THE JOURNAL OF PHYSICAL CHEMISTRY B

Article

pubs.acs.org/JPCB

Characterization of Experimentally Determined Native-Structure Models of a Protein Using Energetic and Entropic Components of Free-Energy Function

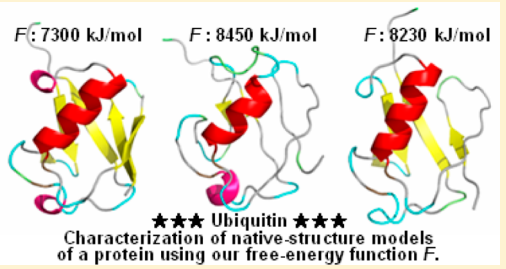
Hirokazu Mishima,[†] Satoshi Yasuda,[†] Takashi Yoshidome,[‡] Hiraku Oshima,[‡] Yuichi Harano,[§] Mitsunori Ikeguchi,[∥] and Masahiro Kinoshita*,[‡]

[†]Graduate School of Energy Science, and [‡]Institute of Advanced Energy, Kyoto University, Uji, Kyoto 611-0011, Japan

[§]Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

[∥]Graduate School of Nanobioscience, Yokohama City University, 1-7-29, Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan

ABSTRACT: We show how to characterize the native-structure models of a protein using our free-energy function F which is based on hydration thermodynamics. Ubiquitin is adopted as an example protein. We consider models determined by the X-ray crystallography and two types of NMR model sets. A model set of type 1 comprises candidate models for a fixed native structure, and that of type 2 forms an ensemble of structures representing the structural variability of the native state. In general, the X-ray models give lower F than the NMR models. There is a trend that, as a model deviates more from the model with the lowest F among the X-ray models, its F becomes higher. Model sets of type 1 and those of type 2, respectively, exhibit two different characteristics with respect to the correlation between the deviation and F. It is argued that the total amount of constraints such as NOEs effectively taken into account in constructing the NMR models can be examined by analyzing the behavior of F. We investigate structural characteristics of the models in terms of the energetic and entropic components of F which are relevant to intramolecular hydrogen bonding and to backbone and side-chain packing, respectively.



INTRODUCTION

The X-ray crystallography and nuclear magnetic resonance (NMR) are popular and important means of investigating the structures of biomolecules such as proteins. A protein folds into its unique native structure (NS) in aqueous solution under the physiological condition. Understanding the NS at the atomic level is essential because it is closely related to the protein function. The NS models are represented by the coordinates of the constituent atoms and registered in Protein Data Bank (PDB). The X-ray crystallography,[1−3] which has been employed for many years, can be performed in an almost established manner once the solvent condition leading to protein crystallization is found, whereas the NMR is still in a developing stage. A variety of solution-NMR approaches[4−6] have been devised to make the NMR applicable even to a large protein by improving its relaxation properties. Solid-state NMR spectroscopy[7,8] has also been proposed as a useful alternative to solution NMR. The NS models are constructed by a structure calculation upon which the structural information experimentally obtained as a set of constraints is imposed.[4−8] Typical constraints are the nuclear Overhauser effect (NOE), residual dipolar coupling (RDC), hydrogen bonding, and dihedral angle restraints. The structure calculation usually yields a number of candidate models which can be significantly different from one another. Thus, the NS models are substantially influenced by the experimental technique and the structure calculation employed.

A great advantage of solution NMR is that it provides insightful information on dynamical properties associated with backbone and side-chain mobilities which are crucially important in discussing the protein function. To elucidate such properties, the native state of a protein needs to be represented not as a single structure but as an ensemble of structures.[9−13] Such a structural ensemble is usually constructed by employing the ensemble refinement protocol[9−11,13] started from an NS model determined by the X-ray crystallography or solution NMR. The refinement is performed in a variety of manners using a computer simulation with all-atom potentials (e.g., an ensemble molecular dynamics (MD) simulation) or a method based on geometric restrictions, which is restrained by NOEs, RDCs, and order parameters S^2 obtained in NMR experiments. The order parameters contain atomic-detailed information about the amplitude of molecular motion. The restraints are enforced not on a single structure but on the average calculated over an ensemble of structures. The NS models thus obtained as a structural ensemble represent accessible structures of a protein in aqueous solution rather than candidate models for a fixed NS. The ensembles are diversified by the details of the ensemble refinement protocol and the restraints effectively taken into consideration.

Received: February 16, 2012
Revised: June 13, 2012

   dx.doi.org/10.1021/jp301541z | J. Phys. Chem. B XXXX, XXX, XXX−XXX

It is not rare that there are many NS models for the same protein: candidate models for a fixed NS and ensembles which represent fluctuating structures of the native state in aqueous solution. Nevertheless, the characteristics of all those NS models are often ambiguous. It is strongly desired that a reliable method, which is based on energetics of a protein accounting for its hydration thermodynamics, be developed for characterizing the NS models. In the present study, we show how to characterize the NS models using our recently developed free-energy function[14−16] F and its energetic and entropic components, Λ and S. Both Λ and $-S$ are positive quantities. Λ is a measure of the assurance of intramolecular hydrogen bonds in the case where donors and acceptors are buried in the interior after the break of hydrogen bonds with water molecules. $-S$ represents the efficiency of backbone and side-chain packing. F has been demonstrated to be far superior to any of the previously reported functions in terms of the performance of discriminating the native fold from misfolded decoys. Ubiquitin with 76 residues is chosen as an example protein. It is known to be stable over a wide range of pH and temperature.[1] It has frequently been adopted as a benchmark protein for which a new NMR approach is illustrated.[4−8] Thus, a comprehensive range of high-quality X-ray and NMR data of ubiquitin are available. For these reasons, ubiquitin is best suited to the present study.

We consider three X-ray models (PDB codes: 1ubq,[1] 1ubi,[2] and 3n32[3]) and a number of models in three solution NMR data (PDB codes: 1d3z,[4] 1g6j,[5] and 2klg[6]) and in two solid-state NMR data (PDB codes: 2jzz[7] and 2l3z[8]) which were obtained as candidate models for a fixed NS. The model sets in 1d3z, 1g6j, 2klg, 2jzz, and 2l3z are categorized as "model sets of type 1". Five ensembles which were constructed for representing the structural variability of the native state (PDB codes: 1xqq,[9] 2nr2,[10] 2k39,[11] 2kn5,[12] and 2kox[13]) are also considered. The model sets in 1xqq, 2nr2, 2k39, 2kn5, and 2kox are categorized as "model sets of type 2". The models belonging to the two types of model sets are referred to as "NMR models" because they were constructed such that the constraints obtained in NMR experiments were satisfied. It should be emphasized that only the models for isolated monomers are selected in the present study.

We propose a set of figures where F, Λ, S, and additional parameters judiciously defined are plotted. The X-ray models resemble one another with respect to structural properties (e.g., the root-mean-square deviation (rmsd) for C_α atoms), F, Λ, and S, whereas the NMR models exhibit diverse properties. In general, the X-ray models give lower F than the NMR models. There is an overall trend that, as a model deviates more from the reference model (the model with the lowest F among the X-ray models) with respect to structural properties, its F becomes higher. Model sets of type 1 and those of type 2, respectively, exhibit two apparently different characteristics in terms of the correlation between the deviation and F. On the whole, the average value of F, F_{av} (the subscript "av" denotes the average value in each model set), tends to become lower as the structures in a model set are better converged. Concerning the correlation between the structural convergence and F_{av}, two apparently different characteristics are exhibited by model sets of type 1 and those of type 2, respectively. There are model sets which depart from the correlation observed, but the physical reasoning for the departure can be made, thus uncovering their features. It is argued that F_{av} becomes lower with the increase in the total amount of constraints effectively incorporated in the structure calculation or the ensemble refinement protocol. Further, by examining Λ and S, we can clarify structural characteristics of each model set with respect to intramolecular hydrogen bonding and packing efficiency of the backbone and side chains.

We are successful in characterizing the NS models of ubiquitin, finding the properties of the X-ray and NMR models, and clarifying the overall features of the structures in a particular model set. It should be emphasized that the results described above are achievable only by a free-energy function like ours capturing essential physics of the structural stability of a protein in aqueous solution. The characterization method thus developed is expected to be useful for the following applications: the evaluation of a set of NS models of ubiquitin or any other protein determined via a new NMR approach by comparing them to the models which are already available; and refinement of an NMR model by rectifying its weak points found. Further, our free-energy function is well suited not only to the selection of the best model from among many candidate NMR models but also to the original construction of the best candidate model or a good structural ensemble for the native state, on the basis of the experimentally obtained constraints such as NOEs and RDCs.

FREE-ENERGY FUNCTION

Definition. We treat a number of different structures of a protein. Our free-energy function F is expressed for the protein with a prescribed structure by[14−16]

$$F = (E_I + \mu)/(k_B T_0), \qquad T_0 = 298\ \mathrm{K} \tag{1}$$

where E_I is the protein intramolecular energy, μ is the hydration free energy (i.e., excess chemical potential) that is the most important thermodynamic quantity of protein hydration, and k_B is the Boltzmann constant. We note that μ is the same irrespective of the protein insertion condition (isobaric or isochoric),[17] and we consider the isochoric condition that is much more convenient in a theoretical treatment. Using the relation

$$\mu = E - TS \tag{2}$$

where E is the hydration energy, S is the hydration entropy, and T is the absolute temperature and defining Λ by

$$\Lambda = E_I + E \tag{3}$$

we obtain

$$F = (\Lambda - TS)/(k_B T_0), \qquad T_0 = 298\ \mathrm{K} \tag{4}$$

Here T is set at T_0 in the present study. Λ is calculated by choosing a fully extended structure as the standard one and referred to as the total dehydration penalty (see "Energetic Component: Total Dehydration Penalty"). S represents a water-entropy loss upon the protein insertion. Both Λ and $-S$ are positive quantities. Λ, S, and F are substantially dependent on the protein structure. The procedures of calculating the entropic component S and the energetic component Λ are briefly described below (more details are given in our earlier publications[14−16]).

Entropic Component: Hydration Entropy. Unlike the previously reported functions where water was regarded as a dielectric continuum, a molecular model is employed in the calculation of S, which would require a heavy burden. However, the calculation is finished quite rapidly by combining the angle-dependent integral equation theory[17−23] (ADIET), a statistical-

dx.doi.org/10.1021/jp301541z | J. Phys. Chem. B XXXX, XXX, XXX−XXX

mechanical theory for molecular liquids, and the morphometric approach (MA).[24−26]

We employ a multipolar model for water:[19] A water is modeled as a hard sphere with diameter d_S = 0.28 nm in which a point dipole and a point quadrupole of tetrahedral symmetry are embedded. In the ADIET the effect of the molecular polarizability is taken into account using the self-consistent mean field (SCMF) theory.[18,19] At the SCMF level the many-body induced interactions are reduced to pairwise additive potentials involving an effective dipole moment. As proved in our earlier work,[23] the ADIET predominates over the reference interaction site model (RISM) and related theories.[27−29] For example, the hydration free energies of nonpolar solutes calculated by the ADIET with the multipolar model are in quantitatively excellent agreement with those from Monte Carlo computer simulations. The dielectric constant of bulk water calculated, which is a good measure of the validity of a theory, is ∼83 that is in good accord with the experimental value ∼78.

The hydration entropy under the isochoric condition is fairly insensitive to the solute-water interaction potential as proved in our earlier works.[30,31] For example, the three quantities, μ, S, and E, under the isochoric condition (T = 298 K) are calculated for a spherical solute with diameter 0.28 nm using the ADIET[17−23] combined with the multipolar water model.[19] For the hard-sphere solute with zero charge, the calculated values are $\mu = 5.95k_BT$, $S = -9.22k_B$, and $E = -3.27k_BT$. When the point charge $-0.5e$ (e is the electronic charge) is embedded at its center, the calculated values are $\mu = -32.32k_BT$, $S = -10.11k_B$, and $E = -42.43k_BT$. Thus, S is fairly insensitive to the solute-water interaction potential, whereas μ and E are largely influenced by it. Further, Imai et al.[30] considered a total of eight peptides and proteins and calculated S using the three-dimensional RISM theory combined with the all-atom potentials. Even when the protein−water electrostatic interactions, which are quite strong, are completely shut off and only the Lennard-Jones (LJ) interactions are retained, S changes only by less than 5%. Therefore, a protein can be modeled as a set of fused hard spheres just for calculating its hydration entropy. (The hydration energy, which is influenced by the protein-water interaction potential, is separately treated in the calculation of Λ.)

The idea of the MA is to express a hydration quantity such as S by the linear combination of only four geometric measures of a solute molecule:[25,26]

$$S/k_B = C_1V_{ex} + C_2A + C_3X + C_4Y \quad (5)$$

Here, eq 5 is referred to as the morphometric form, V_{ex} is the excluded volume, A is the water-accessible surface area, and X and Y are the integrated mean and Gaussian curvatures of the accessible surface, respectively. The water-accessible surface is the surface that is accessible to the centers of water molecules. The volume that is enclosed by this surface is the excluded volume. We note that C_1 is completely independent of the solute−water interaction potential. Though S is influenced by all four terms, C_1V_{ex} is the principal term. This is the reason for the fair insensitivity of S to the solute−water interaction potential. In the MA, the solute shape enters S only via the four geometric measures. Therefore, the four coefficients (C_1−C_4) can be determined in simple geometries. They are calculated from the values of S for hard-sphere solutes with various diameters immersed in our model water. The ADIET[17−23] is employed in the calculation.

The procedure of calculating S of a protein with a prescribed structure comprises the following four steps.

(1) S of a hard-sphere solute with diameter d_U is calculated using the ADIET. The values of S are prepared for sufficiently many different values of d_U ($0 \le d_U \le 10d_S$; d_S is the molecular diameter of water; changing $10d_S$ to $30d_S$, for example, leads to no changes in the result from step (2).)

(2) The four coefficients are determined by the least-squares fitting applied to the following equation for hard-sphere solutes (i.e., eq 5 applied to hard-sphere solutes):

$$S/k_B = C_1(4\pi R^3/3) + C_2(4\pi R^2) + C_3(4\pi R) + C_4(4\pi), \quad R = (d_U + d_S)/2 \quad (6)$$

The most recent method of the least-squares fitting is described in ref 26. The values of the four coefficients thus obtained are the following: $C_1 = -0.1971$ Å^{-3}, $C_2 = 0.06119$ Å^{-2}, $4\pi C_3 = 1.967$ Å^{-1}, and $4\pi C_4 = -2.652$.

(3) The four geometric measures of a protein (V_{ex}, A, X, and Y) with a prescribed structure are calculated by means of an extension[25] of Connolly's algorithm.[32,33] The x−y−z coordinates of the protein atoms are used as part of the input data to account for the polyatomic structure at the atomic level. The diameter of each atom is set at the sigma value of the LJ potential parameters which are taken from the CHARMM22.[34]

(4) S of a protein with a prescribed structure is obtained from eq 5 in which the four coefficients determined in step (2) are used. It should be emphasized that the computation time required for step (4) is only ∼0.1 s on our workstation.

The high reliability of the hybrid of ADIET and MA in calculating S has been demonstrated in the following examples: quantitative reproduction of the experimentally measured changes in thermodynamic quantities upon apoplastocyanin (apoPC) folding;[35] elucidation of the molecular mechanisms of pressure[36] and cold[37,38] denaturating of proteins; and proposal of a reliable measure of the thermal stability of proteins.[39,40]

Energetic Component: Total Dehydration Penalty. Λ defined by eq 3 is calculated in accordance with a simple manner which still accounts for physically the most important factors: intramolecular and protein-water hydrogen bonds. The time required in the calculation of Λ is only ∼0.1 s per protein structure on our workstation. Λ is calculated by choosing a fully extended structure as the standard one. The fully extended structure possesses the maximum number of hydrogen bonds with water molecules but no intramolecular hydrogen bonds. Consequently, Λ corresponds to the total dehydration penalty occurring upon the transition to a more compact structure. Compared to the fully extended structure with Λ = 0, in a more compact structure some donors and acceptors (e.g., N and O, respectively) are buried in the interior after the break of hydrogen bonds with water molecules (CO···W, NH···W, etc.). There is no problem if the intramolecular hydrogen bonds (CO···HN, etc.) are formed. However, such hydrogen bonds are not always formed, giving rise to an energetic penalty and positive Λ. We note that this picture is consistent with the experimental result by Terazima et al.,[35] who reported a significantly large, positive change in enthalpy upon apoPC folding.

dx.doi.org/10.1021/jp301541z | J. Phys. Chem. B XXXX, XXX, XXX−XXX

Our procedure of calculating Λ can be summarized as follows. When a donor and an acceptor are buried in the interior after the break of hydrogen bonds with water molecules, if they form an intramolecular hydrogen bond, we impose no penalty. On the other hand, when a donor or an acceptor is buried with no intramolecular hydrogen bond formed, we impose the penalty of $7k_BT_0$ (T_0 = 298 K). The value, $7k_BT_0$, is based on the energy decrease of $-14k_BT_0$ arising from hydrogen-bond formation between two formamide molecules in a nonpolar liquid.[41] We have to determine whether or not each of the donors and acceptors is buried. The water-accessible surface area is calculated for each of them by means of Connolly's algorithm.[32,33] If it is smaller than 0.001 Å^2, the donor or acceptor is considered buried. We examine all of the donors and acceptors for backbone−backbone, backbone−side chain, and side chain−side chain intramolecular hydrogen bonds and determine if each of the donors and acceptors is buried or not.

The energetic component is not considered for nonpolar groups. This is justifiable because the break of hydrogen bonds with water molecules, when they are not compensated by the intramolecular hydrogen bonds, should be more serious and form a principal component of the total dehydration penalty. The torsion energy is not considered, either. This can also be justified for the following reason: The structures to be treated share the property that the torsion energy is reasonably low (i.e., the structures with unreasonably high torsion energies are not included), and the difference between two structures in the torsion energy makes no essential contribution to the difference in the energetic component.

Performance of Discriminating Native Fold from Misfolded Decoys. We have examined the performance of our free-energy function F in discriminating the native fold from a number of misfolded decoys.[15,16] The examination is carried out for a total of 133 proteins in 8 decoy sets. F is shown to be far superior to any of the previously reported functions. When the NS model is determined by the X-ray crystallography, the discrimination is always successful. In the case of NMR models, as long as a sufficiently good NS model is included in the candidate models, the discrimination is accomplished with 100% accuracy. Thus, the approximations employed in calculating S, Λ, and F can be justified by this success. F captures the features of the NS of a protein such that it is optimized in terms of the sum of the hydration entropy and the total dehydration penalty.

NATIVE-STRUCTURE MODELS CONSIDERED

We consider three X-ray models (PDB codes: 1ubq,[1] 1ubi,[2] and 3n32[3]) and a number of models in three solution NMR data (PDB codes: 1d3z,[4] 1g6j,[5] and 2klg[6]) and two solid-state NMR data (PDB codes: 2jzz[7] and 2l3z[8]) which were obtained as candidate models for a fixed NS. The numbers of the models registered in 1d3z, 1g6j, 2klg, 2jzz, and 2l3z are 10, 32, 20, 20, and 20, respectively. We also consider five sets of models constructed for representing the native state which comprises an ensemble of structures (PDB codes: 1xqq,[9] 2nr2,[10] 2k39,[11] 2kn5,[12] and 2kox[13]). A total of 128, 144, 116, 50, and 640 models are registered in 1xqq, 2nr2, 2k39, 2kn5, and 2kox, respectively. The model sets in 1d3z, 1g6j, 2klg, 2jzz, and 2l3z are categorized as "model sets of type 1". Those in 1xqq, 2nr2, 2k39, 2kn5, and 2kox are categorized as "model sets of type 2".

Three Models Obtained from X-ray Crystallographic Experiments. 1ubq is a model of natural human ubiquitin (hUb) refined at 1.8 Å resolution using a restrained least-squares procedure,[1] while 1ubi is a model of synthetic ubiquitin refined with the coordinates of 1ubq as the starting point of the refinement using the same procedure against the synthetic X-ray data.[2] 3n32 is a model of ubiquitin to which platinum ions bind (Pt_3-hUb)[3] (Pt is omitted in our calculations).

Five Sets of Candidate Models Obtained from NMR Experiments. The model sets of 1d3z, 1g6j, 2klg, 2jzz, and 2l3z give candidate models for a fixed NS. 1d3z was obtained from a structure calculation on the basis of 2727 NOEs, 98 dihedral angle constraints, and 372 RDC restraints.[4] 1g6j was obtained from a structure calculation based on 1291 NOEs, 63 dihedral angle constraints, and 23 hydrogen bonding restraints.[5] A distinctive aspect of 1g6j is the encapsulation of the protein within a reverse micelle. The encapsulation was undertaken for the purpose of improving the relaxation properties even for a large protein. 2klg was obtained via an NMR experiment in which inert paramagnetic molecules were added so that a rather small number of NOEs could be complemented with the restraints from paramagnetic relaxation enhancements (PREs).[6] 2jzz and 2l3z were the products from solid-state NMR experiments intended for obtaining long-range distance information, which was suitable to a large, insoluble protein.[7,8] It should be noted that the environment is considerably different from that in solution NMR. The structure calculation is usually performed using standard methods for generating protein structures satisfying geometric restrictions determined by solution or solid-state NMR. The most popular method is the CYANA program[42] of an MD simulation with the idea of simulated annealing which uses torsion angles instead of Cartesian coordinates as the degrees of freedom.

Five Ensembles of Structures Constructed for Representing Structural Variability of Native State. The models in 1xqq, 2nr2, and 2kox were constructed using the ensemble refinement protocol started from the model in 1ubq.[9,10,13] The ensemble refinement protocol was also employed in 2k39, but the initial structures were generated by amending random coils using the CONCOORD method[43] based on the geometric restrictions of NOEs.[11] The structural ensemble was constructed on the basis of such constraints as NOEs, RDCs, and order parameters S^2 obtained in previously reported NMR experiments.[9−11,13] The construction of the ensemble was made to represent structural fluctuations rather than candidate structures for a fixed NS with the emphasis on protein dynamics in aqueous solution. The ensemble was obtained by employing an all-atom MD simulation with explicit water in 1xqq, 2k39, or 2kox, whereas a hybrid-type approach of MD simulations with explicit water and in vacuum was adopted in 2nr2. The approach in 2kn5 first used unrestrained structural sampling with a Monte Carlo protocol and the Backrub motional model[44] to generate a large set of structures, starting from the model in 1ubq.[12] It then selected an ensemble optimizing the agreement with RDCs. Due to the particular method employed in 2kn5, the rmsd for C_α atoms from the model in 1ubq (and from that in 1ubi) becomes intrinsically small.

Slight Modification of Native-Structure Models. The coordinates of hydrogen atoms cannot be obtained by the X-ray diffraction. We give hydrogen atoms to each model using the CHARMM biomolecular simulation program[45] through the Multiscale Modeling Tools in Structural Biology (MMTSB) program.[46]

dx.doi.org/10.1021/jp301541z | J. Phys. Chem. B XXXX, XXX, XXX−XXX

The LJ potential energy for many of the NS models is positive and quite large due to unrealistic overlaps of protein atoms. Such overlaps are removed by the minimization of the energy function using the CHARMM and MMTSB programs. The minimization is performed so that the original structures can be retained as much as possible. We employ the CHARMM22[34] with the CMAP correction[47] as the force-field parameters. Electrostatic and nonbonded interactions are all evaluated without any cutoff. The generalized-Born (GBMV/SA) approximation[48−50] is employed for the electrostatic part of the hydration free energy. After the minimization, there are no unrealistic overlaps of protein atoms. Moreover, it is verified for each structure that the rmsd for C_α atoms from the structure before the minimization is quite small: 0.5−0.7 Å in 2kn5, 0.1−0.3 Å in 1g6j, 2jzz, and 2l3z, and 0.03−0.06 Å in the others. Each structure is then switched to a set of fused hard spheres in calculating the hydration entropy.

RESULTS AND DISCUSSION

Free-Energy Function Plotted against rmsd or TM-Score. Figure 1 shows a ribbon representation of the X-ray model in 1ubq and NMR models with the highest and lowest values of F in 1d3z, 2klg, 2k39, 2kox, and 2l3z. The three X-ray models, which are almost the same, possess three and one-half turns of α-helix, two short pieces of 3_{10}-helix, a mixed β-sheet containing five strands, and seven reverse turns.[1−3] It is apparent that the secondary structures of the models in 2klg are incomplete. Even the models looking almost indistinguishable can be considerably different from one another in terms of F, S, and Λ (i.e., the details of the secondary structures, backbone and side-chain packing, and intramolecular hydrogen bonding).

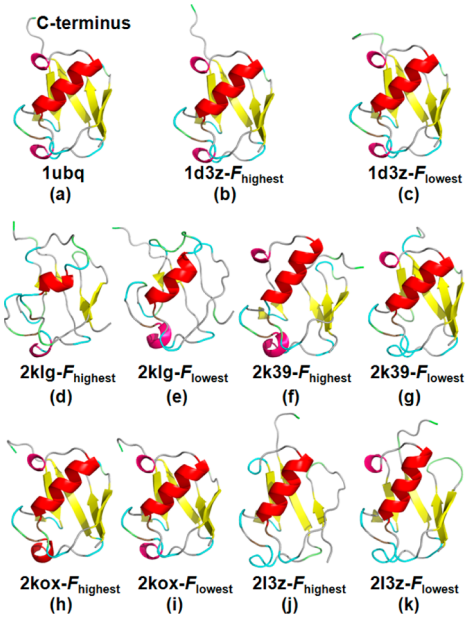


Figure 1. Ribbon representation of an X-ray model and NMR models with the highest and lowest values of F in representative model sets. The models with the highest and lowest values of F in 1d3z, for instance, are denoted by 1d3z-$F_{highest}$ and 1d3z-F_{lowest}, respectively. (a) 1ubq, (b) 1d3z-$F_{highest}$, (c) 1d3z-F_{lowest}, (d) 2klg-$F_{highest}$, (e) 2klg-F_{lowest}, (f) 2k39-$F_{highest}$, (g) 2k39-F_{lowest}, (h) 2kox-$F_{highest}$, (i) 2kox-F_{lowest}, (j) 2l3z-$F_{highest}$, and (k) 2l3z-F_{lowest}. Residues 71−76 on the C-terminus side are significantly flexible. This figure was drawn using PyMol 1.3.

The values of our free-energy function F for the three X-ray models follow the order, 1ubi ≈ 3n32 < 1ubq though the differences are quite small. In general, the X-ray models give lower F than the NMR models: Only 14 models in 2kox and 1 model in 1xqq give lower F than the X-ray model in 1ubq. The model with the lowest F is in 2kox. The X-ray model in 1ubi, whose F is the second lowest, is regarded as the reference model. We consider the template modeling score (TM-score)[51] as well as the rmsd calculated for C_α atoms to look at the similarity between two protein structures. The TM-score indicates the structural difference by a score in the range (0, 1]. The score 1 implies a perfect match of the structures. It can be assumed that a score higher than 0.5 implies significantly high similarity. Unlike the rmsd, the TM-score is more sensitive to the global topology than local variations.

Figures 2 and 3 show the relation between $(F - F_{1ubi})_{av}$ (the subscripts "av" and "1ubi" denote the average value in each model set and the value for the reference model, respectively) and the rmsd or TM-score calculated by choosing the reference model as the standard structure. The standard deviation of $F - F_{1ubi}$, $(F - F_{1ubi})_{sd}$ (the subscript "sd" denotes the standard deviation), and that of the rmsd or TM-score are indicated as error bars. In the calculation of the TM-score and rmsd, all residues (1−76) are considered in Figure 2, whereas only the core region comprising residues 1−70 are considered in Figure

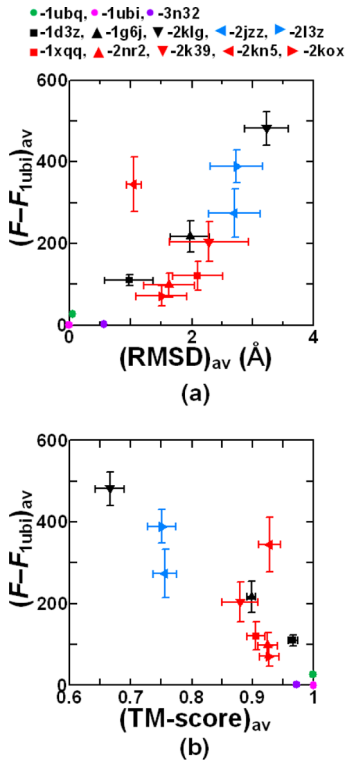


Figure 2. Relation between $(F - F_{1ubi})_{av}$ (the subscripts "av" and "1ubi" denote the average value in each model set and the value for the X-ray model in 1ubi, respectively) and rmsd or TM-score for C_α atoms calculated with the model in 1ubi as the standard structure. The standard deviation of $F - F_{1ubi}$, $(F - F_{1ubi})_{sd}$ (the subscripts "sd" denote the standard deviation), and that of the rmsd or TM-score are indicated as error bars. All residues (1−76) are considered in the calculation of the TM-score and rmsd. (a) rmsd. (b) TM-score. Black: solution NMR in model sets of type 1, blue: solid-state NMR in model sets of type 1, and red: ensembles in model sets of type 2.

dx.doi.org/10.1021/jp301541z | J. Phys. Chem. B XXXX, XXX, XXX−XXX

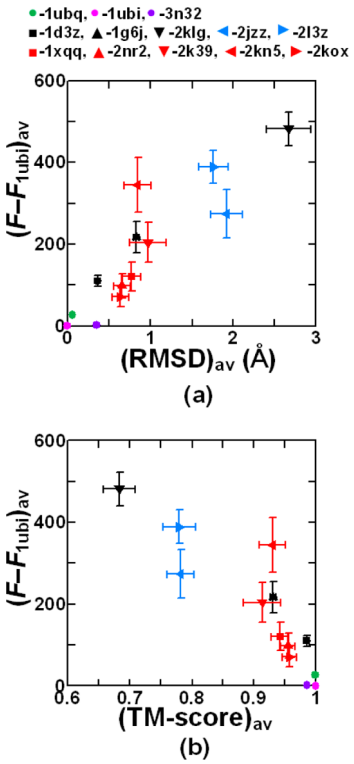


Figure 3. Relation between $(F - F_{1ubi})_{av}$ (the subscripts "av" and "1ubi" denote the average value in each model set and the value for the X-ray model in 1ubi, respectively) and rmsd or TM-score for C_α atoms calculated with the model in 1ubi as the standard structure. The standard deviation of $F - F_{1ubi}$, $(F - F_{1ubi})_{sd}$ (the subscripts "sd" denote the standard deviation), and that of the rmsd or TM-score are indicated as error bars. The core region comprising residues 1−70 are considered in the calculation of the TM-score and rmsd. (a) rmsd. (b) TM-score. Black: solution NMR in model sets of type 1, blue: solid-state NMR in model sets of type 1, and red: ensembles in model sets of type 2.

3. It is known that residues 71−76 which we do not include in the core region are significantly flexible.[11] They are disordered as observed in Figure 1. This is because they do not participate in the packing. Since the correlation between $(F - F_{1ubi})_{av}$ and the rmsd is influenced by these flexible residues, it is not appropriate to discuss the correlation by looking at Figure 2a. As observed in Figures 2b and 3, there is a general trend that $(F - F_{1ubi})_{av}$ becomes higher as the TM-score decreases or the rmsd increases and the model deviates more from the reference model in terms of the backbone structure. The correlation between $(F - F_{1ubi})_{av}$ and the TM-score or rmsd can be represented by a curve (not drawn) for 1d3z, 1g6j, 2klg, 2k39, and 2l3z. The results for 2kox, 2nr2, 1xqq, and 2jzz deviate from the curve in a downward direction, whereas the result for 2kn5 deviates from it in an upward direction.

We first discuss model sets of type 1. On an average, the models in 1d3z are the closest to the reference model in term of the TM-score, rmsd, and F, and those in 1g6j are the second closest. In 2klg, the models exhibit considerably high F and the largest deviation from the reference model in terms of structural properties. Almost the same characteristics are observed in 2l3z. As for model sets of type 2, on an average, the models in 2kox, 2nr2, and 1xqq possess relatively lower F. Compared with the models in 1d3z, those in 2kox, 2nr2, and 1xqq are more different from the reference model with respect to structural properties. However, the values of F for 2kox, 2nr2, and 1xqq are as low as those for 1d3z. We find that the models in 2kn5 give higher F than all the models in 2kox, 2nr2, and 1xqq. Among model sets of type 2, 2kn5 gives the highest F_{av}.

Relation between Behavior of Free-Energy Function and Structural Convergence in Model Set. In order to examine the relation between the behavior of F and the convergence of model structures, we propose the following procedure for each model set: calculate the averaged structure; calculate RMSDs from the averaged structure for heavy atoms in the backbone and side chains; calculate the average value $(\text{rmsd})_{av}$ and standard deviation $(\text{rmsd})_{sd}$; calculate the average value and standard deviation of F, F_{av}, and F_{sd}, respectively; and plot the relation between $(\text{rmsd})_{av}$ and F_{av} as a figure. In the figure, $(\text{rmsd})_{sd}$ and F_{sd} are also indicated as error bars. Residues 71−76 are excluded from the calculation of RMSDs.

Figure 4 shows the plots explained above. A curve (curve 1; not drawn) correlating F_{av} with $(\text{rmsd})_{av}$ is noticed for the three X-ray models and the models in 1d3z, 1g6j, 2klg, 2l3z, and 2kn5. The correlation between F_{av} and $(\text{rmsd})_{av}$ for 2kox, 2nr2, 1xqq, and 2k39 is expressed by another curve (curve 2; not drawn). The result for the models in 2jzz seems to deviate from both of curves 1 and 2: it is in-between. Curve 2 is shifted in a

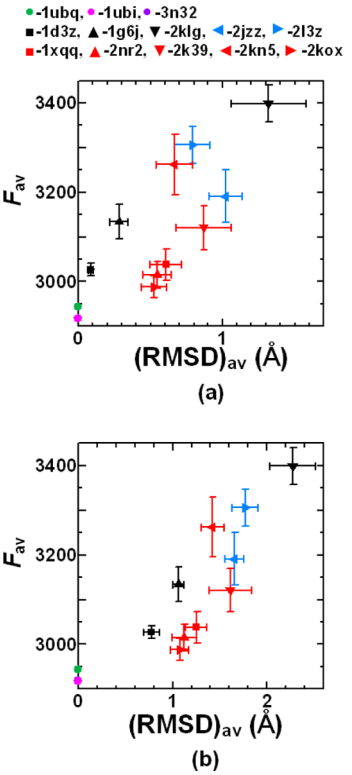


Figure 4. F_{av} plotted against $(\text{rmsd})_{av}$ (the subscript "av" denotes the average value). The core region comprising residues 1−70 are considered. The rmsd, which is for heavy atoms, represents the deviation from the averaged structure calculated for the structures in each model set. Thus, the definition of the rmsd is different from that in Figures 2 or 3. F_{sd} and $(\text{rmsd})_{sd}$ (the subscript "sd" denotes the standard deviation) are indicated as the error bars. (a) Plot for the backbone. (b) Plot for the side chains. Black: solution NMR in model sets of type 1, blue: solid-state NMR in model sets of type 1, and red: ensembles in model sets of type 2. The points for 1ubi and 3n32 are almost indistinguishable.

dx.doi.org/10.1021/jp301541z | J. Phys. Chem. B XXXX, XXX, XXX−XXX

downward direction in comparison with curve 1. Namely, for model sets of type 2 excluding 2kn5, F_{av} is relatively lower for a given value of $(rmsd)_{av}$. In each curve, there is a strong tendency that $(rmsd)_{sd}$, F_{av}, and F_{sd} decrease as $(rmsd)_{av}$ becomes smaller. Smaller $(rmsd)_{av}$ and $(rmsd)_{sd}$ imply better convergence of the model structures, and in such cases F is also better converged. By way of exception, the models in 2kn5 possess very large F_{sd}. Better convergence of F or the model structures is indicative of a larger total amount of constraints effectively taken into account in the structure calculation or the ensemble refinement protocol. It is important to know how effectively the experimentally obtained constraints have been incorporated in the models constructed. However, there are a variety of constraints (e.g., NOEs, RDC, hydrogen bonding, and dihedral angle restraints) and different constraints are used in different amounts. The details of the way employed for imposing the constraints on the model structures are also variable. Nevertheless, impartial comparison among the structures in model sets can be made through the plots like Figure 4 using our free-energy function with regard to the total amount of constraints effectively taken into account.

Among model sets of type 1, 1d3z gives the smallest values of $(rmsd)_{av}$ and F_{av}. The model structures in 1d3z are the best converged. Those in 1g6j are the second best converged. $(rmsd)_{av}$ and F_{av} take the largest values for 2klg. As for model sets of type 2, 2kox, 2nr2, and 1xqq are successful in the construction of fluctuating structures in aqueous solution. This is particularly true for the structural ensemble in 2kox. The success is also reflected on the deviation from the curve in a downward direction observed in Figures 2b and 3. Last, we remark the following: If there is a data set for which $(rmsd)_{av}$ and $(rmsd)_{sd}$ are small though F_{av} and F_{sd} are high, its structures are well converged to a different model, an NS model somewhat vitiated by unfavorable solvent condition (such a case is not found in Figure 4).

Entropic and Energetic Components. We define X and Y as

$$X = \{\Lambda/(k_BT)\} - \{\Lambda/(k_BT)\}_{1ubi} \quad (7)$$

and

$$Y = (-S/k_B) - (-S/k_B)_{1ubi} \quad (8)$$

Λ or X is a measure of the assurance of intramolecular hydrogen bonds for decreasing the total dehydration penalty.[14−16] A higher value of X implies less intramolecular hydrogen bonds formed by donors and acceptors buried in the protein interior. Close packing of the backbone and side chains leads to the reduction of the excluded volume generated for water molecules by a protein followed by the relaxation of water crowding.[31,52−54] $-S$ or Y represents the efficiency of backbone and side-chain packing for making the hydration entropy as small as possible.[14−16,31] Low efficiency results in a higher value of Y.

The plot of Y_{av} against X_{av} is shown for the NMR models in Figure 5: Y_{sd} and X_{sd} are indicated as error bars. The three X-ray models, which share almost the same characteristics with respect to Λ and S, are also included in the plot. There is a general trend that as X_{av} becomes higher, Y_{av} also increases. The points of each data set tend to gather in its own way, reflecting the specifications of the experimental technique and the structure calculation or the ensemble refinement protocol employed. For the models in 2l3z and 2kn5, Y_{av} is relatively higher, indicating less efficient packing of the backbone and/or

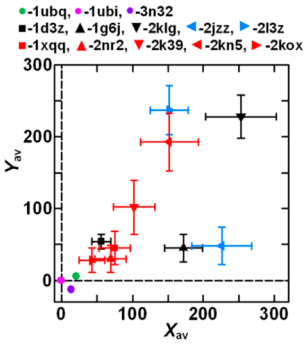


Figure 5. Y_{av} plotted against X_{av} (the subscript "av" denotes the average value). X and Y are defined by eqs 7 and 8, respectively. Y_{sd} and X_{sd} (the subscript "sd" denotes the standard deviation) are indicated as error bars. Black: solution NMR in model sets of type 1, blue: solid-state NMR in model sets of type 1, and red: ensembles in model sets of type 2.

side chains. By contrast, those in 1g6j and 2jzz suffer relatively higher values of X_{av} and less intramolecular hydrogen bonds formed.

Definition of Contributions to Energetic and Entropic Components from Backbone and Side Chains. It is physically insightful to separate the effect of side chains from that of the backbone. The contributions from the backbone and side chains to Λ, which are denoted by Λ_b and Λ_{sc}, respectively, can readily be obtained. X_b and X_{sc} are defined by eq 7 where Λ_b and Λ_{sc} are substituted for Λ, respectively. To perform the separation for the entropic component, we replace all residues in each structure by Gly using the CHARMM and MMTSB programs.[31,45,46] The replacement is carried out after the slight modification of the structure described in "Slight Modification of Native-Structure Models". The structure thus made has essentially no side chains (hereafter, these are referred to as "the structures without side chains"). $-S$ represents the loss of the water entropy upon the insertion of a protein with a prescribed structure. The information on the effect of side chains is contained in "$-S$ of a structure with side chains" − "$-S$ of the corresponding structure without side chains (i.e., with the backbone alone)": The latter is denoted by $-S_b$ and $-S = -S_b + (-S_{sc})$. $-S_b$ and $-S_{sc}$ denote the contributions from the backbone and side chains to $-S$, respectively. Y_b and Y_{sc} are defined by eq 8 where $-S_b$ and $-S_{sc}$ are substituted for $-S$, respectively.

Physical Origin of Superiority or Inferiority of Energetic Component for NMR Models. Figure 6 shows the average values of X_b and X_{sc} (they are denoted by $X_{b,av}$ and

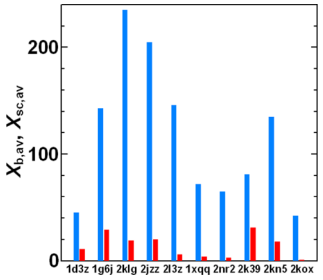


Figure 6. Average values of X_b and X_{sc} (they are denoted by $X_{b,av}$ and $X_{sc,av}$, respectively) for the NMR models. X is defined by eq 7. X_b and X_{sc}, respectively, are the contributions from the backbone and side chains to X. X_b and X_{sc} are marked in blue and red, respectively.

dx.doi.org/10.1021/jp301541z | J. Phys. Chem. B XXXX, XXX, XXX−XXX

$X_{sc,av}$, respectively) for the NMR models. It is observed that X_b is much larger than X_{sc}. In comparison with the X-ray models, the NMR models undergo larger total dehydration penalty in the backbone than in side chains. This is particularly true for 2klg and 2jzz. Among model sets of type 1, $X_{b,av}$ is the smallest for 1d3z and the largest for 2klg. Among those of type 2, $X_{b,av}$ is the smallest for 2kox and the largest for 2kn5.

The contribution from each residue to $\Lambda_{b,av}$ is shown in Figure 7 where the results for 1ubi, 1d3z, and 2klg or those

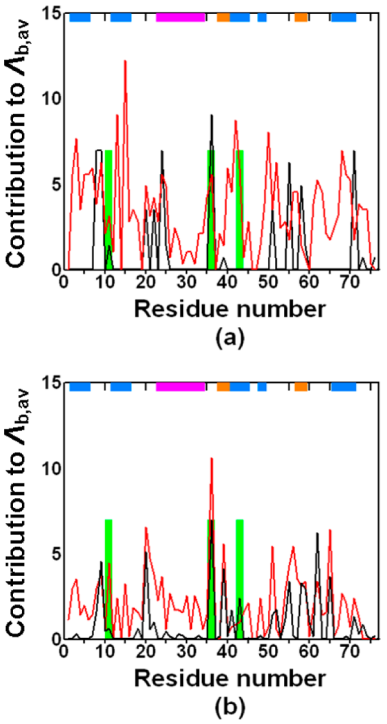


Figure 7. Contribution from each residue to the average value of Λ_b (it is denoted by $\Lambda_{b,av}$). Λ is the total dehydration penalty in eq 4. The results for 1ubi (green), 1d3z (black), and 2klg (red) are compared in panel (a), and those for 1ubi (green), 2kox (black), and 2kn5 (red) are compared in (b). Only residues 11, 36, and 43 undergo the dehydration penalty in 1ubi, and the contribution is shown by a vertical bar. At the top of the figure, the portions of α-helix, β-sheet, and 3_{10}-helix for 1ubi are indicated in pink, blue, and orange, respectively: For example, residues 23−34 form α-helix. These secondary structures are not always complete in 2klg and 2kn5.

from 1ubi, 2kox, and 2kn5 are compared. At the top of the figure, the portions of α-helix, β-sheet, and 3_{10}-helix for 1ubi are indicated in three different colors. These secondary structures are identified using the DSSP program.[55] The three X-ray models share the characteristic that the loop portions as well as the secondary structures in the backbone exhibit only small dehydration penalty and sufficiently many intramolecular hydrogen bonds are formed. On an average, the models in 1d3z and 2kox undergo larger dehydration penalty primarily in the loop portions. However, we find that significantly many models in 2kox possess as many intramolecular hydrogen bonds as the three X-ray models in the secondary structures and the loop portions. As for the models in 2klg and 2kn5, intramolecular hydrogen bonds are not always formed even in the secondary structures, causing considerably large dehydration penalty.

Physical Origin of Superiority or Inferiority of Entropic Component for NMR Models. Figure 8 shows

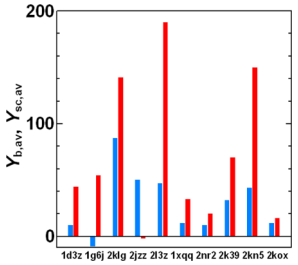


Figure 8. Average values of Y_b and Y_{sc} (they are denoted by $Y_{b,av}$ and $Y_{sc,av}$, respectively) for the NMR models. Y is defined by eq 8. Y_b and Y_{sc}, respectively, are the contributions from the backbone and side chains to Y. Y_b and Y_{sc} are marked in blue and red, respectively.

the average values of Y_b and Y_{sc} (they are denoted by $Y_{b,av}$ and $Y_{sc,av}$, respectively) for the NMR models. Except for the models in 2jzz, Y_{sc} is larger than Y_b. In the NMR models, the deterioration of the packing efficiency is more significant in side chains than in the backbone, which is in contrast to the case of the total dehydration penalty. The side-chain packing is relatively more inefficient for 2l3z, 2kn5, and 2klg. Among model sets of type 2, 2kn5 features appreciably inefficient packing of side chains.

As illustrated in Figure 9, there is an overall trend that as the side-chain packing becomes less efficient, the efficiency of the

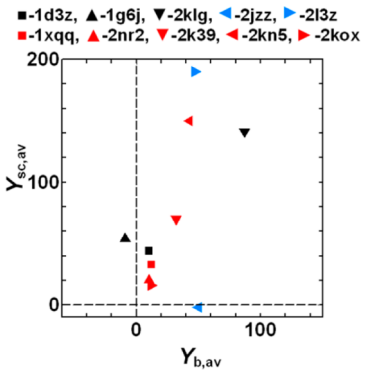


Figure 9. Relation between $Y_{b,av}$ and $Y_{sc,av}$ denoting the average values of Y_b and Y_{sc}, respectively, for the NMR models. Y is defined by eq 8. Y_b and Y_{sc}, respectively, are the contributions from the backbone and side chains to Y. Black: solution NMR in model sets of type 1, blue: solid-state NMR in model sets of type 1, and red: ensembles in model sets of type 2.

backbone packing is also lowered. On an average, the packing of the backbone and side chains for the models in 2kox and 2nr2 is as close as that for the three X-ray models.

Relevance to Specificity of NMR Structures in Model Set. Since any NMR approach has its own advantages, we can never grade the previously reported approaches. Yet, the characteristics of the NMR models can be summarized as follows. We first discuss model sets of type 1. 1d3z merits the largest total amount of constraints effectively taken into account and the best converged model structures, leading to the lowest values of F and Λ and the second lowest value of $-S$ on the whole. Structural properties as well as the values of F, Λ, and $-S$ are close to those of the three X-ray models. From this result, we can conclude that the NS of ubiquitin is influenced

dx.doi.org/10.1021/jp301541z | J. Phys. Chem. B XXXX, XXX, XXX−XXX

neither by crystallization in the X-ray crystallography nor by solvent environment adjusted in the NMR experiment for 1d3z. The characteristics of the models in 1g6j are fairly close to those in 1d3z except the relatively larger total dehydration penalty in the backbone. The structure of ubiquitin remains almost undisturbed upon encapsulation within a reverse micelle. As for the models in 2klg, the total amount of constraints effectively taken into account is not sufficiently large and the models overall suffer the lack in firmly formed intramolecular hydrogen bonds in the backbone. They feature incomplete secondary structures, which can be appreciated even in visualized ribbon representation of the model structures (see Figure 1). It is interesting that the side-chain packing for 2l3z is the least efficient while that for 2jzz is the most efficient despite that both 2l3z and 2jzz were obtained through solid-state NMR. The differences are ascribed probably to the details of the structure calculations employed. The effects of such details on the NS models generated are usually ambiguous, and this is one of the reasons why the characterization of the models becomes useful. The most efficient packing of side chains for 2jzz is the principal reason for the deviations from the correlations among 1d3z, 1g6j, 2klg, and 2l3z observed in Figures 2−4.

We then discuss model sets of type 2. The structural variability of the native state in aqueous solution is well represented by the models in 2kox, 2nr2, and 1xqq. In particular, 2kox provides successful representation of the structures fluctuating around the X-ray structure in 1ubq for the following reasons: 14 models in 2kox give lower F than the X-ray structure; as mentioned above, the structure with the lowest F is included in 2kox; the rmsd and TM-score of this structure calculated for C_α atoms by choosing the reference model as the standard structure (the core region comprising residues 1−70 are considered) are 0.66 and 0.96 Å, respectively; and it is observed in Figure 4 that 2kox merits the largest total amount of constraints effectively taken into account. Figure 10 shows the relation between $F-F_{1ubi}$ and the rmsd or TM-score for all the models in 2k39, 2kn5, and 2kox. The F value in 2k39 varies considerably from model to model, and the maximum difference in F observed is as large as ∼300. 2k39 includes significantly many structures with high F. On the whole, the models in 2kn5 suffer very high F and very large deviation from the reference model primarily due to the rather unfavorable side-chain packing. This is why 2kn5 departs from the correlations among 2kox, 2nr2, 1xqq, and 2k39 as observed in Figures 2−4.

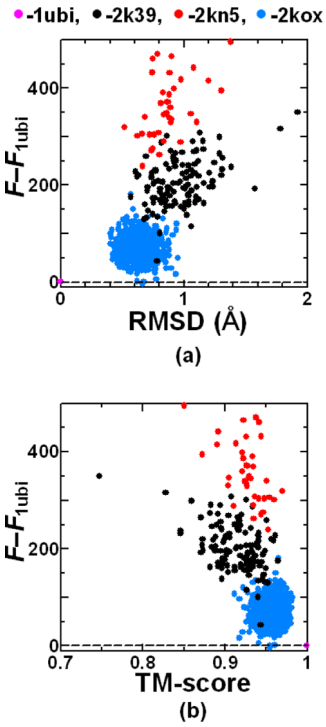


Figure 10. Relation between $F-F_{1ubi}$ and rmsd or TM-score calculated for C_α atoms by choosing the model in 1ubi as the standard structure. The core region comprising residues 1−70 are considered in the calculation of the rmsd or TM-score. All of the models in 2k39, 2kn5, and 2kox are individually treated. (a) rmsd. (b) TM-score.

CONCLUSION

We have developed a reliable method of characterizing the NS models of a protein determined through a variety of routes. It is illustrated for ubiquitin for which a comprehensive range of high-quality X-ray and NMR data are available. The characterization is based on not mere geometric analyses but energetics of a protein accounting for its hydration thermodynamics wherein a molecular model is employed for water. The NS models can be classified into X-ray models[1−3] and two types of NMR model sets. A model set of type 1 comprises candidate models for a fixed NS determined by solution or solid-state NMR with the aid of a structure calculation upon which the constraints experimentally obtained[4−8] (e.g., NOEs, RDCs, hydrogen bonding, and dihedral angle restraints) are imposed. That of type 2 forms an ensemble of structures representing the structural variability of the native state in aqueous solution.[9−13] It is usually constructed by employing the ensemble refinement protocol[9−11,13] started from an NS model determined by the X-ray crystallography or solution NMR. The refinement is performed in a variety of manners using a computer simulation with all-atom potentials or a method based on geometric restrictions which is restrained by NOEs, RDCs, and order parameters S^2 obtained in NMR experiments. Our recently developed free-energy function F and its energetic and entropic components[14−16] (Λ and S, respectively) are shown to be very useful to the characterization.

In general, the X-ray models give lower F than the NMR models. The X-ray models resemble one another with respect to structural properties, F, Λ, and S, whereas the NMR models exhibit diverse properties. These results never deny the usefulness of NMR. In the X-ray crystallography, it may be difficult to find out the solvent condition leading to the crystallization of a protein which is not far from the physiological condition. Moreover, only a smaller amount of information on dynamical properties of a protein structure is provided by the X-ray crystallography.

The model giving the lowest F among the X-ray models is chosen as the reference model in the characterization. If there are only NMR models, the model with the lowest F among them is the best one for the reference model. The plot of $(F - F_{ref})_{av}$ (the subscripts "av" and "ref" denote the average value and the value of the reference model, respectively) against the TM-score[51] calculated for C_α atoms by choosing the reference model as the standard structure allows us to evaluate the quality of each model set. $(F - F_{ref})_{sd}$ (the subscript "sd" denote the standard deviation) and the standard deviation of the TM-score are indicated as error bars. There is a general trend that $(F - F_{ref})_{av}$ becomes higher as the TM-score decreases and the

dx.doi.org/10.1021/jp301541z | J. Phys. Chem. B XXXX, XXX, XXX−XXX

model deviates more from the reference model in terms of the backbone structure (this trend is required in any characterization method). Namely, $(F - F_{ref})_{av}$ is correlated with the TM-score. This correlation can be described by a curve, but some model sets exhibit deviations from the curve. In particular, model sets of type 2, when they successfully represent fluctuating structures of the native state in aqueous solution, deviate from the curve in a downward direction. A good structural ensemble for the native state should include significantly many structures for which F is lower than that for the reference model. When the TM-score is replaced by the rmsd, the correlation between $(F - F_{ref})_{av}$ and the rmsd is observed only if the residues with high flexibility are removed (compare Figures 2a and 3a). The flexible residues, which exhibit large fluctuations, are visualized by superposing the set of structures. Such residues can perfectly be identified by the plots like Figures 2a and 3a together with the superposition. In general, when the rmsd is smaller than 1 Å or the TM-score is larger than 0.9 between two structures, they are considered almost the same. Our free-energy function, which possesses high resolution power, is capable of distinguishing even those structures.

In each model set, the relation between the structural convergence and behavior of the free-energy function can be explored in the following manner: First, calculate the averaged structure; second, calculate RMSDs from the averaged structure for heavy atoms in the backbone and side chains; calculate the average value $(rmsd)_{av}$ and standard deviation $(rmsd)_{sd}$; calculate the average value and standard deviation of F, F_{av}, and F_{sd}, respectively; and plot the relation between $(rmsd)_{av}$ and F_{av} as a figure. In the figure, $(rmsd)_{sd}$ and F_{sd} are also indicated as error bars. The flexible residues which can be identified as described above (see Figure 4) are excluded from the calculation of RMSDs. Model sets of type 1 and those of type 2 can be described by different correlation curves. In each curve, there is a strong tendency that $(rmsd)_{sd}$, F_{av}, and F_{sd} decrease as $(rmsd)_{av}$ becomes smaller. However, F_{av} is lower for a given value of $(rmsd)_{av}$ in model sets of type 2 than in those of type 1. Smaller $(rmsd)_{av}$ and $(rmsd)_{sd}$ imply better convergence of the model structures, and in such cases F is also better converged (this property is required in any characterization method). Even when there is a model set which deviates from the correlation curve, the physical reasoning for the deviation can be made, thus uncovering its features. If we find a data set for which $(rmsd)_{av}$ and $(rmsd)_{sd}$ are small though F_{av} and F_{sd} are high, the structures are well converged to an incorrect NS model, a model somewhat vitiated by unfavorable solvent condition. We remark that impartial comparison among the structures in model sets can be made through the plots like Figure 4 using our free-energy function with regard to the total amount of constraints effectively taken into account.

The plot of X_{av} against Y_{av} ($X = \{\Lambda/(k_BT)\} - \{\Lambda/(k_BT)\}_{ref}$ and $Y = (-S/k_B) - (-S/k_B)_{ref}$) with X_{sd} and Y_{sd} indicated as error bars gives useful information on the assurance of intramolecular hydrogen bonds, efficiency of backbone and side-chain packing, and balance of these two principal factors[14−16] (see Figure 5). By analyzing the contributions from the backbone and side chains to X_{av} and Y_{av} and drawing the figures such as Figures 6−9, we can clarify detailed weak points of the models in an NMR model set in terms of intramolecular hydrogen bonding and packing efficiency in the backbone and side chains.

We emphasize that the results described above are achievable only by a free-energy function capturing essential physics of the structural stability of a protein in aqueous solution. It has been shown that F is certainly this type of function. The characterization method thus developed should be useful for the following applications: the evaluation of a set of NS models of a protein determined via a new NMR approach by comparing them to the models which are already available; and refinement of an NMR model by rectifying its weak points found. Further, our free-energy function is well suited not only to the selection of the best model from among many candidate NMR models but also to the original construction of the best candidate model or a good structural ensemble for the native state, on the basis of the experimentally obtained constraints such as NOEs and RDCs. Works in these directions are in progress.

AUTHOR INFORMATION

Corresponding Author

*E-mail: kinoshit@iae.kyoto-u.ac.jp.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the two NMR experimentalists, Takahisa Ikegami (Institute for Protein Research, Osaka University) and Takashi Nagata (Institute of Advanced Energy, Kyoto University), for their careful reading of our manuscript. The computer program for the morphometric approach was developed with Roland Roth. This work was supported by Grant-in-Aid for Scientific Research on Innovative Areas (Nos. 20118004 and 21118519) from the Ministry of Education, Culture, Sports, Science and Technology of Japan, by Grant-in-Aid for Scientific Research (B) (Nos. 22300100 and 22300102) from Japan Society for the Promotion of Science, by the Grand Challenges in Next-Generation Integrated Nanoscience, MEXT, Japan, and by Kyoto University Global Center of Excellence (GCOE) of Energy Science.

REFERENCES

(1) Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. *J. Mol. Biol.* **1987**, *194*, 531−544.
(2) Alexeev, D.; Bury, S. M.; Turner, M. A.; Ogunjobi, O. M.; Muir, T. W.; Ramage, R.; Sawyer, L. *Biochem. J.* **1994**, *299*, 159−163.
(3) Arnesano, F.; Belviso, B. D.; Caliandro, R.; Falini, G.; Fermani, S.; Natile, G.; Siliqi, D. *Chem.—Eur. J.* **2011**, *17*, 1569−1578.
(4) Cornilescu, G.; Marquardt, J. L.; Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 6836−6837.
(5) Babu, C. R.; Flynn, P. F.; Wand, A. J. *J. Am. Chem. Soc.* **2001**, *123*, 2691−2692.
(6) Madl, T.; Bermel, W.; Zangger, K. *Angew. Chem., Int. Ed.* **2009**, *48*, 8259−8262.
(7) Manolikas, T.; Herrmann, T.; Meier, B. H. *J. Am. Chem. Soc.* **2008**, *130*, 3959−3966.
(8) Huber, M.; Hiller, S.; Schanda, P.; Ernst, M.; Böckmann, A.; Verel, R.; Meier, B. H. *ChemPhysChem* **2011**, *12*, 915−918.
(9) Lindorff-Larsen, K.; Best, R. B.; DePristo, M. A.; Dobson, C. M.; Vendruscolo, M. *Nature* **2005**, *433*, 128−132.
(10) Richter, B.; Gsponer, J.; Várnai, P.; Salvatella, X.; Vendruscolo, M. *J. Biomol. NMR* **2007**, *37*, 117−135.
(11) Lange, O. F.; Lakomek, N. A.; Farès, C.; Schröder, G. F.; Walter, K. F. A.; Becker, S.; Meiler, J.; Grubmüller, H.; Griesinger, C.; de Groot, B. L. *Science* **2008**, *320*, 1471−1475.

(12) Friedland, G. D.; Lakomek, N. A.; Griesinger, C.; Meiler, J.; Kortemme, T. *PLoS Comput. Biol.* **2009**, *5*, e1000393(1−16).
(13) Fenwick, R. B.; Esteban-Martín, S.; Richter, B.; Lee, D.; Walter, K. F. A.; Milovanovic, D.; Becker, S.; Lakomek, N. A.; Griesinger, C.; Salvatella, X. *J. Am. Chem. Soc.* **2011**, *133*, 10336−10339.
(14) Harano, Y.; Roth, R.; Sugita, Y.; Ikeguchi, M.; Kinoshita, M. *Chem. Phys. Lett.* **2007**, *437*, 112−116.
(15) Yoshidome, T.; Oda, K.; Harano, Y.; Roth, R.; Sugita, Y.; Ikeguchi, M.; Kinoshita, M. *Proteins* **2009**, *77*, 950−961.
(16) Yasuda, S.; Yoshidome, T.; Harano, Y.; Roth, R.; Oshima, H.; Oda, K.; Sugita, Y.; Ikeguchi, M.; Kinoshita, M. *Proteins* **2011**, *79*, 2161−2171.
(17) Cann, N. M.; Patey, G. N. *J. Chem. Phys.* **1997**, *106*, 8165−8195.
(18) Kusalik, P. G.; Patey, G. N. *J. Chem. Phys.* **1988**, *88*, 7715−7738.
(19) Kusalik, P. G.; Patey, G. N. *Mol. Phys.* **1988**, *65*, 1105−1119.
(20) Kinoshita, M.; Harada, M. *Mol. Phys.* **1994**, *81*, 1473−1488.
(21) Kinoshita, M.; Iba, S.; Harada, M. *J. Chem. Phys.* **1996**, *105*, 2487−2499.
(22) Kinoshita, M.; Bérard, D. R. *J. Comput. Phys.* **1996**, *124*, 230−241.
(23) Kinoshita, M. *J. Chem. Phys.* **2008**, *128*, 024507(1−14).
(24) König, P. M.; Roth, R.; Mecke, K. R. *Phys. Rev. Lett.* **2004**, *93*, 160601(1−4).
(25) Roth, R.; Harano, Y.; Kinoshita, M. *Phys. Rev. Lett.* **2006**, *97*, 078101(1−4).
(26) Kodama, R.; Roth, R.; Harano, Y.; Kinoshita, M. *J. Chem. Phys.* **2011**, *135*, 045103(1−8).
(27) Hirata, F.; Rossky, P. J. *Chem. Phys. Lett.* **1981**, *83*, 329−334.
(28) Perkyns, J. S.; Pettitt, B. M. *Chem. Phys. Lett.* **1992**, *190*, 626−630.
(29) Perkyns, J. S.; Pettitt, B. M. *J. Chem. Phys.* **1992**, *97*, 7656−7666.
(30) Imai, T.; Harano, Y.; Kinoshita, M.; Kovalenko, A.; Hirata, F. *J. Chem. Phys.* **2006**, *125*, 024911(1−7).
(31) Yasuda, S.; Yoshidome, T.; Oshima, H.; Kodama, R.; Harano, Y.; Kinoshita, M. *J. Chem. Phys.* **2010**, *132*, 065105(1−10).
(32) Connolly, M. L. *J. Appl. Crystallogr.* **1983**, *16*, 548−558.
(33) Connolly, M. L. *J. Am. Chem. Soc.* **1985**, *107*, 1118−1124.
(34) MacKerell, A. D., Jr.; Bashford, D.; Bellott, M.; Dunbrack, R. L., Jr.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E., III; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(35) Yoshidome, T.; Kinoshita, M.; Hirota, S.; Baden, N.; Terazima, M. *J. Chem. Phys.* **2008**, *128*, 225104(1−9).
(36) Harano, Y.; Yoshidome, T.; Kinoshita, M. *J. Chem. Phys.* **2008**, *129*, 145103(1−9).
(37) Yoshidome, T.; Kinoshita, M. *Phys. Rev. E* **2009**, *79*, 030905(R)(1−4).
(38) Oshima, H.; Yoshidome, T.; Amano, K.; Kinoshita, M. *J. Chem. Phys.* **2009**, *131*, 205102(1−11).
(39) Amano, K.; Yoshidome, T.; Oda, K.; Harano, Y.; Kinoshita, M. *Chem. Phys. Lett.* **2009**, *474*, 190−194.
(40) Oda, K.; Kodama, R.; Yoshidome, T.; Yamanaka, M.; Sambongi, Y.; Kinoshita, M. *J. Chem. Phys.* **2011**, *134*, 025101(1−9).
(41) Sneddon, S. F.; Tobias, D. J.; Brooks, C. L., III. *J. Mol. Biol.* **1989**, *209*, 817−820.
(42) Güntert, P.; Mumenthaler, C.; Wüthrich, K. *J. Mol. Biol.* **1997**, *273*, 283−298.
(43) de Groot, B. L.; van Aalten, D. M. F.; Scheek, R. M.; Amadei, A.; Vriend, G.; Berendsen, H. J. C. *Proteins* **1997**, *29*, 240−251.
(44) Davis, I. W.; Arendall, W. B., 3rd; Richardson, D. C.; Richardson, J. S. *Structure* **2006**, *14*, 265−274.
(45) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187−217.
(46) Feig, M.; Karanicolas, J.; Brooks, C. L., III. *J. Mol. Graphics Modell.* **2004**, *22*, 377−395.
(47) MacKerell, A. D., Jr.; Feig, M.; Brooks, C. L., III. *J. Comput. Chem.* **2004**, *25*, 1400−1415.
(48) Lee, M. S.; Salsbury, F. R., Jr.; Brooks, C. L., III. *J. Chem. Phys.* **2002**, *116*, 10606−10614.
(49) Lee, M. S.; Feig, M.; Salsbury, F. R., Jr.; Brooks, C. L., III. *J. Comput. Chem.* **2003**, *24*, 1348−1356.
(50) Chocholoušová, J.; Feig, M. *J. Comput. Chem.* **2006**, *27*, 719−729.
(51) Zhang, Y.; Skolnick, J. *Proteins* **2004**, *57*, 702−710.
(52) Harano, Y.; Kinoshita, M. *Biophys. J.* **2005**, *89*, 2701−2710.
(53) Kinoshita, M. *Front. Biosci.* **2009**, *14*, 3419−3454.
(54) Kinoshita, M. *Int. J. Mol. Sci.* **2009**, *10*, 1064−1080.
(55) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577−2637.