INTERNATIONAL ATOMIC ENERGY AGENCY
UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY, CABLE: CENTRATOM TRIESTE

H4.SMR/642 - 10

College on Methods and Experimental Techniques in Biophysics

28 September - 23 October 1992

Appendix

J.N. ONUCHIC
University of California, U.S.A.

These are preliminary lecture notes, intended only for distribution to participants.

I. Introduction

The prediction of the three-dimensional structure of a protein from its amino acid sequence and ambient thermodynamic conditions is an unsolved problem in several fields of science. A robust predictive method would aid medical science 1) by providing deeper insights into the structure/function relationship in biomolecular activity and 2) by establishing rules for the design of target-specific drugs. In physics, the protein folding problem presents a paradox for modern theories in statistical mechanics. Unlike other heterogeneous "disordered" systems such as glasses, proteins are found in a unique structure. The thermodynamic stability of these "native" conformations has been studied from many experimental and theoretical perspectives, but only a few efforts to understand the long-time kinetics of folding have been undertaken.

The protein folding problem originated from a set of experiments by Anfinsen, *et al*, that demonstrated the reversible denaturation of ribonuclease in vitro. Anfinsen's protocol involved chemically trapping ribonuclease with improper disulfide bonds, then oxidizing the improper bonds to allow folding to occur. The experiments set the foundation for the thermodynamic hypothesis of protein folding: the native conformation is the one in which the Gibbs free energy of the whole system is lowest (Anfinsen, 1973).

A paradoxical argument arose later in which Levinthal presented folding as a deviously simple counting problem (Levinthal, 1968). Assuming that each of the N residues in a protein has two local conformations, a protein has 2^N possible states. Assuming that a protein

could search 10^{12} such states per second (a generous figure) then a 100-mer would require 10^{18} seconds to fold if each conformation is searched with equal probability. Since protein folding takes place on a 1 second time scale, it must be assumed that no more than 10^{-18} of all states are visited by the collapsing polymer. Based on this, some have suggested that proteins must follow a well-defined (if unknown) sequence of events during folding, in apparent contradiction to the thermodynamic hypothesis (Kim & Baldwin, 1983).

The conflicting results of Anfinsen and Levinthal have catalyzed interest in the protein folding as a problem in the basic physics of diffusive molecular self organization. The simplicity of protein folding "theorem" (*if sequence then structure*) entices one to believe that a "proof" should follow from a purely theoretical approach once all relevant quantities have been define. Folding information *must* be contained in the sequence and in environmental parameters. It need not be assumed that most amino acid sequences meet the necessary thermodynamic and kinetic criteria; for most sequences, then, there is no paradox because there is no observed folded structure, and Levinthal's argument is justified. However, biological proteins beat the paradox. A resolution of this paradox must address the kinetics of folding.

It should be mentioned that the standard "theoretical" approach to protein structure and function is molecular dynamics. Based on heavily parameterized potentials stemming from a wide variety of quantum chemical calculations and spectroscopic and thermodynamic data, molecular dynamics has given insights into the temporal evolution of proteins on a nanosecond time scale and shorter. However, in dignifying every conceivable degree of freedom, the Newtonian N-body approach has

rendered itself unfeasible for studies of large proteins or long times. Specifically, since the folding of small globular proteins occurs in about one second, the nanosecond domain of molecular dynamics is not useful for protein folding theory.

The protein folding problem is sufficiently complex that it has spawned two major classes of simplifying formulations: analytical and lattice treatments. Analytical efforts based on mean field theory and spin glasses and have predicted the character of the denaturation transition (Ikegami, 1977; Dill, 1985; Shakhnovich & Finkelstein, 1989; Shakhnovich & Gutin, 1989), nucleation and the existence of multidomain proteins (Dill, 1985; Bryngelson & Wolynes, 1990), a frozen misfolded state (Bryngelson & Wolynes, 1987), and mean first passage times for protein folding (Bryngelson & Wolynes, 1989). Lattice theories, on the other hand, have preserved specific sequence information which is "averaged over" in statistical theories. These models have investigated sequence-energy-structure relations through exact enumerations of compact structures (Chan & Dill, 1989; Chan & Dill, 1990; Chan & Dill, 1991; Shakhnovich & Gutin, 1990; Covell & Jernigan, 1990; Crippen, 1991) and have characterized folding pathways through lattice kinetics simulations of noncompact chains (Skolnick & Kolinski, 1989; Skolnick & Kolinski, 1990; Sikorski & Skolnick, 1989; Miller, *et al*, 1992). The complementary features of analytical and lattice approaches invite attempts to unify the kinetic results of the former with sequence- and structure-specific aspects of the latter.

Analytical protein theory has concentrated on thermodynamic quantities such as temperature of denaturation and heat capacity changes associated with the denaturation transition. The phase transition models

did not address questions of kinetics, which we know must be an important feature in folding. With the recent introduction of stochastic methods (Bryngelson & Wolynes, 1989), kinetic arguments could be made.

The treatment of proteins as spin glasses was first suggested by Stein in a simple model of side chain antiferromagnetic interactions in the compact state (Stein, 1985). Stein's model invoked the fully-frustrated state of antiferromagnetic systems to describe packing and local rearrangement of side chains. Bryngelson & Wolynes (1987) later used the random energy model with superimposed *ferro*magnetic ordering in proposing the principle of minimal frustration. "Frustration," a spin glass term, gives rise to many equivalent minima, but Bryngelson & Wolynes supposed that the uniqueness condition of folding could be obtained if an artificial term would be applied in the Hamiltonian. The result was a phase plot predicting three thermodynamic states of proteins: folded, unfolded, and frozen misfolded. The frozen misfolded phase corresponded to the frustrated limit in which the superimposed ferromagnetic order vanished. In a later effort (Bryngelson & Wolynes, 1989), they formulated a one-dimensional Fokker-Planck equation related to a master equation with Monte Carlo rates, calculated the mean folding time, and showed that in the "frustrated" limit the mean folding time diverged. The divergence of the mean equilibration time with the emergence of frustration, the hallmark of spin glasses, may lead some to think that proteins are *not* spin glasses. Indeed, this is certainly true, but spin glass theories provide a rigorous framework for quantifying frustration -- minimal or not -- where more traditional polymer statistical mechanics has been silent.

In the same vein, Shakhnovich and Gutin analyzed the random

energy model in the absence of a ferromagnetic term and predicted that ten percent of all proteins with a Gaussian-distributed set of random chain-chain contact energies have unique thermodynamically-stable conformations (Shakhnovich & Gutin, 1990).

The two main limitations of the mean field work are the absence of conformational correlation intrinsic to real polymers and the inability to make sequence specific calculations. Not coincidentally, these are the two strongest points of the lattice methods described below.

Trading the atomic resolution of molecular dynamics for longer time simulations, lattice models have proved useful in addressing protein and polymer collapse and kinetic reconfiguration. Two lattice methods have been developed to address the problem: 1) kinetic simulation and 2) exact enumeration of conformations.

Early lattice simulations by Gō demonstrated that the 129-residue enzyme lysozyme can fold from a random coil if native nonbonded contacts are selectively stabilized and all other contacts are not (Ueda *et al*, 1978). More recently Skolnick & Kolinski have shown that a kinetics simulation using a full 20-letter hydropathic coding with no native tertiary preferences but with modest native secondary preferences can obtain the low resolution structure of the 99-residue protein plastocyanin (Skolnick & Kolinski, 1990). In each case, information about the native structure was written into the energy function, but the fact that folding occurred at all is notable in light of Levinthal's famous argument.

The approaches of Gō, *et al*, and Skolnick, *et al*, used knowledge of the native structure in designing their respective parameter sets. Without this knowledge, but on a much simpler problem, Shakhnovich, *et al*, carried out lattice kinetics simulations for the 27-mer parameterized by

random interactions (Shakhnovich, *et al*, 1991). From an exhaustive search of compact conformation space, the minimum energy conformation was known to be a 3x3x3 cube. Shakhnovich, *et al*, have seen that only ten percent of proteins with thermodynamically-stable, randomly-parameterized chains were observed to fold in a kinetics simulation. Coupled with the earlier result that ten percent of all randomly-parameterized proteins have a thermodynamically-stable conformation, it can be surmised that roughly one percent of all randomly parameterized proteins will have a unique, stable, accessible "native" conformation.

A second lattice method involves exact enumeration of lattice structures. Recently, Chan and Dill (1991) have argued using exact enumeration that secondary structure can arise as a result of the compactness phenomenon in proteins. Covell and Jernigan (1990) have applied enumeration in a restricted space in searching for alternative folds of known small globular proteins.

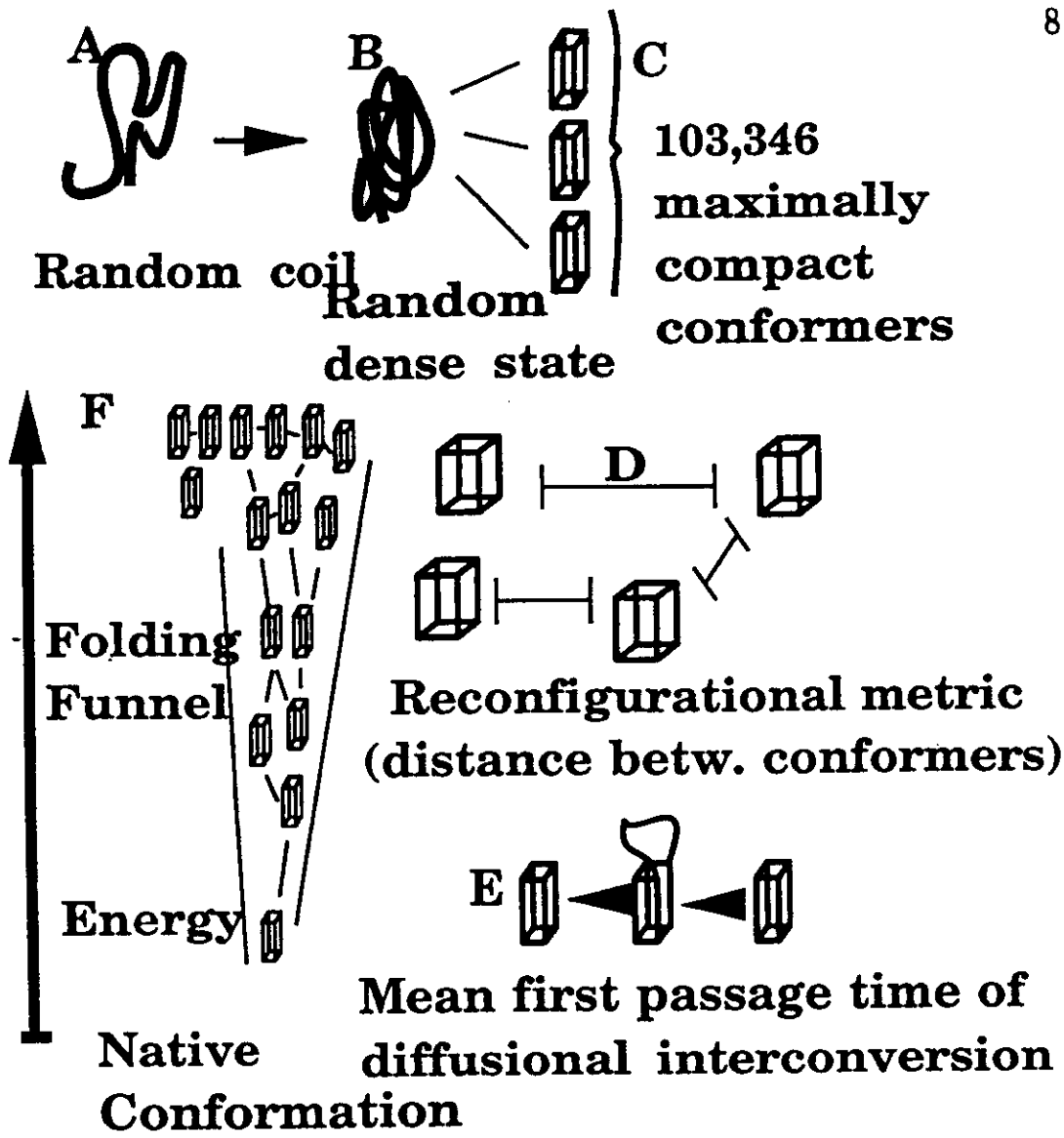
In this thesis, I show that the analytical methods of spin glasses and stochastic processes and the exact enumeration techniques of lattice theories can work hand-in-hand to solve the problem of protein folding kinetics -- in the absence of explicit kinetic simulation trajectories. Lattice kinetics trajectories, like molecular dynamics trajectories, are simulated solutions of temporal equations of motion. For molecular dynamics, the equations of motion are exact. For lattice simulations, the equations of motion are stochastic. Written explicitly, the lattice kinetics equations form a very high-dimensional master equation. This work seeks to formulate the protein folding problem in terms of a master equation for diffusion in a well-characterized, high-dimensional, rough-well potential.

Since the protein is described on a lattice, sequence specificity and excluded-volume are preserved. Since the kinetics are solved in terms of a detailed master equation for a wide variety of conformational interconversions in compactness space, the multiple time scale features of analytical treatments are also maintained.

* * * * *

Since the statement of Levinthal's celebrated paradox, several groups have attempted to salvage the concept of the global minimum by arguing that the requirement of compactness (Dill, 1985) and native state biasing (Zwanzig, *et al*, 1991) reduce the size of conformation space. Neither requirement is fully satisfying, since compactness alone does not solve the paradox and native state biasing cannot model the multiple minimum problem. One must conclude that a resolution of the Levinthal paradox is possible using a kinetic rather than a thermodynamic approach to the folding problem. Folded proteins are not (necessarily) equilibrium entities; rather, they are metastable with lifetimes longer than (or perhaps about equal to) their functionally-significant lifetimes. To overcome the Levinthal problem, proteins must have a broad conformational focussing property to direct folding to a unique, locally-stable, kinetically-accessible conformation.

In this thesis, I introduce the concept of *protein folding funnels*, a kinetic mechanism for understanding the self-organizing principle of the sequence-structure relationship. The concept of a folding funnel follows from a few general considerations (Figure 1). First, proteins fold from a random state by collapsing and reconfiguring; second, reconfiguration occurs diffusively and follows a general drift from higher energy conformations to lower energy conformations; and third, reconfiguration

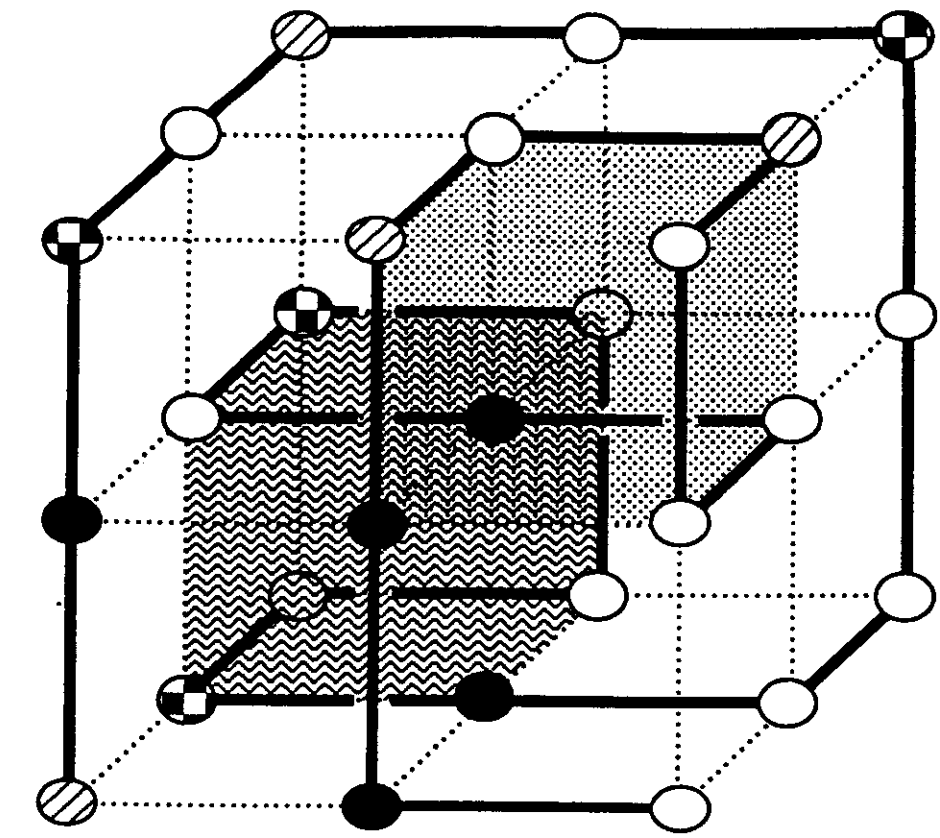


A schematic representation of the folding process. The denatured coil (A) collapses to a random dense structure (B), which is approximated by a set of maximally-compact conformers (C). A reconfiguration distance (D) is defined betweeen compact states and is used to sort pairs of cubes for calculation of the MFPT for interconversion (E). The kinetic structure of conformation space (F) shows a folding funnel leading to a unique, locally-stable, kinetically- accessible state.

Figure 1. Schematic of protein folding.

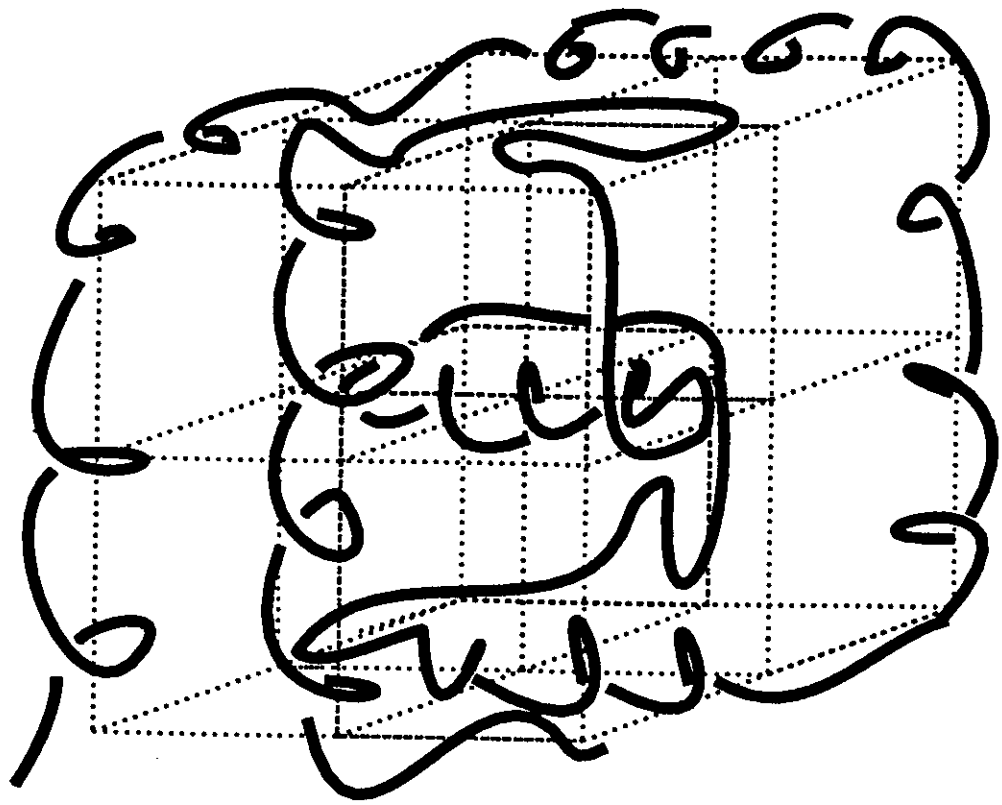
occurs between conformations that are geometrically similar, *i.e.*, global interconversions are energetically prohibitive after collapse, so local interconversions alone should be dominant. I define the folding funnel as a collection of geometrically-similar collapsed structures, one of which is thermodynamically-stable with respect to the rest, though not necessarily with respect to the whole conformation space. Amino acid sequences having only one sizeable folding funnel leading to a unique, stable conformation are said to be *foldable*. Conversely, *non*foldable amino acid sequences may have multiple folding funnels leading to equally stable and accessible conformations.

Three major simplifications are made in order to describe the long-time kinetic behavior of a protein. First, a lattice model is used in order to restrict discussion to the most important degrees of conformational freedom (Figure 2). Second, consistent with the low resolution structural description, the chemical identity of the protein is depicted in a simplified M-letter hydropathic code, where some number M of amino acid types is chosen to enable any degree of heterogeneity for residue-residue interactions. Each residue is located at the vertices of the chain and may be considered to represent clusters of real amino acids (Figure 3). Finally, the dynamics of the folding process can be simplified as a collection of discrete "hopping" processes between local minima. Since conformational dynamics is governed by the same residue-residue interactions that drive the initial conformational collapse, it is possible to divide motion into two types: relatively fast, sterically-constrained chain motions that do not involve the breaking of many chain-chain contacts, and slow activated motion across energy barriers associated with larger structural fluctuations. The time scale separation proposed for the interconversion



The 27-unit heteropolymer is made up of a sequence of amino acid types. Here a four-letter code is used to ascribe chemical identity to each monomer. The conformation exhibits subdomains in the front-upper-right and rear-lower-left octants. Each subdomain can fold separately from the rest of the protein.

Figure 2. A maximally-compact 27-mer "cube".



A depiction of the supposed backbone of a protein whose cubic conformation is given in Figure 2. The cubic model is valuable only as a very low resolution approach to the structure. "Helices" are straight lines of length two; "sheets" are consecutive unit-long segments that reverse the trajectory of the backbone.

Figure 3. A schematic showing how a real protein backbone could be mapped into a cube.

process can be used to divide the set of all collapsed conformations into "compactness cells". Within a compactness cell, conformations interconvert quickly; between compactness cells, conformations interconvert slowly. The rates for slow interconversion can be found by calculating the average time the protein spends in any cell before a thermal fluctuation moves the chain over a conformational barrier into another cell.

The division of conformation space into compactness cells is valid for any model for protein structure, whether on- or off-lattice. A difficulty arises when attempting to define compactness cells for specific sequences. To show how such difficulties can be managed through an approximation technique, I will focus on a simple lattice model. But first we must investigate basic facts of lattice models.

Levinthal's argument shows the simple power of counting methods, which are ubiquitous in polymer theory. In fact, a list of counting techniques provides an excellent organizing tool for understanding the methods and approximations of polymer theory. In the interests of introducing the field and putting this work in context, I will review the various perspectives on enumeration in chapter 2. The important result is that conformational enumeration is possible when constraints are imposed. Also in the second chapter, basic results of lattice polymer representations are reviewed and several useful conformational metrics are defined. Compactness cells are introduced in order to divide compact conformation space into identifiable subsets. In the third chapter, chain energetics and mode of sequence representation are introduced. The concept of compactness cells is extended to provide a low resolution view of the energy landscape of compactness space. In the fourth chapter, the

kinetics of interconversion between compactness cells is defined in terms of a master equation. The mean first passage time for diffusion between compactness cells is calculated and related to long time scale diffusional rates. In the fifth chapter, the results of the kinetic calculations are compared to lattice simulations. In the sixth chapter, several unanswered questions on the protein folding problem are reviewed, including the inverse folding problem, the *M*-letter hydropathic representation, and parametric robustness. In the last chapter, the contributions of this thesis are evaluated.

II. Proteins on Lattices: Geometrical Considerations

In this chapter, counting problems are examined on a more detailed level and the concept of compactness cells and the reconfiguration metric are proposed. Both will be employed later in the search for foldable protein sequences.

A. Counting Problems. There are two main types of counting: sequence and conformation. Sequence counting can be done exactly. Sequences of N-mer proteins represented by an M-letter hydropathic code have exactly M^N manifestations. On the other hand, conformation counting is usually approximated. Motivated by the discrete nature of the dihedral space of real proteins, the conformational counting problem is marked by ν local structural states corresponding to secondary structural features like α-helices, ß-sheets and turns. State or conformation counting draws its intellectual origins from efforts to evaluate partition functions where each monomer has ν possible states ($2 \leq \nu \leq 10$) (Poland & Scheraga, 1970).

State counting can be divided into two categories: noninteracting and interacting. In the first case, there is no monomer-monomer repulsion or attraction, and the counting problem is identical to sequence counting, *viz*, for an N-mer with ν states per monomer, there are ν^N possible conformations. While this model permits nonphysical conformations in which monomers "overlap" in space, it is analytically tractable. In the second case, an explicit monomer-monomer potential affects the number of allowed conformations. The simplest interacting polymer model is the hard-sphere excluded-volume model. Flory showed that there is a special case of the interacting model (the "theta" state) that reproduces the basic

results of the noninteracting model. It should be emphasized that covalent connectivity between monomer i and $i+1$ is assumed in both cases.

Spacial dimensionality of the protein plays an important role in formulating the counting problem. The one-dimensional case may seem pathological because it does not permit the motion of any monomer of the protein. However, it is of interest because it incorporates helix-coil transition theory. When interactions are only permitted between monomers i and $i+1$ and $\nu=2$, the simple Ising model results. For the same case with $\nu=8$, Zimm-Bragg theory is recovered (Zimm & Bragg, 1959). In all Ising-like problems, exact solutions can be obtained.

In two and three dimensions, the general counting problem of the interacting polymer cannot be solved analytically. Lattice models are often used to discretize conformation space, where the dihedral discretization index ν is interpreted as the coordination number of the lattice. Numerical approaches can give exact results for short chains, and several kinds of constraints must be imposed to make enumeration of longer chains feasible. A useful restriction for the protein folding problem is the volume constraint in which the self-avoiding walk must remain within specific volume. For longer chains, functional integration can also be used to evaluate numbers of conformations with a one or two constrained chain-chain (Chan & Dill, 1990a).

For higher-dimensional models, the excluded-volume interaction vanishes. Unfortunately, there is no physical meaning to these models.

In keeping with the spirit of this discussion, it is interesting to note that "zero-dimensional protein" can be associated with a mean-field formulation. The meaning of the "zero dimensions" refers to the averaging

of primary structure information. Contrary to this name, however, mean field theories can be calculated in any dimension, with or without the excluded-volume interaction.

The conformational "trajectory" of a lattice model of a protein is described as a nearest-neighbor walk on a cubic lattice. In the noninteracting case, a simple estimate of the number of conformations of a polymer of length N on a lattice with ν nearest-neighbors is $\Omega^{(0)} = \nu^{N-1}$, since there are $N-1$ subsequent steps after the first monomer has been laid down. For a cube $\nu=6$, so a 10-mer has $6^9 \sim 10^7$ conformations. The superscript $^{(0)}$ indicates that the value of Ω is calculated without considering excluded volume, i.e., two monomers may inhabit the same lattice site. A first correction posits that a more realistic lattice walk may not reverse itself, so each succeeding monomer has ν-1 possible positions and $\Omega^{(1)} = (\nu-1)^{N-1}$. For the 10-mer, $\Omega^{(1)}=5^9=2 \times 10^6$, five times smaller than the non-interacting case. This value still represents an overestimate of the number of conformations because the long-range excluded-volume interaction is ignored; sites i and $i+1$ will not overlap, but sites i and j ($i<j-1$) may coincide. One may imagine that inclusion of higher order corrections may further reduce the effective coordination number. Sykes has shown that the asymptotic expression of Ω for large N is $\Omega^{(sykes)} \sim \mu^{N-1}(N-1)^{\gamma}$, where the base $\mu=\nu-1-\delta$ is the effective coordination number of the lattice. The constant μ is the value to which the base converges after excluding all higher order excluded value interactions and in the limit as N approaches infinity. The dimensionality constant γ is valid for all lattices of a given dimension, independent of coordination number. For the 10-mer example given, $\Omega^{(sykes)} \sim 1.56 \times 10^6$. The exact solution for the 10-mer on the cubic lattice is 1,853,866.

Biologically-important single domain proteins have between 50 and 200 monomers. If these proteins are described using a three dimensional cubic lattice, the Sykes asymptotic formula gives $\Omega^{(sykes)}(50)\sim10^{33}$ and $\Omega^{(sykes)}(200)\sim10^{133}$. Even if lower coordination number lattices are used, the number of conformations is of astronomical proportion: $2^{50}=10^{15}$ and $2^{200}=10^{60}$, motivating Levinthal to pose the famous paradox already described in chapter 1.

However, all folded proteins inhabit a relatively small part of conformation space corresponding to dense, compact lattice walks. The packing density of protein crystals is between 68% - 82%, about the same as the packing density of hard spheres (75%), indicating that compactness is at the limit of the CPK values of atomic/molecular radii. Crystals of small molecules have values between 70% and 80%. Thus it is an experimental fact that folded proteins are dense (Schulz & Schirmer, 1979).

This fact can be used to great advantage in the counting approach to folding. Constraining a self-avoiding walk to lie within a volume greatly reduces the number of conformations. Not surprisingly, constraints also increase the degree to which the long range excluded-volume interaction affects the properties of allowed conformations. Results from compact enumerations give the Sykes-like formulae $\Omega^{(cmpct)} \sim 1.8^{N-1}$. Flory (Flory, 1949) and Sanchez (Sanchez, 1978) have provided estimations of countings based on a mean field approach. Translated into the language of lattices, the excluded-volume-constrained counting gives

$$\Omega(R) = \left(\frac{54}{\pi}\right)^{1/2} (\nu-1)^{M-1} \left(\frac{1}{R}\right)^{3(M-1)} \frac{\Gamma[R^3]}{\Gamma[R^3-(M-1)]} \times \times \frac{R^2}{(M-1)^{3/2}} \exp\left[-\frac{3R^2}{2(M-1)}\right] \quad (1)$$

where R represents the Euclidean distance from the N-terminus to the C-terminus and ν is the coordination number of the lattice. For maximally-compact conformations $R=N^{1/3}$.

These estimates can be checked using explicit numerical enumerations of constrained lattice walks. For a cubic lattice, there are 48 ways of rotating and reflecting a given conformation so as to maintain the same nonbonded contacts but to generate 48 different sets of coordinates for the same geometrical structure. While it is natural for an enumeration to remove this degeneracy, the mean field estimations must be divided by 48 in order to make an accurate comparison.

The success of the mean field counting model is evident for a number of small, exactly-countable geometries. For polymers of length N, the number of conformations is a steep function of the constrained volume. When the constraining volume is exactly equal to the sum of the volumes of each of the monomers, the conformation is said to be maximally compact. In lattice models, each monomer has a volume of one lattice site, so a maximally compact N-mer is enclosed within a (not necessarily unique) volume containing N lattice sites. If N is a perfect cube (8,27,64,....) then the maximally compact walk can fit inside a cube of edge $N^{1/3}$. Otherwise, ambiguities may arise. If N is the product of three similar integers, $N= a \times b \times c$, $a \sim b \sim c$, then maximally compact volume is a parallelopiped of sides a,b,c. If $a>>b\sim c$ or if N is not the product of three non-unitary integers, then the shapes of maximally

compact conformation is ambiguous. We shall not comment on the ambiguities here except to say that the success of the mean field counting method assures us that a reasonable estimate of the number of conformations is possible.

When the constraining volume R^3 is greater than the N lattice sites, the volume is said to possess R^3-N holes, or inclusions. The mean field model predicts an exponential increase in number of configurations with the number of holes. The difficulty in enumerating non-maximally compact conformations poses a serious problem for the analysis of arbitrarily- (but not maximally-) compact polymers. How can we understand the behavior of a random compact polymer if the effort involved in specifying its conformation requires many more degrees of freedom than we believe practical? Without this detailed information, how can two random compact polymers be distinguished? What is the "conformational distance" between them? Before addressing this problem, we must re-examine the importance of compactness in the protein folding question.

B. Models of Folding. As discussed in the last chapter, the basic mechanism of protein folding is still unknown. All models include some form of diffusive driving force, although at different levels of the process. The diffusion-collision school posits that secondary structure is "driven" on a short time scale and tertiary structure arises from the diffusive assembly of secondary structure pieces. On the other hand, the collapse-reconfiguration model suggest that a random hydrophobic collapse to a compact structure precedes a global rearrangement into a unique "native" structure (Dill, 1985). In the framework model, a particularly stable piece of secondary structure acts as a nucleation site and catalyzes the build-up

of secondary and tertiary structure around it (Kim & Baldwin, 1982).

For each of these models, arrival at a compact state occurs at a different stage during folding. In the collapse-reconfiguration model compactness happens first. In the framework model compactness happens during buildup, and in the diffusion-collision model compactness occurs after secondary structure has formed. Consequently, the use of nonbonded interactions as a driving force for folding is different in each case. In the collapse-reconfiguration and (arguably) diffusion-collision models, the number of nonbonded contacts is constant during the folding process. Nonbonded contacts act as the catalysts of folding; bad contacts are exchanged for good, preferentially stabilizing structural elements during fluctuations. In the framework model, the number of nonbonded contacts increases linearly with the size of the nucleation site, stabilizing the parts during the piece-wise transfer of residues from the highly entropic solvent medium to the dense protein medium; nonbonded contacts do not define the energy surface of folding except in the gross sense of preferring the native state (but not other compact states) over the random coil state.

Lattice kinetics simulations on small systems (N=27) indicate that folding is a strong function of non-native contacts and would therefore support the assumptions of the collapse-reconfiguration or renormalized diffusion-collision models. Other simulations of large systems (N~100, Skolnick & Kolinski, 1990) are used to defend the framework model, however it is clear that if the temperature is low enough or the parameter distribution is sufficiently degenerate, nonbonded contacts will favor the collapsed state and folding will proceed on the nonbonded contact energy landscape.

This excursion into protein folding theory has been meant to

sharpen the focus on the problem of compact conformational geometry and its generally overlooked role in defining the features of the folding energy surface. If -- as collapse-reconfiguration models supposes -- collapse happens early in folding, then the whole problem becomes a one of interconversion between compact conformations. If -- as the diffusion-collision model supposes -- secondary structures form first then aggregate, dense phase interconversion occurs on a renormalized level, with helices as primitive units instead of residues. In either case, protein folding dynamics must traverse the energy surface made rough by the heterogeneity of nonbonded interactions and by excluded-volume condition.

Thus it is critical to be able to understand the mathematical properties of compact conformation space, to define distances between its elements, to quantify similarities in structures and to give an overall road map to the geometry of compactness. As usual, all work will be written in the language of lattice models. This effort is a logical precursor to the later work of assigning energies to each of the structures, from which we will then derive kinetics and ultimately, foldability.

References

Anfinsen, C.B. (1973) "Principles that govern the folding of protein chains" *Science*, **181**, 223-230.

Baldwin, R.L. (1990) "Pieces of the folding puzzle" *Nature*, **346**, 409.

Bryngelson, J.D. & Wolynes, P.G. (1987) "Spin glasses and the statistical mechanics of protein folding" *Proc. Natl. Acad. Sci. USA* **84**, 7524-7528.

Bryngelson, J.D. & Wolynes, P.G. (1989) "Intermediates and Barrier Crossing in a Random Energy Model (with applications to protein folding)" *J. Phys. Chem.*, **93**, 6902-6915.

Bryngelson, J.D. & Wolynes, P.G. (1990) "A simple statistical field theory of heteropolymer collapse with application to protein folding" *Biopolymers*, **30**, 177-188.

Chan, H.S. & Dill, K.A. (1990a) "The effects of internal constraints on the configurations of chain molecules" *J. Chem. Phys.*, **92**, 3118-3135.

Chan, H.S. & Dill, K.A. (1990b) "On the Origins of Structure in Globular Proteins" *Proc. Natl. Acad. Sci. USA*, **87**, 6388-6392.

Chan, H.S. & Dill, K.A (1991a) "Sequence Space Soup of Proteins and Copolymers" *J. Chem. Phys.*, **95**, 3775-3787.

Chan, H.S. & Dill, K.A. (1991b) "Polymer Principles in Protein Structure and Stability" *Annu. Rev. Biophys. Biophys. Chem.*, **20**, 447-490.

Covell, D.G. & Jernigan, R.L. (1990) Conformations of Folded Proteins in Restricted Spaces" *Biochemistry*, **29**, 3287-3294.

Creighton, T.E. (1985) "Problem of How and Why Proteins Adopt Folded Conformations" *J. Phys. Chem.*, **89**, 2452-2459.

Crippen, G.M. (1991) "Prediction of Protein Folding from Amino Acid Sequence over Discrete Conformation Spaces" *Biochemistry*, **30**, 4232-4237.

Dill, K.A. (1985) "Theory for the Folding and Stability of Globular Proteins" *Biochemistry*, **24**, 1501-1509.

Dill, K.A., Alonso, D.O.V., & Hutchinson, K. (1989) "Thermal Stabilities of Globular Proteins" *Biochemistry*, **28**, 5439-5449.

Finkelstein, A.V. & Shakhnovich, E.I. (1989) "Theory of Cooperative Transitions in Protein Molecules. II. Phase Diagram for a Protein Molecule in Solution" *Biopolymers*, **28**, 1681-1694.

Flory, P.J. (1949) *J. Chem. Phys.*, **17**, 303-310.

Gō, N. (1983) "Theoretical Studies of Protein Folding" *Annu. Rev. Biophys. Bioeng.*, **12**, 183-210.

Harrison, S.C & Durbin, R. (1985) "Is there a single pathway for the folding of a polypeptide chain?" *Proc. Natl. Acad. Sci. USA*, **82**, 4028-4030.

Ikegami, A. (1977) "Structural Changes and Fluctuations of Proteins I. A statistical thermodynamic model" *Biophysical Chemistry*, **6**, 117-130.

Kanehisa, M.I. & Ikegami, A. (1977) "Structural Changes and Fluctuations of Proteins II. Analysis of the denaturation of globular proteins" *Biophysical Chemistry*, **6**, 131-149.

Kim P.S. & Baldwin, R.L. (1982) "Specific Intermediates in the Folding Reactions of Small Proteins and the Mechanism of Protein Folding" *Annu. Rev. Biochem.* **51**, 459-489 .

Leopold, P.E., Montal, M. & Onuchic J.N. (1992) "Protein folding funnels: Kinetic pathways through compact conformation space" *Proc. Natl. Acad. Sci USA*, submitted.

Levinthal, C. (1968) "Are there pathways for protein folding?" *J. Chim. Phys.*, **65**, 44-45.

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**, 768-780.

Miyazawa, S. & Jernigan, R.L. (1985) "Estimation of Effective Interresidue Contact Energies from Protein Crystal Structures: Quasi-Chemical Approximation" *Macromolecules* **18**, 534-552.

Poland, D. & Scheraga, H.A. (1970) *Theory of Helix Coil Transitions in Biopolymers*, Academic Press, New York.

Rammal, R. Toulouse, G. & Virasoro, M.A. "Ultrametricity for physicists" *Rev. Mod. Phys.*, **58**, 765-788.

Sanchez, I.C. (1979) *Macromolecules*, **12**, 980-988.

Schulz, G.E. & Schirmer, R.H. (1979) *Principles of Protein Structure*, Springer-Verlag, New York.

Shakhnovich, E.I. and Finkelstein, A.V. (1989) "Theory of Cooperative Transitions in Protein Molecules I. Why denaturation of globular protein is a first-order phase transition" *Biopolymers*, **28**, 1667-1680.

Shakhnovich, E.I. & Gutin, A.M. (1989) "Formation of unique structure in polypeptide chains: Theoretical investigation with aid of a replica approach" *Biophysical Chemistry*, **34**, 187-199.

Shakhnovich, E.I. & Gutin, A.M. (1990a) "Enumeration of all compact conformations of copolymers with random sequence of links" *J. Chem. Phys.*, **93**, 5967-5971.

Shakhnovich, E.I. & Gutin, A.M. (1990b) "Implications of thermodynamics of protein folding for evolution of primary sequences" *Nature*, **346**, 773-775.

Shakhnovich, E., Farztdinov, G., Gutin, A.M., & Karplus, M. (1991) "Protein Folding Bottlenecks: A Lattice Monte Carlo Simulation" *Phys. Rev. Lett.*, **67**, 1665-1668.

Sikorski, A & Skolnick, J. (1989) "Monte Carlo simulation of equilibrium globular protein folding: α-Helical bundles with long loops" *Proc. Natl. Acad. Sci. USA* **86**, 2668-2672.

Skolnick, J. & Kolinski, A. (1989) "Computer Simulations of Globular

Protein Folding and Tertiary Structure" *Annu. Rev. Phys. Chem.*, **40**, 207-235.

Skolnick, J. & Kolinski, A. "Simulations of the Folding of a Globular Protein" (1990) *Science*, **250**, 1121-1125.

Sykes, M.F. (1963) *J. Phys. Chem.* **39**, 410.

Tanaka, S. & Scheraga, H.A. (1976) *Macromolecules*, **9**, 945-950.

Ueda, Y., Taketomi, H., Gō, N. (1978) "Studies on Protein Folding, Unfolding, and Fluctuations by Computer Simulation. II. A Three-dimensional Lattice Model of Lysozyme" *Biopolymers*, **17**, 1531-1548.

Weissman, J.S. & Kim, P.S. (1991) "Reexamination of the folding of BPTI: Predominance of Native Intermediates" *Science*, **253**, 1386-1393.

Zimm, B.H. & Bragg, J.K. (1959) "Theory of the Phase Transition between Helix and Random Coil in Polypeptide Chains" *J. Chem. Phys.*, **31**, 526-535.

Zwanzig, R., Szabo, A., and Bagchi, B. (1992) "Levinthal's Paradox" *Proc. Natl. Acad. Sci. USA*, **89**, 20-22.