Proceedings of the 2005 IEEE
Engineering in Medicine and Biology 27th Annual Conference
Shanghai, China, September 1-4, 2005

Protein folding dynamics in lattice model with physical movement

Sëma Kachalo, Hsiao-Mei Lu and Jie Liang*
Department of Bioengineering, MC-063
University of Illinois at Chicago, Chicago, IL, 60305, U.S.A.
{sema,hlu7,jliang}@uic.edu

Abstract— **We study folding dynamcis of protein-like sequences on square lattice by constructing a physically realizable move set that exhausts all possible conformational changes for a structure. By solving the master equation of 16-mer characterized by a 802,075×802,075 transition matrix, we monitor the time-dependent probabilities of occupancy of all conforamtions over 9-orders of time scale from the £rst kinetic move until reaching Boltzmann equlibrium. We £nd that folding rates of protein-like sequences adopting the same ground state conformation differ as much as 200 times, and parameters of the native structures, designability, and thermodynac properties are weak predictors of the folding rates in our model systems. Instead, we show that properties of the kinetic energy landscape de£ned by the connection graph of physical moves can provide excellent account for observed folding rates. Without the approximation of macrostates, we show how transiently accumulating intermediate states can be identi£ed by basin analysis of the kinetic energy landscape.**

I. INTRODUCTION

The dynamics of protein folding has been studied extensively [1, 3–5]; A remarkable empirical observation is that protein folding rates are well correlated with their native structural properties [1]. A native centric view therefore postulates that protein folding rates are largely determined by the topology of its native structure [6]. Theoretical models using Gō potential where only contacts in native structure contribute energetically are very successful in reproducing experimentally observed folding rates [2–5].

However, the extent to which native structure determines 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 some experimental results. As the folding rates of simple single-domain proteins differ by 6 orders of magnitude [6], protein folding rates may be very heterogeneous. A recent experimental study showed that a designed arti£cial protein sequence unseen in nature folds 4,000 faster than natural protein adopting the same structure [7]. The empirical correlation between properties of native protein structures and folding rates may arrise from accumulated biased natural selection rather than from intrinsic physical properties of proteins.

In this paper, we use two-dimensional hydrophobic and polar (HP) lattice model [9] to study the relationship of folding rates, native structure topology, thermodynamic properties, and effects of sequence variation. We study folding dynamics by modeling the physical movement of protein chains. Real protein cannot immediately jump from one speci£c conformation to another arbitrary conformation. Two conformations of the same energy may be well separated kinetically. Protein movement can be regarded 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 master equation to study the folding dynamcis of foldable sequences of length 16. While master equation provides an exact solution of the folding dynamics problem [9, 10], in the past it was necessary to cluster conformations of larger systems into macrostates to reduce the size of the transition matrix [11], therefore making the use of physical moves infeasible. In folding studies, physical moves have been used in conjunction of Monte Carlo sampling of a £nite number of folding trajectories for a few selected sequences [12]. 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 individual conforamtions simultaneously from the £rst kinetic move until reaching Boltzmann equlibrium over 9-orders of time scale.

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

II. MODEL

We use the following energy model for different types of nonbonded HP contacts: $E_{HH} = 1$, $E_{HP} = 0$, and $E_{PP} = 0$. By evaluating the energy level of all 2^{16} sequences of 16-mers on all enumerated $|\Omega| = 802,075$ conformations, we have identi£ed 26 sequences that all fold into the same ground state conformation (Fig. 1.). This set of sequences form the largest protein family, where each sequence adopts the same conformation, and all are connected by a (series

*Corresponding author.

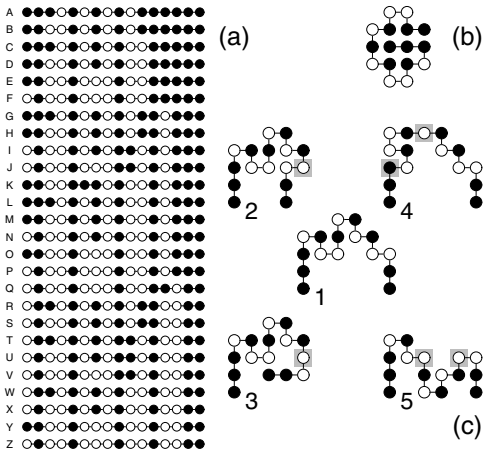

Fig. 1. Protein-like sequences and the set of primitive moves. The largest protein family contains (a) 26 sequences that all fold into (b) the same structure. Filled circles are H residues. (c) The move set includes: among (1, 2, and 3), single point moves rotate around a single point; (b) (1 and 4), generalized corner moves are re¤ections arond a diagonal axis connecting any two residues; (c) (1 and 5), generalized crankshaft moves are re¤ections around a horizontal or vertical axis. Points of rotation are on grey background. For a given N-mer, we exhaustively search all possible position for point moves, all possible pairs of positions for possible generalized corner moves and generalized crankshaft moves.

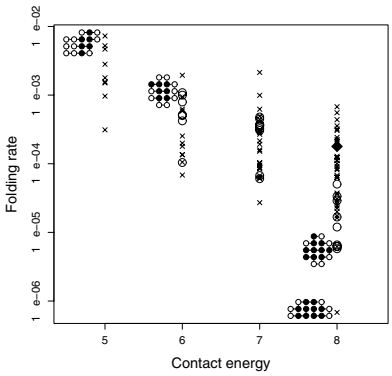

Fig. 2. The correlation of $\log k_f$ and ground state contact energy. A circle represents one of the 26 sequences shown in Fig. 1., and a cross represents one of the 79 singleton sequences. Native conformations for a few sequences are also shown.

of) point mutations. Altogether, there are 1,539 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 move set that includes only physically possible primitive moves (Fig. 1.c). They are generalizations of corner move, crankshaft move, and pivot move. We exhaust all possible occurence of such moves for a given conformation. We veri£ed that this move set is ergodic, *i.e.*, all conforamtions are connected to each other by a series of primitive moves. With this move set, the simple energy scheme of the HP model leads to a complex energy landscape, with numerous local minima for a foldable sequence.

We use Metropolis-type of dynamics to assign the transition rate r_{ij} from conformation i to a neighbor conformation j connected by a move: $r_{ij} = 1$ if $E(j) \leq E(i)$; $r_{ij} = e^{-[E(j)-E(i)]/T}$ if $E(j) > E(i)$; and $r_{ij} = -\sum_{i \neq k} r_{ik}$, if $j = i$. For non-neighbors, $r_{ij} = 0$. Other dynamics such as Glauber dynamics are also possible. For simplicity, we assume the effects of viscosity and friction are negligible, and the transition rate of a pivot move is simply determined by this Metropolis dynamics.

We follow [10, 11] and use a master equation to study protein folding dynamcis. Let $p_i(t)$ be the probability that the HP molecule takes the i-th conformation at the time t, then $dp_i(t)/dt = \sum_{i \neq j}[r_{ji}p_j(t) - r_{ij}p_i(t)]$. Written in vector form, we have: $d\boldsymbol{p}(t)/dt = \boldsymbol{R}\boldsymbol{p}(t)$, where $\boldsymbol{R}$ is the rate matrix whose entries are de£ned by the above expression. We choose temperature $T = 0.2$ in unit of $\Delta E_{HH}/k_B$, which is below the folding temperature T_f when 50% of molecules take the native conformation. For 16-mer HP molecules, folding temperatures vary from ~ 0.2 to ~ 0.5 for different sequences.

A general solution of the master equation can be written as $\boldsymbol{p}(t) = \sum_i C_i \boldsymbol{n}_i e^{-\lambda_i t}$ with $C_i = \boldsymbol{v}_i^T \boldsymbol{P}(0)$, where λ_i is the i-the eigenvalue of the rate matrix $\boldsymbol{R}$, $\boldsymbol{n}_i$ the corresponding right eigenvector, $\boldsymbol{v}_i$ the left eigenvector, and $\boldsymbol{P}(0)$ the initial vector of distribuiton of conformations. In this study, we use the high temperature condition and assign $\boldsymbol{p}(0) = \mathbf{1}/|\Omega|$. Two eigenvalues are of particular interest: $\lambda_0 = 0$ corresponds to the equlibrium Boltzmann distribution, and the smallest none-zero eigenvalue λ_1 determines the slowest mode of relaxation. Following [11], we take λ_1 as the folding rate of the protein. Although the full computation of all eigenvalues and eigenvectors for a $802,075 \times 802,075$ matrix $\boldsymbol{M}$ is infeasible, λ_1 and the corresponding eigenvectors $\boldsymbol{n}_1$ and $\boldsymbol{v}_1$ can be computed by standard inverse power method.

III. THERMODYNAMICS AND FOLDING RATES

Several thermodynamic properties of proteins have been proposed to be determinants of protein folding rates. We found that protein stability as measured by the total contact engery are correlated with $\log k_f$ ($R^2 = 0.71$), *i.e.*, more stable proteins fold faster in general (Fig. 2). However, the folding rates of sequences of the same ground state energy can differ as much as 10^4. The heterogeneity of folding rate was already noted in an earlier study based on the solution of a master equation using macrostate approximation [9]. Here we found that even sequences that fold into the same conformation shown in Fig. 1.a demonstrate a wide range of rates, from 1.1×10^{-3} to 5.8×10^{-6}, which is much larger than the difference between the average folding rates for sequences of different native state energies. Protein stability therefore provide some but not the main explanation of the heterogeneity of folding rates.

Energy gap between ground state and excited staet was thought to be the neccessary and suf£cient determinant of folding rate [14]. For all 1,456 protein-like 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_fs shown in Fig. 2 clearly indicates that energy gap is not a determining

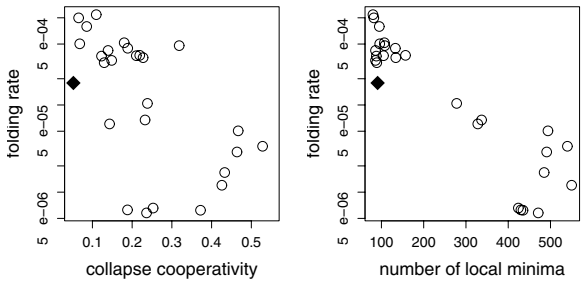

Fig. 3. Examples of the correlation of folding rate with thermodynamic properties and kinetic landscape properties. (a) Folding rate and collapse cooperativity σ have weak correlation ($R^2 = 0.38$). (b) Folding rate has excellent correlation with the number of local minima ($R^2 = 0.85$), a property of the kinetic landscape.

factor for the folding rate. The correlation R^2 for energy gap measured by Z-score is 0.01.

Another thermodynamic property thought to be an important determinant of folding rates is the collapse cooperativity $\sigma = 1 - T_f/T_\theta$ [15], where T_f is the temperature when 50% of the molecules are in the native state, and T_θ is the temperature when heat capacity $C(T)$ reaches its maximum. Fig. 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 variance in observed folding rates exist for sequences of similar collapse cooperativity.

Designability of a conformation, namely, the number of sequences that take this conformation as unique ground state is thought to be correlated with overall protein stability and folding rates [16]. We calculated the folding rate k_f for the 26 HP sequences for the structure of highest designability (Fig. 1.) and k_fs for a group of 79 singleton sequences with no sequence homologs that fold to the same native conformations . The distribution of k_fs for the singleton sequences and the 26 sequences shown in Fig. 2. demonstrate similarly large varation. For our model, designability is not an important determinant of the folding rates.

IV. KINETIC DETERMINANTE OF FOLDING LANDSCAPE

Protein folding kinetics are intrinsically determined by physical movement of molecules. Weak correlations of the folding rate with various thermodynamic properties are not surprising. Thermodynamic properties of an ensemble of conformations for a sequence can be calculated if the 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 differnt transition rules between conformations. In this case, both will have the same thermodynamic properties, but the resulting folding rates for the same sequence will be very different. The kinetic energy landscape of folding is dictated by the connection graph of states de£ned by the move set. Characterizing such kinetically determined energy landscape is therefore essential for studying protein folding dynamics.

Although the energy landscape contain 802,075 conformations, each is connected by the move set to only a limited number (~ 30) of conformations. This is due to the constraint of physical movement. We characterize the kinetic folding landscape by identifying states that are local minima, *i.e.*, all states connected by moves have higher energy. A simple characterization of the kinetic folding energy landscape is then the number count $n_{\min}$ of the local minima. Fig. 3. shows that an excellent correlation of $\log k_f$ and $n_{\min}$ ($R^2 = 0.85$) can be found for the 26 HP sequences that fold into the same conforamtion.

V. TIME EVOLUTION AND BASIN ANALYSIS

Monitoring the exact time evolution of all individual conformations simultaneously until reaching equilibrium during folding is a challenging task. Mathematically, the model of master equation is equivalent to a Markov chain process, where the population vector of conformations at time $t+k\Delta t$ is given by $\boldsymbol{p}(t+k\Delta t) = \boldsymbol{M}^k\boldsymbol{p}(t)$, where $\boldsymbol{M} = \boldsymbol{I} + \boldsymbol{R}\cdot\Delta t$, $\boldsymbol{I}$ being the identitiy matrix. However, the k-time step Markov matrix $\boldsymbol{M}^k$ rapidly becomes a dense matrix, and following the time evolution of folding by a straightforward matrix multiplication of $\mathcal{O}(kN^3)$ steps becomes impossible for a large matrix of size $N = 802,075$ and $k = 10^{10}$. The analytical solution of $\boldsymbol{p}(t) = \sum_i C_i\boldsymbol{n}_i e^{-\lambda_i t}$ through diagonalization is also impractical, as it is only possible to calculate a few eigen vectors 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 $\boldsymbol{R}$, we follow the approach of Sidje [17] and use the analytical solution of matrix exponential:

$$\boldsymbol{p}(t) = e^{\boldsymbol{R}t}\boldsymbol{p}(0), \tag{1}$$

where $e^{\boldsymbol{R}t}$ is de£ned by the Taylor expansion $e^{\boldsymbol{R}t} = \boldsymbol{I} + t\boldsymbol{R} + \frac{t^2}{2}\boldsymbol{R}^2 + \cdots + \frac{t^k}{k!}\boldsymbol{R}^k + \cdots$. However, the Taylor expansion itself is impractical, as it involves again large matrix product of increasing density. In addition, the entries in the matrix terms may have alternating signs and hence cause numerical instability. A better apporach is to expand $e^{\boldsymbol{R}t}\boldsymbol{p}(0)$ in the Krylov subspacce $\mathcal{K}_m$ de£ned as:

$$\mathcal{K}_m(\boldsymbol{R}t, \boldsymbol{p}(0)) \equiv \text{Span}\{\boldsymbol{p}(0), \cdots, (\boldsymbol{R}t)^{m-1}\boldsymbol{p}(0)\}. \tag{2}$$

Denoting $||\cdot||_2$ as the 2-norm of a vector or matrix, our approximation then becomes $\boldsymbol{p}(t) \approx ||\boldsymbol{p}(0)||_2\boldsymbol{V}_{m+1}e^{t\overline{\boldsymbol{H}}_{m+1}}\boldsymbol{e}_1$, where $\boldsymbol{e}_1$ is the £rst unit basis vector, $\boldsymbol{V}_{m+1}$ is a $(m+1)\times(m+1)$ matrix formed by the orthonormal basis of the Krylov subspace, and $\overline{\boldsymbol{H}}_{m+1}$ the upper Hessenberg matrix, both computed from an Arnoldi algorithm. The error can be bounded by $\mathcal{O}(e^{m-t||\boldsymbol{R}||_2}(t||\boldsymbol{R}||_2/m)^m)$. We now only need to compute explicitly $e^{\overline{\boldsymbol{H}}_{m+1}t}$. Because m is much smaller than 802,075, this is a simpler problem. A special form of the well-known Padé rational of polynomials instead of Taylor expansion is

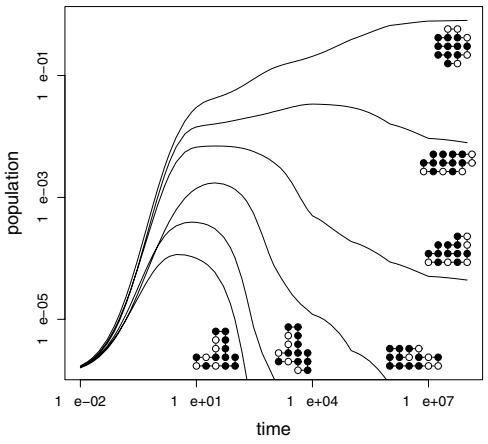

Fig. 4. The time evolution of the native state and several local minima states. The probability of ocupation of native state conformation (top) increases monotonically through a time span of 10^{10}, but local minima conformations go through transiently accumulating intermediate states.

used [17]: $e^{t\overline{\boldsymbol{H}}_{m+1}} \approx N_{pp}(t\overline{\boldsymbol{H}}_{m+1})/N_{pp}(-t\overline{\boldsymbol{H}}_{m+1})$, where $N_{pp}(t\overline{\boldsymbol{H}}_{m+1}) = \sum_{k=0}^{p} c_k(t\overline{\boldsymbol{H}}_{m+1})^k$ and $c_k = c_{k-1} \cdot \frac{p+1-k}{(2p+1-k)k}$. In our calculation, we select $m = 30$.

Fig. 4. shows an example of an HP sequence (c in Fig. 1.a)and the time evolution of its native conformation and several local minima conformations. For this HP sequence, the time evolution of the native conformation shows an initial fast phase upto $t \sim 10^8$ time units. In prinicple, the local minima conformations can follow different kinetic processes: Some could be transiently accumulating, and others either monotopically accumulating or monotopically decreasing. Based on the computed trajectories of time evolution, the dynamic behavior of local minimua conformations can be predicted from *basin analysis* of the kinetic energy landscape. We de£ne the size of the *basin* associated with each local minimum state i by arti£cially making every local minimum an absorption state, *i.e.*, a sink of in£nite depth, such that once reached, no conformation can escape. This is achieved by assigning $r_{ij} = 0$ and $r_{ii} = 1$ for each local minimum state i [19]. $p_i'(t = \infty)$ therefore re¤ects the size of the basin of the i-th local minimum. We de£ne the *accumulation ratio*:

$$\rho = \frac{p_i'(\infty)}{e^{-E_i/T} / \sum_j e^{-E_j/T}}. \quad (3)$$

If $\rho > 1$, state i is most likely a transient accumulating state, *i.e.*, the other conformations in its basin £rst rapidly ¤ow to state i before transiting to conformations outside the basin. If $\rho < 1$, depending on its initial probability of occupancy and the £nal Boltzman factor, state i may be either a monotonically decaying or accumulating state. Among the 493 local minima states for this seqeunce, all except 3 are transiently accumulating, indicating they are responsible for forming transient state ensemble of various time scale.

To conclude, we studied protein folding dynamics using a model based on detailed physical movement and exact solution of the master equation. We found that folding rates vary enormously for sequences of the same length, same energy, same energy gap, and even of the same ground state conformation. In contrast to the thermodynamic parameters which are weak predictor of folding rates, properites of the kinetic landscape de£ned by the phsyical move set provide excellent correlation with folding rates. With the computation of time evolution of individual conformation from the £rst move to reaching equilibrium, we show that many transietly accumulating intermediate states can be identi£ed by landsacpe basin analysis.

VI. ACKNOWLEDGMENTS

We thank Drs. Ken Dill, Bosco Ho, Xiaofan Li, Banu Ozkan, Dev Thirumalai and Jin Wang for helpful discussions. This work is supported by NSF DBI0133856, NIH GM68958, ONR N000140310329, and Whitaker TF-04-0023.

REFERENCES

[1] K. W. Plaxco, K. T. Simons, and D. Baker. Contact order, transition state placement and the refolding rates of single domain proteins. *JMB*, 227:985–994, 1998.

[2] Thomas R. Weikl and Ken A. Dill. Folding rates and low-entropy-loss routes of two-state proteins. *JMB*, 329:585–598, 2003.

[3] O. V. Galzitskaya and A. V. Finkelstein. A theoretical search for golding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl. Acad. Sci. USA*, 96:11299–11304, 1999.

[4] V Munoz and W. A. Eaton. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. USA*, 96:11311–11316, 1999.

[5] E Alm and D. Baker. Prediction of protein-folding mechanisms from free-energy landscapes derived from mative structures. *Proc. Natl. Acad. Sci. USA*, 96:11305–11310, 1999.

[6] Blake Gillespie and Kewin W. Plaxco. Using protein folding rates to test protein folding theories. *Annu. Rev. Biochem.*, 73:837–859, 2004.

[7] Michelle Scalley-Kim and David Baker. Characterization of the folding energy landscapes of computer generated proteins suggests high folding free energy barriers and cooperativity may be consequences of natural selection. *J. Mol. Biol.*, 338:573–583, 2004.

[8] K. F. Lau and K. A. Dill. A lattice statistical model for the conformational and sequence spaces of proteins. *Macromolecules*, 22:3986–3997, 1989.

[9] Hue Sun Chan and Ken A. Dill. Transition states and folding dynamics of proteins and heteropolymers. *J. Chem. Phys.*, 100(12):9238–9257, 1994.

[10] Marek Cieplak, Malte Henkel, Jan Karbowski, and Jayanth R. Banavar. Master equation approach to protein folding and kinetic traps. *Phys. Review Letters*, 80:3654–3657, 1998.

[11] S. B. Ozkan, K Dill, and I. Bahar. Fast folding protein kinetics, hidden intermediates, and the sequential stabilization model. *Protein Sci.*, 11:1958–1970, 2002.

[12] M. Cieplak, M. Henkel, and J. R. Banavar. Master equation approach to protein folding. *Condensed Matter Physics*, 2:369–378, 1999.

[13] Roy J. Glauber. Time-dependent statistics of the ising model. *Journal of Math. Phys.*, 4:294–307, 1963.

[14] A. Šali, E. I. Shakhnovich, and M. Karplus. How does a protein fold? *Nature*, 369:248–251, 1994.

[15] D. K. Klimov and D. Thirumalai. Criterion that determines the foldability of proteins. *Phys. Rev. Lett.*, 76:4070–4073, 1995.

[16] R. Melin, H. Li, N. Wingreen, and C. Tang. Designability, thermodynamic stability, and dynamics in protein folding: a lattice model study. *J. Chem. Phys.*, 110:1252–1262, 1999.

[17] Roger B. Sidje. Expokit: a software package for computing matrix exponentials. *ACM Trans. Math. Softw.*, 24(1):130–156, 1998.

[18] *Numerical linear algebra and applications*. Brooks/Cole Publishing Company, 1995.

[19] Iksoo Chang, Marek Cieplak, Jayanth R. Banavar1, and Amos Maritan. What can one learn from experiments about the elusive transition state? *Protein Science*, 13:2446–2457, 2004.