Folding a Protein with Equal Probability of Being Helix or Hairpin

Chun-Yu Lin,¹ Nan-Yow Chen,‡ and Chung Yu Mou†‡*†
¹Department of Physics, National Tsing Hua University, Hsinchu, Taiwan; ²National Center for High-Performance Computing, Hsinchu, Taiwan; ³Institute of Physics, Academia Sinica, Nankang, Taiwan; and ⁴Physics Division, National Center for Theoretical Sciences, Hsinchu, Taiwan

ABSTRACT We explore the possibility for the native structure of a protein being inherently multiconformational in an ab initio coarse-grained model. Based on the Wang-Landau algorithm, the complete free energy landscape for the designed sequence 2DX4: INYWLAHAKAGYIVHWTA is constructed. It is shown that 2DX4 possesses two nearly degenerate native structures: one is a helix structure with the other a hairpinlike structure, and their free energy difference is ≈2% of that of local minima. Two degenerate native structures are stabilized by an energy barrier of ≈10 kcal/mol. Furthermore, the hydrogen-bond and dipole-dipole interactions are found to be two major competing interactions in transforming one conformation into the other. Our results indicate that two degenerate native structures are stabilized by subtle balance between different interactions in proteins. In particular, for small proteins, balance between the hydrogen-bond and dipole-dipole interactions happens for proteins of sizes being ~18 amino acids and is shown to the main driving mechanism for the occurrence of degeneracy. These results provide important clues to the study of native structures of proteins.

INTRODUCTION

Solving the protein-folding problem has tremendous implications. Among possible applications, the solution to the problem makes it possible to design drugs theoretically, which would result in the greatest impact to the biological science. Nonetheless, despite much effort being devoted during the past, the problem continues to be one of the most basic unsolved problems. To solve the folding problem completely, it is generally believed that to be able to predict the protein structure for a given sequence of amino acids is the key step. Following the classical Anfinsen’s work (1), it is known that the native state of a globular protein would lie at the minimum of the free energy; hence, the problem of structure prediction reduces to the problem of finding the minimum of the free energy.

During the past decades, it has become evident that the free energy landscape for a given segment of amino acids is more complicated than was previously thought and may possess local minima exhibited as metastable states. Such evidence has been often exhibited as the conformation switch of proteins. For instance, the bovine β-lactoglobulin protein is a predominantly β-sheet protein but it has been observed to go through a remarkable α→β transition during the folding process (2,3). In the effort of unraveling the mechanism for protein misfolding and aggregation, which are known to be causes for perplexing diseases such as Alzheimer’s disease and the prion encephalopathies, it is found that even though aggregates found in the patients of Alzheimer’s disease comprise extended β-sheet structures, the building block of the aggregates (the amyloid-β monomer) adopts a random coil structure in aqueous solution (4,5) or predominantly α-helix structure in membrane-mimicking environments (6,7). It is thus rational to postulate that an α→β or a random coil→β transition occurs during the early aggregation process (8).

Typically, a conformation switch of proteins can be induced by changing external conditions such as the pH value, the ionic strength (9,10), the temperature (11), the solvent polarity (12), or by mutating a few amino acids. Kabsch and Sander (13) found a pentapeptide sequence that could adopt an α-helix or a β-sheet conformation in different proteins. Cohen et al. (14) extended this work to hexapeptides. Minor and Kim (15) have conducted an experiment showing that an 11-amino-acid sequence can be transformed into an α-helix or a β-sheet in protein G. Such chameleon-like sequences have their cooperative local interactions competing against long-range interactions of sequence environment. The fragmental propensity of secondary structures is found to be overwhelmed by larger structures. It is also shown that proteins may evolve from one structure to another by mutating single or several amino acids in sequence (16,17). The general assumption behind this is that the key mutation would destabilize the original structure, and favor another propensity.

The above facts indicate that there may exist nearby competing states to the native state of a given protein. Therefore, given appropriate conditions, the native state of a given sequence of amino acids can be changed. To elucidate the real mechanism that causes the conformation change, a de novo protein has recently been designed by Araki and Tamura (18). They reported a modified sequence INYWLAHAKAGYIVHWTA deposited in the Protein Data Bank (19) (PDB IDs 2DX3 and 2DX4; we shall term this simply as 2DX4 hereafter) was identified to have equal populations of α-helical or β-sheet in an aqueous solution.

*Correspondence: mou@phys.nthu.edu.tw
Editor: Michael Levitt.
© 2012 by the Biophysical Society
0006-3495/12/07/0099/10 $2.00

doi: 10.1016/j.bpj.2012.05.029
Although it is well recognized that protein solutions are in equilibrium with intermediate peptides, the dual native structures are rarely reported in the literature. Furthermore, it is shown that the conformational transformation of 2DX4 is not induced by any environmental conditions or binding motifs. These facts make 2DX4 a valuable target to study. In particular, folding 2DX4 would be a crucial test for any viable approach for solving the protein-folding problem.

On the theoretical side, all-atom simulation is the most comprehensive approach for understanding the folding processes; however, the requirement of computational resources tends to be realistically unaffordable (20). Itoh et al. (21) have combined all-atom molecular-dynamics simulation with multicanonical multioverlap algorithm to simulate 2DX4. From the limited phase space obtained, they investigated possible pathways for the $\alpha \rightarrow \beta$ transition. In particular, three local minima in free energy are identified. However, only partial $\alpha$-helices or $\beta$-hairpins are found in the structures associated with these local minima. The mechanism that is responsible for the possibility of two native structures of 2DX4 thus remains unclear. On the other hand, there has been much effort in developing coarse-grained models to predict protein structures (22). In these models, effects of water molecules are implicitly included in effective interacting potentials between amino acids. The required computational resources are much reduced and it enables the prediction of protein structures feasible. Indeed, progress have recently been made in predicting structures of wild-type proteins of sizes from 12 to 56 amino acids by using realistic and unbiased potentials between amino acids (23). To further check the validity of coarse-grained models, folding proteins such as 2DX4 would be an ideal test.

In this work, based on an ab initio coarse-grained model constructed in Chen et al. (23), we constructed the complete free energy landscape for 2DX4. It is shown that in agreement with the experimental observation, there are only two native structures associated with local minima of the free energy: $\alpha$-helix- and $\beta$-hairpin-like structures. Moreover, within the accuracy of the coarse-grained model, it is found that whereas local minima are degenerate in the case of 2DX4, the $\beta$-hairpin-like structure is higher in energy for the DP3 protein that results from the mutation of one amino acid of 2DX4 and was reported to have zero population of hairpin structure (18). In addition, the pathways between the helix and hairpin configurations are simulated by Monte Carlo (MC) algorithm in high temperatures. By analyzing a detailed free-energy profile, we find that the hydrogen-bond and dipole-dipole interactions are two major competing mechanisms in transforming one conformation into the other. Our results indicate that, generally, degenerate native structures are stabilized by subtle balance between different interactions in proteins. For small proteins, the balance between the hydrogen-bond and dipole-dipole interactions can occur for sizes of proteins being ~18 amino acids or 40 amino acids. These results provide important clues to the study of the native structures of proteins.

**THEORY AND METHODS**

**Ab initio coarse-grained potentials**

We shall first recapture essentials of the coarse-grained model constructed in Chen et al. (23). In this model, residues are coarse-grained, as spheres are centered at C$_\alpha$ atoms but complete structures are kept in backbones. Bond angles and bond lengths are fixed between these atoms to increase folding efficiency; the only variables are dihedral angles $\phi$ and $\psi$ on the C$_\alpha$ atom-hinging, two-amide planes. Water molecules are not included explicitly, but their effects are incorporated in effective potentials among side chains and backbones. In these representations and with all energies being in unit of kcal/mol, the total energy can be written as

$$E_{\text{total}} = E_{\text{Steric}} + E_{\text{DD}} + E_{\text{HB}} + E_{\text{MJ}} + E_{\text{NP}} + E_{\text{SA}}.$$  

(1)

Here each energy term is a weighted potential energy with $E_i = \epsilon_i V_i$, where $\epsilon_i$ is the weighting factor to be determined later and $V_i$ is the corresponding potential energy. Among these energy terms, $E_{\text{Steric}}$ is to enforce the structural constraints such as hard-core potentials to avoid unphysical contacts, whereas $E_{\text{SA}}$ is the solvent-accessible surface energy in proportion to the area of each side chain that is exposed to water and is primarily responsible for stabilizing the tertiary structure. The remaining terms are three ingredients for the formation of the secondary structures, with $E_{\text{HB}}$ being the hydrogen-bonding between any nonneighboring NH and CO pair, $E_{\text{DD}}$ being the summation of screened dipole-dipole interaction at large distance (global dipole interaction, $E_{\text{DDC}}$) and local dipole-dipole interaction between dipoles on the backbones, and $E_{\text{MJ}} + E_{\text{NP}}$ accounting for the interactions due to hydrophobicity or the charge state of the amino acids. Except for $E_{\text{MJ}} + E_{\text{NP}}$, all the potentials are based on realistic and bare values of parameters obtained from experimental data. The potential, $E_{\text{MJ}} + E_{\text{NP}}$ was based on simple generalizations of the Miyazawa-Jernigan matrix (24,25) by using a 12-6 Lennard-Jones potential modified by effects due to the sizes of water molecules (23). To include realistic effects due to hydrophobicity or the charge state of the amino acids, we shall construct the corresponding potentials by statistical methods so that $E_{\text{MJ}}$ generalizes the Miyazawa-Jernigan matrix (24,25) to finite large distances between amino acids, whereas $E_{\text{NP}}$ generalizes the $V_{\text{localHP}}$ in Chen et al. (23) and is the statistical energy that characterizes the propensity (to $\alpha$ or $\beta$) of amino acids in nearest neighbors.

With these potentials, the weighting factor values $\epsilon_i$ are calibrated based on a few proteins of known structures (22). Details of calibration are given in the next subsection. Typical values of $\epsilon_i$ are $\epsilon_{\text{DG}} = 0.21$, $\epsilon_{\text{EM}} = 2.0$, $\epsilon_{\text{HB}} = 4.8$, $\epsilon_{\text{SA}} = 1.35$, and $\epsilon_{\text{MJ}} = 0.85$. For helix and sheet propensity energies, we get $\epsilon_{\text{DG}}^\alpha = 6.4$ and $\epsilon_{\text{DG}}^\beta = 16$. These calibrated parameters are then used to fold various target proteins. Note that there are ranges of parameters that allow successful folding of target proteins. In our model, success of folding target proteins requires a strong hydrogen-bonding: the upper bound of a hydrogen bound is 4.8 kcal/mol (the magnitude of the hydrogen bond by taking the vacuum as a reference point); the lower bound of the energy for the hydrogen bond is 3.84 kcal/mol. The lower bound is larger than the value of 3.1 kcal/mol obtained in careful studies of the hydrogen bond (26,27).

For ordered states, because it is the relative strengths between different energy terms that determine the native structures, relative ratios of energy terms are more important. These ratios are fixed by calculating the weighting factor $\epsilon_i$. In the allowed ranges of parameters, the lower bound of the ratio of the hydrogen bond to a typical bonding in $E_{\text{MJ}}$ (taking the interaction between Leu and Leu as an example) is 3.84/1.02 $= 3.77$, which is about the same scale as 3.44(= 3.1/0.9) that was adopted in the literature.
Therefore, even though the absolute magnitude of the hydrogen bond used is strong, the lower bounds of relative strengths of the hydrogen bond to other energy terms are about the same scales adopted in the literature.

To extend the Miyazawa-Jernigan matrix to finite distances, we perform extended statistical analysis by first writing

$$E_{MJ} = \epsilon_{MJ} \sum_{i,j} V_{ij,MJ}(r)(1 - S_A)(1 - S_A),$$

(2)

where \(S_A\) and \(S_A\) are the solvent accessibilities for \(i\)th and \(j\)th residues, respectively. The quantity, \(V_{ij,MJ}(r)\), is the statistical potential between the \(i\)th and \(j\)th residues obtained by counting number \(n_{ij}\) of the corresponding \(i\)-type and \(j\)-type residues separating by \(r\), that appears in the PDB. Fundamentally, \(V_{ij,MJ}(r)\) is the generalization of the pair distribution function \((28)\) and its relation to \(n_0(r)\) is given by the Boltzmann’s statistics

$$\exp(-V_{ij,MJ}(r))$$

$$= \frac{\sum \rho \cdot \prod (n_{i,x,p}(r_k) + n_{i,y,p}(r_k))}{\sum \rho \cdot \prod (n_{i,x,p}(r_k) + n_{i,y,p}(r_k))},$$

(3)

where \(A(r_k)\) is a normalization factor to be determined later, numbers with the index \(p\) denote the corresponding statistical values that belong to one specific protein \(p\), \(S_A\) represents the solvent group, and \(r_k\) is the radius of the \(k\)th spherical shell centered at \(i\)-type residue. Note that different amino acids have a different occurrence frequency in real proteins and this is normalized by the denominator in Eq. 3. Furthermore, homology of sequence bias was eliminated by the sequence alignment method in combination with the weighting matrix used by Miyazawa and Jernigan (25). Here \(2n_{i,x,p}(r_k)\) for \(i \neq j\) and \(n_{i,x,p}(r_k)\) are the counts when the \(i\)-type residue is at the origin and the \(j\)-type residue is in the \(k\)th distance \(r_k\), whereas \(n_{i,x}\) is the total count of the \(i\)th residue

$$n_{i,x}(r_k) = \sum j n_{i,x,p}(r_k).$$

The value \(n_{i,y,p}\) counts events taking place between the \(i\)-type residue and solvent group 0,

$$n_{i,y,p}(r_k) = \frac{1}{2} q_i(r_k) n_{i,x}(r_k) - n_{i,x}(r_k),$$

(5)

where \(q_i\) is the coordinate number of the \(i\)-type residue in the \(k\)th spherical shell and \(n_i\) is the total number of the \(i\)-type residues in protein \(p\). The values \(n_{i,x}\) and \(n_{i,y}\) are summations of \(n_{i,x}\) and \(n_{i,y}\) over \(i\)-type residue, respectively,

$$n_{i,x}(r_k) = \sum i n_{i,x}(r_k),$$

(6)

$$n_{i,y}(r_k) = \sum i n_{i,y}(r_k).$$

(7)

Finally, the normalization factor \(A(r_k)\) is defined by

$$A(r_k) = \frac{\text{total number of shells}}{\frac{4}{3} \pi \left(\frac{r_k + \Delta r}{2}\right)^3 - \left(\frac{r_k - \Delta r}{2}\right)^3},$$

(8)

where \(\Delta r\) is the width of each spherical shell. The effective potential as a continuous function of \(r\), \(V_{ij,EMJ}(r)\) is then interpolated from \(V_{ij,MAJ}(r)\). As a demonstration, in Fig. 1, we show a typical effective potential obtained by the above statistical analysis. We see that similar to the pair-distribution function for liquid molecules \((28)\), \(V_{ij,MIJ}(r)\) exhibits oscillations similar to those of the desolvation model \((29)\). Clearly, an energy barrier exists before two amino acids get closer to the repulsive core. The desolvation-like barrier was not included in many implicit-solvent potentials \((30)\). However, it has been pointed out that the barrier favors the \(\beta\)-sheet rather than the \(\alpha\)-helix and may play significant roles in the formation of secondary structure \((31)\). Note that the origin of the energy barrier shown in Fig. 1 is not purely contributed by solvent molecules, thus it is different from the desolvation mechanism. Nevertheless, the insertion of effective solvent groups \(\text{(defined by Miyazawa and Jernigan \((24)\)) and utilized here)}\) cooperating with residue contacts can represent an effective liquid-phase potential.

In the linear regression analysis of \(V_{ij,MAJ}(r)\), the correlation of the first and the second minimum positions \(r_1\) and \(r_2\) is 0.925 and can be represented by

$$r_2 = 1.21 r_1 + 0.34 (l_{\alpha,i} + l_{\alpha,j}) + a_0.$$

Here \(r_1\) is found to be the summation of averaged radius of residues \(i\) and \(j\), \(r_1 = a_i + a_j\), \(l_{\alpha,i} \) is the maximum excess length of residue \(i\) over the averaged radius \((\text{ranges from 1 to 5})\), and \(a_0 = 1.61\) is the size of an effective solvent molecule. Furthermore, even though there are structures in proteins, there is no indication of any ordering in \(V_{ij,MAJ}(r)\). The effective \(V_{ij,MAJ}(r)\) is only valid for large enough distances. For residues in nearest neighbors, due to the steric constraints, the pair distribution function starts to deviate from the desolvation model. To extend \(EMJ\) to characterize interactions of residues in nearest neighbors, \(ENP\) is introduced to account for the statistical energy between nearest-neighbor residues. The interactions among nearest-neighbor residues are best characterized by dihedral angles \(\phi\) and \(\psi\) of the corresponding amide planes. Because \(V_{ij,MAJ}(r)\) does not cover distances of three successive residues, \(ENP\) needs to characterize
three successive residues in the protein, labeled by $i-1$, $i$, and $i+1$. Using the corresponding dihedral angles shown in Fig. 2 a, $E_{NP}$ can be written as

$$E_{NP} = \sum_i \sum_{k=a,\beta} e_{k,NP} \left[ V_{1m}^{k}(\psi_{i-1}, \psi_{i}) + V_{m}^{k}(\psi_{i}, \psi_{i+1}) \right] \times V_{m}(\phi_{i}, \psi_{i}),$$

(9)

where $l$, $m$, and $n$ are indices for type of residues, $V_{m}$ is a one-body potential that depends on $\psi$, and $\phi_{i}$ of the amide planes connecting to the $m$-type residue, and $V_{1m}$ (also $V_{m}$) is a two-body energy that depends on dihedral angles of $l$-type and $m$-type residues in nearest neighbors. According to the Ramachandran plot, it is known that $\phi$ and $\psi$ are statistically concentrated at particular regions, which are either in the $\alpha$-helix configuration or $\beta$-sheet configuration. To ensure the relative magnitudes of $\alpha$-helix and $\beta$-sheet part are not biased by the database, different weighting factors with $k = a$ and $\beta$ are introduced in Eq. 9. The one-body angular potential $V_{m}$ is obtained by first analyzing the bare potential $v_{m}$ defined by

$$\exp(-v_{m}(\phi, \psi)) = \frac{n_{m}(\phi, \psi)}{\int n_{m}(\phi, \psi) d\phi d\psi},$$

(10)

where $n_{m}$ is the number density taken over the whole PDB for type-$m$ residues with dihedral angles being $(\phi, \psi)$. To account for the preference or nonpreference of $\alpha$- or $\beta$-structures, we set

$$V_{m}(\phi_{i}, \psi_{i}) = \theta(\Lambda - v_{m}(\phi_{i}, \psi_{i})),\,$$

with $\theta$ being the step function and $\Lambda$ being a negative threshold energy level so that $V_{m}$ is either 1 or 0.

The bare two-body potential is constructed by

$$\exp\left(-v_{m}^{k}(\psi_{i-1}, \phi_{i})\right) =\frac{n_{l}(\psi_{i-1}, \phi_{i}) \int n_{m}(\psi_{i-1}, \phi_{i}) d\psi_{i-1} d\phi_{i}}{\int n_{l}(\psi_{i-1}, \phi_{i}) d\psi_{i-1} d\phi_{i} \int n_{m}(\psi_{i-1}, \phi_{i}) d\psi_{i-1} d\phi_{i}},$$

(11)

where $n_{lr}$, $n_{mr}$, and $n_{rr}$ are defined in the same way as those in Eqs. 4 and 6, except that they are specialized to the dihedral angle $(\psi_{i-1}, \phi_{i})$. The expression $V_{m}^{k}(\psi_{i-1}, \phi_{i})$ is then defined by rescaling $u_{m}$ with respect to the average value of $v_{m}$,

$$V_{m}^{k}(\psi_{i-1}, \phi_{i}) = \frac{(A_{m}^{1} - A_{m}^{0}) u_{m}(\psi_{i-1}, \phi_{i})}{A_{m}},$$

(12)

where $A_{m}$ is the minimum of $u_{m}$ over all possible $(\psi_{i-1}, \phi_{i})$ values and $A_{m}$ is the average value of $A_{m}$ over all possible pairs of amino acids. A typical $V_{m}$ is shown in Fig. 2 b. It is clear that $V_{m}(\psi_{i-1}, \phi_{i})$ does not vanish only in particular regions, in which local structures of proteins are either $\alpha$-helices or $\beta$-sheets.

**Calibration of energy weighting factors**

The weighting factors $\epsilon_{i j}$ are calibrated by comprehensively searching valid values within specific ranges in a selected set of reference proteins, which are 1NJO, 1DJF, 1GB4, 1PIQ, and 2NOU. Specifically, a set of decoy conformations of the reference proteins is selected and the weighting factors have to be in the physical region in which total energies of decoy conformations are greater than those of native structures. To adjust the weighting factors toward the physical region, a cost function is defined by

$$\text{cost} = \sum_{i} \frac{\Delta E_{\text{total}}^{i} f_{i}}{D_{\text{RMS}}^{i} f_{i}}.$$

(13)

Here $\Delta E_{\text{total}} = E_{\text{total}} - E_{\text{native}}^{i}$ is the total energy of the native state for the corresponding reference protein. The factor $f_{i}$ gives high score to the negative values so that weighting factors in the unphysical region can be identified: $f_{i} = 1$ for $\Delta E_{\text{total}} > 0$, otherwise $f_{i} = \epsilon(1 + \Delta E_{\text{total}})$ with $\epsilon$ being an arbitrary small number chosen as $10^{-6}$. $D_{\text{RMS}}^{i}$ measures deviation of the decoy conformation from the native structures and is the relative root mean-square distance defined by Betancourt and Skolnick (32). The set of weighting factors that corresponds to the most positive cost value will be selected. The optimization results and allowed ranges of weighting factors are listed in Table 1. Changing one factor from the default value within the allowed range will not cause serious misfolding. For efficiency, only five reference proteins are used for optimization. Furthermore, during the calibration, if the reference proteins end up with any wrong conformations, in the newly launched Monte Carlo simulation with the selected factors, the misfolded
TABLE 1 Weighting factors of energy terms and its valid range

<table>
<thead>
<tr>
<th>Term</th>
<th>Default</th>
<th>Lower limit</th>
<th>Upper limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>e_{DG}</td>
<td>0.21</td>
<td>0.00</td>
<td>2.56</td>
</tr>
<tr>
<td>e_{DN}</td>
<td>2.00</td>
<td>2.00</td>
<td>2.60</td>
</tr>
<tr>
<td>e_{HB}</td>
<td>4.80</td>
<td>3.84</td>
<td>4.80</td>
</tr>
<tr>
<td>e_{IP}</td>
<td>6.40</td>
<td>4.48</td>
<td>8.32</td>
</tr>
<tr>
<td>e_{SP}</td>
<td>16.0</td>
<td>14.4</td>
<td>19.2</td>
</tr>
<tr>
<td>e_{MJ}</td>
<td>0.85</td>
<td>0.43</td>
<td>1.45</td>
</tr>
<tr>
<td>e_{SA}</td>
<td>0.35</td>
<td>0.54</td>
<td>3.78</td>
</tr>
</tbody>
</table>

conformations will be added into the decoy sets and rerun the process iteratively.

Although the reference proteins are not plentiful, the emergent weighting factors will be examined further by folding a larger pool of target proteins. These targets comprise secondary structures of α-helix (PDB ID: 1DJF, 1DN3, 1DNG, 1DU1, 1EMZ, 1EQX, 1FAC, 1GFH, 1HU6, 1JYC, 1KZ2, 1LBI, 1O53, 1ODP, 1ODQ, 1QG9, 1X0Q, 1X0P, 2A1C, 2AP7, 2B0Y, 2BBL, 2DCL, 2FQ5, 2FXY, 2IM2, 2MYJ, 2RJL, 2RJH, and 1S4W); β-sheet (1B03, 1E0Q, 1E0N, 1J4M, 1K43, 1U6U, 2ESZ, 2ORU, and 1NJ0); and mixed α/b structures (1FSV, 1PSV). At the end, all the target proteins are correctly folded.

Wang-Landau Monte Carlo algorithm

Given the ab initio coarse-grained potential obtained, one can determine the free energy landscape by using the Wang-Landau algorithm (33). The density of states is estimated by random sampling on energy space via the transition probability

$$P(E_1 \rightarrow E_2) = \min \left( \frac{g(E_1)}{g(E_2)} ; 1 \right),$$

where $g(E)$ is the density function of energy $E$. Although this algorithm was first demonstrated on Ising model of spin array, it is portable to molecular systems with continuous energy value (34,35). Specific implementations adapted in our work are the following steps:

1. Define a density function $g(E, X)$ and histogram $H(E, X)$ with $X$ values being any variables other than energy. Set initial values: $g(E, X) = 1$ and $H(E, X) = 0$ for all $E$ and $X$.
2. Generate an initial conformation randomly and calculate its energy $E_1$.
3. Generate a new conformation by making a small change (e.g., the dihedral angles). Calculate the new energy $E_2$, and the transition to the new conformation is determined by the transition probability $P(E_1, X_1 > E_2, X_2) = \min[g(E_1, X_1)/g(E_2, X_2)]$.
4. If the system stays in the original $E_1$ state, $g(E_1, X_1)$ is replaced by $g(E_2, X_2) \times f$ and $H(E, X)$ is accumulated through $H(E, X_1) + 1$. Otherwise, one sets $g(E_2, X_2) = g(E_1, X_1) \times f$ and $H(E, X_2) = H(E_2, X_2) + 1$. The factor $f$ is initially set to $e_{\alpha}$. 
5. After each MC step, check if $< 2\%$ of sites in $H$ are smaller than flat threshold, which is defined to be 10% of averaged $H(E, X)$. If this is satisfied, the histogram is flat and one then sets $f = \sqrt{f / H(E, X)} = 0$ and goes to Step 2. When $f < \exp(-10^{3})$ is satisfied, one exits the procedure.

All the above steps are identical to Wang-Landau’s scheme except for the flat histogram criteria in Step 5, which is modified to accommodate enormous states involved for proteins so that sampling can be done in finite computation time. Once the density of states is constructed, the free energy landscape can be calculated as

$$F(E, X) = E - k_b T \log[g(E, X)],$$

where $k_b$ is the Boltzmann constant and $T$ is the absolute temperature. The variable space $X$ is not restricted to be one dimension and has to be chosen to exhibit the landscape.

RESULTS

Propensity analysis and Monte Carlo simulation

To investigate the energy landscape of 2DX4, we first analyze its propensity. Past studies (36,37) have indicated that each amino acid has its propensity of secondary structure. By using the constructed statistical potential $V_{\alpha}$ (see Theory and Methods), we summarize the nearest-neighbor propensity of 2DX4 in Fig. 3. Here amino acids in nearest neighbors are classified according to the tendency of corresponding amino acids being in α-helix, β-sheet, dual, or neutral. The dual propensity implies the residue pair can adopt either α- or β-structure. By contrast, the neutral propensity implies that the residue pair is free to rotate in dihedral angles and it is often that a turn region of antiparallel β-sheet is developed. From the propensity analysis, it is clear that even though there is no absolute global tendency for 2DX4 being α-helix or β-sheet, by including residues with neutral and dual propensities, there are more residues in favor of α-helix. Nonetheless, the high β-sheet propensity near the C-terminal, containing amino acids V, H, and W, indicates the possibility of switching 2DX4 between helix and hairpin structures. Because each of these three amino acids has larger side-chain radius than the averaged radius of others, it is more difficult for the segment to curl into part of the helix structure. As a result, the strand formed by residues 14–18 regularly dangles in solvent and deposits a nucleation seed to transform from α-helix to β-sheet.

To investigate the stability of α-helix due to residue 14–18, an MC simulation of 2DX4 by starting from an all-helix conformation is conducted. Because the expanding of the strand affects the size of 2DX4, we record the radius of gyration ($R_g$) for structure resembling the α-helix. Larger $R_g$ represents structures with extended strands, whereas smaller $R_g$ represents structures that are closer to the standard α-helix. Because each $R_g$ interval may contains several helix structures with different energy values, the internal energy $U$, defined by the Boltzmann statistics
\[ U = \sum_{E} E \exp(-\beta E), \]

is evaluated as a function of \( R_g \). In Fig. 4, we show the plot of \( U \) versus \( R_g \). It is seen that the lowest energy state is not a complete \( \alpha \)-helix. In general, hydrogen bonds and long-range dipole energy favor helix structures \((23)\). In the case of 2DX4, nearest-neighbor interactions \( V_{NP} \) compete with these helix-favored energies and result in the lowest total energy state with partial helix and partial strand structure. The native \( \alpha \)-helix structure found in our MC simulation is identical to results obtained by the experiment \((18)\) and other simulations \((21)\), indicating the credibility of the coarse-grained potentials described in Eq. 1.

To clarify the final fate of \( \alpha \)-helix, we perform full MC simulations by starting from the initial state of a straight line with all dihedral angles \( \varphi \) and \( \psi \) being equal to 180°. Indeed, the \( \alpha \)-helix and \( \beta \)-hairpin-like structures are found to be two configurations with lowest energies and root mean-square deviation of positions being 3.74 Å and 4.40 Å, respectively. Furthermore, similar to the \( \alpha \)-helix structure, the \( \beta \)-hairpin also has variants in addition to the standard hairpin structure (see the next subsection for more details). The simulations take \( 4 \times 10^8 \) MC steps and ended on either helix or hairpin states. Furthermore, starting from an \( \alpha \)-helix at 400 K \((RT = 0.8 \text{ kcal/mol})\), the \( \alpha \)-helix is transformed into a \( \beta \)-sheet and vice versa. All of the transitions occurred successfully in our MC simulations. However, the helix-to-hairpin transition takes \( 2 \times 10^8 \) more MC steps than that for the transition from hairpin to the helix. A hairpin-to-helix transition finished in \( \sim 5 \times 10^7 \) MC steps, where the reverse process took \( 10^8 \) MC steps or longer. Although number of MC steps does not reflect the physical folding time quantitatively, qualitatively, the obtained asymmetry of transition probability does suggest that the sheet formation of 2DX4 is a slower process than the formation of helix.

**Free energy landscape**

To make sure the helix and hairpin structures found in MC simulations are the only two native structures, we calculate the free energy by employing the Wang-Landau algorithm. To characterize the energy landscape, we use the contact ratios \( Q \) and \( Q^b \) as coordinates. Here \( Q \) is defined by ratios of contact number of a given state to that of the minimum \( \alpha \)-helix structure and \( Q^b \) is defined similarly with the reference structure being the perfect \( \beta \)-hairpin structure. The free energy \( F \) is thus a function of \( Q \) and \( Q^b \), both of which range from 0 to 1. In the calculation, to insure that all regions can be accessed, a trial run with \( 4 \times 10^8 \) MC steps is first performed to identify regions with scarce probability. In the latter runs, free energy density in these regions will be computed separately.

Fig. 5a shows the resulting complete free energy landscape for 2DX4. It demonstrates that the free energy has minima at helix and hairpin states. As we can see, similar to the \( \alpha \)-helix structure, in addition to the standard hairpin structure, the \( \beta \)-hairpin also has a variant structure labeled by \( \beta^\prime \), whose turn is shifted by one side chain in comparison to that of the hairpin. Both the hairpin \( \beta \) and \( \beta^\prime \) structures have the same contact number \( Q \) and the same energy; hence they can be considered collectively as the hairpin-like structures. The difference of free energies for the helix and hairpin-like structures is \( \sim 0.17 \text{ kcal/mol} \) at room temperature, which clearly demonstrates that 2DX4 is a protein with two stable native structures. In Fig. 5b, the one-dimensional free energy curves \( F(Q) \) are deduced from the density of states \( g(E,Q,Q') \) via the formula

\[ \exp\left(\frac{-F(Q)}{k_B T}\right) = \sum_{E,Q'} g(E,Q,Q') \exp\left(\frac{-E}{k_B T}\right). \]

A free-energy barrier at \( \sim 10 \text{ kcal/mol} \) exists between helix and hairpin structures. Because the energy barrier is much larger than typical energy fluctuations \( k_B T \), it stabilizes both the helix and hairpin structures. The free energy landscape also depends on temperature. At temperature \( k_B T = 0.8, 400 \text{ K} \), the minimum at helix side expands from \( Q = 1 \) to \( Q = 0.65 \) with residues 1–10 being kept in helix conformation. In other words, half of the peptide on N-terminal is thermally stable in helix, and residues 11–14 are free to denature at high temperatures.
Folding a Protein: Helix or Hairpin

As a comparison, we examine energy landscapes of mutated 2DX4–Y12S mutation, which are labeled as DP3 and DP5 in the previous experiment (18). It is reported that DP3 has zero population of hairpin formation in the sense that even though there is minor intrastand signal, there is no interstrand signal for the hairpin structure. It is therefore important to examine native structures of DP3 in our model. Fig. 5a reveals that for DP3, the helix region gets expanded, whereas hairpin regions get reduced. This indicates that the helix structure is more stable for DP3. Indeed, Fig. 5b shows that the free energy of the helix state is less than that of the hairpin state by 1.1 kcal/mol at room temperature. In addition, we find that this energy difference is sensitive to temperature and becomes 1.4 kcal/mol at 100 K.

In contrast, for DP5, the free energy of the helix state is found to be fixed at 100–298 K, suggesting that helical structure is thermally more stable in DP5 than in DP3, in agreement with experimental observation (18). Note that it was presumed (18) that absence of π–π interaction of Tyr12–His7 near the turn region is the cause for the absence of hairpin in DP3. In addition, the sequence propensity is also relevant. By closely inspecting the neighboring propensity energy $E_{NP}$, the G11-Y12-I13 peptide has $-0.5$ kcal/mol and $-1.9$ kcal/mol in helix and sheet conformations, respectively. In contrast, the G11-S12-I13 peptide in DP3 has $-1$ kcal/mol and 0 kcal/mol in helix and sheet conformations. Namely, the three successive residues in DP3 have a helical propensity; however, in DP5 these residues have a sheet propensity without losing a helix propensity. These observations are in accordance with experimental results that DP3 has only helical population and DP5 are stable both in helix and sheet conformations.

**Mechanism of degeneracy**

The mechanism for the existence of degenerate native structures can be explored by analyzing changes of different energy terms when 2DX4 changes between the helix and the hairpin structures. In Fig. 6, we shows changes of different energies along one of the paths that connects the helix and the hairpin structures. Because the route is chosen such that 2DX4 is not fully stretched on the route, the energy changes of the sequence-dependent terms, $E_{NP}$, $E_{M2}$, and $E_{SA}$ are small and may appear to play minor roles during the folding process. Nonetheless, by taking conformation 8 as an example, we find that the total energy is $-150$ kcal/mol and $E_{NP} + E_{M2} + E_{SA} = -55$ kcal/mol. Hence the sequence-dependent energy is 36% of the total energy and plays a major role in the folding. The obtained percentage of the sequence-dependent free energy is generally consistent with experimental observations in which it is explicitly demonstrated that whereas specific proteins, such as protein G and protein L, may have 75% difference in their sequences, the free energy released during folding differs only by ~28% (38,39).

From Fig. 6, it is clear that there is a large compensation between hydrogen-bond energy (HB) and local dipole energy in going from the helix structure to the hairpin structures and vice versa. In the inset of Fig. 6, it is seen that even though there is also a large change of the distribution of $E_{NP}$ between α-propensity (NPα) and β-propensity (NPβ), the net change of $E_{NP}$ is small. Hence the main driving mechanism is the compensation between hydrogen bond and the local dipole energy. Physically, it is known that the helix structure has more hydrogen bonds (23) and hence one loses energy in hydrogen bonds by going from the helix structure to the hairpin structure. On the other hand, β-sheets contain...
large antiparallel dipoles on nearest-neighboring amide planes, which lowers the local dipole interaction energy. The competition of hydrogen-bond energy and local dipole energy depends on the length of the protein.

To see why 2DX4 is special, we examine the difference of hydrogen-bond energy and local dipole energy for α-helix and β-sheet versus number of side chains. The energy difference is optimized with respect to the number of β-strands. Fig. 7 shows the computed optimized difference of hydrogen-bond energy and local dipole energy for α-helix and β-sheet versus number of side chains. It is seen that the difference of hydrogen-bond energy and local dipole energy vanishes at number of side chains being around 18 amino acids, which is precisely the number of side chains in 2DX4. Therefore, our results show that although differences in other energy changes in 2DX4 contribute 2–3 kcal/mol, the major compensation in energy comes from the hydrogen-bond energy and local dipole energy and it leads to the degeneracy of the helix and hairpin structures.

**DISCUSSION AND CONCLUSION**

In conclusion, the possibility for the existence of degenerate native states provides what to our knowledge is new insight into the folding mechanism of proteins. Our results show that the possibility is realized in the designed 2DX4, which possesses two nearly degenerate native structures: one has a helix structure, whereas the other has a hairpin-like structure. The two degenerate native structures of 2DX4 are shown to be separated by an energy barrier of 10 kcal/mol. Based on the usage of the Arrhenius form for the kinetic rate, $k = A e^{-E/RT}$, where the preexponential factor $A$ is the attempt rate and can be estimated as inverse of typical fold time, $10^3$/s, we find that the kinetic rate for the transformation between two native states is $\sim 6 \times 10^{-5}$/s. The transformation rate is thus a slow process. As a result, two degenerate structures are stabilized, consistent with experiments (18) in which no apparent transitions between two degenerate structures are observed.

Our results further indicate that the existence of two degenerate native structures in 2DX4 is driven by large compensation between the hydrogen-bond energy and the local dipole energy. The length study of the difference between hydrogen-bond energy and local dipole energy for α-helix and β-hairpin shows that 2DX4 is special in that it has 18 amino acids, which is exactly the number required for balancing the hydrogen-bond energy and local dipole energy.
dipole energy. Therefore, although differences of other energy terms in 2DX4 do contribute, the major energy compensation in going from α-helix and β-hairpin is determined by the hydrogen-bond energy and local dipole energy, which leads to the observed degeneracy of the helix and hairpin structures. If the length study is further extended to a larger number of side chains, we find that the next balance between hydrogen-bond energy and local dipole energy for α-helix and β-sheet could occur for the number of side chains being ~40. Although it does not mean that degenerate binary native structures will necessarily occur, our results provide important clues for the study of native structures of proteins, especially for proteins with possibly degenerate native states.

SUPPORTING MATERIAL

One figure is available at http://www.biophysj.org/biophys/supplemental/S0006-3495(12)00582-6.

We thank Prof. Chia-Ching Chang for helpful discussions.

This work was supported by the National Science Council and National Tsing Hua University of Taiwan.

REFERENCES


