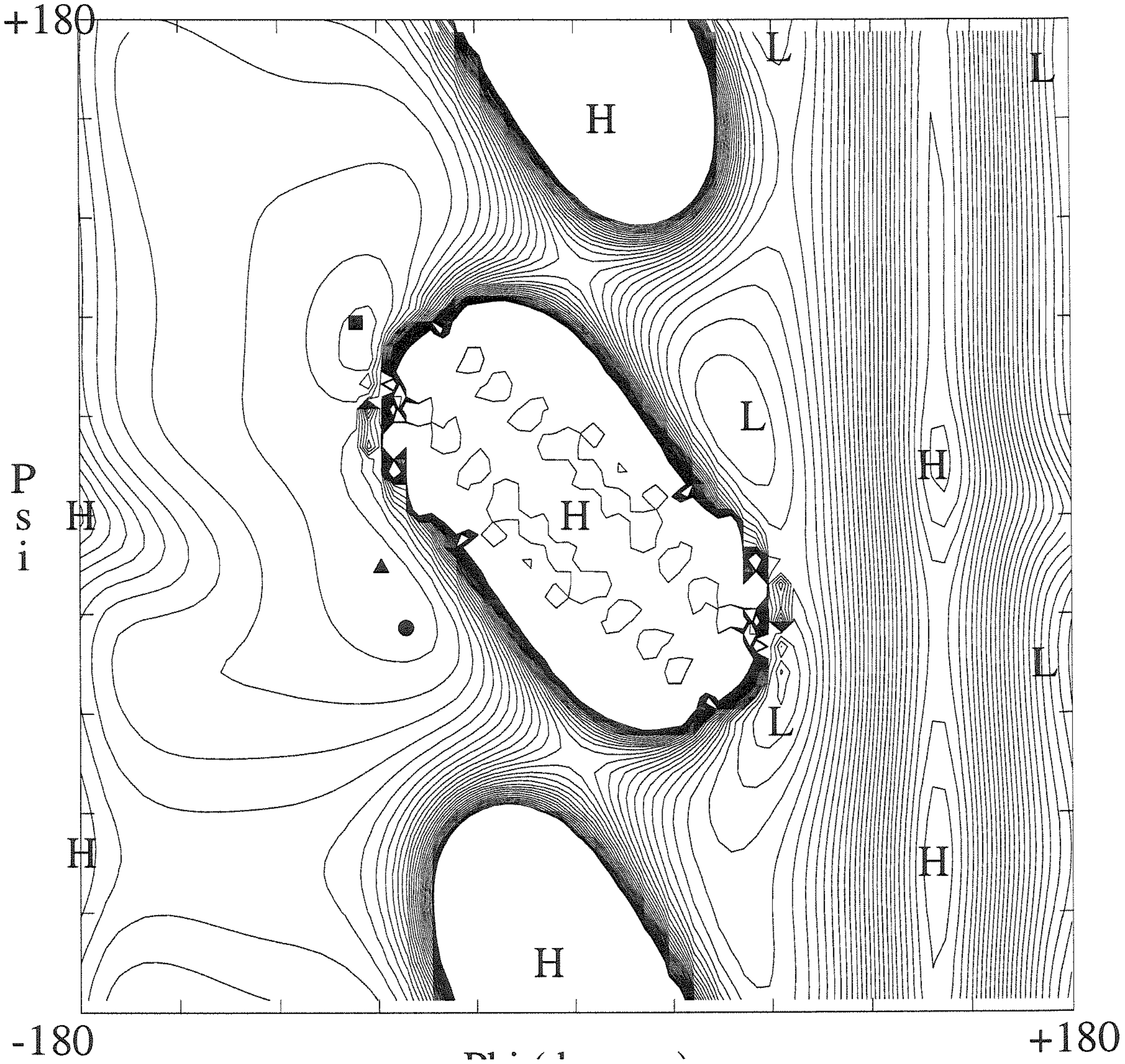


Figure 4.14: A Ramachandran contour plot of the energy of trialanine in AMBER (1984). "H"s represent energy maxima, and "L"s represent energy minima. Each contour line represents an energy change of 1 kcal/mole, and the line surrounding the black circle at (-62°, -41°) is at 0 kcal/mole. The black square identifies (-80°, +70°), in the C7 well. The black circle at (-62°, -41°) labels the average coordinates of α-helices in the PDB. The black triangle marks (-71°, -18°), the average location of 3_{10} helices in the PDB. A constant was subtracted from the energies to zero the energy of the α-helix. To conform to $(Ala)_{20}$ simulation conditions, terminal residues had net neutral charges and nonbonded interactions were cutoff at 15 Å cutoff as described in Chapter 6, Section 6.2. Figure modified from NCAR Graphics.

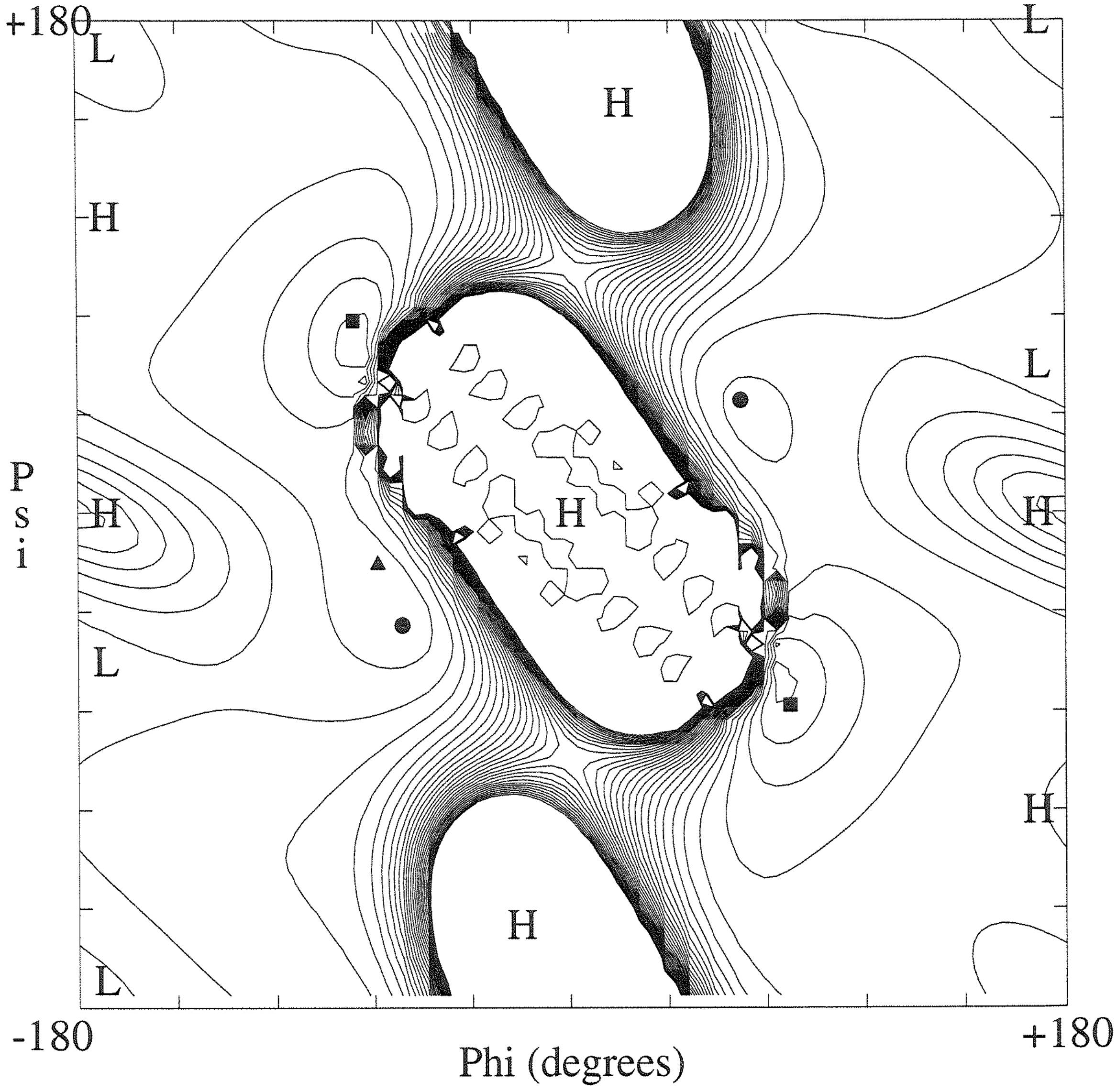


Figure 4.15: A Ramachandran contour plot of the energy of triglycine in AMBER (1984). "H"s represent energy maxima, and "L"s represent energy minima. Each contour line represents an energy change of 1 kcal/mole, and the line surrounding the black circle at (-62°, -41°) is at 0 kcal/mole. The black square identifies (-80°, +70°), in the C7 well. The black circle at (-62°, -41°) labels the average coordinates of α-helices in the PDB. The black triangle marks (-71°, -18°), the average location of 3_{10} helices in the PDB. The C7_G and the α_G region are also marked by a square at (80°, -70°) and a circle at (62°, 41°). A constant was subtracted from the energies to zero the energy of the α-helix. To conform to $(\mathrm{Gly})_{20}$ and $(\mathrm{Ala})_{20}$ simulation conditions, terminal residues had net neutral charges and nonbonded interactions were cutoff at 15 Å cutoff as described in Chapter 6, Section 6.2. Figure modified from NCAR Graphics.

Chapter 5 Discussion

5.1 Conclusions of the kinetics

5.1.1 $i, i+2$ HBs

The C7 conformation appears prominently in all simulations tested from 300K to 500K, whether on polyalanine or polyglycine.

Figure 5.1 on p. 107 illustrates the position of the C7 region with respect to predominate structures in the Protein Data Base (PDB) (Bernstein *et al.,* 1977). According to PROCHECK (Laskowski, 1993), the C7 conformation lies in an "allowed" region, a region where good quality X-ray structures do not expect to place more than 10% of their residues. In other words, protein residues do not favor this C7 conformation. The simulations did not favor it either—83% moved out of it and turned into a helix. The other 17% formed disordered balls whose (ϕ, ψ) angles mostly moved out of the C7 region to sample other parts of the Ramachandran plot. However, the C7 region showed relatively high density, presumably because AMBER has an energetic minimum at (~-80°, ~+60°).

Although the $(\phi, \psi \approx -80°, +70°)$ orientation is not highly favored by native protein structures at equilibrium, $i, i+2$ HBs have been observed both in small peptides in nonaqueous solutions and in crystallographic structures of proteins.

$i, i+2$ HBs exist in small peptides in nonaqueous solutions, presumably because each peptide can form stronger intramolecular HBs to itself than intermolecular bonds to the solvent. For example, dialanine forms C7 HBs in carbon tetrachloride (Avignon *et al.*, 1969; Bystrov *et al.,* 1969).

According to Milner-White (1990), there are numerous, weak $i, i+2$ HBs in the PDB, and he calls the ones that do not reverse the direction of the protein chain

Figure 5.1: p. 108. Comparison of the (ϕ, ψ) angles of predominant configurations in polyalanine simulations to the (ϕ, ψ) angles of structures in the Protein Data Base (PDB) (Bernstein, 1977). Figure on p. 107.

The shading. The most darkly shaded regions (regions A, B, and L) are where good quality X-ray structures expect to place over 90% of their residues. Less densely shaded regions represent less favorable regions for residues. (Plot modified from PROCHECK (Laskowski *et al.*, 1993).) The "P" and the "N" represent the positions of regular parallel and antiparallel β-sheets, respectively.

The squares and the pentagon. The filled squares represent a polyalanine in the C7 configuration. The white square is approximately at the center of the C7 region in our simulations, i.e., at $(\phi, \psi) \approx (-80°, +75°)$. The pentagon approximately encompasses the (ϕ, ψ) angles of $i, i+2$ HBs in the Protein Data Base (Milner-White, 1990; Bernstein *et al.*, 1977). See Section 5.1.1, $i, i+2$ kinetics.

The triangles. The black triangles represent residues in 3_{10} helices. The white triangle represents the average position of residues in 3_{10} helices in the PDB (Barlow & Thornton, 1988).

The circles. The small empty circles represent a real α-helix. The small white circle represents the mean position of α-helices in proteins (Barlow & Thornton, 1988). Finally, the large empty circle defines an α-helical residue in this simulation. A residue was determined to be α-helical if its (ϕ, ψ) angles lay within the pictured circle with radius 30°, centered at $-57°, -47°$. (See Section 6.7, Definition of Helix Formation.) The small, black circle to the upper left of the small white circle has no special significance; the black circle is actually a superposition of an unfilled circle and two triangles.

The sources of the (ϕ, ψ) angles. The empty squares originate from the polyalanine structure at 3ps in Figure 2.1. The shaded triangles are from 1lrv (Peters *et al.*, 1996). The small empty circles are from helix 3 of an enolase from *Saccaromyces cerevisiae* (PDB entry 4enl, Lebioda & Stec, 1989).

inverse γ-turns. They have ϕ, ψ angles in the pentagon in Figure 5.1. Frequently they lie in β-strands, and many of these lie in contiguous chains of γ-turns called compound γ-turns. The energies of most of the HBs in compound γ-turns are less than 1 kcal/mole per HB by the Kabsch and Sanders definition (1983). The amide-hydrogen bond and the carbonyl are often almost parallel. Thus, these interactions may not have genuinely overlapping orbitals. Nonetheless, Milner-White argues that these compound inverse γ-turns may, *en masse*, help stabilize both β-strands in β-

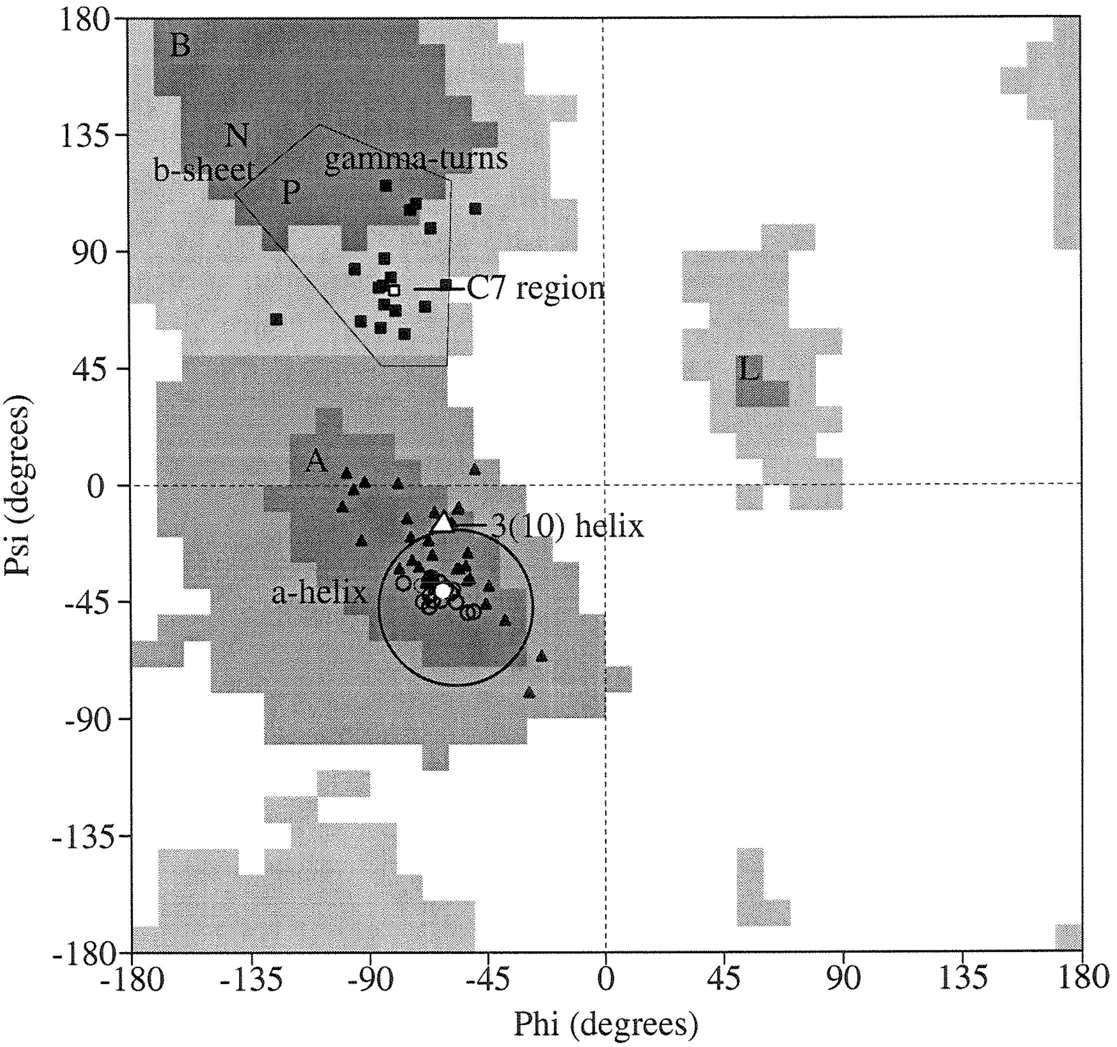


Figure 5.1: Comparison of the ϕ, ψ angles of predominant configurations in polyalanine simulations to the ϕ, ψ angles of structures in the Protein Data Base (PDB) (Bernstein, 1977). See caption p. 107.

sheets and β-strands that are on the way to becoming β-sheets.[1]

The $i, i+2$ HBs in these simulations appear significantly stronger than the γ-turns Milner-White catalogs from the Protein Data Base. The average energy of the simulated, C7 HBs is -1.5 kcal/mole ± 0.3 kcal/mole using the Kabsch & Sander definition

[1] To avoid confusing the $(\phi, \psi) \approx (-80°, +70°)$ HBs seen in our simulations with the weak, unoptimally oriented HBs in compound γ-turns, we call the $i, i+2$ HB the C7 structure. Not only is the boat shape of many of the simulated $i, i+2$ HBs similar to that of the IR data describing the C7 structure (Avignon *et al.*, 1969; Bystrov *et al.*, 1969), but most of the simulated $i, i+2$ HBs have (ϕ, ψ) angles closer to $(-80°, +75°)$ than to the β-sheet dihedrals.

(1983). The strength of the simulated $i, i+2$ HBs may be exaggerated with respect to the $i, i+2$ HBs in the Protein Data Base due to (1) the force field and/or (2) the lack of intermolecular hydrogen bonding opportunities for a molecule simulated without explicit solvent. The AMBER force field has a minimum at $(\phi, \psi) = (-80°, +60°)$, in the center of the C7 region. (See Figures 4.14 and 5.1.) In contrast, analyses of protein crystal structures in the PDB suggest that proteins more commonly place the dihedral angles of β-sheets and other structures in the "B" region of Figures 4.2 and 5.1 (Laskowski *et al.*, 1993). This may indicate that the AMBER force field is biased towards the C7 region over the "B" region. Thus, the force field may overestimate the importance of $i, i+2$ HBs and the C7 region, but the simulations still indicate they play a role in the denatured states. Perhaps further experimental work on the denatured states will elucidate the prevalence and strength of $i, i+2$ HBs in proteins. In addition, perturbing the energy of the C7 well could influence the kinetics of helix formation, if the C7 configuration is an intermediate for a large percentage of folding helices.

5.1.2 $i, i+3$ and $i, i+5$ intermediates

We observed many $i, i+3$ and $i, i+5$ HBs during the nucleation and propagation of α-helices. Formation of $i, i+3$ and/or $i, i+5$ HBs often preceded the formation of $i, i+4$ HBs. The $i, i+3$ HBs and $i, i+5$ were sometimes interspersed in the middle of α-helical segments but usually on the edges of them. Sometimes stretches of 3_{10} helix were as long as ten contiguous residues.

The observations on 3_{10} helices and $i, i+3$ HBs are consistent with other research. Sung found that $i, i+3$ HBs formed during helix propagation and at the fraying ends of helices during his Monte Carlo simulations of $(\mathrm{Ala})_{16}$ using AMBER (Weiner *et al.,* 1984) at 274K (Sung, 1994). Floriano *et al.* (1997) find that during pressure denaturation of myoglobin, HBs in α-helices convert from $i, i+4$ HBs to $i, i+3$ and then to $i, i+2$ HBs. Similarly, Hirst & Brooks found that $i, i+3$ HBs replaced $i, i+4$

HBs during unfolding simulations of apomyoglobin (1995).

In contrast to the $i, i+3$ HBs, $i, i+5$ HBs never formed π-helices (composed of $i, i+5$ HBs) longer than six contiguous residues. Also, $i, i+5$ HBs formed less frequently than $i, i+3$ HBs.

Of course, as the polymers matured into solidly α-helices, the $i, i+3$ and $i, i+5$ HBs disappeared. This is not surprising since both 3_{10} and π-helices are energetically less stable in AMBER than α-helices, and π-helices are less stable than 3_{10}.

5.1.3 Propagation is rate-limiting

Helix-coil transition theory assumes that nucleating a helix (i.e., forming the first helical, $i, i+4$ HB) is less probable than propagating a helix (i.e., forming a subsequent $i, i+4$ HB adjacent to the first one). In contrast, we find that nucleating a segment of helix is more probable than propagating it.[2]

This agrees with the results of Daggett *et al.* (1991) and Sung (1994). Daggett *et al.* found many short helical segments in their MD simulation of an $(\text{Ala})_{20}$ in AMBER (1986) at 400K. At the end of the 4 ns simulation, 45% of the polyalanine structures had more than one helical segment. The average number of helices per peptide was 1.5, and the maximum number of helices per peptide was 4 (Daggett *et al.*, 1991). In contrast, helix-coil transition theory predicts that the polyalanine would have to be 153 residues long before an ensemble at equilibrium will have an average of 1.5 helices per polymer molecule at 0K.[3] In concert with Daggett *et al.*, Sung found

[2] Note that helix-coil transition theory was developed for infinitely long polymers, not eicosamers (20-mers).

The simulation results do not dispute the ability of helix-coil transition theory to predict equilibrium helicities for different amino acid sequences. These predictions are made possible by measuring the average helicities for a multitude of sequences, fitting the results to theory, and getting σ and s parameters for each amino acid. In combination with the experimentally derived parameters, the theory is very successful at predicting helix propensity. The simulations in this thesis only contradict the popular understanding that helix-coil transition theory assumes that nucleation is rate-limiting for the formation of α-helices from 10–20 residues long.

[3] This estimate was derived using Equation 10 from Qian & Schellman (1992), Equations 20-52, 20-55, and 20-61 from Cantor & Schimmel (1980), and the parameters v and s from Chakrabartty *et al.* (1994). Both Mathematica and Maple confirmed the result.

that "very short helical segments" of an $(\mathrm{Ala})_{16}$ "were often not very stable" in his Monte Carlo simulations in AMBER (Weiner *et al.,* 1984) at 274K (Sung, 1994). If propagation were more probable than nucleation, the short helical segments would lengthen instead of uncoiling, and we would only observe long helices or completely unfolded polymers.

The recent simulations by Sung and Wu confirm that nucleation is more probable than propagation. Sung and Wu carried out MD on a polyalanine-based peptide with three glutamines using a modified AMBER force field (Sung & Wu, 1996). Fifty to 60% (50%–60%) of the peptide conformations had only one helical segment; 17–30% had two segments; and less than 2% had three or more.

5.1.4 Nonlocal hydrogen bonds retard helix formation

Because I found that helix formation involves a competition between the formation of local and nonlocal HBs, there should be an optimal polymer length for helix formation. This is consistent with the conclusions of helix-coil transition theory, and we plan to test these conclusions. Note that helix-coil transition theory applies to equilibrium ensembles while our results derive from kinetic behavior.

Since β-sheets consist mostly of nonlocal HBs, I expect β-sheet formation to follow principles different from those governing α-helix formation.

5.1.5 The "folding funnel" is extraordinarily rugged

Because I find that $(\mathrm{Ala})_{20}$ gets trapped in nonhelical balls of nonlocal HBs from which it does not escape within the simulation time, I do not see evidence for a perfectly smoothly draining "folding funnel" for $(\mathrm{Ala})_{20}$. Recall that some $(\mathrm{Ala})_{20}$ trajectories reduced their internal energies by compacting and forming energetically favorable, nonlocal HBs. The nonlocal HBs often trapped the $(\mathrm{Ala})_{20}$ in nonhelical conformations for the duration of the simulation time. Thus, in funnel parlance, the $(\mathrm{Ala})_{20}$ frequently fell into mountain valleys, from which it could not escape in 500 ps. The valleys were deep enough that 17% of the 155 runs did not form an α-helix in 500 ps. See Section 4.3, p. 4.3.

5.2 Limitations in the calculations

5.2.1 Solvent

One of the most serious approximations in these simulations is the lack of an explicit, aqueous solvent. However, the AMBER force field was developed for implicit solvent calculations, and it partially compensates for this lack (Weiner *et al.*, 1988). Explicit solvent can be included in the simulations, but it increases the time scale for the molecular dynamics (MD).

There are no experiments reported on polyalanine in water because it is not sufficiently soluble. Nonetheless, the simulated polyalanine dissolves in simulated water because the single polymer has no other choice.

5.2.2 Temperature

These simulations were run at a dynamic temperature of 450K, indicating that the average thermal energy of $(\text{Ala})_{20}$ is $\frac{1}{2}gkT = 53$ kcal/mole, where the number of degrees of freedom, g, is 59. This does not necessarily correspond to the laboratory temperature of 177°C since NEIMO keeps the bonds and angles frozen (and gives their motions zero temperature). This increases the torsional barrier and hence increases the effective temperature required for conformational transitions.

5.2.3 Time scale

The estimated half-live for α-helix formation in $(\text{Ala})_{20}$ was 209 ps, and the average time was 147 ps for the 129 helix-forming simulations of $(\text{Ala})_{20}$. However, this does not necessarily mean that $(\text{Ala})_{20}$ forms an α-helix in either 147 ps or 209 ps of real time. These numbers are likely to be lower limits on experimental time scales because eliminating the bond and angle motions in NEIMO decreases internal friction. In contrast to a NEIMO simulation, an experimental trajectory going over a conformational barrier will vacillate at the barrier due to interactions with explicit solvent and bond and angle motion. Thus, it is difficult to estimate the effect of friction on this time scale, but a factor in the range of 10 to 1000 seems plausible. This suggests that helix formation might take from 1.5 to 210 nanoseconds.

Other research also concludes that helix formation occurs on the nanosecond time scale. Brooks (1996) predicted ~50 ns for α-helix formation, based on solutions of the rate laws of helix formation derived from helix-coil transition theory. (See Section 1.5.2, p. 31.) Ballew *et al.* (1996) measured a possible helix-coil transition occurring at ~250 ns as apomyoglobin folds. Dyer *et al.* (1996) calculated ~50 ns for helix folding after observing helix unfolding in apomyoglobin. (See Section 1.5.1, p. 1.5.1.)

5.3 Helix formation

Despite the above caveats, these simulations reproduce many experimental facts. Most important, the simulations show that polyalanine forms a stable, α-helix while polyglycine does not.

5.3.1 Natural polyalanine forms an α-helix

It is well-known that alanine is an excellent helix former. Fiber diffraction studies of poly-L-alanine indicate that it can form an α-helix (Brown & Trotter, 1956). Crystallographic studies have also observed a polyalanine α-helix embedded in a protein: A decamer of polyalanine was mutagenically incorporated in T4 lysozyme, and the decamer formed an α-helix in the folded protein (Heinz *et al.,* 1992). Helix-coil transition theory, based on the circular dichroic (CD) signals of known polypeptide sequences, predicts that an $(\text{Ala})_{20}$ in water at 0°C would be 77% helical.[4]

There have been recent discussions of whether strong CD signals at 222 nm indicate the presence of α- or 3_{10} helices (Miick *et al.*, 1992, 1995). 3_{10} helices are helices composed of $i, i+3$ HBs. The average (ϕ, ψ) angles of 3_{10} helices in the crystallographic data base is $(-71°$, -18°). In comparison, the average (ϕ, ψ) angles of α-helices in the PDB is (-62°, -41°) (Barlow & Thornton, 1988). See Figure 5.1 on p. 107. Although the (ϕ, ψ) angles are very close, recent research suggests that equilibrium helices are not 3_{10} helices but α-helices (Tirado-Rives *et al.*, 1993; Smythe *et al.*, 1995). We will discuss in Section 5.1.2 whether 3_{10} helices could be intermediates on the pathway to α-helices.

5.3.2 Simulated polyalanine forms a stable α-helix

After formation, the persistent α-helices that formed in 83% of our simulations had an average helix content of 91% ± 10% (1 σ standard deviation. I defined 100% helical content as the maximum percentage of helical residues circular dichroism can detect in a polymer (Dyson, 1991). This is the percentage of residues in a block of three consecutive residues whose (ϕ, ψ) angles are within 30° of the ideal α-helical configuration, $(-57°, -47°)$ (Daggett *et al.*, 1991; Sung & Wu, 1996). Our results compare well with Daggett *et al.* (1991) who found an overall helicity of 84% (using the same definition) during equilibrium MD of $(\text{Ala})_{20}$ with the AMBER (1986) force field at 400K.

[4] This estimate was derived using Equation 10 from Qian & Schellman (1992), Equations 20-52, 20-55, 20-56, 20-57, and 20-59 from Cantor & Schimmel (1980), and the parameters v and s from Chakrabartty *et al.* (1994). Maple was used to solve the equations both assuming and not assuming that n, i.e., the number of residues in the polymer, is large enough to simplify the mathematics (Equations 20-56 and 20-57 in Cantor & Schimmel). Maple produces 77% helicity not assuming large n and 76% helicity assuming it. This agrees with the 76% result from the program "helix" by Schellman *et al.* (1993).

5.3.3 Polyglycine does not form a helix

As discussed above, we can conclude that entropy prevents polyglycine from forming α-helices more often. None of the simulations of polyglycine formed a full α-helix,[5] although the helical state is more favorable energetically. Our results are similar to those of Sung (1994) using Monte Carlo simulations on a $(\text{Gly})_{16}$ in a modified AMBER force field. The α-helical state of Sung's polyglycine was more stable energetically than the states the polymer sampled during the MD, yet his simulation of polyglycine did not form an α-helix while his simulations of polyalanine did under analogous conditions.

[5] The longest α-helix that $(\text{Gly})_{20}$ formed (pictured on p. 101 in Figure 4.11) was shorter than any of the helices the $(\text{Ala})_{20}$ simulations formed.

Chapter 6 Methods of Calculation

6.1 The force field

All simulations used the AMBER force field (Weiner *et al.,* 1984). This force field treats the methyl group of alanine as a single united atom (with implicit hydrogen atoms) so that each alanine has six atoms. The heterohydrogens on electronegative atoms like N and O are treated explicitly to allow for HBs. Partial charges from the AMBER force field gave each residue a net zero charge. This AMBER force field was developed for use with implicit solvent calculations.

6.2 Nonbond interactions

The nonbond interactions (Coulomb and van der Waals) were calculated out to a radius of $R_{cut} = R_{inner} = 14.0$Å from each atom, and then smoothly decreased to zero at $R_{outer} = 14.5$Å using a cubic spline function. However, the nonbond list was constructed based on $R_{cutoff} = 15.0$ Å. The dielectric constant was linearly dependent on distance in order to simulate solvent. The dielectric constant was $\epsilon(R) = \epsilon_0 * R$ with $\epsilon_0 = 1$.

6.3 NEIMO molecular dynamics

The Newton-Euler Inverse Mass Operator (NEIMO) MD method was used for inertial coordinate dynamics. NEIMO fixes all valence bonds and angles, leading to motion only in the three backbone dihedral angles ϕ, ψ, and ω. This reduces the eicosamer (20-mer) from

$$18N - 6 = 354$$

to

$$3(N-1)+2=59$$

degrees of freedom where $N = 20$. Of those degrees of freedom, only $3(N-1)$ were ϕ, ψ, and ω backbone dihedrals. The other two degrees of freedom were terminal dihedrals at residues 1 and 20.

NEIMO dramatically reduces the folding time of a protein. Fluctuations in bonds and angles from their equilibrium positions create huge forces in a protein while torsional modes create much smaller forces. NEIMO shortens folding time by eliminating the strong forces inherent in bond and angle motion and allowing the more smaller torsional forces to guide the protein to the native state.

The process of protein folding is somewhat analogous to a steel ball in a concrete, half-cylinder trough or track that runs down a mountain. If the ball is put at the upper edge of the track at the top of the mountain, it will oscillate quickly between the opposite edges of the trough, running first down one side and then up the other. Very slowly the ball will start falling down the mountain. This is because the gradients, and hence the forces, on the sides of the trough are much stronger than the gradient, and the force, pointing down the mountain. However, the ball will race speedily down the track to the bottom of the mountain if the ball is put in the base of the trough at the top of the mountain instead of on an upper edge. Placing the ball at the base of the trough eliminates the high frequency oscillations impeding the descent of the ball. Analogously, NEIMO allows the protein to find its native state quickly because the algorithm eliminates strong bond and angle forces and their resulting high frequency motions.

Bond and angle forces thus impede protein folding. The bond and angle forces create a high friction Brownian system for the torsions because these bond and angle forces are so much larger than the torsional ones. This leads to very slow diffusion in the torsional degrees of freedom. When the dynamics is run with only torsional degrees of freedom, these high frictional Browning terms are eliminated. Friction still exists from the nonbond interactions and from kinematic coupling between different torsions. However, the net effect is that the time scale for folding may be decreased by orders of magnitude.

Constrained torsional dynamics reduces folding time in another way too. Because high frequency motion is eliminated, the molecule is not expected to change positions as quickly, and longer time steps can be used. NEIMO allows stable dynamics for timesteps as large as 20 fs on $(\text{Ala})_{20}$. In contrast, traditional Cartesian dynamics is stable only with time steps of 1–2 fs. (Jain *et al.*, 1993; Mathiowetz *et al.*, 1994).

Previously existing constrained torsional dynamics algorithms are computationally expensive. Conventional (Cartesian) MD solves the equations of motion as

$$m_a \ddot{x}_{ai} = F_{ai}$$
$$\ddot{x}_{ai} = \tfrac{1}{m_a} F_{ai}$$

The subscripts a and i refer to the atom a and the coordinate i. Solving for the acceleration is easy because inverting the mass, a scalar, is easy. However, in torsional dynamics, the equation of motion is

$$I_{\alpha\beta} \ddot{\theta}_\beta = T_\alpha$$
$$\ddot{\theta}_\beta = I^{-1}_{\alpha\beta} T_\alpha$$

T is a vector of torques; $\ddot{\theta}$, the angular acceleration, is also a vector. α and β refer to the number of degrees of freedom. To solve for the angular accelerations, the moment of inertia matrix must be inverted, and this is computationally expensive. In traditional algorithms, the cost scales as N^3 where N is the number of degrees of freedom. But, by using a spatial operator algebra, Jain *et al.* (1993) determined how to make the inversion of the mass matrix scale linearly with N, i.e., as N.

Unless otherwise noted, simulations were run for 500 ps with a timestep of 0.010 ps. Initial momenta were selected randomly from a Gaussian (Boltzmann) distribution.

Nosè thermostat. To properly describe the canonical distribution of conformations at constant temperature, volume, and number of particles, we used the Nosè-Hoover thermostat formulation (Vaidehi *et al.,* 1996).

6.4 The Polyalanine model

Eicosamers (20-mers) of polyalanine were built using the standard Peptide Builder of POLYGRAF with standard geometric parameters. The amino terminus was protonated but given a net neutral charge to simulate a long protein. The carboxy terminus was described as an unprotonated carboxylate with a net zero charge.

6.5 Temperature

See Section 5.2.2. Pilot NEIMO-Hoover simulations were carried out at different temperatures in the range of 300K to 500K to determine how thermal energy affects the probability of α-helix formation. The temperature affects the balance between kinetic effects trying to expand the polymer and potential energy effects trying to compact it. As discussed in Section 5.2.1, we used an implicit solvent. This modified the energy scale and, hence, we cannot be sure which dynamic temperatures would simulate a system at room temperature. Note that the molecule cannot decompose at these temperatures because the bonds and angles are rigid.

Temperatures between 300K and 375K did not appear to allow the polymer enough kinetic energy to break free of nonhelical conformations. Instead, the polymers stayed trapped in conformations with multiple, nonlocal HBs. Although a helix was more stable than the irregular conformations, the cold temperatures did not allow the polymer to break free.

At 500K, the kinetic energy in the polymer appeared too great to allow it to form the close, tight, strong HBs necessary for helix formation.

Helices formed with high probability from 400K to 450K, with the highest probability at 450K. Since our goal has been to elucidate the steps in helix formation, we selected 450K for our studies.

6.6 Optimal conditions

We carried out 155 simulations on polyalanine at 450K. Each run started with the fully extended conformation ($\phi = \psi = 180°$) but used a different set of random initial atomic velocities. Each set of velocities reflected a Maxwell-Boltzmann distribution about the assigned temperature. The polyalanine formed a helix within 500 ps in 83.2% of the runs (± 6.0% with 95% confidence).

6.7 Definition of helix formation

The polyalanine was declared a helix if the average over the last 5 ps of the percent helicity was greater than 75%. Percentage helicity was defined in one of two ways:

(a) as the percentage of residues which have either carbonyls or amides participating in $i, i+4$ HBs, or

(b) as the percentage of residues which are in a block of three consecutive residues whose (ϕ, ψ) angles are within 30° of the ideal α-helical angles $(-57°, -47°)$. Figure 5.1 (Laskowski *et al.*, 1993) illustrates the position of the 30° radius with respect to dominant conformations in the PDB (Bernstein *et al.*, 1977).

Both definitions resulted in the same two sets of helix-forming simulations and non-helix-forming simulations.

Chapter 7 References in the Simulation Chapters

AMBER, 1984. See Weiner *et al.*, 1984.

AMBER, 1986. See Weiner *et al.*, 1986.

Avignon, M., Huong, P.V., Lascombe, J. (1969) Étude, par Spectroscopie infra-rouge, de la Conformation de quelques Composés peptidiques modèles. *Biopolymers 8*: 69–89.

Ballew, R.M., Sabelko, J., Gruebele, M. (1996) Direct observation of fast protein folding: The initial collapse of apomyoglobin. *PNAS USA 93(12)*: 5759–5764.

Barlow, D.J. & Thornton, J.M. (1988) Helix Geometry in Proteins. *J. Mol. Biol. 201*: 601–619.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. (1977) The Protein Data Bank: A computer based archival file for macromolecular structures. *J. Mol. Bio. 122*: 535–542.

Brown, L. & Trotter, I.F. (1956) X-ray Studies of Poly-L-Alanine *Transactions of the Faraday Society 52*: 537–548.

Bystrov, V.F., Portnova, S.L., Tsetlin, V.I., Ivanov, V.T., Ovchinnikov, Y.A. (1969) Conformational studies of peptide systems: The rotational states of the NH-CH fragment of alanine dipeptides by nuclear magnetic resonance. *Tetrahedron 25*: 493–515.

Cantor, C.R. & Schimmel, P.R. (1980) "Conformational equilibria of polypeptides and proteins: The Helix-coil transition," in *Biophysical Chemistry: The Behavior of Biological Macromolecules*, vol. III, (San Francisco, W.H. Freeman & Comp.), pp. 1063–1066.

Chakrabartty, A., Kortemme, T., Baldwin, R.L. (1994) Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. *Protein Science 3*: 843–852.

Daggett, V., Kollman, P.A., Kuntz, I.D. (1991) A Molecular dynamics simulation of polyalanine: An analysis of equilibrium motions and helix-coil transitions. *Biopolymers 31*: 1115–1134.

Dyer, R.B., Williams, S., Woodruff, W.H. (1996) The Earliest events in protein folding: Helix dynamics in proteins and model peptides. *Abstracts of Papers of the American Chemical Society 212(2)*: 150.

Dyson, H.J. & Wright, P.E. (1991) Defining solution conformations of small linear peptides. *Ann. Rev. Biophys. Biophys. Chem. 20*: 519–38.

Elove, G.A., Chaffotte, A.F., Roder, H., Goldberg, M.E. (1992) Early steps in cytochrome c folding probed by time-resolved circular dichroism and fluorescence spectroscopy. *Biochemistry 31*: 6876–6883.

Floriano, W.B., Nascimento, M.A.C., Domont, G., Goddard, W.A. III (1997) Pressure effects on the structure of myoglobin. In progress.

Heinz, D.W., Baase, W.A., Matthew, B.W. (1992) Folding and function of a T4 lysozyme containing 10 consecutive alanines illustrate the redundancy of information in an amino acid sequence. *PNAS 89*: 3751–3755.

Hirst, J.D. & Brooks, C.L. III (1995) Molecular dynamics simulations of isolated helices of myoglobin. *Biochemistry 34*: 7614–7621.

Jaenicke, R. (1991) Protein folding: Local structures, subunits, and assemblies. *Biochemistry 30*: 3147–3161.

Jain, A., Vaidehi, N., Rodriguez, G. (1993) A Fast recursive algorithm for molecular dynamics simulations. *J. Computational Physics 106*: 258–268. The NEIMO-Hoover algorithm was incorporated into POLYGRAF by N. Vaidehi.

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

Kholodenko, Y., Volk, M., Gooding, E., Hochstrasser, R.M. (1996) Early events in protein folding studied by photoinitiation of alpha-helix formation in de-novo peptides. *Abs. of Papers of the American Chemical Society 212(2)*: 346.

Kiefhaber, T., Rudolph, R., Kohler, H.-H., Buchner, J. (1991) Protein Aggregation *in vitro* and *in vivo*: A quantitative model of the kinetic competition between folding and aggregation. *Bio Tech 9*: 825-829.

King, J. (1996) personal communication.

Laskowski, R.A., MacArthur, M.W., Moss, D.S., Thornton, J.M. (1993) PROCHECK: A Program to Check the Stereochemical Quality of Protein Structures. *J Appl. Cryst. 26*: 283–291.

Lebioda, L. & Stec, B. (1989) Crystal structure of holoenzyme refined at 1.9 Å resolution: Trigonal-Bipyramidal geometry of the cation binding site. *JACS 111*: 8511–8513.

Mathematica. (1993) Version 2.2 for the X Window System. Wolfram Research, Inc., 100 Trade Center Drive, Champagne, IL 61820-7237.

Mathiowetz, A.M., Jain, A., Karasawa, N., Goddard, W.A. (1994) Protein simulations using techniques suitable for very large systems: The Cell multipole method for nonbond interactions and the Newton-Euler inverse mass operator method for internal coordinate dynamics. *PROTEINS: Structure, Function, and Genetics 20*:227–247.

Miick, S.M., Martinez, G.V., Fiori, W.R., Todd, A.P., Millhauser, G.L. (1992) Short alanine-based peptides may form 3_{10}-helices and not α-helices in aqueous solution. *Nature 359*:653–655.

Miick, S.M., Martinez, G.V., Fiori, W.R., Todd, A.P, Millhauser, G.L. (1995) Correction to “Short alanine-based peptides may form 3(10)-helices and not alpha-helices in aqueous-solution ((1992) *Nature 359*: 653–655). *Nature 377*: 257.

Milner-White, J.E. (1990) Situations of gamma-turns in proteins: Their relation to alpha-helices, beta-sheets and ligand binding sites. *J Mol Biol 216*:385–397.

Mines, G.A., Pascher, T., Lee, S.C. (1996) Cytochrome *c* folding triggered by electron-transfer. *Chemistry & Biology 3(6)*: 491–497.

Mitraki, A., Fane, B., Haase-Pettingell, C., Sturtevant, J., King, J. (1991) Global suppression of protein folding defects and inclusion body formation. *Science 253*: 54–58.

NCAR Graphics (1989) Copyright by University Corporation for Atmospheric Research. Published by National Center for Atmospheric Research, Scientific Computing Division, P.O. Box 3000, Boulder CO, 80307-3000. Version 3.00 for UNIX.

Nölting, B., Golbik, R., Fersht, A.R. (1995) Submillisecond events in protein folding. *PNAS USA 92*: 10668–10672.

Pascher, T., Chesick, J.P., Winkler, J.R., Gray, H.B. (1996) Protein-folding triggered by electron-transfer. *Science 271*: 1558–1560.

Peters, J.W., Stowell, M.H.B., Rees, D.C. (1996) A leucine-rich repeat variant with a novel repetitive protein structural motif. *Nature Structural Biology 3(12)*: 991–994.

POLYGRAF/BIOGRAF. Molecular Simulations, Incorporated, San Diego, CA.

Qian, H. & Schellman, J.A. (1992) Helix-coil theories: A comparative study for finite length polypeptides. *J. Phys. Chem. 96*: 3987–3994.

Radford, S.E., Dobson, C.M., Evans, P.A. (1992) The folding of hen lysozyme involves partially structured intermediates and multiple pathways. *Nature 358*: 302–307.

Rock, R.S. & Chan, S.I. (1996) Synthesis and photolysis properties of a photolabile linker based on 3-methoxybenzoin. *J. of Organic Chemistry 61(4)*: 1526–1529.

Schellman, J., Qian, H., Rohl, C., Kortemme, T. (1993) FORTRAN program "hcontent.for" The N-capping and C-capping parameters of Chakrabartty *et al.*, 1994, were used.

Siobhan, M.M., Gary, V.M., Wayne, R.F., Todd, A.P., Milhauser, G.L. (1992) Short alanine-based peptides may form 3_{10}-helices and not alpha-helices in aqueous solution. *Nature 359*: 653–655.

Smythe, M.L., Nakaie, C.R., Marshall, G.R. (1995) Alpha-helical versus 3_{10}-helical conformation of alanine-based peptides in aqueous solutions: An Electron spin resonance investigation. *J Am Chem Soc 117*: 10555–10562.

Sung, S.-S. (1994) Helix folding simulations with various initial conformations. *Biophysical Journal 66*:1796–1803.

Sung, S.-S. (1995) Folding simulations of alanine-based peptides with lysine residues. *Biophysical Journal 68*:826–834.

Sung, S.-S. & Wu, X.-W. 1996. Molecular dynamics simulations of synthetic peptide folding. *PROTEINS: Structure, Function, and Genetics 25*:202–214.

Tannor, D.J., Marten, B., Murphy, R., Friesner, R.A., Sitkoff, D., Nicholls, A., Ringnalda, M., Goddard, W.A. III, Honig, B. (1994) Accurate First Principles Calculation of Molecular Charge Distributions and Solvation Energies from Ab Initio Quantum Mechanics and Continuum Dielectric Theory. *J. Am. Chem. Soc. 116(26)*: 11875–11882.

Tirado-Rives, J., Maxwell, D.S., Jorgensen, W.L. (1993) Molecular dynamics and Monte Carlo simulations favor the alpha-helical form for alanine-based peptides in water. *J. Am. Chem. Soc. 115*:11590–11593.

Truong, H.T., Pratt, E.A., Rule, G.S., Hsue, P.Y., Ho, C. (1991) Inactive and temperature-sensitive folding mutants generated by tryptophan substitutions in the membrane-bound D-lactate dehydrogenase of Escherichia coli. *Biochemistry 30*: 10722–10729.

Vaidehi, N., Jain, A., and Goddard III, W. A. (1996) Constant Temperature Constrained Molecular Dynamics- The Newton-Euler Inverse Mass Operator Method, *J. Phys. Chem. 100(25)*: 10508–10517.

Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Gio, C., Alagona, G., Profeta, S., Jr., Weiner, P. (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc. 106:*:765–784.

Weiner, S.J., Kollman, P.A., Nguyen, D.T., Case, D.A. (1986) An all atom force-field for simulations of proteins and nucleic acids. *J. of Computational Chemistry 7(2)*: 230–252.

Wetzel, R. & Chrunyk, B.A. (1993) "Mutational effects on inclusion body formation." In *Biocatalyst design for stability and specificity*, ed. by Himmel M.E. & Georgiou, G. (Washington, D. C.; American Chemical Society) pp. 116–125,

Winkler, J.R. & Gray, H.B. (1996) Response. *Science 274*: 629.

Yu, M.-H. & King, J. (1988) Surface amino acids as sites of temperature-sensitive folding mutations in the P22 tailspike protein. *J Biol Chem 263*: 1424–1431.

Appendix A Definitions of Basic Terms

A.1 Dihedral angles ϕ and ψ

Two angles that give proteins their characteristic shapes are the ϕ and ψ dihedral angles along the protein backbone. The cosine of ϕ is the dot product of the O=C—N bond and the vector along the CHR—NH bond, and the cosine of ψ is the dot product of the angle between the vector along the next HN–CHR and the vector along the next O=C—NH bond. See Figure A.1. Because the dihedral angles ϕ and ψ govern the bend of the backbone of the protein, they determine how the protein curves, coils, and folds. Each type of secondary structure has its own characteristic ϕ, ψ angles. Thus, Ramachandran plots, graphs with ϕ on the x axis and ψ on the y axis, are extremely useful.

A.2 Molecular dynamics (MD)

Molecular dynamics is a type of simulation. A molecule is placed in a force field determined by some equations of motion. The force on each atom is calculated and then the accelerations. By integrating the acceleration over a timestep, often 1 fs, the new positions of the atoms are calculated. Then, the forces and accelerations on the atoms in the new positions are calculated, and the entire cycle begins again.

Molecular dynamics is considered a more realistic method of simulation than, for example, Monte Carlo dynamics. MD simulates the way nature moves molecules: nature has them move in response to forces. As long as the force field accurately mimics nature, the simulation should too. The validity of the force field is considered the biggest problem in MD.

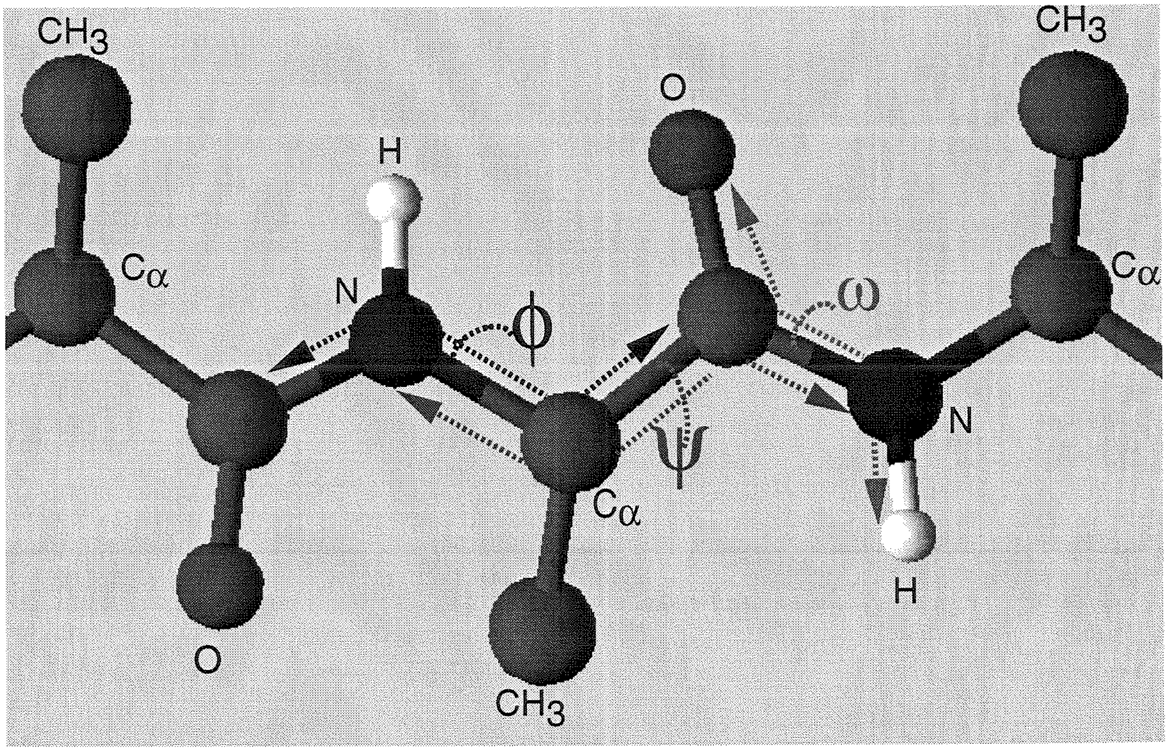


Figure A.1: Definition of backbone dihedral angles ϕ, ψ, and ω: A dipeptide of polyalanine (plus an extra C_α and methyl) with the ϕ, ψ, and ω angles marked. ϕ is the angle between the vector along the O=C—N bond and the vector along the CHR—NH bond, and ψ is the angle between the vector along the next HN–CHR and vector along the next O=C—NH bond. The cosine of the dihedral ω is the dot product of the vector along the C=O bond and the N—H bond.

A.3 Proteins

Proteins are polymers found in biological systems. Any protein polymer is made up of monomer units called amino acids. The general formula for an amino acid is:

$$H_3N—CHR—COOH$$

Usually in solution it looks like this:

$$^{+}H_4N—CHR—COO^{-}$$

The COO^{-} gives it the name "acid," and the NH_3 gives it the "amino." Biological systems rely on about 20 amino acids to produce their proteins. In this thesis, you only have to remember two of them: alanine and glycine. See Figure A.2. Alanine has a methyl (—CH_3) group for the "R" above. Thus, alanine is chiral, i.e., bending it to the right is different from bending it to the left. Glycine has only an H for the "R," and thus, it is achiral.

Each residue is one monomer unit in the protein chain. Once an amino acid has become part of a protein chain, it is technically no longer an "amino acid" because it lacks both the amino and the acid. It is then called a "residue" or, sometimes, an "amino acid residue." There are 20 residues in each of the proteins that I simulate, i.e., in $(Ala)_{20}$ and $(Gly)_{20}$.

In the polymer, the residues look like this:

$$^{+}H_4N—CHR—C{=}O—NH—CHR—C{=}0—NH-\ldots—COO^{-}$$

N–terminus ⟶ C–terminus

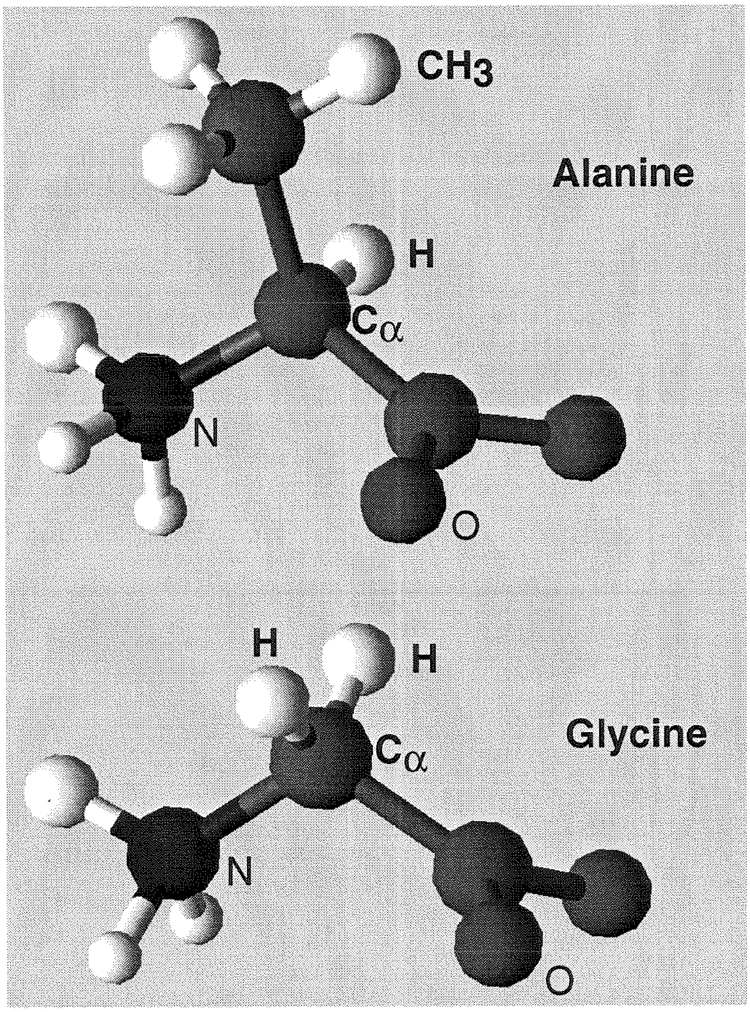


Figure A.2: Comparison of the structures of L-alanine and L-glycine.

Appendix B C7 vs γ-turn terminology

$i, i+2$ HBs have been called both γ-turns and C7 structures (Milner-White, 1990; Avignon; Bystrov). We choose to call the structures we see in the first few picoseconds of simulation, occurring at $(\phi, \psi) \approx (\sim\text{-}80°, \sim+75°)$, C7 structures instead of γ-turns because the C7 configuration (Avignon *et al.*, 1969; Bystrov *et al.*, 1969) is specific only to very strong HBs in that region of the Ramachandran diagram. The boat-shaped structure characteristic of the strongest simulated $i, i+2$ HBs matches that of the C7 conformation.

In contrast to the C7 nomenclature, γ-turns encompass a large variety of $i, i+2$ turns. First, there are both inverse γ-turns, where the direction of the chain does not change, and "classic" γ-turns, which usually occur at the end of beta-hairpins (Milner-White, 1988). Second, inverse γ-turns occur anywhere within the pentagon in Figure B.1. These multiple, contiguous $i, i+2$ HBs are found in the middle of some β-strands in β-sheets. The Kabsch and Sanders (1983) definition of a hydrogen bond,

$$E = q_1 q_2 (\frac{1}{r_{0N}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} - \frac{1}{r_{CN}}) \times f$$

where $q_1 = 0.42e$, $q_2 = 0.20e$, e is the unit electron charge, r_{AB} is the distance from A to B in angstroms, and f is 332 $\frac{\text{kcal}}{\text{mole}}\frac{\AA}{e^2}$, predicts interaction energies less than 1 kcal/mole per HB. Not only are these very weak HBs, but they may not even be hydrogen bonds at all. The carbonyl bond of residue i and the amide nitrogen bond of residue $i+2$ are sometimes nearly parallel. In such cases, it is hard to see how the orbitals of the amide proton and the carbonyl oxygen could overlap. Since the Kabsch and Sanders definition of a HB does not take geometries into account, it does not reject this interaction. Milner-White agrees that these interactions are very weak, but he describes them because their electrostatic attractions may be important in

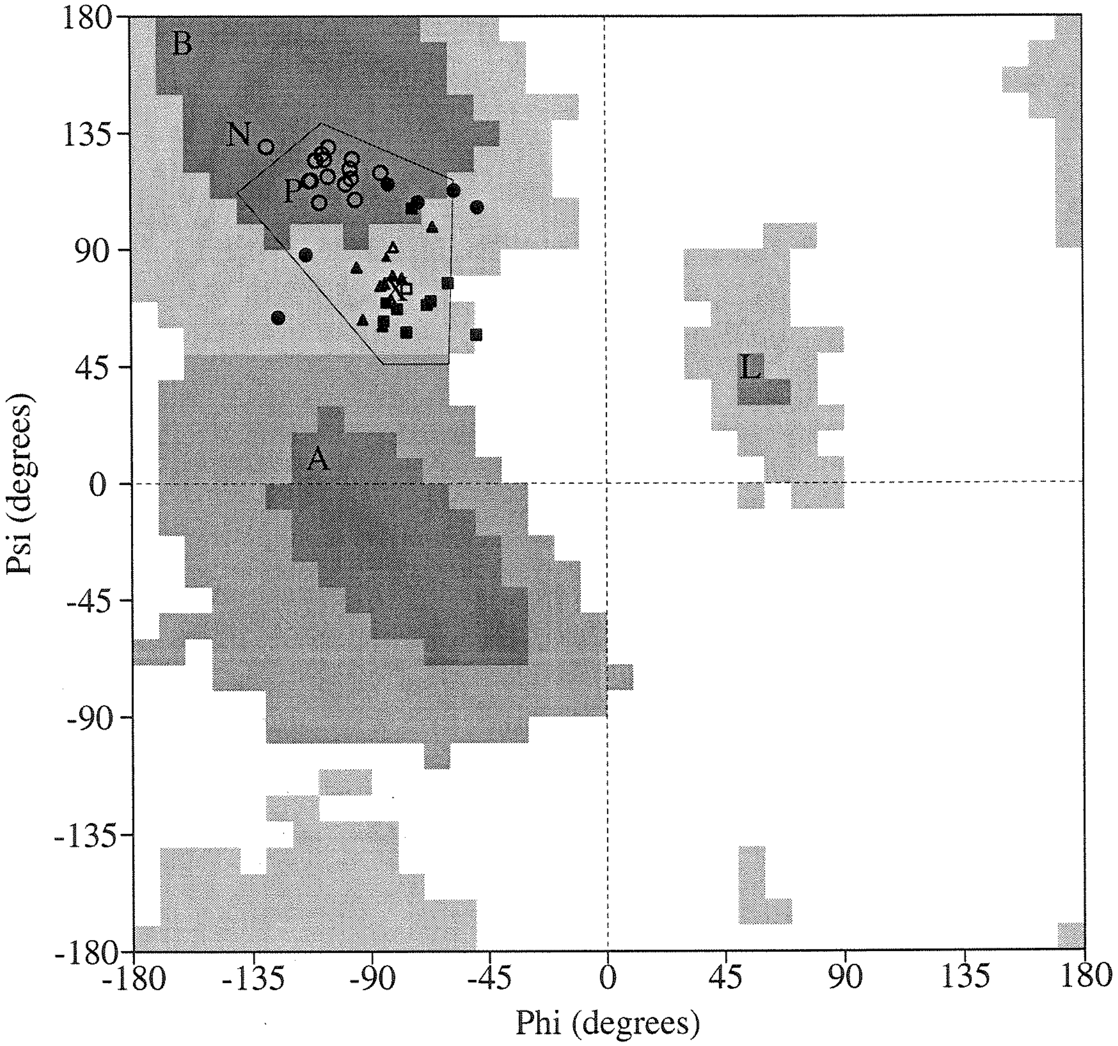


Figure B.1: A comparison of the energies and dihedral angles of typical $i, i+2$ HBs from the PDB and the $i, i+2$ HBs from the first few picoseconds of the simulations of polyalanine. Experimental residues are shown as unfilled polygons whereas simulated dihedrals are shown in filled shapes. Circles represent $i, i+2$ HBs with interaction energies weaker than -1 kcal/mole; triangles represent interactions weaker than -2 kcal/mole, and squares weaker than -3 kcal/mole. The "P" and the "N" represent the positions of regular parallel and antiparallel β-sheets, respectively. The "X" designates the (ϕ, ψ) coordinates $(-80°, +75°)$. The $i, i+2$ hydrogen bonding residues from the PDB were: residues 31–36 and 102 from structure 3rp2, 58–62 from 3dfr, 70–74 from 2pab, and residue 211 from 2fb4 (Milner-White, 1990). The simulated residues represented residues 2–19 at 3.0 ps of trajectory pAnc-450K-ad-15A. The background of this figure is courtesy of PROCHECK (Laskowski *et al.*, 1993). For an explanation of the shading, read the caption to Figure 5.1, p. 108.

stabilizing a β-strand in the process of forming a β-sheet (Milner-White, 1990).

We choose to call the $i, i+2$ HBs in our simulations C7 to distinguish them from inverse γ-turns in β-sheets. C7 HBs are stronger than most inverse γ-turns and are more prevalent in our simulations. Figure B.1 plots a sampling of $i, i+2$ HBs Milner-White found in the PDB and against the ϕ, ψ angles from all of the residues after 3 ps of one simulation. The PDB residues are shown as unfilled circles, triangles, and squares while the simulated residues are depicted as the analogous, filled polygons. Circles represent $i, i+2$ HBs weaker than -1 kcal/mole; triangles HBs weaker than -2 kcal/mole, and squares weaker than -3 kcal/mole. Regular parallel and antiparallel β-sheets lie at the "P" and the "N," respectively. The figure demonstrates that the strongest $i, i+2$ HBs fall closer to the ($\sim$-80°, $\sim$+75°) region than the β-sheet region and that most of the simulated HBs are strong.

Appendix C The Search for Optimal Folding Conditions

Conditions were sought that would be more conducive to helix formation. It is easier to study successful helix formation if it happens with high probability. A trajectory that forms a helix under conditions unlikely to produce a helix may be suspect because it may have formed merely because random momenta predominated over the force field. However, if the simulation conditions usually produce a helix, real forces must drive polyalanine there and helix formation could not be an improbable, random outcome.

I systematically tried a range of simulation temperatures and cutoff radii for nonbonded interactions. In addition, some pilot simulations explored the role of the dielectric constant and of charges on the termini. As stated in Chapter 6, the best conditions for helix formation from the extended state appeared to be 450K, 1ϵ(r) dielectric constant, 15 Å nonbonded cutoff radius, and neutral charges at the terminal residues.

Table C.1, p. 134. The effect of varying the temperature from 300K to 500K on the probability of helix formation. Electrostatic cutoff radii from 6 Å to 10 Å were used. At each set of conditions, each simulation used a different seed to randomize the initial velocities. Although the first run of each set of conditions used seed 12345, the rest of the runs used seeds unique in that set of conditions. See Table C.1.4 for the seeds. The column entitled "Percentage Helix Formation" does not attempt to describe the level of confidence in the percentage.

Table C.1: The Effect of varying the temperature from 300K to 500K on the probability of helix formation.

300K				**350K**				**375K**				**400K**			
Non-Bond Cutoff (A)	No. of Helices Formed	No. of Runs	% Helix Formation	Non-Bond Cutoff (A)	No. of Helices Formed	No. of Runs	% Helix Formation	Non-Bond Cutoff (A)	No. of Helices Formed	No. of Runs	% Helix Formation	Non-Bond Cutoff (A)	No. of Helices Formed	No. of Runs	% Helix Formation
6A	0	1	0%	**6A**	0	3	0%	**6A**	0	2	0%	**6A**	0	3	0%
8A	1	1	100%	**8A**	0	1	0%	**8A**				**8A**	0	2	0%
9A	0	1	0%	**9A**	0	3	0%	**9A**	0	3	0%	**9A**	3	6	50%
10A	1	1	100%	**10A**	0	1	0%	**10A**				**10A**	1	2	50%

425K				**450K**				**475K**				**500K**			
Non-Bond Cutoff (A)	No. of Helices Formed	No. of Runs	% Helix Formation	Non-Bond Cutoff (A)	No. of Helices Formed	No. of Runs	% Helix Formation	Non-Bond Cutoff (A)	No. of Helices Formed	No. of Runs	% Helix Formation	Non-Bond Cutoff (A)	No. of Helices Formed	No. of Runs	% Helix Formation
6A	0	3	0%	**6A**	0	4	0%	**6A**	0	1	0%	**6A**	0	3	0%
8A	1	2	50%	**8A**	1	5	20%	**8A**	1	1	100%	**8A**	0	3	0%
9A	1	3	33%	**9A**	4	6	67%	**9A**	1	1	100%	**9A**	2	6	33%
10A	1	2	50%	**10A**	4	6	67%	**10A**	1	1	100%	**10A**	1	3	33%

C.1 Systematic searches for optimal folding conditions

The temperature and nonbonded cutoff radius were systematically varied to determine the values which produce helices most often. Instead of testing only one simulation under each condition, approximately six simulations, differing only in initial velocities, were run under each set of conditions. The statistical significance of the data was considered before choosing the best conditions.

C.1.1 Optimal folding temperature

Table C.1 shows the effects of temperature on helix formation propensities. 450K promised the highest propensity to form α-helices, but Figure C.1 illustrates that results based on a maximum of six trials lack much confidence.

As stated in Chapter 5, 450K in the simulations does not necessarily mean 351°F or 177°C in the laboratory. 450K is a constant used in the molecular dynamics simulation; since other parameters in the simulation, e.g., the strength of hydrogen bond interactions, are not necessarily perfect reflections of reality, the temperature itself is not either. However, the error from each constant hopefully counteracts errors in the others so that the simulation *in toto* is a useful model of reality.

Figure C.1: p. 136. Probabilities of helix formation at different temperatures. The y axis represents the relative degree of confidence that the x value is the "true" probability of helix formation, i.e., the fraction that would form a helix if a trillion or more simulations were performed under the same conditions. The legend contains the number of successful, helix-forming runs out of the number of runs attempted at the listed temperature.

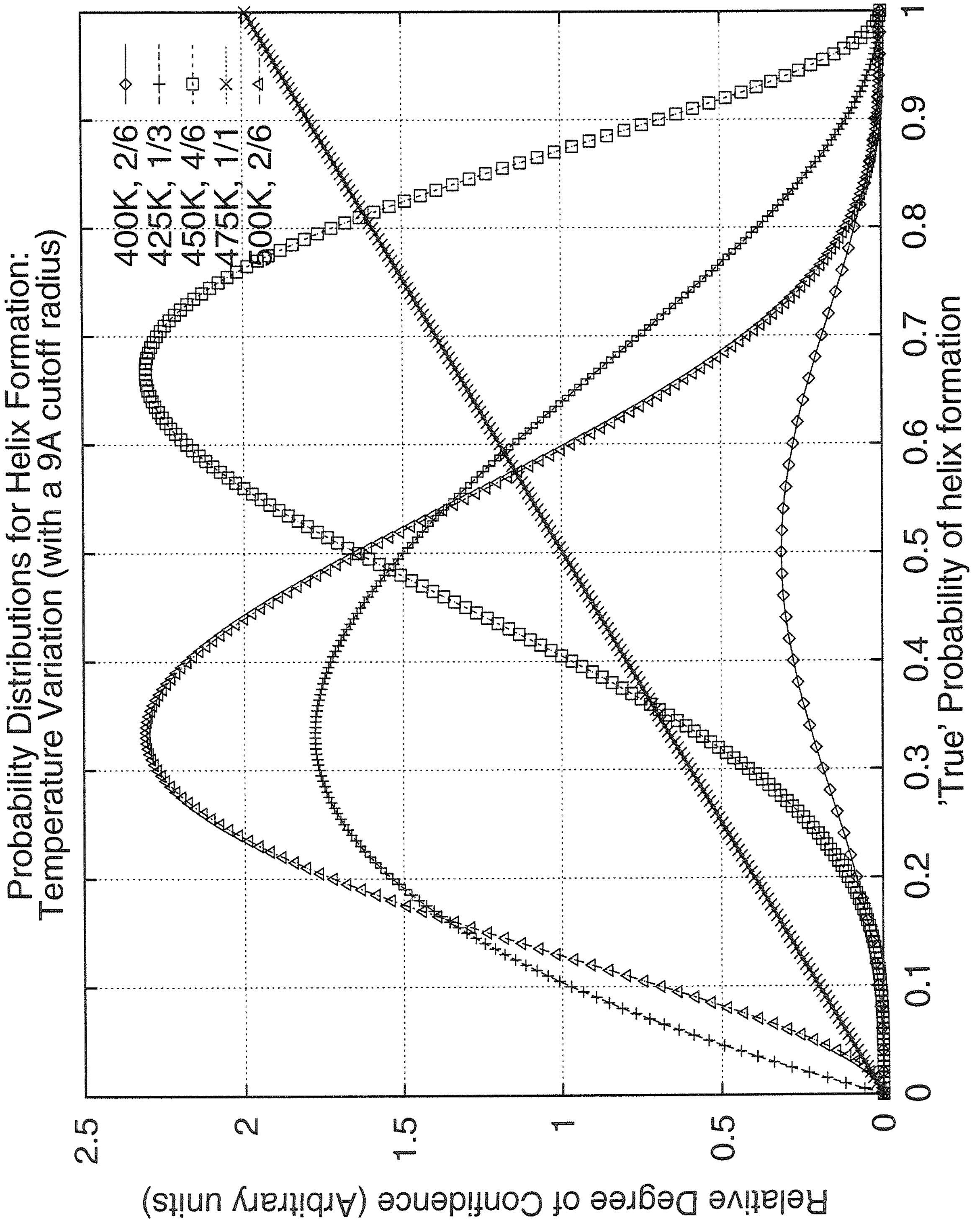


Figure C.1: Caption on p. 135.

C.1.2 Optimal cutoff radius for nonbonded interactions

Different cutoff radii for nonbonded interactions were tried. Because the cutoff radius determines the possible range of van der Waals and electrostatic forces, the cutoff radius is probably an important variable in mimicking solvent. Table C.2 summarizes the results of using different nonbonded cutoff radii at 400K, 450K, and 500K. After initial results on no more than six different simulations at each set of conditions, the 15 Å cutoff radius was chosen because five out of six of its runs formed helices. Further work supported the choice of 15 Å. Figure C.1.2 illustrates the relative confidence levels of the results at different cutoff radii.

A few simulations used an "infinite" cutoff radius. In other words, these simulations calculated every possible van der Waals and electrostatic interaction, no matter how distant and weak. This is also called calculating the nonbonded terms "explicitly." Most most simulations used finite cutoff radii, radii beyond which all nonbonded interactions were ignored.

When using a cutoff radius, a splines function attenuated the interactions in the spherical shell from the last 1 Å to the last 0.5 Å. For example, "a cutoff radius of 15 Å" means in this paper that interactions were assumed to be zero from 14.5 Å to 15.0 Å and a splines function attenuated interactions between 14 Å and 14.5 Å.

Table C.2: The Effect of varying the cutoff radius for nonbonded interactions on the probability of helix formation.

400K				450K				500K			
Non-bond cutoff (A)	No. of helices formed	No. of runs	% helix formation	Non-bond cutoff (A)	No. of helices formed	No. of runs	% helix formation	Non-bond cutoff (A)	No. of helices formed	No. of runs	% helix formation
6A	0	3	0%	**6A**	0	4	0%	**6A**	0	3	0%
8A	0	2	0%	**8A**	1	5	20%	**8A**	1	3	33%
9A	2	6	33%	**9A**	4	6	67%	**9A**	2	6	33%
10A	1	2	50%	**10A**	4	6	67%	**10A**	2	3	66%
15A				**15A**	129	155	83.2% +/- 6.0%	**15A**			
30A				**30A**	2	6	33%	**30A**			
inf		1	0%	**inf**	4	6	67%	**inf**			

In all simulations, the dielectric constant was $1\epsilon(r)$; the simulations lasted 500–1000 ps; the polyalanine had neutral terminal residues, and it was initially extended. Each run in one set of conditions used a different seed to randomize the initial velocities. Just as with Table C.1, the first run under all of the conditions used seed 12345; but the rest of the runs used seeds unique to that set of conditions. See Table C.1.4 for the seeds. The column entitled "Percentage Helix Formation" does not attempt to describe the confidence level in the percentage. "inf" means infinity. In other words, "inf" represents an infinitely long cutoff radius where the interactions were calculated explicitly.

Figure C.2: Page 140. Probabilities of helix formation at different nonbonded cutoff radii. The y axis represents the relative degree of confidence that the x value is the "true" probability of helix formation, i.e., the fraction that would form a helix if a trillion or more simulations were performed under the same conditions. The legend contains the number of successful, helix-forming runs out of the number of runs attempted at the listed cutoff radius. Although 15 Å was chosen based on the preliminary results of 5/6 runs, further results (129/155 runs) confirmed the choice. "No cutoff" means all nonbonded terms were calculated.

C.1.3 Summary of optimal conditions

The best conditions for helix formation from the extended state appeared to be 450K, 1ϵ(r) dielectric constant, 15 Å nonbonded cutoff radius, and neutral charges at the terminal residues. Under these conditions, polyalanine formed a helix within 500 ps with 83.2% probability (± 6.0% with 95% confidence), depending on the initial conditions. Refer to Table C.3. The degree of confidence under these conditions is compared to those with other nonbonded cutoff radii in Figure C.1.2.

Table C.3: Probability of helix formation at 450K and 15 Å cutoff radius for nonbonded interactions.

Total No. of Simulations	No. of Helices Formed	No. of Failures in 500 ps	Percentage Helix Formation	Error at 95% Confidence Level
155	129	26	83.2%	6.0%

C.1.4 Filenames and seeds of the above simulations

The filenames, seeds, and outcomes of the above simulations are summarized in Table C.1.4.

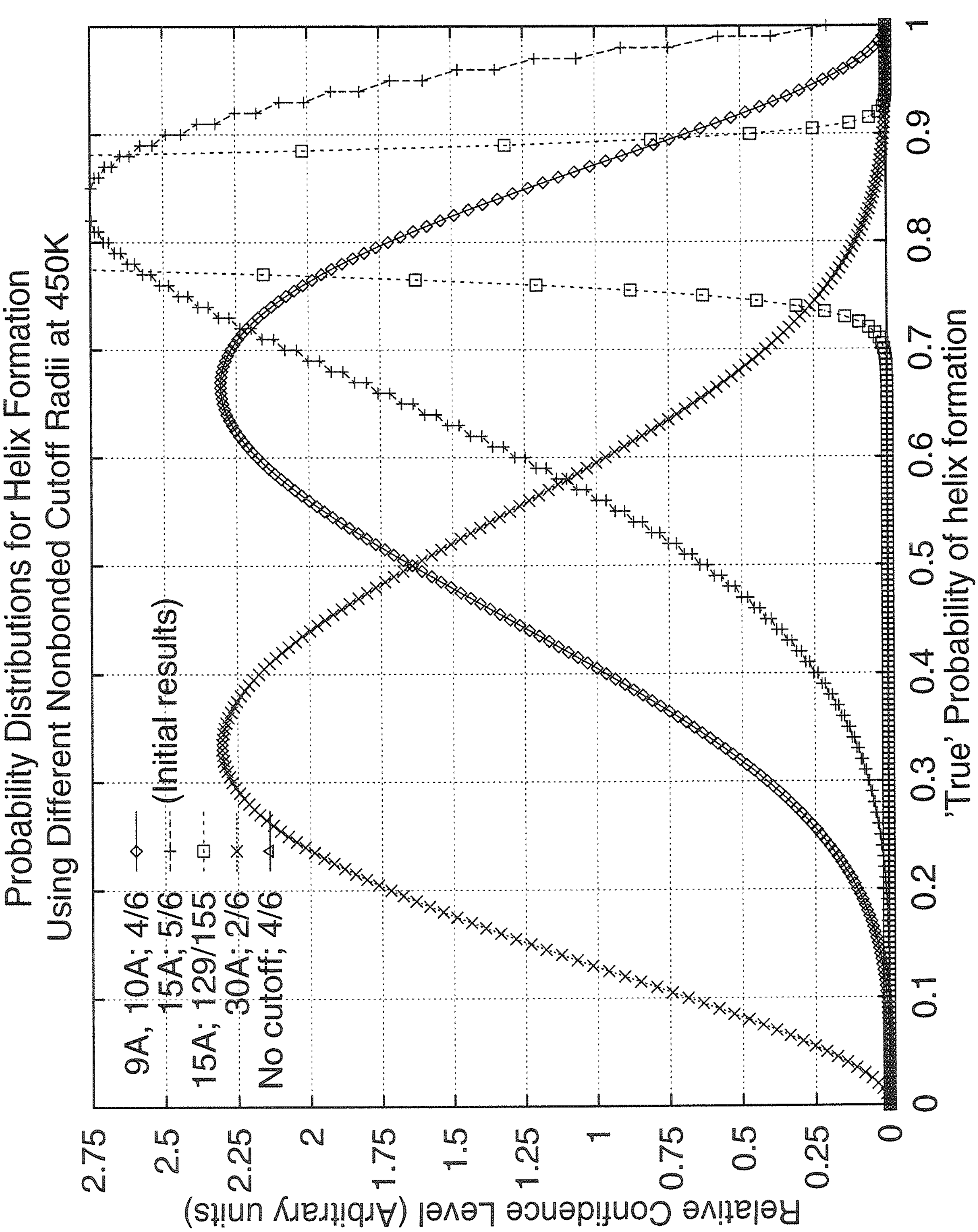


Figure C.2: Caption on p. 139.

Table C.4: pp. 142–144. Summary of runs: Filenames, random number seeds, and cutoff radii for nonbonded interactions for the simulations exploring temperature and nonbonded cutoff radius. Filenames and their seeds are listed at the top of each box. On the left column in each box is the nonbonded cutoff radius. If the number (in angstroms) has a slash through it, the run at that cutoff radius did not form an α-helix in the simulation time. The simulation time was 500 ps except where stated on the right-hand column within each box. Numbers on the right-hand column refer to picoseconds unless otherwise stated.

300K		350K	375K	400K	425K	450K	475K	500K
pAnc-300K 12345 ~~6~~ 8 30-140 ~~9~~ 10 180-340 except res 7		pAnc-350K 12345 ~~6~~ ~~8~~ ~~9~~ ~~10~~	pAnc-375K 12345 ~~9~~	pAnc-400K 12345 ~~6~~ ~~8~~ 9 100-230 ~~10~~ ~~in~~ <= 1000 ~~f~~ ps very trapped	pAnc-425K 12345 ~~6~~ ~~8~~ DARTS ~~9~~ ~~10~~ DARTS	pAnc-450K-2 12345 ~~8~~ 9 412 ~~10~~ 15 < 50 ps ~~30~~ ~~in~~ ~~f~~	pAnc-475K 12345 ~~6~~ 8 <300 9 ~20-200 10 ~130 2 1/2s first	pAnc-500K 12345 ~~6~~ ~~8~~ ~~9~~ -1500 ps ~~10~~ in 280ps f
		pAnc-350K-b 575797 ~~6~~ ~~9~~	pAnc-375K-b 414637 ~~6~~ ~~9~~	pAnc-400K-b 693053 ~~6~~ 9 60-~200	pAnc-425K-b 294875 ~~6~~ 8 <150 9 ~30-260 10 <60	pAnc-450K-b 706439 ~~6~~ ~~8~~ restart? 9 ~60-230 10 30-70 15 25-75 30 13-60 ~~in~~ ~~f~~		pAnc-500K-b 567499 ~~6~~ ~~8~~ 9 ~100-225 except 1-2 r 10 50-250 C term 1st

300K		350K	375K	400K	425K	450K	475K	500K
		pAnc-350K-c 736987 ~~6~~ ~~9~~	pAnc-375K-c 415003 ~~6~~ ~~9~~	pAnc-400K-c 240455 9 Kinked ~700 ps	pAnc-425K-c 137907 ~~6~~ ~~9~~ Candi- date for higher temp	pAnc-450K-c 567081 ~~6~~ ~~8~~ ~~9~~ glob 10 30-40 15 35-130 30 300?-440 in ~30-145 f		pAnc-500K-c 737897 ~~9~~ -1500ps
				pAnc-400K-d 137443 ~~6~~ ~~8~~ ~~9~~ 10 40-130		pAnc-450K-d 737553 ~~8~~ res 11-20 < 50 ps ~~9~~ -1000ps restart? 10 30-80 15 14-100 ~~30~~ in 12-88 f		pAnc-500K-d 914541 ~~6~~ ~~8~~ restart? res 1-10 <250 ~~9~~ glob 10 30-70

300K		350K	375K	400K	425K	450K	475K	500K
				pAnc-400K-e 412723 ~~9~~ -2000ps		pAnc-450K-e 43753 ~~6~~ 8 ~50 9 50-142 ~~10~~ ~~15~~ ~~30~~ in ~30-415 f alpha helix		pAnc-500K-e ~~9~~ why not?
				pAnc-400K-f 137527 ~~9~~ -1000ps		pAnc-450K-f 415121 ~~6~~ 1/2helix 4,11,10 not restart? 9 80-270 10 30-330 15 30-100 ~~30~~ in 10-105 f		pAnc-450K-d-500K 737553 9 100-350

C.2 Explorations of the effect of the terminal charges and the dielectric constant

Some preliminary data was collected on the effects of charged terminal residues and different dielectric constants. None of these explorations attempted helix formation with more than one set of initial velocities. Thus, the statistical significance of these results may be scant. However, all of these runs used the same relative initial velocities, scaled to the temperature of the simulation, by using the same initial seed, 12345, to assign the initial momenta.

C.2.1 The effect of terminal charge

The effect of charge on the terminal residues is summarized in Table C.2.1. The runs had +1 and -1 charges on the amino and carboxy termini, respectively. There were no counterions to dampen the electrostatic field (6-27-95), and the dielectric constant was $1\epsilon(\mathrm{r})$. Having uncompensated charges did not prevent the simulation from forming an α-helix. Comparing the 400K simulation of the charged $(\mathrm{Ala})_{20}$ to an identical simulation of a neutral $(\mathrm{Ala})_{20}$ indicates that charges accelerated the formation of the helix. In the charged simulation, a helical loop formed in 60 ps; by 100 ps residues 5 through 15 were in an α-helix. The terminal residues, i.e., residues 1–4 and residues 16–20, took longer to settle into a helix. In contrast, by 100 ps the neutral simulation (pAnc-400K-15A) was just beginning to form its first helical loop.

C.2.2 The effect of the dielectric constant

Energy calculations did not indicate that changing the dielectric constant in the range of $2\epsilon(\mathrm{r})$ to $0.5\epsilon(\mathrm{r})$ would improve the probability of helix formation. Figure C.3 compares the total energy of several key structures that $(\mathrm{Ala})_{20}$ adopts during most of the simulations to changes in the dielectric constant. The structures compared were the extended structure, the α-helix, and the "C7-well" (Avignon *et al.*, 1969) where

Table C.5: Simulations with charged terminal residues.

Run Name	Date	Temperature	Formed a Helix?	Comments
pA-10fs-300K-500ps.[12]	6-19-95	300K	no	The magnitude of the forces crashed some atoms into a potential wall. Simulation aborted.
pA-10fs-400K-500ps.3	6-19-95	400K	yes	Residues 5–15 formed a helix in less than 110 ps. The terminal residues remained disordered for longer.

Simulations were run 6-19-95 with +1 and -1 charges on the amino and carboxy terminal residues, respectively. The cutoff radii for nonbonded interactions was 9 Å; the dielectric constant was $1\epsilon(r)$. Because the seed initializing the random velocities was 12345 in all cases, the simulations started with the same relative velocities, scaled for the temperature of the simulation. (Macros, torsion, trajectory, and output files exist for both runs. The bgf file was pA-amber.bgf. Doubling was not used. The initial temperature was the same as the final temperature.)

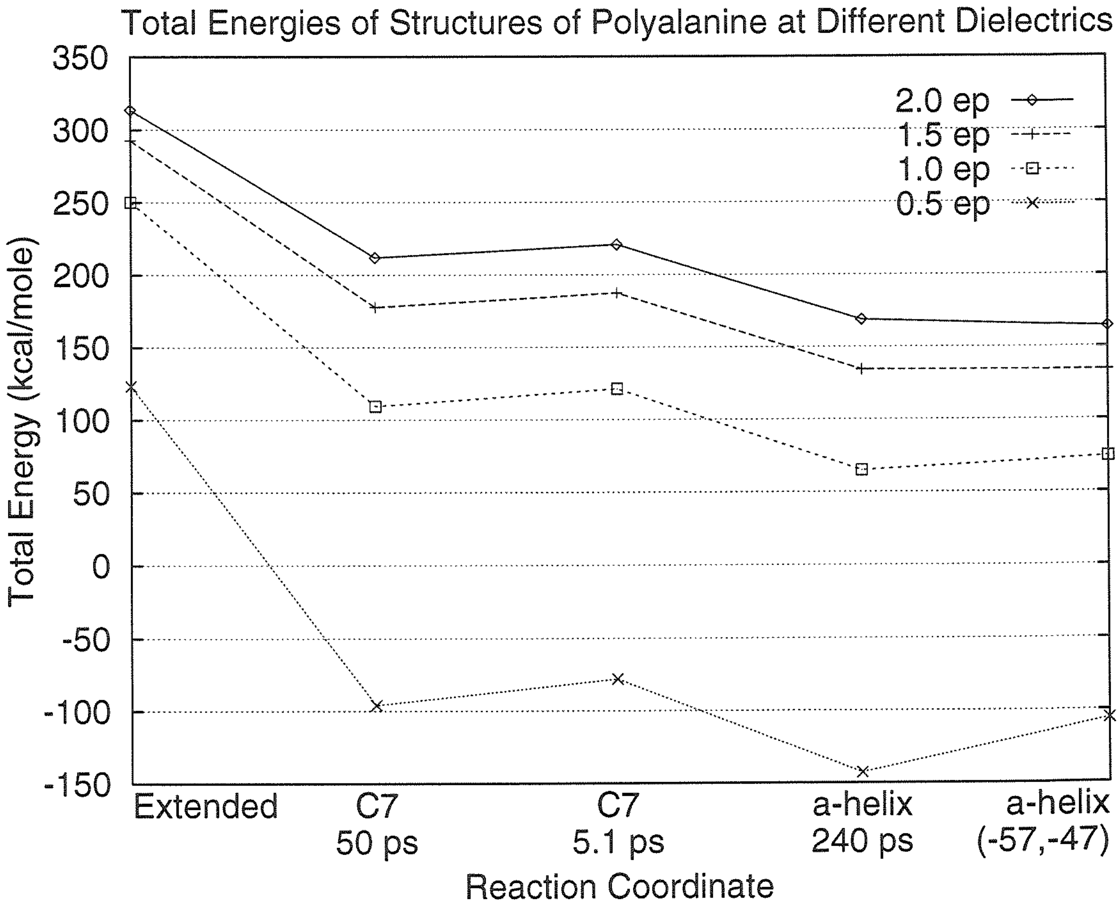


Figure C.3: Total energies of structures of $(\text{Ala})_{20}$ at different dielectrics. The "extended" structure is where $(\phi, \psi) = (-180°, +180°)$. "C7" refers to $i, i{+}2$ hydrogen bonding structures at 5.1 ps and 50 ps of simulation pAnc-400K-9A. "a-helix" refers to an α-helix. The first α-helix plotted is the structure of trajectory pAnc-400K-9A at 240 ps when it was a fully formed helix. The second structure is an ideal structure with $(\phi, \psi) = (-57°, -47°)$.

Table C.6: Simulations at 300K and 400K with $10\epsilon(\mathrm{r})$ dielectric constant.

Run	Date	Temp	Formed a Helix?	Duration of Simulation	Comments
pAnc-300K-10ep	11-15-95	300K	no	500 ps	
pAnc-400K-10ep	11-15-95	400K	no	513 ps	Lots of structure, but none persisted. H bonds were not tight. Large radius of gyration.

Both runs used neutral terminal residues. The cutoff radii for nonbonded interactions was 9 Å; the dielectric constant was $10\epsilon(\mathrm{r})$. The seed initializing the random velocities was 12345 in all cases.

every residue is stably hydrogen bonded to its $i, i+2$th neighbor. See Section 4.2.1.

Tables C.6 and C.7 summarize the effect of altering the magnitude of the dielectric constant while keeping its distance dependence. Figure C.3 graphs the relative energies of the polyalanine in the extended state, the γ-well, and the α-helix for different magnitudes of ϵ. As expected, the energy of the C7 region, shown by $(\mathrm{Ala})_{20}$ at 5.1 ps, is substantially lower than that of the extended structure in all cases. The difference in energy between the C7 structure and the α-helix is as large or larger in the 1.0ϵ force field as in the other dielectrics tried. Thus, there is no indication a different dielectric would increase the propensity to move from the C7 region to the α-helix.

Table C.7: Runs with different dielectric constants.

Run	Date	Dielectric	Helix?	Duration of Simulation or Time of Helix Formation
pAnc-450K-0.5ep	12-12-95	0.5e(r)	no	1000 ps
pAnc-450K-2	12-1-95	1e(r)	yes	412 ps
pAnc-450K-1.25ep	12-13-95	1.25e(r)	NA	
pAnc-450K-1.5ep	12-13-95	1.5e(r)	No	1000 ps
pAnc-450K-2ep	12-12-95	2e(r)	No	1000 ps

All simulations were run at 450K with a 9 Å cutoff radius for nonbonded interactions. The seed initializing the random velocities was 12345 in all cases. “NA” means “not available.”

Appendix D Summary of (Ala)$_{20}$ and (Gly)$_{20}$ Runs

D.1 (Ala)$_{20}$ simulations

Table D.1: Each (Ala)$_{20}$ simulation and its statistics. These are the records from file /ul/rbertsch/bgf/prog/pAnc-450K-xx-15A.sum and describes the results from all simulations of (Ala)$_{20}$ at 450K and 15 Å cutoff radius for nonbonded interactions. Column one is a shorthand for the name of the simulation. Column two is the name the trajectory file. Column three lists the outcome of the run, i.e., if it formed a helix or not. "he" means it formed a helix; "N" means it had not by 500 ps, and "NA" means not determined. The fourth column is sometimes empty, but it lists the number of residues that lie outside the α-helical range defined in /ul/rbertsch/bin/runs/helix.awk. Column five is the time each trajectory first formed an α-helix, determined by the total number of $i, i+4$ HBs. Column six is the maximum number of nonlocal HBs occurring at any one time but at any point during the trajectory. Column seven is the time of the structure with the most nonlocal hydrogen bonds in the trajectory. Column eight sometimes lists the helix formation time, defined by the percentage of α-helical residues visible to circular dichroism (CD) and calculated by /ul/rbertsch/bin/percent-helicity.f.

#		ou	No.	HB-type			% hel
#	Run Name	tc	out-	Time of	Max	Time of	Time
#		om	lyers	Helix	No	Max No	Helix
#		e	500ps	Formatn	Reds	of Reds	Formtn
a	pAnc-450K-a-15A.trj	he		58.000	1	24.000	
aa	pAnc-450K-aa-15A.trj	he		160.000	0	0.000	
ab	pAnc-450K-ab-15A.trj	he		265.000	8	86.400	
ac	pAnc-450K-ac-15A.trj	N		500.000	6	495.772	
ad	pAnc-450K-ad-15A.trj	he		80.000	1	60.296	
ae	pAnc-450K-ae-15A.trj	he		90.000	0	0.000	
af	pAnc-450K-af-15A.trj	he		310.000	4	25.101	
ag	pAnc-450K-ag-15A.trj	he		170.000	6	36.700	
ah	pAnc-450K-ah-15A.trj	N		500.000	7	232.651	

ai	pAnc-450K-ai-15A.trj	he		60.000	0	0.000	
aj	pAnc-450K-aj-15A.trj	N		500.000	12	133.606	
ak	pAnc-450K-ak-15A.trj	he		50.000	2	36.600	
al	pAnc-450K-al-15A.trj	he		150.000	5	59.896	
am	pAnc-450K-am-15A.trj	N		500.000	8	128.809	
an	pAnc-450K-an-15A.trj	N		500.000	9	95.002	
ao	pAnc-450K-ao-15A.trj	N		500.000	12	128.309	
ap	pAnc-450K-ap-15A.trj	he		90.000	3	22.700	
aq	pAnc-450K-aq-15A.trj	he		60.000	2	26.101	
ar	pAnc-450K-ar-15A.trj	he		70.000	3	45.798	
as	pAnc-450K-as-15A.trj	he		120.000	5	28.301	
at	pAnc-450K-at-15A.trj	he		80.000	5	42.999	
au	pAnc-450K-au-15A.trj	N		500.000	8	387.467	
av	pAnc-450K-av-15A.trj	N		500.000	11	241.147	
aw	pAnc-450K-aw-15A.trj	he		100.000	4	44.399	
ax	pAnc-450K-ax-15A.trj	he		80.000	1	57.796	
ay	pAnc-450K-ay-15A.trj	he		350.000	9	64.895	
az	pAnc-450K-az-15A.trj	he		130.000	7	94.202	
b	pAnc-450K-b-15A.trj	he		100.000	3	38.400	
ba	pAnc-450K-ba-15A.trj	he		110.000	7	50.498	
bb	pAnc-450K-bb-15A.trj	N		500.000	9	425.804	
bc	pAnc-450K-bc-15A.trj	he		80.000	3	54.897	
bd	pAnc-450K-bd-15A.trj	N		500.000	12	145.599	
be	pAnc-450K-be-15A.trj	he		105.000	0	0.000	
bf	pAnc-450K-bf-15A.trj	he		130.000	5	73.797	
bg	pAnc-450K-bg-15A.trj	he	2	100.000	3	53.797	
bh	pAnc-450K-bh-15A.trj	he		100.000	6	58.396	
bi	pAnc-450K-bi-15A.trj	he		80.000	2	42.599	
bj	pAnc-450K-bj-15A.trj	he		160.000	10	46.498	
bk	pAnc-450K-bk-15A.trj	he		190.000	6	58.496	
bl	pAnc-450K-bl-15A.trj	N		500.000	8	365.846	
bm	pAnc-450K-bm-15A.trj	he		305.000	8	43.399	
bn	pAnc-450K-bn-15A.trj	he		135.000	3	29.401	170.00
bo	pAnc-450K-bo-15A.trj	he	2	500.000	9	186.976	
bp	pAnc-450K-bp-15A.trj	N		500.000	14	336.517	
bq	pAnc-450K-bq-15A.trj	he		90.000	2	18.400	
br	pAnc-450K-br-15A.trj	he		70.000	2	27.601	
bs	pAnc-450K-bs-15A.trj	he		400.000	9	96.102	
bt	pAnc-450K-bt-15A.trj	he		130.000	5	53.397	
bu	pAnc-450K-bu-15A.trj	he		100.000	2	30.501	
bv	pAnc-450K-bv-15A.trj	N	15	500.000	10	149.297	
bw	pAnc-450K-bw-15A.trj	he	2	250.000	7	38.999	270.00
bx	pAnc-450K-bx-15A.trj	N	11	500.000	8	150.097	
by	pAnc-450K-by-15A.trj	he		100.000	0	0.000	
bz	pAnc-450K-bz-15A.trj	N		500.000	10	292.774	
c	pAnc-450K-c-15A.trj	he		140.000	5	65.896	

ca	pAnc-450K-ca-15A.trj	he		50.000	0	0.000	
cb	pAnc-450K-cb-15A.trj	he	2	70.000	2	45.898	
cc	pAnc-450K-cc-15A.trj	he	1	50.000	2	21.600	
cd	pAnc-450K-cd-15A.trj	he		100.000	4	21.500	
ce	pAnc-450K-ce-15A.trj	he	2	150.000	5	61.996	175.0
cf	pAnc-450K-cf-15A.trj	he		230.000	10	113.206	
cg	pAnc-450K-cg-15A.trj	he	0	50.000	0	0.000	55.0
ch	pAnc-450K-ch-15A.trj	he		210.000	9	55.797	
ci	pAnc-450K-ci-15A.trj	he		500.000	7	237.948	
cj	pAnc-450K-cj-15A.trj	he		100.000	2	44.199	
ck	pAnc-450K-ck-15A.trj	he		450.000	6	55.897	
cl	pAnc-450K-cl-15A.trj	he	1	210.000	7	23.500	220.0
cm	pAnc-450K-cm-15A.trj	he		220.000	7	27.601	
cn	pAnc-450K-cn-15A.trj	he		70.000	5	20.400	
co	pAnc-450K-co-15A.trj	N	10	500.000	8	78.098	
cp	pAnc-450K-cp-15A.trj	he		70.000	3	34.900	
cq	pAnc-450K-cq-15A.trj	he	1	130.000	3	46.898	
cr	pAnc-450K-cr-15A.trj	he	3	105.000	5	21.600	115.0
cs	pAnc-450K-cs-15A.trj	he		110.000	2	31.501	
ct	pAnc-450K-ct-15A.trj	he		100.000	4	36.300	
cu	pAnc-450K-cu-15A.trj	he		80.000	2	58.796	
cv	pAnc-450K-cv-15A.trj	he	0	170.000	2	158.092	
cw	pAnc-450K-cw-15A.trj	he	2	80.000	0	0.000	
cx	pAnc-450K-cx-15A.trj	he		70.000	8	34.400	
cy	pAnc-450K-cy-15A.trj	he	2	90.000	4	34.200	
cz	pAnc-450K-cz-15A.trj	he		75.000	0	0.000	
d	pAnc-450K-d-15A.trj	he		100.000	4	36.900	
da	pAnc-450K-da-15A.trj	N	12	500.000	9	184.378	
db	pAnc-450K-db-15A.trj	he	1	270.000	7	56.397	
dc	pAnc-450K-dc-15A.trj	he		35.000	0	0.000	
dd	pAnc-450K-dd-15A.trj	he	2	75.000	1	46.698	
de	pAnc-450K-de-15A.trj	he	3	70.000	3	35.900	
df	pAnc-450K-df-15A.trj	he	0	150.000	4	24.101	
dg	pAnc-450K-dg-15A.trj	he	2	50.000	1	28.201	
dh	pAnc-450K-dh-15A.trj	he	6	150.000	6	47.998	
di	pAnc-450K-di-15A.trj	he	0	120.000	3	18.900	135.0
dj	pAnc-450K-dj-15A.trj	he	2	340.000	6	103.904	
dk	pAnc-450K-dk-15A.trj	N	17	500.000	10	99.603	
dl	pAnc-450K-dl-15A.trj	he	1	70.000	3	34.600	
dn	pAnc-450K-dn-15A.trj	N	9	500.000	10	95.302	
do	pAnc-450K-do-15A.trj	he	0	130.000	2	36.800	
dp	pAnc-450K-dp-15A.trj	he	1	90.000	0	0.000	
dq	pAnc-450K-dq-15A.trj	he	0	150.000	6	43.299	
dr	pAnc-450K-dr-15A.trj	he	1	75.000	2	37.900	
ds	pAnc-450K-ds-15A.trj	he	0	90.000	5	54.897	
dt	pAnc-450K-dt-15A.trj	he	0	150.000	5	58.696	

du	pAnc-450K-du-15A.trj	he	0	70.000	1	43.999	
dv	pAnc-450K-dv-15A.trj	he	0	175.000	1	59.496	
dw	pAnc-450K-dw-15A.trj	he	0	70.000	2	25.701	
dx	pAnc-450K-dx-15A.trj	he		120.000	9	31.701	
dy	pAnc-450K-dy-15A.trj	he	0	75.000	4	51.897	
dz	pAnc-450K-dz-15A.trj	he	2	130.000	2	15.100	
e	pAnc-450K-e-15A.trj	N		500.000	9	144.900	
ea	pAnc-450K-ea-15A.trj	he	0	360.000	8	185.078	
eb	pAnc-450K-eb-15A.trj	he	1	90.000	5	46.898	
ec	pAnc-450K-ec-15A.trj	he	1	120.000	0	0.000	
ed	pAnc-450K-ed-15A.trj	he	0	140.000	4	46.298	
ee	pAnc-450K-ee-15A.trj	he	2	140.000	9	53.497	
ef	pAnc-450K-ef-15A.trj	he	1	375.000	8	31.501	
eg	pAnc-450K-eg-15A.trj	N	7	500.000	8	23.300	
eh	pAnc-450K-eh-15A.trj	he	2	110.000	4	66.396	
ei	pAnc-450K-ei-15A.trj	he	0	225.000	7	30.201	
ej	pAnc-450K-ej-15A.trj	he	0	60.000	0	0.000	
ek	pAnc-450K-ek-15A.trj	N	14	500.000	7	122.108	
el	pAnc-450K-el-15A.trj	N	17	500.000	9	300.482	
em	pAnc-450K-em-15A.trj	he	2	80.000	6	31.501	
en	pAnc-450K-en-15A.trj	he	0	75.000	1	11.800	
eo	pAnc-450K-eo-15A.trj	he	1	50.000	1	23.000	
ep	pAnc-450K-ep-15A.trj	he	1	100.000	1	38.999	120
eq	pAnc-450K-eq-15A.trj	he	0	120.000	1	93.602	
er	pAnc-450K-er-15A.trj	he	1	200.000	6	158.892	
es	pAnc-450K-es-15A.trj	he	1	50.000	2	20.100	
et	pAnc-450K-et-15A.trj	he	1	75.000	3	26.901	
eu	pAnc-450K-eu-15A.trj	he	0	125.000	0	0.000	
ev	pAnc-450K-ev-15A.trj	he	1	60.000	1	38.700	
ew	pAnc-450K-ew-15A.trj	he	0	90.000	0	0.000	
ex	pAnc-450K-ex-15A.trj	he	2	150.000	8	41.899	
ey	pAnc-450K-ey-15A.trj	he	0	80.000	0	0.000	
ez	pAnc-450K-ez-15A.trj	he	1	200.000	7	55.697	
f	pAnc-450K-f-15A.trj	he		100.000	5	36.200	
g	pAnc-450K-g-15A.trj	he		210.000	9	74.097	
h	pAnc-450K-h-15A.trj	he		90.000	1	69.597	
i	pAnc-450K-i-15A.trj	he		80.000	0	0.000	
j	pAnc-450K-j-15A.trj	he		75.000	0	0.000	
k	pAnc-450K-k-15A.trj	N		500.000	11	85.900	
l	pAnc-450K-l-15A.trj	he		275.000	6	116.707	
m	pAnc-450K-m-15A.trj	he		140.000	3	31.201	145.0
n	pAnc-450K-n-15A.trj	he		185.000	6	23.800	170.0
o	pAnc-450K-o-15A.trj	he		100.000	4	47.398	
p	pAnc-450K-p-15A.trj	he		240.000	6	57.496	
q	pAnc-450K-q-15A.trj	he		420.000	7	41.599	
r	pAnc-450K-r-15A.trj	N		500.000	10	471.649	

s	pAnc-450K-s-15A.trj	he	200.000	7	34.600	
t	pAnc-450K-t-15A.trj	N	500.000	6	34.100	
u	pAnc-450K-u-15A.trj	he	500.000	6	53.397	
v	pAnc-450K-v-15A.trj	he	60.000	2	28.401	
w	pAnc-450K-w-15A.trj	he	425.000	9	114.806	
x	pAnc-450K-x-15A.trj	he	210.000	6	40.599	
y	pAnc-450K-y-15A.trj	he	80.000	2	25.800	100.0
z	pAnc-450K-z-15A.trj	he	310.000	7	169.986	
#						

Table D.2: Filenames, seeds, and dates of all (Ala)$_{20}$ runs. This table is part of file /ul/rbertsch/bgf/prog/pAnc-450K-xx-15A.results.txt. The "Directory Name" can be converted into the prefix of the file name by adding the prefex "pA." For example, directory /ul/rbertsch/bgf/nc-450K-a-15A contains file pAnc-450K-a-15A.trj and pAnc-450K-a-15A.tor. The seed is the random number seed that generates the unique set of initial velocities. The outcome is whether or not the run formed a helix within 500 ps. "he" means the run formed a helix, and "N" means it did not.

Directory Name	Seed	Outcome	Directory Name	Seed	Outcome
nc-450K-a-15A	12345	he	nc-450K-au-15A	748336	N
nc-450K-aa-15A	564027	he	nc-450K-av-15A	55106	N
nc-450K-ab-15A	341122	he	nc-450K-aw-15A	89565	he
nc-450K-ac-15A	904865	N	nc-450K-ax-15A	485160	he
nc-450K-ad-15A	698606	he	nc-450K-ay-15A	734162	he
nc-450K-ae-15A	417488	he	nc-450K-az-15A	864876	he
nc-450K-af-15A	649117	he	nc-450K-b-15A	706439	he
nc-450K-ag-15A	278881	he	nc-450K-ba-15A	809043	he
nc-450K-ah-15A	630224	N	nc-450K-bb-15A	67342	N
nc-450K-ai-15A	954549	he	nc-450K-bc-15A	162648	he
nc-450K-aj-15A	438310	N	nc-450K-bd-15A	934857	N
nc-450K-ak-15A	484127	he	nc-450K-be-15A	321390	he
nc-450K-al-15A	726423	he	nc-450K-bf-15A	153740	he
nc-450K-am-15A	938500	N	nc-450K-bg-15A	256653	he
nc-450K-an-15A	569813	N	nc-450K-bh-15A	12439	he
nc-450K-ao-15A	122233	N	nc-450K-bi-15A	166736	he
nc-450K-ap-15A	784741	he	nc-450K-bj-15A	162222	he
nc-450K-aq-15A	141509	he	nc-450K-bk-15A	418626	he
nc-450K-ar-15A	604230	he	nc-450K-bl-15A	170174	N
nc-450K-as-15A	935806	he	nc-450K-bm-15A	181527	he
nc-450K-at-15A	456563	he	nc-450K-bn-15A	135687	he

nc-450K-bo-15A	728254	he	nc-450K-dg-15A	834665	he
nc-450K-bp-15A	314697	N	nc-450K-dh-15A	873505	he
nc-450K-bq-15A	375571	he	nc-450K-di-15A	912751	he
nc-450K-br-15A	434919	he	nc-450K-dj-15A	915661	he
nc-450K-bs-15A	494455	he	nc-450K-dk-15A	916023	N
nc-450K-bt-15A	560219	he	nc-450K-dl-15A	954465	he
nc-450K-bu-15A	618303	he	nc-450K-dn-15A	302459	N
nc-450K-bv-15A	684729	N	nc-450K-do-15A	312725	he
nc-450K-bw-15A	753217	he	nc-450K-dp-15A	318071	he
nc-450K-bx-15A	0	N	nc-450K-dq-15A	908445	he
nc-450K-by-15A	884371	he	nc-450K-dr-15A	957565	he
nc-450K-bz-15A	656737	N	nc-450K-ds-15A	961169	he
nc-450K-c-15A	567081	he	nc-450K-dt-15A	964115	he
nc-450K-ca-15A	745773	he	nc-450K-du-15A	33609	he
nc-450K-cb-15A	824977	he	nc-450K-dv-15A	38247	he
nc-450K-cc-15A	910417	he	nc-450K-dw-15A	85933	he
nc-450K-cd-15A	971937	he	nc-450K-dx-15A	0	he
nc-450K-ce-15A	13463	he	nc-450K-dy-15A	116039	he
nc-450K-cf-15A	0	he	nc-450K-dz-15A	125669	he
nc-450K-cg-15A	92891	he	nc-450K-e-15A	43753	N
nc-450K-ch-15A	130109	he	nc-450K-ea-15A	214103	he
nc-450K-ci-15A	170651	he	nc-450K-eb-15A	214241	he
nc-450K-cj-15A	67595	he	nc-450K-ec-15A	214361	he
nc-450K-ck-15A	67869	he	nc-450K-ed-15A	344805	he
nc-450K-cl-15A	150367	he	nc-450K-ee-15A	187439	he
nc-450K-cm-15A	158183	he	nc-450K-ef-15A	190579	he
nc-450K-cn-15A	235081	he	nc-450K-eg-15A	397947	N
nc-450K-co-15A	242001	N	nc-450K-eh-15A	401209	he
nc-450K-cp-15A	297333	he	nc-450K-ei-15A	481903	he
nc-450K-cq-15A	307323	he	nc-450K-ej-15A	482213	he
nc-450K-cr-15A	337419	he	nc-450K-ek-15A	567889	N
nc-450K-cs-15A	362653	he	nc-450K-el-15A	572211	N
nc-450K-ct-15A	378433	he	nc-450K-em-15A	580513	he
nc-450K-cu-15A	378651	he	nc-450K-en-15A	618917	he
nc-450K-cv-15A	419373	he	nc-450K-eo-15A	658575	he
nc-450K-cw-15A	448315	he	nc-450K-ep-15A	695581	he
nc-450K-cx-15A	455797	he	nc-450K-eq-15A	727999	he
nc-450K-cy-15A	462009	he	nc-450K-er-15A	728103	he
nc-450K-cz-15A	493183	he	nc-450K-es-15A	734429	he
nc-450K-d-15A	737553	he	nc-450K-et-15A	775439	he
nc-450K-da-15A	744349	N	nc-450K-eu-15A	806167	he
nc-450K-db-15A	744483	he	nc-450K-ev-15A	807335	he
nc-450K-dc-15A	744569	he	nc-450K-ew-15A	821189	he
nc-450K-dd-15A	793767	he	nc-450K-ex-15A	869243	he
nc-450K-de-15A	854687	he	nc-450K-ey-15A	883373	he
nc-450K-df-15A	854739	he	nc-450K-ez-15A	884927	he

nc-450K-f-15A	415121	he	nc-450K-q-15A	396746	he
nc-450K-g-15A	314159	he	nc-450K-r-15A	214512	N
nc-450K-h-15A	265359	he	nc-450K-s-15A	893471	he
nc-450K-i-15A	979323	he	nc-450K-t-15A	352354	N
nc-450K-j-15A	846264	he	nc-450K-u-15A	35413	he
nc-450K-k-15A	338327	N	nc-450K-v-15A	546627	he
nc-450K-l-15A	950288	he	nc-450K-w-15A	446092	he
nc-450K-m-15A	419716	he	nc-450K-x-15A	877159	he
nc-450K-n-15A	939937	he	nc-450K-y-15A	421310	he
nc-450K-o-15A	660498	he	nc-450K-z-15A	427937	he
nc-450K-p-15A	471339	he			

D.2 (Gly)$_{20}$ simulations

Table D.3: Directory names, seeds, and outcomes of (Gly)$_{20}$ simulations. The fourth column is the number of residues at 500 ps that lay outside the helical region defined in /ul/rbertsch/bin/runs/helix.awk. Columns are labeled identically to Table D.2.

Directory Name	Seed	Outcome	No. Outlyers
pG/pG-450K-a-15A	12345	N	13
pG/pG-450K-b-15A	706439	N	8
pG/pG-450K-c-15A	567081	N	7
pG/pG-450K-d-15A	737553	N	16
pG/pG-450K-e-15A	43753	N	15
pG/pG-450K-f-15A	415121	N	16
pG/pG-450K-g-15A	314159	N	10
pG/pG-450K-h-15A	265359	N	18
pG/pG-450K-i-15A	979323	N	16
pG/pG-450K-j-15A	846264	N	17
pG/pG-450K-k-15A	338327	N	18
pG/pG-450K-l-15A	950288	N	17
pG/pG-450K-m-15A	419716	N	16
pG/pG-450K-n-15A	939937	N	18
pG/pG-450K-o-15A	660498	N	17
pG/pG-450K-p-15A	471339	N	18
pG/pG-450K-q-15A	396746	N	15
pG/pG-450K-r-15A	214512	N	18
pG/pG-450K-s-15A	893471	N	17
pG/pG-450K-t-15A	352354	N	17

Appendix E The BIOGRAF Molecule

E.1 $(\mathrm{Ala})_{20}$

The BIOGRAF file specifying the atoms, their initial positions, and charges was /ul/rbertsch/bgf/pAnc2.bgf for all simulations on $(\mathrm{Ala})_{20}$ except for several in Appendix C, Section C.2.1 that used terminal charges. The charges in the pAnc2.bgf file were modified from the default BIOGRAF, AMBER values to have neutral terminal amino acids. In the future, simulations will be run with the terminal oxygens sharing the -0.5 charge. The file pAnc2.bgf is included as Figure E.1.

E.2 $(\mathrm{Gly})_{20}$

The $(\mathrm{Gly})_{20}$ analog of pAnc2.bgf, pG.bgf (/ul/rbertsch/bgf/pG/pG.bgf), is included as Figure E.2, starting p. 165.

Figure E.1: The BIOGRAF file for $(\mathrm{Ala})_{20}$.

```
BIOGRF  332
DESCRP pAnc2
REMARK phi, psi originally at +180, +180
REMARK (Ala)20 Extended Fully
REMARK unminimized
REMARK charges corrected
REMARK Created by rbertsch @ sgi1 on 11/14/95   11:55:11
FORCEFIELD AMBER
FORMAT ATOM   (a6,1x,i5,1x,a5,1x,a3,1x,a1,1x,a5,3f10.5,1x,a5,i3,i2,1x,f8.5,i2,i4,f10.5)
ATOM       1  N    ALA     1   -0.66300  -0.42900   0.56900 N      4 0 -0.52000 0    0
ATOM       2 HN    ALA     1   -0.64379  -1.41544   0.78501 H      1 0  0.08200 0    0
ATOM       3 HN    ALA     1   -0.65948  -0.05715   1.50805 H      1 0  0.08300 0    0
ATOM       4 HN    ALA     1   -1.58300  -0.34700   0.39500 H      1 0  0.08300 0    0
ATOM       5  CA   ALA     1    0.30609   0.08967  -0.38434 CH     4 0  0.21500 0    0
ATOM       6  C    ALA     1    1.71548  -0.25130   0.03694 C      3 0  0.52600 0    0
ATOM       7  O    ALA     1    2.44037  -0.94377  -0.69287 O      1 2 -0.50000 0    0
ATOM       8  CB   ALA     1   -0.00891   1.56167  -0.70034 C3     4 0  0.03100 0    0
ATOM       9  N    ALA     2    2.10437   0.23100   1.20816 N      3 0 -0.52000 0    0
ATOM      10 HN    ALA     2    1.45079   0.79350   1.75298 H      1 0  0.24800 0    0
ATOM      11  CA   ALA     2    3.43541  -0.01740   1.74076 CH     4 0  0.21500 0    0
ATOM      12  C    ALA     2    3.61250   0.64567   3.08578 C      3 0  0.52600 0    0
ATOM      13  O    ALA     2    2.67399   1.27883   3.54858 O      1 2 -0.50000 0    0
ATOM      14  CB   ALA     2    4.49172   0.26674   0.65939 C3     4 0  0.03100 0    0
ATOM      15  N    ALA     3    4.79396   0.48220   3.66289 N      3 0 -0.52000 0    0
ATOM      16 HN    ALA     3    5.50646  -0.06834   3.18368 H      1 0  0.24800 0    0
ATOM      17  CA   ALA     3    5.10585   1.06418   4.95945 CH     4 0  0.21500 0    0
ATOM      18  C    ALA     3    6.51523   0.72321   5.38073 C      3 0  0.52600 0    0
ATOM      19  O    ALA     3    7.19554   0.03812   4.62992 O      1 2 -0.50000 0    0
ATOM      20  CB   ALA     3    3.97068   0.76402   5.95301 C3     4 0  0.03100 0    0
ATOM      21  N    ALA     4    6.90389   1.20491   6.55228 N      3 0 -0.52000 0    0
ATOM      22 HN    ALA     4    6.25014   1.76699   7.09733 H      1 0  0.24800 0    0
ATOM      23  CA   ALA     4    8.23489   0.95643   7.08492 CH     4 0  0.21500 0   0
ATOM      24  C    ALA     4    8.41169   1.61877   8.43034 C      3 0  0.52600 0   0
ATOM      25  O    ALA     4    7.47314   2.25182   8.89320 O      1 2 -0.50000 0   0
ATOM      26  CB   ALA     4    9.29130   1.24136   6.00386 C3     4 0  0.03100 0   0
ATOM      27  N    ALA     5    9.59296   1.45484   9.00771 N      3 0 -0.52000 0   0
ATOM      28 HN    ALA     5   10.30552   0.90444   8.52841 H      1 0  0.24800 0   0
ATOM      29  CA   ALA     5    9.90455   2.03608  10.30466 CH     4 0  0.21500 0   0
ATOM      30  C    ALA     5   11.31377   1.69469  10.72617 C      3 0  0.52600 0   0
ATOM      31  O    ALA     5   11.99398   1.00936   9.97549 O      1 2 -0.50000 0   0
ATOM      32  CB   ALA     5    8.76905   1.73557  11.29773 C3     4 0  0.03100 0   0
ATOM      33  N    ALA     6   11.70239   2.17632  11.89776 N      3 0 -0.52000 0   0
ATOM      34 HN    ALA     6   11.04873   2.73863  12.44268 H      1 0  0.24800 0   0
ATOM      35  CA   ALA     6   13.03324   1.92745  12.43061 CH     4 0  0.21500 0   0
ATOM      36  C    ALA     6   13.21004   2.58979  13.77603 C      3 0  0.52600 0   0
ATOM      37  O    ALA     6   12.27148   3.22285  14.23889 O      1 2 -0.50000 0   0
ATOM      38  CB   ALA     6   14.08990   2.21201  11.34970 C3     4 0  0.03100 0   0
ATOM      39  N    ALA     7   14.39130   2.42586  14.35340 N      3 0 -0.52000 0   0
ATOM      40 HN    ALA     7   15.10386   1.87546  13.87411 H      1 0  0.24800 0   0
ATOM      41  CA   ALA     7   14.70290   3.00711  15.65036 CH     4 0  0.21500 0   0
ATOM      42  C    ALA     7   16.11211   2.66572  16.07187 C      3 0  0.52600 0   0
ATOM      43  O    ALA     7   16.79232   1.98038  15.32119 O      1 2 -0.50000 0   0
ATOM      44  CB   ALA     7   13.56740   2.70659  16.64342 C3     4 0  0.03100 0   0
ATOM      45  N    ALA     8   16.50073   3.14734  17.24346 N      3 0 -0.52000 0   0
ATOM      46 HN    ALA     8   15.84708   3.70966  17.78838 H      1 0  0.24800 0   0
ATOM      47  CA   ALA     8   17.83158   2.89847  17.77631 CH     4 0  0.21500 0   0
ATOM      48  C    ALA     8   18.00838   3.56081  19.12173 C      3 0  0.52600 0   0
ATOM      49  O    ALA     8   17.06983   4.19386  19.58459 O      1 2 -0.50000 0   0
ATOM      50  CB   ALA     8   18.88824   3.18302  16.69541 C3     4 0  0.03100 0   0
ATOM      51  N    ALA     9   19.18964   3.39687  19.69910 N      3 0 -0.52000 0   0
ATOM      52 HN    ALA     9   19.90220   2.84648  19.21981 H      1 0  0.24800 0   0
ATOM      53  CA   ALA     9   19.50124   3.97812  20.99606 CH     4 0  0.21500 0   0
ATOM      54  C    ALA     9   20.91045   3.63672  21.41757 C      3 0  0.52600 0   0
ATOM      55  O    ALA     9   21.59066   2.95139  20.66689 O      1 2 -0.50000 0    0
ATOM      56  CB   ALA     9   18.36573   3.67761  21.98912 C3     4 0  0.03100 0    0
ATOM      57  N    ALA    10   21.29907   4.11835  22.58916 N      3 0 -0.52000 0    0
ATOM      58 HN    ALA    10   20.64527   4.68030  23.13428 H      1 0  0.24800 0    0
ATOM      59  CA   ALA    10   22.62992   3.86947  23.12201 CH     4 0  0.21500 0    0
ATOM      60  C    ALA    10   22.80672   4.53181  24.46743 C      3 0  0.52600 0    0
ATOM      61  O    ALA    10   21.86816   5.16486  24.93030 O      1 2 -0.50000 0    0
ATOM      62  CB   ALA    10   23.68658   4.15403  22.04111 C3     4 0  0.03100 0    0
ATOM      63  N    ALA    11   23.98798   4.36786  25.04481 N      3 0 -0.52000 0    0
ATOM      64 HN    ALA    11   24.70039   3.81711  24.56571 H      1 0  0.24800 0    0
ATOM      65  CA   ALA    11   24.29958   4.94910  26.34176 CH     4 0  0.21500 0    0
ATOM      66  C    ALA    11   25.70879   4.60771  26.76328 C      3 0  0.52600 0    0
ATOM      67  O    ALA    11   26.38899   3.92237  26.01260 O      1 2 -0.50000 0    0
ATOM      68  CB   ALA    11   23.16407   4.64859  27.33483 C3     4 0  0.03100 0    0
ATOM      69  N    ALA    12   26.09741   5.08933  27.93487 N      3 0 -0.52000 0    0
ATOM      70 HN    ALA    12   25.44376   5.65164  28.47979 H      1 0  0.24800 0    0
ATOM      71  CA   ALA    12   27.42826   4.84045  28.46773 CH     4 0  0.21500 0    0
ATOM      72  C    ALA    12   27.60505   5.50278  29.81314 C      3 0  0.52600 0    0
ATOM      73  O    ALA    12   26.66663   6.13615  30.27583 O      1 2 -0.50000 0    0
ATOM      74  CB   ALA    12   28.48492   5.12500  27.38682 C3     4 0  0.03100 0    0
ATOM      75  N    ALA    13   28.78618   5.33850  30.39071 N      3 0 -0.52000 0    0
ATOM      76 HN    ALA    13   29.49864   4.78786  29.91154 H      1 0  0.24800 0    0
ATOM      77  CA   ALA    13   29.09776   5.91969  31.68769 CH     4 0  0.21500 0    0
ATOM      78  C    ALA    13   30.50680   5.57787  32.10943 C      3 0  0.52600 0    0
ATOM      79  O    ALA    13   31.18690   4.89229  31.35888 O      1 2 -0.50000 0    0
ATOM      80  CB   ALA    13   27.96201   5.61958  32.68059 C3     4 0  0.03100 0    0
ATOM      81  N    ALA    14   30.89539   6.05942  33.28107 N      3 0 -0.52000 0    0
ATOM      82 HN    ALA    14   30.24183   6.62196  33.82587 H      1 0  0.24800 0    0
ATOM      83  CA   ALA    14   32.22607   5.81014  33.81414 CH     4 0  0.21500 0    0
ATOM      84  C    ALA    14   32.40287   6.47248  35.15955 C      3 0  0.52600 0    0
ATOM      85  O    ALA    14   31.46445   7.10585  35.62225 O      1 2 -0.50000 0    0
ATOM      86  CB   ALA    14   33.28299   6.09433  32.73339 C3     4 0  0.03100 0    0
ATOM      87  N    ALA    15   33.58400   6.30819  35.73712 N    3 0 -0.52000 0   0
ATOM      88 HN    ALA    15   34.29631   5.75719  35.25815 H    1 0  0.24800 0   0
ATOM      89  CA   ALA    15   33.89558   6.88939  37.03410 CH   4 0  0.21500 0   0
ATOM      90  C    ALA    15   35.30462   6.54757  37.45584 C    3 0  0.52600 0   0
ATOM      91  O    ALA    15   35.98472   5.86199  36.70529 O    1 2 -0.50000 0   0
ATOM      92  CB   ALA    15   32.75982   6.58928  38.02700 C3   4 0  0.03100 0   0
ATOM      93  N    ALA    16   35.69321   7.02912  38.62748 N    3 0 -0.52000 0   0
ATOM      94 HN    ALA    16   35.03965   7.59166  39.17227 H    1 0  0.24800 0   0
ATOM      95  CA   ALA    16   37.02390   6.77984  39.16054 CH   4 0  0.21500 0   0
ATOM      96  C    ALA    16   37.20069   7.44218  40.50596 C    3 0  0.52600 0   0
ATOM      97  O    ALA    16   36.26226   8.07555  40.96865 O    1 2 -0.50000 0   0
ATOM      98  CB   ALA    16   38.08081   7.06402  38.07978 C3   4 0  0.03100 0   0
ATOM      99  N    ALA    17   38.38182   7.27790  41.08352 N    3 0 -0.52000 0   0
ATOM     100 HN    ALA    17   39.09428   6.72727  40.60436 H    1 0  0.24800 0   0
ATOM     101  CA   ALA    17   38.69339   7.85910  42.38050 CH   4 0  0.21500 0   0
ATOM     102  C    ALA    17   40.10243   7.51729  42.80225 C    3 0  0.52600 0   0
ATOM     103  O    ALA    17   40.78254   6.83171  42.05170 O    1 2 -0.50000 0   0
ATOM     104  CB   ALA    17   37.55765   7.55898  43.37341 C3   4 0  0.03100 0   0
ATOM     105  N    ALA    18   40.49102   7.99884  43.97388 N    3 0 -0.52000 0   0
ATOM     106 HN    ALA    18   39.83745   8.56137  44.51868 H    1 0  0.24800 0   0
ATOM     107  CA   ALA    18   41.82170   7.74957  44.50695 CH   4 0  0.21500 0   0
ATOM     108  C    ALA    18   41.99849   8.41191  45.85237 C    3 0  0.52600 0   0
ATOM     109  O    ALA    18   41.06007   9.04527  46.31506 O    1 2 -0.50000 0   0
ATOM     110  CB   ALA    18   42.87863   8.03376  43.42620 C3   4 0  0.03100 0   0
ATOM     111  N    ALA    19   43.17962   8.24763  46.42994 N    3 0 -0.52000 0   0
ATOM     112 HN    ALA    19   43.89223   7.69736  45.95058 H    1 0  0.24800 0   0
ATOM     113  CA   ALA    19   43.49119   8.82883  47.72692 CH   4 0  0.21500 0   0
ATOM     114  C    ALA    19   44.90023   8.48702  48.14867 C    3 0  0.52600 0   0
ATOM     115  O    ALA    19   45.58034   7.80145  47.39812 O    1 2 -0.50000 0   0
ATOM     116  CB   ALA    19   42.35544   8.52871  48.71982 C3   4 0  0.03100 0   0
ATOM     117  N    ALA    20   45.28881   8.96857  49.32030 N    3 0 -0.52000 0   0
ATOM     118 HN    ALA    20   44.63524   9.53111  49.86510 H    1 0  0.24800 0   0
ATOM     119  CA   ALA     20  46.61950   8.71930  49.85338 CH     4 0  0.21500 0    0
ATOM     120  C    ALA     20  46.79628   9.38164  51.19880 C      3 0  0.52600 0    0
ATOM     121  O    ALA     20  45.85785  10.01501  51.66148 O      1 2 -0.50000 0    0
ATOM     122  CB   ALA     20  47.67642   9.00349  48.77264 C3     4 0  0.03100 0    0
ATOM     123  OXT  ALA     20  47.82390   9.33251  51.87322 O2     1 2  0.00000 0    0
FORMAT CONECT (a6,14i6)
FORMAT ORDER (a6,i6,13f6.3)
CONECT     1     4     5     2     3
CONECT     2     1
ORDER      2     0
CONECT     3     1
ORDER      3     0
CONECT     4     1
CONECT     5     1     8     6
CONECT     6     7     5     9
ORDER      6     2     0     0
CONECT     7     6
ORDER      7     2
CONECT     8     5
CONECT     9     6    11    10
CONECT    10     9
CONECT    11    14     9    12
CONECT    12    13    11    15
ORDER     12     2     0     0
CONECT    13    12
ORDER     13     2
CONECT    14    11
CONECT    15    12    17    16
CONECT    16    15
CONECT    17    20    15    18
CONECT    18    19    17    21
ORDER     18     2     0     0
```

```
CONECT    19    18
ORDER     19     2
CONECT    20    17
CONECT    21    18    23    22
CONECT    22    21
CONECT    23    26    21    24
CONECT    24    25    23    27
ORDER     24     2     0     0
CONECT    25    24
ORDER     25     2
CONECT    26    23
CONECT    27    24    29    28
CONECT    28    27
CONECT    29    32    27    30
CONECT    30    31    29    33
ORDER     30     2     0     0
CONECT    31    30
ORDER     31     2
CONECT    32    29
CONECT    33    30    35    34
CONECT    34    33
CONECT    35    38    33    36
CONECT    36    37    35    39
ORDER     36     2     0     0
CONECT    37    36
ORDER     37     2
CONECT    38    35
CONECT    39    36    41    40
CONECT    40    39
CONECT    41    44    39    42
CONECT    42    43    41    45
ORDER     42     2     0     0
CONECT    43    42
ORDER     43     2
CONECT    44    41
CONECT    45    42    47    46
CONECT    46    45
CONECT    47    50    45    48
CONECT    48    49    47    51
ORDER     48     2     0     0
CONECT    49    48
ORDER     49     2
CONECT    50    47
CONECT    51    48    53    52
CONECT    52    51
CONECT    53    56    51    54
```

```
CONECT    54    55    53    57
ORDER     54     2     0     0
CONECT    55    54
ORDER     55     2
CONECT    56    53
CONECT    57    54    59    58
CONECT    58    57
CONECT    59    62    57    60
CONECT    60    61    59    63
ORDER     60     2     0     0
CONECT    61    60
ORDER     61     2
CONECT    62    59
CONECT    63    60    65    64
CONECT    64    63
CONECT    65    68    63    66
CONECT    66    67    65    69
ORDER     66     2     0     0
CONECT    67    66
ORDER     67     2
CONECT    68    65
CONECT    69    66    71    70
CONECT    70    69
CONECT    71    74    69    72
CONECT    72    73    71    75
ORDER     72     2     0     0
CONECT    73    72
ORDER     73     2
CONECT    74    71
CONECT    75    72    77    76
CONECT    76    75
CONECT    77    80    75    78
CONECT    78    79    77    81
ORDER     78     2     0     0
CONECT    79    78
ORDER     79     2
CONECT    80    77
CONECT    81    78    83    82
CONECT    82    81
CONECT    83    86    81    84
CONECT    84    85    83    87
ORDER     84     2     0     0
CONECT    85    84
ORDER     85     2
CONECT    86    83
CONECT    87    84    89    88
CONECT    88    87
CONECT    89    92    87    90
CONECT    90    91    89    93
ORDER     90     2     0     0
CONECT    91    90
ORDER     91     2
CONECT    92    89
CONECT    93    90    95    94
CONECT    94    93
CONECT    95    98    93    96
CONECT    96    97    95    99
ORDER     96     2     0     0
CONECT    97    96
ORDER     97     2
CONECT    98    95
CONECT    99    96   101   100
CONECT   100    99
CONECT   101   104    99   102
CONECT   102   103   101   105
ORDER    102     2     0     0
CONECT   103   102
ORDER    103     2
CONECT   104   101
CONECT   105   102   107   106
CONECT   106   105
CONECT   107   110   105   108
CONECT   108   109   107   111
ORDER    108     2     0     0
CONECT   109   108
ORDER    109     2
CONECT   110   107
CONECT   111   108   113   112
CONECT   112   111
CONECT   113   116   111   114
CONECT   114   115   113   117
ORDER    114     2     0     0
CONECT   115   114
ORDER    115     2
CONECT   116   113
CONECT   117   114   119   118
CONECT   118   117
CONECT   119   122   117   120
CONECT   120   121   123   119
ORDER    120     2     1     0
CONECT   121   120
ORDER    121     2
CONECT   122   119
CONECT   123   120
END
```

Figure E.2: Page 165. The BIOGRAF file for $(\mathrm{Gly})_{20}$.

```
BIOGRF  321
DESCRP pG.bgf
REMARK pG with amberv321 autotype.
REMARK terminal residues are neutral.  20 glycines
REMARK modified by rab 9/25/96 11:58 am
REMARK Created by vaid @ degas on  9/25/96    9:32:12
FORCEFIELD AMBER
FORMAT ATOM   (a6,1x,i5,1x,a5,1x,a3,1x,a1,1x,a5,3f10.5,1x,a5,i3,i2,1x,f8.5,f10.5)
ATOM       1  N    GLY     1   -0.70900  -0.41100   0.36100 N      4 0 -0.52000  14.00670
ATOM       2 HN    GLY     1   -0.71275  -1.38002   0.64576 H      1 0  0.08200   1.00800
ATOM       3 HN    GLY     1   -0.72750   0.02577   1.27149 H      1 0  0.08300   1.00800
ATOM       4 HN    GLY     1   -1.62200  -0.32800   0.15300 H      1 0  0.08300   1.00800
ATOM       5  CA   GLY     1    0.29683   0.02531  -0.59554 CH     4 0  0.24600  14.02700
ATOM       6  C    GLY     1    1.68948  -0.20327  -0.05852 C      3 0  0.52600
ATOM       7  O    GLY     1    2.26996  -1.28156  -0.25326 O      1 2 -0.50000
ATOM       8  N    GLY     2    2.22512   0.80451   0.61460 N      3 0 -0.52000  14.00670
ATOM       9 HN    GLY     2    1.68550   1.66182   0.73387 H      1 0  0.24800   1.00800
ATOM      10  CA   GLY     2    3.55968   0.72538   1.18881 CH     4 0  0.24600  14.02700
ATOM      11  C    GLY     2    3.92005   2.00818   1.89920 C      3 0  0.52600
ATOM      12  O    GLY     2    3.12785   2.93911   1.94440 O      1 2 -0.50000
ATOM      13  N    GLY     3    5.12315   2.04725   2.45296 N      3 0 -0.52000  14.00670
ATOM      14 HN    GLY     3    5.73036   1.23145   2.37451 H      1 0  0.24800   1.00800
ATOM      15  CA   GLY     3    5.60444   3.21853   3.16955 CH     4 0  0.24600  14.02700
ATOM      16  C    GLY     3    7.00142   2.99405   3.69694 C      3 0  0.52600
ATOM      17  O    GLY     3    7.58181   1.93670   3.49344 O      1 2 -0.50000
ATOM      18  N    GLY     4    7.53431   3.99914   4.37625 N      3 0 -0.52000  14.00670
ATOM      19 HN    GLY     4    6.98990   4.85187   4.50618 H      1 0  0.24800   1.00800
ATOM       20  CA   GLY      4   8.87246   3.92341   4.94250 CH    4 0  0.24600  14.02700
ATOM       21  C    GLY      4   9.22853   5.20204   5.66251 C     3 0  0.52600
ATOM       22  O    GLY      4   8.43068   6.12743   5.72040 O     1 2 -0.50000
ATOM       23  N    GLY      5  10.43437   5.24384   6.21008 N     3 0 -0.52000  14.00670
ATOM       24 HN    GLY      5  11.04632   4.43269   6.12098 H     1 0  0.24800   1.00800
ATOM       25  CA   GLY      5  10.91209   6.41168   6.93463 CH    4 0  0.24600  14.02700
ATOM       26  C    GLY      5  12.31334   6.19142   7.45241 C     3 0  0.52600
ATOM       27  O    GLY      5  12.89933   5.13970   7.23622 O     1 2 -0.50000
ATOM       28  N    GLY      6  12.84351   7.19375   8.13789 N     3 0 -0.52000  14.00670
ATOM       29 HN    GLY      6  12.29424   8.04159   8.27884 H     1 0  0.24800   1.00800
ATOM       30  CA   GLY      6  14.18519   7.12151   8.69619 CH    4 0  0.24600  14.02700
ATOM       31  C    GLY      6  14.53705   8.39586   9.42580 C     3 0  0.52600
ATOM       32  O    GLY      6  13.73365   9.31555   9.49637 O     1 2 -0.50000
ATOM       33  N    GLY      7  15.74556   8.44046   9.96722 N     3 0 -0.52000  14.00670
ATOM       34 HN    GLY      7  16.36217   7.63408   9.86747 H     1 0  0.24800   1.00800
ATOM       35  CA   GLY      7  16.21980   9.60475  10.69970 CH    4 0  0.24600  14.02700
ATOM       36  C    GLY      7  17.62522   9.38884  11.20790 C     3 0  0.52600
ATOM       37  O    GLY      7  18.21670   8.34288  10.97903 O     1 2 -0.50000
ATOM       38  N    GLY      8  18.15273  10.38832  11.89955 N     3 0 -0.52000  14.00670
ATOM       39 HN    GLY      8  17.59901  11.23149  12.05076 H     1 0  0.24800   1.00800
ATOM       40  CA   GLY      8  19.49787  10.31966  12.44992 CH    4 0  0.24600  14.02700
ATOM       41  C    GLY      8  19.84558  11.58961  13.18913 C     3 0  0.52600
ATOM       42  O    GLY      8  19.03675  12.50345  13.27234 O     1 2 -0.50000
ATOM       43  N    GLY      9  21.05672  11.63707  13.72440 N     3 0 -0.52000  14.00670
ATOM       44 HN    GLY      9  21.67789  10.83559  13.61404 H     1 0  0.24800   1.00800
ATOM       45  CA   GLY      9  21.52752  12.79773  14.46482 CH    4 0  0.24600  14.02700
ATOM       46  C    GLY      9  22.93703  12.58626  14.96347 C     3 0  0.52600
ATOM       47  O    GLY      9  23.53391  11.54622  14.72195 O     1 2 -0.50000
ATOM       48  N    GLY     10  23.46194  13.58283  15.66128 N     3 0 -0.52000  14.00670
ATOM       49 HN    GLY     10  22.90370  14.42103  15.82310 H     1 0  0.24800   1.00800
ATOM       50  CA   GLY     10  24.81047  13.51783  16.20375 CH    4 0  0.24600  14.02700
ATOM       51  C    GLY     10  25.15411  14.78327  16.95253 C     3 0  0.52600
ATOM    52  O   GLY    10   24.33993  15.69112  17.04839 O    1 2 -0.50000
ATOM    53  N   GLY    11   26.36782  14.83368  17.48166 N    3 0 -0.52000  14.00670
ATOM    54 HN   GLY    11   26.99348  14.03723  17.36070 H    1 0  0.24800   1.00800
ATOM    55  CA  GLY    11   26.83525  15.99061  18.23001 CH   4 0  0.24600  14.02700
ATOM    56  C   GLY    11   28.24879  15.78370  18.71910 C    3 0  0.52600
ATOM    57  O   GLY    11   28.85094  14.74974  18.46497 O    1 2 -0.50000
ATOM    58  N   GLY    12   28.77114  16.77729  19.42305 N    3 0 -0.52000  14.00670
ATOM    59 HN   GLY    12   28.20847  17.61040  19.59545 H    1 0  0.24800   1.00800
ATOM    60  CA  GLY    12   30.12300  16.71607  19.95761 CH   4 0  0.24600  14.02700
ATOM    61  C   GLY    12   30.46266  17.97688  20.71594 C    3 0  0.52600
ATOM    62  O   GLY    12   29.64326  18.87859  20.82440 O    1 2 -0.50000
ATOM    63  N   GLY    13   31.67890  18.03031  21.23894 N    3 0 -0.52000  14.00670
ATOM    64 HN   GLY    13   32.30894  17.23902  21.10740 H    1 0  0.24800   1.00800
ATOM    65  CA  GLY    13   32.14304  19.18343  21.99519 CH   4 0  0.24600  14.02700
ATOM    66  C   GLY    13   33.56051  18.98120  22.47474 C    3 0  0.52600
ATOM    67  O   GLY    13   34.16784  17.95345  22.20801 O    1 2 -0.50000
ATOM    68  N   GLY    14   34.08036  19.97174  23.18481 N    3 0 -0.52000  14.00670
ATOM    69 HN   GLY    14   33.51334  20.79963  23.36778 H    1 0  0.24800   1.00800
ATOM    70  CA  GLY    14   35.43548  19.91438  23.71149 CH   4 0  0.24600  14.02700
ATOM    71  C   GLY    14   35.77124  21.17047  24.47935 C    3 0  0.52600
ATOM    72  O   GLY    14   34.94671  22.06588  24.60038 O    1 2 -0.50000
ATOM    73  N   GLY    15   36.98995  21.22698  24.99623 N    3 0 -0.52000  14.00670
ATOM    74 HN   GLY    15   37.62429  20.44096  24.85414 H    1 0  0.24800   1.00800
ATOM    75  CA  GLY    15   37.45085  22.37620  25.76036 CH   4 0  0.24600  14.02700
ATOM    76  C   GLY    15   38.87218  22.17875  26.23040 C    3 0  0.52600
ATOM    77  O   GLY    15   39.48457  21.15736  25.95112 O    1 2 -0.50000
ATOM    78  N   GLY    16   39.38958  23.16617  26.94658 N    3 0 -0.52000  14.00670
ATOM    79 HN   GLY    16   38.81831  23.98873  27.14009 H    1 0  0.24800   1.00800
ATOM    80  CA  GLY    16   40.74789  23.11276  27.46540 CH   4 0  0.24600  14.02700
ATOM    81  C   GLY    16   41.07984  24.36401  28.24275 C    3 0  0.52600
ATOM    82  O   GLY    16   40.25030  25.25299  28.37632 O    1 2 -0.50000
ATOM    83  N   GLY    17   42.30097  24.42368  28.75353 N    3 0 -0.52000  14.00670
ATOM      84 HN    GLY    17   42.93952  23.64306  28.60092 H     1 0  0.24800   1.00800
ATOM      85  CA   GLY    17   42.75872  25.56891  29.52551 CH    4 0  0.24600  14.02700
ATOM      86  C    GLY    17   44.18382  25.37636  29.98605 C     3 0  0.52600
ATOM      87  O    GLY    17   44.80117  24.36147  29.69425 O     1 2 -0.50000
ATOM      88  N    GLY    18   44.69883  26.36058  30.70833 N     3 0 -0.52000  14.00670
ATOM      89 HN    GLY    18   44.12341  27.17769  30.91235 H     1 0  0.24800   1.00800
ATOM      90  CA   GLY    18   46.06027  26.31122  31.21929 CH    4 0  0.24600  14.02700
ATOM      91  C    GLY    18   46.38848  27.55752  32.00612 C     3 0  0.52600
ATOM      92  O    GLY    18   45.55405  28.43992  32.15222 O     1 2 -0.50000
ATOM      93  N    GLY    19   47.61199  27.62042  32.51081 N     3 0 -0.52000  14.00670
ATOM      94 HN    GLY    19   48.25465  26.84533  32.34769 H     1 0  0.24800   1.00800
ATOM      95  CA   GLY    19   48.06664  28.76156  33.29063 CH    4 0  0.24600  14.02700
ATOM      96  C    GLY    19   49.49544  28.57401  33.74170 C     3 0  0.52600
ATOM      97  O    GLY    19   50.11764  27.56580  33.43737 O     1 2 -0.50000
ATOM      98  N    GLY    20   50.00811  29.55497  34.47007 N     3 0 -0.52000  14.00670
ATOM      99 HN    GLY    20   49.42861  30.36649  34.68459 H     1 0  0.24800   1.00800
ATOM     100  CA   GLY    20   51.37261  29.50974  34.97319 CH    4 0  0.24600  14.02700
ATOM     101  C    GLY    20   51.69717  30.75097  35.76949 C     3 0  0.52600
ATOM     102  O    GLY    20   50.85793  31.62664  35.92809 O     1 2 -0.50000
ATOM     103 OXT   GLY    20   52.80207  30.89504  36.27455 O2    2 1  0.00000
FORMAT CONECT (a6,12i6)
CONECT     1     4     5     2     3
CONECT     2     1
CONECT     3     1
CONECT     4     1
CONECT     5     1     6
CONECT     6     7     5     8
ORDER      6     2     1     1
CONECT     7     6
ORDER      7     2
CONECT     8     6    10     9
CONECT     9     8
CONECT    10     8    11
CONECT    11    12    10    13
ORDER     11     2     1     1
CONECT    12    11
ORDER     12     2
CONECT    13    11    15    14
CONECT    14    13
CONECT    15    13    16
CONECT    16    17    15    18
ORDER     16     2     1     1
CONECT    17    16
ORDER     17     2
CONECT    18    16    20    19
CONECT    19    18
CONECT    20    18    21
CONECT    21    22    20    23
ORDER     21     2     1     1
CONECT    22    21
ORDER     22     2
CONECT    23    21    25    24
CONECT    24    23
CONECT    25    23    26
CONECT    26    27    25    28
ORDER     26     2     1     1
CONECT    27    26
ORDER     27     2
CONECT    28    26    30    29
CONECT    29    28
CONECT    30    28    31
CONECT    31    32    30    33
ORDER     31     2     1     1
CONECT    32    31
ORDER     32     2
CONECT    33    31    35    34
CONECT    34    33
CONECT    35    33    36
CONECT    36    37    35    38
ORDER     36     2     1     1
CONECT    37    36
ORDER     37     2
CONECT    38    36    40    39
CONECT    39    38
CONECT    40    38    41
CONECT    41    42    40    43
ORDER     41     2     1     1
CONECT    42    41
```

```
ORDER     42     2
CONECT    43    41    45    44
CONECT    44    43
CONECT    45    43    46
CONECT    46    47    45    48
ORDER     46     2     1     1
CONECT    47    46
ORDER     47     2
CONECT    48    46    50    49
CONECT    49    48
CONECT    50    48    51
CONECT    51    52    50    53
ORDER     51     2     1     1
CONECT    52    51
ORDER     52     2
CONECT    53    51    55    54
CONECT    54    53
CONECT    55    53    56
CONECT    56    57    55    58
ORDER     56     2     1     1
CONECT    57    56
ORDER     57     2
CONECT    58    56    60    59
CONECT    59    58
CONECT    60    58    61
CONECT    61    62    60    63
ORDER     61     2     1     1
CONECT    62    61
ORDER     62     2
CONECT    63    61    65    64
CONECT    64    63
CONECT    65    63    66
CONECT    66    67    65    68
ORDER     66     2     1     1
CONECT    67    66
ORDER     67     2
CONECT    68    66    70    69
CONECT    69    68
CONECT    70    68    71
CONECT    71    72    70    73
ORDER     71     2     1     1
CONECT    72    71
ORDER     72     2
CONECT    73    71    75    74
CONECT    74    73
CONECT    75    73    76
CONECT    76    77    75    78
ORDER     76     2     1     1
CONECT    77    76
ORDER     77     2
CONECT    78    76    80    79
CONECT    79    78
CONECT    80    78    81
CONECT    81    82    80    83
ORDER     81     2     1     1
CONECT    82    81
ORDER     82     2
CONECT    83    81    85    84
CONECT    84    83
CONECT    85    83    86
CONECT    86    87    85    88
ORDER     86     2     1     1
CONECT    87    86
ORDER     87     2
CONECT    88    86    90    89
CONECT    89    88
CONECT    90    88    91
CONECT    91    92    90    93
ORDER     91     2     1     1
CONECT    92    91
ORDER     92     2
CONECT    93    91    95    94
CONECT    94    93
CONECT    95    93    96
CONECT    96    97    95    98
ORDER     96     2     1     1
CONECT    97    96
ORDER     97     2
CONECT    98    96   100    99
CONECT    99    98
CONECT   100    98   101
CONECT   101   102   103   100
ORDER    101     2     1     1
CONECT   102   101
ORDER    102     2
CONECT   103   101
END
```

Appendix F Trajectory Generation: Examples of Input and Output Files

An example of a macro file that generated one of the 155 simulations of $(Ala)_{20}$ is file pAnc-450K-a-15A.macro (/ul/rbertsch/bgf/nc-450K-a-15A/pAnc-450K-a-15A). It is included here as Figure F. This generates two data files: a trajectory file called "pAnc-450K-a-15A.trj" and a "tor" file called "pAnc-450K-a-15A.tor."

The trajectory file. pAnc-450K-a-15A.trj is a binary file with the coordinates of each atom, the energies, temperature, and other parameters every 0.1 ps. Its average size for 500 ps of simulation of $(Ala)_{20}$ was 7.9 MB, after compressing with "gzip." Both BIOGRAF and Cerius(2)[1] read, animate, and analyze the trajectory file. The trajectory file can be read by converting it from binary to ASCII via the BIOGRAF program rdtrj330_iris. It is available in the Goddard group by typing

```
$BG_EXE/rdtrj330_iris
```

A FORTRAN subroutine that reads the binary file is /ul/rbertsch/bin/HB/HB-read-traj.f, modified from a subroutine courtesy of Molecular Simulations Inc[2]. FORTRAN77 programs that read and analyze the trajectory files are listed in Appendix G.

The tor file. The tor file is short for "torsions." It is an ASCII file of every torsional degree of freedom at each time point that is written out. In the case of the $(Ala)_{20}$ and $(Gly)_{20}$ runs, the tor files contained 59 ϕ, ψ, and ω dihedrals at each 0.1 ps. FORTRAN programs that analyze the tor file are listed in Appendix G.

[1] Cerius^2, Version 3.0. Copyright 1997 by Molecular Simulations Inc., 9685 Scranton Road, San Diego, CA 92121. http://www.msi.com/.

[2] See Footnote 1.

Figure F.1: A Macro file to generate a trajectory of $(\text{Ala})_{20}$ folding.

```
beginmacro
%
% version      : 3.21
% version date : 21:49:53 4/30/93
% link date    : 16:33:53 11/17/93
%
% Macro created on 11/19/93   10:18:09
%
% program started with an initialization file
%
Top menu/in-out
   In-Out/read
   File types/BioDesign
     "pAnc2.bgf"
   In-Out/return
Top menu/simulate
   Simulate/update eex
   Simulate/defaults
    Defaults/neimo var
%        Restart Neimo
%        The name of the dihedral file of velocities
%        "pAnc-450K-15A.dhv"
%        The exact time of the starting place
%         "195.9716"
%        Rescale the velocities to some other temperature?
%          "1"
         mod random seed
           "12345"
%          Input 0 for the computer to select the random seed
         Write Torsions
          "pAnc-450K-15A.tor"
%        Writes dihedral angles and their velocities
%        The command has to be 'Write dih vel' if run in /v/ubio_run
%        but should be 'Write dih' if run under ~vaid/ubio_run
%         Write dih vel
%           "pAnc-450K-15A.dhv"
          Return
      defaults/misc var
        misc var/NEIMO dynamic
      misc var/return
      Defaults/return
   Simulate/defaults
      Defaults/nonbond var
         Nonbond var/Nonbond cut:
         Nonbond var/nblst cut
           "15.0"
         Nonbond var/splin on
           "14.0"
         Nonbond var/splin of
           "14.5"
         Nonbond var/return
      Defaults/return
   Simulate/dynamics
        Dynamics/canonical (TVN)
         Canonical/write traject
         Canonical/tau(can)
           "0.100"
         Canonical/temperatur var
            Temperatur var/temp assignmnt
            Temperatur var/initial temp
              "450"
            Temperatur var/final temp
              "450"
            Temperatur var/return
         Canonical/dynamics var
            Dynamics var/time step
               "0.010"
            Dynamics var/return
            Canonical/time
             "500.0"
         Canonical/execute
             "pAnc-450K-15A.trj"
             " electrostatic and VdW interactions cutoff at 15 A"
             " 12345 seed "
             " "
  Canonical/return
Dynamics/return
Simulate/return
Top menu/exit
  "OK"
%
endmacro
```

Appendix G Analysis Programs and Scripts

Because neither BIOGRAF nor Cerius^2 have accessible, efficient analyze programs specifically for proteins, I wrote a number of off-line FORTRAN programs to do very specialized analyses. The important FORTRAN programs are summarized in Table G.1.

Note that FORTRAN programs using the trajectory file for input are specific to $(\mathrm{Ala})_{20}$ because the atoms numbers are hardcoded to match the bgf file. To make the code general for all sequences, a "read-bgf.f" subroutine should be incorporated. Code relying on the torsion file as input is specific to any eicosamer (20-mer) with only three degrees of freedom per residue, i.e., only the ϕ, ψ, and ω dihedrals. In other words, code analyzing the tor file will work for any combination of 20 alanines and glycines.

Table G.2 charts the important scripts I used. The scripts in this table usually initiate a FORTRAN program to make an ASCII data file and then write a file instructing Gnuplot to turn the data into a graph (a *.dem file). Table G.3 documents the scripts written to analyze or summarize multiple runs.

In addtion, there are other scripts located in ~/bin, one of its subdirectories, or ~/bgf/prog. In general, if I had to do something, I probably wrote a script for it. If you think I have done something, try searching for a script called *.awk, *.nawk, or *.csh to do the job.

Table G.1: FORTRAN programs to analyze the simulations.

Example	Description	Executable	Source Code	Directory
	Analyses from the beginnings of Chapters 2 and 3			
Figure 2.2, Figure 3.2	Scroll of α-helical residues	helix-coil-v-time.e	helix-coil-v-time.f	~/bin
Figure 3.3, Figure 2.3	Stacked plots of each residue	phi-psi-v-time.e	phi-psi-v-time.f	~/bin
Figure 2.4, Figure 3.4	Radius of gyration End-to-end distance	anal-rad-inert.e	Make file make-anal-rad-inert	bin
Figure 2.7, Figure 3.7	Energies vs time	anal-rad-inert.e	Make file make-anal-rad-inert	bin
	Ramachandran plots			
Figure 2.5, Figure 3.5	All 20 residues at one time	phi-psi-sort.e	phi-psi-sort.f	~/bin
Figure 2.6, Figure 3.6	Trajectory of the (ϕ, ψ) angles of one residue	phi-psi-by-aa.e	phi-psi-by-aa.f	~/bin
	Hydrogen bonds			
Figure 2.9, Figure 2.10	Energies of a single HB	HB.e, 'res' option	See HB-main.f & make-HB-2 & make-HB-2-works	~bin/HB
Figure 2.11, Figure 3.8	Totals of each type of HB	HB.e, 'typ' option	HB-main.f, etc.	~/bin/HB

Table G.2: Scripts and Gnuplot instructions for the FORTRAN programs.

Description	Executable	Script[a]	Gnuplot File[b]
Analyses from the beginnings of Chapters 2 and 3			
Scroll of α-helical residues	helix-coil-v-time.e	None written.	Not necessary.
Stacked plots of each residue	phi-psi-v-time.e	~/bin/analyze	a50vt.dem
Radius of gyration	anal-rad-inert.e	~/bin/analyze	~/bgf/nc-450K-ad-15A/ener/ pAnc-450K-ad-15Adistps.dem
End-to-end distance			~/bgf/nc-450K-ad-15A/ener pAnc-450K-ad15Adist.tex
Energies vs time	anal-rad-inert.e	~/bin/analyze	pAnc.ener.dem ener.tex
Ramachandran plots			
All 20 residues at one time	phi-psi-sort.e	~/bin/analyze	ppsort.dem ppsort-750.tex
Trajectory of the (ϕ, ψ) angles of one residue	phi-psi-by-aa.e	~/bin/analyze	a10ppa.dem
Hydrogen bonds			
Energies of a single HB	HB.e, 'res' option	~/bin/csh/HB.csh, ~/bin/csh/make-AMB.csh	AMB.dem
Totals of each type of HB	HB.e, 'typ' option	~/bin/csh/make-HB-types-demo.csh, ~/bin/csh/make-HB-types-tex.csh	HB-type.dem

[a] A script that runs the FORTRAN program and prepares the Gnuplot graphs. *.csh scripts will be in the directory with the FORTRAN code or in /ul/rbertsch/bin/csh.

[b] Contained in the~/bgf/demos directory, unless specified otherwise. Sometimes a script combines the Gnuplot files using a TEX file. See Appendix H.

Table G.3: Scripts to run, analyze, and summarize multiple runs.

Script	Directory	Description
startbgf	~/bin	Starts and finishes a BIOGRAF job
analyze	~/bin	Runs five sets of analyzes for a given simulation[a]
lpr-sum.csh	~/bin	Prints results of analyze script
analyze-runs2.csh, compile-results.csh	~/bgf/prog	Reads lists of runs and their statistics from file ~/bgf/prog/pAnc-450K-xx-15A.sum, iterates one type of analysis on all the simulations in the file, and then outputs a summary
run_queue.csh	~/bin/runs	Keeps computers busy running my jobs
check-runs.csh, small-check-runs.csh[b]	~/bin/runs	Displays data on multiple runs

[a] See Table G.

[b] These are early scripts and best ignored. Try analyze-runs2.csh and compile-results.csh instead.

Appendix H How the Figures Were Created

H.1 Figures of molecules produced in Showcase

Figures of molecules against grey backgrounds were probably produced using a combination of BIOGRAF and Showcase. This method is especially good at shading the molecule and giving it depth, even without stereo viewing. Examples of figures done this way are Figures 2.1, 3.1, 4.3, 4.4, and 4.11.

The figures of the protein were first made in BIOGRAF, then turned into an *.rgb file via snapshot, imported into Showcase, and finally output as an unencapsulated postscript file.

1. BIOGRAF.[1] Conformations of $(\mathrm{Ala})_{20}$ were first extracted from the trajectory and then output as a *.bgf file. The *.bgf file was modified so that hydrogen bonds could easily be illustrated in black and white. The goal was to make hydrogens white, oxygens dark, carbons and nitrogens dark grey, and the background light grey. The background cannot be white in order for the nitrogens to be visible. The *.bgf file was modified with the sed script /ul/rbertsch/bin/sed/convert-bgf-to-photo-bgf.sed. See Figure H.1. The script changes the atom labels and types, so do not use the resulting file for simulations.

   The new *.bgf file, usually named *-NCON.bgf, was then loaded into BIOGRAF and "rendered" under the "visualize" menu as "cylinders" of scale 0.2

[1] Version 330 of BIOGRAF was used. This is a version the Goddard group produced and is not commercially available. See William A. Goddard, III, Materials Simulation Center, Beckman Institute, California Institute of Technology, Pasadena, CA 91125.

with "half-bonds." The background color was set in the utilities menu to 0.80 grey. Snapshot then turned the image into an rgb file I called *-NCON.rgb.

Note that stereo images are not set to the default mode. The default mode claims it is "distal," but in fact the default mode is for crossed-eyed stereo. To show the molecule in relaxed-eyed stereo, click the stereo/mode button until the mode reads "proximal."

2. Showcase.[2] The rgb files are loaded into Showcase using the "insert image" option under the file menu. Most annotations were done in Helvetica, 18 point, black text. Hydrogen bonds were illustrated using a line width of 0.5 points and the first set of dashes (not dots) on the "Master Gizmo." After creating the entire figure, the background was made grey by importing an rgb file of the grey background created in BIOGRAF and putting the rgb file at the bottom of the stack of other images in the Showcase file. I used the file /ul/rbertsch/bgf/grey-background.rgb.

   To incorporate the figures into LaTeX2$_\epsilon$, postscript files were output, one Showcase page per postscript file. The Showcase ps files were then modified with the following sed command (typed all on one line):

```
sed -e 's/newpath clippath pathbbox/0.0 0.0 612.38 792.022
                     % newpath clippath pathbbox/g'
                     showcase_file.ps > LaTeX2e-able_file.ps
```

   Table 2 lists the versions of Showcase that were used to build the figures.

H.2 Sets of graphs from Gnuplot

Figures 2.6, 2.5, and 4.13 were multiple Gnuplot[3] graphs combined into one postscript file using the "special" command in TeX and dvips555 on sgi1. The TeX files are in-

[2] See Table 2.

[3] Gnuplot, UNIX version 3.5, patch level 3.50.1.17, August 27, 1993, copyright 1993 by Kelley, C. & Williams, T.

Table H.1: The versions of Showcase used to build the figures.

SGI machine	Version of Showcase	Version of IRIX
teijin	IRIS Showcase 3.3.3	IRIX 5.3
impacts	IRIS Showcase 3.3.3	IRIX 6.2
octane1	IRIS Showcase 3.4	IRIX 6.4

Figure H.1: Sed script to change colors on *.bgf files.

```
#!/usr/bin/sed -f

# Converts a bgf file into something that BIOGRAF can make pretty
# pictures out of.  Eliminates the Nitrogens to reduce one color.
# Important for black and white 3-D rendering as cylinders at 0.2
# thickness.

s/ N     / C     /g
s/ O2 / N  /g
s/ O     / N     /g
```

cluded in each subdirectory where the single postscript files were created. "dvips555" is a wag system alias for dvips, version 5.55.[4]

H.3 Graphs from KaleidaGraph

Two figures, Figures 2.11 and 3.8, were made in KaleidaGraph.[5] See Table H.3 for approximate guides to line styles.

The graphs were saved to a file in the print. Under "postscript job," encapsulated postscript "no preview" was chosen. The resulting postscript file was modified with the following command:

```
perl -pe 's/\r/\n/g'
```

[4] dvips, version 5.55. Copyright 1986 & 1994 by Radical Eye Software.

[5] KaleidaGraph, version 3.0.2. Copyright 1993 by Abelbeck Software.

Table H.2: KaleidaGraph line styles for Figures 2.11 and 3.8.

Type of HB	Line Style[a]	Line Width[b]
$i, i+2$	5	1
i, i+3–5	1	3
nonlocal	10	2

[a] Counting from the top
[b] As numbered by KaleidaGraph

Finally, the postscript file was included into the LaTeX2ε document using the includegraphics* command. Remember that the directory in which the KaleidaGraph files were produced is /ul/rbertsch/mac_files/thesis.

H.4 Graphs from Microsoft Word

Tables output as postscript files by Microsoft Word[6] had to be modified to be included in the LaTeX2ε document. Usually, the following perl[7] command was sufficient.

```
perl -pe 'if (/^%%BeginSetup/ .. /^%%EndSetup/)
                { $_ = ``%$_''; }'
                MSWord_file.ps > LaTeX2e-able_file.ps
```

If that does not work on a particular version of UNIX, the script~/bin/perl/MSWord-2-Latex.perl should work. See Figure H.2. Then, the new postscript file can then be included into the LaTeX2ε document with an includegraphics* command.

[6] Microsoft Word 97. Copyright 1983, 1996 by Microsoft Corporation.
[7] perl, version 4.0. Revision 4.0.1.8, 1993. Copyright 1991 by Larry Wall.

Figure H.2: Perl script to convert MSWord postscript files to files LaTeX2$_\epsilon$ can interpret.

```
#!/usr/local/bin/perl

# Convert a ps file from MS Word, version 97 to something
# the includegraphics* command in LaTeX2e can process
# Should work on all architectures with perl version 5.0 and
# higher

while ($_ = <>) {
    if (/^%%BeginSetup/ .. /^%%EndSetup/) { $_ = "%$_"; }
    print $_;
}
```

H.5 NCAR graphics

Contour plots such as Figures 4.15 and 4.14 were produced by NCAR graphics.[8] See the MSC documentation files on NCAR graphics for instructions on making a contour plot without annotations. I added landmarks and "H"s and "L"s by editting the postscript file manually.

H.6 Location of the graphs

To edit or reproduce the graphs in the thesis, find the original directories in which they were created. The /ul/rbertsch/thesis/Figs directory and subdirectories has soft links from the names of the figures as listed in the thesis LaTeX2$_\epsilon$ file to the directories in which the data and the Gnuplot[9] scripts reside and where the graphs were assembled.

[8] NCAR Graphics (1989) Copyright by University Corporation for Atomspheric Research. Published by National Center for Atmospheric Research, Scientific Computing Division, P.O. Box 3000, Boulder CO, 80307-3000. Version 3.00 for UNIX.

[9] See Footnote 3.

Index