From laws of inference to protein folding dynamics

Chih-Yuan Tseng*
Graduate Institute of Systems Biology and Bioinformatics, National Central University, 320 Chungli, Taiwan

Chun-Ping Yu†
Department of Physics, National Central University, 320 Chungli, Taiwan

H. C. Lee
Graduate Institute of Systems Biology and Bioinformatics and Department of Physics, National Central University, 320 Chungli, Taiwan

(Received 4 November 2009; revised manuscript received 12 May 2010; published 18 August 2010)

Protein folding dynamics is one of major issues constantly investigated in the study of protein functions. The molecular dynamic (MD) simulation with the replica exchange method (REM) is a common theoretical approach considered. Yet a trade-off in applying the REM is that the dynamics toward the native configuration in the simulations seems lost. In this work, we show that given REM-MD simulation results, protein folding dynamics can be directly derived from laws of inference. The applicability of the resulting approach, the entropic folding dynamics, is illustrated by investigating a well-studied Trp-cage peptide. Our results are qualitatively comparable with those from other studies. The current studies suggest that the incorporation of laws of inference and physics brings in a comprehensive perspective on exploring the protein folding dynamics.

DOI: 10.1103/PhysRevE.82.021914 PACS number(s): 87.15.Cc, 87.15.hm, 87.10.Ca, 87.10.Tf

I. INTRODUCTION

Protein folding dynamics is one of major issues constantly investigated in the study of protein functions. Because the protein folding process involves complicated many-body interactions, MD simulation is a common theoretical approach considered. However, one issue hinders the practical usage of MD simulation in studying protein folding processes. As it is recognized from energy landscape theory, protein folding is a series of processes that starts with many possible states and goes through a rough potential energy surface created by many-body interactions [1]. It then ends with a few possible states associated with native structures. However, proteins may be trapped in one of local energy minima on the energy surface during simulations. To resolve this issue, the replica exchange method (REM) has been proposed [2]. However, the introduction of the Monte Carlo aspect in REM seems to lose dynamical information of the folding process. Juraszek and Bolhuis' recent studies suggest that the dynamics is not lost and is merely hidden beneath the sampling space [3]. To reveal the dynamics, they propose to integrate appropriate sampling techniques such as transition pathway sampling (TPS) [4–6] in MD simulation. By studying Trp-cage peptide folding dynamics, they found two folding trajectories in their simulations and were found to be consistent with the experimental results [3].

In this work, we tackle the folding dynamics problem differently by asking "Can we reveal folding dynamics from pure REM-MD simulation results directly? And if so, how?" Because protein folding primarily associates slow processes such as the backbone movement compare to fast atomic motions, the approach hinges on the idea of developing a dynamical law that specifically takes information relevant to slow folding processes into account. Because the common procedure to develop such physical laws is normally started with the establishment of a mathematical formalism, upon which one then tries to append an interpretation, it is difficult to develop a dynamical law of many bodies, which only takes specific information such as many body interactions into account, based on the procedure.

However, a reverse procedure, in which one constructs a physical theory by first deciding what the subject is and what one wants to accomplish, and then designing an appropriate mathematical formalism, provides a solution. Because our goal is to study dynamics of many-body systems by processing the corresponding dynamical information directly, the appropriate formalisms are found to be laws of inference, consistency, objectivity, universality, and honesty. They are sufficiently constraining that they lead to a unique set of rules for processing information: rules of probability theory and the method of maximum entropy (ME) [7,8]. Furthermore, Caticha argues that information geometry is a convenient tool to proceed. An information manifold is constructed based on independent parameters that characterize the system. The probability distributions of the system at specific states are treated as points in the manifold. The evolution of probability distributions then is simply represented by that a point object "moves" in the manifold. Caticha shows that the dynamics of a physical system can be derived directly from laws of inference [7,9,10]. He therefore termed this approach the entropic dynamics. It should be noted that information geometry was originally proposed as a method of applying differential geometry to study statistical estimation (please refer to [11] for details). It has been successfully applied to

*Corresponding author; Department of Oncology, University of Alberta, Edmonton, AB T6G 1Z2, Canada; FAX: 1-780-6434380; chih-yuan.tseng@ualberta.ca

†Present address: Graduate Institute of Systems Biology and Bioinformatics, National Central University, Chungli, Taiwan 320

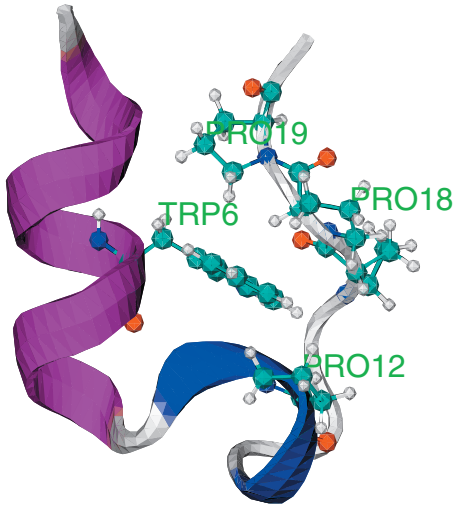

FIG. 1. (Color) The NMR structure of Trp-cage (PDB code is 1L2Y). The structure shows Trp^6 is caged inside the hydrophobic pocket formed by three proline.

different disciplines such as fluctuation theory in statistical physics [12], phase transitions [13], model selection [14], and neuroscience [15].

Following Caticha's studies [7,9,10], we argue that protein folding dynamics also can be derived from laws of inference given REM-MD simulation results and propose entropic folding dynamics. A well-studied Trp-cage peptide, the native structure is shown in Fig. 1 generated by the software, visual molecular dynamics (VMD) [16], is then studied to illustrate its applicability.

II. METHODS

A. Entropic dynamics: From laws of inference to the dynamical laws of physics

In entropic dynamics, dynamical laws follow from recognizing the appearance of changes from one point to a neighboring point in the manifold. According to the ME principle, the preferred neighboring point is the one at the state of maximum entropy.

Consider the microstates of a physical system that are defined by parameters x, and let $m(x)dx$ be the number of microstates within dx. Furthermore, consider that a macrostate of the system is defined by the expectation values A^α of some n_A variables $\{a^\alpha(x);\alpha=1,2,\cdots,n_A\}$, $\langle a^\alpha(x)\rangle = \int dx p(x) a^\alpha(x) = A^\alpha$, where $p(x)$ is the probability distribution function (PDF) of the system at microstate x. Given this constraint equation, the method of ME indicates that the probability distribution of the system at x updated from some prior probability distribution $m(x)$ is given by $p(x|A^\alpha) = \frac{1}{Z} m(x) e^{-\lambda_\alpha a^\alpha(x)}$, where the partition function and Lagrangian multipliers are given by $Z = \int dx m(x) e^{-\lambda_\alpha a^\alpha(x)}$ and $-\frac{\partial \log Z}{\partial \lambda_\alpha} = A^\alpha$.

According to information geometry, a convenient way of distinguishing two states A^α and $A^\alpha + dA^\alpha$ is to treat each as a point in the space of states, the information manifold with coordinates A^α. One can then show that the difference between the two states is given by the distance dl between $p(x|A^\alpha)$ and $p(x|A^\alpha + dA^\alpha)$ by

$$dl^2 = g_{\alpha\beta} dA^\alpha dA^\beta, \tag{1}$$

where a general expression of $g_{\alpha\beta}$ is given by

$$g_{\alpha\beta} = \int dx p(x|A^\alpha) \frac{\partial \log p(x|A^\alpha)}{\partial A^\alpha} \frac{\partial \log p(x|A^\beta)}{\partial A^\beta}, \tag{2}$$

which is the Fisher-Rao metric [17–19], which is the only Riemannian metric that adequately reflects the underlying statistical nature of the manifold of distributions $p(x|A^\alpha)$ Namely, This result indicates that when the probability $p(x|A^\alpha)$ is assigned to each point A^α, it automatically provides the space of states with a metric structure. Note that the coordinates of the manifold need not be the expected values. Here, because the coordinates are chosen as the expectation values A^α, one can show that an alternative expression for the Fisher-Rao metric is

$$g_{\alpha\beta} = -\frac{\partial^2 S(A)}{\partial A^\alpha \partial A^\beta}. \tag{3}$$

Having determined the metric structure, one can tackle the question of how the system evolves from one state to a nearby one by a small amount dl. Because there are many states that lie on the surface of n_A dimensional sphere of radius dl centered at A^α, Caticha shows the preferred one is given by the method of ME, which moves along the entropy gradient with a changing rate,

$$\frac{dA^\alpha}{dl} = \dot{A}^\alpha = \frac{1}{\sigma} g^{\alpha\beta} \frac{\partial S(A)}{\partial A^\beta} = \frac{1}{\sigma} g^{\alpha\beta} \lambda_\beta, \tag{4}$$

where $\lambda_\beta = \frac{\partial S[A^\beta + dA^\beta]}{\partial A^\beta}$, $g^{\alpha\beta}$ is the inverse of $g_{\alpha\beta}$, and one also can show entropy gradient along the trajectory is given by $\sigma = dS[A^\alpha + dA^\alpha]/dl = (\lambda_\alpha \lambda^\alpha)^{1/2}$ and $\lambda^\alpha = g^{\alpha\beta}\lambda_\beta$. The gradient vector λ_β refers to a direction in which there is a maximum increase *per unit distance*. Note that one cannot talk about the gradient vector without introducing a metric at first place. Eq. (4) shows that the system evolves according to its own clock, the intrinsic time $d\tau$. Caticha argues that one convenient choice of intrinsic time is the distance of the space of macrostates, i.e., dl. However, the absolute speed dl/dt, the ratio of the intrinsic time dl and external time of real world dt, remains unknown. Entropic dynamics may be a reasonable theory but it is not yet "physical" because it does not take all dynamical information such as the absolute speed into account [7,10]. When a conformal factor that codifies the information of the motions and the interactions of particles and defines the absolute speed is introduced, Caticha and Cafaro [10] recently show that the equation of motions of particles can then be derived purely from entropic dynamics.

B. Ordering points to identify the clustering structure

As it is recognized that two issues always hinder the interpretation of protein simulation results. The first issue is that when there are no crystal protein structures available as

the references, it becomes obscure to evaluate the structural results from simulations. The second issue is how one determines the native structure or a structure close to native state in REM-MD simulations. The cluster analysis is commonly considered to provide partial solutions for these two issues. To cluster samples without a reference, we consider a density-based clustering method, ordering points to identify the clustering structure (OPTICS) [20]. It applies five criteria: (1) The neighborhood within an empirically defined radius ε of a given protein is called the ε-neighborhood of the protein (here we set $\varepsilon=4$ Å); (2) If the ε neighborhood of a protein contains at least an empirically defined minimum number of proteins, MinPts (here we set the number to 7 proteins), then the protein is called a core protein; (3) Given a set of proteins D, a protein p is directly density reachable from a core protein q, if p is within the ε neighborhood of q; (4) A protein p is density reachable from protein q with respect to ε and MinPts in D set, if there is a chain of proteins $p_1,\ldots p_n$, where $p_1=q$ and $p_n=p$ such that p_{i+1} is directly reachable from p_i for $1\leq i\leq n$; (5) A protein p is density connected to protein q in D set, if there is a protein $o \in D$ such that both p and q are density reachable from o. Note that for our studies here, the radius ε is defined by the root-mean-square deviation (RMSD) of any two protein structures in the sampling space. The density-based clustering regards a cluster as an union of ε neighborhood of core proteins and each two proteins within the cluster are density connected.

OPTICS further defines two quantities, the core distance r_c and reachability distance $\hat{r}$ to indicate density distributions within a cluster and to reveal similarities of the proteins. First, the core distance of a core protein p is defined as the shortest radius ε, the RMSD of any two protein structures within the ε neighborhood of protein p. The short core distance of a core protein indicates a high dense ε neighborhood. We then can measure the difference between two core proteins q and p by a quantity, reachability distance, which is defined by the greater value of the core distance of p and the Euclidean distance between p and q. Based on these definitions, OPTICS first ranks the sets of proteins in the order of density or reachability distance $\hat{r}$. Furthermore, the decreased similarity of proteins within the set is also ranked in the increased order of $\hat{r}$. Therefore, the decreased similarity of protein structures is ranked according to the increased order of reachability distance $\hat{r}$. Note that more explanations on the practical implementation of OPTICS in protein studies will be given in the section of Results.

C. Potential energy that associates with folding process

Because we are interested in the structural changes of Trp-cage along the time line rather than the equilibrium states, potential energy provides sufficient information to address it. However, the fast stochastic atomic motions, which have minor effects on folding processes, always results in large energy fluctuations. These may veil the energy distribution that primarily associates with the slow folding process. We propose a two-step approach to extract the energy components that are associated with the slow folding process

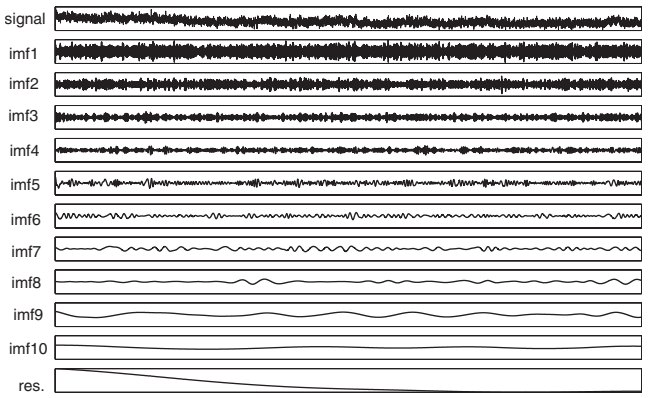

FIG. 2. An example of applying EMD to decompose the potential energy profile of a 40 ns long Trp-cage MD simulation. The first row is the raw potential energy profile. The last row "res" is the residual component after the decomposition.

from the potential energy profile of Trp-cage obtained from REM-MD simulations. The method first decomposes the potential energy profile into several components by an adaptive signal separation method, empirical mode decomposition (EMD) [21]. The EMD is specifically designed to decompose a nonlinear and nonstationary signal into several components, intrinsic mode functions (IMF), with different instantaneous frequencies. An IMF satisfies two conditions: (1) that the number of extrema and the number of zero crossings must either equal or differ at most by one in the whole data set; and (2) that the mean value of the envelope defined by the local maximum and the envelope defined by the local minima at any point is zero [21]. Because IMFs are adaptive and locally determined, they have physical representations of underlying processes [21,22]. In addition, IMFs form an orthogonal set, and they can be used as the basis to represent the data.

We then select the components that primarily associate with the slow folding process according to the selection criterion determined based on the *in silico* studies on dynamics of Trp-cage folding conducted by Hu *et al.* [23]. Their MD simulations show that Trp-cage collapses into a partially organized globular state within a very short time. Particularly, it takes around 0.8 ns to form the α helix. One of protein folding theories suggests before proteins fold into tertiary structures, it will first form secondary structures including α helix or β sheet [24]. Therefore, we consider the time required to form an α helix as the minimum criterion in Trp-cage to identify all folding processes that requires the same or longer time. Namely, the IMFs with mean periods equal to or longer than 0.8 ns are likely the components of the folding processes such as the formation of the alpha-helix and the movement of backbone. Practically, because the EMD cannot exactly separate the IMFs with the mean period 0.8 ns from the energy profile, we set the criterion as 0.5 ns to include a margin of uncertainty.

As an example, we consider an EMD analysis of the potential energy of a 40 ns long Trp-cage MD simulation at room temperature with the same initial structure and simulation settings as introduced previously. The first row of Fig. 2 shows the original potential energy. The rest of the rows

shows the IMF 1 to 10 with a mean period of 0.015, 0.031, 0.059. 0.107, 0.217, 0.406, 0.837, 1.605, 3.043, and 10.71 ns respectively obtained from the EMD. The last row, labeled "res," represents the residual trend of the energy profile. Note that because the sampling rate for outputting energy profile is set to 5 ps, the energy profile can only present the motions with relaxation time longer than 5 ps. Thus, the superposition of IMF 7 to 10 and "res" will be considered to primarily associate with the slow folding process, and be defined as the smoothed potential energy $\bar{u}$.

D. Target protein

Trp-cage is a twenty amino acids long minipeptide (NLYIQWLKDGGPSSGRPPPS) designed by Neidigh *et al.* [25] as a target for investigating protein folding problems. This peptide folds spontaneously and cooperatively from a random coil into this native structure in about 4 μs [26]. Qiu *et al.* further experimentally showed that the folding process is a two-state folding [26]. The native structure of Trp-cage contains three key secondary structures: a short α-helix in residues 2–9; a 3_{10}-helix in residues 11–14 and a C-terminal polyproline II helix after residue 14. The hydrophobic residue Trp^{6} is caged inside the hydrophobic pocket formed by polyproline II helix and Pro^{12} as marked by the green labels in Fig. 1.

E. REM-MD simulations

The AMBER 8 package is utilized for MD simulations [27], which are performed on our 24-node cluster. We use the AMBER 2003 force field and the generalized Born model for mimicking the effects of solvents. The minimum time step is 2 fs. The initial structure of Trp-cage is set to be an extended state. The REM is applied using the multisander REM module of AMBER 8. Twenty four replicas are simulated over a range of temperatures from 276 to 507 K including 276, 283, 290, 298, 305, 313, 321, 329, 337, 346, 355, 364, 373, 382, 392, 402, 414, 426, 439, 451, 465, 478, 492, and 507 K. The REM attempts to swap replicas every 4.0 ps. The protein structures are recorded every 2.0 ps. The accumulated simulation time for each replica is 140 ns and the total CPU time is around 49 days. Totally, seventy thousand protein structures are recorded, which forms the Trp-cage sampling space to be used in this work.

III. RESULTS

A. Two-dimensional information manifold for Trp-cage

The first step of entropic folding dynamics is to construct the information manifold that codifies the Trp-cage structural and dynamical information. Two macroscopic quantities that characterize protein structures and the slow folding process are considered as two coordinates.

For the first quantity, instead of directly utilizing RMSD to characterize structures in the Trp-cage sampling space, we utilize a density-based clustering analysis method, OPTICS [20]. It quantifies structural differences through ranking proteins in the order of similarity defined by reachability distance without any reference structures. The seventy thousand

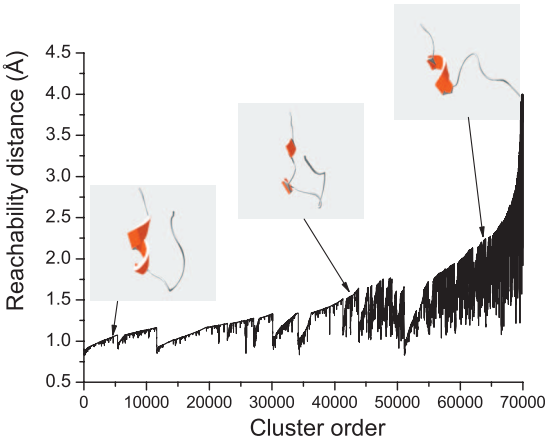

FIG. 3. (Color) The ranking scheme of seventy thousand Trp-cage structures obtained from OPTICS. The definition of reachability distance is given in the text. To illustrate the differences of Trp-cage structures at different ranks, three cartoon representations generated from Swiss-DeepView [28] are shown.

Trp-cage structures then are clustered according to the cluster order and the first structure is chosen to has the lowest core distance as shown in Fig. 3. Hereafter, the same cluster order will be used for energy and probability analysis. In Fig. 3, we also show three examples of Trp-cage's cartoon representations generated by SWISS-DeepView [28] at cluster order 5000, 45 000, and 65 000 to illustrate the structural differences of Trp-cage in the sampling space.

For the second quantity, as shown in the gray line of Fig. 4, the potential energy of our REM-MD simulation results for Trp-cage fluctuates from −300 to −360 kcal/mol. The plot includes seventy thousand structures, ranked according to the reachability distance proposed above. The large fluctuations raise the difficulty to reveal the slow folding process. Although the energy profile of seventy thousand Trp-cage structures is a function of the cluster order rather than the time, we argue in the followings that the proposed smoothing approach still can be applied to extract the energy component associated with the slow folding process. Be-

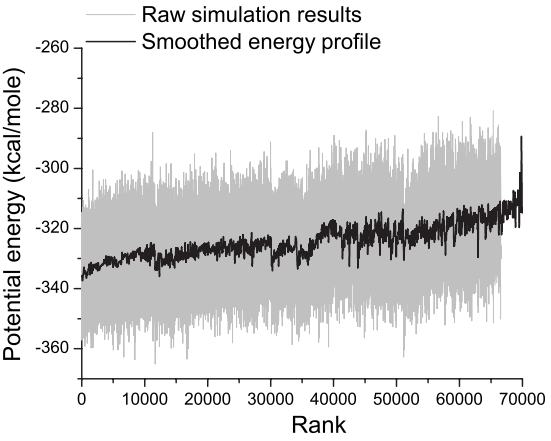

FIG. 4. The potential energy profile of seventy thousand Trp-cage structures, which is ranked according to the ranking scheme obtained from OPTICS (the gray line). The dark line represents the smoothed potential energy obtained from the EMD-based method.

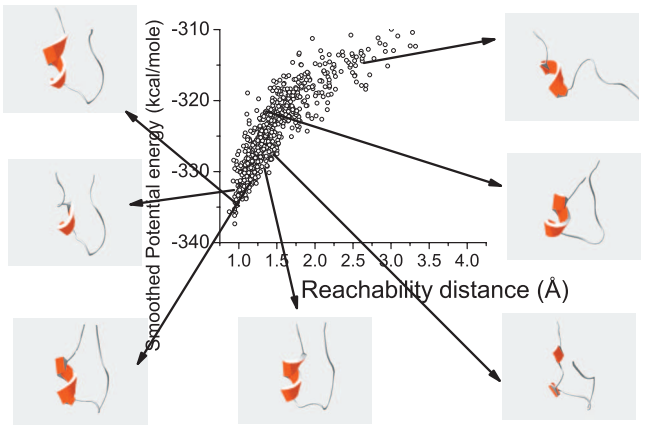

FIG. 5. (Color) Seventy thousand Trp-cage structures are plotted in reachability distance vs smoothed potential energy diagram. Seven cartoon representations of Trp-cage structures generated from Swiss-DeepView [28] at different locations in the diagram are presented to show the structural differences at various locations of the diagram.

cause each structure in the cluster ordering scheme represents a state of Trp-cage at a specific time in one of the REM-MD simulations, the cluster order can be treated as an implicit time dependent parameter. Furthermore, the total accumulated simulation time to obtain seventy thousand structures is 140 ns. In addition, because the ranking scheme ranks the Trp-cage structures from a random state, cluster order 70 000, to a folded state, cluster order 1, we may analog this scheme to the result from performing a 140 ns long MD simulation with the sampling interval 2 ps. Consequently, the frequency spectrum of this energy profile can still reflect the atomic motions during the folding process. Therefore, the same selection criterion proposed above is still applicable. The smoothed potential energy $\bar{u}$ based on the proposed smoothing method is then defined by the superposition of components of potential energy with mean periods longer than 0.5 ns as represented by the dark line in Fig. 4.

Given the reachability distance $\hat{r}$ and smoothed potential energy $\bar{u}$, we plot the sampling space of seventy thousands structures in $(\bar{u},\hat{r})$ space as shown in Fig. 5 to present the range of the sampling space generated from REM-MD simulations and to demonstrate the structural differences at different locations in that space. Each dot labeled by $\{b_i^\alpha;\ \alpha=1$ denotes for $\hat{r}_i$ and $\alpha=2$ for $\bar{u}_i\}$ where subscript i denotes the rank as well as the ith microstate of this Trp-cage statistical system represents a Trp-cage structure at a specific time and temperature in the simulations. We also present seven Trp-cage's cartoon representations at different locations from large to small $\bar{u}$ and $\hat{r}$ to show the structural differences.

Next suppose all possible Trp-cage folding trajectories can be described by $(\bar{u},\hat{r})$, we can treat a transition state along the folding trajectories as an average structure of all possible structures in the phase space with specific probabilities, $P(i)$. Therefore, we define two macrostates,

$$B^\alpha = \sum_{i=1}^{N} P(i) b_i^\alpha, \tag{5}$$

where N is the total number of protein structures, B^1 denotes the expectation reachability distance and B^2 is the expectation smoothed potential energy as the two coordinates for the information manifold. Note that ideally, all possible structures in the folding process are expected to be equally generated from the REM-MD simulations. When there are no constraints provided, the most honest choice is to assign an equal probability to each structure as *a priori*. Furthermore, even though one of the constraints, smoothed potential energy, refers to quasistatic equilibrium state, Jaynes proved that the method of maximum entropy is still applicable [29]. Therefore, given a uniform prior probability and the constraint equation, Eq. (5), ME PDF $P(i)$ is given by

$$P(i|B) = \frac{1}{Z}\exp(-\gamma_\alpha b_i^\alpha), \tag{6}$$

where the partition function is $Z=\Sigma_{i=1}^{N}\exp(-\gamma_\alpha b_i^\alpha)$. Note that in tensor notation, $\gamma_\alpha b_i^\alpha = \gamma_1 \hat{r}_i + \gamma_2 \bar{u}_i$, is used. The Lagrangian multipliers γ_α are numerically determined by using MATLAB scripts [30]. Furthermore, the entropy of the Trp-cage statistical system at state B^α is given by

$$S(B) = -\sum_{i=1}^{N} P(i|B)\log P(i|B) = \log Z + \sum_{n=1}^{2} \gamma_n B_n. \tag{7}$$

Finally, according to information geometry, when a probability distribution is defined, a metric space for the information manifold is created naturally. Therefore, the action of choosing the expectation reachability-distance and smoothed potential energy as the coordinate system defines the metric tensor, Eq. (3). Practically, the component of the metric tensor, function of B, is calculated through

$$g_{nm}(B) = -\frac{\partial^2 S(B)}{\partial B_n\, \partial B_m} = -\frac{\partial \gamma(B)}{\partial B}, \tag{8}$$

where either n or m equals to 1 or 2. By applying the chain rule, $\frac{\partial \gamma(B)}{\partial B}\frac{\partial B}{\partial \gamma(B)}=1$ in Eq. (8), we have

$$g_{nm}(B) = -\left[\frac{\partial B}{\partial \gamma(B)}\right]^{-1} = -(B_n B_m - \langle b_n b_m\rangle)^{-1}. \tag{9}$$

The entropy gradient is given by $\lambda_\alpha = \frac{\partial S(B)}{\partial B^\alpha} = \gamma_\alpha$.

B. Folding dynamical equation of Trp-cage

After the metric space is determined, the evolution of the Trp-cage from a given initial macrostate to the final macrostate in the information manifold is determined by the ME principle. The Trp-cage statistical system will evolve from the jth macrostate $B^\alpha(j)$ to neighboring state $B^\alpha(j+1)$ via

$$B^\alpha(j+1) = B^\alpha(j) + dB^\alpha(j). \tag{10}$$

The change, $dB^\alpha(j)$ calculated from Eq. (4),

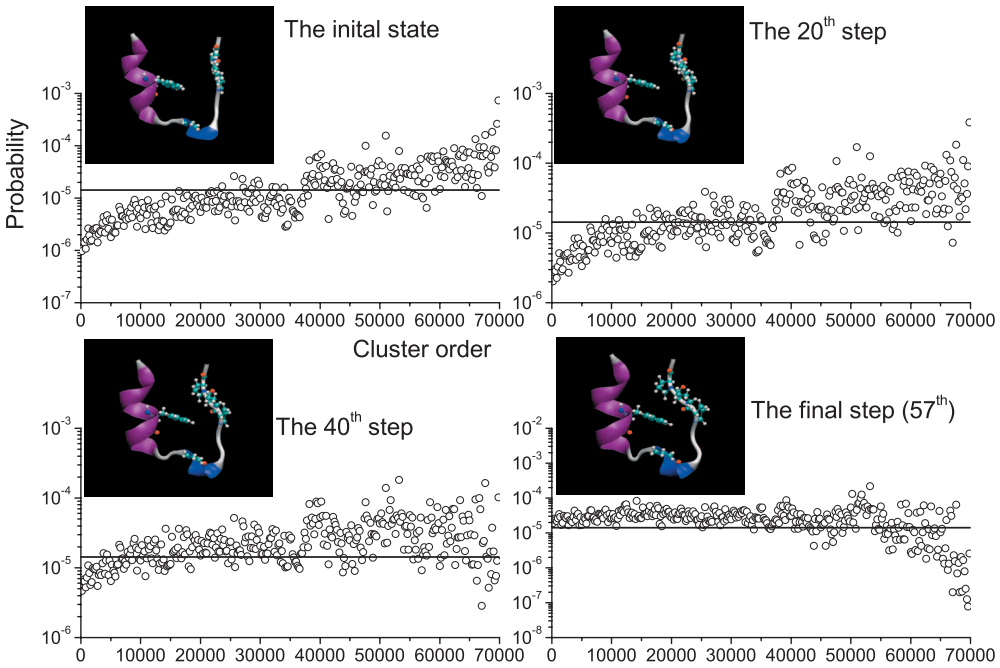

FIG. 6. (Color) The evolution of ME PDF, Eq. (6), given an initial state $B^1=2.2$ Å and $B^2=-312$ kcal/mole. The plot only shows the results at the initial, 20th, 40th, and the final step along with the corresponding averaged Trp-cage cartoon representations generated by VMD [16]. The figure shows that the probabilities at lower ranks are gradually increased while at higher ranks are gradually decreased along the trajectory.

$$dB^{\alpha}(j) = \frac{g^{\alpha\beta}(j)\gamma_{\beta}(j)}{[\gamma_{\alpha}(j)\gamma^{\alpha}(j)]^{1/2}}dl, \tag{11}$$

where $\gamma^{\alpha}(i)=g^{\alpha\beta}(i)\gamma_{\beta}(i)$ and $g^{\alpha\beta}(j)g_{\alpha\beta}(j)=1$. However, Eq. (10) is not yet a physical law without introducing dynamical information to constrain dl, the minimum distance between two macrostates.

To constrain dl, as is recognized in many studies and is also demonstrated in Figs. 3 and 4, proteins tends to fold in a direction that decreases the reachability distance and potential energy globally. We thus constrain the "direction" of changes with regard to the potential energy and reachability distance $dB^{\alpha}(j)$. We set a negative absolute speed dl/dt when the rate of the ith macrostate change $dB^{\alpha}(j)/dl$ is positive. On the other hand, if $dB^{\alpha}(j)/dl<0$, this indicates that the protein evolves in the right direction. The absolute speed will then be set to positive to maintain the direction of changes. According to Eq. (4), the magnitude of absolute speed influences the resolution of two neighboring states only and not the folding trajectories. Furthermore, when the change of potential energy $dB^2(j)$ approaches a threshold (such as $dB^2=10^{-6}$ kcal/mole in our studies), Eq. (10) is considered to be converged. The system then reaches the maximum entropy state. The corresponding average protein structure will be the preferred final structure. Note that this final structure needs not be the native structure of Trp cage. It only represents the maximum entropy state of the Trp-cage sampling space. When more samples and more dynamical information are included, one can expect the preferred structure to coincide with the native structure.

C. Evolution of ME PDF

We further investigate the evolution of the probability distribution $P(i|B^{\alpha})$, Eq. (6), along entropic folding trajectories when an initial state $B^1=2.2$ Å and $B^2=-312$ kcal/mole is considered. Furthermore, we simply set $dl/dt=0.01$ and $dt=1$ unit time. Note that we only use 35 000 structures sampled from the ranked raw data with the rate 0.5 cluster order^{-1} in the following studies. The probability distribution and the corresponding cartoon representation of the average structure of this initial state are shown in left upper panel of Fig. 6. This structure contains a loose helix and a bend around residue 11–19, and is structureless after residue 19. Furthermore, Trp^6 slightly points outward from the paper. This structure qualitatively agrees with Ahmed *et al.*'s experimental studies, in which they show the appearance of an α-helical structure in the denatured state [31]. The system evolves from this initial state toward the maximum entropy state with 57 steps. We select three steps in the trajectory as examples to show the evolution of the PDF along the trajectory in Fig. 6, in which the corresponding average structures of Trp-cage generated by VMD [16] are also presented. The figure shows that the probabilities of the system at the first 10 000 microstates, structures, gradually increase from below 10^{-5} to around it while the probabilities of the system at rank from 60 000 to 70 000 decrease from above 10^{-4} to less than 10^{-5} in the final step. Note that if the 70 000 structures are equally likely, the probability of observing one structure is $1/70\,000 \approx 1.43\times10^{-5}$. Furthermore, we mark this probability by a horizontal line to emphasize to what extent the probability of each microstate at specific step along the trajectory differs from the uniform probability. Figure 6 shows

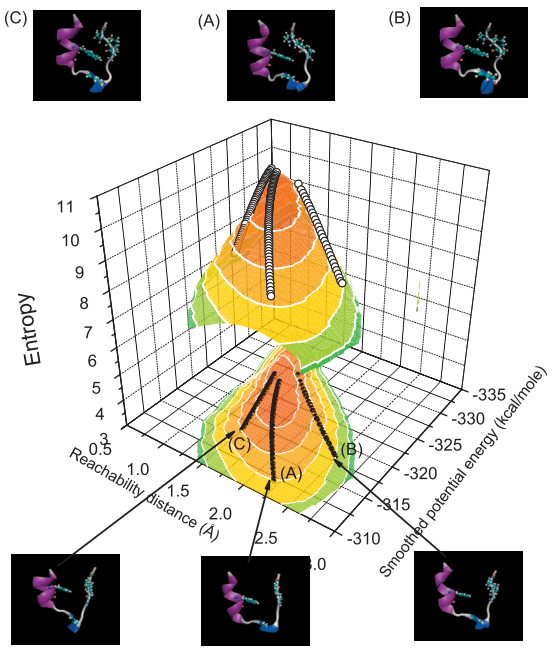

FIG. 7. (Color) Entropic folding trajectories of Trp-cage given three various initial starting structures. The three trajectories are plotted on the three-dimensional entropy surface. In the bottom, the same trajectories are projected on the two dimensional entropy contour map. The averaged initial (bottom) and final (top) structures generated by VMD [16] are also presented for comparisons.

the first 30 000 microstates in the final step, which are shown to have reachability distance around 1 Å in Fig. 3, likely to have similar probabilities. However, the probabilities of having the microstates after the cluster order 30 000 start decreasing and fluctuating. It indicates the maximum entropy state, final average structure, is primarily contributed from the first 30 000 structures. This final average structure presents several key features of the native structure of Trp-cage including a solid alpha helix, 310 helix and polyproline II helix. This evolution corresponds to the formation process of the solid α helix, 3_{10} helix and polyproline II helix. Furthermore, it shows that the Trp^6 is gradually rotated and buried in the hydrophobic proline pocket at almost the same time.

D. Two features of Trp-cage entropic folding trajectories

Finally, we investigate the properties of the folding trajectories when the initial states are given differently with $dl/dt=0.01$. The three initial states, including the one used previously, are (A) $B^1=2.2$ Å; $B^2=-312$ kcal/mole, (B) $B^1=2.5$ Å; $B^2=-318$ kcal/mole, and (C) $B^1=1.5$ Å; $B^2=-316$ kcal/mole. The initial state (B) has the lowest smoothed potential energy, and yet has the largest reachability distance. In contrast, the initial state (C) has the shortest reachability distance. Both initial states are in the vicinity of the upper and lower boundaries of the sampling space shown in Fig. 5. Figure 7 shows how Trp-cage evolves toward the maximum entropy state on the three-dimensional entropy surface through three trajectories. It also shows the steepness of this entropy surface. Furthermore, we project the same trajectories on two dimensional entropy contour map in the bottom to show the differences of the trajectories in terms of smoothed potential energy and reachability distance. Six cartoon representations of averaged Trp-cage structure at initial (bottom) and final (top) steps are presented in the same figure. The entropy of the system is calculated from Eq. (7) for each state B^α and its magnitudes are denoted by the color scale. Note that the entropy contour map is plotted to roughly cover the region of the sampling space shown in Fig. 5. It takes 29 and 48 steps for state (B) and (C) to reach the maximum entropy state, respectively.

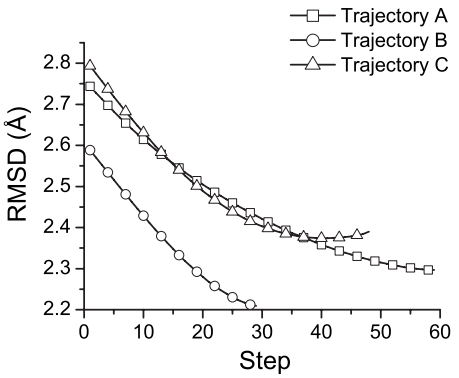

FIG. 8. The RMSD of averaged structures along the three trajectories and the NMR structure of Trp-cage. The lines with hollow square, circle and triangle are from trajectories (a), (b), and (c), respectively. Note that we only label the data points by symbols at every 4 steps.

Furthermore, the projection of trajectories onto reachability distance vs potential energy surface in Fig. 7 shows two features of folding trajectories. The first feature, revealed by either trajectory (B) or (C), is that the system likely evolves through a straight route. There is a linear relation between smoothed potential energy and reachability distance. The second feature, revealed by trajectory (A), shows a curve type route. This type of route first shows the system evolves with the reachability distance having a decreasing rate faster than the smoothed potential energy does. After around $B^1=1.75$ Å and $B^2=-316$ kcal/mole, the system switches to evolve with the decreasing rate for the reachability distance being slower than the decreasing rate for the smoothed potential energy. In general, when the initial state of the system is in the vicinity of the sampling space, the system tends to evolve through a straight route. Otherwise, a curve type route will be expected.

To quantitatively analyze the structural changes along the trajectories, we calculate the RMSD of the C_α atoms of the average structures at each step of the three trajectories and the NMR structure of Trp-cage. The results are shown in Fig. 8. The RMSD of final structures and the NMR structure are 2.17, 2.17, and 2.26 Å, respectively. For both trajectories A and B, the RMSD gradually decrease to around to 2.13 Å after 45 and 20 steps and climb back a little bit to 2.17 Å, respectively. This suggests that when the system evolves to the lower left region in Fig. 7, the corresponding average structure is likely to be equilibrated and approaches to the native structure.

IV. DISCUSSION

The features of folding trajectories discussed previously suggest that the folding process begins with quickly collapsing the protein structure (the reachability distance is decreased faster than the changes of the potential energy). It forms a partially folded and loose α helix and a bend structure around residue 11–14. Afterward, the structure will undergo a fine tuning process to the native structure (the smoothed potential energy is decreased faster than the changes of the reachability distance). In the process of the fine tuning, the helicity of the partially folded α helix is gradually increased. In the meantime, the 3_{10} helix is formed and the Trp^6 is packed within the polyproline hydrophobic pocket. Furthermore, our results also indicate that when the system evolves to the lower left region (reachability distance <1.6 Å and smoothed potential energy <-320 kcal/mol), the averaged structures of Trp-cage from different trajectories are slightly different in the 3_{10} helix only. Although the RMSD of the structure at maximum entropy state and Trp-cage crystal structure in Fig. 8 is still too large to indicate that the maximum entropy state is the native state (normally, two structures are considered similar when their RMSD is less than 1 Å), we cannot simply conclude that the entropic folding trajectory fails to reach the native state. Several factors are likely to cause large RMSD. For example, the numbers of samples used for calculations may not provide sufficient statistics. Furthermore, because the potential energy difference between two neighboring states when the system at that region is insignificant, it also suggests a broad global energy minima in the Trp-cage energy landscape.

This folding process qualitatively agrees with the results of other studies such as Juraszek and Bolhuis [3]. Juraszek and Bolhuis propose a modified REM-MD simulation, in which the transition pathway sampling technique [4–6] is integrated, to solve sampling issue to study the folding dynamics [3]. Their results first show that Trp-cage is a two-step folder, which agrees with the results of a free energy landscape study [32]. In addition, they discover twofolding pathways, LN and IN, where L stands for the loop state, I stands for the intermediate state and N for the native state. In both pathways, Trp-cage starts by undergoing either a fast initial collapse or the formation of the helical structure. Next, the protein will either form a loop structure in state L, or the helical structure in state I. Both states will eventually reach the native state. Their studies further show that the occurrence of the LN pathway is four times more likely than that of the IN pathway. In addition, there are switching events between these two pathways. Despite the results of Juraszek and Bolhuis differ from those from an all-atom Go model presented by Linhananta *et al.* [33], Juraszek and Bolhuis showed both LN and IN pathways agree with experimental results [31,34].

Our studies also present an advantage and one insight with regard to the folding dynamics. The advantage is that there is no need to modify the current REM-MD simulation protocol, such as by integrating it with some sampling techniques. Instead, one can directly apply the proposed approach to the systems that have investigated using either REM-MD or Monte Carlo simulations as long as "right" coordinate systems are defined to construct the information manifold. For the insight, as expected, our results also indicate that the folding dynamics is driven by ME principle. One may dismiss this conclusion by arguing that this is merely a consequence of thermodynamic second law. However, one cannot talk about these folding trajectories without mapping the systems into information manifold at beginning.

V. CONCLUSIONS

In this work, we show that given REM-MD simulation, protein folding dynamics can be directly derived from laws of inference. The crux hinges on appropriately codifying information relevant to the dynamics of many-body systems into an information manifold. There are no restrictions in applying this method to different systems. However, the evolution trajectory may not be the correct pathway, but instead is merely a route preferred over all possible routes based on the information provided. When the system is appropriately characterized and dynamical information, the absolute speed, relevant to it is included, the preferred route is then likely to coincide with the correct one.

To illustrate the proposed approach, we study the folding dynamics of Trp-cage. Two quantities, reachability distance and smoothed potential energy, are used as coordinates of the two dimensional information manifold for Trp-cage. A preferred folding trajectory is then derived and found to qualitatively comparable with those from other studies.

Despite the promising results in this work, there are still some fundamental issues that need to be investigated further. For example, we only consider two parameters to construct the information manifold in this work, but it is unknown if these are sufficient for all different proteins. Moreover, is there any other dynamical information that can be included as the constraints? One can expect that a more complete entropic approach that will better tackle the dynamical problems of more complicated biological systems such as binding problems will be available when these issues have been appropriately investigated. Nevertheless, our current studies still suggest that laws of inference are not merely the principles of information processing. In searching the link between physics and information in Wheeler's late works [35], the current studies provide a strong evidence to show that the incorporation of laws of inference and physics brings in a comprehensive perspective on exploring the nature. Furthermore, our works suggest that the introduction of information geometry likely to be an appropriate tool to bridge physics and information.

ACKNOWLEDGMENTS

This work is partially supported by Grant No. NSC 93-3112-B-008-001 from the National Science Council, Taiwan, ROC (to C.-P.Y.). C.Y.T. is gratitude to A. Caticha for many discussions and suggestions.

[1] J. N. Onuchic, N. D. Socci, Z. Luthey-Schulten, and P. G. Wolynes, Folding Des. **1**, 441 (1996).
[2] Y. M. Rhee and V. S. Pande, Biophys. J. **84**, 775 (2003).
[3] J. Juraszek and P. G. Bolhuis, Proc. Natl. Acad. Sci. U.S.A. **103**, 15859 (2006).
[4] C. Dellago, P. G. Bolhuis, F. S. Csajka, and D. Chandler, J. Chem. Phys. **108**, 1964 (1998).
[5] P. G. Bolhuis, D. Chandler, C. Dellago, and P. I. Geissler, Annu. Rev. Phys. Chem. **53**, 291 (2002).
[6] C. Dellago, P. G. Bolhuis, and P. I. Geissler, Adv. Chem. Phys. **123**, 1 (2002).
[7] A. Caticha, in *Maximum Entropy and Bayesian Methods in Science and Engineering*, AIP Conference Proceedings, edited by A. Mohammad-Djafari (AIP, New York, 2001), Vol. 568, p. 72.
[8] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
[9] A. Caticha, in *Maximum Entropy and Bayesian Methods in Science and Engineering*, AIP Conference Proceedings, edited by R. L. Fry (AIP, New York, 2002), Vol. 617, p. 302.
[10] A. Caticha and C. Cafaro, in *Maximum Entropy and Bayesian Methods in Science and Engineering*, AIP Conference Proceedings, edited by K. H. Knuth, A. Caticha, J. L. Center, A. Giffin, and C. C. Rodriguez (AIP, New York, 2007), Vol. 954, p. 165.
[11] S. Amari and H. Nagaoka, *Methods of Information Geometry* (Oxford Press, New York, 2000).
[12] G. Ruppeiner, Rev. Mod. Phys. **67**, 605 (1995).
[13] W. Janke, D. A. Johnston, and R. Kenna, Physica A **336**, 181 (2004).
[14] C. Rodriguez, in *Maximum Entropy and Bayesian Methods in Science and Engineering*, AIP Conference Proceedings, edited by K. H. Knuth, A. E. Abbas, R. D. Morris, and J. P. Castle (AIP, New York, 2005), Vol. 803, p. 80.
[15] H. Nakahara and S. Amari, Neural Comput. **14**, 2269 (2002).
[16] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).
[17] R. A. Fisher, Proc. Cambridge Philos. Soc. **22**, 700 (1925).
[18] C. R. Rao, Bull. Calcutta Math. Soc. **37**, 81 (1945).
[19] S. Amari, *Differential-Geometrical Methods in Statistics* (Springer-Verlag, New York, 1985).
[20] J. Han and M. Kamber, *Data Mining: Concept and Techniques*, 2nd ed. (Morgan Kaufmann, San Francisco, CA, 2006).
[21] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, Proc. R. Soc. London, Ser. A **454**, 903 (1998).
[22] S. Kizhner, T. P. Flatley, N. E. Huang, K. Blank, and E. Conwell, in *2004 IEEE Aerospace Conference Proceedings* (IEEE, New York, 2004), p. 1961.
[23] Z. Hu, Y. Tang, H. Wang, X. Zhang, and M. Lei, Arch. Biochem. Biophys. **475**, 140 (2008).
[24] C. Branden and J. Tooze, *Introduction to Protein Structure*, 2nd ed. (Garland Science, New York, 1999).
[25] J. W. Neidigh, R. M. Fesinmeyer, and H. H. Andersen, Nat. Struct. Biol. **9**, 425 (2002).
[26] L. Qiu, S. A. Pabit, A. E. Roitberg, and S. J. Hagen, J. Am. Chem. Soc. **124**, 12952 (2002).
[27] D. A. Case, T. E. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, Jr., A. Onufriev, S. Simmerling, B. Wang, and R. Woods, J. Comput. Chem. **26**, 1668 (2005).
[28] N. Guex and M. C. Peitsch, Electrophoresis **18**, 2714 (1997).
[29] E. T. Jaynes, in *The Maximum Entropy Formalism*, edited by R. D. Levine and M. Tribus (MIT Press, Cambridge, MA, 1979), p. 15.
[30] A. Mohammad-Djafari, in *Maximum Entropy and Bayesian*, edited by T. W. Grandy (Kluwer, Dordrecht/Academic, New York, 1991), p. 221.
[31] Z. Ahmed, I. A. Beta, A. V. Mikhonin, and S. A. Asher, J. Am. Chem. Soc. **127**, 10943 (2005).
[32] R. Zhou, Proc. Natl. Acad. Sci. U.S.A. **100**, 13280 (2003).
[33] A. Linhananta, J. Boer, and I. Mackay, J. Chem. Phys. **122**, 114901 (2005).
[34] H. Neuweiler, S. Doose, and M. Sauer, Proc. Natl. Acad. Sci. U.S.A. **102**, 16650 (2005).
[35] J. Wheeler, in *Complexity, Entropy and the Physics of Information*, edited by W. H. Zurek (Addison-Wesley, Redwood City, CA, 1990), p. 1.