View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Kyoto University Research Information Repository

KURENAI

Kyoto University Research Information Repository

KYOTO UNIVERSITY

Title	Free-energy function for discriminating the native fold of a protein from misfolded decoys.
Author(s)	Yasuda, Satoshi; Yoshidome, Takashi; Harano, Yuichi; Roth, Roland; Oshima, Hiraku; Oda, Koji; Sugita, Yuji; Ikeguchi, Mitsunori; Kinoshita, Masahiro
Citation	Proteins : structure, function, and genetics (2011), 79(7): 2161-2171
Issue Date	2011-05-09
URL	http://hdl.handle.net/2433/197178
Right	This is the peer reviewed version of the following article: Yasuda, S., Yoshidome, T., Harano, Y., Roth, R., Oshima, H., Oda, K., Sugita, Y., Ikeguchi, M. and Kinoshita, M. (2011), Free-energy function for discriminating the native fold of a protein from misfolded decoys. Proteins, 79: 2161–2171, which has been published in final form at http://dx.doi.org/10.1002/prot.23036
Type	Journal Article
Textversion	author

Kyoto University

Free-Energy Function for Discriminating the Native Fold of a Protein from Misfolded Decoys

Satoshi Yasuda[1], Takashi Yoshidome[2], Yuichi Harano[3], Roland Roth[4], Hiraku Oshima[2], Koji Oda[5], Yuji Sugita[6], Mitsunori Ikeguchi[7], and Masahiro Kinoshita[2]*

[1]Graduate School of Energy Science, Kyoto University, Uji, Kyoto 611-0011, Japan

[2]Institute of Advanced Energy, Kyoto University, Uji, Kyoto 611-0011, Japan

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

[4]Institut für Theoretische Physik I, Staudtstrasse 7, 91058 Erlangen, Germany

[5]Taisho Pharmaceutical Co., Ltd., Yoshino-cho, Kita-ku, Saitama 331-9530, Japan

[6]RIKEN Advanced Science Institute, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

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

TITLE RUNNING HEAD: Free-energy Function for Proteins

*Correspondence author.
E-mail: kinoshit@iae.kyoto-u.ac.jp.
Address: Institute of Advanced Energy, Kyoto University, Uji, Kyoto 611-0011, Japan.

Key words: protein folding; structure prediction; decoy recognition; hydration entropy; dehydration; molecular liquid; integral equation theory; morphometric approach

Abstract

We investigate our free-energy function (FEF) for discriminating the native fold of a protein from misfolded decoys. It is a physics-based function using an all-atom model which comprises the hydration entropy (HE) and the total dehydration penalty (TDP). The HE is calculated using a hybrid of a statistical-mechanical theory applied to a molecular model for water and the morphometric approach. The energetic component is suitably taken into account in a simple manner as the TDP. On the basis of the results from a careful test of the FEF, which have newly been performed for 118 proteins in some representative decoy sets, we show that its performance is distinctly superior to that of any other function. By our FEF which precisely captures the features of the native structure, some important findings are made possible. For instance, our FEF varies largely from model to model for the candidate models obtained from nuclear magnetic resonance experiments. We can select the best model that is optimized in terms of the sum of the two components, HE and TDP. A decoy set is not suited to the test of a free-energy or potential function in cases where a protein isolated from a protein complex is considered and the structure in the complex is employed as the model NS of the isolated protein without any change or where portions of the terminus sides of a protein are removed and the percentage of the secondary structures lost due to the removal is significantly high.

Introduction

Predicting the native structure (NS) of a protein from its amino-acid sequence is one of the most challenging subjects in molecular biology, biophysics, and biochemistry.[1] As the first step toward the prediction, the development of a free-energy (or potential) function which takes the lowest value for the NS is highly desired. Up to now, there have been many attempts to develop such a function.[2-10] The function is usually tested as the so-called scoring function for discriminating the native fold from misfolded decoys. If its performance in the discrimination is sufficiently high, it is applied to the prediction of the NS for a practical purpose. The previously reported potential functions can be categorized into the following two types: knowledge-based[2-7] and physics-based[8-10] functions. The former functions are based on statistical analysis of known protein structures while the latter functions are developed on the basis of quantum chemistry and chemical physics.

We have recently developed a free-energy function (FEF) possessing the following features[11]: (1) The function comprises two components, the hydration entropy (HE) and the total dehydration penalty (TDP); (2) the HE, which is strongly dependent on details of the protein polyatomic structure, is calculated using a hybrid of the angle-dependent integral equation theory, a statistical-mechanical theory for molecular liquids,[12-16] and the morphometric approach[17,18]; (3) the roles of water as a molecular ensemble are fully taken into consideration, which is in marked contrast with the other physics-based functions[8-10] where water is regarded as a dielectric continuum and the hydrophobic effect is taken into account from the conventional viewpoint[19] through the solvation free energy evaluated using the solvent-accessible surface areas; and (4) the calculation of the function is accomplished quite rapidly (the computation time required per structure is ~0.1 sec on our workstation). The HE represents a water-entropy loss upon the protein insertion with a fixed structure. In our function, the HE arises primarily from the entropic excluded-volume effect[20,21]: Upon the insertion, the total volume available to the translational displacement of the coexisting water molecules decreases, leading to decreases in the number of accessible configurations of the water and in the water entropy. The TDP corresponds to “the protein intramolecular (Coulomb and Lennard-Jones terms) plus the hydration energy (not the hydration free energy)” of a given structure minus the same quantity of a fully extended structure.[11] The TDP is calculated using a simple method in which the physical essence is suitably incorporated. The meaning of the dehydration penalty is as follows. When a protein takes a more compact structure, “CO” and “HN”, for example, are buried after the break of hydrogen bonds with water molecules. There is no problem if intramolecular hydrogen bonding (CO···HN) is formed. However, the formation is not always attained, leading to the

dehydration penalty. With smaller HE and/or TDP, the structure is more stable in terms of the free energy.

In our earlier work,[11] the FEF was tested for the 4state_reduced,[22] fisa,[23] and fisa_casp3[23] decoy sets obtained from the database "Decoys 'R' Us''.[24] The total number of proteins considered was 15. The 100% success was achieved in the discrimination of the native fold by the function of Lu *et al.*[7], a knowledge-based function which has once been shown to be the most successful among the functions available in literature, as well as by ours. We notice that there are more difficult decoy sets where the former function fails for significantly many proteins. In the present work, we challenge the following decoy sets: Rosetta,[25] lattice_ssfit,[26,27] lmds,[28] and semfold[29] decoy sets (lattice_ssfit, lmds, and semfold decoy sets are taken from the database "Decoys 'R' Us''). Any of the decoy sets tested consists of two data sets where the native structures are obtained from X-ray crystallographic experiments and from nuclear magnetic resonance (NMR) experiments, respectively. For the Rosetta decoy set, we refer to the two data sets as Rosetta(X-ray) and Rosetta(NMR), respectively. We can test a variety of proteins (the total number of proteins considered reaches 118) with these decoy sets. In the Rosetta(X-ray), Rosetta(NMR), and lmds decoy sets, there are proteins for which the function of Lu *et al.* ends with failure: The success rates are, respectively, "37/41, 90%", "8/10, 80%", and "18/51, 35%". The success rate is quite low for Rosetta(NMR). As for our FEF, it is shown to be distinctly superior to the function of Lu *et al.* in terms of the performance. It is almost as successful as in our prior work. The NS is optimized in terms of the sum of the two components, HE and TDP. In the course of the test, a significant amount of new information is obtained as described in the next two paragraphs, thanks to our FEF that precisely captures the features of the NS.

We cannot exactly know the *true* NS of a protein. What we have is a model NS represented by structural data obtained via a specific route. In some of the decoy sets, however, the model NS employed is incomplete and not always close to the *true* NS for diverse reasons. In the lmds[28] decoy set, there is an example where a protein isolated from a protein complex is considered and the structure in the complex is employed as the model NS of the isolated protein without any change. The decoy structures are constructed for the isolated protein. The consequence is that there are a number of decoy structures whose FEF is lower than that of the model NS. However, we find that the structure of the protein isolated in aqueous solution has been determined in an experiment: It is considerably different from the model NS mentioned above. The model NS is then replaced by the experimentally determined structure in the isolated form with the result that the FEF becomes the lowest for the model NS. There are decoy sets in which portions in the terminus sides of a protein are removed. There is a strong trend that the performance of our FEF becomes higher as the percentage of the secondary

structures lost due to the removal decreases. This is because the resultant model NS becomes closer to the *true* NS. As explained so far, a model NS is often disqualified for representing the *true* NS.

When the NMR technique is employed in experimental determination of the NS, there are usually a lot of candidate models of the NS satisfying the experimental data. However, such NS models are not necessarily physically reasonable. This is because, in our view, a molecular model is not employed for water and the incorporation of the water-entropy effect is insufficient in the conventional procedures of determining the NS models. The problem is that it is not simple to select the best NS model. This is why the proteins whose native structures were determined through the NMR are often not considered in testing the potential or free-energy function. We find that our FEF as well as its two components, HE and TDP, varies largely from model to model. Some of the models are substantially different from the best model for which our FEF takes the lowest value. It is argued that when the best NS model is employed, the discrimination of the native fold becomes successful (i.e., its FEF becomes lower than that of any decoy structure). This result is crucially important for future works in the NMR-experimental research community.

Method

Hydration entropy (HE)

The hydration free energy (i.e., excess chemical potential), which is the most important thermodynamic quantity of hydration, consists of two components, hydration entropy (HE) and hydration energy. Unlike the two components, the hydration free energy is the same irrespective of the protein insertion condition: isobaric or isochoric.[30,31] We consider the isochoric condition that is much more convenient in a theoretical treatment. Since the HE is fairly insensitive to the protein-water interaction potential,[32,33] the protein can be modeled as a set of fused hard spheres. The hydration energy, which is influenced by the protein-water interaction potential, is treated in the TDP. Hereafter, the HE is denoted by S_{VH}.

We have developed a method which enables us to obtain S_{VH} with minor computational effort.[11] In this method, S_{VH} for a protein in a fixed structure is calculated using the angle-dependent integral equation theory[12-16] applied to a multipolar water model[12,13] (a hard sphere with diameter 0.28 nm in which a point dipole and a point quadrupole of tetrahedral symmetry are embedded) combined with the morphometric approach.[17,18] In the angle-dependent integral equation theory the effect of the molecular

polarizability is taken into account using the self-consistent mean field (SCMF) theory.[12,13] At the SCMF level the many-body induced interactions are reduced to pairwise additive potentials involving an effective dipole moment.

The idea of the morphometric approach is to predict the HE using a linear combination of only four geometrical measures for a protein with a prescribed structure: the excluded volume, the water-accessible surface area, and the integrated mean and Gaussian curvatures of the accessible surface, respectively. Though the excluded-volume term is the principal one, the other three terms also influence S_{VH}. The water-accessible surface is the surface that is accessible to the centers of water molecules. The excluded volume is the volume that is enclosed by the surface area. The four coefficients in the linear combination are determined in simple geometries. They are calculated from the values of S_{VH} for hard-sphere solutes with various diameters immersed in our model water. The angle-dependent integral equation theory is employed in the calculation for incorporating the orientational correlations. The *x*-*y*-*z* coordinates of the protein atoms, which characterize each structure at the atomic level, are used as part of the input data for calculating the four geometric measures.

The dielectric constant of bulk water calculated using the angle-dependent integral equation theory combined with the multipolar water model is ~83 that is in good agreement with the experimental value ~78.[15] As proved in our earlier work,[15] the angle-dependent integral equation theory gives a quantitatively accurate value of the hydration free energy of a nonpolar solute. It also gives a successful result in elucidating the hydrophilic hydration.[16] However, due to the mathematical complexity its extension to complex solute molecules like proteins is rather difficult. This problem is overcome by combing it with the morphometric approach as described above. The high reliability of the morphometric approach in calculating the HE has been demonstrated in our earlier publications. For example, the experimentally measured changes in thermodynamic quantities upon apoplastocyanin folding are quantitatively reproduced by a hybrid of the angle-dependent integral equation theory combined with the multipolar water model and the morphometric approach.[34] Moreover, great progresses have been made in elucidating the molecular mechanism of pressure,[35] cold,[36,37] and thermal[38] denaturations of proteins by the hybrid.

Total dehydration penalty (TDP)

A fully extended structure possesses the maximum number of hydrogen bonds with water molecules and no intramolecular hydrogen bonds. The protein intramolecular energy plus the hydration energy, when the fully extended structure is chosen as the standard one, corresponds to the TDP occurring upon the transition to a more compact

structure.[11] Let Λ denote the TDP. 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, leading to the dehydration penalty.

Our basic strategy for calculating the TDP is as follows.[11] 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 result obtained by a molecular dynamics simulation[39] performed for hydrogen-bond formation between two formamide molecules in a nonpolar liquid.

We examine all the donors and acceptors for backbone-backbone, backbone-side chain, and side chain-side chain intramolecular hydrogen bonds and calculate Λ. It is necessary to determine if each of the donors and acceptors is buried or not. The water-accessible surface area is calculated for each of them by means of Connolly's algorithm[40,41] (the TINKER program package[42] is used). If it is smaller than a threshold value A_0, the donor or acceptor is considered buried. A_0 is set at 0.001 Å^2. To determine if an intramolecular hydrogen bond is formed or not, we use the criteria proposed by McDonald and Thornton.[43]

Free-energy function (FEF)

Our FEF F is expressed by[11]

$$F=(\Lambda-TS_{VH})/(k_BT_0),\ T_0=298\ \text{K}. \quad (1)$$

S_{VH} is negative while Λ is positive, and they are strongly dependent on the protein structure. In the present study, T is set at T_0. In what follows, we investigate the properties of our FEF exhibited when it is applied to the discrimination of the native fold of a protein from misfolded decoys.

Decoy sets tested

We test the Rosetta,[25] lattice_ssfit,[26,27] lmds,[28] and semfold[29] decoy sets. Among them, there are decoy sets for which a protein taken from a protein complex is considered. The structure of the protein is assumed to remain unchanged even if it is

isolated in aqueous solution, and the structure is regarded as the model NS of the protein. The decoy structures are often constructed for a protein whose portions in the two terminus sides are removed (or a portion in a terminus side is removed). In such cases, the same removal is made for the NS as well for impartial comparison with the result that some of the secondary structures are lost. We are concerned with the percentage of the secondary structures thus lost, Ω. The DSSP program[44] is employed in the calculation of Ω.

Each of the decoy structures is slightly modified to eliminate the unrealistic overlaps as described in "Refinement of protein structures" of Appendix (the modification is made for some of the native structures as well). Calcium (Ca^{2+}) or zinc (Zn^{2+}) ion is entangled in the native structures of some proteins. There are proteins whose native structures are characterized by the heme binding (i.e., covalent heme linkages). On the other hand, neither the ion nor heme is included in the decoy structures. We cope with this problem as explained in "Treatment of ions and heme" of Appendix.

When the NMR technique is employed in experimental determination of the NS, there are usually a lot of candidate models of the NS. It has been found that our FEF varies largely from model to model. We select the model for which our FEF takes the lowest value. More details are described in the subsection, "Selection of the best NMR model of the native structure (NS)".

Results

Discrimination of the native structure (NS) from decoys

A measure of the performance of a free-energy or potential function is the *Z*-score defined by

$$Z=(F_{\mathrm{Native}}-\langle F\rangle)/F_{\sigma}, \qquad (2)$$

where $\langle F\rangle$ is the function averaged over all decoy structures of a protein in a decoy set and F_{σ} the standard deviation of F for the decoy structures. The performance is higher if the *Z*-score takes a larger, negative value (i.e., if the *Z*-score is negative and its absolute value is larger). The performances of our FEF and the potential functions previously proposed by two research groups are compared in Table 1, in terms of the number of successful proteins and the average *Z*-score for each decoy set. The 4state_reduced,[22] fisa,[23] and fisa_casp3[23] decoy sets tested in our earlier work[11] are also included in the table.

We remark that the function of Lu *et al.*,[7] a knowledge-based function, has once been shown to give the best result among the functions available in literature. As observed in the table, our FEF is successful in discriminating the native fold from misfolded decoys with 100% accuracy except in the Rosetta(NMR)[25] and lmds[28] decoy sets (the nonsuccess in these decoy sets can be justified as discussed in the subsections, "In cases where a protein taken from a protein complex is considered" and "Selection of the best NMR model of the native structure (NS)", respectively). The predominance of our FEF over the function of Lu *et al.* is appreciable in Table 1, especially for the Rosetta(NMR) decoy set. Our FEF (success rate: "46/51, 90%") is far superior to the latter (success rate: "18/51, 35%"). The semfold[29] decoy set is characterized by the largest average number of decoy structures, 12,900. For this decoy set, the success rate of the function of Miyazawa and Jernigan is "4/6, 67%" while ours is always successful. In Figure 1, as a representative case, the plot of $F-F_{\mathrm{Native}}$ (the subscript "Native" denotes the value for the NS) against the root mean square displacement (RMSD) for C_α atoms from the NS is shown for the protein 1khm (this is the PDB code) in the semfold[29] decoy set. For this protein, the number of decoys reaches 21,080. Despite the large number of decoys, our FEF is capable of discriminating the NS from the decoys. In the subsection, "In cases where a protein taken from a protein complex is considered", we explain the meaning of the value put within parentheses in "9/10 (10/10)" or "–6.29 (–6.79)" for the lmds decoy set (see Table 1).

Characteristics of the native structure (NS)

We decompose $F-F_{\mathrm{Native}}$ into the two components, X and Y, defined by[11]

$$X=\Lambda/(k_{\mathrm{B}}T_0)-\{\Lambda/(k_{\mathrm{B}}T_0)\}_{\mathrm{Native}}, \tag{3}$$

and

$$Y=-TS_{\mathrm{VH}}/(k_{\mathrm{B}}T_0)-\{-TS_{\mathrm{VH}}/(k_{\mathrm{B}}T_0)\}_{\mathrm{Native}},\ T=T_0. \tag{4}$$

The plot of Y against X is shown in Figure 2 for the protein 1khm in the semfold[29] decoy set. There are significantly many structures with $X<0$ or $Y<0$. However, there are no structures causing $X+Y=F-F_{\mathrm{Native}}<0$. The NS is optimized in terms of the sum of the two components, HE and TDP. This result is consistent with the finding in our prior work.[11]

Discussion

In cases where a protein taken from a protein complex is considered

In the lmds[28] decoy set, there is an example where one of the two proteins (i.e., chain C of PDB structure with code 1fc2) forming a protein complex is taken as illustrated in Figure 3(a). The structure of the protein is assumed to remain unchanged even if it is isolated in aqueous solution, and the structure is regarded as the model NS of the protein. The decoy structures are constructed not for the complex but for the isolated protein. As shown in Figure 3(b), it is predicted that there are a number of decoy structures whose FEF is lower than that of the model NS and the *Z*-score is 0.76. This prediction is identified as a failure and reflected in "9/10" and "−6.29" for the lmds decoy set in Table 1. However, the structure of the protein (chain C of PDB structure with code 1fc2) isolated in aqueous solution has been determined by the NMR (its PDB code is 1bdc): It is considerably different from the model NS as illustrated in Figure 4(a) (it has one more α-helix). The 1bdc structure must be defined as the NS of the protein. With this definition, our FEF takes the lowest value for the NS as shown in Figure 4(b) and the *Z*-score is −4.15. This alteration is identified as a success and reflected in "(10/10)" and "(−6.79)" for the lmds decoy set in Table 1. These results indicate that the features of the *true* NS (e.g., it is optimized in terms of the sum of the two important factors, HE and TDP) are precisely captured by our FEF.

Correlation between percentage of secondary structures lost in the model native structure (NS) and *Z*-score

For most of the decoy structures in the Rosetta[25] decoy set, portions in the two terminus sides of a protein are removed (or a portion in a terminus side is removed). The same removal is made for the NS as well with the result that some of the secondary structures (i.e., important constituents of the NS) are lost. The problem is that as the percentage of the secondary structures thus lost (this percentage is denoted by Ω) increases, the resultant model NS becomes less similar to the *true* NS. The *Z*-score is plotted against Ω in Figure 5. There is an apparent correlation between the two quantities for both Rosetta(X-ray) and Rosetta(NMR). The lower-limit value of the *Z*-score decreases as Ω becomes smaller. Namely, as Ω decreases, the model NS becomes closer to the *true* NS, and the performance of our FEF becomes higher. This result gives another evidence that the features of the *true* NS are precisely captured by our FEF.

Selection of the best NMR model of the native structure (NS)

When the NMR technique is employed in experimental determination of the NS, there are usually a lot of candidate models of the NS satisfying the experimental data. We find that the values of $-S_{VH}/k_B$, $\Lambda/(k_BT_0)$, and FEF change largely from model to model. As an example, we consider the protein 1khm in the semfold[29] decoy set. There are a total of twenty models, Models 1 through 20, of the NS and Models 11 and 20 give the highest and lowest values of our FEF, respectively. $-S_{VH}/k_B$ and $\Lambda/(k_BT_0)$ in Model 11 are larger than those in Model 20 by ~77 and ~70, respectively. The rank of the model native structures and the FEF-value relative to that for Model 20 are collected in Table 2. With Models 3 and 20, our FEF is lower for the model NS than for any of the decoy structures. Our FEF is capable of selecting the best model which captures the features of the *true* NS the most precisely. (We attribute this capability of our FEF to the thorough incorporation of the water-entropy effect using a molecular model for water.) In other words, our FEF can be applied to the refinement of low-resolution protein structure models, which have been derived from the NMR, to atomic-level accuracy.[45]

For five proteins in Rosetta(NMR),[25] our FEF is not successful in discriminating the native fold from misfolded decoys. However, this nonsuccess can be justified as follows. For two of the unsuccessful proteins, the structures stabilized under acidic conditions (pH=3.5 and 4.5) are regarded as the native structures. They should be significantly different from the *true* native structures stabilized under physiological conditions. For our FEF to become applicable to the structures stabilized under acidic conditions, the evaluation method for the TDP is to be modified. This is because significantly many of the side chains are positively charged and the TDP effect is larger than that evaluated in our FEF. For the other three unsuccessful proteins, portions of the terminus sides are removed and the percentages of the secondary structures thus lost are 25%, 35%, and 100% which are quite high (see Fig. 5): The ranks of the native structures are 4, 11, and 21 among 998, 997, and 1000 structures, respectively.

We find that some of the model native structures determined by the NMR undergo unreasonably large TDP. In one of the model native structures of the protein whose PDB code is 1btb, for example, the number of hydrogen bonds is fewer than that in most of the decoy structures. As shown in Figure 6, the model NS is inferior to most of the decoy structures in terms of the TDP: It is not suitable as a model of the *true* NS optimized in terms of the sum of the HE and the TDP (i.e., for which the TDP as well as the HE should be sufficiently small). The model NS is to be refined so that more complete intramolecular hydrogen bonds can be formed, and our FEF should be applicable to such refinement.

Correlation between number of residues and *Z*-score

Here, we examine the characteristics of decoy structures constructed for the test of a free-energy or potential function. The quality of a decoy set can be assessed by analyzing the dependence of the *Z*-score on the number of residues N_r. For the Rosetta[25] decoy set, the *Z*-score is plotted against N_r in Figure 7. There is an apparent trend that the *Z*-score becomes better and the performance of the FEF becomes higher as N_r increases with the correlation coefficient R=–0.68. On the other hand, we find that there is no appreciable correlation between N_r and any of the three quantities, F_{Native}/N_r, $\langle F\rangle/N_r$, and $(F_{Native}-\langle F\rangle)/N_r$: F_{Native}, $\langle F\rangle$, and $F_{Native}-\langle F\rangle$ are almost proportional to N_r. By contrast, F_σ/N_r decreases with increasing N_r, i.e., F_σ increases only less than in proportion to N_r. This implies that the variation of the FEF for the decoys becomes unreasonably smaller for a larger protein.

In summary, for larger proteins the structural space of decoys is not widely explored and the distribution of the FEF for the decoys becomes unreasonably narrow. There are fewer decoy structures whose FEF is close to the FEF of the NS, and the *Z*-score becomes better. Thus, for large proteins the *artificial* construction of good competitors of the NS is considerably difficult. We note that this finding can be made possible only by a FEF which precisely captures the features of the NS and our FEF is this type of function. (A function with low performance is often unsuccessful even when the decoy structures are not good competitors of the NS.)

Decoy set which is not suited to test of free-energy or potential function

The results described so far suggest that the following decoy sets are not suited to the test of a free-energy or potential function: (i) those where a protein isolated from a protein complex is considered and the structure in the complex is employed as the model NS of the isolated protein without any change; (ii) those where a fragment taken from a protein is considered and its structure is assumed to remain unchanged even when it is isolated in aqueous solution; (iii) those where portions in the two terminus sides are removed (or a portion in a terminus side is removed) and the percentage of the secondary structures thus lost is significantly high; and (iv) those where a monomer taken from a homo-oligomer is considered and the structure in the homo-oligomer is regarded as the model NS of the isolated monomer. In (ii), the structure regarded as the NS is often significantly different from the *true* NS of the fragmental protein. In (iii), the structure used as the NS is no more close to the *true* NS of the modified protein. Similar statements can be given for (i) and (iv).

Conclusion

We have investigated the properties of our free-energy function[11] (FEF) exhibited when it is applied to the discrimination of the native fold of a protein from misfolded decoys. It is based on an all-atom model and comprises two components, the hydration entropy (HE) and the total dehydration penalty (TDP). Upon protein insertion, the total volume available to the translational displacement of the coexisting water molecules decreases, leading to a decrease in the number of accessible configurations of the water.[20,21] Primarily from this effect, a water-entropy loss occurs. In order to fully account for the water-entropy loss, the HE is calculated using a statistical-mechanical theory applied to a molecular model for water[12-16] combined with the morphometric approach.[17,18] The TDP corresponds to the sum of the hydration energy and the protein intramolecular energy when a fully extended structure, which possesses the maximum number of hydrogen bonds with water molecules and no intramolecular hydrogen bonds, is chosen as the standard one. When a donor and an acceptor (e.g., N and O, respectively) are buried in the interior after the break of hydrogen bonds with water molecules, if they form an intramolecular hydrogen bond, no penalty is imposed. When a donor or an acceptor is buried with no intramolecular hydrogen bond formed, an energetic penalty is imposed. We examine all the donors and acceptors for backbone-backbone, backbone-side chain, and side chain-side chain intramolecular hydrogen bonds and calculate the TDP.

The new, original aspects of the present study which were not found in our prior work[11] are as follows:

(1) In the 4state_reduced,[22] fisa,[23] and fisa_casp3[23] decoy sets tested in our prior paper, as observed from Table 1, the 100% success was achieved by the function of Lu *et al.*[7] as well as by ours. In the present study, by contrast, the function of Lu *et al.* fails for some proteins in the Rosetta(X-ray)[25] (success rate: 37/41, 90%) and lmds[28] (success rate: 8/10, 80%) decoy sets while ours always gives success. For the Rosetta(NMR)[25] decoy set, the performance of our function (success rate: 46/51, 90%) is much higher than that of the function of Lu *et al.* (success rate: 18/51, 35%). The semfold[29] decoy set is characterized by the largest average number of decoy structures, 12,900. In this decoy set, the function of Miyazawa and Jernigan[6] fails for some proteins (success rate: 4/6, 67%) while ours is always successful. Thus, the decoy sets tested in the present study are far more difficult than those tested in our prior work. Nevertheless, our FEF provides almost the same success.

(2) When the NMR technique is employed in experimental determination of the native structure (NS), there are usually a lot of candidate models of the NS satisfying the

experimental data. However, these models are not always physically reasonable in terms of the two components, HE and TDP. In fact, our FEF as well as the two components varies largely from model to model. However, we certainly find models (or a model) for which our FEF becomes lower than any decoy structure (except when the NS was measured under acidic conditions which are significantly different from physiological conditions or when an unreasonably large percentage of secondary structures is lost due to the removal of portions of the terminus sides: for the five proteins in Rosetta(NMR)). It follows that our FEF is capable of selecting the best model among the candidate models. This capability is attributable to the thorough incorporation of the water-entropy effect in our FEF using a molecular model for water. The result mentioned above is crucially important for future works in the NMR-experimental research community.

(3) The number of proteins in our prior paper was only 15 while that in the present study is 118. In particular, there are 92 proteins in the Rosetta(X-ray) and Rosetta(NMR) decoy sets. Thanks to this large number, the plots in Figures 5 and 7 become significant, and the following significant conclusions have been drawn: For a protein whose portions in the terminus sides are removed, as the percentage of the secondary structures lost due to the removal decreases, the resultant model NS becomes closer to the *true* NS (see the next paragraph) and the performance of our FEF becomes higher (from Figure 5); and for large proteins the *artificial* construction of good competitors of the NS is considerably difficult (from Figure 7).

(4) We have found that some of the decoy sets are unsuitable to the test of the potential or free-energy function and how such unsuitable decoy sets are characterized, which was never argued in previously reported works.

We emphasize that these findings can be made possible only by a FEF which precisely captures the features of the NS and our FEF is this type of function.

We cannot exactly know the *true* NS of a protein. What we have is a model NS represented by structural data obtained via a specific route. However, such a model NS is not sufficiently close to the true NS in the following examples:

(A) When a protein isolated from a protein complex is considered and the structure in the complex is employed as the model NS of the isolated protein without any change, the model NS is often quite different from the true NS.

(B) When a fragment taken from a protein is considered and its structure is assumed to remain unchanged even if it is separately immersed in aqueous solution, the structure is often significantly different from a well-qualified model NS of the fragmental protein.

(C) When portions of the terminus sides of a protein are removed, the resultant NS can be worse than some of the decoy structures, because it is no more close to the structure of the modified protein which would actually be formed.
(D) When a monomer taken from a homo-oligomer is considered and the structure in the homo-oligomer is regarded as the model NS of the isolated monomer, it is disqualified as a physically good model of the NS.
(E) Some of the candidate models based on NMR experiments are disqualified as physically good models of the NS.

In the above examples, the test of the free-energy or potential function cannot correctly be performed. (Since the decoy structures of large proteins are not good competitors of the NS, a decoy set of very large proteins may not be quite suited to the test of a free-energy or potential function.)

Our FEF and its calculation method are best suited to selecting the most stable structure from among the candidate structures. The number of the candidate structures is allowed to be huge, because the function is calculated with minor computational effort. Our FEF should be applicable to the refinement of protein structure models obtained from the NMR, comparative modeling, and *de novo* modeling approaches.[45] Further, it may be possible to develop a practical tool for predicting the NS of a protein from its amino-acid sequence, by combining our free-energy function with the techniques which can generate a variety of candidate structures. The function and its calculation method are capable of handling much larger proteins than those considered in this article and can also be extended to analyses of protein-protein interaction and protein aggregation. Works in these directions are in progress.

Appendix

Treatment of ions and heme

Calcium (Ca^{2+}) or zinc (Zn^{2+}) ion is entangled in the native structures of some proteins. The cation, which is bound to negatively charged atoms, should be included in the model NS. The HE of the cation is added to that of a decoy in which the cation is not included, and the sum is regarded as the HE of the decoy. As for the TDP for the model NS, considering the cation as a donor heavy atom and the coordinating atoms as acceptors, we calculate the TDP with the hydrogen-bond criteria for the usual hydrogen bonds.

There are proteins whose native structures are characterized by the heme binding

(i.e., covalent heme linkages). Heme should be included in the model NS because it plays important roles in the stabilization of the NS.[46] The HE of heme is added to that of a decoy in which heme is not included, and the sum is regarded as the HE of the decoy. In the calculation of the TDP for the model NS, the intramolecular hydrogen bonds between heme and the apo structure are not considered: The dehydration penalty is always imposed when a donor or an acceptor is buried, and the model NS undergoes unreasonably large TDP. Nevertheless, our FEF gives the lowest value to the model NS.

Refinement of protein structures

The Lennard-Jones (LJ) potential energy for many of the decoy structures and some of the native structures is positive and quite large due to the unrealistic overlaps of protein atoms. Such overlaps are removed by the minimization of the energy function using the CHARMM biomolecular simulation program[47] through the Multi-scale Modeling Tools in Structural Biology (MMTSB) program.[48] The minimization is performed so that the original structures can be retained as much as possible. We employ the CHARMM22[49] with the CMAP correction[50] as the force-field parameters. Electrostatic and non-bonded interactions are all evaluated without any cut-off. The Generalized-Born (GBMV/SA) approximation[51-53] is employed for the electrostatic part of the hydration energy. After the minimization, there are no unrealistic overlaps of protein atoms. Moreover, it is verified that the RMSD for C_α atoms before and after the minimization is quite small. Each structure is then switched to a set of fused hard spheres in calculating the HE as described in the subsection, “Hydration entropy (HE)’’.

Acknowledgments

This work was supported by Grants-in-Aid for Scientific Research on Innovative Areas (Nos. 20118004 and 21118519), that on Priority Areas (No. 18074004), and that on (B) (Nos. 22300100 and 22300102) from the Ministry of Education, Culture, Sports, Science and Technology of Japan, by the Grand Challenges in Next-Generation Integrated Simulation of Nanoscience and Living Matter, a part of the Development and Use of the Next-Generation Supercomputer Project of MEXT, by Kyoto University Global Center of Excellence (GCOE) of Energy Science, and by Kyoto University Pioneering Research Unit for Next Generation.

References

1. Dill KA, Ozkan SB, Shell MS, Weikl TR. The protein folding problem. Annu Rev Biophys 2008;37:289-316.
2. Toby D, Elber R. Distance-dependent, pair potential for protein folding: Results from linear optimization. Proteins 2000;41:40-46.
3. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11:2714-2726.
4. Onizuka K, Noguchi T, Akiyama Y, Matsuda H. Using data compression for multidimensional distribution analysis. Control Intell Syst 2002;17:48-54.
5. Zhang C, Liu S, Zhou H, Zhou Y. An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. Protein Sci 2004;13:400-411.
6. Miyazawa S, Jernigan RL. How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins? J Chem Phys 2005;122:024901(1-18).
7. Lu M, Dousis AD, Ma J. OPUS-PSP: An orientation-dependent statistical all-atom potential derived from side-chain packing. J Mol Biol 2008;376:288-301.
8. Dominy BN, Brooks CLIII. Identifying native-like protein structures using physics-based potentials. J Comput Chem 2002;23:147-160.
9. Lee MC, Duan Y. Distinguish protein decoys by using a scoring function based on a new AMBER force field, short molecular dynamics simulations, and the generalized born solvent model. Proteins 2004;55:620-634.
10. Feig M, Brooks CLIII. Evaluating CASP4 predictions with physical energy functions. Proteins 2002;49:232-245.
11. Yoshidome T, Oda K, Harano Y, Roth R, Sugita Y, Ikeguchi M, Kinoshita M. Free-energy function based on an all-atom model for proteins. Proteins 2009;77:950-961.
12. Kusalik PG, Patey GN. On the molecular theory of aqueous electrolyte solutions. I. The solution of the RHNC approximation for models at finite concentration. J Chem Phys 1988;88:7715-7738.
13. Kusalik PG, Patey GN. The solution of the reference hypernettedchain approximation for water-like models. Mol Phys 1988;65:1105-1119.
14. Kinoshita M, Bérard DR. Analysis of the bulk and surface-induced structure of electrolyte solutions using integral equation theories. J Comput Phys 1996;124:230-241.
15. Kinoshita M. Molecular origin of the hydrophobic effect: Analysis using the

angle-dependent integral equation theory. J Chem Phys 2008;128:024507(1-14).
16. Kinoshita M, Yoshidome T. Molecular origin of the negative heat capacity of hydrophilic hydration. J Chem Phys 2009;130:144705(1-11).
17. König PM, Roth R, Mecke KR. Morphological thermodynamics of fluids: Shape dependence of free energies. Phys Rev Lett 2004;93:160601(1-4).
18. Roth R, Harano Y, Kinoshita M. Morphometric approach to the solvation free energy of complex molecules. Phys Rev Lett 2006;97:078101(1-4).
19. In the conventional concept, only the water in the close vicinity of a protein surface is considered when the effects of the water entropy are discussed: The effects are argued primarily in terms of the surface-water orientational correlations, enhanced hydrogen-bonding network of water, and restriction of the rotational freedom of water molecules. We remark that the entropic effect emphasized in the present study (i.e., the excluded-volume effect), which reaches a far larger length scale, is substantially different from the conventionally argued one.
20. Kinoshita M. Roles of translational motion of water molecules in sustaining life. Front in Biosci 2009;14:3419-3454.
21. Kinoshita M. Importance of translational entropy of water in biological self-assembly processes like protein folding. Int J Mol Sci 2009;10:1064-1080.
22. Park B, Levitt M. Energy functions that discriminates X-ray and near native folds from well-constructed decoys. J Mol Biol 1996;258:367-392.
23. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. J Mol Biol 1997;268:209-225.
24. Samudrala R, Levitt M. Decoys ‘R’ Us: A database of incorrect protein conformations to improve protein structure prediction. Protein Sci 2000;9:1399-1401.
25. Simons, KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins 1999;37:171-176.
26. Samudrala R, Xia Y, Levitt M, Huang ES. A combined approach for ab initio construction of low resolution protein tertiary structures from sequence. Proceedings of the Pacific Symposium on Biocomputing, 1999;4:505-516.
27. Xia Y, Huang ES, Levitt M, Samudrala R. *Ab initio* construction of protein tertiary structures using a hierarchical approach. J Mol Biol, 2000;300:171-185.
28. Keasar C, Levitt M. A Novel approach to decoy set generation: Designing a Physical Energy Function Having Local Minima with Native Structure Characteristics. J Mol Biol 2003;329:159-174.
29. Samudrala R, Levitt M. A comprehensive analysis of 40 blind protein structure predictions. BMC Structural Biology 2002;2:3-18.

30. Cann NM, Patey GN. An investigation of the influence of solute size and insertion conditions on solvation thermodynamics. J Chem Phys 1997;106: 8165-8195.
31. Kinoshita M, Harano Y, Akiyama R. Changes in thermodynamic quantities upon contact of two solutes in solvent under isochoric and isobaric conditions. J Chem Phys 2006;125:244504(1-7).
32. Imai T, Harano Y, Kinoshita M, Kovalenko A, Hirata F. A theoretical analysis on hydration thermodynamics of proteins. J Chem Phys 2006;125:024911(1-7).
33. Yasuda S, Yoshidome T, Oshima H, Kodama R, Harano Y, Kinoshita M. Effects of side-chain packing on the formation of secondary structures in protein folding. J Chem Phys 2010;132:065105(1-10).
34. Yoshidome T, Kinoshita M, Hirota S, Baden N, Terazima M. Thermodynamics of apoplastocyanin folding: Comparison between experimental and theoretical results. J Chem Phys 2008;128:225104(1-9).
35. Harano Y, Yoshidome T, Kinoshita, M. Molecular mechanism of pressure denaturation of proteins. J. Chem. Phys. 2008;129:145103(1-9).
36. Yoshidome T, Kinoshita M. Hydrophobicity at low temperatures and cold denaturation of a protein. Phys Rev E 2009;79:030509(1-4).
37. Oshima H, Yoshidome T, Amano K, Kinoshita M. A theoretical analysis on characteristics of protein structures induced by cold denaturation. J Chem Phys 2009;131:205102(1-11).
38. Amano K, Yoshidome T, Harano Y, Oda K, Kinoshita M. Theoretical analysis on thermal stability of a protein focused on the water entropy. Chem Phys Lett 2009;474:190-194.
39. Sneddon SF, Tobias DJ, Brooks CLIII. Thermodynamics of amide hydrogen bond formation in polar and apolar solvents. J Mol Biol 1989;209:817-820.
40. Connolly ML. Analytical molecular surface calculation. J Appl Crystallogr 1983;16:548-558.
41. Connolly ML. Computation of molecular volume. J Am Chem Soc 1985;107: 1118-1124.
42. Ponder JW, Richards FM. An efficient Newton-like method for molecular mechanics energy minimization of large molecules. J Comput Chem 1987;8:1016-1024.
43. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. J Mol Biol 1994;238:777-793.
44. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577-2637.
45. Qian B, Raman S, Das R, Bradley P, McCoy AJ, Read RJ, Baker D. High-resolution structure prediction and the crystallographic phase problem. Nature 2007;450:259-264.

46. Oda K, Kodama R, Yoshidome T, Yamanaka M, Sambongi Y, Kinoshita M. Effects of heme on the thermal stability of mesophilic and thermophilic cytochromes *c*: Comparison between experimental and theoretical results. J Chem Phys 2011;134:025101(1-9).
47. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. A program for macromolecular energy, minimization, and dynamics calculations. J Comput Chem 1983;4:187-217.
48. Feig M, Karanicolas J, Brooks CLIII. MMTSB tool set: enhanced sampling and multiscale modeling methods for applications in structural biology. J Mol Graphics 2004;22:377-395.
49. MacKerell ADJr, Bashford D, Bellott M, Dunbrack RLJr, Evanseck JD, Field MJ, *et al*. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 1998;102:3586-3616.
50. MacKerell ADJr, Feig M, Brooks CLIII. Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J Comput Chem 2004;25:1400-1415.
51. Lee MS, Salsbury FRJr, Brooks CLIII. Novel generalized Born methods. J Chem Phys 2002;116:10606-10614.
52. Lee MS, Feig M, Salsbury FRJr, Brooks, CLIII. New analytic approximation to the standard molecular volume definition and its application to generalized Born calculations. J Comput Chem 2003;24:1348-1356.
53. Chocholoussšová J, Feig M. Balancing an accurate representation of the molecular surface in generalized born formalisms with integrator stability in molecular dynamics simulations. J Comput Chem 2006;27:719-729.

Figure captions

Figure 1. $F-F_{\mathrm{Native}}$ plotted against the root mean square displacement (RMSD) for C_α atoms from the native structure, for the protein 1khm (this is the PDB code) in the semfold[29] decoy set.

Figure 2. Y plotted against X for the protein 1khm in the semfold[29] decoy set. The three straight lines drawn represent X=0, Y=0, and X+Y=0, respectively.

Figure 3. (a) One of the two proteins (chain C of PDB structure with code 1fc2) forming the protein complex is taken in the lmds[28] decoy set. It has two α-helices. They were drawn by the cartoon representation using the PDBjViewer. (b) $F-F_{\mathrm{Native}}$ plotted against the root mean square displacement (RMSD) for C_α atoms from the native structure. Chain C of PDB structure with code 1fc2 is employed.

Figure 4. (a) Structure of the protein taken in Figure 3(a) when it is isolated in aqueous solution: Its PDB code is 1bdc. It has three α-helices. It was drawn by the cartoon representation using the PDBjViewer. (b) $F-F_{\mathrm{Native}}$ plotted against the root mean square displacement (RMSD) for C_α atoms from the native structure. The structure of 1bdc is employed.

Figure 5. Z-score plotted against the percentage of the secondary structures lost in the model native structure, Ω. The plot is made for the proteins in the Rosetta[25] decoy set in which the model native structures are determined by either X-ray crystallographic experiments or nuclear magnetic resonance (NMR) experiments.

Figure 6. Y plotted against X for 1btb in the Rosetta[25] decoy set. The three straight lines drawn represent X=0, Y=0, and X+Y=0, respectively. The model native structure is determined by the nuclear magnetic resonance (NMR).

Figure 7. Z-score plotted against the number of residues N_{r} for the Rosetta[25] decoy set.

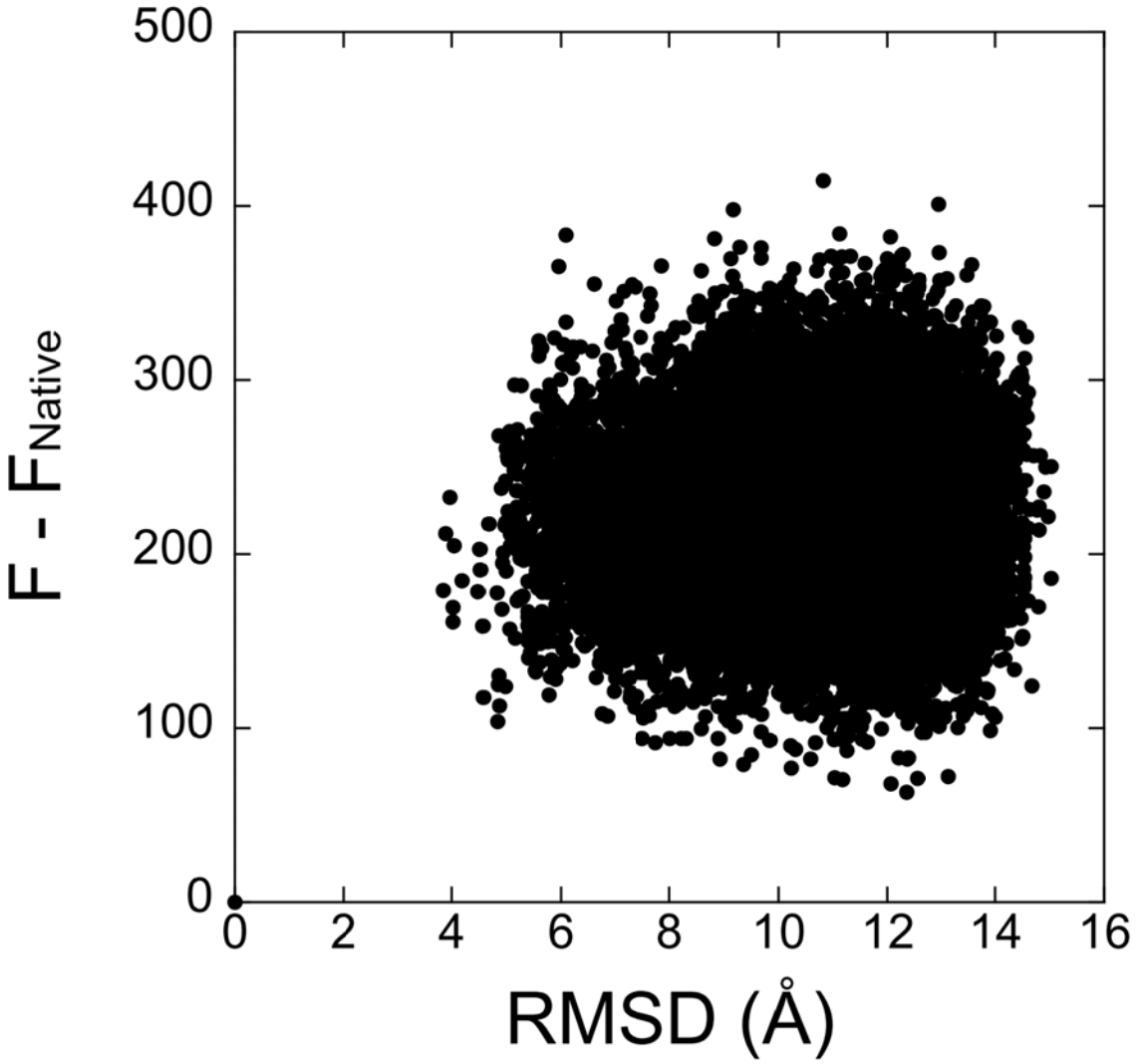


Fig. 1

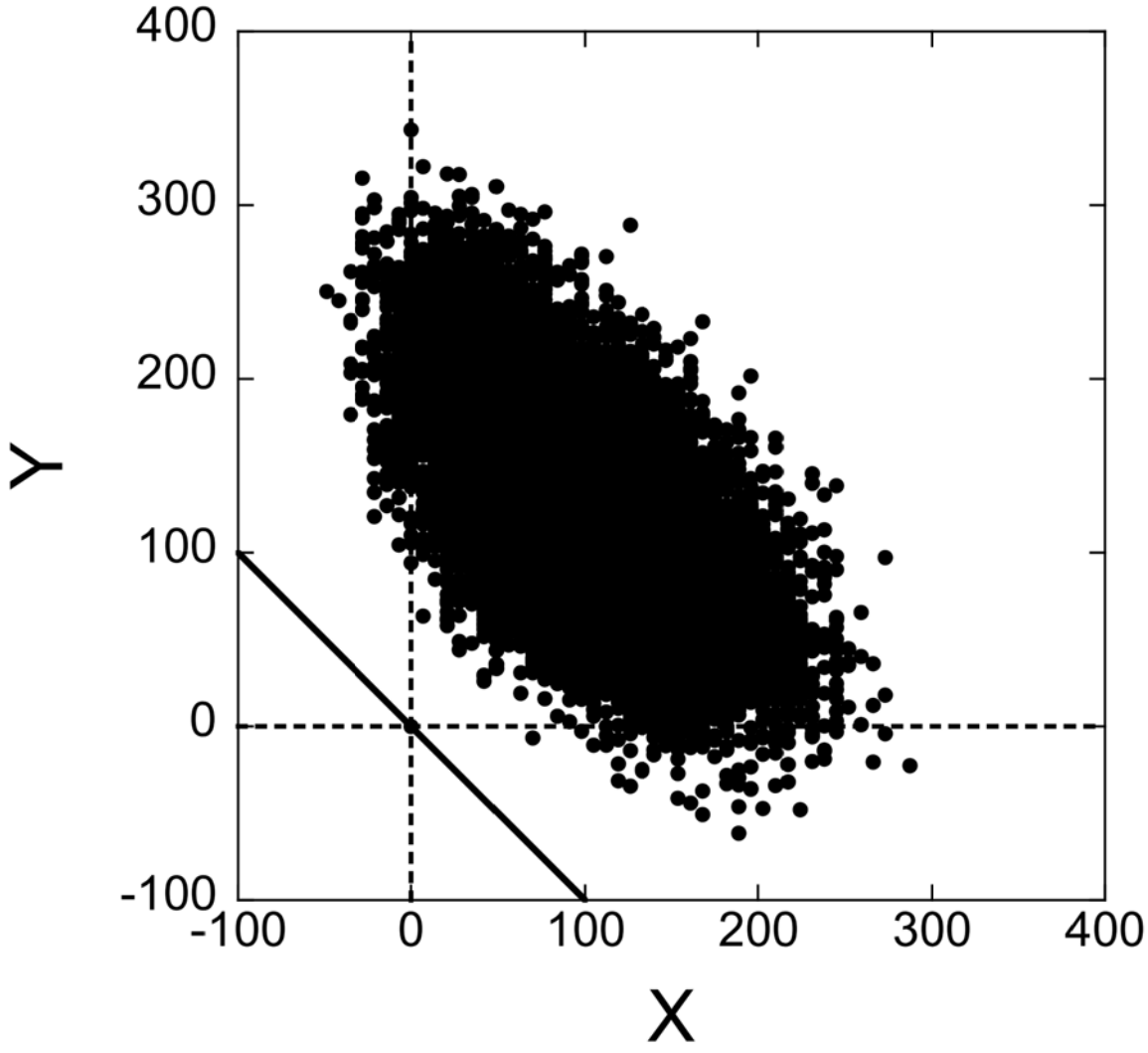
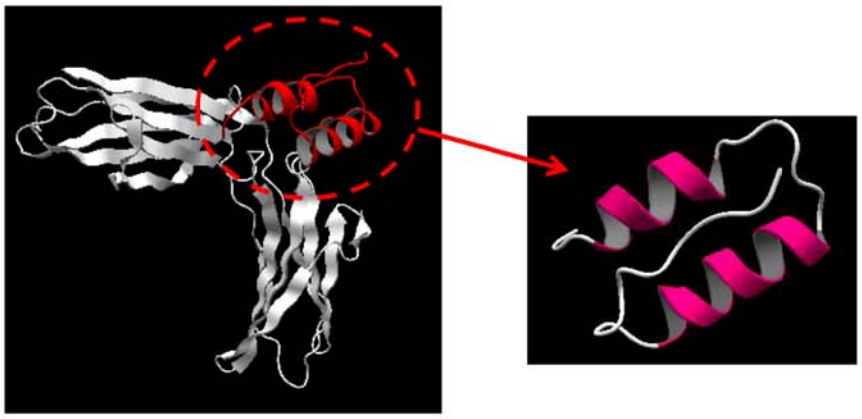


Fig. 2

(a)



(b)

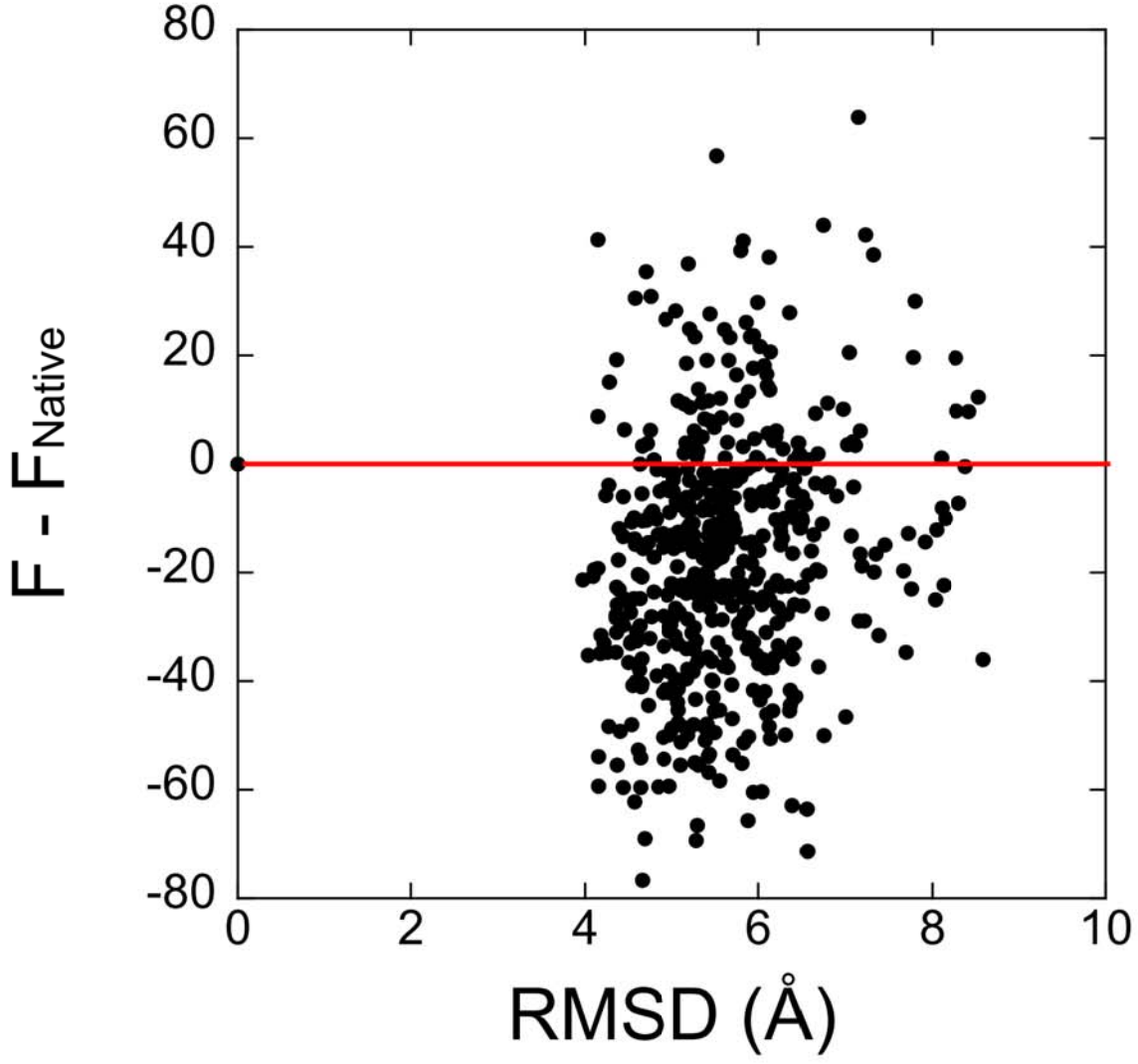
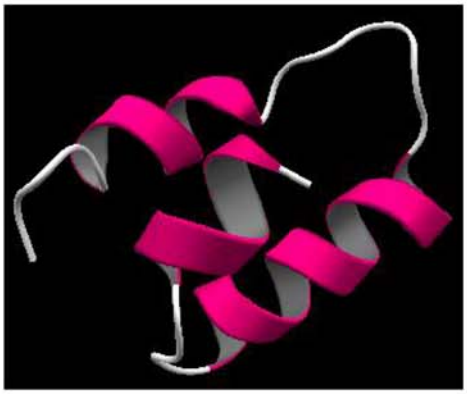


Fig. 3

(a)



(b)

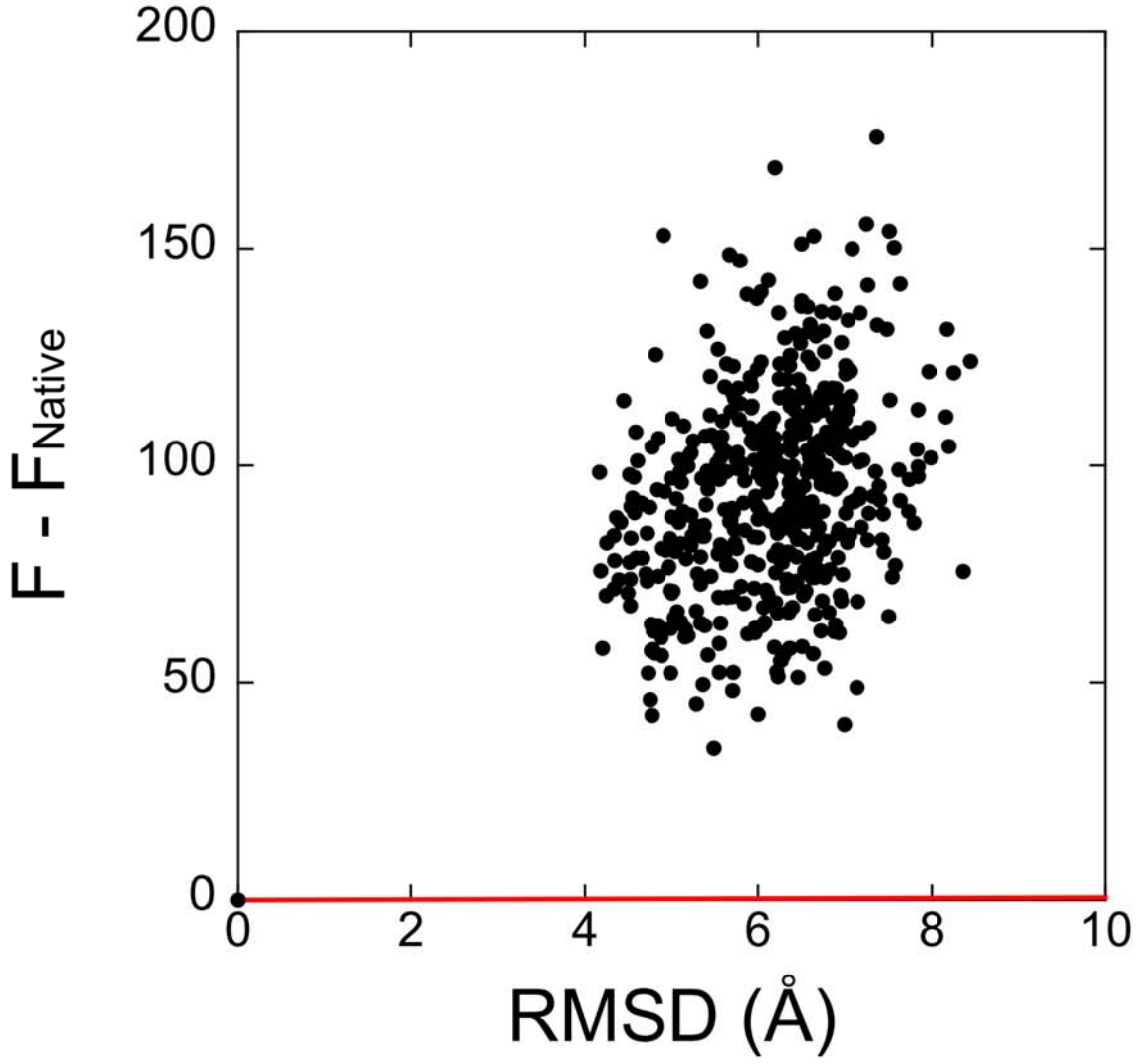


Fig. 4

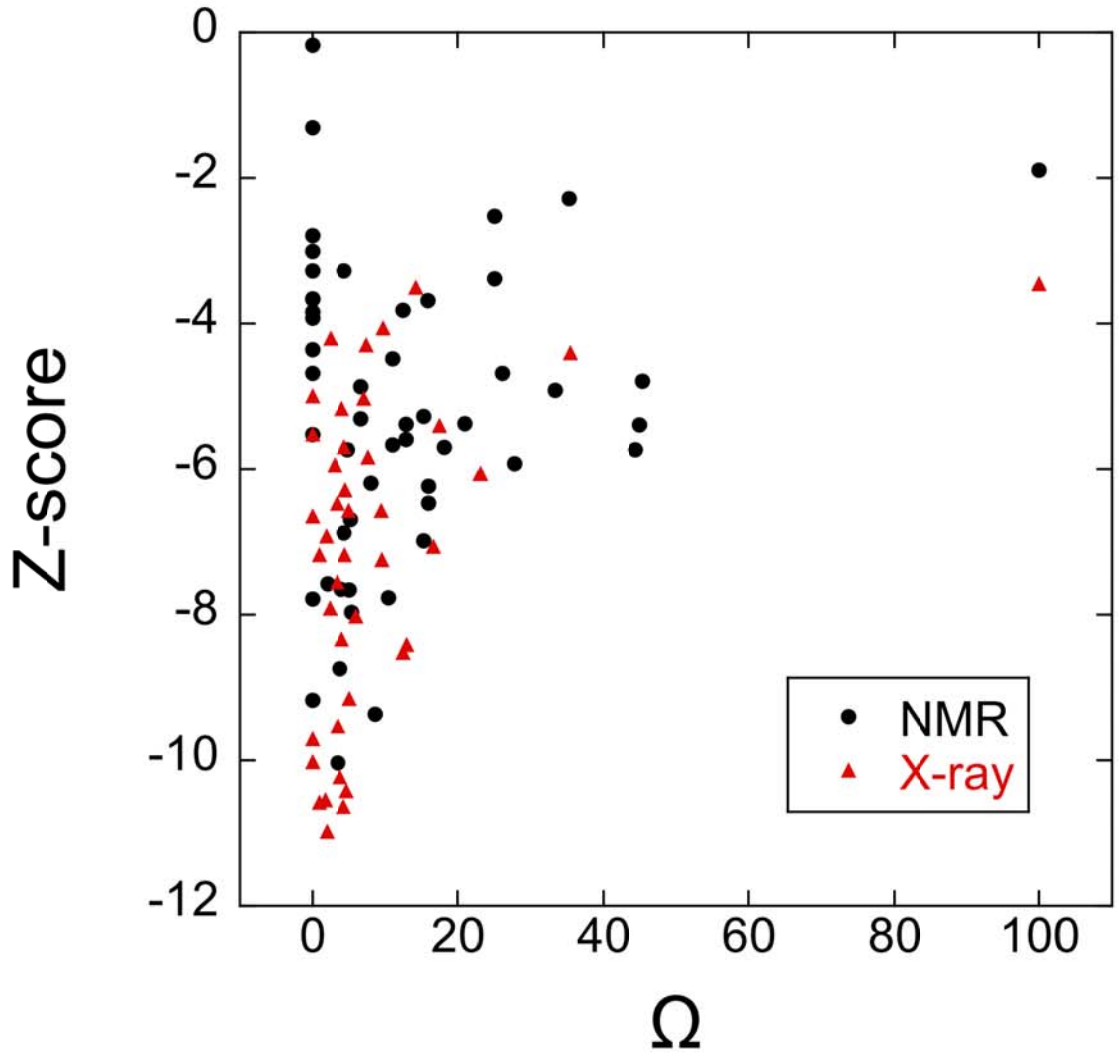


Fig. 5

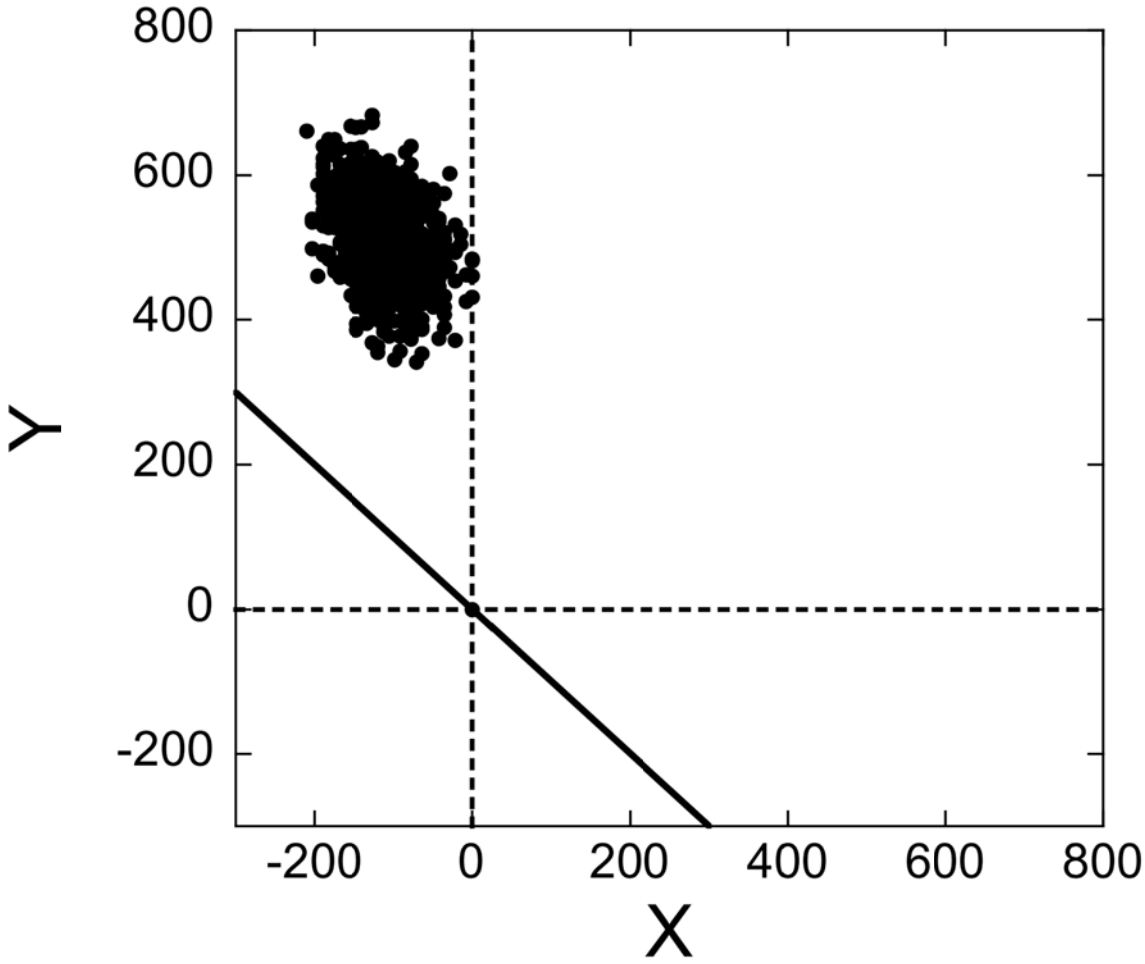


Fig. 6

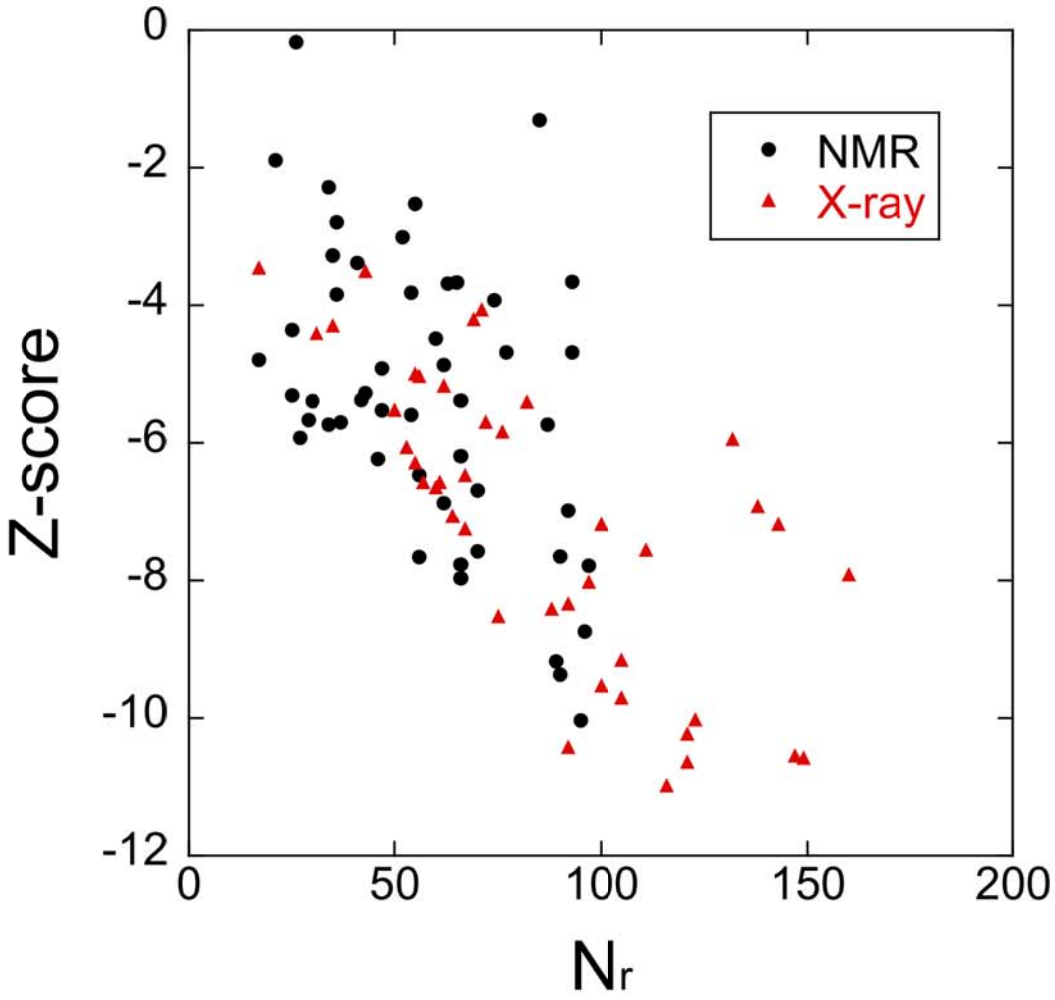


Fig. 7