Effective Potentials for Folding Proteins

Nan-Yow Chen,[1] Zheng-Yao Su,[2,3] and Chung-Yu Mou[2,4]

[1]*Institute of Physics, Academia Sinica, Nankang, Taipei 11529, Taiwan*
[2]*Department of Physics, National Tsing Hua University, Hsinchu, Taiwan*
[3]*National Center for High-Performance Computing, Hsinchu, Taiwan*
[4]*Physics Division, National Center for Theoretical Sciences, P. O. Box 2-131, Hsinchu, Taiwan*
(Received 18 September 2004; published 22 February 2006)

A coarse-grained off-lattice model that is not biased in any way to the native state is proposed to fold proteins. To predict the native structure in a reasonable time, the model has included the essential effects of water in an effective potential. Two new ingredients, the dipole-dipole interaction and the local hydrophobic interaction, are introduced and are shown to be as crucial as the hydrogen bonding. The model allows successful folding of the wild-type sequence of protein G and may have provided important hints to the study of protein folding.

DOI: 10.1103/PhysRevLett.96.078103 PACS numbers: 87.15.Aa, 36.20.−r, 87.14.Ee

The problem of predicting the native structure of a protein for a given sequence has been of great interest due to its relevancy to many fields in biology. In the crudest level, lattice models are proposed and have provided important insights [1,2]; however, due to the oversimplification, they are far from real applications. On the other hand, all-atom simulations deliver more details for the folding process, but the requirement of computational resources tends to be realistically unaffordable [3]. Developing models of coarse graining thus becomes the next step. For this purpose, off-lattice models [4] using Gō-type [5] potentials have been used to explore the folding dynamics. Since the relevant interactions are based on native structures, the Gō-type potentials cannot be used to predict structures. There are also models that succeeded in separately folding helix bundles or folding beta hairpins [6]. Nevertheless, the interacting potentials employed are also biased towards the native states. So far, there is no model that can fold proteins using realistic potentials; it is therefore desirable to construct a coarse-grained model that can fold proteins without being biased in any way to the native state.

In this Letter, based on microscopic considerations, we propose a coarse-grained model with realistic potentials. The model has been tested successfully on more than 16 small proteins, of sizes from 12 to 56 amino acids [7]. For most examples, even without particularly optimizing our code, the computing time is reasonably short and is within the order of hours on ordinary desktop computers. Here, instead of exploring its predicting ability, we shall be focusing on only one protein (one of the protein G families with protein data bank (PDB) ID:1GB4) to illustrate the folding mechanism embedded in the proposed model. A brief summary of other important proteins is given in [7].

In our model, side chains are coarse grained as spheres but explicit structures are kept in backbones [8]. On the other hand, water molecules are not included explicitly but their effects are incorporated in effective potentials among side chains and backbones. The hydrophobic (HP) interaction has been known as the most important effect due to water. Recently, it is realized that the length scale of water molecules has to be kept at short distances to prevent proteins collapsing prematurely [9]. Therefore, the desolvation model [9] combined with the Miyazawa-Jernigan (MJ) matrix [10] is employed to describe the interaction among the side chains. Furthermore, since the MJ potential is a non-neighboring interaction, its extension to include nearest neighbors (NN) along the sequence is needed. Similar to the spirit of the HP model [2], a *local hydrophobic potential*, $V_{\mathrm{Local\,HP}}$, is implemented by assigning potential energies to any successive pairs of amino acids according to their hydrophobicity. On the other hand, the hydrogen bonding (HB) has long been thought of as the key molecular interaction [6]. However, for small proteins, it is known that HB prefers the helix structure over the beta sheet because the former has a larger number of HBs. Thus it hints to include a second molecular interaction. Indeed, analysis on the MJ matrix indicates that the electric dipole-dipole interaction dominates in the pairwise interaction among side chains [11]. Microscopically, there is also charge imbalance in the CO-NH group on the amide plane with the magnitude of the dipole being estimated to be $p = 1.15 \times 10^{-19}$ Cm. Simple analyses reveal that the directions of these dipoles have strong correlation with the secondary structure [12]: in the alpha helix, successive dipoles on the backbone tend to be in parallel; while in the beta sheet, they tend to change directions alternately (see Fig. 1, for example). In order to capture relevant energetics, we explicitly introduce the *dipole-dipole interaction* V_{DD} among the backbone elements. The potentials $V_{\mathrm{Local\,HP}}$ and V_{DD} are the main ingredients that make our model different from early models. Remarkably, our simulations indicate that these two interactions and the hydrogen bonding form the key interactions for determining the secondary structure. Specifically, we find that while the hydrogen bonding is essential to the formation of the alpha

 © 2006 The American Physical Society

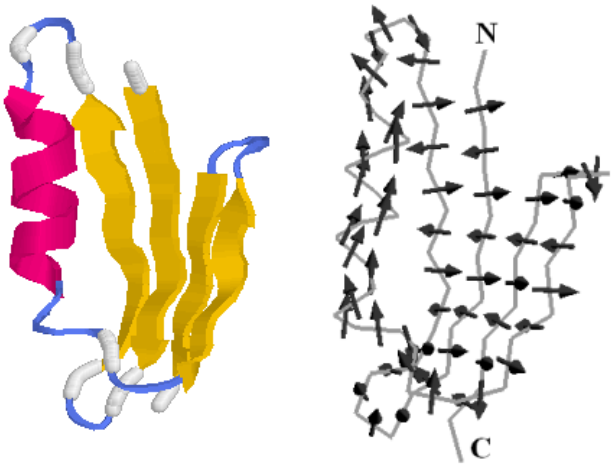

FIG. 1 (color online). Schematic plot of the simulated protein G—1GB4: native conformation and corresponding dipole configuration.

helix, to fold the beta sheet, both V_{DD} and $V_{\mathrm{Local\ HP}}$ are indispensable.

The potential is constructed in a renormalized fashion: except for global multiplicative scales (denoted by ϵ_α in the following, with α representing different contributions), interactions (such as V_{DD} and V_{MJ}, see below) at large distances take the usual form; while for interactions (such as $V_{\mathrm{Local\ HP}}$ and V_{ND}, see below) at successive neighbors (short distances), since the variation of distance is unimportant, only angle variables are kept. The parameters employed in the potentials are adopted from experimental data [8], while the scales ϵ_α's are calibrated based on a few proteins of known structures [13].

The degrees of freedom for backbones are two Ramachandran angles ϕ and ψ [14]. Since the peptide bond on any amide plane is partially double bonded, the angle ω around the peptide bond is fixed to be 180° so that it corresponds to the *trans*-conformation. The spheres that represent side chains are centered at C^β and are attached to C^α atoms rigidly, and different effective radii are assigned, consistent with the geometric structures [15]. In these representations and with all energies being in units of kcal/mol, the potential can be written as

$$V_{\mathrm{total}} = V_{\mathrm{steric}} + V_{\mathrm{HB}} + V_{\mathrm{DD}} + V_{\mathrm{MJ}} + V_{\mathrm{Local\ HP}} + V_A. \tag{1}$$

Here V_{steric} enforces structural constraints such as hard-core potentials to avoid unphysical contacts. V_{HB} accounts for the hydrogen bonding between any non-neighboring NH (labeled by i) and CO (j) pair and is implemented as $V_{\mathrm{HB}} = \epsilon_{\mathrm{HB}}\sum_{n,i,j} u(r_{ij})v(\theta_{n,ij})$, where r_{ij} is the distance between H_i and O_j and $u(r)$ is the standard 12-10 Lennard-Jones potential with the equilibrium distance being set to the averaged experimental value 1.738 Å [8]. The angle function v imposes the directional nature of HB, parametrized by three angles ($n = 1, 2, 3$): $\pi - \angle C_iO_iH_j$, $C_iO_i \wedge N_jH_j$, and $\pi - \angle O_iH_jN_j$. Their values are confined to the averaged experimental data [8], respectively: 26.77°, 11.60°, and 17.98°. To increase the efficiency of HB formation, certain uncertainty $\Delta\theta$ is allowed. Empirically, $\Delta\theta = 60°$ is most efficient.

The dipole term V_{DD} at large distances takes the ordinary form

$$V_{\mathrm{DG}} = \epsilon_{\mathrm{DG}}\sum_{i,j}\left[\frac{\vec{p}_i\cdot\vec{p}_j}{r_{ij}^3} - \frac{3(\vec{p}_i\cdot\vec{r}_{ij})(\vec{p}_j\cdot\vec{r}_{ij})}{r_{ij}^5}\right], \tag{2}$$

where $\vec{p}_i$ and $\vec{p}_j$ are dipoles of either CO or NH, and the summation excludes successive dipoles. When dipoles are in successive neighbors, it is given by

$$V_{\mathrm{DN}} = \epsilon_{\mathrm{DN}}\sum_i \frac{1}{2}\left(\frac{\vec{p}_i\cdot\vec{p}_{i+1}}{p_ip_{i+1}} - 1\right). \tag{3}$$

V_{MJ} is the extension of the MJ matrix with the form $V_{\mathrm{MJ}} = \epsilon_{\mathrm{MJ}}\sum_{i,j}[V_{\mathrm{LJ}}(r_{ij}) + V_{\mathrm{G1}}(r_{ij}) + V_{\mathrm{G2}}(r_{ij})]$. Here V_{LJ} is the MJ matrix element ϵ_{ij} multiplied by the usual 12-6 Lennard-Jones potential with the equilibrium distance being the sum of radii of two side chains. $V_{\mathrm{G1}} + V_{\mathrm{G2}}$ represents the potential obtained numerically in the desolvation model [9]. For numerical purpose, however, we find that it is more convenient to use the following approximately analytic forms: $V_{\mathrm{G1}} = \epsilon_1 \times \exp[-\sigma_w \times (r_{ij} - r_b)^2]$ is a Gaussian fit to the desolvation barrier with r_b being the position of desolvation barrier and σ_w being the size of the water molecule, while $V_{\mathrm{G2}} = \epsilon_2 \times \exp[-\sigma_w \times (r_{ij} - r_w)^2]$ is an inverted Gaussian fit to the metastable minimum at r_w due to water molecules. Here for the best fit, $\epsilon_1 \sim 5|\epsilon_{ij}|/9$ and $\epsilon_2 \sim -|\epsilon_{ij}|/3$.

The potential $V_{\mathrm{Local\ HP}}$ acts only on successive pairs of side chains

$$V_{\mathrm{Local\ HP}} = \sum_i V_{q_i,q_{i+1}}. \tag{4}$$

Here q_i represents the hydrophobicity or the charge state of the ith side chain. Following Ref. [12], q_i are classified into hydrophobic (H), polar (P), neutral (N), positive charged (+), and negative charged (−). In this classification, N is regarded as a referential type such that whenever $q_i = N$ or $q_{i+1} = N$, $V_{q_i,q_{i+1}} = 0$. Furthermore, when charged side chains encounter other noncharged ones, they are considered as polar. Therefore, the only nontrivial potential energies are ($V_{\mathrm{HH}}, V_{\mathrm{PP}}, V_{+-}$) (attractive) and ($V_{\mathrm{HP}}, V_{++}$) (repulsive). To implement the hydrophobic effects, an attractive pair acquires a negative energy $-\epsilon_{q_i,q_{i+1}}$ when their $C^\alpha C^\beta$ lines are parallel to each other, and when in other orientation, no energy is assigned; while for repulsive pairs, a negative energy $-\epsilon_{q_i,q_{i+1}}$ is assigned when their $C^\alpha C^\beta$ lines are antiparallel. In practice, a smooth function is used to interpolate between finite $V_{q_i,q_{i+1}}$ and zero.

Finally, V_A is an on-site potential in proportion to the area of each side chain that is exposed to water. The proportional constant is $\epsilon_{ii} - \langle \epsilon_{ii} \rangle$ with i being the index for the side chain. The existence of V_A has already been found in the analysis of the MJ matrix [11] and it helps to further contrast the hydrophobicity of each side chain.

The Monte Carlo method is employed to fold proteins. After careful calibration [13], the global scales are found to be $\epsilon_{\rm HB} \approx 4.8$, $\epsilon_{\rm DG} \approx 0.2$, $\epsilon_{\rm DN} \approx 2.1$, $\epsilon_{\rm MJ} \approx 0.2$, and for $V_{\rm Local\,HP}$, $\epsilon_{\rm HH} = \epsilon_{\rm PP} = \epsilon_{\rm HP} \approx 5.0$, $\epsilon_{++} = \epsilon_{+-} \approx 5.0$. The same scales are adopted to simulate the protein 1GB4, which is a wild-type protein with one alpha helix and two beta hairpins. Figure 1 shows its spatial arrangement and corresponding dipole arrangement of our simulated energy ground state, while Fig. 2 shows the contact map. The native contact number ratio (Q) for simulated ground state is 0.6, while the RMSD is 2.97 Å. Clearly, our simulation is in good agreement with the experiment while the computing time is only a few hours on a P4-3.0 GHz PC. Note that the ground state energy is -545 kcal/mol and the nearest local minimum is about 34 kcal/mol higher in energy. Furthermore, both the helix and the beta sheet are formed only when correct scales ϵ's and appropriate temperature are employed. The portability of these scales (and our model) to other proteins are tested in 15 proteins. The results are briefly summarized in Ref. [7]. Our results are generally in good agreement with experiments with the tolerance of ϵ's being about 0.5. Occasionally, the accuracy is not good. However, in that case, the cause is due to the metal ion not being included in our simulation [7].

To clarify the roles of $V_{\rm DD}$ and $V_{\rm Local\,HP}$, the alpha helix (A24 to D37) and the beta hairpin with C terminus (G42 to E57) are extracted. The energy versus Q along the folding is then monitored for different strengths of the potentials. Figures 3(a) and 3(b) show the effect of $V_{\rm DD}$ for three different strengths. Clearly, we see that the native conformation ($Q = 1$) stays at the minimum for the helix, while for the beta hairpin, it gradually moves away from the minimum. When $V_{\rm DD}$ is completely turned off, the beta sheet is no longer the ground state. Similar analyses are done by tuning $V_{\rm Local\,HP}$ as shown in Figs. 4(a) and 4(b). We see that although affecting the formation of the alpha helix, both $V_{\rm Local\,HP}$ and $V_{\rm DD}$ have stronger effects on the formation of the beta sheet and can change the ground state completely. Similar behaviors also occur for the beta hairpin with N terminus and other 15 proteins. For 1GB4, if we turn off $V_{\rm Local\,HP}$ and $V_{\rm DD}$, only segments of helices are formed. Therefore, both $V_{\rm DD}$ and $V_{\rm Local\,HP}$ are responsible for the formation of the beta sheet.

It should be noted that the above analyses are done with fixed $V_{\rm HB}$, and the vanishing alpha helix in Fig. 4(a) can be restablized by increasing $V_{\rm HB}$. However, similar restablization does not occur to the beta sheet due to the fact that the helix has more HBs. Therefore, when $V_{\rm HB}$ is large enough, the helix conformation always wins, and even a beta sheet will be turned into a helix. On the other hand, because successive dipoles in a helix tend to have unfavorable parallel orientations, sufficient strong $V_{\rm DD}$ can stabilize the beta sheet over the helix. Therefore, in the intermediate strength of $V_{\rm HB}$, a beta sheet could form if the deficient energy due to smaller number of HBs is compensated by the energy gain of $V_{\rm DD}$.

Similar analysis on the MJ potential shows that instead of deciding the secondary structure explicitly, $V_{\rm MJ}$ plays a crucial role in making its formation more efficiently. In the

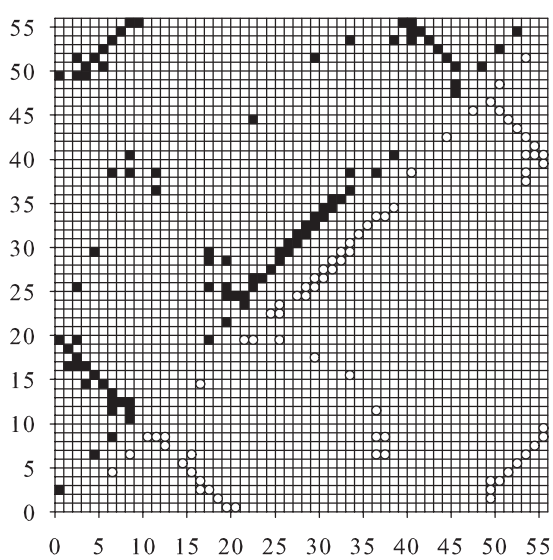

FIG. 2. Contact map of energy ground states (native structures) for 1GB4 at $k_BT = 0.8$ kcal/mol with open circle being our simulation result ($Q = 0.6$) and solid square being the data from PDB.

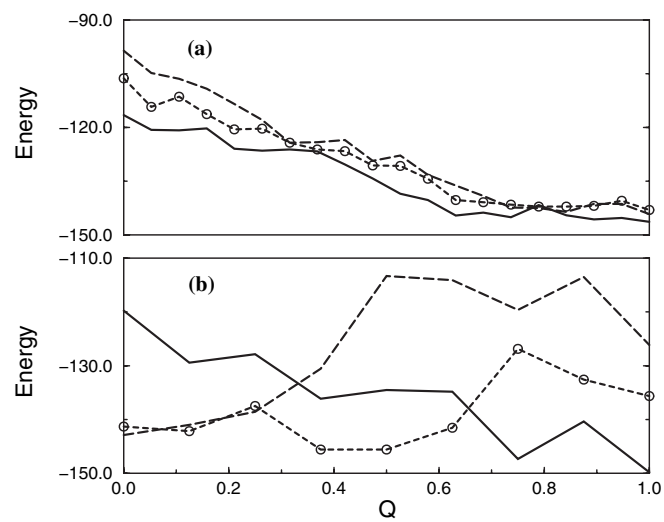

FIG. 3. Effects of different strengths of $V_{\rm DD}$ on the formation of the alpha helix (a) and the beta hairpin with C terminus (b). The corresponding strengths: solid line—$V_{\rm DD}$, circles—$0.5V_{\rm DD}$, and long dashed line— 0. Q is the native contact number ratio with $Q = 1$ corresponding to the native conformation.

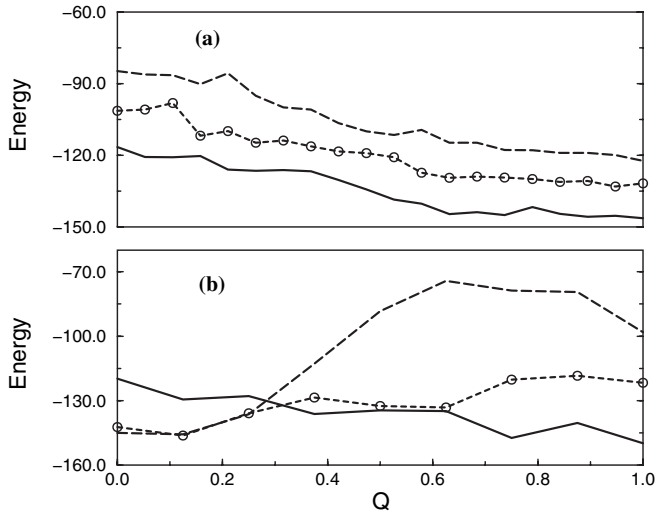

FIG. 4. Effects of different strengths of $V_{\text{Local HP}}$ on the formation of the alpha helix (a) and the beta hairpin with C terminus (b). The corresponding strengths: solid line—$V_{\text{Local HP}}$, circles—$0.5V_{\text{Local HP}}$, and long dashed line—0.

early stages of folding, V_{MJ} collapses all residues into a compact space. Only when the collapsing happens, interactions of shorter ranges could function. If the initial collapsing does not go in the right direction or happens too fast, the final protein structure may become disordered. After the initial collapsing, the potentials V_{DD}, $V_{\text{Local HP}}$, and V_{HB} start to dominate. At this point, an obvious question remains to be addressed: since both V_{DD} and V_{HB} are sequence independent, then for a given sequence, what determines that it should fold into a helix or a beta sheet? This is where $V_{\text{Local HP}}$ comes into play because it forces successive neighboring side chains to be either on the same side or on the opposite side of the backbone according to their hydrophobicity. Thus different sequences result in different local spatial arrangements of side chains, and only when the arrangement is correct, the protein can be compacted into the correct secondary structure. Finally, our analysis shows that even though the native state is still the ground state in the absence of V_A, incorrect strength of V_A would result in itinerant motion of the secondary structures. Therefore, V_A is primarily responsible for stabilizing the tertiary structure.

In conclusion, an effective potential that can fold proteins without being biased to the native state is constructed and tested. All testing peptides can fold to their native states in acceptable computing time. By systematically tuning relative strengths of interactions in the potential, the dipole-dipole interaction V_{DD} and the local hydrophobic interaction $V_{\text{Local HP}}$ are shown to be as crucial as the hydrogen bonding V_{HB}. While V_{HB} prefers the helix structures, V_{DD} tends to form sheetlike structures. Only when a subtle balance between these two interactions holds, the helix and sheet structures can coexist. The sequence-dependent potential $V_{\text{Local HP}}$ is then responsible for the final selection of either a helix or a beta sheet forming.

We thank Professor C. C. Chang and Professor T. K. Lee for useful discussions. This research was supported by NSC of Taiwan.

[1] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, Biochemistry **33**, 10 026 (1994); D. K. Klimov and D. Thirumalai, J. Chem. Phys. **109**, 4119 (1998).
[2] K. A. Dill *et al.*, Protein Science **4**, 561 (1995).
[3] Y. Duan and P. A. Kollman, Science **282**, 740 (1998).
[4] H. Nymeyer, A. E. Garcia, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **95**, 5921 (1998); Y. Zhou and M. Karplus, Nature (London) **401**, 400 (1999); C. Clementi, P. A. Jennings, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **97**, 5871 (2000).
[5] N. Gō and H. Taketomi, Proc. Natl. Acad. Sci. U.S.A. **75**, 559 (1978).
[6] Z. Guo and D. Thirumalai, J. Mol. Biol. **263**, 323 (1996); G. Favrin, A. Irbäck, and S. Wallin, Proteins: Struct., Funct., Genet. **47**, 99 (2002); A. Irbäck and F. Sjunnesson, Proteins: Struct., Funct., Genet. **56**, 110 (2004); S. Takada, Z. L. Schulten, and P. G. Wolynes, J. Chem. Phys. **110**, 11 616 (1999); J.-E. Shea, Y. D. Nochomovitz, Z. Guo, and C. L. Brooks, III, J. Chem. Phys. **109**, 2895 (1998); D. K. Klimov and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **97**, 2544 (2000).
[7] N. Y. Chen, Ph.D. thesis, National Tsing Hua University, 2004. Most of the examples are real proteins from the PDB. Some of them are often simulated in literature such as 1NJ0, 1IBN, 1ZDD, and 1VII with corresponding root-mean-square distances (RMSDs) in our simulations: 2.94, 3.53, 4.19, and 5.39 (Å) and the native contact number ratios (Q): 0.47, 0.71, 0.70, and 0.54. Here the RMSD is evaluated according to E. A. Coutsias *et al.*, J. Comput. Chem. **25**, 1849 (2004).
[8] Experimental data such as bond lengths and angles in the backbones can be found in G. Solomons and C. Fryhle, *Organic Chemistry* (Wiley, New York, 2000), 7th ed. Other data such as the average direction of the HB are from the PDB data.
[9] M. S. Cheung, A. E. Garcia, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **99**, 685 (2002).
[10] S. Miyazawa and R. L. Jernigan, J. Mol. Biol. **256**, 623 (1996).
[11] Z.-H. Wang and H. C. Lee, Phys. Rev. Lett. **84**, 574 (2000).
[12] C. Branden and J. Tooze, *Introduction to Protein Structure* (Garland, New York, 1998), 2nd ed.
[13] The calibration is based on proteins simulated in the first two references of Ref. [6] and S. R. Griffiths-Jones *et al.*, J. Mol. Biol. **292**, 1051 (1999). The scales ϵ's are chosen such that the energies of the above peptides are minima.
[14] G. N. Ramachandran and V. Sasisekharan, Adv. Protein Chem. **23**, 283 (1968).
[15] A. Finkestein and O. B. Ptitsyn, *Protein Physics* (Academic, New York, 2002), p. 118.