On Side-Chain Conformational Entropy of Proteins

Jinfeng Zhang, Jun S. Liu*

Department of Statistics, Harvard University, Cambridge, Massachusetts, United States of America

The role of side-chain entropy (SCE) in protein folding has long been speculated about but is still not fully understood. Utilizing a newly developed Monte Carlo method, we conducted a systematic investigation of how the SCE relates to the size of the protein and how it differs among a protein's X-ray, NMR, and decoy structures. We estimated the SCE for a set of 675 nonhomologous proteins, and observed that there is a significant SCE for both exposed and buried residues for all these proteins—the contribution of buried residues approaches ∼40% of the overall SCE. Furthermore, the SCE can be quite different for structures with similar compactness or even similar conformations. As a striking example, we found that proteins' X-ray structures appear to pack more “cleverly” than their NMR or decoy counterparts in the sense of retaining higher SCE while achieving comparable compactness, which suggests that the SCE plays an important role in favouring native protein structures. By including a SCE term in a simple free energy function, we can significantly improve the discrimination of native protein structures from decoys.

Introduction

Side-chains of amino-acid residues encode the information governing a protein's three-dimensional fold. In a typical X-ray crystal structure, each residue's side-chain is represented by a fixed configuration, and most side-chain modelling methods assume that each buried side-chain takes only one fixed conformation among all possible rotameric states (rotamers) [1–4]. Recent studies, however, have shown that many different self-avoiding side-chain packing (called the side-chain conformation of a backbone structure henceforth) may exist for a given native backbone structure [5–7]. It is also well-recognized that the so-called “native protein structure” is an ensemble of structures instead of a single structure as normally seen from X-ray crystallography [8–11]. Ensemble properties of a protein are thus important for characterizing its structure and function.

Estimating ensemble properties such as entropy or free energy has been a long-standing difficult task in structure modelling and simulations [12,13]. In general, side-chain entropy (SCE) can be divided into the vibrational and the conformational [12]. Studies have shown that vibrational entropy is invariant in folded and unfolded states [14]. Therefore, most studies including ours focus on the estimation of conformational SCE [12]. Throughout this article, the term “SCE” actually refers to the conformational. Because of computational limitations, most of our current understanding of SCE is based on an aggregation of entropic effects such as rotamer counts of individual amino-acid residues [4,15–17], which has been shown to significantly overestimate the true SCE [18,19]. With the aid of a new Monte Carlo method, we can now accurately estimate the SCE of proteins based on a realistic model with all heavy atoms explicitly represented.

Results

A Large-Scale Analysis of Side-Chain Conformational Entropy

We computed SCE for a set of 675 nonhomologous proteins obtained from the PISCES database [20]. These proteins are selected under requirements that they have no missing residues; their structural resolutions are better than 1.6 Å; and no pairs have more than 20% sequence identity. The largest protein in the set has 839 residues. Figure 1A plots the SCE of proteins versus their chain lengths, showing that the SCE increases nearly linearly with the chain length. It also demonstrates that the SCE computation is insensitive to the use of two different scales for atom radius (see Methods). Furthermore, we estimated each individual residue's marginal SCE based on our weighted Monte Carlo samples, and observed that the fraction of SCE contributed by all the buried residues (defined as the one with less than 25% of its surface area accessible to solvent) of a protein approaches 40%–50% as the chain's length grows (Figure 1B).

Side-Chain Entropy of the Native and Decoy Structures of Proteins

We considered all the 24 distinct monomeric proteins in five decoy sets (e.g., 4state__reduced, fisa, fisa_casp3, lattice__ssfit, and lmds) of the Decoys ‘R’ Us database [21], in which each protein has a few hundred to ∼2,000 decoy structures and its native structure has been solved by X-ray crystallography. All decoy structures have been minimized using some physical force fields to reach a local energy minimum. Most of the decoy structures have large RMSD to the corresponding native structure (>3 Å).

We plot SCE (S_v) of the native and decoy structures of protein lecf against the corresponding radii of gyration (R_g) in Figure 2A, and against the number of residue contacts (N_c) in Figure 2B.
Synopsis

Side-chains of amino acids determine a protein’s three-dimensional structure. The flexible nature of side-chains introduces a significant amount of conformational entropy associated with both protein folding and interactions. Despite many studies, the role that this side-chain entropy (SCE) plays in the process of folding and interactions has not been fully understood. Some basic questions about SCE have not been systematically studied. In this study, Zhang and Liu developed an efficient sequential Monte Carlo strategy to accurately estimate the SCE of proteins of arbitrary lengths with a given potential energy function. Using this novel tool, they studied how the SCE scales with the length of the protein, and how the SCE differs among a protein’s X-ray, NMR, and decoy structures. They observed that X-ray structures pack more “smartly” than the corresponding decoy and NMR structures: with the same compactness, X-ray structures tend to have larger SCE. A combination of an SCE term with a contact potential energy significantly improved the discrimination between native and decoy structures. The implication of this study is that the SCE contributes so significantly to protein stability that it should be included explicitly in tasks such as structure prediction, protein design, and NMR structure refinement.

The SCE difference between X-ray and NMR structures, $\Delta S_{XX} = S_{c,x-ray} - S_{c,NMR}$, displays a much different behaviour. As shown in Figure 3B, magnitudes of $\Delta S_{XX}$ between proteins’ X-ray structure and their corresponding multiple NMR structures are much larger than $|\Delta S_N|$’s (2 versus 8 kBT unit on average). Although each chosen X-ray structure is very similar to its corresponding NMR structures with small RMSD [25], X-ray structures generally have higher SCE than the corresponding NMR structures. To see how this is related to their packing, we show in Figure 4 the average $\Delta S_{XX}$ of a protein versus $\Delta R_g = R_{g,x-ray} - R_{g,NMR}$, the average difference of the radius of gyration of backbone atoms between X-ray and NMR structures, for all the 24 proteins. Clearly, X-ray structures have comparable $R_g$ to the corresponding NMR structures. For many proteins (“+” in Figure 4), their X-ray structures have much higher SCE than the corresponding NMR structures with similar $R_g$. Some proteins’ X-ray structures (“−”) gain considerable SCE by packing a little looser. Two X-ray structures (“•”) pack tighter than NMR structures but with comparable SCE. Small proteins (“−”) tend to have small $\Delta R_g$ and $\Delta S_{XX}$, while large proteins tend to have large $\Delta S_{XX}$ (see also Figure 5). This is expected since NMR experiments tend to be more accurate for small proteins.

Incorporation of Side-Chain Entropy in Free Energy Functions

Since native structures tend to have higher SCE than computer-generated decoys at the same level of compactness, incorporating SCE into free energy functions should improve modelling accuracy. We tested this idea on all 24 distinct proteins and their decoys in the Decoys ‘R’ Us database. We use a statistical contact potential [26] based on the pairwise distances of $G_{bb}$ atoms, which can be easily computed from a protein’s backbone structure. Following the equation of Gibbs free energy, the free energy of ensemble structures represented by a backbone structure is defined as: $G_{bb} = H_{bb} - TS_{bb}$, where $H_{bb}$ is the potential energy defined by the backbone conformation, $S_{bb}$ is the side-chain entropy, and $T$ is the temperature. Since we use here a statistical potential, temperature $T$ has no physical meaning and can be freely adjusted. We set $T$ to 1 in this study without optimization. We use the rank of the native structure among all the decoys to evaluate the discrimination performance. Table 1 shows that for most proteins, the measures based on free energy $G_{bb}$ significantly improved the discrimination power compared with those using potential energy $H_{bb}$. For a few proteins, the discrimination performances under $G_{bb}$ and $H_{bb}$ are comparable, and in only one case $G_{bb}$ performed slightly worse than $H_{bb}$. It is possible that some proteins are stabilized mainly by enthalpy and other entropic terms instead of by side-chain conformational entropy. For example, the energy of a couple of disulfide bonds may be enough to stabilize a small protein so that other factors become insignificant.

We note here that an all-atom potential function, which differentiates different side-chain conformations, can also be accommodated by our Monte Carlo method. In particular, free energy $G_{bb}$ can be estimated using the formula $G_{bb} = -k_BT \ln Q_{bb}$, where $Q_{bb}$ is the partition function of the ensemble side-chain conformations of a backbone structure, which can be estimated by our Monte Carlo method.

The SCE difference between X-ray and NMR structures, $\Delta S_{XX} = S_{c,x-ray} - S_{c,NMR}$, displays a much different behaviour. As shown in Figure 3B, magnitudes of $\Delta S_{XX}$ between proteins’ X-ray structure and their corresponding multiple NMR structures are much larger than $|\Delta S_N|$’s (2 versus 8 kBT unit on average). Although each chosen X-ray structure is very similar to its corresponding NMR structures with small RMSD [25], X-ray structures generally have higher SCE than the corresponding NMR structures. To see how this is related to their packing, we show in Figure 4 the average $\Delta S_{XX}$ of a protein versus $\Delta R_g = R_{g,x-ray} - R_{g,NMR}$, the average difference of the radius of gyration of backbone atoms between X-ray and NMR structures, for all the 23 proteins. Clearly, X-ray structures have comparable $R_g$ to the corresponding NMR structures. For many proteins (“+” in Figure 4), their X-ray structures have much higher SCE than the corresponding NMR structures with similar $R_g$. Some proteins’ X-ray structures (“−”) gain considerable SCE by packing a little looser. Two X-ray structures (“•”) pack tighter than NMR structures but with comparable SCE. Small proteins (“−”) tend to have small $\Delta R_g$ and $\Delta S_{XX}$, while large proteins tend to have large $\Delta S_{XX}$ (see also Figure 5). This is expected since NMR experiments tend to be more accurate for small proteins.

Incorporation of Side-Chain Entropy in Free Energy Functions

Since native structures tend to have higher SCE than computer-generated decoys at the same level of compactness, incorporating SCE into free energy functions should improve modelling accuracy. We tested this idea on all 24 distinct proteins and their decoys in the Decoys ‘R’ Us database. We use a statistical contact potential [26] based on the pairwise distances of $G_{bb}$ atoms, which can be easily computed from a protein’s backbone structure. Following the equation of Gibbs free energy, the free energy of ensemble structures represented by a backbone structure is defined as: $G_{bb} = H_{bb} - TS_{bb}$, where $H_{bb}$ is the potential energy defined by the backbone conformation, $S_{bb}$ is the side-chain entropy, and $T$ is the temperature. Since we use here a statistical potential, temperature $T$ has no physical meaning and can be freely adjusted. We set $T$ to 1 in this study without optimization. We use the rank of the native structure among all the decoys to evaluate the discrimination performance. Table 1 shows that for most proteins, the measures based on free energy $G_{bb}$ significantly improved the discrimination power compared with those using potential energy $H_{bb}$. For a few proteins, the discrimination performances under $G_{bb}$ and $H_{bb}$ are comparable, and in only one case $G_{bb}$ performed slightly worse than $H_{bb}$. It is possible that some proteins are stabilized mainly by enthalpy and other entropic terms instead of by side-chain conformational entropy. For example, the energy of a couple of disulfide bonds may be enough to stabilize a small protein so that other factors become insignificant.

We note here that an all-atom potential function, which differentiates different side-chain conformations, can also be accommodated by our Monte Carlo method. In particular, free energy $G_{bb}$ can be estimated using the formula $G_{bb} = -k_BT \ln Q_{bb}$, where $Q_{bb}$ is the partition function of the ensemble side-chain conformations of a backbone structure, which can be estimated by our Monte Carlo method.

The SCE difference between X-ray and NMR structures, $\Delta S_{XX} = S_{c,x-ray} - S_{c,NMR}$, displays a much different behaviour. As shown in Figure 3B, magnitudes of $\Delta S_{XX}$ between proteins’ X-ray structure and their corresponding multiple NMR structures are much larger than $|\Delta S_N|$’s (2 versus 8 kBT unit on average). Although each chosen X-ray structure is very similar to its corresponding NMR structures with small RMSD [25], X-ray structures generally have higher SCE than the corresponding NMR structures. To see how this is related to their packing, we show in Figure 4 the average $\Delta S_{XX}$ of a protein versus $\Delta R_g = R_{g,x-ray} - R_{g,NMR}$, the average difference of the radius of gyration of backbone atoms between X-ray and NMR structures, for all the 23 proteins. Clearly, X-ray structures have comparable $R_g$ to the corresponding NMR structures. For many proteins (“+” in Figure 4), their X-ray structures have much higher SCE than the corresponding NMR structures with similar $R_g$. Some proteins’ X-ray structures (“−”) gain considerable SCE by packing a little looser. Two X-ray structures (“•”) pack tighter than NMR structures but with comparable SCE. Small proteins (“−”) tend to have small $\Delta R_g$ and $\Delta S_{XX}$, while large proteins tend to have large $\Delta S_{XX}$ (see also Figure 5). This is expected since NMR experiments tend to be more accurate for small proteins.

Incorporation of Side-Chain Entropy in Free Energy Functions

Since native structures tend to have higher SCE than computer-generated decoys at the same level of compactness, incorporating SCE into free energy functions should improve modelling accuracy. We tested this idea on all 24 distinct proteins and their decoys in the Decoys ‘R’ Us database. We use a statistical contact potential [26] based on the pairwise distances of $G_{bb}$ atoms, which can be easily computed from a protein’s backbone structure. Following the equation of Gibbs free energy, the free energy of ensemble structures represented by a backbone structure is defined as: $G_{bb} = H_{bb} - TS_{bb}$, where $H_{bb}$ is the potential energy defined by the backbone conformation, $S_{bb}$ is the side-chain entropy, and $T$ is the temperature. Since we use here a statistical potential, temperature $T$ has no physical meaning and can be freely adjusted. We set $T$ to 1 in this study without optimization. We use the rank of the native structure among all the decoys to evaluate the discrimination performance. Table 1 shows that for most proteins, the measures based on free energy $G_{bb}$ significantly improved the discrimination power compared with those using potential energy $H_{bb}$. For a few proteins, the discrimination performances under $G_{bb}$ and $H_{bb}$ are comparable, and in only one case $G_{bb}$ performed slightly worse than $H_{bb}$. It is possible that some proteins are stabilized mainly by enthalpy and other entropic terms instead of by side-chain conformational entropy. For example, the energy of a couple of disulfide bonds may be enough to stabilize a small protein so that other factors become insignificant.

We note here that an all-atom potential function, which differentiates different side-chain conformations, can also be accommodated by our Monte Carlo method. In particular, free energy $G_{bb}$ can be estimated using the formula $G_{bb} = -k_BT \ln Q_{bb}$, where $Q_{bb}$ is the partition function of the ensemble side-chain conformations of a backbone structure, which can be estimated by our Monte Carlo method.
Figure 1. Side-Chain Entropy (SCE), $S_{sc}$, of 675 Nonhomologous Proteins in the PDB

(A) Side-chain entropy versus chain length. Two results with $\alpha = 0.6$ (red crosses) and 0.8 (green circles) are shown.
(B) Percentage of SCE contributed by buried residues versus chain length.

Discussion

In this study, we systematically investigated the SCE of a large set of protein structures and its difference among X-ray, NMR, and decoy structures. Our findings do not contradict the traditional view that SCE is an opposing factor for protein folding from extended states to compact native states, but our findings on the systematic difference of the SCE among the folded conformations with similar compactness suggest that the SCE plays an important role in protein stability and should be included in tasks such as protein structure prediction, protein design, and NMR structure refinement.

The Nuts-and-Bolts model states that proteins pack quite randomly, thus giving rise to many internal voids [19,22,27,28]. In contrast, the Jigsaw-Puzzle model alleges that proteins pack like a jigsaw puzzle with side-chains closely interlocked [29–33]. It is conceivable that side-chain packing in protein cores is not completely random, as some regularities and specific residue interactions have been observed [32,33]. However, such specific interactions are sparse among all interacting residue pairs [34]. Our observation that buried residues of a protein contribute significantly to its overall SCE suggests that the interior of a protein’s native structure is unlikely to pack in a jigsaw-puzzle mode. However, we also found that the SCE of individual buried residues vary greatly, with some having comparable entropy to exposed ones while others have almost zero entropy, which is consistent with observed local packing in proteins [35]. This indicates that the packing of the protein core is likely heterogeneous, with parts forming a jigsaw puzzle to gain specificity and other parts resembling nuts and bolts to maintain entropy and gain robustness against mutations.

Structures solved by X-ray crystallography are generally more reliable than the corresponding NMR structures, which lack the quality measurement for solved structures. It has been found that NMR structures tend to pack poorly [36]. Such poor packing is mainly due to the nature of experimental data and computational methods employed instead of a reflection of the difference between the solution and crystal states. Indeed, experimental NMR observables agree better with structures calculated from high-resolution crystals than those from the corresponding NMR structures [37]. Our findings suggest that the SCE difference found between X-ray and NMR structures may account for some of the poor packing of NMR structures, and thus, incorporating SCE in the energy functions used in computational methods of NMR experiments, may improve the quality of NMR structures.

Both decoy and NMR structures were obtained by structural optimization under some potential functions. The backbone conformational entropy has been suggested as a stabilizing factor for native proteins. [38] Observations made in this study indicate that ignoring SCE by those optimization techniques produces significant deviations from characteristic packing and interaction of native proteins, which suggest that atom-level modelling of protein structures and interactions should take approaches with more emphasis on ensemble sampling rather than on optimization. Our preliminary study on the incorporation of SCE in an empirical free energy function shows a significant improvement in discrimination of native structure against decoys.

We used in SCE estimation a very simplified energy function, which focuses only on the excluded volume effect. It is somewhat surprising to us that, just with excluded volume effect, the SCE can already differentiate well native X-ray structures from NMR and decoy ones. We also experimented with another energy function considering rotamer probabilities, which reduces the SCE by 10% on average, and observed that the results reported here hold well. It remains to be seen how the reported results will be affected when a more realistic potential energy function is
Proteins 1btv, 1vre, and 1ah2, ranges of average RMSDs of the X-ray and NMR structure pairs. For NMR structures (B) Box plot for distributions of the SCE difference between X-ray and indicate different ranges of average RMSDs of the structure pairs. As shown in Figure 2, we believe that an energy function thus producing many false positives for this type of complex.

Accuracy of this and any other type of protein complex in incorporating an SCE term should improve the prediction of protein–protein complexes based on CAPRI [40] results, which indicated that one type of difficult prediction difficulties of protein–protein complexes based on the discrete nature of side-chain rotamers. A recent study has assessed roles in protein interactions. A recent study has assessed suggest that conformational entropy may play important

The compactness measurement $N_c$ is defined as number of pairwise $C_B$ (or $C_O$ of Glycine) atoms with their distance of less than 7.5 Å.

Side-chain modelling. Each residue’s side-chain conformation is modelled as a rotamer with a finite number of discrete states [42]. The rotamer library used is developed by Lovell et al. [43], as recommended by Dunbrack [42] for the study of entropy. The rotamer library of Dunbrack and Cohen [44] was also applied to some of the proteins studied here and similar results were observed. To account for the excluded volume effect (or self-avoiding requirement), we took the approach of Russell et al. [6], in which a pair of atoms $i$ and $j$ is considered to be a hard clash if $r_{ij} < \pi \times (r(i) + r(j))$, where $r(i)$ is their distance, $\pi$ is a scaling coefficient to account for the discrete nature of side-chain rotamers, and $r(i)$ and $r(j)$ are the van der Waals radii of the two atoms. We tested three $\pi$ values at 0.6, 0.7, and 0.8, respectively, and found that they gave qualitatively similar results (Figure 2). Lower $\pi$ values give higher entropy and diminish the side-chain entropy differences among different structures, whereas higher $\pi$ values give lower entropy and cause some structures to have no valid self-avoiding side-chain conformations, which were discarded in the analysis. All results on the comparison of X-ray, decoy, and NMR structures were obtained with $\pi$ equal to 0.8 unless otherwise stated.

The SCE is defined as: $S_{sc} = -k_B \sum_i p_i \ln(p_i)$, where $k_B$ is the Boltzmann constant and $p_i = e^{-E_i/k_B T} / \sum_j e^{-E_j/k_B T}$ is the probability of a self-avoiding side-chain conformation. When the $p_i$’s are all equal or $T$ is very high, we have $S = k_B \ln(n)$, where $n$ is the number of self-avoiding side-chain conformations for the given backbone structure. The compactness measurement $N_c$ is defined as number of pairwise $C_B$ (or $C_O$ of Glycine) atoms with their distance of less than 7.5 Å.

Sequential Monte Carlo method. The Sequential Monte Carlo method (SMC) is a generalization of the Rosenbluth’s chain growth method [15] and has been applied previously in studying problems ranging from protein-packing behaviour, effect of amino acid chirality, side-chain flexibility, protein folding, and near-native structures of proteins [19,22,46–48]. In this work, we made two design modifications to further improve the SMC’s efficiency: (a) we used. For example, if a Van der Waals interaction term is to be added, the discrete rotamer formulation adopted in this article’s research has to be adequately refined so as to accommodate the continuous nature of the protein side-chain positions. Otherwise, the SCE could be seriously distorted when a few atoms are not placed very well due to the discrete nature of side-chain rotamers.

Interfacial regions in protein–protein complexes have been shown to be less flexible than other parts of the protein surface [23]. It has also been suggested that conserved polar residues at the binding interfaces have higher rigidity so that the entropic cost is minimized on binding, whereas surrounding residues form a flexible cushion [39]. These studies suggest that conformational entropy may play important roles in protein interactions. A recent study has assessed prediction difficulties of protein–protein complexes based on CAPRI [40] results, which indicated that one type of difficult complex has a small interface area and a weak binding energy [41]. Existing computational docking algorithms typically favor interaction conformations with large interface areas, thus producing many false positives for this type of complex. As shown in Figure 2, we believe that an energy function incorporating an SCE term should improve the prediction accuracy of this and any other type of protein complex in

![Figure 3. SCE of NMR and X-Ray Structures](image)

(A) Box plot for distributions of the absolute pairwise SCE difference ($\Delta S_{sc}$) of NMR structures of 23 proteins. Different coloured boxes indicate different ranges of average RMSDs of the structure pairs. (B) Box plot for distributions of the SCE difference between X-ray and NMR structures ($\Delta S_{sc}$) for 23 proteins. Different colours indicate different ranges of average RMSDs of the X-ray and NMR structure pairs. For proteins 1btv, 1vre, and 1ah2, $\pi = 0.7$ was used for both X-ray and NMR structures.

doi:10.1371/journal.pcbi.0020168.g003

doi:10.1371/journal.pcbi.0020168.g004

Figure 4. SCE of NMR and X-Ray Structures versus $R_g$

Average SCE difference between X-ray and NMR structures ($\Delta S_{sc}$) versus the average difference of radius of gyration between X-ray and NMR backbones ($\Delta R_g$) for the 23 proteins. $\times$, proteins whose X-ray structures have much higher SCE than but similar $R_g$ to the corresponding NMR structures. $\triangle$, proteins whose X-ray structures gain considerable SCE by packing a little looser. $\circ$, proteins whose X-ray structures pack tighter than NMR structures but with comparable SCE. $\times$, small proteins of which both $\Delta R_g$ and $\Delta S_{sc}$ are small.

doi:10.1371/journal.pcbi.0020168.g004

which SCE contributes significantly in the binding free energy.

Materials and Methods

Side-chain modelling. Each residue’s side-chain conformation is modelled as a rotamer with a finite number of discrete states [42]. The rotamer library used is developed by Lovell et al. [43], as recommended by Dunbrack [42] for the study of entropy. The rotamer library of Dunbrack and Cohen [44] was also applied to some of the proteins studied here and similar results were observed. To account for the excluded volume effect (or self-avoiding requirement), we took the approach of Russell et al. [6], in which a pair of atoms $i$ and $j$ is considered to be a hard clash if $r_{ij} < \pi \times (r(i) + r(j))$, where $r(i)$ is their distance, $\pi$ is a scaling coefficient to account for the discrete nature of side-chain rotamers, and $r(i)$ and $r(j)$ are the van der Waals radii of the two atoms. We tested three $\pi$ values at 0.6, 0.7, and 0.8, respectively, and found that they gave qualitatively similar results (Figure 2). Lower $\pi$ values give higher entropy and diminish the side-chain entropy differences among different structures, whereas higher $\pi$ values give lower entropy and cause some structures to have no valid self-avoiding side-chain conformations, which were discarded in the analysis. All results on the comparison of X-ray, decoy, and NMR structures were obtained with $\pi$ equal to 0.8 unless otherwise stated.

The SCE is defined as: $S_{sc} = -k_B \sum_i p_i \ln(p_i)$, where $k_B$ is the Boltzmann constant and $p_i = e^{-E_i/k_B T} / \sum_j e^{-E_j/k_B T}$ is the probability of a self-avoiding side-chain conformation. When the $p_i$’s are all equal or $T$ is very high, we have $S = k_B \ln(n)$, where $n$ is the number of self-avoiding side-chain conformations for the given backbone structure. The compactness measurement $N_c$ is defined as number of pairwise $C_B$ (or $C_O$ of Glycine) atoms with their distance of less than 7.5 Å.

Sequential Monte Carlo method. The Sequential Monte Carlo method (SMC) is a generalization of the Rosenbluth’s chain growth method [15] and has been applied previously in studying problems ranging from protein-packing behaviour, effect of amino acid chirality, side-chain flexibility, protein folding, and near-native structures of proteins [19,22,46–48]. In this work, we made two design modifications to further improve the SMC’s efficiency: (a) we...
use of a recently developed stratified resampling technique [19,47], and (b) we take advantage of the fact that the sampling order of each residue’s conformation can be arranged arbitrarily. A brief description of the method is given below. More details about the general method can be found in [19,22,46,48].

Table 1. Discrimination of Native Structures Using a Free Energy Function

| Protein ID | Rank by $H_{bb}$ | Rank by $G_{bb}$ | Protein ID | Rank by $H_{bb}$ | Rank by $G_{bb}$ |
|------------|------------------|------------------|------------|------------------|------------------|
| 1ctf (A)   | 6                | 1                | 1beo (D)   | 67               | 2                |
| 1r69 (A)   | 24               | 5                | 1ctf (D)   | 10               | 1                |
| 1sn3 (A)   | 86               | 10               | 1dkt-A (D) | 588              | 5                |
| 2cro (A)   | 63               | 5                | 1fc-a (D)  | 136              | 10               |
| 3icb (A)   | 19               | 25               | 1nkl (D)   | 217              | 3                |
| 4pti (A)   | 143              | 83               | 1pgb (D)   | 12               | 1                |
| 4nnn (A)   | 14               | 7                | 1bun-b (E) | 114              | 104              |
| 1fc2 (B)   | 7                | 5                | 1ctf (E)   | 13               | 4                |
| 1hind-C (B)| 10               | 5                | 1dk (E)    | 1                | 1                |
| 2cro (B)   | 47               | 17               | 1fc2 (E)   | 32               | 3                |
| 4icb (B)   | 1                | 1                | 1gld (E)   | 159              | 6                |
| 1b10 (C)   | 851              | 4                | 1shf-A (E) | 1                | 1                |
| 1eh2 (C)   | 995              | 3                | 2cro (E)   | 19               | 2                |
| 1jwe (C)   | 288              | 1                | 2ovo (E)   | 1                | 1                |
| 3icd (C)   | 266              | 1                | 4pti (E)   | 1                | 1                |

Ranks are produced according to the energy value of the native structure relative to those of its decays (the smaller the better). For most of the proteins, the scaling factor $\alpha = 0.7$ was used for both native and their decay structures. In the case that more than half of the structures in a decay set do not have a self-avoiding side-chain conformation, $\alpha = 0.6$ was used for both the native structure and its decays.

(a) Letters in parentheses represent the particular decay set in the database, where A stands for ʻstate_reduced, B for ʻfisa, C for ʻfisa_casp3, D for ʻlattice_sifit, and E for ʻlmds.

(b) Protein 1bba in this decay set is an NMR structure and thus excluded from this study.

doi:10.1371/journal.pcbi.0020168.t001

Figure 5. Performance of the Sequential Monte Carlo Method

(A) Comparison of the SMC estimation with exhaustive enumeration for fragments of proteins 2ovo and 3ebx.

(B) Standard deviation of the SMC estimation for four different sample sizes, 100, 500, 1,000, and 2,000, respectively, calculated from 20 independent SMC runs. The first number in each parentheses pair is the number of residues of the protein, and the second number the average SCE of 20 runs with 1,000 samples in each run.

doi:10.1371/journal.pcbi.0020168.g005
with the exact answers. As seen from Figure 5A, the estimates using SMc are indistinguishable from those obtained by exhaustive enumeration. For example, the total number of self-avoiding side-chain conformations for the fragment of 3exb, residue 1–17, is 396,325,923,840, and our SMc estimate is 4.01 × 10^{11} with the Monte Carlo sample size M = 1,000. Figure 5B shows the standard deviations of these estimates against the sample size M used by SMc. We found that a single run of SMc with M = 1,000 is enough to give accurate estimates of the SCE for all the proteins we studied. The running time of SMc with M = 1,000 samples and α = 0.6, on a Linux machine with a CPU of 1.4 GHz, was 3.1 s for protein 4rnt (104 residues); 6.4 s for protein 1svn (269 residues); and 81 seconds for protein 1epw (1,287 residues), the longest protein we have tried.

Supporting Information

Protocol S1. Side-Chain Entropy and Packing of Native and Decoy Structures

Found at doi:10.1371/journal.pcbi.0020168.s001 (3.3 MB PDF).

Accession Numbers

PDB names of NMR structures of the 23 proteins in Figure 3A are (with PDB names of X-ray structures and protein lengths in parentheses): 1erc (2erl, 40), 11tr (2ovo, 56), 1ff2 (1ffx, 58), 1fra (3exb, 62), 1r63 (1r69, 63), 3mef (1mjc, 69), 2abd (1hb6, 86), 1af (1icb, 214), 1bmw (1who, 94), 1fra (3ebx, 62), 1ah2 (2erl, 159), 1xoa (2tir, 108), 1ycp (1hka, 117–121).

Acknowledgments

We thank Dr. Rong Chen and Dr. Jie Liang for helpful discussions. Author contributions. JZ and JL conceived and designed the experiments and wrote the paper. JZ performed the experiments, analyzed the data, and contributed reagents/materials/analysis tools.

Funding. This research was supported in part by US National Science Foundation grants DMS-0244676 and DMS-0244658.

Competing interests. The authors have declared that no competing interests exist.