Unraveling protein folding mechanism by analyzing the hierarchy of models with increasing level of detail

Tomohiko Hayashi, Satoshi Yasuda, Tatjana Škrbić, Achille Giacometti, and Masahiro Kinoshita

Citation: The Journal of Chemical Physics 147, 125102 (2017); doi: 10.1063/1.4999376
View online: http://dx.doi.org/10.1063/1.4999376
View Table of Contents: http://aip.scitation.org/toc/jcp/147/12
Published by the American Institute of Physics
Unraveling protein folding mechanism by analyzing the hierarchy of models with increasing level of detail

Tomohiko Hayashi,1 Satoshi Yasuda,1,2,3 Tatjana Škrbić,4 Achille Giacometti,4 and Masahiro Kinoshita1,a)

1Institute of Advanced Energy, Kyoto University, Uji, Kyoto 611-0011, Japan
2Graduate School of Science, Chiba University, 1-33 Yayoi-cho, Inage, Chiba 263-8522, Japan
3Molecular Chirality Research Center, Chiba University, 1-33 Yayoi-cho, Inage, Chiba 263-8522, Japan
4Dipartimento di Scienze Molecolari e Nanosistemi, Università Ca’ Foscari Venezia, Edificio Alfa Campus Scientifico, Via Torino 155, Venezia-Mestre I-3010, Italy

(Received 7 August 2017; accepted 7 September 2017; published online 25 September 2017)

Taking protein G with 56 residues for a case study, we investigate the mechanism of protein folding. In addition to its native structure possessing α-helix and β-sheet contents of 27% and 39%, respectively, we construct a number of misfolded decoys with a wide variety of α-helix and β-sheet contents. We then consider a hierarchy of 8 different models with increasing level of detail in terms of the number of entropic and energetic physical factors incorporated. The polyatomic structure is always taken into account, but the side chains are removed in half of the models. The solvent is formed by either neutral hard spheres or water molecules. Protein intramolecular hydrogen bonds (H-bonds) and protein-solvent H-bonds (the latter is present only in water) are accounted for or not, depending on the model considered. We then apply a physics-based free-energy function (FEF) corresponding to each model and investigate which structures are most stabilized. This special approach taken on a step-by-step basis enables us to clarify the role of each physical factor in contributing to the structural stability and separately elucidate its effect. Depending on the model employed, significantly different structures such as very compact configurations with no secondary structures and configurations of associated α-helices are optimally stabilized. The native structure can be identified as that with lowest FEF only when the most detailed model is employed. This result is significant for at least the two reasons: The most detailed model considered here is able to capture the fundamental aspects of protein folding notwithstanding its simplicity; and it is shown that the native structure is stabilized by a complex interplay of minimal multiple factors that must be all included in the description. In the absence of even a single of these factors, the protein is likely to be driven towards a different, more stable state. Published by AIP Publishing. https://doi.org/10.1063/1.4999376

I. INTRODUCTION

A protein folds into its own unique three-dimensional (3D) structure called “native structure” in aqueous solution under physiological conditions. Protein folding is one of the most fundamental self-assembly processes in biological systems. Unraveling its mechanism is a central issue for understanding life at the molecular level and expected to provide new physical insight into the other self-assembly processes as well. However, the native structure is formed by a complex interplay of a number of physical factors, and despite an enormous amount of experimental and theoretical effort, there are still lots of uncertain and controversial aspects.1 Molecular dynamics (MD) simulations based on all-atom models have been popular tools for investigating protein folding.2–5 Still, it is difficult within this framework to elucidate how and to what extent each factor contributes to the stabilization of the native structure. Moreover, it was shown that MD results obtained are strongly dependent on the force fields employed even at the qualitative level.4,5 For a polypeptide given, some force fields stabilize the α-helix whereas others stabilize the β-hairpin or intermediate structures.5 Within this scenario, an approach combining concepts based on statistical thermodynamics with minimal models could then prove more convenient to gain insights addressing specific issues. This study sets in this perspective.

The factors promoting protein folding are as follows: (a) the hydrophobic effect; (b) decrease in protein intramolecular electrostatic (ES) interaction energy; and (c) decrease in intramolecular van der Waals (vdW) interaction energy. The factors hindering protein folding are as follows: (d) the conformational-entropy loss of the protein; (e) increase in protein-water ES interaction energy; and (f) increase in protein-water vdW interaction energy. Factor (e) accompanies a decrease in water-water ES interaction energy due to the structural reorganization of water released to the bulk upon protein folding: About half of the increase due to factor (e) is cancelled by this decrease, but the net energy change is positive. Hereafter, this positive change in net energy is referred to as factor (e). Likewise, factor (f) accompanies a change
in water-water vdW interaction energy due to the structural reorganization of water released to the bulk. However, the net energy change is positive, and this positive change in net energy is referred to as factor (f).

Here, we comment on factor (a) whose physical origin is the most controversial. In what follows, we describe our concept of the hydrophobic effect.6–9 The presence of a solute in water generates an excluded space which the centers of water molecules cannot enter. The presence of a water molecule also generates an excluded space for the other water molecules in the system. Water molecules are thus entropically correlated, causing “water crowding.” When a solute changes its structure to a more compact one, for example, the volume of the total excluded space decreases, leading to an increase in the total volume available to the translational displacement of water molecules, which is followed by an increase in the translational, configurational entropy of water. The entropy increase is ascribed primarily to the reduction of water crowding. This effect acting primarily to mitigate the water crowding is the physical origin of the hydrophobic effect. We refer to this effect as “entropic excluded-volume (EV) effect.”7–9 Note that water molecules in the entire system (not limited to the water molecules near the solute) contribute to the effect. Obviously, modeling water as dielectric continuum fails to account for the effect. Several different views on the hydrophobic effect have been suggested,10 but they are not capable of elucidating the following features: the structure formed by a self-assembly process driven by the hydrophobic effect is collapsed at elevated pressures and the power of the structure formation is weakened at low temperatures as evidenced by pressure and cold denaturations of a protein.7–9 Kinoshita and co-workers have been arguing that these pressure and temperature dependencies of the hydrophobic effect can be explicated only by the concept mentioned above.7–9 We emphasize that the concept is not relevant to the entropically unfavorable water molecules arising from the so-called ice-berg structure near a nonpolar group or from the formation of protein-water H-bonds. Unfortunately, the concept has often been misinterpreted in the literature.7,11

The native structure of a protein always possesses the secondary structures formed by a portion or portions of the backbone. To explore the formation mechanism of the secondary structures, tube-like models were employed for a protein.12,13 On the other hand, Kinoshita and co-workers emphasized the crucial role played by the side chains in a protein and remarked that the formation mechanism of the secondary structures as well as the native structure must be investigated by incorporating the side chains and geometric properties of each side chain in the model.14,15 Recently, Giacometti and co-workers16,17 studied the mechanism of protein folding using a coarse-grained model where each amino acid was represented by two beads: One of them is for the backbone and the other is for a side chain. In addition to steric interactions common in both of them, non-consecutive backbone beads experienced short range attractions, mimicking hydrophobic interactions promoting compactification of the chain. The Cα-Cα distance and vdW radii associated with the backbone beads were set at the experimental values for proteins (∼3.8 Å and ∼5 Å, respectively), meaning that consecutive backbone beads could partially interpenetrate (but non-consecutive beads could not). Side chains experienced only steric interactions and their sizes were adjusted to an average value of the vdW spheres of amino acids. In the studies mentioned above,12–17 the backbone and side chains have been modeled in a variety of ways.

In this study, we wish to account for only essential factors in our physical picture of the stability of a folded protein. First, we consider factor (a) based on our concept (the entropic EV effect) mentioned above. Second, only protein intramolecular (P-P), protein-water (P-W), and water-water (W-W) hydrogen bonds (H-bonds) are considered in factors (b) and (e) as principal contributors to the ES energies. Upon the burial of donors and acceptors for P-P H-bonds, the increase in ES energy caused the break of P-W H-bonds cannot necessarily be compensated by the decrease in ES energy brought by the formation of P-P H-bonds. By contrast, it can be assumed that the increase in vdW energy can be cancelled to a significant extent by the decrease in vdW energy: Third, (c) and (f) are assumed to be compensating and not taken into account. Since we treat only very compact structures of a protein (i.e., the native fold and a number of misfolded decoys), factor (d) can be neglected. The free-energy function (FEF) developed by Kinoshita and co-workers18,19 is best suited to this study because it is based on factors (a), (b), and (e) redefined as explained above.

While factor (a), the entropic EV effect, is omnipresent in any solvent, its detailed mechanism is expected to depend on the solvent specificities. Factor (e) is crucial in water but absent in a nonpolar solvent. Unlike synthetic homo- and heteropolymers, a protein is characterized by side chains whose steric hindrance significantly constrains possible conformations achieved upon protein folding. It would then be quite insightful to be able to underpin the role of the solvent properties (water versus a nonpolar solvent), as well as the role played by the presence of side chains. In this paper, we pave the way along these lines by testing a total of 8 different models with different combinations of the above ingredients, investigating how they affect the stability of the native structure. In all cases, the geometric properties of the backbone and side chains are taken into account at atomistic level. In what follows, “HS” in the model name denotes that the solvent is nonpolar and formed by neutral hard spheres whose particle diameter and bulk number density are set at those of water at 298 K and 1 atm, and “WT” denotes that the solvent is water. “HB” represents that factor (b) is incorporated for a protein and factor (e) is incorporated for water. When factors (b) and (e) are both neglected, “HB” is omitted. “SC” signifies that the side chains are fully accounted for, while the omission of “SC” means that every residue is replaced by glycine (i.e., the side chains are removed) to highlight the contribution solely from the backbone to the protein structural stability. The 8 models are the following (also see Fig. 1):

Model HS. A protein, whose side chains are removed, is modeled as a set of fused, neutral hard spheres and immersed in the hard-sphere solvent.

Model WT. The hard-sphere solvent in model HS is replaced by water possessing W-W H-bonds.
Model HS-SC. A protein with the backbone and side chains is modeled as a set of fused hard spheres and immersed in the hard-sphere solvent. Comparing the results from models HS and HS-SC, we can analyze the effect of side chains on the protein structural stability in the hard-sphere solvent.

Model WT-SC. The hard-sphere solvent in model HS-SC is replaced by water possessing W-W H-bonds. In models HS, WT, HS-SC, and WT-SC, only factor (a) is taken into consideration and the analysis is focused on the entropic EV effect.

Model HS-HB. A protein, whose side chains are removed, is immersed in the hard-sphere solvent as in model HS, but P-P H-bonds described by factor (b) are taken into account.

Model WT-HB. The hard-sphere solvent in model HS-HB is replaced by water and thereby P-P, P-W, and W-W H-bonds described by factors (b) and (e) are also taken into account.

Model HS-SC-HB. A protein with the backbone and side chains is immersed in the hard-sphere solvent as in model HS-SC, but P-P H-bonds described by factor (b) are taken into account.

Model WT-SC-HB. The hard-sphere solvent in model HS-SC-HB is replaced by water and thereby P-P, P-W, and W-W H-bonds described by factors (b) and (e) are also taken into account. This model is the most realistic.

In our fused-sphere protein model, the protein structure is characterized by the \((x, y, z)\) coordinates of centers and the diameters of all of the protein atoms including hydrogen atoms. The 20 amino-acid residues are distinguished in the atomic detail and the sequence information is fully taken into account.

II. MODEL AND THEORY

A. Entropic excluded-volume effect

As explained in the third paragraph of Introduction, protein folding is crucially driven by the entropic EV effect as illustrated in Fig. 2. The presence of the backbone generates an EV. The formation of \(\alpha\)-helix by a portion of the backbone [Fig. 2(a)] or that of \(\beta\)-sheet by portions of the backbone [Fig. 2(b)] leads to a reduction of the total EV followed by a corresponding gain of solvent entropy.\(^8,9,14,15,21\) The
presence of a side chain also generates an EV. The close packing of side chains [Fig. 2(c)] leads to a reduction of the total EV and a solvent-entropy gain. The entropic EV effect is incorporated in the FEF to its full extent.

B. Free-energy function

Kinoshita and co-workers have developed the free-energy function $F$ for a protein. It is defined by

$$F(k_BT_0) = (Λ − TS)/(k_BT_0) = Λ/(k_BT_0) − S/k_B,$$

$$T = T_0 = 298 \text{ K}.$$

Here, $T$ is the absolute temperature and set at $T_0$, $k_B$ is the Boltzmann constant, $Λ$ is the sum of protein intramolecular energy and solvation energy, and $S$ is the solvation entropy calculated under the isochoric condition. When $F$ is applied to only quite compact structures, the protein conformational entropy need not be incorporated in $F$. $Λ/(k_BT_0)$ and $S/k_B$ are energetic and entropic components, respectively. $Λ$, $S$, and $F$ are calculated for a fixed structure and they vary to a large extent from structure to structure. $F$ is also dependent on the model employed. For instance, when “HB” is omitted in the name of a model, $Λ = 0$ and $F = −T_0S$. $ΔZ$ ($Z = −S/k_B$, $F(k_BT_0)$) is defined as “$Z$ for a decoy structure” minus “$Z$ for the native structure.”

$$ΔZ = Z(\text{a decoy structure}) − Z(\text{the native structure});$$

$$Z = −S/k_B, F(k_BT_0).$$

Negative $ΔZ$ for a decoy structure implies that the decoy structure is more stable than the native one within the framework of the model employed.

C. Calculation of entropic component

We adopt a molecular model for water. A water molecule is modeled as a hard sphere with diameter $d_S = 0.28$ nm ($ρ_Sd_S^3 = 0.7317$; $ρ_S$ is the bulk number density) in which a point dipole and a point quadrupole of tetrahedral symmetry are embedded. For a protein with a prescribed structure, $S$ is calculated by a hybrid method combining the integral equation theory (IET) and the morphometric approach (MA). The IET for the hard-sphere solvent is the radial-symmetric version and that for water is the angle-dependent version explicitly accounting for the dependence of a correlation function on the orientations of water molecules. The influence of molecular polarizability of water is taken into account by employing the self-consistent mean field theory. Therefore, the protein can be modeled as a set of fused, neutral hard spheres, in which case the diameter of each protein atom is set at the corresponding value of a Lennard-Jones potential parameter $σ$.

The geometric characteristics of the polyatomic structure can be taken into account by only its four geometric measures with sufficient accuracy, which is the basic idea of the MA. The four measures are the EV denoted by $V_{ex}$, solvent-accessible surface area $A$, and integrated mean and Gaussian curvatures of the accessible surface, $X$ and $Y$, respectively. $S$ is expressed as the linear combination of the four measures referred to as “morphometric form”

$$S/k_B = C_1V_{ex} + C_2A + C_3X + C_4Y.$$

The four coefficients ($C_1–C_4$), which are dependent only on the solvent species (e.g., the hard-sphere solvent or water) and its thermodynamic state, are determined beforehand for the simplest geometries: isolated hard-sphere solutes with various diameters. The solute-water correlations are both translational and orientational. The orientational contribution is also included in $S$ when the angle-dependent integral equation theory (IET) is employed, though it is much smaller than the translational contribution. The four coefficients for water are somewhat different from those for the hard-sphere solvent. This difference reflects only the effect of water-water (W-W) H-bonds on the translational contribution because the particle diameter and bulk number density of the hard-sphere solvent are set at those of water at 298 K and 1 atm.

The calculation comprises the following steps:

1. Calculate $S$ of an isolated hard-sphere solute ($S_{IHSS}$) with diameter $d_U$ using the IET. Consider different values of $d_U$ in the range, $0.6 \leq d_U/d_S \leq 10$, to obtain a sufficiently large set of data for $S$ and $d_U$.
2. Determine $C_1–C_4$ by applying the least-squares method to the following equation:

$$S_{IHSS}/k_B = C_1(4πR^3/3) + C_2(4πR^2) + C_3(4πR) + C_4(4π), R = (d_U + d_S)/2.$$
Equation (4) is the morphometric form for isolated spherical solutes.

(3) Calculate $V_{ex}$, $A$, $X$, and $Y$ of the protein using an extended version of Connolly’s algorithm. The input data are the $(x, y, z)$ coordinates of the center of each protein atom and its diameter $D$. $D$ is set at $\sigma$ taken from the CHARMM22 parameters.

(4) Calculate $S$ of the protein from Eq. (3) to which $C_1-C_4$ determined in step (2) are substituted.

The high accuracy of the MA has been demonstrated in earlier studies of Kinoshita and co-workers.9,26

D. Calculation of energetic component

First, we discuss the water case [see Fig. 3(a)]. The calculation of $\Lambda$ is performed by choosing a fully extended structure as the reference. The fully extended structure possesses the maximum number of P-W H-bonds but no P-P H-bonds. Compared to the fully extended structure with $\Lambda = 0$, in a compact structure significantly many donors and acceptors for P-P H-bonds (nitrogen and oxygen atoms) are buried in the interior after the break of P-W H-bonds. The burial of a donor and an acceptor followed by the formation of a P-P H-bond is expressed by, for example, CO···W + NH···W → CO···HN + W···W where W denotes a water molecule. It can be assumed that all of the four H-bonds in this reaction share the same energy. Namely, when a donor and an acceptor are buried but a P-P H-bond is formed, there is no energy change. The effect of the structural reorganization of water is implicitly taken into account by introducing “W···W” to the reaction. On the other hand, when a donor or an acceptor is buried and no P-P H-bond is formed, an energy increase of $E$ is imposed. From the thermodynamic cycle illustrated in Fig. 3(a), the formation of a P-P H-bond in the interior accompanies an energy decrease of $-2E$.

What value should $-2E$ be set at? The value of $E$ calculated using quantum chemistry for the formation of an H-bond in gas phase is $-10k_BT_0$ ($T_0 = 298$ K).33 If a donor-acceptor pair was in vacuum, $2E$ could be set at $10k_BT_0$. However, it is in the environment where atoms with positive and negative partial charges are present. $2E$ can be regarded as the potential of mean force between the pair in such environment. In this respect, $2E$ should be significantly smaller than $10k_BT_0$ (factor 1). On the other hand, protein folding is accompanied by the lowering of interaction energy, but this lowering arises from not only the intramolecular hydrogen bonding but also the other electrostatic interaction. When this lowering is incorporated in $2E$, the resulting value should be much larger than $10k_BT_0$ (factor 2). Further, the free-energy decrease brought by the formation of an H-bond between two formamide molecules in a nonpolar solvent, which is considered to mimic the environment within protein interior, was calculated to be $-14k_BT_0$: $2E = 14k_BT_0$.34 This value implicitly includes the contribution from entropic gain of the nonpolar solvent (i.e., that of protein atoms) arising from the donor-acceptor contact forming an H-bond. Following our earlier studies that were successful in solving several important problems,8,15,18,19 we set $2E$ at $14k_BT_0$. It has been verified that the qualitative aspects of our conclusions are not likely to be altered by the uncertainty of this value.

The water-accessible surface areas of all the donors and acceptors in the backbone and side chains are calculated using Connolly’s algorithm30,31 (the TINKER program package35 is employed). When a donor or an acceptor has a water-accessible surface area that is smaller than 0.001 Å², it is considered that the donor or acceptor is buried and the water oxygen-donor or -acceptor H-bond is broken. Otherwise, the water oxygen-donor or -acceptor H-bond is maintained. To determine if a P-P H-bond is formed or not, we use the criteria proposed by McDonald and Thornton.36 In the criteria, it is considered to be formed when all of the following conditions are satisfied: The distance between centers of D and A (D is a donor and A is an acceptor) is shorter than 3.9 Å; the distance between centers of the hydrogen atom (H) and A is shorter than 2.5 Å; and the angle formed by centers of D–H···A is larger than 90°. We examine the donors and acceptors not only for backbone-backbone but also for backbone-side chain and side chain-side chain P-P H-bonds for calculating $\Lambda$.

Second, we discuss the hard-sphere solvent case [see Fig. 3(b)]. The protein forms no H-bonds with the solvent. The formation of a P-P H-bond, regardless that it is buried or not, surely leads to an energy decrease of $-2E$. Hence, $\Lambda$ can be calculated simply by counting the number of P-P H-bonds in the protein with a prescribed structure. Note that the values of $\Lambda$ in water and in the hard-sphere solvent are positive and negative, respectively. The formation of a P-P H-bond is examined using the criteria proposed by McDonald and Thornton.36

E. Preparation of the native and decoy structures

For the native structure of protein G, we adopt the most recent data obtained by an NMR experiment (PDB CODE: 3GB1).37 A total of 32 structures (structural models 1–32) are registered. Their secondary structures are quite well converged and only the loop portions slightly differ from model to model. The root mean square deviation (RMSD) from model 1 (the $\alpha$-helix and $\beta$-sheet contents are 27% and 39%),
respectively) in terms of the C$_\alpha$ atoms is in the narrow range 0.19–0.34 Å. The secondary-structure contents are determined by the DSSP program. A thermodynamic quantity of the native structure is set at the average of 32 values calculated for structural models 1–32. When model WT-SC-HB is employed, for example, $-S/k_B = 2140$ and $F/(k_BT_0) = 2161$ for model 1 and their average values are $-S/k_B = 2134$ and $F/(k_BT_0) = 2165$.

We generate significantly many decoy structures using a recently developed methodology called “3Drobot.” This is a computer program that generates an arbitrary number of compact decoy structures of a protein from an input structure. 3Drobot is an extension of the fragment assembly simulation protocol which starts from multiple structure scaffolds identified from the input structure. Unlike the other existing protocols, 3Drobot does not scarify the P-P H-bonds and the compactness of the input structure. In this study, model 1 in 3GB1 is used as the input structure and the RMSD cutoff for the output decoy structures is set at 15 Å: Their values of RMSD from the input structure are smaller than 15 Å. We then obtain 1000 decoy structures (structures of type A). Five representative structures are shown in Fig. 4(a). The 3Drobot tends to yield the structures whose α-helix and β-sheet contents are not far from those of the input native structure: The content of α-helix is in the range 12.5%–37.5% and that of β-sheet is in the range 0%–54%. The structures generated include those which are quite similar to the native structure. This is understandable because the structures from the 3Drobot are intended for checking the ability of a FEF or a force field in discriminating the native structure from the decoy structures.

To assure a wider variety of α-helix and β-sheet contents, we construct additional decoy structures in the following manner. Starting from the homo-polymer chain obtained using a coarse-grained model, which possesses the relative contents of α-helix and β-sheet characteristic of each of the categories discussed below, we first try to achieve a maximally compact configuration both by increasing the range of attraction and by minimizing the gyration radius $R_g$ with a suitably devised annealing scheme. We then re-insert the atomistic content of the backbone and side chains using the Phyre2 backbone reconstruction tool and the Scwrl4 side-chain reconstruction tool. Finally, we perform a short MD run for the NPT ensemble in explicit water using the GROMACS software package. We categorize the structures thus constructed into five types with respect to the α-helix and β-sheet contents. Types B–F are shown in Figs. 4(b)–4(f), respectively. In type B, the contents of (α-helix, β-sheet) of the five structures are (25%, 32%), (23%, 25%), (25%, 32%), (25%, 36%), and (27%, 14%), respectively. The content of secondary structures is significantly high and in the range 41%–61%. Structures of type C possess only the α-helix and its content is in the range 46%–70%. Those of type D also possess only the α-helix and its content is in the range 32%–39%. Those of type E possess only the β-sheet and its content is in the range 11%–29%. The content of secondary structures in those of type F is significantly low and in the range 0%–21%. Types A through E are written as “type A-3Drobot,” “type B-α–β,” “type C-α-higher,” “type D-α-lower,” and “type E-β,” respectively. Structures of type F are fairly compact (see the next paragraph).

Further, very compact structures are generated using an MD simulation in vacuum. In this simulation, the initial structure is an extended one and only the protein intramolecular vdW interaction energy and the bonded energy are incorporated. Four structures are chosen and tested (they are referred to as “structures of type G”). Their secondary-structure contents are only 0%, 0%, 7%, and 14%, respectively. One of them is shown in Fig. 4(g). The average values of $R_g$ and EV are 10.8 Å and 12 049 Å$^2$ for the native structure, 10.3 Å and 12 119 Å$^2$ for structures of type F, and 10.0 Å and 11 442 Å$^2$ for structures of type G, respectively. Structures of type G are even more compact than those of type F. Types F and G are written as “type F-compact-less” and “type G-compact-more,” respectively. We also prepare a complete α-helix (a single helix), “all-α,” and associated two α-helices, “all-α-2,” which are shown in Figs. 4(h) and 4(i), respectively.

The slight, unrealistic overlaps of the constituent atoms occurring in a protein structure are removed by the local minimization of the energy function using the CHARMM and MMTSB programs. They are based on the CHARMM22 parameters combined with the CMAP correction and the GBMV implicit solvent model.

FIG. 4. Representation of decoy structures in ribbon models. (a) Five of 1000 structures of type A-3Drobot generated using the 3Drobot. The first structure is very close to the native structure (these two structures are compared in Fig. 11). (b) Five structures of type B-α–β. (c) Six structures of type C-α-higher. (d) Seven structures of type D-α-lower. (e) Four structures of type E-β. (f) Eight structures of type F-compact-less. (g) One of structures of type G-compact-more. (h) All-α (a single helix). (i) All-α-2 (associated two α-helices).
III. RESULTS AND DISCUSSION

A. Models HS and WT

In this section, we analyze the contribution from the backbone to the structural stability of a protein either in the hard-sphere solvent or in water. From each of the structures prepared as described above, the side chains are removed (i.e., every residue is replaced by glycine) and treated in the analysis. The backbone structure thus obtained from the native fold is referred to as “native structure.” Note that the FEF does not possess the energetic component ($\Lambda = 0$).

Figure 5(a) shows the relation between the RMSD from the native structure (model 1 in 3GB1$^{37}$) in terms of the $C_\alpha$ atoms and $-\Delta S/k_B$ for model HS where the solvent is the hard-sphere one. There are a number of structures that are more stable than the native structure. As observed in Fig. 5(b), there is a strong tendency that the stability becomes higher as the secondary-structure content increases. This result is consistent with Figs. 2(a) and 2(b) and suggests that the formation of $\alpha$-helix and $\beta$-sheet can lead to a significant reduction in the total EV. Among the backbone structures tested, the all-$\alpha$ (a single helix) is the most stable, and the all-$\alpha$-2 (associated two $\alpha$-helices) is also fairly stable, due to their very high secondary-structure contents.

Even when the solvent is replaced by water and model WT is employed, the qualitative aspects of the results do not significantly change. However, the differences among the structures in terms of the values of $-\Delta S/k_B$ are somewhat magnified as observed in Fig. 5(c) when compared to Fig. 5(b), which is indicative that the entropic EV effect depends on the protein structure more strongly in water than in the hard-sphere solvent.

B. Models HS-HB and WT-HB

Next, we additionally incorporate the effect of H-bonds as illustrated in Fig. 3. Figure 6(a) shows the relation between the RMSD and $\Delta F/(k_B T_0)$ for model HS-HB. There is a strong tendency that the stability becomes higher as the secondary-structure content increases [see Fig. 6(b)]. Except for structures of type C-$\alpha$-higher, the qualitative aspects of the results are not affected by the incorporation of P-P H-bonds. Structures of type C-$\alpha$-higher, which are characterized by a high $\alpha$-helix content, become more stable than the native structure despite their large RMSD values exceeding 10 Å. Again, the all-$\alpha$ is the most stable and the all-$\alpha$-2 is also considerably stable. Most of the donors and accepters participate in the formation of P-P H-bonds in the $\alpha$-helix, whereas there remain significantly many donors and accepters without forming P-P H-bonds in the $\beta$-sheet: The $\alpha$-helix is capable of forming more P-P H-bonds than the $\beta$-sheet. When the formation of P-P H-bonds is the most important, the $\alpha$-helix is more advantageous than the $\beta$-sheet.

We then replace the hard-sphere solvent by water and employ model WT-HB. Figure 7(a) shows the relation between the RMSD and $\Delta F/(k_B T_0)$. In water, the effect of breaking P-W
H-bonds (and recovering W-W H-bonds) comes into play. Further, the entropic EV effect on the structural stability becomes stronger. For these reasons, the formation of as many P-P H-bonds as possible is less essential in water. As a consequence, the result from model WT-HB is qualitatively similar to that from model WT [compare Fig. 7(b) to Fig. 5(c)]. An appreciable change resulting from the replacement of the hard-sphere solvent by water is that structures of type C-α-higher are no more stable than the native structure.

C. Models HS-SC and WT-SC

We analyze the entropic EV effect on the structural stability of a protein with the backbone and side chains. Note that the FEF does not possess the energetic component ($\Lambda = 0$). For model HS-SC, $-\Delta S/k_B$ is related to the RMSD as shown in Fig. 8(a). Comparing Fig. 8(a) with Fig. 5(a), we notice that the result exhibits a drastic change when the side chains are incorporated in the model. The all-α and the all-α-2 are not stable and the tendency that the stability becomes higher as the secondary-structure content increases is not observed [see Fig. 8(b)]. The very compact structures, structures of type G-compact-less, are the most stable. One of the compact structures in type F-compact-less is also significantly stable. It is thus suggested that the solvent-entropy gain originating from the close packing of side chains is substantially larger than that brought by the formation of secondary structures of the backbone. Still, there are several structures which are more stable than the native structure.

We then replace the hard-sphere solvent by water and employ model WT-SC. The qualitative aspects of the results do not significantly change. However, the differences among the structures in terms of the values of $-\Delta S/k_B$ are remarkably magnified [compare Fig. 8(c) with Fig. 8(b)]. The entropic EV effect depends on the protein structure more strongly in water than in the hard-sphere solvent, but this becomes more prominent when the side chains are incorporated in the model.

D. Model HS-SC-HB

In this model, not only the entropic EV effect in the hard-sphere solvent but also P-P H-bonds described by factor (b) are taken into account. Figure 9(a) shows the relation between the RMSD and $\Delta F/(k_B T_0)$ for model HS-SC-HB. The number of the structures that are much more stable than the native structure decreases to a significant extent. There is a tendency that the stability becomes higher as the secondary-structure content increases [see Fig. 9(b)]. In particular, it is the increase in the α-helix content of the structure that tends to increase the stability, as is observed in Fig. 9(c). It is interesting that the all-α-2 is now the most stable among the structures tested. Two of the structures in type C-α-higher having large RMSD values and with considerably high α-helix contents are also significantly stable. The very compact structures of type G-compact-more are no longer stable.
FIG. 9. (a) $\Delta F/(k_B T_0)$ plotted against the RMSD from the native structure in terms of the $C_\alpha$ atoms for model HS-SC-HB. (b) $\Delta F/(k_B T_0)$ plotted against the secondary-structure content for model HS-SC-HB. (c) $\Delta F/(k_B T_0)$ plotted against the $\alpha$-helix content for model HS-SC-HB.

The close packing of side chains and the formation of as many P-P H-bonds as possible are both crucially important in this case. Let us compare Fig. 9(a) with Figs. 6(a) and 8(a). In Figs. 6(a) and 8(a), a few structures in type C-$\alpha$-higher are as stable as the all-$\alpha$-2. In Fig. 9(a), on the other hand, they are significantly less stable than the all-$\alpha$-2. Hence, the remarkably high stability of the all-$\alpha$-2 is attributed to both efficient side-chain packing and high content of P-P H-bonds in the backbone. It is also attributed to P-P H-bonds in the side chains between two $\alpha$-helices in this case. The all-$\alpha$, which lacks the side-chain packing, becomes less stable than the all-$\alpha$-2, whereas the type C-$\alpha$-higher are also significantly stable and expected to become progressively more stable upon increasing the number of amino acids.

Model HS-SC-HB corresponds to a protein in nonpolar environment. If the number of residues of a protein is sufficiently large (much larger than that of protein G), it should take a structure of several associated $\alpha$-helices which look similar to that of a membrane protein immersed in hydrocarbon groups constituting nonpolar chains of the lipid bilayer. The entropic EV effect that arises from the translational displacement of hydrocarbon groups is surely present. Further, our result should be relevant to the experimental observations evidencing that alcohol tends to induce a protein to form $\alpha$-helices. The power of inducing $\alpha$-helices is strengthened as the hydrocarbon group of an alcohol molecule becomes more bulky. Alcohol addition makes the environment less aqueous and more nonpolar. When the environment becomes sufficiently nonpolar, a protein is driven to change its structure to the one with high $\alpha$-helix content.

E. Model WT-SC-HB

We now discuss the result from the most realistic model for a protein in aqueous environment, model WT-SC-HB. Figure 10(a) shows the relation between the RMSD and $\Delta F/(k_B T_0)$. A few structures with small RMSD values in type A-3Drobot are slightly more stable than the native structure. However, they are essentially the same as the native structure. The most stable structure in type A-3Drobot is compared with model 1 of the native structure in Fig. 11. As observed in the figure, only the loop portions indicated by the solid arrows are slightly different and the secondary structures are indistinguishable. Therefore, it can be claimed that model WT-SC-HB is successful in identifying the native structure as the one with lowest $\Delta F$. As observed in Fig. 10(b), there is a tendency that...
the stability becomes higher as the secondary-structure content increases. However, when \( \Delta F/(k_B T_0) \) is plotted against the \( \alpha \)-helix content [see Fig. 10(c)], such a tendency is no more appreciated. This is consistent with the result that neither the all-\( \alpha \)-2 nor the all-\( \alpha \) is stable at all.

When the \( \alpha \)-helix or the \( \beta \)-sheet is formed, not only a water-entropy gain occurs but also the energetically unfavorable break of P-W H-bonds is compensated with the formation of P-P H-bonds. Hence, these secondary structures (\( \alpha \)-helix and \( \beta \)-sheet) are fundamental structural units to be formed as much as possible in protein folding. The packing of side chains is also critically required. Let us consider the close packing of side chains and the formation of as many P-P H-bonds as possible. In water, the relative importance of the former is significantly larger than in the hard-sphere solvent. To achieve sufficiently close packing of side chains favored by the water entropy, high \( \alpha \)-helix content is not necessarily suitable. There are many cases where the \( \beta \)-sheet should be chosen. Taken together, the structural stability is largely influenced by the water-entropy effect which makes the close packing of side chains particularly important. The contents of \( \alpha \)-helix, \( \beta \)-sheet, and sum of these contents (i.e., the content of secondary structures) are optimized in terms of the close side-chain packing and the formation of P-P H-bonds.

IV. CONCLUSIONS

Some of the important results are recapitulated as follows:

(I) The formation of \( \alpha \)-helix and \( \beta \)-sheet (secondary structures) as well as the close packing of side chains is accompanied by a gain of solvent entropy through the entropic EV effect. When only this effect is incorporated in the model, the protein atoms constituting the backbone and side chains are driven to be packed as closely as possible regardless of the secondary-structure content. The formation of \( \alpha \)-helices and \( \beta \)-sheets itself is not significant at all, and the structures like those of type G-compact-more are the most stabilized.

(II) The entropic EV effect is substantially larger in water than in the hard-sphere solvent (the entropy of water is more sensitive to the protein structure than that of the hard-sphere solvent). This is particularly true when the side chains are incorporated in the model.

(III) When the entropic EV effect and the energetic component related to H-bonds are both taken into account in the model, the secondary structures need to be formed as much as possible. In the hard-sphere solvent, the formation of \( \alpha \)-helices and \( \beta \)-sheets is strongly favored because it leads to a large decrease in energy due to the formation of protein intramolecular (P-P) H-bonds. In water, the formation is required to compensate for the energetic dehydration caused by the break of protein-water H-bonds. In both the hard-sphere solvent and water, the formation of secondary structures also leads to a solvent-entropy gain.

(IV) The entropic EV effect and the energetic component related to H-bonds are both essential. However, the former is relatively more important than the latter in water, whereas the opposite is true in the hard-sphere solvent.

(V) In the hard-sphere solvent, a structure with very high \( \alpha \)-helix content is most stabilized because the \( \alpha \)-helix is more advantageous than the \( \beta \)-sheet in terms of the number of P-P H-bonds formed. The entropic EV effect also acts, and a structure of associated \( \alpha \)-helices accompanying the side-chain packing and still retaining sufficiently many P-P H-bonds is most stabilized.

(VI) In water, the close packing of the backbone and side chains (in particular, the close side-chain packing) is strongly required by the water-entropy effect, and high \( \alpha \)-helix content is not necessarily suitable: It is often that the \( \beta \)-sheet is preferable to the \( \alpha \)-helix or the content of secondary structures needs to be reduced to a significant extent. The contents of \( \alpha \)-helix, \( \beta \)-sheet, and sum of these contents (i.e., the content of secondary structures) are optimized in terms of the close side-chain packing and the formation of P-P H-bonds. For protein G, the structure with \( \alpha \)-helix and \( \beta \)-sheet contents of 27% and 39%, respectively (i.e., the content of secondary structures is 66%), is the most stabilized.

Figure 12 represents a graphical summary of the above findings and clearly illustrates the fact that the native state of protein G is the most stable only when the model with the highest complexity (model WT-SC-HB) is employed [see panel (b)]. If water is replaced by a nonpolar fluid, a conformation of associated \( \alpha \)-helices [see panel (a)] is the most stable. Other misfolded decoys become more stable when different models are used: A very compact one with no secondary structures can be the most stable. Minimal multiple factors, whose complex interplay stabilizes the native state, are incorporated in model WT-SC-HB. In the absence of even a single of
The combined effect of the presence of side chains and the two physical factors discussed above. If the amino-acid sequences are arbitrarily constructed, only an extremely small percentage of them can achieve the close packing of side chains with sufficiently many P-P H-bonds that surpass the physical factors opposing to protein folding. The amino-acid sequences, which are capable of achieving them, are perpetuated by natural selection and sustaining life. We plan to extend the analysis of this study to proteins with native states inherently different from protein G (e.g., those possessing only β-sheets) in the near future.

ACKNOWLEDGMENTS

The computer program for the morphometric approach was developed with R. Roth and Y. Harano. This work was supported by Grant-in-Aid for Scientific Research (B) (No. 17H03663) from Japan Society for the Promotion of Science (JPSJ) to M.K. and by MIUR PRIN-COFIN2010-2011 (Contract No. 2010LKE4CC) to A.G. The use of the SCSCF multiprocessor cluster at the Università Ca’ Foscari Venezia is gratefully acknowledged.

FIG. 12. Cartoon illustrating the structure of protein G giving lowest value to the free-energy function (FEF), which is dependent on the model employed. (a) In HS-models, the solvent is formed by neutral hard spheres whose particle diameter and bulk number density are set at those pertinent to water. In WT-models, the solvent is water (a molecular model is employed). The backbone structures are obtained by removing the side chains from the configurations already prepared: Among these backbone structures, the single α-helix gives lowest value to the FEF.