Protein Folding Dynamics via Quantification of Kinematic Energy Landscape

Séma Kachalo, Hsiao-Mei Lu, and Jie Liang*

Department of Bioengineering, MC-063, University of Illinois at Chicago, Chicago, Illinois 60608, USA

(Received 18 August 2005; published 8 February 2006)

We study folding dynamics of protein-like sequences on a square lattice using a physical move set that exhausts all possible conformational changes. By analytically solving the master equation, we follow the time-dependent probabilities of occupancy of all 802,075 conformations of 16 mers over 7 orders of time span. We find that (i) folding rates of these protein-like sequences of the same length can differ by 4 orders of magnitude, (ii) folding rates of sequences of the same conformation can differ by a factor of 190, and (iii) parameters of the native structures, designability, and thermodynamic properties are weak predictors of the folding rates; rather, a basin analysis of the kinematic energy landscape defined by the moves can provide an excellent account of the observed folding rates.

DOI: 10.1103/PhysRevLett.96.058106

PACS numbers: 87.15.He, 87.15.Cc, 87.15.Aa

The dynamics of protein folding has been studied extensively [1,2]. A remarkable observation is that protein folding rates are well correlated with their native structural properties [1]. A native-centric view postulates that protein folding rates are largely determined by the topology of their native structure [3]. Theoretical models using the Go potential, where only native contacts contribute energetically, are very successful in reproducing observed folding rates [2,4].

However, the extent to which native structure determines the folding rate remains unclear. By the native-centric view, different sequences for the same protein structural fold would all have very similar folding rates, as they share essentially the same native structure topology. However, this is not consistent with experimental results. As the folding rates of simple single-domain proteins differ by 6 orders of magnitude [3], folding rates may be very heterogeneous. A recent experimental study showed that a designed artificial protein with no homologous sequence in nature that adopts the same structure as a natural protein can fold 4000 times faster [5]. A distinct possibility is that the empirical correlation between properties of native protein structures and folding rates may arise from inadequate sampling in the sequence space due to accumulated biased natural selection and limited genetic drift, rather than from intrinsic physical properties of proteins.

In this Letter, we use the two-dimensional hydrophobic and polar (HP) lattice model [6] to study the relationship of folding rates, native structure topology, thermodynamic properties, and effects of sequence variation. We model the physical movement of protein chains. Real proteins cannot immediately jump from one conformation to another arbitrary conformation. Two conformations of the same energy may be well separated kinetically. We regard protein movement as a sequence of successive conformational changes, each represented by a physically realizable primitive move. The physical move set we developed exhausts all possible conformational changes for a structure. We use the master equation to study the folding dynamics of foldable sequences of length 16. While the master equation provides an exact solution [6,7], in the past it was necessary to cluster conformations of larger systems into macrostates to reduce the size of the transition matrix [8], therefore making the use of physical moves infeasible. Here we directly solve the eigenvalue problem of the 802,075 × 802,075 transition matrix and develop a method to monitor the time-dependent probability of occupancy of all conformations simultaneously from the first kinetic move until reaching half-equilibrium concentration over 7 orders of time scale.

Our results show that the properties of native structure, designability, and thermodynamic properties are inadequate to explain protein folding dynamics in our model systems. We found that protein-like sequences can fold into the same native structure with folding rates differing as much as 190 times and sequences of the same length and energy gap can differ by 4 orders of magnitude in folding rate. Instead of thermodynamic properties, we show that properties of the move-connected energy landscape defined by the connection graph of physical moves can provide an excellent account for observed folding rates.

Model.—We use the following energy model for different types of nonbonded HP contacts: 

\[
E_{\text{HH}} = -1, \quad E_{\text{HP}} = 0, \quad \text{and} \quad E_{\text{PP}} = 0.
\]

By evaluating the energy level of all 216 sequences of 16 mers on all enumerated \(|\Omega| = 802,075\) conformations, we have identified 26 sequences that all fold into the same ground state conformation (Fig. 1). This set of sequences forms the largest protein family, where each sequence adopts the same conformation, and all are connected by (a series of) point mutations. Altogether, there are 1539 foldable sequences with unique ground state conformations. There are 456 conformations that are the unique ground state for 1 or more foldable sequences.

We develop a set of physically possible primitive moves [Fig. 1(c)]. They are generalizations of corner move, crankshaft move, and pivot move. We exhaust all possible occurrence of such moves for every conformation. We verified that this move set is ergodic, i.e., all conformations are connected to each other by a series of primitive moves. With this move set, the simple energy scheme of the HP
Thermodynamics and folding rates.—Several thermodynamic properties have been proposed to be determinants of protein folding rates. We found that protein stability as measured by the total contact energy is correlated with log $k_f$ ($R^2 = 0.71$); i.e., more stable proteins fold slower in general (Fig. 2). Because stable proteins have lower ground state energy, some local minima will also have relatively deep energy traps. As a result, more stable proteins will have slower folding rates because they can be trapped in such local minima. However, the folding rates of sequences of the same ground state energy can still differ as much as $10^3$. The heterogeneity of the folding rate was already noted in an earlier study using the macrostate approximation [6]. Here we found that even sequences that fold into the same conformation shown in Fig. 1(a) demonstrate a wide range of rates, from $1.1 \times 10^{-3}$ to $5.8 \times 10^{-6}$, which is much larger than the difference between the average folding rates for sequences of different native state energies. Protein stability therefore provides part, but not the main explanation of the heterogeneity of folding rates.

The energy gap between ground state and excited state was thought to be the necessary and sufficient determinant of folding rate [9]. For all 1456 proteinlike sequences of $N = 16$, the energy gap between the lowest state and the next state is $\Delta E = 1$. The diversity in folding rate $k_f$ shown in Fig. 2 clearly indicates that energy gap is not a determining factor for the folding rate. The correlation $R^2$ between log $k_f$ and the energy gap normalized by standard deviation is 0.01.

Another thermodynamic property thought to be an important determinant of folding rates is the collapse cooperativity $\sigma = 1 - T_f/T_0$ [10], where $T_f$ is as defined
earlier, and $T_\theta$ the temperature when heat capacity $C(T)$ reaches its maximum. Figure 3 shows that for the 26 sequences that fold to the same native structure in Fig. 1(b), there is a weak correlation ($R^2 = 0.38$) between collapse cooperativity and $\log k_f$. Large variances in observed folding rates exist for sequences of similar collapse cooperativity.

The number of sequences that take a specific conformation as the unique ground state is thought to be correlated with overall protein stability and folding rates [11]. We calculated, in addition, $k_f$ for a group of 79 singleton sequences with no sequence homologs that fold to the same native conformations. The distribution of $k_f$s for the singleton sequences and the 26 sequences shown in Fig. 2 demonstrate similarly large variation. For our model, designability is not an important determinant of the folding rates.

The inverse participation ratio $I$ is commonly used to characterize the localization of eigenvectors. It is defined as $I = \sum_k v_k^4$, where $v_k$ is the $k$th coefficient of the normalized eigenvector. The correlation between $I$ for the equilibrium eigenvector and the folding rate for the 26 sequences is rather poor ($R^2 = 2 \times 10^{-3}$).

Kinematic determinants of folding landscape.—Protein folding kinetics are intrinsically determined by physical movement of molecules. Weak correlations of the folding rate with thermodynamic properties are not surprising. Thermodynamic properties of a sequence can be calculated if its complete set of conformations are enumerated. Such properties are not affected by the kinetic connections between conformations. A smooth energy landscape ensuring fast folding can be easily permuted into a rugged landscape by assuming different transition rules between conformations. Both will have the same thermodynamic properties, but the resulting folding rates for the same sequence will be very different. The energy landscape of folding is dictated by the connection graph of states defined by the move set. Characterizing such a kinematic energy landscape is therefore essential for studying protein folding dynamics.

Although the energy landscape contains 802,075 conformations, each is connected by the move set to only a limited number (~30) of conformations. We can identify states that are local minima, i.e., all the states connected to which by moves have higher energy. A simple characterization of the kinematic energy landscape is then the number count $n_{\text{min}}$ of the local minima. Figure 3(b) shows that an excellent correlation of $\log k_f$ and $n_{\text{min}}$ ($R^2 = 0.85$) can be found for the 26 HP sequences that fold into the same conformation.

Our conclusions are not sensitive to temperature $T$. When $T$ is raised from 0.20 to 0.21 (equivalent to raising $T$ from 300 to 315 K), we found that the folding rate $k_f$ of the 26 sequences all increases. Although $k_f$ for a slow folder increases more (by a factor of 2.0 versus a factor of 1.4 for fast folders), $k_f$ at $T = 0.21$ is well correlated with $k_f$ at $T = 0.20$. The correlation coefficients of $\log k_f$ with the number of local minima, collapse cooperativity (Fig. 3), and other thermodynamic parameters are essentially unchanged.

Time evolution and basin analysis.—Monitoring the exact time evolution of all conformations simultaneously until reaching equilibrium during folding is a challenging task. Mathematically, the model of the master equation is equivalent to a Markov process, where the population vector of conformations at time $t + k\Delta t$ is given by $p(t + k\Delta t) = M^k p(t)$, where $M = I + R \cdot \Delta t$, $I$ being the identity matrix. However, the $k$ time step Markov matrix $M^k$ rapidly becomes a dense matrix, and following the time evolution of folding by a straightforward matrix multiplication of $O(|\Omega|^3 \log k)$ steps, becomes impossible for a large matrix of size $|\Omega| = 802,075$ and $k = 10^6 - 10^{10}$. The analytical solution of $p(t) = \sum_i C_i n_i e^{-\lambda_i t}$ through diagonalization is also impractical, as it is only possible to calculate a few eigenvectors and eigenvalues for a large matrix.

We seek an accurate solution without the approximation of macrostates. Taking advantage of the sparsity of the rate matrix $R$, we follow the approach of Sidje [12] and use the analytical solution of matrix exponential: $p(t) = e^{Rt} p(0)$, where $e^{Rt}$ is defined by the Taylor expansion $e^{Rt} = I + tR + \frac{t^2}{2}R^2 + \cdots + \frac{t^k}{k!}R^k + \cdots$. This expansion itself is impractical, as it also involves a large matrix product of increasing density. Plus, the entries in the matrix terms may have alternating signs and hence cause numerical instability. A better approach is to expand $e^{Rt} p(0)$ in the Krylov subspace $K_m$ defined as

$$K_m(R, p(0)) \equiv \text{Span}(p(0), \ldots, (Rt)^{m-1} p(0)).$$

Denoting $\| \cdot \|_2$ as the 2 norm of a vector or matrix, our approximation then becomes $p(t) = \|p(0)\|_2 V_{m+1} e^{H^{m+1}} e_1$, where $e_1$ is the first unit basis vector, $V_{m+1}$ is a $(m+1) \times (m+1)$ matrix formed by the orthonormal basis of the Krylov subspace, and $H_{m+1}$ the upper Heisenberg matrix, both computed from an Arnoldi algorithm. The error can be bounded by $O(e^{-m-t} R_{m+1} e_1 (\|R\|_2/(m+1))^m)$. We now only need to compute...
explicitly $e^{\hat{H}_{m+1}}$. Because $m$ is much smaller than 802 075, this is a simpler problem. A special form of the Padé rational of polynomials instead of the Taylor expansion is used for this \cite{12}: $e^{\hat{H}_{m+1}} = N_{pp}(\hat{H}_{m+1})/N_{pp}(-t\hat{H}_{m+1})$, where $N_{pp}(\hat{H}_{m+1}) = \sum_{k=0}^{p} c_{k}(t\hat{H}_{m+1})^{k}$ and $c_{k} = c_{k-1}^{p+1-k}/(p+1-k)!$. In our calculation, we select $m = 30$.

Figure 4 shows an example of an HP sequence \cite{sequence C in Fig. 1(a)} and the time evolution of its native conformation and several local minima conformations. The time evolution of the native conformation shows an initial fast phase up to $t \approx 50$ time units. In principle, the local minima conformations can follow different kinetic processes: Some could be transiently accumulating, and others either monotonically accumulating or monotonically decreasing. Based on the computed trajectories of time evolution, we find that the dynamic behavior of local minima conformations can be predicted from basin analysis of the move-connected energy landscape. We define the size of the basin associated with each local minimum state $i$ computationally by artificially making every local minimum an absorption state, i.e., a sink of infinite depth, such that once reached, no molecule can escape. This is achieved by assigning $r_{ij} = 0$ and $r_{ii} = 1$ for each local minimum state $i$ \cite{13}, $p_{i}(t = \infty)$ therefore reflects the size of the basin of the $i$th local minimum. We define the accumulation ratio as $Q = e^{-p_{i}(0)/\sum_{j} e^{-p_{j}}}$, where $Q > 1$, state $i$ is most likely a transient accumulating state, i.e., the other conformations in its basin first rapidly flow to state $i$ before transiting to conformations outside the basin. If $Q < 1$, depending on its initial probability of occupancy and the final Boltzmann factor, state $i$ may be either a monotonically decaying or accumulating state. We find that among the 493 local minima states for this sequence, all except 3 are transiently accumulating, indicating they are responsible for forming transient state ensemble of various time scale.

To understand whether the formation of certain native contacts facilitate folding, we examine the time evolution of each of the 8 native contacts (a)–(h) in Fig. 1(b) for the 26 sequences. We found that fast folders have larger amount native contact $d$ ($R^{2} = 0.74$–0.81 with log$\xi$), and contact $c$ at the transient time of 50–100 (Fig. 4), indicating that these contacts are critical for folding by restricting favorably the conformational search space. The formation of other native contacts seems not to be directly related to folding rates.

To conclude, we studied protein folding dynamics using a model based on detailed physical moves and the exact solution of the master equation. We found that folding rates vary enormously for sequences of the same length, energy, energy gap, and even of the same ground state conformation. In contrast to the thermodynamic parameters which are weak predictors of folding rates, properties of the kinematic landscape defined by the physical moves provide excellent correlation with folding rates. With the computation of time evolution of individual conformation from the first move to the half time of equilibrium, we show that many transiently accumulating intermediate states can be identified by basin analysis.

We thank Dr. Ken Dill, Dr. Bosco Ho, Dr. Xiaofan Li, Dr. Banu Ozkan, Dr. Dev Thirumalai, Dr. Jin Wang, and Dr. Weitao Yang for helpful discussions. This work is supported by NSF DBI0133856, NIH GM68958, and Whitaker TF-04-0023.

*Corresponding author.
Electronic address: jliang@uic.edu

\begin{thebibliography}{99}
\bibitem{13} I. Chang \textit{et al.}, Protein Science \textbf{13}, 2446 (2004).
\end{thebibliography}