INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

RICE UNIVERSITY

A Flexible Lattice Model to Study Protein Folding

by

NICOLE LYNNE NUNES

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE
DOCTOR OF PHILOSOPHY

APPROVED, THESIS COMMITTEE:

John S. Hutchinson, Director
Associate Professor of Chemistry

John L. Margrave
Professor of Chemistry

Steven J. Cox
Associate Professor of Computational and
Applied Mathematics

Houston, Texas

April, 1996

UMI Number: 9631077

UMI Microform 9631077
Copyright 1996, by UMI Company. All rights reserved.
This microform edition is protected against unauthorized
copying under Title 17, United States Code.

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

A Flexible Lattice Model to Study Protein Folding

by

NICOLE LYNNE NUNES

ABSTRACT

The biological activity of protein molecules is central to life. It has been known for decades that this biological activity is dependent on the protein molecule achieving a particular geometric conformation. Simple lattice models have been developed to investigate the protein folding pathway since all-atom molecular dynamics simulations on the time scale of folding are beyond the current capabilities of computers. We present a new Monte Carlo lattice model to study the folding of heteropolymer chains. Previous lattice model studies of two-dimensional chains have been performed on square grids using pre-defined "move sets" of allowed moves. The motion of the polymer chain in these models is thus highly constrained. In order to add greater flexibility, we use a triangular lattice and allow the chain to choose its own moves. Physically unrealistic moves are prevented by including kinetic energy effects in the Metropolis algorithm. By looking at the results, we are able to characterize all of the possible one and two particle moves in the two-dimensional model and sort them according to relative importance. This information will be used to guide simplified molecular dynamics studies. We also find that the initial phase of the folding process is a rapid collapse to a relatively compact state which is entropically driven. The model has been extended to three dimensions. The increased difficulty of working in three-dimensions as well as the preliminary results will be discussed.

Acknowledgments

One does not get to this point alone. Many thanks to

My advisor, Dr. John Huthchinson, for letting me be a part of his research group and for allowing me the flexibility to complete my MAT as well. You are a true mentor.

Kaiqi Chen for his help in the development of this model. As well as the other members of the group, Karin Wright and Martin Holt, all of whom have been excellent co-workers.

Larry Lewis for sharing my vision of what this department could be and for working with me to help us get there.

Cass Sackett, Greg Odom, Mark Anderson, and Ian McAlexander for discussing my research with me and offering helpful suggestions and support.

Jenifer and Colin Monks for being my expert biologists as well as providing support and encouragement along with Marie Villarba and Tish Poland.

Contents

Abstract	ii
Acknowledgments	iii
List of Figures	v
1 Introduction	1
2 Development of a New Model	8
2.1 Chain Motion	11
2.2 Chain Energy	14
3 Two-Dimensional Model	18
3.1 The Model	18
3.2 Results and Discussion	19
3.2.1 Chain Motion	20
3.2.2 Folding Time	26
3.2.3 Entropy	27
4 Three-Dimensional Model	33
4.1 The Model	33
4.2 Results and Discussion	34
5 Conclusion	37
Bibliography	39

List of Figures

1.1	A small portion of an extended protein molecule[1b].	2
1.2	The Framework Model and the Collapse Model for protein folding[6].	5
2.1	A sequence used in Dill's model[9].	8
2.2	The three allowed moves in Dill's 2D square lattice model[9].	9
2.3	Examples of chain motion showing a one monomer move, a two monomer move and a seven monomer move.	13
3.1	The native state of sequence S1.	18
3.2	The directions of allowed motion.	19
3.3	The percentage of N monomer moves accepted for sequence S1 in 2D.	21
3.4	The two types of one monomer moves.	21
3.5	The possible two monomer moves in 2D.	22
3.6	Percentage of each type of two monomer move accepted for sequence S1 in 2D.	24
3.7	The percentage of moves accepted as a function of time for sequence S1 in 2D.	25
3.8	Folding time as a function of temperature for sequence S1 in 2D. . . .	26
3.9	The configuration enumeration scheme used to calculate the entropy.	27
3.10	The degeneracy of states for sequence S1 in 2D.	29
3.11	The entropy of sequence S1 in 2D.	29
3.12	The native state of sequence S1a in 2D.	31
3.13	The degeneracy of states for sequence S1a in 2D.	32
3.14	The entropy of sequence S1a in 2D.	32
4.1	The native state of sequence S7 in 3D.	33

4.2 The percentage of N monomer moves accepted for sequence S7 in 3D.	35
---	----

Chapter 1

Introduction

Protein molecules are central to life[1]. Not only are proteins the building blocks of our bodies, they also have numerous active roles that allow us to live. Collagen is a fibrous protein that gives cartilage and bone its strength. Actin and myosin are the major proteins involved in muscle movement. $\text{Na}^+ \text{-K}^+$ ATPase is a membrane transport protein which is used by nerve cells to pass signals throughout the body. Immune system proteins such as the major histocompatibility complex and immunoglobins allow the body to recognize foreign substances and thus fight off disease. Hemoglobin and myoglobin are globular proteins that can bind oxygen and transport it to our cells. These are just a few of the functions that protein molecules perform in our bodies. The key to a protein molecule's ability to perform its specific function is its three dimensional structure.

A protein is a very large molecule typically containing thousands of atoms. It is a polymer composed of amino acids. There are twenty different amino acids which make up all protein molecules. This sequence of amino acids making up a protein is called the primary structure. Amino acids all have some structure in common which makes up the protein backbone. The side chains are the part of the amino acids that distinguish them from one another and give them their unique chemical properties. A small segment of a protein molecule is shown in Figure 1.1. The shaded region represents the protein backbone. The general spatial arrangement of the protein's backbone atoms, such as alpha helices and beta sheets, is called the secondary structure. The three-dimensional structure of the entire protein is called the tertiary structure. Some proteins are made up of several separate chains, or subunits, that associate with each other. The arrangement of these subunits is the quaternary structure. The atomic interactions that are important in protein structure

Figure 1.1 A small portion of an extended protein molecule[1b].

are the noncovalent interactions, especially hydrogen bonding and solvent effects.

The functionality of a protein is governed by the specific three-dimensional structure it adopts, the tertiary structure. This functional state of the protein is referred to as the "native state". The mechanism by which a protein molecule adopts this native state is called protein folding. It is not known whether the native state is the lowest energy state (and therefore thermodynamically determined) or whether it is a local minimum (kinetically determined). Most proteins spontaneously fold into the native state conformation under physiological conditions. This has been shown through denaturation/renaturation experiments. For example, in 1957 Christian Anfinsen [1b] completely unfolded the RNase A protein in a solution of urea and 2-mercaptoethanol. In this solution, the protein is inactive. After returning the protein to physiological conditions, the enzymatic activity of the protein returns. Note that physiological conditions are necessary for protein activity. This illustrates the importance of the solvent in protein folding. Additional proof of the spontaneity of protein folding was shown in experiments in which RNase A was chemically synthesized from the amino acids and was found to be enzymatically active under physiological conditions.

The proteins that cannot refold after denaturation are proteins that are modified after folding, such as insulin. The initial protein that leads to insulin is called proin-

sulin. Once proinsulin has reached its folded state, a segment made up of thirty-three amino acids is removed from the protein by an enzyme, and insulin is left. If the insulin molecule is denatured, it can not refold into the native configuration.

The fact that proteins spontaneously fold has led to the two main questions that are at the forefront of protein folding research. How does the primary sequence encode the three-dimensional structure of a protein, and what is the mechanism by which a protein molecule folds into its native conformation?

These questions are difficult to answer for a variety of reasons. First, protein molecules are composed of thousands of atoms so any attempt at a direct quantum mechanical calculation is not solvable. It is not obvious how to write a potential for a protein molecule, especially since the solvent is an integral part of the system and must therefore be taken into account. Also, the stabilization energy of the native state of a protein does not exceed 50 KJ/mol[2]. So, to give an accurate result, the potential must be extremely accurate, since it requires summing of a very large number of interactions whose energies are of the same order of magnitude as that of the final stabilization. Computing power is also a limitation. An all-atom molecular dynamics simulation of a protein made up of one hundred amino acids would require approximately 10^7 years to complete since the time scale of protein folding is on the order of seconds whereas step-size is approximately 10^{-15} seconds[3]. Finally, proteins have been evolutionarily selected, and so are probably "special" sequences. It is important to keep this in mind when working with random sequences.

Despite these difficulties, progress has been made on both questions. Much effort has been focused on developing a program which could correctly predict the three-dimensional structure of a protein given the primary structure. Two recent programs, LINUS and the Sun-Thomas-Dill program, have been developed which predict folded protein structures[4]. LINUS, a program developed at Johns Hopkins, predicts secondary structure very well, but is not able to correctly predict tertiary structure. The

Sun-Thomas-Dill program does just the opposite. It accurately predicts tertiary structure, but has problems predicting secondary structure, so it uses known secondary structure. These models use an energy scheme based on hydrophobic interactions and hydrogen bonding. The success of these models shows that hydrophobic interactions and hydrogen bonding do indeed play a major role in the folding of proteins. However, we are still quite a ways away from being able to predict tertiary structure from the amino acid sequence. Additionally, these programs do not attempt to explain the mechanism by which the protein reaches the native state.

The fact that a special mechanism for protein folding must exist was concluded by Cyrus Levinthal[1b] who showed that it is impossible for folding to occur via a random search of all possible conformations because it would take even a small size protein an outrageously long time, on the order of billions of years, to search all of the conformations. Protein molecules fold with a time scale on the order of seconds, so there must be some mechanism that allows the protein to reach the native state quickly. An understanding of the protein folding process would lead to great advances in medicine, since new drugs and vaccines could be developed that are modelled on proteins that already exist and function within the body. Additionally, amyloid diseases, such as Alzheimer's, are thought to be caused by proteins that misfold and form aggregates instead of folding into the usual active structure[5]. An understanding of the folding process could lead to treatments for such diseases. Great advances in chemistry could also result from an understanding of this process, especially because of the insight it would give into noncovalent interactions. This would be important in the nanotechnology area since proteins are the prime example of nano-machines — molecules that self-assemble and perform very specific functions.

There have been two main theories for the mechanism of protein folding[6]. The pathways predicted by these theories are shown in Figure 1.2. The Framework Model is based on the idea that local interactions dominate and secondary structure for-

Figure 1.2 The Framework Model and the Collapse Model for protein folding[6].

mation precedes collapse. In the Collapse Model, the nonlocal interactions dominate and collapse drives secondary structure formation.

As in the case of three-dimensional structure prediction, simple models have also been developed to investigate the mechanism of protein folding. The most simple models are the lattice models[6-16]. In these models, the protein is represented as a chain of monomers, where each monomer represents an amino acid. The bond lengths between amino acids are fixed, and each monomer is allowed to occupy one lattice site on a two or three dimensional grid. Usually the grid is square[7, 8] or cubic[9, 10], although work has been done on developing other types of lattices that accurately fit the known geometries of real protein molecules[11, 12]. There is also a variety of definitions of monomer types. The simplest model involves only two types of monomers, hydrophobic (H) and hydrophilic (P)[6, 7, 8]. Other models have been developed which include neutral monomers as well [12]. Some models use twenty different types of monomers, corresponding to the twenty amino acids[16]. Once the chain is defined, it is allowed to move on the lattice via a Monte Carlo simulation.

The moves are chosen randomly from a predefined set of allowed moves called a “move set”. The probability of acceptance for a move is determined using the Metropolis criteria[17] which is shown in Equation 1.1.

$$Prob = \begin{cases} 1 & \text{if } \Delta E \leq 0 \\ e^{-\Delta E/kT} & \text{if } \Delta E > 0 \end{cases} \quad (1.1)$$

All moves that decrease the energy are accepted, while moves that increase the energy are accepted with a probability based on the magnitude of the energy change and the temperature.

These simple lattice models are extremely useful in the study of protein folding. The main reason is that for the simplest models it is possible to enumerate all of the possible chain configurations. This allows thermodynamic information to be calculated directly from the partition function. The thermodynamic information can be related to the kinetic information obtained from the Monte Carlo simulations. Information gained from these models gives insight into the protein folding problem. Also, lattice models can be easily modified to study other aspects of proteins such as the effects of side chains on protein folding[18] or the effect of ligand binding[19].

In order for lattice models to give meaningful results, it is important that they are physically realistic. Obviously, lattice models are very simplified models of proteins and protein motion. For example, a real protein molecule has much more freedom of motion than is allowed in a simple lattice model using a move set. Also, the energy of a real protein molecule is a very complex function, involving many different types of interactions.

With this in mind, we set out to develop a new lattice model that would take these concerns into account, while still remaining computationally accessible. To create more flexibility of motion, we allow the chain to choose its own moves on a triangular lattice. To prevent unrealistic moves, we have included kinetic energy effects in the Metropolis algorithm. In addition, the solvent is an integral part of protein folding, so we have derived an energy scheme which takes this into account.

These changes have resulted in several significant and interesting results. By allowing the chain flexibility of motion, we have been able to investigate the types of motions that are predominate as the chain moves, as well as how the chain motion changes as a function of time. We also find that the changes we have made to the energy function have resulted in a more physical entropy function which shows that the initial collapse of the chain is entropically driven.

The following chapters discuss the lattice model that we have developed in order to investigate the important molecular motions involved in protein folding. A description of how we developed our model is given in Chapter 2. The details and results for the two-dimensional case and the three-dimensional case are given in Chapters 3 and 4 respectively. A summary and the conclusions are given in Chapter 5.

Chapter 2

Development of a New Model

The starting point from which we developed our model was the two-dimensional HP model developed by Dill and co-workers[8]. In their model, the protein is represented as a chain of H and P monomers. Each monomer is allowed to occupy one lattice site on a square lattice. The native state of one of their sequences is shown in Figure 2.1. The energy of a given configuration is given by the number of HH contacts.

Figure 2.1 A sequence used in Dill's model[9].

A contact occurs when two monomers that are not covalently linked occupy adjacent lattice sites. The chain is allowed to move on a two dimensional square lattice using the move set shown in Figure 2.2. Note that the motion of the chain is quite limited.

Figure 2.2 The three allowed moves in Dill's 2D square lattice model[9].

A completely unfolded linear chain would only be able to begin motion from either end of the chain. Intuitively, it seems that the motion of a real protein molecule could be initiated from any monomer in the chain. Additionally, the results of the simulations are affected by the choice of move set. As Dill states, “exact studies also show that energy landscapes inferred from the models are strongly dependent on the choice of move sets, which are arbitrary constructs, and thus Monte Carlo dynamics must be interpreted with caution”[8].

To avoid these problems as well as to allow more flexibility of motion, we envisioned a model in which the chain was allowed to choose its own moves. When thinking about how motion takes place, we realized that kinetic energy has an important role to play in chain motion, and it has been neglected in these models. Using

potential energy considerations alone, the probability of a move taking place is independent of the number of monomers moving. It is physically unreasonable to assume that the probability for a single monomer to move is equal to the probability for many monomers to move together since a large amount of kinetic energy is required for a large number of monomers to move together. This is especially true in modelling the dynamics of protein folding, since additional kinetic energy is required for a monomer to move solvent molecules out of the way. Also, chain motion is very limited on a square lattice. A lattice with a higher coordination number allows more flexibility, so we perform our simulation on a triangular lattice.

The second area of concern for us was the energy function. Only including the HH contacts gives the impression that there is an attractive force between hydrophobic monomers. Although it is true that hydrophobic polymers collapse when placed in water, this effect, called the hydrophobic effect, is not due to an attraction between the hydrophobic monomers. The driving force of the effect is based on the solvent-solvent interactions. Water molecules have a strong tendency to hydrogen bond with each other. It is the fact that the strongest interaction in the protein/solvent system is the hydrogen bonding of the water molecules together which leads to the collapse. The water would like to have the polymer out of its way in order to maintain the maximum number of hydrogen bonding interactions. To address this point, we wished to develop an energy scheme that gives a clearer picture of driving force of collapse. We include all of the different types of interactions in the system: hydrophobic-hydrophobic, hydrophilic-hydrophilic, hydrophobic-hydrophilic, hydrophobic-solvent, hydrophilic-solvent, and solvent-solvent, to derive our energy function.

These changes can be made with a minimal cost in computational effort. The details of how these changes were made are explained below.

2.1 Chain Motion

To incorporate kinetic energy effects into the chain motion algorithm, we introduce for each monomer j a pseudomomentum vector, $\vec{p}_j$, pointing from the previous lattice site of the monomer, $\vec{r}_j^o$, to its current lattice site, $\vec{r}_j$:

$$\vec{p}_j = \vec{r}_j - \vec{r}_j^o \quad (2.1)$$

Defined in this way, the pseudomomentum is simply a memory of the previous move. As described in detail shortly below, we use the pseudomomentum to incorporate kinetic energy into the Monte Carlo move sampling.

We also use the pseudomomentum to include inertial effects as the chain moves. A particle in solution, including a monomer in a protein chain, which has just completed a move in one direction should preferentially continue moving in that same direction and should preferentially not reverse direction immediately. However, due to collisions of monomers with solvent molecules, it is possible for the monomer to change its direction of motion. We can incorporate this into our model by biasing or weighting the move sampling with the pseudomomentum.

We begin a single Monte Carlo move of the chain by randomly choosing a monomer. This monomer can move to a nearby lattice site, where “nearby” is defined as the first, second, and third nearest neighbors. The exact locations to which a monomer may move are shown for the two dimensional case in chapter 3. In order to choose which site the monomer moves to, we define a move vector, $\vec{r}_n$, for each of the possible directions of movement of the selected monomer. Considering the importance of inertial effects, we take the probability of the monomer moving in the direction $\vec{r}_n$ to depend on the projection of the monomer’s pseudomomentum vector onto the unit vector $\vec{r}_n / |\vec{r}_n|$. Specifically, we take a sampling weight for moving in the direction $\vec{r}_n$ as given by

$$W_n = \frac{\vec{p} \cdot \vec{r}_n}{|\vec{r}_n|^2} \quad (2.2)$$

Then the probability of moving in direction $\vec{r}_n$ is obtained from Equation 2.3, where D is the number of possible directions of motion[20].

$$P_n = \frac{A + W_n}{\sum_{m=1}^D (A + W_m)} \quad (2.3)$$

We include the adjustable parameter, A , to set the amount of bias for continuing in the same direction: a larger value of A produces a smaller bias. We wish to include only a slightly enhanced probability of continuing motion in the direction of the pseudomomentum vector since there are many solvent molecules that could collide with the monomer, randomizing its pseudomomentum. An appropriate value of A was taken to be 5.

Once the initial monomer has moved to a new lattice site, other monomers in the chain move in response as necessary to accommodate the initial move. A move of the chain is not complete until all of the covalent linkages of the chain are intact. This process can be understood more clearly with the aid of a few examples from the two-dimensional case as shown in Figure 2.3.

In the first example, shown in the top panel of Figure 2.3, monomer two was chosen to move “east.” Once this move has taken place, the covalent linkages of the chain are intact, so the move is complete.

In the second example, monomer seven was chosen to move “northwest” as shown in the middle panel of Figure 2.3. Once this move has occurred, the covalent linkages of the chain are no longer intact. Monomer six must move to accommodate the motion of monomer seven, and then the move is complete.

The final example is shown in the bottom panel of Figure 2.3. In this case, monomer six is chosen to move “north.” To restore the chain linkage, six other monomers must move, resulting in a seven particle move. Because of kinetic energy effects, we would expect the probability of this type of move to be quite small compared to the previous two examples.

Once a move is complete, the pseudomomentum vectors are recalculated, and the

Figure 2.3 Examples of chain motion showing a one monomer move, a two monomer move and a seven monomer move.

move is accepted or rejected based on the Metropolis criteria. However, we now use the total energy in the criteria instead of potential energy. In other words, the move is accepted immediately if the total energy is decreased by the move. If instead the total energy increases by an amount ΔE , then the move is accepted randomly with a probability given by

$$P_{i \rightarrow j} = e^{-\Delta E_{i \rightarrow j}/kT} \quad (2.4)$$

The total energy consists of two parts, potential energy and kinetic energy. The kinetic energy K is calculated from the pseudomomentum vectors according to[21]

$$\frac{K}{kT} = \sum_{i=1}^N p_i^2 \quad (2.5)$$

2.2 Chain Energy

The potential energy of a selected conformation of the chain is determined by the numbers of different types of “contacts.” A contact occurs when two non-linked monomers occupy adjacent lattice sites. In addition, each monomer adjacent to a lattice site which is not occupied by another monomer is assumed to be in contact with solvent molecules at that site. Thus, including three types of possible occupants of any lattice site (hydrophobic monomer H, hydrophilic monomer P, or solvent S), there are six possible pair-wise interactions. Assuming that the interaction potential is fixed for each type of contact, then the total potential energy, V_{tot} , of a given conformation is given by

$$V_{tot} = N_{HH}E_{HH} + N_{HP}E_{HP} + N_{PP}E_{PP} + N_{HS}E_{HS} + N_{PS}E_{PS} + N_{SS}E_{SS} \quad (2.6)$$

where N_{ij} is the number of $i - j$ interactions and E_{ij} is the interaction energy of the $i - j$ pair.

This equation can be simplified so that the solvent interactions are included implicitly. To do this, it is necessary to examine the properties of the lattice and the

chain. For a given lattice, the total number of interacting pairs, N_{tot} , is independent of the chain. So the following equation must be true, where L is the length of the chain, so $(L - 1)$ is the number of covalent bonds in the chain.

$$N_{tot} = N_{HH} + N_{HP} + N_{PP} + N_{HS} + N_{PS} + N_{SS} + (L - 1) \quad (2.7)$$

For a given chain, the number of hydrophobic monomers, n_H , and the number of hydrophilic monomers, n_P , are fixed. The number of interactions that each monomer can participate in is fixed by the number of nearest neighbors that it has, the coordination number, c , of the lattice. In the case of a two-dimensional triangular lattice, $c = 6$. Each monomer in the middle of the chain is bonded to two monomers, and can participate in 4, or $c - 2$, interactions. Monomers at the end of the chain can participate in one additional interaction since they are only bonded to one monomer. These relationships are shown in Equation 2.9.

$$\begin{aligned} 2N_{HH} + N_{HP} + N_{HS} &= (c - 2)n_H + n_H^{end} \\ 2N_{PP} + N_{HP} + N_{PS} &= (c - 2)n_P + n_P^{end} \end{aligned} \quad (2.8)$$

The number of H monomers at the end of the chain is given by n_H^{end} and the number of P monomers at the end of the chain is given by n_P^{end} .

Applying these conditions to Equation 2.6, the stabilization energy can be rewritten as

$$\begin{aligned} V_{tot} &= N_{HH}(E_{SS} + E_{HH} - 2E_{HS}) \\ &+ N_{HP}(E_{SS} + E_{HP} - E_{HS} - E_{PS}) \\ &+ N_{PP}(E_{SS} + E_{PP} - 2E_{PS}) \\ &+ E_{HS}[(c - 2)n_H + n_H^{end}] + E_{PS}[(c - 2)n_P + n_P^{end}] \\ &- E_{SS}[(c - 2)n_H + n_H^{end} + (c - 2)n_P + n_P^{end}] \\ &+ N_{tot}E_{SS} - (L - 1)E_{SS} \end{aligned} \quad (2.9)$$

Note that the last five terms depend only on the coordination number of the lattice, c , the number of monomers of each type, n_H and n_P , the number of monomers at the end of the chain of each type, n_H^{end} and n_P^{end} , the total number of interactions, N_{tot} , and the length of the chain, L . All of these values are constant for a given chain and lattice, so these terms can be omitted since we are concerned with the relative stabilization energies of different conformations of a given chain. Therefore, Equation 2.9 can be written as

$$V = N_{HH}E_{HH}^{eff} + N_{HP}E_{HP}^{eff} + N_{PP}E_{PP}^{eff} \quad (2.10)$$

where E_{ij}^{eff} is an “effective” interaction energy arising from an $i - j$ contact:

$$\begin{aligned} E_{HH}^{eff} &= E_{SS} + E_{HH} - 2E_{HS} \\ E_{HP}^{eff} &= E_{SS} + E_{HP} - E_{HS} - E_{PS} \\ E_{PP}^{eff} &= E_{SS} + E_{PP} - 2E_{PS} \end{aligned} \quad (2.11)$$

Note that the effective energies include solvent interactions as well as monomer interactions. In particular, for an $H - H$ contact, a significant part of the effective interaction arises from an increased number of $S - S$ contacts permitted.

Thus, the values of the effective energies, E_{HH}^{eff} , E_{HP}^{eff} , and E_{PP}^{eff} , are determined by the six different interaction energies, E_{HH} , E_{HP} , E_{PP} , E_{HS} , E_{PS} , and E_{SS} [14, 22]. Since the monomers represent hydrocarbons, the dominant interaction between them is the van der Waals interaction. Hydrophilic monomers have the ability to hydrogen bond, so there is a small additional stabilization energy due to hydrogen bonding. Therefore, the interaction energies between the monomers can be described as follows.

$$|E_{PP}| > |E_{HP}| \approx |E_{HH}| \quad (2.12)$$

When considering the solvent, it is important to note that water molecules are much smaller than amino acids, so each unoccupied lattice site represents a group of

water molecules. The dominant interaction for solvent molecules is hydrogen bonding. There are many opportunities for hydrogen bonding within a group of solvent molecules, so E_{SS} is quite large. Hydrophilic monomers may make a single hydrogen bond with the solvent, while hydrophobic monomers are unable to hydrogen bond at all. Therefore, the solvent-monomer interaction is dominated by van der Waals forces as in the case of the monomer-monomer interactions. Thus, the relative values of the interaction energies are related in the following manner.

$$|E_{SS}| \gg |E_{PS}| > |E_{PP}| > |E_{HS}| \approx |E_{HP}| \approx |E_{HH}| \quad (2.13)$$

We define the interaction energies as

$$\begin{aligned} E_{SS} &= -130 \\ E_{PS} &= -25 \\ E_{PP} &= -21 \\ E_{HS} &= -5 \\ E_{HP} &= -5 \\ E_{HH} &= -5 \end{aligned} \quad (2.14)$$

Substituting the values for the interaction energies in the energy equations, Equation 2.11, results in the following values for the effective energies.

$$\begin{aligned} E_{HH}^{eff} &= -125 \\ E_{HP}^{eff} &= -105 \\ E_{PP}^{eff} &= -101 \end{aligned} \quad (2.15)$$

All three effective energies are large due to the dominant effect of E_{SS} . This is consistent with the general view that solvent interactions are a driving force for protein folding. The variation amongst the values results in a reduction of the degeneracy of the lowest energy state as was seen in work done by O'Toole et. al[14].

Chapter 3

Two-Dimensional Model

3.1 The Model

The protein is modeled as a chain made up of fourteen monomers, where each monomer represents a single amino acid, and is taken to be either hydrophobic (H) or hydrophilic (P). Each monomer is allowed to occupy one site on a triangular lattice, with covalently bonded monomers constrained to occupy adjacent lattice sites. An example sequence, S1, is shown in its native state in Figure 3.1. The small dots

Figure 3.1 The native state of sequence S1.

are unoccupied lattice sites, and the circles represent monomers. H monomers are shown as shaded circles and P monomers are shown as open circles. Covalent bonds

are indicated by gray lines. The native state is the state of lowest energy, and was determined by enumeration of all of the configurations.

When a monomer is chosen to move, it is allowed to move to one of eighteen positions in twelve directions around its initial position as shown in Figure 3.2. The

Figure 3.2 The directions of allowed motion.

shaded circles represent the eighteen possible positions to which the center monomer may move. The arrows show the twelve possible directions of motion. The directions are numbered clockwise starting from $\vec{r}_1$.

3.2 Results and Discussion

The modifications described in Chapter 2 provide some significant insights into protein folding. For example, the greater flexibility of motion of the chain allows

us to do move set sampling, that is, to determine what types of concerted motions are most important for chain folding. Moreover, we can do move set sampling as a function of time. Additionally, our definition of the energy function produces a more physically meaningful entropy function compared to models which only include HH contacts in the energy.

3.2.1 Chain Motion

A primary purpose in developing a new Monte Carlo sampling procedure was to determine how a molecular chain will move when it is allowed to choose its own moves, rather than to sample a preselected move set. For the sequence S1, we sampled the dynamics of folding from an “unfolded” conformer to the native state (see Figure 3.1) with 242 runs, and then counted the number of successful moves corresponding to one monomer moves, two monomer moves, etc. The percentage of moves accepted is given as a function of the number of monomers moved in Figure 3.3 for sequence S1. We note that moves involving several monomers (more than three) are exceedingly rare: over half of the moves that occur are two monomer moves. Although virtually all moves involve three monomers or fewer, moves involving up to eight monomers are found to occur.

Taking a closer look, we find that two types of one monomer moves are possible. These are shown in Figure 3.4. The “end move” accounts for seventy-five percent of the one monomer moves, while the other twenty-five percent are “corner moves”.

Two monomer moves are the predominant moves in the simulations. By observing the chain motion resulting from several simulations, we defined the thirteen types of two monomer moves shown in Figure 3.5. A is named the “end” move. B is a “wrinkle end” move. C is an “unwrinkle end” move. D is a “corner” move. E is a “small crankshaft” move. F is a “large crankshaft” move. G is a “zig crankshaft” move. H is a “crank to corner” move. I is a “shift” move. J is a “line shift” move. K

Figure 3.3 The percentage of N monomer moves accepted for sequence S1 in 2D.

Figure 3.4 The two types of one monomer moves.

Figure 3.5 The possible two monomer moves in 2D.

is a “zig shift” move. L is a “slink” move, and M is an “unslink” move. The forward and reverse moves are labelled separately if the number of contacts changes as a result of the move.

Note the great variety of two monomer moves which are possible on the triangular lattice. This contrasts with the square lattice where movement is much more constrained.

It is important to remember that these moves were not predefined for selection:

rather they occurred randomly, and we developed a method for systematically recognizing when these moves occur. The method is a brute force algorithm in which the relative positions of the monomers are determined both before and after the move occurs and are compared to the positions shown in Figure 3.5. When the type of move that matches the one that occurred is found, the move is identified, and the number of each type of move that occurs can be counted.

Since we were able to count the number of times that each type of move occurred during the simulations, we can determine which type of two monomer motion is predominant in the simulations. The relative importance of each type of two monomer move is shown in Figure 3.6.

Over half of the two monomer moves that occur are “shift” moves (I, J, or K in Figure 3.5). These moves are concerted motions of two monomers. We expect that further study into how such cooperative motion occurs in a chain molecule will provide greater insight into the mechanism of protein folding. It is interesting to note that the “crankshaft” moves (E, F, and G in Figure 3.5), the only type of two monomer move included in the move set for the square lattice model[8], do not occur very often. This we believe is significant supporting evidence for the need for an algorithm such as ours which permits cooperative motions to be chosen independently of a predefined move set.

Another advantage of our algorithm is the opportunity to look at the types of moves that occur as a function of time. This allows us to identify the types of motion that are important during different stages of the protein folding process. To illustrate, we studied the motion of sequence S1 as a function of time by counting the number of times that each type of move occurred over a predefined time segment for over two hundred simulations. We then calculated the percentage of time that each move was selected. The results are shown in Figure 3.7.

We find that the first move that occurs is always an end move, which is expected

Figure 3.6 Percentage of each type of two monomer move accepted for sequence S1 in 2D.

since we choose a linear chain as our initial configuration. As the simulation progresses, the end moves become less important. After the first five hundred moves, a steady state is reached in which the distribution of moves selected remains constant over time. This continues until the final five hundred moves. The steady state distribution was presented in Figures 3.3 and 3.6. As the chain nears the native state, the “shift”, “zig shift”, and “slink” moves (I, K, and L in Figure 3.5) increase in importance, and, remarkably, over half of the simulations end with a “zig shift” move.

Figure 3.7 The percentage of moves accepted as a function of time for sequence S1 in 2D.

3.2.2 Folding Time

With the new model, we see folding times similar to previous models[6, 23]. We use the first time the simulation reaches the native state, the first passage time, as the folding time. A plot of the mean folding time versus temperature for sequence S1 is shown in Figure 3.8. The maximum number of steps allowed in the simulation

Figure 3.8 Folding time as a function of temperature for sequence S1 in 2D.

was 10^7 , so points at the top of the graph correspond to chains that have not folded. At low temperature, it appears that the chain gets trapped in local minima and is unable to find the native state. At higher temperatures, it takes longer for the chain

to reach the native state.

3.2.3 Entropy

The entropy of our lattice model was calculated as a function of conformational energy by complete enumeration of all possible conformations. The method we have developed to calculate the entropy is as follows. Since our chain is 14 monomers long, each chain configuration can be labeled with a 12 digit number which completely identifies the configuration of the chain. For example, the configuration shown in Figure 3.9 corresponds to the number 410303314013. Each digit of this number

Figure 3.9 The configuration enumeration scheme used to calculate the entropy.

represents the bond angle of three monomers in the chain. The angle formed by monomers 1, 2 and 3 corresponds to the angle labeled number 4 on the chart as shown in Figure 3.9. The second digit is 1 which corresponds to the angle formed by monomers 2, 3 and 4. This process continues on down the chain. This process allows us to save computing time as we enumerate all of the conformations because we can calculate only the unique conformations without calculating all of the rotational isomers and mirror images. This is done by enumerating only those configurations which have a first digit of a 0, a 1 or a 2, as well as requiring that as long as the

chain is linear, further monomers may only be placed at angles of 0, 1 or 2. As each configuration is enumerated, the energy is calculated and the degeneracy, $\Omega(E)$, of that energy level is increased by one. Once all of the configurations have been enumerated, the entropy is calculated using Equation 3.1

$$S(E) = \ln \Omega(E) \quad (3.1)$$

We used this procedure to calculate the degeneracy for sequence S1, which is shown as a function of energy in Figure 3.10. The entropy is shown in Figure 3.11. Notice that the degeneracy reaches a maximum at an energy of about -450 corresponding to partially folded conformers. This result contrasts to that of previous models in which the energy is determined by the number of HH contacts.[7, 8, 6] In these models, the degeneracy increases monotonically with the degree of unfolding, so that the perfectly unfolded chain has the greatest degeneracy.

It is our expectation that the entropy function shown in Figure 3.11 is a more reasonable entropy function. It should be true for a realistic chain that there are very few conformations for which a chain will have absolutely no interaction between its monomers, and therefore zero energy. Therefore, the maximum degeneracy and entropy should correspond to a partially folded chain, and our energy function provides this result. Significantly, this reveals that the initial collapse of the chain to a partially folded state is entropically driven.

The “clustering” of conformations as a function of energy in Figure 3.10 is a result of the still comparatively simple model. Essentially, with only 14 monomers and effectively only three types of interaction energy, the states group themselves into families relatively simply according to the number of contacts. We were interested in how this would change with longer chain length and greater variety of pairwise interaction energies, as would be available in the limit of a real protein. When S1 was extended to sixteen monomers, we obtained an entropy function very similar to that of the fourteen monomer chain with the maximum entropy occurring at an energy of

Figure 3.10 The degeneracy of states for sequence S1 in 2D.

Figure 3.11 The entropy of sequence S1 in 2D.

-550.

To study the effect of a larger number of interaction energies, we added two additional monomer types, G and Q, where G is hydrophobic and Q is hydrophilic. The new interaction energies were defined as follows.

$$\begin{aligned}
 E_{HH} &= -125 \\
 E_{HG} &= -123 \\
 E_{GG} &= -120 \\
 \\
 E_{HP} &= -114 \\
 E_{HQ} &= -111 \\
 E_{GP} &= -108 \\
 E_{GQ} &= -105 \\
 \\
 E_{QQ} &= -103 \\
 E_{PQ} &= -102 \\
 E_{PP} &= -101
 \end{aligned} \tag{3.2}$$

Sequence S1 was modified to include all four types of monomers, resulting in the new sequence S1a, shown in Figure 3.12.

The degeneracy for sequence S1a is shown in Figure 3.13. We see that the maximum has moved to an energy of about -100. The entropy is shown in Figure 3.14. Although the curves have become much smoother, we still see families of configurations that occur periodically with the energy.

Since the initial collapse of a chain appears to be entropically driven, it would be useful to extend this study even further. However, a fourteen monomer chain is too short to incorporate more than four monomer types. Increasing the length of the

Figure 3.12 The native state of sequence S1a in 2D.

chain significantly increases the amount of computing time needed to calculate the entropy, so a longer chain with more monomer types is currently out of our reach.

Overall, our work supports the collapse model of protein folding as described in Chapter 1 and shown in Figure 1.2. We see a fast initial collapse, about twenty Monte Carlo steps, that is entropically driven, and then a slow rearrangement process that eventually leads to the native state. We find that the predominant motions in the simulation are two monomer moves and that the types of moves that are important changes during the simulation.

Figure 3.13 The degeneracy of states for sequence S1a in 2D.

Figure 3.14 The entropy of sequence S1a in 2D.

Chapter 4

Three-Dimensional Model

4.1 The Model

We were interested in extending our model to three-dimensions in order to more easily compare chain motion with molecular dynamics simulations.

As in the two-dimensional case, the chain is made up of fourteen monomers, where each monomer is either hydrophobic(H) or hydrophilic(P). The chain is allowed to move on a face centered cubic lattice based on the algorithm discussed in Chapter 2. In the three dimensional case, there are forty-two directions in which a monomer may move, and fifty-four positions available. These include the first, second, and third nearest neighbors. An example sequence, S7, is shown in one of the two lowest energy states in Figure 4.1. The H monomers are shown as large spheres and the P

Figure 4.1 The native state of sequence S7 in 3D.

monomers are shown as small spheres. The native state energies were determined by exhaustive enumeration of all of the completely compact configurations. Sequence S7 has the lowest degeneracy of any of the sequences we have tried thus far, with

two distinct lowest energy states. Sequence S1, which we used predominately in the two-dimensional model, has 35 lowest energy states in three dimensions.

4.2 Results and Discussion

Three-dimensional models are much more computationally intensive which results in some difficulties. First, it is no longer feasible to enumerate all possible chain configurations. This also means that we are unable to calculate the partition function and therefore the entropy function. However, we are still able to calculate the energy of the native state by enumerating all of the completely compact states.

We also find that our chains do not fold up in a reasonable time, especially sequences such as S7 where the degeneracy of the native state is quite small. Possibly, a fourteen monomer chain is too short to collapse completely in three-dimensions, because the energetic driving force is too small. However, moving to a longer chain would greatly increase the computing time. Despite these difficulties, we can still gain some insight about the motion of the three dimensional chain.

As in the two-dimensional case, we sampled the dynamics of folding starting from an “unfolded” linear conformer. Instead of sampling motion until the chain reaches the native state, we let each simulation run until 50,000 or 100,000 moves have been accepted. Then, we count the number of successful moves corresponding to one monomer moves, two monomer moves, etc. for several simulations. The percentage of moves accepted is shown as a function of the number of monomers moved in Figure 4.2 for sequence S7. As in the two-dimensional case, moves involving several monomers are exceedingly rare: over half of the moves that occur are two monomer moves. Although moves of up to eight monomers do occur. If we compare this result to the result in two-dimensions, we note that the percentage of one monomer moves has decreased, while the percentage of two and three monomer moves has increased. As you can see, two monomer moves are even more predominant in this case than in

Figure 4.2 The percentage of N monomer moves accepted for sequence S7 in 3D.

the two dimensional case.

We are still working on being able to sort all of the possible two monomer moves. In three dimensions the number of possible two monomer moves has increased, and it is harder to recognize patterns of motion while watching the chain move. A knowledge of what the moves are is necessary in order to sort the moves.

To avoid these problems, we are developing a sorting scheme that will take advantage of the moves that we have already identified from the two-dimensional simulations. We can divide the moves into four categories: two-dimensional moves, rotations, rotations plus two-dimensional moves, and three-dimensional moves.

As in the two-dimensional case, each two monomer move is described by the positions of the two monomers that move as well as the position of the monomers that they are bound to. Two monomer end moves only involve three monomers, and can be identified quickly. All other two monomer moves can be identified by looking

at the relative positions of these four monomers before and after the move.

For example, if the four monomers are all in a plane before the move occurs, and they are all in the same plane after the move occurs, a two-dimensional two monomer move has occurred. In this case, the two-dimensional move sorting routine can be used to identify the move.

The crankshaft moves in the two-dimensional model are best described as rotational moves. By moving to three-dimensions, rotations of angles other than 180 degrees are allowed. As well as the crankshaft type moves, other types of rotations may be present. A method needs to be developed to separate out rotational moves.

Additionally, there may be moves that are a combination of a rotation and a two-dimensional move. When these moves occur, all four monomers involved in the move will initially be in a plane. After the move occurs, the monomers will still be planar, but they will be in a different plane. If we rotate the new plane onto the original plane, we can identify the type of move that occurred with the two-dimensional move sorting routine.

Any moves that do not fit into these categories will be new types of three dimensional moves. Such moves will have to be identified as in the two-dimensional case, by watching the moves occur and trying to identify patterns. Hopefully the number of moves in this category will be small.

More work needs to be done in order to identify all of the three-dimensional two monomer moves. Once this has been achieved, we will be able to sort the moves as in the two-dimensional case, to determine the types of moves that occur most often as well as how the types of moves that occur change as a function of time. Future work will also be done on the entropy calculation, since such a calculation may be possible through the use of parallel computing. Future work in our group will address these issues.

Chapter 5

Conclusion

We have presented a new simple lattice model for the kinetics and thermodynamics of protein folding, which we believe more accurately represents physically realistic proteins than have previous models. This model improves on the kinetic aspects of protein folding by allowing the chain greater flexibility of motion. First, a triangular lattice is used to increase the number of possible motions. Second, the chain is allowed to define and choose its own moves based on kinetic energy considerations, as opposed to using a predefined set of moves selected solely on the basis of potential energy considerations. A move set free model brings one closer to investigating the kinds of motion that would freely occur in a protein molecule.

The thermodynamics of the model have been improved by using a more detailed description of the monomer-monomer pairwise interactions. Three interaction energies, HH, HP, and PP, are incorporated, as opposed to simply the HH contact energy arising from the “hydrophobic effect.” In the two-dimensional model, this results in a more physically realistic density of states, and therefore a more physical entropy function which, as a function of energy, displays a maximum instead of a monotonic increase.

This work supports the Collapse Model shown in Figure 1.2, since we see a fast initial collapse that is entropically driven, followed by a slow process that eventually leads to the native state.

Our goal is to use Monte Carlo simulations to guide molecular dynamics simulations of small subunits of proteins to investigate the motion involved in protein folding in greater detail. We have extended the model to three-dimensions. Although the three-dimensional problem presents some difficulties, preliminary results have been obtained. Further work will be done to obtain more information about the kinetics

and thermodynamics of the three-dimensional model to further guide our molecular dynamics studies.

Bibliography

1. General information about proteins can be found in a variety of biology and biochemistry textbooks. See for example, (a) Donald Voet and Judith G. Voet, *Biochemistry*, John Wiley & Sons, New York, 1990 and (b) William K. Purves, Gordon H. Orians and H. Craig Heller, *Life, The Science of Biology*, Sinauer Associates, Massachusetts, 1995.
2. Peter L. Privalov, "Physical Basis of the Stability of the Folded Conformations of Proteins", in *Protein Folding*, T. E. Creighton, Ed., p. 123, W. H. Freeman, New York, 1992.
3. M. Karplus and E. Shakhnovich, "Protein folding: theoretical studies of thermodynamics and dynamics", in *Protein Folding*, T. E. Creighton, Ed., pp. 127–196, W. H. Freeman, New York, 1992.
4. Stu Borman, *C & E News* **73**, 20 (1995).
5. G. Taubes, *Science* **271**, 1493 (1996).
6. K. A. Dill, S. Bromberg, K. Yue, K. M. Fieberg, D. P. Yee P. D. Thomas, and H.S. Chan, *Protein Sci.* **4**, 561 (1995).
7. C. J. Camacho and D. Thirumalai, *Proc. Natl. Acad. Sci. USA* **90**, 6369 (1993).
8. R. Miller, C. A. Danko, M. J. Fasolka, A. C. Balazs, H. S. Chan, and K. A. Dill, *J. Chem. Phys.* **96**, 768 (1992).
9. J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, *Proteins Struct. Funct. Genet.* **21**, 167 (1995).
10. A. Šali, E. Shakhnovich, and M. Karplus, *J. Mol. Biol.* **235**, 1614 (1994).

11. A. Kolinski and J. Skolnick, *Proteins Struct. Funct. Genet.* **18**, 338 (1994).
12. M. H. Hao and H. A. Scheraga, *J. Phys. Chem.* **98**, 4940 (1994).
13. Thomas E. Creighton, Ed., *Protein Folding*, Dover, New York, 1992.
14. E. M. O'Toole and A. Z. Panagiotopoulos, *J. Chem. Phys.* **97**, 8644 (1992).
15. D. G. Covell, *J. Mol. Biol.* **235**, 1032 (1994).
16. E.I. Shakhnovich, *Phys. Rev. Lett.* **72**, 3907 (1994).
17. David Chandler, *Introduction to Modern Statistical Mechanics*, Oxford, New York, 1987.
18. S. Bromberg and K. A. Dill, *Protein Sci.* **3**, 997 (1994).
19. K. A. Dill, "Simple Models of Proteins", presentation at the Protein Structure and Folding Symposium, Oct 10, 1995.
20. In the unlikely event that the calculated probability is less than zero, the value is assigned to be zero.
21. Note that the kinetic energy is scaled by kT , so both sides of the equation are unitless.
22. K. Chen, "The Dynamics of a Model Chain by Constrained Molecular Dynamics", Master's thesis, Rice University, 1995.
23. N. D. Socci and J. N. Onuchic, *J. Chem. Phys* **101**, 1519 (1994).