Simple Models of Protein Folding

Andrew Blanchard

May 11, 2012

Abstract

Lattice models with simplified interaction potentials have been used to analyze the ability of certain amino acid sequences to adopt a unique configuration in space [1, 2, 3]. Furthermore, phenomenological models have been used to predict protein folding kinetics amongst a subset of energetically favorable states [4, 5]. In the following, I will specifically discuss both the use of two dimensional lattice models and simple rate matrices to describe the transition of disordered proteins to a unique native state (or subset of states). Furthermore, I will discuss the use of both molecular dynamics simulations and experimental techniques to observe specific pathways for protein folding and provide direct connections to theory.

Contents

1 Introduction

The central dogma of molecular biology describes the typical cellular use of genetic information in the construction of proteins. Construction proceeds from DNA $\rightarrow$ RNA through a process known as transcription and then from RNA $\rightarrow$ Protein in a process known as translation. Translation is performed by the ribosome, a cellular organelle, which sequentially converts the nucleotide sequence of RNA into a growing protein chain of amino acids. The shape and function of nascent proteins are then ultimately determined by the specific sequence of amino acids [6] . Thus, the diverse phenomena of the cell are largely explained by the translation of genetic information, in the form of a nucleotide sequence, into proteins, in the form of a distinct amino acid sequence.

1.1 Brief Overview of Protein Folding

A common starting point for providing a theoretical explanation of protein folding is the assumption that an amino acid sequence specifies a distinct three dimensional native configuration [5, 7]. Furthermore, we may refine the assumption by adding that a protein will adopt the configuration which minimizes overall free energy of the system and that this configuration coincides with the native state of the protein under physiological conditions [8]. For an extensive study of the thermodynamics of protein folding, please see [9]. Utilizing the previous assumptions, the problem of protein folding is simplified to predicting the final structure given a unique sequence of amino acids.

Under physiological conditions, a protein not only needs to reach a repeatable native state, but it must do so on an appropriate timescale. Thus, cells need a way to regulate folding kinetics of nascent proteins to ensure proper function. Experiments have shown that appropriate enzymes may greatly accelerate the folding process towards the native state [8]. In general, enzymes that aid in proper folding are known as chaperones [6]. Chaperones likely play a large role in both ensuring proper protein placement and regulating folding kinetics within the cell [6].

1.2 Personal Interest

The problem of protein folding lies at the intersection of both biology and physics. On the biological side, proteins are responsible for practically all cellular processes. On the physical science side, the energy landscape of protein confirmations contains many metastable states, and is largely susceptible to solvent and temperature effects. Furthermore, the apparent simplicity of the cellular machinery for construction of proteins (the reading of a four letter code to produce a polymer of amino acids [6]) is amazing considering the diverse functions carried out by proteins. Thus, the protein folding problem lies at the heart of understanding how nature can use simple building blocks to synthesize exceedingly complex systems with robust functionality.

1.3 Model Systems

To characterize both the existence of a unique native state and the kinetics of protein folding, several model systems have been employed. Specifically, lattice models with simplified interaction potentials and mean field theory have been used to determine the ability of sequences of amino acids to adopt a compact structure [1, 10]. Furthermore, rate matrices have been used to model folding kinetics around a native state [4]. Another, slightly more complicated, approach to modeling protein folding involves using molecular dynamics simulations. Current advances in computing technologies have allowed all-atom simulations to probe several microseconds over a folding trajectory [11]. With the three aforementioned techniques in mind, the following methods section will give an explicit example of each approach and then conclude with a short description of experimental techniques to capture protein folding on a short (μs) timescale.

2 Methods

2.1 2-D Lattice Model

One of the simplest models for protein folding consists of a two letter amino acid code (hydrophobic or hydrophilic) placed on a two dimensional square lattice [1]. The main advantage of this model is that for small polymer chains the entire space of different sequences and spatial configurations can be enumerated. To consider the implications of the 2-D lattice model I will follow the results and notation of [1] and consider the case of a ten monomer chain.

For the model system, we fill each lattice position with either solvent, a hydrophilic residue, or a hydrophobic residue and we only consider interactions amongst non-sequential hydrophobic residues. Thus, for a system of ten residues, we have the following partition function:

H1	H2	P3	P4	S
S	S	S	H5	P6
S	S	S	H8	P7
S	S	S	P9	S
S	S	S	P10	S

Example of a ten monomer protein on a two dimensional lattice with one non-sequential hydrophobic interaction between H5,H8

$$Z = \sum_{m=0}^{s} g(m) e^{(s-m)\epsilon} \tag{1}$$

In the above equation, m is the number of non-sequential hydrophobic interactions; g(m) is the degeneracy of the state with m interactions. ϵ is the energy (divided by $k_B T$) of the hydrophobic interactions. Notice that the $e^{s\epsilon}$ term is arbitrary and is used to make the state

with the most possible hydrophobic interactions the zero energy state. Now, we wish to look for states that meet two conditions: 1. They have a minimum configurational energy; 2. They are maximally compact (where compact refers to a state where there are no solvent sites interior to the polymer). We expect the existence of maximally compact, minimum energy states to suggest a native configuration for the protein. The measure of compactness for the system is given by:

$$\rho = \frac{m+u}{t_{max}} \tag{2}$$

Here, m is the number of non-sequential hydrophobic interactions; u is the number of other non-sequential interactions. t_{max} is the number of non-sequential interactions for a maximally compact state.

Now that we have established the partition function and the measure of compactness, we consider the average compactness for the states and the most likely value of compactness as a function of ϵ.

$$< \rho > = \frac{1}{Z} \sum_{m=0}^{s} \frac{m+u}{t_{max}} g(m) e^{(s-m)\epsilon} \tag{3}$$

$$\rho^* = (\frac{m^*+u}{t_{max}}) \tag{4}$$

$$m^* = max_m[g(m)e^{(s-m)\epsilon}] \tag{5}$$

We look for amino acid sequences that adopt a value of compactness near 1. These sequences can be identified as having a well defined folded state (or subset of states). Notice that as $\epsilon \to -\infty$, we expect that ρ^* samples states with minimum configurational energy.

From the data [1], there are three distinct cases which depend on amino acid sequence. The value for $< \rho >$ may never approach 1, which implies no subset of folded configurations. Otherwise, the value of $< \rho >$ may approach 1, but the value for ρ^* may approach 1 through one or two steps, corresponding to a first-order folding transition or transition through a folding intermediate respectively. Now, specifically for the ten monomer chain, there are 1024 possible sequences. Of these sequences, only 259 yield values of $< \rho >$ that tend to 1, suggesting a folded state. Furthermore, the majority of these sequences do not have a unique minimum energy state, implying a subset of possible folded configurations [1].

Results consistent with the two dimensional lattice model have also been obtained using a mean field approach [10]. In the mean field theory, the fraction of hydrophobic residues in the interior of a protein were calculated again using a non-sequential hydrophobic interaction energy. The mean field theory did not, however, impose strict constraints on the sequential ordering of the amino acids, which yielded greater folding probabilities for chains with few hydrophobic residues [1, 10].

2.2 Rate Matrix Kinetics

Another approach to modeling protein folding involves using a Hamiltonian that assigns an energy to all native (folded state) contacts between amino acids and a different energy to all non-native contacts. The term contact refers to the juxtaposition of two non-sequential amino acids in space. Following the notation of [4], the energy for a state α is given by:

$$H_\alpha = \epsilon_N \sum_{ij} C_{ij}^\alpha C_{ij}^N + \epsilon_{NN} \sum_{ij} C_{ij}^\alpha (1 - C_{ij}^N) \quad (N \textit{ is native structure}) \tag{6}$$

$$C_{ij}^\alpha = \begin{cases} 1 & \textit{contact between residues } i \textit{ and } j \textit{ exists in structure } \alpha \\ 0 & \textit{otherwise} \end{cases} \tag{7}$$

Furthermore, transitions amongst states are modeled with the use of a rate matrix k, which governs the equation:

$$\frac{dp_\alpha}{dt} = \sum_\beta k_{\alpha\beta} p_\beta \tag{8}$$

In the above equation, p_α is the probability that the protein is in state α. Using the model Hamiltonian in (6) and Kramer's approximation (see [4]), the coefficients of k can in principle be computed.

To discuss the qualitative implications of a rate matrix model, we will now assume a specific form of k and consider the effects for certain rate coefficients. Following [4], we assume k is composed of n blocks (representing n intermediate states) and one extra row representing the folded state. Let k have the form:

$$k_{\alpha\beta} = \begin{cases} k_1 & \textit{within a intermediate block} \\ k_0 & \textit{beween intermediate blocks} \\ k_{0N} & \textit{from intermediate to native} \\ k_{N0} & \textit{from native to intermediate} \end{cases} \tag{9}$$

The diagonal terms of k are fixed so that the rate equation (8) obeys conservation of probability.

Now, we can discuss the implications for the kinetics of folding due to different coefficients for k. Specifically, consider the case where $k_{0N} >> k_0$. In this case, the protein will likely transition through the native state as it samples other intermediates. Thus, the native state acts as "kinetic hub" (using the language of [4]). In the opposite case, where $k_0 >> k_{0N}$, we have a different scenario. In this case, the protein will likely sample the intermediate states, with a small probability to fold into the native state [4].

2.3 Molecular Dynamics Simulations

Classical molecular dynamics simulations use Newton's equations of motion ($\vec{f} = m\vec{a}$) to approximate the trajectory of a system through phase space. Concerning protein folding, a

common system would be comprised of water molecules, ions, and a protein configuration (usually taken from an x-ray crystal structure). The force on each atom in the system, typically including electrostatics, dispersion, and bond constraints, is computed and the system is moved forward in time by integrating the equations of motion. A correct parameterization of forces on each atom and an adequate time step are thus pivotal to a reasonable representation of the system. An explanation of force field parameters can be found in reference [12]. For the rest of this section, we will focus on a specific result using the molecular dynamics software NAMD [13].

The typical folding time for even small amino acid segments may extend into the millisecond range. For a molecular dynamics simulation with a typical time-step of 1 fs, the millisecond timescale poses a major feat for both computation time and storage of data. To overcome these hurdles, researchers have used enhanced sampling methods to force a system to explore phase space in more computationally feasible (μs) times. Specifically, in a recent study involving the folding of a five-helix bundle fragment of λ-repressor, the temperature of the system was stochastically varied as a function of time [11]. Exposing the system to high temperatures allows the protein to move out of possible metastable traps in the folding trajectory.

Results from the trajectory of λ-repressor with enhanced sampling showed the protein adopt the native state (as seen from the crystal structure) twice over a 10 μs trajectory. In contrast, a 100 μs simulation of the denatured protein with a constant temperature failed to show full folding [11]. The inability to observe folding over the longer trajectory likely points to a free energy profile for the protein that has many metastable wells. Thus, enhanced sampling methods provide a useful tool to examine the full range of intermediates along a folding pathway but fail to provide detailed kinetic information about folding events due to the large perturbation of the system at each time-step.

2.4 Experimental Fast Folding

The timescale for protein folding after denaturation from different stimuli, such as temperature or pressure, may vary greatly. Specifically, for the aforementioned λ-repressor, the folding timescale from a pressure denatured state is known to be around 2 μs [14]. This result allows for direct comparison with molecular dynamics simulations and can be used to refine protein folding force fields. Furthermore, for fast folding proteins, the denatured state may provide an important intermediate along the normal folding pathway.

To resolve protein folding on the microsecond time scale, researchers use fluorescence techniques to monitor the local environment of key amino acids in the protein. Upon a pressure jump, fluorescence measurements can confirm the transition to some denatured state. Furthermore, the fluorescence data can be used to measure the fluctuations along the folding pathway to the native state upon return to normal pressure [14]. Through a series of temperature or pressure jump denaturations, experimental data can be compared directly to the enhanced sampling methods used in molecular dynamics.

3 Discussion

3.1 2-D Lattice Model

An obvious oversimplification of the two dimensional lattice model is the exclusion of entropic effects, which may be even more important in three dimensions. However, short proteins in two dimensions do maintain a surface-to-volume ratio similar to long proteins in three dimensions [1]. Furthermore, the simplicity of two dimensions allows an exact characterization of short polymer sequences. For more complicated three dimensional lattice models, we refer the reader to the references [3, 15].

Another simplification that needs consideration concerns the reduction of amino acid types to hydrophobic or hydrophilic. In a normal cell, amino acids may be polar, acidic, basic, or hydrophobic; side chains for each type may vary greatly in size [6]. To explore the ramifications of this simplification, a model has been proposed that considers four types of amino acids, namely hydrophobic and hydrophilic each with a spin value 0 or 1. The energy between non-sequential residues is then determined based on both the spin and type of residue [2]. One of the main qualitative differences presented by the increased complexity of the four amino acid type model is the existence of a larger set of minimum energy structures compared to the two type model. This is to be expected due to the increased interactions possible, which yield more diverse folded structures [2].

A final major consideration involves the use of a more complicated interaction potential. The most simple two dimensional lattice model only involves interactions amongst non-sequential hydrophobic residues, which neglects favorable interactions amongst hydrophilic residues and solvent and unfavorable interactions amongst hydrophobic residues and solvent. To use the most realistic interaction potentials, standard all-atom molecular dynamics simulations are often employed. Complex potentials, however, drastically slow the calculation of trajectories due to the computation of forces for all atoms in the system. For a review of the effects of multiple simple lattice potentials, the reader is referred to [3].

Despite the drastic simplifications of the two dimensional lattice model, several qualitative aspects of the protein folding problem are reproduced. First of all, for a subset of the overall sequences, a first order folding transition occurs as the interaction energy is increased [1]. Notice that physically a change in interaction energy may be the result of a change in local solvent content or the presence of a chaperone protein. Furthermore, many states do not have a distinct free energy minimum, corresponding to a subset of long-lived intermediate states [1]. This result may partially be due to the simplicity of the interaction potential, but molecular dynamics simulations of actual proteins do indeed show slow folding pathways indicative of a glassy free energy surface [11].

3.2 Rate Matrix Kinetics

The major simplification of the rate matrix model presented [4] concerns the uniformity of the rate constants. We expect that transitions amongst intermediate states are not uniform and that a realistic protein may spend the duration of its existence in the cell exploring a confined region of configuration space. The experimental evidence for the previous statement is clearly seen by considering pressure denatured proteins [14]. Fast folding after a pressure

jump implies that the denatured state does not explore the majority of possible intermediate states. However, the simplification of uniform rate constants allows the the model to be exactly solved and thus provides interesting insights into possible kinetics.

Perhaps the most interesting result of the rate matrix approach is the existence of two qualitatively different kinetic theories, namely the long-lived folded state and the "kinetic hub" [4]. Both behaviors could be used in regulatory roles in the cell. For example a long-lived folded state may be necessary when a chaperone is needed for proper folding. The protein would likely not fold on its own, as transitions between intermediate states are favored, but with the help of a chaperone would form a stable state to be used for a specific function. The "kinetic hub" scenario may be necessary for proteins that spontaneously form in solution, as the folded state would predominate, allowing a necessary function to be stochastically achieved. Thus, coupled with a simplistic lattice model, the rate matrix approach provides a useful theoretical tool to probe possible protein interactions leading to a native state.

3.3 Molecular Dynamics Simulations

Molecular dynamics simulations enable researches to mine data from a complex system using a set of fundamental stochastic differential equations [13]. As noted earlier, the accurate reproduction of physical phenomena during a simulation relies on the parameterization of the forces amongst atoms [12]. Thus, the correct parameterization of molecular models is of utmost importance in producing accurate protein folding simulations. To ensure correct parameterization of model systems, direct comparison with experiment is pivotal. Hence, folding simulations should be closely tied to experimental folding on a μs timescale (see [11, 14]).

To simulate folding of large proteins and/or to reach long timescales, several coarse-grained models have been proposed (for a thorough review see [16]). A typical coarse-grained model represents a group of atoms in a system as a bead. The bead then interacts with the system through a redefined interaction potential. Coarse-graining can thus drastically lower the number of particles in a system, and the dynamics are typically much faster than all-atom simulations [16]. Coarse-graining, however, is highly sensitive to parameterization. Furthermore, parameters must be tweaked to reproduce all-atom phenomena, because a direct experimental measurement of a coarse-grained parameter may not be feasible. We do expect coarse-grained models to work well for systems comprised of only a few different types of monomers, including lipids and certain simple proteins [16], because much conformational information about individual monomers is lost in the coarse-graining process.

Another important consideration for simulations involves the representation of solvent degrees of freedom. Several water models are commonly employed in molecular dynamics simulations that yield different diffusion coefficients and radial distribution functions (for a review of all-atom water models and their properties, see [17]). The choice of water model again reflects the need to closely tie simulation to experimental data. Simulations can give useful extrapolation to novel systems using parameterized force fields, but the results should be compared to a known structure or pathway for verification.

Even with the possible problems inherent in using a parameterized force field, molecular dynamics simulations are currently the pinnacle of accurately representing protein folding

through a lattice model. Furthermore, with the experimental capabilities of observing folding pathways on a μs timescale, simulations provide an ideal setting to probe specific intermediate states between known configurations. To reliably proceed to the study of folding amongst protein complexes or large proteins, advances in both computational resources and force field development will be necessary.

4 Further Considerations

4.1 Theory

Thus far, we have considered protein configurations to consist of a set of coordinates for each amino acid on a lattice. In this section, we will briefly discuss a different viewpoint, namely the characterization of the energy spectrum for folding. We take as our starting point the perhaps naive assumption that the "energies of different conformations should be considered independent random variables [5]." Furthermore, a specific sequence of amino acids is taken to be a specific realization of disorder for the system [5]. These assumptions correspond to the random-energy model [18]. The random-energy model provides a standard starting point for more complicated theories of the protein folding energy landscape.

We will now state a few of the interesting properties of the random energy model (for a more complete discussion and derivations, see the appendix in [5]). The following properties are paraphrased from [5].

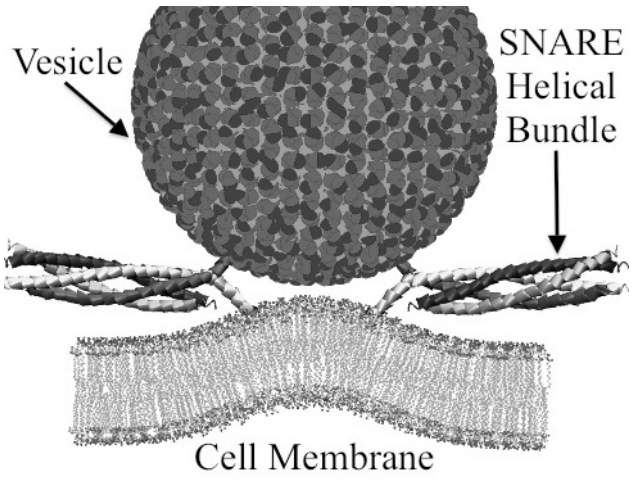
1. The energy spectrum contains both a quasicontinuous portion and a discrete portion
2. Above a certain temperature (T_f), the system explores the quasicontinuous portion
3. Below T_f, the system explores the discrete portion

Thus, the random energy model qualitatively reproduces the folding transition at a certain threshold temperature. The transition implies that entropic considerations dominate above T_f and become negligible below T_f. This result is consistent with the findings in the two dimensional lattice model, where a transition occurred for a specific value of the amino acid interaction energy. Lowering the temperature in the random-energy model corresponds to raising $|\epsilon|$ in the lattice model. The random-energy model also makes an important connection between protein folding and theories of disordered systems (see [5] for more discussion).

4.2 SNARE Complex Formation

We now discuss a specific system where the formation of a helical bundle through the folding of multiple proteins is observed. The explanation of the folding pathway for such systems is at the frontier of current molecular dynamics simulations. The soluble N-ethylmaleimide sensitive factor attachment protein receptor (SNARE) complex is vital in synaptic transmission [19]. The SNARE complex is comprised of three proteins that are thought to undergo a drastic transition from a disordered state to an ordered four-helix bundle [19].

To understand how the SNARE complex is regulated, a sequential ordering of folding events must be determined. Thus, a basic understanding of synaptic transmission will necessitate a deeper understanding of the transition for a set of proteins between a continuous (or quasicontinuous spectrum) and a native subset of states. A major challenge for molecular dynamics simulations will be to develop the methods necessary to observe multi-protein complex formation on a computationally feasible timescale.



4.3 Closing Thoughts

The theoretical and experimental techniques covered in this paper are a very small subset of the ongoing research into protein folding. For the interested reader, please see the following references, which are mostly theoretical in nature [5, 9, 15, 20]. Also, multiple groups at the University of Illinois at Urbana-Champaign are currently doing research on protein folding. We have specifically discussed two examples, namely Dr. Klaus Schulten [11] and Dr. Martin Gruebele [14] who are active in molecular dynamics simulations and experimental fast folding respectively. Finally, I would like to thank Dr. Nigel Goldenfeld for a great semester in Phase Transitions.

References

[1] Lau, K. F.; Dill, K. A. *Macromolecules* **1989**, *22*, 3986-3997

[2] Liu, H.; Tang, L. *Physical Review E* **2006**, *74*, 051918-1-8

[3] Qin, M.; Wang, J.; Tang, Y.; Wang, W. *Physical Review E* **2003**, *67*, 061905-1-8

[4] Pande, V.S. *Physical Review Letters* **2010**, *105*, 198101-1-4

[5] Pande, V.S.; Grosberg, A.Y.; Tanaka, T. *Reviews of Modern Physics* **2000**, *72*, 259-314

[6] Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. *Molecular Biology of the Cell* **2008**, Fifth Edition, Garland Science, Taylor and Francis Group Publishing

[7] Mirsky, A. E.; Pauling, L. *Proc. Natl. Acad. Sci. USA* **1936**, *22*, 439-447

[8] Anfinsen, C.B. *Science* **1973**, *181*, 223-230

[9] Privalov, P.L. *Advances in Protein Chemistry* **1979**, *33* 167-241

[10] Dill, K. A. *Biochemistry* **1985**, *24*, 1501-1509

[11] Liu, Y.; Strumpfer, J.; Freddolino, P.L.; Gruebele, M.; Schulten, K. *The Journal of Physical Chemistry Letters* **2012**, *3*, 1117-1123

[12] Becker, O.M.; MacKerell, A.D.; Roux, B.; Watanabe, M. *Computational Biochemistry and Biophysics* **2001**, Marcel Dekker, Inc. New York, p. 7-38

[13] Phillips, J.C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.D.; Kale, L.; Schulten, K. *Journal of Computational Chemistry* **2005**, *26*, 1781-1802

[14] Dumont, C.; Emilsson, T.; Gruebele, M. *Nature Methods* **2009**, *6*, 515-520

[15] Bryngelson, J.D.; Onuchic, J.N.; Socci, N.D.; Wolynes, P.G. *Proteins: Structure, Function, and Genetics* **1995**, *21*, 167-195

[16] Voth, G.A. *Coarse-graining of Condensed Phase and Biomolecular Systems* **2009**, Boca Raton, CRC Press

[17] Mark, P.; Nilsson, L. *Journal of Physical Chemistry A* **2001**, *105*, 9954-9960

[18] Derrida, B. *Physical Review Letters* **1980**, *45*, 79-82

[19] Brunger, A.T.; Weninger, K.; Bowen, M.; Chu, S. *Annual Review of Biochemistry* **2009**, *78*, 903-928

[20] Zhuravlev, P.I.; Papoian, A. *Quarterly Reviews of Biophysics* **2010**, *43*, 295-332