A New Efficient Method for Generating Conformations of Unfolded Proteins with Diverse Main-Chain Dihedral-Angle Distributions

Yasutaka Seki,[†] Yudai Shimbo,[‡,§] Takamasa Nonaka,[†] and Kunitsugu Soda*[,§]

[†]School of Pharmacy, Iwate Medical University, Yahaba, Iwate 028-3694, Japan

[‡]Department of Bioengineering, Nagaoka University of Technology, Nagaoka, Niigata 940-2188, Japan

[§]Laboratory for Computational Molecular Design, Center for Computational Life Science, RIKEN, Kobe, Hyogo 650-0047, Japan

ABSTRACT: A new method for generating polypeptide-chain conformations has been developed for studying structural characteristics of unfolded proteins. It enables us to generate a large number of conformations very rapidly by avoiding atomic collisions efficiently with the use of main-chain dihedral-angle distributions derived from a crystal-structure database of proteins. In addition, combining main-chain dihedral-angle distributions for the amino acid residues incorporated in different secondary structures, we can obtain diverse conformational ensembles with different structural features. Structural characteristics of proteins denatured in high-concentration denaturant solution were analyzed by comparing predictions from this method with results from solution X-ray scattering (SXS) measurement. Analysis of the dependence of the mean square radius (R_{sq}) of protein on the number of residues and the shape of its Kratky profile has confirmed that the highly denaturing solvent serves as a good solvent in accordance with previous reports. It was also found that, in order for a conformational ensemble to reproduce experimental data, the percentage in which main-chain dihedral angles are found in the α region must be in the range of 20−40%. It agrees with studies on the $^3J_{HN\alpha}$ coupling constant using the multidimensional NMR method. These results confirm that our method for generating diverse conformations of polypeptide chains is very useful to the conformational analysis of unfolded protein, because it enables us to analyze comprehensively both of the local structural features obtained from NMR and the global ones obtained from SXS.

■ INTRODUCTION

The detailed molecular mechanism of protein folding and the physical design principle of the 3D structure of protein have been the most important targets to be studied in protein science. In order to solve the problem, it is necessary to clarify structural characteristics of protein in the completely or partially unfolded state as well as those in the natively folded state, because the non-native state appears in the initial or intermediate state of folding processes.[1,2] In recent years, attention has been given to the function and its expression mechanism of proteins that are natively unfolded or have an intrinsically disordered region.[3] Presumably, these proteins do not have a definite conformation but have diverse ones in their isolated state not bound to another protein and/or a ligand molecule. Hence, it is necessary to analyze detailed structural characteristics of unfolded proteins also for elucidating the molecular mechanism of expressing their functions. Under these circumstances, studies for analyzing structural features of proteins in the unfolded state have become increasingly important.[4,5]

Solution X-ray and neutron scatterings (SXS and SNS) and spectroscopic techniques such as multidimensional NMR are useful methods that give mutually complementary information on the structure of unfolded proteins in solution. Application of the SXS method to a protein in solution yields information of its global structure such as the mean square radius (R_{sq}) from a Guinier analysis, the molecular shape from a Kratky plot, and the distance distribution function.[2] Especially for unfolded proteins, the statistical feature of an ensemble of chain-like conformations has been analyzed on the basis of the scaling law that characterizes the dependence of R_{sq} on the number of residues, N_r.[6,7] From analyses of SXS data, many studies have yielded an exponent of about 0.6 on the scaling law for the denatured state of proteins in high-concentration denaturant solution.[6,7] It is assumed from comparison of the exponents estimated theoretically and experimentally that the conformational characteristic of the denatured state of a protein under highly denaturing conditions can be well approximated by a random-flight chain with a finite excluded-volume effect. Kratky-profile analysis also supports qualitatively the view that the conformation of a highly denatured protein can be virtually described by a random-flight chain.[8,9] Wang et al.[10] took into account the solvation effect in their method of modeling unfolded protein structures by assuming that the solvation energy of each conformation is proportional to its accessible surface area (ASA). They showed that the ensemble of conformations generated by their method reproduces well the experimental scaling-law exponent and SXS profile. The scaling-law analysis was also applied to the intermediate observed experimentally at the initial stage of the refolding process, which indicated that the statistical-structure analysis of chain molecules is useful for examining the molecular mechanism of protein folding.[11] On the other hand, the probability distribution function was estimated for the main-chain dihedral angle φ in the fully unfolded state from measurements of the $^3J_{HN\alpha}$ coupling constant by multidimensional NMR.[12,13] In addition, the content of residual secondary structures was evaluated from chemical shifts of proteins in the partially unfolded state[14] or peptides.[15] The information on these local

Received: December 9, 2010
Published: June 02, 2011

structures is complementary to the information on global structures obtained from the SXS method. In recent years, NMR and CD (circular dichroism) measurements revealed that short peptides can form a polyproline II helix in aqueous solution.[16,17] It was also found that residual dipolar couplings (RDCs) of NMR for staphylococcal nuclease,[18] apomyoglobin,[19] and some other proteins[20] under high-concentration denaturant conditions show amino acid sequence dependence clearly different from that of a simple bell-shaped form from the calculation for a homopolypeptide.[20] Initially, controversial interpretations were presented for these RDC data. However, it is now accepted[21] that (a) proteins under such conditions basically form no specific residual structure such as secondary structures and (b) unique RDC data result from the fact that the occurrence probability of main-chain dihedral angles varies with amino acid species and sequences.

We can see from the above that it is essential to elucidate characteristics of unfolded-protein structures consistent with the information on both of the local and global structures. So, it would be very effective to use a molecular modeling method where both kinds of information can be accurately taken into account with a high-performance computer. Several analyses have already been made from a viewpoint such as above. Previous works[22,23] revealed that the R_{sq} of unfolded proteins with residual secondary structures shows dependence on N_r similar to that of completely denatured proteins under high-concentration denaturant conditions. Jha et al.[20] reproduced experimental data of RDCs and the exponent of the dependence of R_{sq} on N_r, but the calculated R_{sq} differs significantly from the experimental one. The analysis[10] reproduced well experimental SXS data, but no comparative study with local structural analysis as NMR has been performed yet. Thus, no molecular modeling analysis of unfolded proteins in which both of the global SXS properties and the local NMR ones are reproduced has been carried out yet. One of the reasons for the difficulty of such an analysis is that conventional methods have not yet enabled us to generate a sufficiently large number of model structures with both kinds of information and reasonable conformational energies being incorporated.

To overcome the problem, we developed a new method for rapidly generating model structures of unfolded protein, where polypeptide conformations are produced by using various probability distribution functions for main-chain dihedral angles determined with the crystal structures of native globular proteins. Using this method, we can generate the structure of unfolded protein very rapidly. Moreover, in most cases, it is possible to reach a nearby local minimum-energy structure after a small number of computational steps, because interatomic collisions are eliminated in all of the initially generated structures. Diverse ensembles of polypeptide conformations with different structural characteristics can be generated by combining various main-chain dihedral-angle distribution functions derived from the ensembles of residues incorporated in different local structures of native proteins. We carried out an SXS analysis of the ensemble of conformations generated by using this method and considering the solvent effect in the ASA approximation. From our analysis, we could elucidate the relations among the scaling-law exponent of R_{sq}, the SXS profile, the distribution function of main-chain dihedral angles, and the solvent effect. On the basis of the results, it is demonstrated that detailed physical properties of unfolded protein can be analyzed by combining our modeling method with SXS measurement.

METHODS

Method for Generating Conformations of Unfolded Proteins. A new computer program was developed for generating conformations of unfolded proteins. This program is written in Fortran90 on a Linux computer. It accepts protein−structure data with any of the formats of Protein Data Bank, AMBER,[24] and GROMACS,[25] while it reads in the structural parameters such as bond lengths and bond angles from an input structure file. In this study, we use structural data with the GROMACS format and structural parameters of the AMBER99[26] package. Atomic collisions are decided using the set of atomic radii from Tsai et al.[27] Generated conformations are energy-minimized using a package in GROMACS 4.0[28] and the force field of AMBER 99. Solvent is implicitly considered by assuming its dielectric constant to be that of liquid water in evaluating conformational energies. The bond length of the hydrogen atom is held constant using the LINCS algorithm. All of the van der Waals and Coulomb interactions were evaluated within a cutoff distance of 1.2 nm. The steepest descent method was used for energy minimization.

We analyzed 12 proteins: horse cytochrome *c* (104 aa), bovine α-lactalbumin (123), bovine ribonuclease A (124), hen egg white lysozyme (129), sperm whale myoglobin (153), avian sarcoma virus integrase core (162), dihydrofolate reductase (191), MutY catalyic domain (225), triosephosphate isomerase (250), *Eco*Rl endonuclease (276), UDP-galactose 4-epimerase (338), and creatine kinase (379). It is assumed here that cytochrome *c* and myoglobin are of the apo form and all proteins are free of disulfide bonds, because the covalent structures of the same protein must be the same between computational and experimental[7] analyses to compare the results on the dependence of R_{sq} on the number of residues, N_r.

Determination of the Dihedral-Angle Distributions for Main and Side Chains. The main-chain dihedral-angle distributions (MCDAD) necessary for generating unfolded protein conformations were determined through the following procedure: First, the 379 native structures satisfying the three requirements below are chosen from 526 folds in the *Dali Domain Dictionary 2.0*:[29,30] (a) The number of residues is larger than 50. (b) The resolution of X-ray structure analysis is better than 0.24 nm. (c) The probability of occurrence of main-chain dihedral angles in the "core region[31]" is higher than 0.94. Next, all of the main-chain dihedral angles are calculated, and the secondary structures in all of the structures of target proteins are determined by using the DSSP algorithm.[32] Resulting data of the main-chain dihedral angles are classified by amino acid species and secondary-structure types to obtain the six MCDAD models denoted as (all), (α + t + c), (β + t + c), (t + c), (β + c) and (c) or (coil) for each of the 20 residues. Here, the four letters, i.e. α, β, t and c (or coil), refer to all of the helices (H, G, and I in DSSP), β strands (B and E in DSSP), turns with a H bond (T and S in DSSP), and coils other than those above (all except secondary structures in DSSP). For example, the (β + t + c) model denotes a MCDAD model constructed from the MCDADs of the residues not incorporated in any helical structure. All MCDAD models are assumed to contain the coil model, and all of the other combinations are taken for analysis, though the (α + β + c) and (α + c) models are excluded because they show MCDADs very similar to those of the (all) and (α + t + c) models, respectively.

As an example, the 3D plot of MCDAD for the (all) model is shown in Figure 1. Around the peak of distribution in the

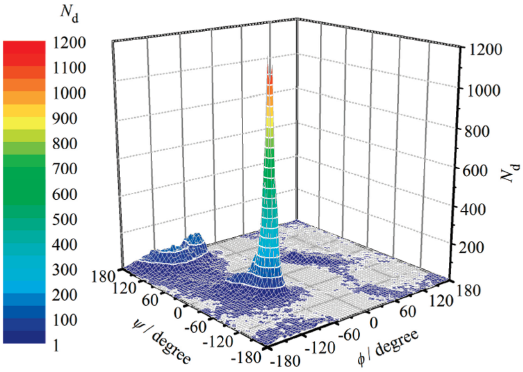

Figure 1. Example of the 3D distribution of occurrence frequency on the (φ, ψ) plane for a MCDAD model, (all). A total of 55 012 sets of (φ, ψ) data are obtained from 379 native-protein structures. All pairs of (φ, ψ) data are sorted with a division of $\Delta\varphi = \Delta\psi = 5°$, and the number of data, N_d, in each unit area is 3D plotted. The color of the 3D surface is changed with a division of $\Delta N_d = 50$, while the areas with $N_d = 0$ are not colored.

α region with the highest occurrence frequency, the number of occurrences in the unit area with $\Delta\varphi = \Delta\psi = 5°$ exceeds 1000. Hence, the relative standard error of the number of occurrences in this area is 3.2%, showing high accuracy in this region. The definition of the α region as well as those of β and p regions are given in the literature.[33] On the other hand, the number of occurrences is only slightly larger than 100 around the peak of distribution in the β or p region, indicating about 10% error or 3 times lower accuracy than that in the α region. However, the occurrence frequency in the β or p region is more widely distributed than that in the α region. The percentage with its standard error that a pair of main-chain dihedral angles occurs in either the β or p region is 43.6 ± 0.3%. It is comparable with that for the α region, 44.0 ± 0.3%. For the other regions, the number of occurrences in a unit area is smaller than 50, and the relative error is still larger. However, the total occurrence probability for these regions is not so high. For example, the total percentage that a pair of dihedral angles occurs in some area having a number of occurrences lower than 10 is only 11%. Thus, MCDADs obtained in this study acquire enough accuracy to consider the probability of occurrence in the α region or the β and p regions.

The side-chain dihedral angles are chosen as follows: The dihedral angle of rotation around a bond between two sp^3 atoms is randomly chosen at the three angles of 60, 180, and 300° with a range of ±15°. The dihedral angle for sp^3 and sp^2 atoms is randomly chosen from the whole range of 0−360°. That for two sp^2 atoms is fixed to their optimum value. Exceptionally, those of χ_2, χ_3, and χ_4 for lysine and χ_2, χ_3, and χ_5 for arginine are randomly chosen within a range of 180 ± 15°.

Calculation of the SXS Profile, R_{sq}, and ASA. The SXS profile of a protein molecule is calculated from the atomic coordinates derived from its generated conformation using the following equation:

$$I(K) = \sum_{i=1}^{N_a} \{f_i(K)\}^2 + 2 \times \sum_{i>j} f_i(K) f_j(K) \frac{\sin Kr_{ij}}{Kr_{ij}} \qquad (1)$$

where K is the magnitude of the scattering vector, $f_i(K)$ is the scattering factor of atom i, r_{ij} is the distance between atoms i and j, and N_a is the number of atoms. Strictly, eq 1 is an expression for calculating the SXS profile of a molecule in a vacuum. An experimental SXS profile includes both the contrast effect of the solvent with finite electron density and the hydration effect due to differences in spatial distribution between hydration water and bulk water. As these effects of solvent water are not taken into account in eq 1, this formula cannot be applied to compare quantitatively with an experimental profile in the range of K where they have significant contributions. However, we could confirm that, for a protein with unfolded conformations generated in this study, the SXS profiles calculated by using eq 1 virtually agree with those obtained when including the two solvent effects[34,35] in the range of $K < 1.5$ nm^{-1} (data not shown). Hence, we can assume that eq 1 gives a good approximation for the SXS profile of unfolded protein. In this study, it is essential to calculate SXS profiles for protein conformations of a large number of 1 million per model. So, we limited the K range of analysis to the above and adopted eq 1, which requires the shortest time for estimating SXS profiles. For the same reason, the mean square radius (R_{sq}) is calculated using the following equation:

$$R_{sq} = \sqrt{\sum_{i=1}^{N} n_{e,i} r_i^2 \Big/ \sum_{i=1}^{N} n_{e,i}} \qquad (2)$$

where $n_{e,i}$ is the number of electrons in atom i and r_i is the position vector of atom i from the center of gravity of the protein.

The accessible surface area (ASA) of the protein is calculated analytically[36] using an originally developed program, "cgp". Details on the program cgp will be published elsewhere. We employed the set of atomic radii by Tsai et al.[27] and assumed the radius of probe water to be 0.14 nm.

RESULTS

Development of the Program for Generating Conformations of Unfolded Proteins. Previous methods for generating conformations of unfolded proteins are roughly classified into two types by the way of choosing its main-chain dihedral angles. In one method, main-chain dihedral angles are randomly chosen from the whole region of (φ, ψ).[37,38] After that, a possible high-energy conformation due to atomic collisions is relaxed and then brought into energy minimization. In the other method, some main-chain dihedral-angle distribution (MCDAD) derived from the database of native structures is utilized for generating unfolded conformations.[20] The former is inefficient for generating conformations, because the structural relaxation and energy minimization of a conformation with many atomic collisions requires a fairly long computational time, and its conformational energy often does not converge. The latter enables us to effectively obtain conformations with no intraresidue atomic collisions by excluding pairs of φ and ψ that lead to atomic collisions such as a pair of nearly zero φ and ψ values. As a result, the energy minimization in the latter is carried out far more easily compared with the former, because the generated conformation can have only inter-residue atomic collisions. Taking these into account, we have employed the following criteria for our new method of generating unfolded-protein conformations: (1) The all-atom model is adopted where all atomic species are explicitly considered. (2) All of the bond lengths and bond angles are fixed to their respective optimum values. (3) The occurrence probability of a main-chain dihedral angle is assumed to be given by a MCDAD derived from an ensemble of native-protein structures, except that all ω's are fixed to 180° and the φ of proline is fixed to −75°. (4) The side-chain dihedral angle is randomly chosen from the conformations near its stereochemically most stable structure (see the Methods

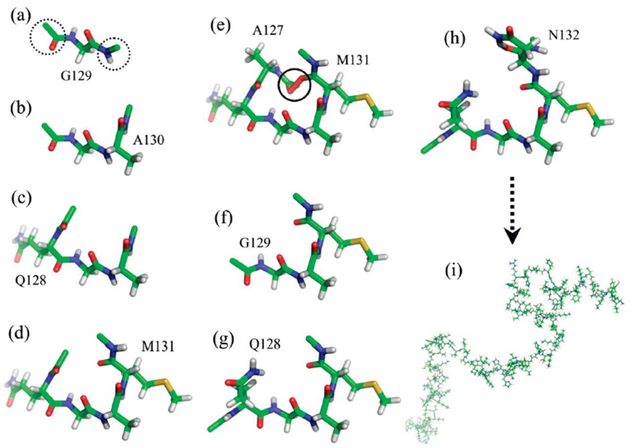

Figure 2. Algorithm for generating unfolded protein structures. An example for apomyoglobin is shown in Figure 2a–i. (a) G129 is chosen as the first residue. The coordinates of main-chain atoms of the residues neighboring the first residue, i.e., C_α, C, and O on the N-terminal side and N, H, and C_α on the C-terminal side, are calculated simultaneously, as indicated by dotted circles. These atoms determine the dihedral angles of rotation, ω, around the peptide bonds on both sides. They are generated also when the neighboring residues are generated. The neighboring residues are connected by superimposing the doubly generated atoms. (b) A130 is chosen as the next residue and added to the first residue. (c and d) Q128 and M131 are added at steps of c and d, respectively. (e) The addition of A127 has failed due to the atomic collision between the two main-chain carbonyl oxygen atoms of D126 and M131, as indicated by a circle. If the number of regeneration trials exceeds a limit, the process is restarted with returning to the step before adding several residues. In this study, the former limit is set at 1000 trials, and the latter limit is set at three residues. We confirmed that changes in the two technical parameters affect the rate of generating conformations but have practically no effect on the structural features of generated conformations. As the number of regeneration trials reached 1000 in the case of e, G129, Q128, and A127 were deleted, as shown in f. (f, g, and h) The conformations of G129 and Q128 are regenerated, and N132 is added at the respective steps. (i) The resultant whole structure is shown.

section for details). (5) All atomic collisions are eliminated between the two non-hydrogen atoms separated from each other by more than three chemical bonds. Criteria 1 and 2 are reasonable and have also been employed in many previous methods. We propose a new algorithm for eliminating atomic collisions to satisfy criterion 5.

The details of our algorithm are described below, and some examples of generated conformations are shown in Figure 2a–i. First, the first residue is randomly chosen from among all of the residues in the given sequence with equal probability. Next, main- and side-chain dihedral angles of the first residue are chosen by using the method described above. Then, the Cartesian coordinates of atoms are calculated from these dihedral angles, bond lengths, and angles. Here, the conformation of the first residue is examined for atomic collisions. In the case of some atomic collisions, the step of choosing dihedral angles is repeated again. An example of the conformation thus obtained is shown in Figure 2a. Next, a residue on either of the N- and C-terminal sides is chosen for adding to the first residue. Coordinates of the atoms in the added residue are calculated and examined for intraresidue atomic collisions. The generated residue is connected to the first residue so as to form a peptide bond (see the caption of Figure 2). Possible atomic collisions between the added and first residues are examined. If no atomic collision is found, we move on to choosing a new residue to be added. As shown in Figure 2b–d, the peptide chain is grown alternately toward either the N-terminal or C-terminal side. When the newly added residue is either the N- or C-terminal residue of the protein, the next residue is naturally chosen so as to grow the chain only in the direction of the segment not yet generated. In this way, the conformation of a polypeptide chain is generated so that it can grow toward both termini. As shown in Figure 2e, when there are atomic collisions between the added residue and the existing chain, the process is restarted from the step of choosing dihedral angles of the added residue. In the case when atomic collisions can never be eliminated by the above means, the process is restarted with returning to the step before adding several residues. Eventually, as shown in Figure 2i, the conformation of the unfolded protein is obtained, and it meets the five criteria described above.

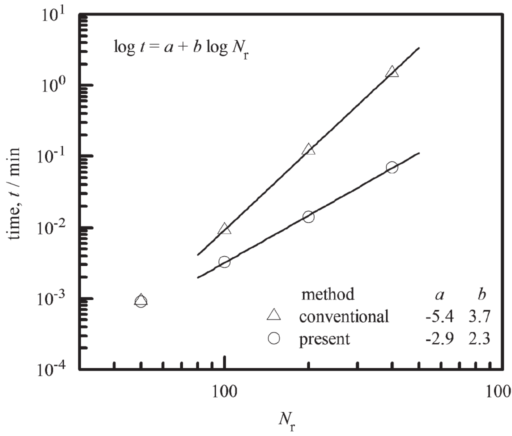

Figure 3. Comparison of dependences of the mean time, t, required for generating one conformation of an unfolded alanine-homopolypeptide on the number of residues, N_r, between the present method (○) and the conventional one (△). A function of the form of $\log t = a + b \log N_r$ is least-squares fitted to each set of data for N_r= 100, 200, and 400. The values of the parameters obtained are shown in the figure. The standard errors of b are 0.02 and 0.05 for the conventional and present methods, respectively. R^2 is greater than 0.999 for both methods. MCDAD model, "all"; CPU, Intel Xeon processor (2.8 GHz).

A unique feature of the present algorithm is that the peptide chain grows by residue from an arbitrarily chosen residue. Another feature is that only a small number of residues are remodeled when encountering atomic collisions. As a result, it is expected that the computational cost for examining atomic collisions and remodeling chain conformations is minimized. To test the performance of our method, we examined how much the computational time for generating conformations of unfolded protein is decreased by using this algorithm. A conventional method not using the growing-chain algorithm was adopted as a reference method. Specifically, the mean computational times were compared between the cases where the respective methods were employed for generating unfolded chains of polyalanine with four different chain lengths of N_r = 50, 100, 200, and 400. In the conventional method, the process of eliminating atomic collisions is carried out after main-chain dihedral angles of the whole chain are chosen. Except for this point, there is no distinction between the two methods. The results obtained are shown in Figure 3. The two computational times are nearly the same for a chain of 50 residues. The computational time with our method, however, is much shorter than that with the conventional one for the other longer chains: Those with the

Table 1. Fractions (%) of the α and β Regions in MCDADs from the Database and Generated Conformations

	α region[a]/%		β region[a]/%	
MCDAD	database[b]	model[c]	database[b]	model[c]
(α + t + c)	55.1	48.8	30.0	36.3
(all) or (α + β + t + c)	44.0	37.5	43.6	50.5
(t + c)	24.4	24.0	52.2	57.0
(β + t + c)	17.3	16.8	65.5	69.1
(coil) or (c)	6.9	7.6	79.5	80.1
(β + c)	4.5	4.7	86.9	87.2

[a] The definition of the (φ, ψ) regions is given in ref 33. The β region in this paper is a combination of the β and p regions in ref 33. [b] These are estimated from 55 012 sets of (φ, ψ) values obtained from 379 native-protein structures. [c] These are estimated from all of the conformations generated with 12 unfolded proteins. One million conformations are generated for each set of a MCDAD model and a protein species.

former and the latter increase with an increase in the number of residues, N_r, with exponents of 2.2 and 3.6, respectively. For example, generating 1 million conformations of 400-residue polyalanine with the conventional method requires a 20 times longer computational time than that with our method. Hence, the difference in the process avoiding atomic collisions results in a great difference in the speed of generating conformations.

We confirmed that (a) almost all of the potential energies of generated conformations reach their local minima after steps smaller than 500 and (b) the average difference in conformation between those before and after energy minimization is smaller than 0.025 nm in RMSD. These small conformational changes lead to practically no significant changes in the conformational parameters such as R_{sq} and ASA in this study. Hence, we analyzed the conformation of unfolded proteins, skipping the step of energy minimization. Consequently, our newly developed method can generate reasonable conformations without energy minimization. In conclusion, it enables us to generate a large number of conformations of an unfolded protein with reasonable potential energy very rapidly by eliminating atomic collisions efficiently.

Structural Features of Unfolded Proteins. In the following, we will call each of the different MCDADs a MCDAD model. To analyze the structural features of chains derived from different MCDAD models, we generated 1 million unfolded conformations for each of the 12 proteins with different N_r's using six MCDAD models. As a result, we obtained 72 different ensembles, each of which consists of 1 million unfolded conformations. The ensemble of conformations generated with a MCDAD model will hereafter be called briefly the MCDAD ensemble. The fractions of α and β regions in each MCDAD and those for the ensemble of conformations generated from 12 unfolded proteins are shown in Table 1. Here, the α-region fraction is defined as the probability that main-chain dihedral angles are chosen in the α region.[33] The β-region fraction is defined similarly to the α-region fraction, but the β region in this study is defined as the sum of the β and p regions in the literature.[33] The α-region fraction for generated unfolded proteins is hereafter denoted as η_α (in percent). Each of the generated MCDADs reproduces well the α-region fraction for the corresponding model derived from PDB except for the (α + t + c) and (all) models. These two MCDADs having a high η_α give about a 6% lower α-region fraction than those from the database. Nevertheless, they have a significantly higher η_α's than the other MCDAD models for structural analysis. Such consequences for these two models are caused by the fact that atomic collisions occur more frequently when main-chain dihedral angles are chosen in the α region than in the other regions. As a result, the number of dihedral angles chosen from the other regions increases for avoiding atomic collisions. To cope with this problem, an algorithm for further exploring a dihedral angle without atomic collisions in the α region must be introduced into our method. There is a trade-off between the effectiveness of this means and the speed of generating conformations. In this study, we have taken no precautions to meet the problem, taking account of the speed of generating conformations. To minimize the effect of this problem, not the α-region fraction from the database but the value of η_α from the generated ensemble is used for comparing calculations with experimental results. Thus, the value of η_α obtained from analysis correctly reflects structural features of target proteins. The *ASA* (accessible surface area), R_{sq} (mean square radius), and SXS (solution X-ray scattering) profile were calculated for all of the conformations in each MCDAD ensemble.

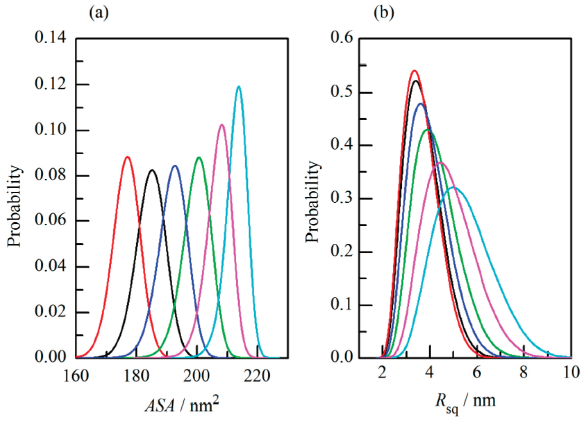

Figure 4. Probability distribution of the (a) *ASA* and (b) R_{sq} of unfolded apomyoglobin. MCDAD model: red, (α + t + c); black, (all); blue, (t + c); green, (β + t + c); purple, (coil); light blue, (β + c). Each of the six MCDAD ensembles consists of 1 million conformations. Division of data: (a) ΔASA = 0 nm^2, (b) ΔR_{sq} = 0.2 nm. Estimated error of probability density: (a) 3×10^{-4}, (b) 2×10^{-3}. The curves in the figure are drawn using the cubic spline interpolation.

To see structural features of the generated unfolded conformations, probability distributions of the *ASA* and R_{sq} in the respective MCDAD ensembles for unfolded apomyoglobin are shown in Figure 4.

We can see from Figure 4a that the bell-shaped ASA distributions for different models have their respective different peak positions, and their tails mutually overlap with those of nearby models. The values of *ASA* for all of the MCDAD models are distributed all over the range of 170−220 nm^2. The ASA of fully extended apomyoglobin is 213 nm^2, which is found in its range of the (β + c) ensemble. To compare our estimate with the previous one for the *ASA* of unfolded protein, we evaluated the *ASA* of unfolded horse apomyoglobin using the method of estimating its upper and lower limits described by Creamer et al.[39,40] It yielded estimates for the lower and upper limits of 158 nm^2 and 203 nm^2, respectively. This lower limit is found in the minimum *ASA* region of the (α + t + c) ensemble that gives the smallest *ASA* among our conformational ensembles. Considering that the lower limit is estimated from the secondary-structure segment of the crystal structure in their method, we can

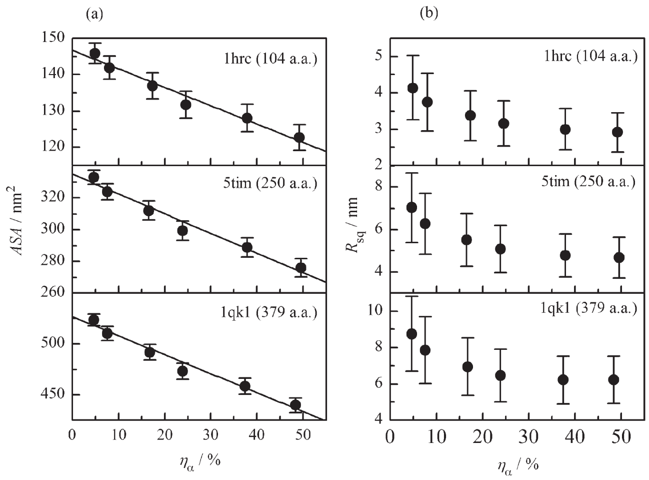

Figure 5. Dependences of the (a) mean *ASA* and (b) mean R_{sq} on the α-region fraction, η_α, of the protein. A least-squares fit of a linear function, $ASA(\eta_\alpha) = ASA(0) - a \times \eta_\alpha$, is made on each data point of a. The two obtained parameters, *ASA*(0) and *a*, their standard errors, and R^2 are listed in Table 2.

safely assume that the unfolded conformation generated by us gives the lower limit of *ASA* for the unfolded protein having no local structure such as secondary structures. On the other hand, we can see that our ensemble covers broader conformational space than theirs, as the upper limit is found near the peak of the (coil) ensemble.

Evidently, the R_{sq} also covers conformational space extensively, as seen from its distribution shown in Figure 4b. However, the dependence on the MCDAD ensemble of the R_{sq} distribution differs greatly from that of the *ASA* distribution: All of the MCDAD ensembles have a common range of R_{sq} between 3 and 6 nm, and with an increase in the R_{sq} at the peak position, the peak height of the R_{sq} distribution decreases and the width of distribution increases. It is also notable that the two R_{sq} distributions for the (α + t + c) and the (all) ensembles are very similar to each other, though their *ASA* distributions are completely separated. It is strongly suggested that the lower limit of R_{sq} for unfolded apomyoglobin is near 2 nm because all of the ensembles show almost the same lower limit of about 2 nm.

To see the relation of the global-structure parameters of *ASA* and R_{sq} to the MCDAD in more detail, dependences of the mean *ASA* and R_{sq} for the six MCDAD ensembles on their α-region fraction, η_α, are shown in Figure 5. As seen from Figure 5a, the mean ASA decreases monotonically with an increase in η_α. Applying the least-squares fit of a linear function to this plot, we can see that the standard errors of the fitting parameters for each protein are sufficiently small, as shown in the caption of Figure 5. It confirms that the dependence of the mean *ASA* on η_α is well approximated by a linear function. In addition, the slope of this line, i.e., the reduction rate of *ASA* with increasing η_α, is well proportional to N_r, and the determination coefficient, R^2, obtained from the least-squares fit of this relation is greater than 0.999. We can see from the above that the mean *ASA* of the unfolded protein decreases by 0.49 ± 0.06 nm^2 per 100 residues with a 1% increase in η_α. On the other hand, we found that, though the mean R_{sq} decreases with increasing η_α for the MCDAD ensemble having a η_α lower than 30%, it is hardly changed with an η_α higher than 30%. The same dependence was observed for all sizes of proteins taken in this study. This result strongly suggests that there is a lower limit of R_{sq} for each unfolded protein.

Table 2. ASA(0) and *a*, Their Standard Errors, and R^2

PDB ID	a/nm^2 %$^{-1}$	ASA(0)/nm^2	R^2
1hrc	−0.49 ± 0.04	146 ± 1	0.963
5tim	−1.19 ± 0.10	333 ± 3	0.966
1qk1	−1.81 ± 0.15	524 ± 4	0.966

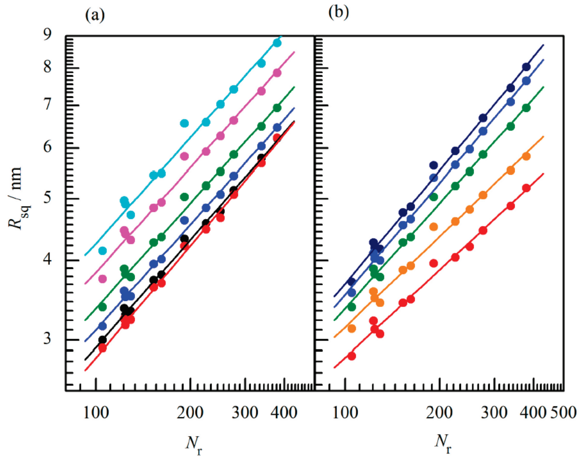

Figure 6. Dependence of the mean R_{sq} of unfolded proteins on N_r. (a) Variation of the mean R_{sq} with the MCDAD ensemble. The correspondence of ensembles to colors is the same as in Figure 4. (b) Variation of the weighted mean R_{sq} for the (*β* + t + c) ensemble with solvation effects. The weighted mean R_{sq} is obtained using the Boltzmann factor at *T* = 298.15 K including solvation free energy, where it is assumed that the solvation free energy is proportional to the ASA. Proportional constants, *σ*: −10 (deep blue), −5 (blue), 0 (green), +5 (orange), +10 (red) J mol^{-1} Å^{-2}. The straight lines are best fitted to respective data by using eq 3. The obtained parameters, standard errors, and R^2 in a and b are listed in Table 3. Results of a fitting analysis are shown in Figure 8 for the combinations of MCDAD and *σ*, including those above.

The scaling law, i.e., the dependence on N_r, of R_{sq} is a very important relation for characterizing the structure of an unfolded protein. For random-coil chains consisting of $N_r(= N + 1)$ units (*N* is the number of virtual bonds), the following equation holds with a sufficiently large *N*:[41]

$$R_{sq} = R_0 N^{\nu} \approx R_0 {N_r}^{\nu}, N_r \gg 1 \tag{3}$$

where *ν* is the scaling exponent. The value is 0.5 for ideal random-flight chains without an excluded volume effect. The value of *ν* is theoretically estimated to be about 0.6 for random-coil chains with an excluded volume effect.[42,43] The R_0 is a quantity with a dimension of length.

To examine the scaling parameters for our MCDAD ensembles, dependences of the mean R_{sq} on N_r are shown in Figure 6. As seen from Figure 6a, the values of *ν* for the respective models are near the range of 0.55−0.57. These values are intermediate between those for ideal random-flight chains (*ν* = 0.5) and random-coil chains with an excluded-volume effect (*ν* ≃ 0.6). As no significant correlation is observed between the vales of η_α and *ν*, we can see that the value of *ν* does not depend effectively on MCDAD. On the other hand, the value of R_0 varies significantly with MCDAD and decreases monotonically with increasing η_α.

These results agree completely with the prediction from polymer theory.[41,42]

The dependence of R_{sq} on N_r is shown in Figure 6b, where the values of R_{sq} for the respective conformations are averaged by weighting them with the solvation effect in the *ASA* approximation. It is assumed here that the solvation energy is proportional to the *ASA* of the protein with a proportional constant, σ. The statistical-thermodynamically averaged R_{sq} is obtained by weighting the occurrence probability of each generated conformation with the Boltzmann factor at T = 298.15 K for the solvation energy. As seen from Figure 6b, the exponent ν, which is given by the slope of the line, decreases from 0.6 to 0.45 with the above increase in σ. It shows clearly that solvation significantly affects the scaling exponent ν in our MCDAD model, which agrees with the result of Wang et al.[10] on the scaling law analysis.

The calculated Kratky profiles of SXS, which provide important information on the molecular shape, are shown in Figure 7.

Table 3. Parameters, Standard Errors, and R^2 for (a) Variation of the Mean R_{sq} with the MCDAD Ensemble and (b) Variation of the Weighted Mean R_{sq} for the (β + t + c) Ensemble with Solvation Effects

(a) Variation of the Mean R_{sq} with the MCDAD Ensemble			
MCDAD	ν	R_0/nm	R^2
(α + t + c)	0.58 ± 0.02	0.20 ± 0.02	0.996
(all)	0.56 ± 0.02	0.21 ± 0.02	0.996
(t + c)	0.55 ± 0.02	0.24 ± 0.02	0.994
(β + t + c)	0.55 ± 0.02	0.31 ± 0.03	0.991
(coil)	0.57 ± 0.03	0.48 ± 0.05	0.983
(β + c)	0.57 ± 0.04	0.64 ± 0.06	0.976

(b) Variation of the Weighted Mean R_{sq} for the (β + t + c) Ensemble with Solvation Effects			
σ/J mol^{-1} Å^{-2}	ν	R_0/nm	R^2
−10	0.59 ± 0.02	0.24 ± 0.03	0.994
−5	0.58 ± 0.02	0.24 ± 0.03	0.993
0	0.55 ± 0.02	0.26 ± 0.03	0.991
5	0.48 ± 0.02	0.35 ± 0.05	0.985
10	0.46 ± 0.02	0.34 ± 0.04	0.987

As seen from Figure 7a, the difference in MCDAD has considerable effects on the Kratky profile of the unfolded protein. There are significant effects not only in the small K region of K < 0.5 nm^{-1} essential for R_{sq} but also in the intermediate K region of 0.5 < K < 2.0 nm^{-1}. In these K regions, a MCDAD ensemble with a higher value of η_α yields a higher scattering intensity. On the other hand, the solvation effect on the Kratky profile is very different from the effect of MCDAD, as shown in Figure 7b. Introduction of the solvation effect considerably changes the estimate of scattering intensity at 0.5 < K < 1.0 nm^{-1}. Especially, there appears a peak in the profile under poor solvent conditions of σ > 0. On the other hand, the difference in scattering intensity is fairly small at 1.5 < K < 2.0 nm^{-1}. All of the Kratky profiles for apomyoglobin nearly coincide with each other at K = 1.8−2.0 nm^{-1}. Naturally, we can see that the changes of MCDAD and solvation give qualitatively different effects on the Kratky profile.

Structural Analysis of Highly Denatured Proteins. As described above, we could confirm that the differences among various MCDAD models adopted here are strongly reflected in the value of R_0 in the scaling law and the shape of the Kratky profiles, to which attention had hardly been paid previously. It is expected that the true MCDAD model can be predicted from comparing quantitatively the calculated Kratky profiles with the experimental ones by using the above result. As a test for the possibility, we will compare quantitatively experimental data on a protein in the highly unfolded state in an aqueous concentrated solution of denaturant with our predictions from various MCDAD models. The solid and the broken lines in Figure 8 indicate the scaling-law parameter and the range of estimated errors respectively obtained from an analysis of the reported experimental values of R_{sq} for proteins in the highly unfolded state.[7] As evident from Figure 8a, only the ensembles weighted with negative σ, i.e., assuming good solvent conditions, can reproduce the experimental value of ν. In addition, the experimental value of ν can be reproduced independently of the values of η_α. On the other hand, as shown in Figure 8b, not only the solvation effect but also the value of η_α must be limited to reproduce the experimental value of R_0. The prediction for scaling parameters from an ensemble of generated conformations agrees with experimental results within experimental error only when the η_α of its MCDAD is higher than about 20%. In other words,

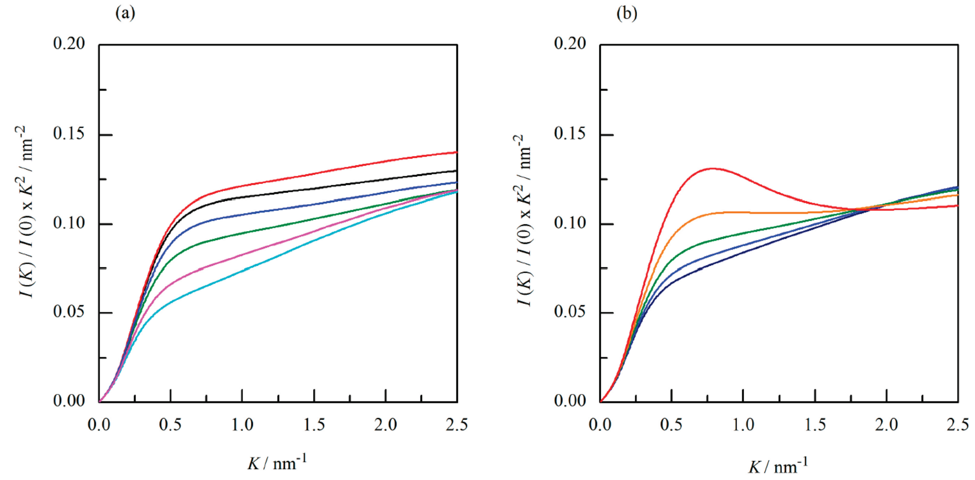

Figure 7. Kratky profiles for random chain models of apomyoglobin. The colors of lines in a and b correspond to those in Figure 6a and b, respectively.

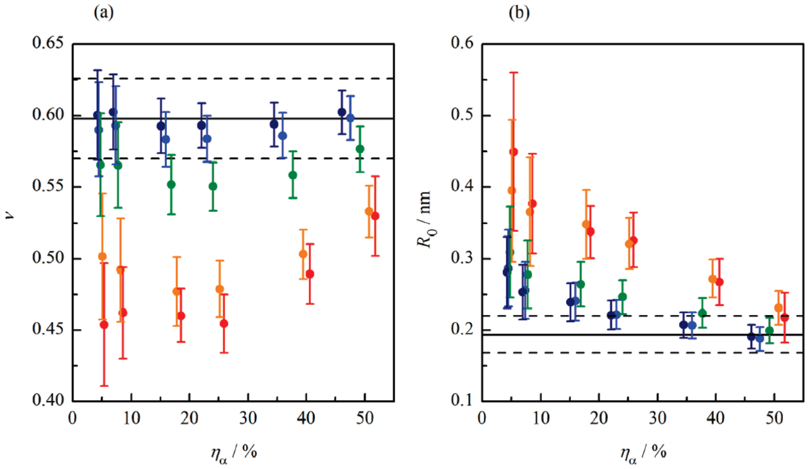

Figure 8. Comparison of experimental and calculated scaling parameters. Variation of the dependence of (a) ν and (b) R_0 on the α-region fraction, η_α, with the solvation effect, where ν and R_0 are obtained from the scaling-law analysis of MCDAD ensembles. The color of the data points corresponds to each value of σ characterizing the solvation effect, as shown in Figure 6b. The experimental values[7] of ν and R_0 and their respective standard deviations for highly unfolded proteins are shown by solid and broken lines: $\nu = 0.598 \pm 0.028$ and $R_0 = 0.192^{+0.271}_{-0.238}$ Å.

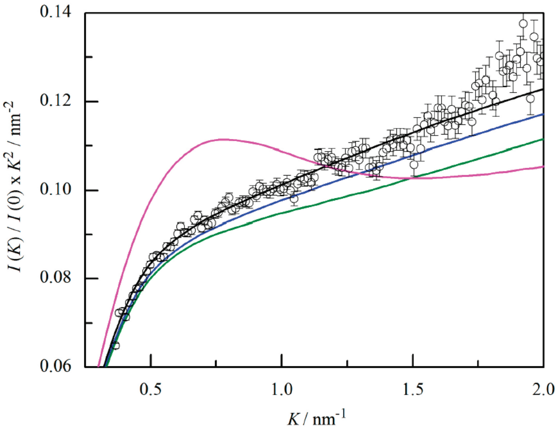

Figure 9. Comparison of the experimental and calculated SXS profiles for unfolded apomyoglobin. Open circles indicate the experimental profile under 5 M urea conditions,[9] and solid lines indicate the profiles estimated with various MCDAD ensembles. The MCDAD model and solvation energy for each estimated profile are as follows: black, (all) and $\sigma = -10$; blue, (t + c) and $\sigma = -5$; green, (β + t + c) and $\sigma = 0$; purple, (coil) and σ = 10 J mol^{-1} Å^{-2}.

the ensembles with lower values of η_α such as the (β + c), (coil), and (β + t + c) models cannot reproduce experimental results.

To obtain more detailed information about the protein structure in the highly unfolded state, the Kratky profile for the urea unfolded state of apomyoglobin[9] and those for the generated ensembles of the conformation are compared in Figure 9. The figure shows only profiles for the four ensembles that yield R_{sq} close to the experimental one among the 30 ensembles generated by combining six MCDAD models with five kinds of solvation effects. The experimental profile lies between those for the (all$_{\sigma=-10}$) and (t = c$_{\sigma=-5}$) ensembles at $K < 1.5$ nm^{-1}. We cannot make a decision on the validity of a model by comparing experimental and computational profiles at $K > 1.5$ nm^{-1}, because the present method of analysis cannot estimate the solvent effect on SXS intensity with enough accuracy. The Kratky profile for the (β + t + c$_{\sigma=-0}$) ensemble, which has an η_α lower than those of the two ensembles above, has a similar shape to the experimental profile but deviates downward in the whole K region as its R_{sq} is larger than the experimental one. Adding a weakly poor solvation effect of $\sigma > 0$ to decrease the R_{sq} of the ensemble, the shape of the Kratky profile deviates from the experimental one to approach that of the (coil$_{\sigma=10}$) ensemble. To summarize the above, the ensemble reproducing the experimental Kratky profile for urea unfolded apomyoglobin is a MCDAD ensemble for either the (all) or (t + c) model under good solvent conditions of $\sigma = -10$ to -5 kJ mol^{-1} Å^{-2}. These results indicate that the possible MCDAD model needs to be further limited to reproduce the SXS profile in addition to the scaling law. Thus, we could predict not only the value of R_{sq} but also the solvation effect and MCDAD model of an unfolded protein more definitely by comparing the estimated Kratky profile with the experimental one.

DISCUSSION

In the algorithm developed in this study for generating unfolded-protein conformations, we make use of the probability distribution of main- and side-chain dihedral angles in native proteins and completely eliminate the atomic collisions between neighboring residues. As a result, it is expected that the conformational distribution of amino acid residues and the conformational correlation among them are both adequately considered. It is for this reason that we could obtain good agreement between the computation and experimental results for the parameters reflecting the structural characteristics of a protein at a short distance. On the other hand, the effect of cooperative interactions between residues at long distances, which is essential to the higher-order structure formation in a protein, cannot be taken into account. For example, an α helix is generated by its nucleation accompanying a large entropy loss followed by the cooperative incorporation of helical residues accumulating small enthalpy gains. However, such a structure can hardly be generated by our method. It assumes implicitly that the target is limited to an ensemble of conformations having practically no cooperative local structure, such as those of unfolded proteins in a concentrated denaturant solution. Even if the solvation effect is included, we can hardly generate any cooperative structure by this method. It is almost inapplicable to modeling transiently denatured structures such as those in the intermediate state of protein folding, because they are formed only under the solution condition that the equilibrium is strongly biased toward higher-order structure formation. Similar situations are also expected with denatured structures at high temperatures, high pressures, and extreme pH. To generate conformations for a protein denatured under these conditions, it will be necessary to develop an algorithm that can explicitly consider the formation of higher-order structures.[44] Conversely, it is expected that this method can be applied effectively to the structural analysis of natively unfolded proteins that can hardly form a higher-order structure by itself.

While both *ASA* and R_{sq} are important parameters characterizing the structure of the unfolded protein, the α-region fraction, η_α, is a quantity reflecting the characteristics of each MCDAD. Analyzing the interrelation between the former two and the latter one, we found that the change in η_α is reflected sensitively in the change in *ASA*. As shown in Figure 5a, the mean *ASA* decreases with an increase in η_α. We will discuss the reason why this relation holds in the following. As the bond length and bond angle of a peptide chain, and the dihedral angle of its peptide bond, ω, are nearly constant, the angle formed by two neighboring virtual bonds is predominantly determined by a pair of main-chain dihedral angles, φ and ψ. When a pair of main-chain dihedral angles of a residue is in the α region, the virtual-bond angle formed by the neighboring residues is about 90°, which is smaller than 120°, i.e., the corresponding angle for a residue in either the β or p region. Therefore, the magnitude of η_α is reflected in the degree that the main-chain is sharply bent. It is easily understood that a residue with its main-chain sharply bent tends to approach neighboring residues and decrease the mean *ASA* of the protein. In fact, we can see from Figure 5a that the decrease in *ASA* with an increase in η_α is in very good proportion to the N_r of the protein. We can conclude from the above that the difference in mean *ASA* between different MCDAD ensembles in the highly unfolded state is mainly brought about by local effects of the difference in the degree of spatial proximity between neighboring residues.

On the other hand, the relation between R_{sq} and η_α is not so straightforward. A change in MCDAD directly brings about a change in the local structure, but it must affect indirectly the values of global structural parameters such as R_{sq}. Interestingly, with regard to the scaling law of R_{sq}, the change in η_α does not affect the value of ν only to change the value of R_0. As shown in Figure 8b, it decreases monotonically with increasing η_α under the same solvent conditions. This dependence of R_0 on η_α will reasonably be explained as follows: As an increase in the η_α of a polypeptide chain increases, the degree of its chain bending, the R_{sq} for an ensemble of chains with a higher probability of sharply bent conformations is smaller than those with lower probability. This change in R_{sq} only affects the value of R_0 in the low scaling, because ν hardly depends on η_α under the same solvent conditions. As the chains in any MCDAD ensemble have a value of ν larger than 0.5 and are not ideal random-flight ones, it is difficult to have more detailed, quantitative discussions about the R_0 for MCDAD ensembles. Thus, the effect of differences in MCDAD appears only in R_0 on the scaling law of R_{sq}. This is in contrast to the incorporation of solvation effects by the *ASA* approximation: It modifies the occurrence probability of each conformation according to the value of its *ASA*, which changes the distribution of the spatial extension of chains to result in a change in the value of ν.[10]

We have analyzed the structure of a highly unfolded protein to confirm the usefulness of conformational ensembles generated from diverse MCDADs. In practice, comparison has been made between experimental and computational estimates for the parameters, ν and R_0, of the scaling law of R_{sq} for highly unfolded proteins and the SXS profile for urea unfolded apomyoglobin. As a result, it has become evident that the highly denaturing solvent is a good solvent with $\sigma = -10$ to -5 kJ mol^{-1} Å^{-2}, and the η_α for the MCDAD of highly unfolded proteins is in the range of 20−40%. The result of our analysis on the solvation effect agrees qualitatively with that of Wang et al.[10] It is also in accord with the widely accepted view that denaturant molecules interact directly with the peptide groups of a protein molecule.[45,46] However, the present method incorporates solvation effects only through the *ASA* of the whole protein molecule not taking into account the variety of its polar atomic groups. Hence, with this method, we cannot discuss the detailed molecular mechanism of the interactions involved in protein denaturation such as the hydrogen bonding of main-chain polar groups[38] and the hydrophobic interaction of side-chain nonpolar groups[47] with water and denaturant molecules. In addition, even if the difference in the magnitude of σ between the polar and nonpolar groups is explicitly considered, it is very difficult to estimate their respective magnitudes of σ with high accuracy, because the differences in the proportion of their respective surface areas are fairly small among the target proteins taken in this study.

On the other hand, it is necessary to discuss more carefully the α-region fraction. The interrelation between the $^3J_{HN\alpha}$ coupling constant and the MCDAD has been analyzed for unfolded proteins. Smith et al.[12] determined the correlation coefficient, γ, between two $^3J_{HN\alpha}$ coupling constants of each residue, which were measured experimentally and estimated using MCDAD. They concluded that the structural ensemble of the "COIL" MCDAD (γ = 0.92) reproduces experimental coupling constants for short peptides having no secondary structure better than the ensemble of the "ALL" MCDAD (γ = 0.81). Their MCDAD models are constructed from the probability distribution of dihedral angles of native proteins similarly to our models. They reported that the values of η_α for the COIL and ALL models are 28% and 45%, respectively. These estimates for the η_α agree fairly well with our estimates of 20−40% in this study. Analyzing ubiquitin denatured in aqueous 8 M urea solution, Peti et al.[13] confirmed that the ensemble for the COIL model reproduces experimental results with a high correlation coefficient of γ = 0.96. Thus, the value of η_α estimated from our analysis for proteins highly unfolded in concentrated denaturant solution agrees with the result of analysis of NMR coupling constants. These results confirm that our analysis gives reasonable estimates for the η_α of the unfolded protein. However, the information on MCDAD derived from our analysis and NMR coupling constants is inadequate in particular with respect to the dependence on amino acid sequence. It will be necessary to verify the validity of MCDAD models by using more efficient methods such as the residual dipolar couplings method.[19,20]

From the viewpoint of molecular biophysics, the structure of a biomolecule can be described not by mechanics based on the minimum potential energy but by statistical thermodynamics based on the minimum free energy, including both enthalpic and entropic contributions. It is a well-known fact that all of the real, physical structures are continually fluctuating by various degrees. In this sense, the structure of a protein molecule consists of an ensemble of multiple structures not only in the unfolded state but also in the intermediate and native states. Especially, an unfolded protein molecule, which is the target of this study, has remarkable structural characteristics such as those above. It seems to be impossible to characterize the structure of proteins in the highly unfolded state. In spite of it, our study revealed that the value of η_α for seemingly random-flight polypeptide chains in high-concentration denaturant solution is confined to a fairly narrow range. This result shows clearly that even the highly unfolded protein exhibits structural features inherent in proteins. So, it is expected that this type of study will serve as a step for clarifying the local structural characteristics of proteins in the unfolded state.

Our method of analysis using the conformational ensemble of various MCDAD models has the potential to develop into a method for more detailed analysis of unfolded proteins. For realizing the possibility, it is necessary to study various factors involved in determining the SXS profile in the intermediate and large K region of $0.5 < K < 2.0\ \mathrm{nm}^{-1}$. To expand the possibility of comparing computation with experimental results quantitatively, it is necessary to obtain the following: (1) precise SXS data in the wide range of scattering angles including the forward scattering intensity[9] and (2) a method of accurately estimating SXS profiles that is applicable to unfolded proteins and incorporates solvent effects explicitly.[34,35]

■ AUTHOR INFORMATION

Corresponding Author

*Tel.: +81-78-940-5645. Fax: +81-78-304-4958. E-mail: soda@riken.jp.

■ ACKNOWLEDGMENT

Y.S. and K.S. express sincere thanks to Drs. Nobuhiro Go and Mitiko Go for their continuous support for our research work. Y.S. is very grateful to Dr. Shun-ichi Kidokoro for his support. This work was supported by a Grant-in-Aid for Young Scientists (B) (number 21700322) and for Scientific Research on Innovative Areas "Intrinsically Disordered Proteins" (number 22113507) of the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

■ REFERENCES

(1) Dyson, H.; Wright, P. Insights into the structure and dynamics of unfolded proteins from nuclear magnetic resonance. *Adv. Protein Chem.* **2003**, *62*, 311.

(2) Kataoka, M.; Goto, Y. X-ray solution scattering studies of protein folding. *Fold. Des.* **1996**, *1*, R107.

(3) Tompa, P. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett.* **2005**, *579*, 3346.

(4) Mittag, T.; Forman-Kay, J. D. Atomic-level characterization of disordered protein ensembles. *Curr. Opin. Struct. Biol.* **2007**, *17*, 3.

(5) Eliezer, D. Biophysical characterization of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **2009**, *19*, 23.

(6) Wilkins, D. K.; Grimshaw, S. B.; Receveur, V.; Dobson, C. M.; Jones, J. A.; Smith, L. J. Hydrodynamic Radii of Native and Denatured Proteins Measured by Pulse Field Gradient NMR Techniques. *Biochemistry* **1999**, *38*, 16424.

(7) Kohn, J.; Millett, I.; Jacob, J.; Zagrovic, B.; Dillon, T.; Cingel, N.; Dothager, R.; Seifert, S.; Thiyagarajan, P.; Sosnick, T.; Hasan, M.; Pande, V.; Ruczinski, I.; Doniach, S.; Plaxco, K. Random-coil behavior and the dimensions of chemically unfolded proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 12491.

(8) Semisotnov, G.; Kihara, H.; Kotova, N.; Kimura, K.; Amemiya, Y.; Wakabayashi, K.; Serdyuk, I.; Timchenko, A.; Chiba, K.; Nikaido, K.; Ikura, T.; Kuwajima, K. Protein globularization during folding. A study by synchrotron small-angle X-ray scattering. *J. Mol. Biol.* **1996**, *262*, 559.

(9) Seki, Y.; Tomizawa, T.; Hiragi, Y.; Soda, K. Global structure analysis of acid-unfolded myoglobin with consideration to effects of intermolecular Coulomb repulsion on solution X-ray scattering. *Biochemistry* **2007**, *46*, 234.

(10) Wang, Y.; Trewhella, J.; Goldenberg, D. P. Small-Angle X-ray Scattering of Reduced Ribonuclease A: Effects of Solution Conditions and Comparisons with a Computational Model of Unfolded Proteins. *J. Mol. Biol.* **2008**, *377*, 1576.

(11) Uzawa, T.; Kimura, T.; Ishimori, K.; Morishima, I.; Matsui, T.; Ikeda-Saito, M.; Takahashi, S.; Akiyama, S.; Fujisawa, T. Time-resolved Small-angle X-ray Scattering Investigation of the Folding Dynamics of Heme Oxygenase: Implication of the Scaling Relationship for the Submillisecond Intermediates of Protein Folding. *J. Mol. Biol.* **2006**, *357*, 997.

(12) Smith, L. J.; Bolin, K. A.; Schwalbe, H.; MacArthur, M. W.; Thornton, J. M.; Dobson, C. M. Analysis of Main Chain Torsion Angles in Proteins: Prediction of NMR Coupling Constants for Native and Random Coil Conformations. *J. Mol. Biol.* **1996**, *255*, 494.

(13) Peti, W.; Hennig, M.; Smith, L.; Schwalbe, H. NMR spectroscopic investigation of ψ torsion angle distribution in unfolded ubiquitin from analysis of $^3J(\mathrm{C}\alpha,\mathrm{C}\alpha)$ coupling constants and cross-correlated $\Gamma^{\mathrm{C}}_{\mathrm{HNN,C}\alpha\mathrm{H}\alpha}$ relaxation rates. *J. Am. Chem. Soc.* **2000**, *122*, 12017.

(14) Eliezer, D.; Yao, J.; Dyson, H.; Wright, P. Structural and dynamic characterization of partially folded states of apomyoglobin and implications for protein folding. *Nat. Struct. Biol.* **1998**, *5*, 148.

(15) Weinstock, D. S.; Narayanan, C.; Baum, J.; Levy, R. M. Correlation between $^{13}\mathrm{C}^{\alpha}$ chemical shifts and helix content of peptide ensembles. *Protein Sci.* **2008**, *17*, 950.

(16) Shi, Z.; Olson, C.; Rose, G.; Baldwin, R.; Kallenbach, N. Polyproline II structure in a sequence of seven alanine residues. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 9190.

(17) Shi, Z.; Chen, K.; Liu, Z.; Kallenbach, N. Conformation of the backbone in unfolded proteins. *Chem. Rev.* **2006**, *106*, 1877.

(18) Shortle, D.; Ackerman, M. Persistence of native-like topology in a denatured protein in 8 M urea. *Science* **2001**, *293*, 487.

(19) Mohana-Borges, R.; Goto, N. K.; Kroon, G. J.; Dyson, H.; Wright, P. E. Structural Characterization of Unfolded States of Apomyoglobin using Residual Dipolar Couplings. *J. Mol. Biol.* **2004**, *340*, 1131.

(20) Jha, A.; Colubri, A.; Freed, K.; Sosnick, T. Statistical coil model of the unfolded state: Resolving the reconciliation problem. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 13099.

(21) Jensen, M. R.; Markwick, P. R.; Meier, S.; Griesinger, C.; Zweckstetter, M.; Grzesiek, S.; Pau, B.; Blackledge, M. Quantitative Determination of the Conformational Properties of Partially Folded and Intrinsically Disordered Proteins Using NMR Dipolar Couplings. *Structure* **2009**, *17*, 1169.

(22) Fitzkee, N.; Rose, G. Reassessing random-coil statistics in unfolded proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 12497.

(23) Wang, Z.; Plaxco, K.; Makarov, D. Influence of local and residual structures on the scaling behavior and dimensions of unfolded proteins. *Biopolymers* **2007**, *86*, 321.

(24) Case, D. A.; Cheatham, T. E., III; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668.

(25) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A.; Berendsen, H. GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **2005**, *26*, 1701.

(26) Wang, J.; Cieplak, P.; Kollman, P. A. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.* **2000**, *21*, 1049.

(27) Tsai, J.; Taylor, R.; Chothia, C.; Gerstein, M. The packing density in proteins: standard radii and volumes. *J. Mol. Biol.* **1999**, *290*, 253.

(28) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435.

(29) Holm, L.; Sander, C. Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.* **1998**, *26*, 316.

(30) Holm, L.; Sander, C. Dictionary of recurrent domains in protein structures. *Proteins* **1998**, *33*, 88.

(31) Morris, A. L.; MacArthur, M. W.; Hutchinson, E. G.; Thornton, J. M. Stereochemical quality of protein structure coordinates. *Proteins* **1992**, *12*, 345.

(32) Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577.

(33) Swindells, M, B.; MacArthur, M, W.; Thornton, J, M. Intrinsic phi, psi propensities of amino acids, derived from the coil regions of known structures. *Nat. Struct. Biol.* **1995**, *2*, 596.

(34) Soda, K.; Miki, Y.; Nishizawa, T.; Seki, Y. New method for incorporating solvent influence into the evaluation of X-ray scattering intensity of Proteins in Solution. *Biophys. Chem.* **1997**, *65*, 45.

(35) Seki, Y.; Tomizawa, T.; Khechinashvili, N. N.; Soda, K. Contribution of solvent water to the solution X-ray scattering profile of proteins. *Biophys. Chem.* **2002**, *95*, 235.

(36) Richmond, T. J. Solvent accessible surface area and excluded volume in proteins. Analytical equations for overlapping spheres and implications for the hydrophobic effect. *J. Mol. Biol.* **1984**, *178*, 63.

(37) Goldenberg, D. Computational simulation of the statistical properties of unfolded proteins. *J. Mol. Biol.* **2003**, *326*, 1615.

(38) Gong, H.; Rose, G. D. Assessing the solvent-dependent surface area of unfolded proteins using an ensemble model. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 3321.

(39) Creamer, T.; Srinivasan, R.; Rose, G. Modeling unfolded states of peptides and proteins. *Biochemistry* **1995**, *34*, 16245.

(40) Creamer, T.; Srinivasan, R.; Rose, G. Modeling unfolded states of proteins and peptides II. Backbone solvent accessibility. *Biochemistry* **1997**, *36*, 2832.

(41) Flory, P. *Statistical Mechanics of Chain Moleclues*; Interscience: New York, 1969.

(42) Flory, P. *Principles of Polymer Chemistry*; Cornell Uni. Press: NewYork, 1953.

(43) Le Guillou, J. C.; Zinn-Justin, J. Critical Exponents for the n-Vector Model in Three Dimensions from Field Theory. *Phys. Rev. Lett.* **1977**, *39*, 95.

(44) Kamatari, Y. O.; Ohji, S.; Konno, T.; Seki, Y.; Soda, K.; Kataoka, M.; Akasaka, K. The compact and expanded denatured conformations of apomyoglobin in the methanol-water solvent. *Protein Sci.* **1999**, *8*, 873.

(45) Tirado-Rives, J.; Orozco, M.; Jorgensen, W. L. Molecular dynamics simulations of the unfolding of barnase in water and 8 M aqueous urea. *Biochemistry* **1997**, *36*, 7313.

(46) Bennion, B.; Daggett, V. Counteraction of urea-induced protein denaturation by trimethylamine N-oxide: A chemical chaperone at atomic resolution. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 6433.

(47) Ikeguchi, M.; Nakamura, S.; Shimizu, K. Molecular Dynamics Study on Hydrophobic Effects in Aqueous Urea Solutions. *J. Am. Chem. Soc.* **2001**, *123*, 677.